Introducción al

Razonamiento Aproximado

F. J. Díez
Dpto. Inteligencia Artificial
UNED

Primera edición: Octubre 1998
Revisión: Noviembre 2005

A mi hija Nuria

Prefacio

Esta *Introducción al Razonamiento Aproximado* está destinada principalmente a los estudiantes de *Razonamiento y Aprendizaje*, asignatura optativa del tercer curso de la Ingeniería Técnica de Informática de Sistemas, de la UNED. La redacción inicial se basó en la tesis doctoral del autor y en las transparencias del curso de doctorado *Razonamiento Aproximado*. El capítulo 4, especialmente en lo relativo a la teoría de la confimación (secs. 4.1.2 y 4.4.1), se basó también en un documento redactado para esta obra por Enrique Nell, quien ha aportado además referencias bibliográficas y comentarios muy acertados.

La primera edición apareció en octubre de 1998. Desde entonces cada año hemos publicado una nueva versión corregida y a veces aumentada. En esta labor hemos contado con la ayuda de Carlos Cabezas, Ildefonso Bellón, Flavio Cuéllar, Ismael Falcón, José Antonio Fernández, José Melgar, Eva Millán, Enrique Nell, David Penas, Lourdes Pérez, José Rabaneda, Montserrat Sans, Óscar Sanz y Xavier Torres, quienes nos han señalado un buen número de erratas.

El autor y los lectores de futuras ediciones, especialmente los alumnos que tendrán que esforzarse por comprender y aprender su contenido, agradecen sinceramente todas las correcciones y sugerencias recibidas hasta la fecha y las que se reciban en el futuro: erratas detectadas, puntos que no están claros, omisiones importantes, cuestiones que conviene matizar, o incluso errores conceptuales, que es posible que los haya. Todos los comentarios serán bien recibidos.

En la página de Internet `http://www.ia.uned.es/~fjdiez/libros/razaprox.html` pondremos información actualizada sobre este texto (fe de erratas, versiones actualizadas, etc.), material complementario y enlaces de interés.

Francisco Javier Díez Vegas
UNED, Madrid, noviembre de 2005

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Razonamiento aproximado en Inteligencia Artificial

El tratamiento de la incertidumbre constituye uno de los campos fundamentales de la inteligencia artificial, pues afecta en mayor o menor medida a todos los demás. En particular, una de las propiedades esenciales de los sistemas expertos, y a la vez una de las más complejas, es el tratamiento de la incertidumbre. En este capítulo enumeramos y clasificamos las fuentes de incertidumbre habituales, tomando como ejemplo el campo de la medicina, con el fin de mostrar la importancia del tema. Hacemos también un breve resumen de la evolución histórica del razonamiento incierto que, como comentaremos más adelante, cuando se realiza mediante métodos numéricos, suele denominarse **razonamiento aproximado**.

1.1 Fuentes de incertidumbre

Observando la historia de los sistemas expertos, y en particular de los métodos de razonamiento incierto, se comprueba que casi todos los primeros (cronológicamente) y muchos de los más importantes, se han desarrollado en el campo de la **medicina**. Si tratamos de averiguar el porqué, descubrimos que éste es un campo donde se dan todos los tipos de incertidumbre. A grandes rasgos, podemos clasificar las fuentes de incertidumbre en tres grupos:

—deficiencias de la información,
—características del mundo real y
—deficiencias del modelo.

Veamos algunos ejemplos:

- *Información incompleta.* En muchos casos la historia clínica completa no está disponible, y el paciente es incapaz de recordar todos los síntomas que ha experimentado y cómo se ha desarrollado la enfermedad. Además, en otras ocasiones, las limitaciones prácticas impiden contar con todos los medios que deberían estar disponibles, por lo que el médico debe realizar su diagnóstico con la información que posee, aunque ésta sea muy limitada.

- *Información errónea.* En cuanto a la información suministrada por el paciente, puede que éste describa incorrectamente sus síntomas e incluso que trate de mentir deliberadamente al médico. También es posible que el diagnóstico anterior, contenido en la historia clínica, haya sido erróneo. Y tampoco es extraño que las pruebas de laboratorio den

falsos positivos y falsos negativos. Por estas razones, el médico siempre debe mantener una duda razonable frente toda la información disponible.

- *Información imprecisa.* Hay muchos datos en medicina que son difícilmente cuantificables. Tal es el caso, por ejemplo, de los síntomas como el dolor o la fatiga. Incluso en un método tan técnico como la ecocardiografía hay muchas observaciones que en la práctica deben ser cuantificadas subjetivamente, como son el prolapso valvular ("caída" o desplazamiento excesivo de una válvula al cerrarse) o la aquinesia ventricular (falta de movimiento de un ventrículo).

- *Mundo real no determinista.* A diferencia de las máquinas mecánicas o eléctricas, cuyo funcionamiento se rige por leyes deterministas, los profesionales de la medicina comprueban a diario que cada ser humano es un mundo, en que las leyes generales no siempre resultan aplicables. Muchas veces las mismas causas producen efectos diferentes en distintas personas, sin que haya ninguna explicación aparente. Por ello, el diagnóstico médico debe estar siempre abierto a admitir la aleatoriedad y las excepciones.

- *Modelo incompleto.* Por un lado, hay muchos fenómenos médicos cuya causa aún se desconoce. Por otro, es frecuente la falta de acuerdo entre los expertos de un mismo campo. Finalmente, aunque toda esta información estuviera disponible, sería imposible, por motivos prácticos, incluirla en un sistema experto.

- *Modelo inexacto.* Por último, todo modelo que trate de cuantificar la incertidumbre, por cualquiera de los métodos que existen, necesita incluir un elevado número de parámetros; por ejemplo, en el caso de las redes bayesianas, necesitamos especificar todas las probabilidades a priori y condicionales. Sin embargo, una gran parte de esta información no suele estar disponible, por lo que debe ser estimada de forma subjetiva. Es deseable, por tanto, que el método de razonamiento empleado pueda tener en cuenta las inexactitudes del modelo.

Hemos escogido el campo de la medicina como ejemplo paradigmático de dominio incierto, aunque todas estas fuentes de incertidumbre pueden darse, y de hecho se dan, en cualquier otro campo de las ciencias naturales, la ingeniería, el derecho, las humanidades... y muy especialmente en los problemas de reconocimiento del lenguaje natural, tanto hablado como escrito, donde la información implícita, la polisemia, la ambigüedad y la imprecisión, hacen imprescindible el tratamiento de la incertidumbre. En realidad, ésta es una necesidad que no sólo incumbe a los sistemas expertos y a los problemas de lenguaje natural, sino a todas las ramas de la inteligencia artificial, como el aprendizaje, la visión artificial, la robótica, los interfaces inteligentes, la recuperación de información, los juegos complejos (no sólo los juegos de azar, sino también juegos como el ajedrez, donde no se conocen con certeza las preferencias del contrario), etc., etc.

En resumen, el tratamiento de la incertidumbre es, junto con la representación del conocimiento y el aprendizaje, uno de las problemas fundamentales de la inteligencia artificial. Por ello no es extraño que casi desde los orígenes de este campo se le haya prestado tanta atención y hayan surgido tantos métodos, motivados por los distintos problemas que se han ido planteando. Vamos a hablar de ello en la próxima sección.

1.2 Breve historia del tratamiento de la incertidumbre

Los métodos de razonamiento incierto se clasifican en dos grandes grupos: **métodos numéricos** y **métodos cualitativos**. Cuando el razonamiento incierto se realiza mediante métodos numéricos suele hablarse de **razonamiento aproximado** (aunque tampoco es una cuestión en la que haya acuerdo unánime, pues algunos autores, al hablar de "razonamiento aproximado", piensan sobre todo en la lógica difusa y en modelos afines, como la teoría de la posibilidad).

Entre los **métodos cualitativos** para el tratamiento de la incertidumbre, destacan los basados en lógicas no monótonas, tales como los modelos de *razonamiento por defecto* (el más conocido es el de Reiter [52]), los sistemas de *suposiciones razonadas* (originalmente llamados *truth maintenance systems*, aunque sería más correcto denominarlos *reason maintenance systems*) de Doyle [17] y la *teoría de justificaciones* (*theory of endorsements*) de Cohen y Grinberg [8, 9]. Estos métodos consisten en que, cuando no hay información suficiente, se hacen suposiciones, que posteriormente podrán ser corregidas al recibir nueva información. El problema principal que presentan se debe a su naturaleza cualitativa, por lo que no pueden considerar los distintos grados de certeza o incertidumbre de las hipótesis. Suelen presentar además problemas de explosión combinatoria. En consecuencia, se estudian más por su importancia teórica (fundamentación de la inteligencia artificial) que por las aplicaciones prácticas a que puedan dar lugar.

En cuanto a los **métodos numéricos**, que son los que vamos a estudiar en este texto, el primero que surgió fue el tratamiento probabilista. En efecto, ya en el siglo XVIII, Bayes y Laplace propusieron la probabilidad como una medida de la creencia personal hace 200 años. A principios del siglo XX surgen las interpretaciones de la probabilidad como la frecuencia (a largo plazo) asociada a situaciones o experimentos repetibles; en esta línea, destacan especialmente los trabajos estadísticos de Fisher. A principios de los años 30, en cambio, debido sobre todo a los trabajos de L. J. Savage y B. de Finetti, entre otros muchos, se redescubre la probabilidad como medida de la creencia personal.

Unos años más tarde, se inventan las computadoras y poco después surge la inteligencia artificial (suele tomarse como punto de referencia el año 1956, en que se celebró la Conferencia de Darmouth, aunque otros autores sitúan el origen de la inteligencia artificial en 1943, el año en que se publicaron dos trabajos eminentes [39, 53]). En aquella época, los ordenadores habían superado ampliamente la capacidad de cálculo de cualquier ser humano, pero estaban muy lejos del denominado "comportamiento inteligente". Precisamente por eso la inteligencia artificial se centraba en la resolución de problemas simbólicos y se esforzaba en distinguirse de los métodos algorítmicos dedicados sobre todo al cálculo numérico [55]. Ésta es una de las razones por las que inicialmente no se prestó atención al estudio de la probabilidad como rama o al menos como herramienta de la inteligencia artificial.

Sin embargo, al enfrentarse a problemas de diagnóstico médico, era inevitable tener que tratar la incertidumbre, por las razones expuestas en la sección anterior, y en aquellos años la única técnica disponible, aun con todas sus limitaciones, era el método probabilista clásico (a veces llamado *Bayes ingenuo*, o *naïve Bayes*, en inglés); con él se construyeron los primeros sistemas de diagnóstico médico, como veremos en el próximo capítulo, que obtuvieron un éxito razonable en problemas que hoy nos parecen pequeños en tamaño, pero que en aquella época eran imposibles de abordar de ninguna otra forma.

No obstante, el método probabilista clásico presentaba dos inconvenientes principales: el

primero de ellos era la dificultad de obtener las probabilidades condicionales necesarias para construir el modelo. La aplicación del teorema de Bayes "en bruto" requería un número exponencial de parámetros (cf. sec. 2.4), por lo que se hacía necesario introducir hipótesis simplificadoras, que eran básicamente dos: la exclusividad de los diagnósticos y la independencia condicional de los hallazgos. Aún así, el número de parámetros seguía siendo relativamente elevado, sobre todo teniendo en cuenta que raramente había bases de datos a partir de las cuales se pudieran obtener las probabilidades objetivas, por lo que en la mayor parte de los casos se hacía necesario recurrir a estimaciones subjetivas, poco fiables. Además —y éste es el segundo inconveniente grave del modelo— las hipótesis eran poco verosímiles, sobre todo la de independencia condicional, sobre la que se escribieron páginas y páginas en los años 70. Por estos motivos, la mayor parte de los investigadores estaban de acuerdo en que la probabilidad no era un método adecuado para la inteligencia artificial [38].

Por otro lado, el éxito obtenido por el sistema experto DENDRAL, considerado por muchos como el primer sistema experto, mostró las grandes ventajas de la programación mediante reglas (cap. 4). Por ello, los creadores de MYCIN buscaban un método de computación eficiente que pudiera adaptarse al razonamiento mediante encadenamiento de reglas. Los problemas mencionados anteriormente y la incapacidad de los métodos probabilistas para encajar en este esquema llevaron a los responsables del proyecto a desarrollar un método propio, consistente en asignar a cada regla un factor de certeza. Este modelo, aunque inspirado lejanamente en el cálculo de probabilidades, a través de la teoría de la confirmación de Carnap, en la práctica no tenía ninguna relación con la teoría de la probabilidad, ni siquiera con su interpretación subjetiva.

El éxito de MYCIN fue muy grande, pues en un campo tan complejo y tan incierto como el de las enfermedades infecciosas, fue capaz de conseguir diagnósticos y recomendaciones terapéuticas al menos tan buenos como los de los mejores expertos de su especialidad. Sin embargo, los propios creadores del modelo estaban insatisfechos con él, y por ello encargaron a un matemático, J. B. Adams un estudio, el cual demostró que en el método de combinación convergente de reglas había unas hipótesis implícitas tan fuertes como la independencia condicional exigida por el método probabilista, pero aún más difíciles de justificar. En los años siguientes surgieron nuevas críticas cada vez más fuertes contra la validez del modelo de factores de certeza (sec. 4.4).

Cuando los creadores de MYCIN tenían puestos sus ojos en la teoría de Dempster-Shafer como tabla de salvación del modelo de factores de certeza (estamos hablando de principios de los años 80), ocurrió un acontecimiento que cambió completamente el escenario: la aparición de las redes bayesianas, un modelo probabilista inspirado en la causalidad, cuya virtud principal consiste en que lleva asociado un modelo gráfico en que cada nodo representa una variable y cada enlace representa, generalmente, un mecanismo causal.[1] El extraordinario desarrollo experimentado por las redes bayesianas en esa década y, a ritmo más moderado pero constante, en los años 90, ha permitido construir modelos de diagnóstico y algoritmos eficientes para problemas de tamaño considerable, a veces con cientos de variables, o incluso con miles

[1] Conviene señalar que ésta es una cuestión muy discutida. Más aún, la propia existencia de la causalidad ha sido seriamente negada en algunas épocas: los ataques más conocidos y virulentos son los del joven Bertrand Russel, quien luego evolucionó hacia una oposición más moderada. Recientemente (a partir de 1993) se han publicado varios artículos y algún libro dedicados al estudio de la causalidad y, en particular, a su relación con los modelos gráficos probabilistas. Sin embargo, éste es un punto aún muy debatido y, a nuestro juicio, aún no se estudiado suficientemente en el papel esencial que desempeña la causalidad en las redes bayesianas; éste es el tema de un trabajo que tenemos en preparación (cf. sec. 3.2.4).

de variables en algunos problemas de genética. Por ello, algunos de los antiguos detractores del uso de la probabilidad en inteligencia artificial son hoy en día defensores entusiastas de los modelos gráficos probabilistas. Prácticamente todas las universidades más importantes de Estados Unidos y las empresas punteras de la informática tienen grupos de investigación dedicados a este tema. Microsoft, por ejemplo, creó en 1992 un grupo formado por algunos de los investigadores más prestigiosos del área, los cuales se han organizado recientemente en tres subgrupos, especializados en distintos aspectos de la aplicación de las redes bayesianas a la informática; de hecho, la inclusión de estos métodos y modelos en Windows 95/98 y Office 97/2000 ha hecho que las redes bayesianas sean la aplicación de la inteligencia artificial que ha llegado, con diferencia, a mayor número de usuarios. Otras empresas líderes de la informática, como Digital, Hewlett-Packard, IBM, Intel, Siemens, SRI, etc., cuentan igualmente con equipos de investigación en este campo. También en España hay un buen número de investigadores que, aunque muy dispersos por toda la geografía nacional, han empezado a trabajar de forma coordinada para abordar proyectos de mayor envergadura (se puede obtener más información a través de las páginas de Internet que indicamos más adelante); de hecho, creemos que, después de Estados Unidos, España es el país en que más universidades investigan sobre redes bayesianas.

En paralelo con esta evolución histórica de crisis y resurgimiento de la probabilidad, se desarrolló la *teoría de los conjuntos difusos*, frecuentemente llamada —con cierta impropiedad— *lógica difusa*.[2] La motivación inicial no fue el estudio de la incertidumbre, sino el estudio de la vaguedad, que es algo diferente. Por ejemplo, si sabemos que Juan mide 1'78 m., no podemos decir con rotundidad que es alto, pero tampoco podemos decir que no lo es: se trata de una cuestión de grado; en este caso hay vaguedad intrínseca, pero no hay incertidumbre, con lo que se demuestra que son dos conceptos en principio independientes, aunque existe una cierta relación en el sentido de que si recibimos una información imprecisa (por ejemplo, si nos dicen que Juan es alto, pero sin decirnos su estatura exacta) tenemos una cierta incertidumbre.

En realidad, la necesidad de tratar la vaguedad surge de una antigua paradoja, que podríamos expresar así: una persona que sólo tiene un céntimo de euro es sumamente pobre, indudablemente; ahora bien, si a una persona que es sumamente pobre le damos un céntimo, sigue siendo sumamente pobre; aplicando esta regla repetidamente, llegamos a la conclusión de que una persona que tiene 10 millones de euros es sumamente pobre. La solución a esta paradoja es que el concepto de "pobre" o "sumamente pobre" no tiene un límite completamente definido, sino que a medida que le damos a esa persona un céntimo tras otro, hasta llegar a los 10 millones de euros (en el supuesto de que tuviéramos esa cantidad de dinero), el grado de pobreza va disminuyendo paulatinamente: no hay un único céntimo que le haga pasar de ser pobre a ser rico.

Por eso, la brillante idea de Lofti Zadeh —considerado como el "padre" de la lógica difusa, no sólo por haber tenido la idea original, sino también por la gran cantidad de líneas que ha abierto en el campo desde entonces— consiste en permitir que el *grado de pertenencia* a algunos conjuntos sea un número entre 0 y 1, de modo que, por ejemplo, para quien no tiene más que dos pesetas, su grado de pertenencia al conjunto de personas pobres es 1, mientras que para quien tiene 1.500 millones de pesetas es 0; en cambio, para una persona que tiene

[2]Hay quienes traducen el adjetivo anglosajón *fuzzy logic* como *borroso*, mientras que otros lo traducimos como *difuso*. El motivo de preferir expresiones como *conjuntos difusos* en vez de *conjuntos borrosos*, es que el hecho de que no tengan una frontera bien definida es una propiedad intrínseca, mientras que el término *borroso* sugiere que se trata de un fenómeno de observación. En cualquier caso, conviene que el lector se acostumbre a encontrar ambos términos indistintamente.

500.000 pesetas ahorradas el grado de pertenencia podría ser 0'4 o 0'5 (cf. fig. 5.2, en la pág. 131). Lamentablemente, el punto más débil de la lógica difusa es la carencia de una definición operativa que permita determinar objetivamente el grado de pertenencia,[3] con lo que toda la teoría queda coja desde su nacimiento; esto no ha impedido el extraordinario desarrollo de la lógica difusa, con miles de artículos, libros, revistas y congresos dedicados al tema.

Al igual que la aplicación de la teoría de la probabilidad a los sistemas expertos y el surgimiento del modelo de factores de certeza de MYCIN vinieron motivados por la necesidad de abordar problemas médicos, la mayor parte de las aplicaciones de la lógica difusa se han desarrollado en el campo de la ingeniería y la industria, especialmente en Japón, donde el control difuso se está utilizando desde hace varios años en la supervisión de procesos de fabricación, en el guiado de ferrocarriles, en pequeños electrodomésticos, en cámaras de fotos, etc., etc. También en este campo es España uno de los países punteros, tanto por la importancia de las aportaciones teóricas como por las aplicaciones a la medicina, entre las que destaca el sistema experto MILORD.

Los cuatro métodos que acabamos de comentar corresponden a los cuatro capítulos siguientes de este texto:

- Capítulo 2: método probabilista clásico
- Capítulo 3: redes bayesianas
- Capítulo 4: modelo de factores de certeza
- Capítulo 5: lógica difusa.

Es importante señalar que, mientras las redes bayesianas y la lógica difusa son temas de gran actualidad, como lo prueba la intensa labor investigadora que se está realizando en cada uno de ellos, el método probabilista clásico y el modelo de factores de certeza se consideran temas "muertos" desde el punto de vista de la investigación, por razones diversas. En cuanto al método probabilista clásico, se trata en realidad de un caso particular de red bayesiana, que en la mayor parte de los problemas reales ha de ser sustituido por una red bayesiana general (un poliárbol o una red con bucles). En cuanto a los factores de certeza, Heckerman demostró en 1986 [26] no sólo que el modelo de MYCIN contiene graves incoherencias, sino que es imposible construir un modelo coherente de factores de certeza, salvo para casos sumamente simples.

Sin embargo, los motivos por los que los vamos a estudiar en este texto no son solamente históricos. Por un lado, el método probabilista clásico, al ser una red bayesiana muy simple, ayuda a entender mejor las redes bayesianas (se trata, por tanto de un motivo pedagógico); por otro, el mismo hecho de ser una red bayesiana muy simple permite aplicarle métodos de aprendizaje que no son válidos para redes bayesianas generales, con lo que se obtienen algunas ventajas [37]. En cuanto al modelo de factores de certeza, a pesar de sus graves

[3] Una de las formas propuestas —en la literatura, no en aplicaciones prácticas— para la asignación de grados de pertenencia consiste en hacer una encuesta, de modo que si el 80% de los encuestado opina que x pertenece al conjunto difuso A, tendríamos $\mu_A(x) = 0'8$. Sin embargo, esta interpretación del grado de pertenencia contradice la forma en que se aplican habitualmente las reglas de composición de conjuntos, mediante normas y conormas. Dado el carácter introductorio de este texto, no vamos a entrar en tales críticas, que son, por cierto, parte de nuestro trabajo de investigación en la actualidad.

inconsistencias se sigue utilizando en muchos de los sistemas expertos de la actualidad; de hecho —algunas herramientas comerciales destinadas a construir sistemas expertos, como GoldWorks, lo incorporan "de serie"— pues la única alternativa que existe para el razonamiento aproximado mediante reglas es la lógica difusa, con lo que resulta mucho más difícil construir el modelo (ya no es tan sencillo como asignar un factor de certeza a cada regla) y aumenta considerablemente el coste computacional del encadenamiento de reglas.

Como conclusión, queremos señalar que el debate sobre cuál es el método más adecuado para representar la incertidumbre sigue abierto hoy en día. Por un lado, está el grupo de los bayesianos (en el que se encuentran los autores de este texto), que defienden —algunos de ellos con gran vehemencia— que la teoría de la probabilidad es el único método correcto para el tratamiento de la incertidumbre. Por otro, están quienes señalan que los modelos probabilistas, a pesar de sus cualidades, resultan insuficientes o inaplicables en muchos problemas del mundo real, por lo que conviene disponer además de métodos alternativos.

Dado el carácter introductorio de esta obra no vamos a entrar aquí en el debate entre los defensores de una y otra postura, sino que nos limitamos a exponer de forma lo más objetiva posible los métodos más utilizados, con el fin de que el alumno que está a punto de convertirse en Ingeniero Técnico en Informática conozca estas herramientas por si algún día le pueden ser útiles en su labor profesional.

1.3 Bibliografía recomendada

De momento nos vamos a limitar a recomendar sólo un par de libros generales, y al final de cada uno de los capítulos siguientes daremos bibliografía específica. Como libro introductorio de nivel asequible recomendamos el de Krause y Clark [36], que además de los temas que hemos mencionado en este capítulo, incluye otros, como la teoría de Dempster-Shafer y algunos métodos cualitativos. Como obra de referencia, recomendamos el libro de Shafer y Pearl [56], en que se recogen, clasificados y comentados, la mayor parte de los artículos más relevantes sobre razonamiento aproximado publicados hasta la fecha (aunque, sorprendentemente, no hay ninguno dedicado a la lógica difusa).

Capítulo 2

Método probabilista clásico

El objetivo central de este capítulo es estudiar el método probabilista clásico. Para ello, debemos introducir primero los conceptos fundamentales sobre probabilidad: variables aleatorias, probabilidades conjunta, marginal y condicionada, etc. (sec. 2.1). Presentaremos después dos secciones independientes entre sí: una dedicada a los conceptos de independencia, correlación y causalidad (2.2) y otra al teorema de Bayes (2.3). Con esta base, podremos por fin estudiar el método probabilista clásico en la sección 2.4.

2.1 Definiciones básicas sobre probabilidad

Una exposición correcta de la teoría de la probabilidad debe apoyarse en la teoría de conjuntos, concretamente, en la teoría de la medida. Sin embargo, dado que en este capítulo vamos a tratar solamente con variables discretas, podemos simplificar considerablemente la exposición tomando como punto de partida el concepto de variable aleatoria.

▷ **Variable aleatoria.** Es aquélla que toma valores que, a priori, no conocemos con certeza.

En esta definición, "a priori" significa "antes de conocer el resultado de un acontecimiento, de un experimento o de una elección al azar". Por ejemplo, supongamos que escogemos al azar una persona dentro de una población; la edad y el sexo que va a tener esa persona son dos variables aleatorias, porque antes de realizar la elección no conocemos su valor.

Para construir un modelo matemático del mundo real —o, más exactamente, de una porción del mundo real, que llamaremos "sistema"— es necesario seleccionar un conjunto de variables que lo describan y determinar los posibles valores que tomará cada una de ellas. Los valores asociados a una variable han de ser *exclusivos* y *exhaustivos*. Por ejemplo, a la variable edad podemos asociarle tres valores: "menor de 18 años", "de 18 a 65 años" y "mayor de 65 años". Estos valores son *exclusivos* porque son incompatibles entre sí: una persona menor de 18 años no puede tener de 18 a 65 años ni más 65, etc. Son también *exhaustivos* porque cubren todas las posibilidades. En vez de escoger tres *intervalos* de edad, podríamos asignar a la variable edad el número de años que tiene la persona; en este caso tendríamos una *variable numérica*.

Es habitual representar cada variable mediante una letra mayúscula, a veces acompañada por un subíndice. Por ejemplo, podemos representar la variable edad mediante X_1 y la variable sexo mediante X_2. Los valores de las variables suelen representarse con letras minúsculas.

Por ejemplo, podríamos representar "menor de 18" mediante x_1^j, "de 18 a 65" mediante x_1^a y "mayor de 65" mediante x_1^t. (Hemos escogido los superíndices j, a y t como abreviaturas de "joven", "adulto" y "tercera edad", respectivamente.) Si en vez de representar un valor concreto de los tres queremos representar un valor genérico de la variable X_1, que puede ser cualquiera de los anteriores, escribiremos x_1, sin superíndice. Los dos valores de la variable sexo, X_2, que son "varón" y "mujer", pueden representarse mediante x_2^v y x_2^m, respectivamente.

Cuando tenemos un conjunto de variables $\{X_1, \ldots, X_n\}$, lo representaremos mediante $\bar{X}$. La n-tupla $\bar{x} = (x_1, \ldots, x_n)$ significa que cada variable X_i toma el correspondiente valor x_i. En el ejemplo anterior, el par (x_1^a, x_2^m) indicaría que la persona es una mujer adulta (entre 18 y 65 años).

En la exposición de este capítulo y de siguiente vamos a suponer que todas las variables son discretas. Nuestro punto de partida para la definición de probabilidad será el siguiente:

▷ **Probabilidad conjunta.** Dado un conjunto de variables discretas $\bar{X} = \{X_1, \ldots, X_n\}$, definimos la *probabilidad conjunta* como una aplicación que a cada n-tupla $\bar{x} = (x_1, \ldots, x_n)$ le asigna un número real no negativo de modo que

$$\sum_{\bar{x}} P(\bar{x}) = \sum_{x_1} \cdots \sum_{x_n} P(x_1, \ldots, x_n) = 1 \tag{2.1}$$

Recordemos que, según la notación que estamos utilizando, $P(x_1, \ldots, x_n)$ indica la probabilidad de que, para cada i, la variable X_i tome el valor x_i. Por ejemplo, $P(x_1^a, x_2^m)$ indica la probabilidad de que la persona escogida por cierto procedimiento aleatorio sea una mujer de entre 18 y 65 años.

▷ **Probabilidad marginal.** Dada una distribución de probabilidad conjunta $P(x_1, \ldots, x_n)$, la *probabilidad marginal* para un subconjunto de variables $\bar{X}' = \{X'_1, \ldots, X'_{n'}\} \subset \bar{X}$ viene dada por

$$P(\bar{x}') = P(x'_1, \ldots, x'_{n'}) = \sum_{x_i \,|\, X_i \notin \bar{X}'} P(x_1, \ldots, x_n) \tag{2.2}$$

El sumatorio indica que hay que sumar las probabilidades correspondientes a todos los valores de todas las variables de X que no se encuentran en X'. Por tanto, la distribución marginal para una variable X_i se obtiene sumando las probabilidades para todas las configuraciones posibles de las demás variables:

$$P(x_i) = \sum_{x_j \,|\, X_j \neq X_i} P(x_1, \ldots, x_n) \tag{2.3}$$

Proposición 2.1 Dada una distribución de probabilidad conjunta para X, toda distribución de probabilidad marginal obtenida a partir de ella para un subconjunto $X' \subset X$ es a su vez una distribución conjunta para X'.

Demostración. A partir de la definición anterior es fácil demostrar que $P(x'_1, \ldots, x'_{n'})$ es un número real no negativo; basta demostrar, por tanto, que la suma es la unidad. En efecto, tenemos que

$$\sum_{\bar{x}'} P(\bar{x}') = \sum_{x'_1} \cdots \sum_{x'_n} P(x'_1, \ldots, x'_{n'}) = \sum_{x_i \,|\, X_i \in \bar{X}'} \left[\sum_{x_i \,|\, X_i \notin \bar{X}'} P(x_1, \ldots, x_n) \right]$$

Como las variables son discretas, el número de sumandos es finito, por lo que podemos reordenar los sumatorios de modo que

$$\sum_{\bar{x}'} P(\bar{x}') = \sum_{x_1} \cdots \sum_{x_n} P(x_1, \ldots, x_n) = 1 \tag{2.4}$$

con lo que concluye la demostración. □

Corolario 2.2 La suma de las probabilidades de los valores de cada variable ha de ser la unidad:

$$\sum_{x_i} P(x_i) = 1 \tag{2.5}$$

Ejemplo 2.3 Supongamos que tenemos una población de 500 personas cuya distribución por edades y sexos es la siguiente:

N	Varón	Mujer	TOTAL
<18	67	68	135
18-65	122	126	248
>65	57	60	117
TOTAL	246	254	500

Realizamos un experimento que consiste en escoger una persona mediante un procedimiento aleatorio en que cada una de ellas tiene la misma probabilidad de resultar elegida. En este caso, la probabilidad de que la persona tenga cierta edad y cierto sexo es el número de personas de esa edad y ese sexo, dividido por el total de personas en la población: $P(x_1, x_2) = N(x_1, x_2)/N$. Por tanto, la tabla de probabilidad será la siguiente:

P	Varón	Mujer	TOTAL
<18	$P(x_1^j, x_2^v) = 0'134$	$P(x_1^j, x_2^m) = 0'136$	$P(x_1^j) = 0'270$
18-65	$P(x_1^a, x_2^v) = 0'244$	$P(x_1^a, x_2^m) = 0'252$	$P(x_1^a) = 0'496$
>65	$P(x_1^t, x_2^v) = 0'114$	$P(x_1^t, x_2^m) = 0'120$	$P(x_1^t) = 0'234$
TOTAL	$P(x_2^v) = 0'492$	$P(x_2^m) = 0'508$	$1'000$

Las probabilidades marginales se obtienen sumando por filas (para X_1) o por columnas (para X_2), de acuerdo con la ec. (2.2). Observe que estas probabilidades marginales también se podrían haber obtenido a partir de la tabla de la población general. Por ejemplo: $P(x_1^j) = N(x_1^j)/N = 135/500 = 0'270$. Naturalmente, la suma de las probabilidades de los valores de cada variable es la unidad. □

▷ **Probabilidad condicional.** Dados dos subconjuntos disjuntos de variables, $\bar{X} = \{X_1, \ldots, X_n\}$ e $\bar{Y} = \{Y_1, \ldots, Y_m\}$, y una tupla $\bar{x}$ (es decir, una asignación de valores para las variables de $\bar{X}$) tal que $P(\bar{x}) > 0$, la *probabilidad condicional* de $\bar{y}$ dado $\bar{x}$, $P(\bar{y}|\bar{x})$, se define como

$$P(\bar{y} \,|\, \bar{x}) = \frac{P(\bar{x}, \bar{y})}{P(\bar{x})} \tag{2.6}$$

□

El motivo de exigir que $P(\bar{x}) > 0$ es que $P(\bar{x}) = 0$ implica que $P(\bar{x}, \bar{y}) = 0$, lo que daría lugar a una indeterminación.

Ejemplo 2.4 Continuando con el ejemplo anterior, la probabilidad de que un varón sea mayor de 65 años es la probabilidad de ser mayor de 65 años (x_1^t) dado que sabemos que es varón (x_2^v): $P(x_1^t \,|\, x_2^v) = P(x_1^t, x_2^v)/P(x_2^v) = 0'114/0'492 = 0'23171$. Observe que, como era de esperar, este resultado coincide con la proporción de varones mayores de 65 años dentro del grupo de varones: $N(x_1^t, x_2^v)/N(x_2^v) = 57/246 = 0'23171$. En cambio, la probabilidad de que una persona mayor de 65 años sea varón es $P(x_2^v \,|\, x_1^t) = P(x_1^t, x_2^v)/P(x_1^t) = 0'114/0'234 = 0'48718$. Se comprueba así que, en general, $P(x_1 \,|\, x_2) \neq P(x_2 \,|\, x_1)$. Igualmente, se puede calcular la probabilidad de que una persona mayor de 65 años sea mujer: $P(x_2^m \,|\, x_1^t) = P(x_1^t, x_2^m)/P(x_1^t) = 0'120/0'234 = 0'51282$. Por tanto, $P(x_2^v \,|\, x_1^t) + P(x_2^m \,|\, x_1^t) = 0'48718 + 0'51282 = 1$, como era de esperar, pues toda persona mayor de 65 años ha de ser o varón o mujer, y no hay otra posibilidad. □

Este resultado se puede generalizar como sigue:

Proposición 2.5 Dados dos subconjuntos disjuntos de variables, $\bar{X}$ e $\bar{Y}$, y una tupla $\bar{x}$ tal que $P(\bar{x}) > 0$, se cumple que

$$\forall \bar{x}, \quad \sum_{\bar{y}} P(\bar{y} \,|\, \bar{x}) = 1 \tag{2.7}$$

Demostración. Aplicando las definiciones anteriores,

$$\sum_{\bar{y}} P(\bar{y} \,|\, \bar{x}) = \sum_{\bar{y}} \frac{P(\bar{x}, \bar{y})}{P(\bar{x})} = \frac{1}{P(\bar{x})} \sum_{\bar{y}} P(\bar{x}, \bar{y}) = \frac{1}{P(\bar{x})} P(\bar{x}) = 1 \tag{2.8}$$

□

Observe que esta proposición es el equivalente de la ecuación (2.4) para probabilidades condicionadas.

Ejercicio 2.6 Como aplicación de este resultado, comprobar que $\sum_{x_2} P(x_2|x_1) = 1$ para todos los valores de x_1 en el ejemplo 2.3 (antes lo hemos demostrado sólo para x_1^t). También se puede comprobar que $\sum_{x_1} P(x_1|x_2) = 1$, tanto para x_2^v como para x_2^m.

Teorema 2.7 (Teorema de la probabilidad total) Dados dos subconjuntos disjuntos de variables, $\bar{X}$ e $\bar{Y}$, se cumple que

$$P(\bar{y}) = \sum_{\bar{x} \,|\, P(\bar{x})>0} P(\bar{y} \,|\, \bar{x}) \cdot P(\bar{x}) \tag{2.9}$$

Demostración. Por la definición de probabilidad marginal,

$$P(\bar{y}) = \sum_{\bar{x}} P(\bar{x}, \bar{y})$$

Ahora bien, $P(\bar{x}) = 0$ implica que $P(\bar{x}, \bar{y}) = 0$, por lo que sólo es necesario incluir en la suma las tuplas cuya probabilidad es positiva:

$$P(\bar{y}) = \sum_{\bar{x} \,|\, P(\bar{x})>0} P(\bar{x}, \bar{y})$$

Basta aplicar ahora la definición de probabilidad condicional para concluir la demostración. □

Este resultado se puede generalizar como sigue (observe que la proposición siguiente no es más que el teorema de la probabilidad total, con condicionamiento):

Proposición 2.8 Dados tres subconjuntos disjuntos de variables, $\bar{X}$, $\bar{Y}$ y $\bar{Z}$, si $P(\bar{z}) > 0$, se cumple que

$$P(\bar{y} \mid \bar{z}) = \sum_{\bar{x} \mid P(\bar{x} \mid \bar{z})>0} P(\bar{y} \mid \bar{x}, \bar{z}) \cdot P(\bar{x} \mid \bar{z}) \tag{2.10}$$

Demostración. Por la definición de probabilidad condicional,

$$P(\bar{y} \mid \bar{z}) = \frac{P(\bar{y}, \bar{z})}{P(\bar{z})} = \frac{1}{P(\bar{z})} \sum_{\bar{x}} P(\bar{x}, \bar{y}, \bar{z})$$

Al igual que en el teorema anterior, basta sumar para aquellas tuplas $\bar{x}$ tales que $P(\bar{x} \mid \bar{z}) > 0$, que son las mismas para las que $P(\bar{x}, \bar{z}) > 0$, pues

$$P(\bar{z}) > 0 \Longrightarrow \{P(\bar{x} \mid \bar{z}) = 0 \Leftrightarrow P(\bar{x}, \bar{z}) = 0\}$$

Por tanto

$$\begin{aligned} P(\bar{y} \mid \bar{z}) &= \frac{1}{P(\bar{z})} \sum_{\bar{x} \mid P(\bar{x} \mid \bar{z})>0} P(\bar{x}, \bar{y}, \bar{z}) = \sum_{\bar{x} \mid P(\bar{x} \mid \bar{z})>0} \frac{P(\bar{x}, \bar{y}, \bar{z})}{P(\bar{z})} \\ &= \sum_{\bar{x} \mid P(\bar{x} \mid \bar{z})>0} \frac{P(\bar{x}, \bar{y}, \bar{z})}{P(\bar{x}, \bar{z})} \cdot \frac{P(\bar{x}, \bar{z})}{P(\bar{z})} = \sum_{\bar{x} \mid P(\bar{x} \mid \bar{z})>0} P(\bar{y} \mid \bar{x}, \bar{z}) \cdot P(\bar{x} \mid \bar{z}) \end{aligned}$$

Ejemplo 2.9 (Continuación del ejemplo 2.3) La probabilidad de ser varón dentro de cada intervalo de edad es $P(x_2^v|x_1^j) = 0'49630$, $P(x_2^v|x_1^a) = 0'49194$ y $P(x_2^v|x_1^t) = 0'48718$. Aplicando el teorema de la probabilidad total,

$$P(x_2^v) = \sum_{x_1} P(x_2^v|x_1) \cdot P(x_1) \tag{2.11}$$

$$= 0'49630 \cdot 0'270 + 0'49194 \cdot 0'496 + 0'48718 \cdot 0'243 \tag{2.12}$$

$$= 0'134 + 0'244 + 0'114 = 0'492 \tag{2.13}$$

que es el valor que ya conocíamos. □

Finalmente, enunciamos una proposición que se deduce fácilmente de la definición de probabilidad condicional, pero que nos será de gran utilidad en este capítulo y en el siguiente.

Proposición 2.10 (Factorización de la probabilidad conjunta) Dado un conjunto de variables $\bar{X}$ y una partición $\{\bar{X}_1, \ldots, \bar{X}_k\}$ de $\bar{X}$, se cumple que

$$P(\bar{x}) = \prod_{i=1}^{k} P(\bar{x}_i \mid \bar{x}_{i+1}, \ldots, \bar{x}_k) \tag{2.14}$$

Demostración. Por la definición de probabilidad condicional

$$\begin{aligned} P(\bar{x}) = P(\bar{x}_1, \ldots, \bar{x}_k) &= P(\bar{x}_1 \mid \bar{x}_2, \ldots, \bar{x}_k) \cdot P(\bar{x}_2, \ldots, \bar{x}_k) \\ P(\bar{x}_2, \ldots, \bar{x}_k) &= P(\bar{x}_2 \mid \bar{x}_3, \ldots, \bar{x}_k) \cdot P(\bar{x}_3, \ldots, \bar{x}_k) \\ &\vdots \\ P(\bar{x}_{k-1}, \bar{x}_k) &= P(\bar{x}_{k-1} \mid \bar{x}_k) \cdot P(\bar{x}_k) \end{aligned}$$

Basta sustituir cada igualdad en la anterior para concluir la demostración. □

Ejemplo 2.11 Sea $\bar{X} = \{A, B, C, D, E\}$. Para la partición $\{\{A, D\}, \{C\}, \{B, E\}\}$ tenemos

$$P(a, b, c, d, e) = P(a, d \,|\, b, c, e) \cdot P(c \,|\, b, e) \cdot P(b, e)$$

Del mismo modo, para la partición $\{\{B\}, \{D\}, \{C\}, \{A\}, \{E\}\}$ tenemos

$$P(a, b, c, d, e) = P(b \,|\, a, c, d, e) \cdot P(d \,|\, a, c, e) \cdot P(c \,|\, a, e) \cdot P(a \,|\, e) \cdot P(e)$$

2.2 Independencia, correlación y causalidad

2.2.1 Independencia y correlaciones

▷ **Valores independientes.** Dos valores x e y de dos variables X e Y, respectivamente, son independientes sii $P(x, y) = P(x) \cdot P(y)$.

▷ **Valores correlacionados.** Dos valores x e y de dos variables X e Y, respectivamente, están correlacionados sii no son independientes, es decir, sii $P(x, y) \neq P(x) \cdot P(y)$. Cuando $P(x, y) > P(x) \cdot P(y)$, se dice que hay correlación positiva. Cuando $P(x, y) < P(x) \cdot P(y)$, se dice que hay correlación negativa.

Ejemplo 2.12 (Continuación del ejemplo 2.3) Entre ser varón y ser menor de 18 años hay correlación positiva, porque $P(x_1^j, x_2^v) = 0'134 > P(x_1^j) \cdot P(x_2^v) = 0'270 \cdot 0'492 = 0'13284$, aunque es una correlación débil. Igualmente, hay una débil correlación positiva entre ser mujer y mayor de 65 años: $P(x_1^t, x_2^m) = 0'120 > P(x_1^t) \cdot P(x_2^m) = 0'234 \cdot 0'508 = 0'118872$. En cambio, entre ser varón y mayor de 65 años hay correlación negativa, pues $P(x_1^t, x_2^v) = 0'114 < P(x_1^t) \cdot P(x_2^v) = 0'234 \cdot 0'492 = 0'115128$.

Consideramos ahora una tercera variable, X_3, el color de los ojos, de modo que x_3^{az} indica "ojos azules". Supongamos que la probabilidad de tener los ojos de un cierto color es la misma para cada edad: $P(x_3|x_1) = P(x_3)$; entonces, dados dos valores cualesquiera x_1 y x_3, han de ser independientes. Del mismo modo, si la probabilidad de tener los ojos de un cierto color es la misma para cada sexo, entonces x_2 y x_3 han de ser independientes [para todo par (x_2, x_3)]. □

De los conceptos de independencia y correlación **entre valores** podemos pasar a los de independencia y correlación **entre variables**.

▷ **Variables independientes.** Dos variables X e Y son independientes sii todos los pares de valores x e y son independientes, es decir, sii

$$\forall x, \forall y, \quad P(x, y) = P(x) \cdot P(y) \tag{2.15}$$

▷ **Variables correlacionadas.** Dos variables X e Y están correlacionadas sii no son independientes, es decir, sii

$$\exists x, \exists y, \quad P(x, y) \neq P(x) \cdot P(y) \tag{2.16}$$

Hemos visto anteriormente que, cuando dos valores están correlacionados, la correlación ha de ser necesariamente o positiva o negativa. Sin embargo, en el caso de dos variables correlacionadas la cuestión es bastante más compleja. Intuitivamente, podemos decir que

entre dos variables X e Y hay correlación positiva cuando los valores altos de una están correlacionados positivamente con los valores altos de la otra y negativamente con los valores bajos de ella; por ejemplo, dentro de la población infantil hay correlación positiva entre la edad y la estatura. Por tanto, la primera condición para poder hablar del signo de la correlación entre dos variables es que ambas sean ordinales; cuando una de ellas no lo es (por ejemplo, el sexo y el color de ojos no son variables ordinales), no tiene sentido buscar el signo de la correlación. Además, es necesario establecer una definición matemática precisa, lo cual encierra algunas sutilezas en las que no vamos a entrar, dado el carácter introductorio de esta obra, por lo que nos quedamos con la definición intuitiva anterior.

Estas definiciones de correlación e independencia se pueden generalizar inmediatamente de dos variables X e Y a dos conjuntos de variables $\bar{X}$ e $\bar{Y}$, y de dos valores x e y a dos tuplas $\bar{x}$ e $\bar{y}$.

2.2.2 Independencia condicional

▷ **Valores condicionalmente independientes.** Sean tres valores x, y y z de las variables X, Y y Z, respectivamente, tales que $P(z) > 0$; x e y son condicionalmente independientes dado z sii $P(x, y|z) = P(x|z) \cdot P(y|z)$.

▷ **Variables condicionalmente independientes.** Las variables X e Y son condicionalmente independientes dada una tercera variable Z sii todo par de valores x e y es condicionalmente independiente para cada z tal que $P(z) > 0$; es decir, sii

$$\forall x, \forall y, \forall z, \quad P(z) > 0 \implies P(x, y \,|\, z) = P(x \,|\, z) \cdot P(y \,|\, z) \tag{2.17}$$

▷ **Separación.** La variable Z *separa* las variables X e Y sii éstas dos últimas son condicionalmente independientes dada Z.

Estas definiciones son igualmente válidas para conjuntos de variables $\bar{X}$, $\bar{Y}$ y $\bar{Z}$.

Proposición 2.13 Sea un conjunto de variables $\bar{Y} = \{Y_1, \ldots, Y_m\}$ y una tupla $\bar{x}$ de $\bar{X}$ que separa el conjunto $\bar{Y}$, de modo que $P(\bar{y}|\bar{x}) = \prod_{j=1}^{m} P(y_j|\bar{x})$. Para todo subconjunto $\bar{Y}'$ de $\bar{Y}$ se cumple que:

$$\forall \bar{y}', \; P(\bar{y}'|\bar{x}) = \prod_{j|Y_j \in \bar{Y}'} P(y_j|\bar{x}) \tag{2.18}$$

Demostración. Por la definición de probabilidad marginal,

$$\begin{aligned} P(\bar{y}' \,|\, \bar{x}) &= \sum_{y_j \,|\, Y_j \notin \bar{Y}'} P(\bar{y} \,|\, \bar{x}) = \sum_{y_j \,|\, Y_j \notin \bar{Y}'} \prod_{j=1}^{m} P(y_j \,|\, \bar{x}) \\ &= \left[\prod_{j \,|\, Y_j \in \bar{Y}'} P(y_j \,|\, \bar{x}) \right] \cdot \left[\sum_{y_j \,|\, Y_j \notin \bar{Y}} \; \prod_{j \,|\, Y_j \notin \bar{Y}'} P(y_j \,|\, \bar{x}) \right] \end{aligned}$$

Aplicando la propiedad distributiva de la suma y el producto recursivamente dentro del segundo corchete, se van "eliminando" variables, con lo que al final se obtiene la unidad. El siguiente ejemplo ilustra el proceso.

Ejemplo 2.14 Sean $\bar{Y} = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$ e $\bar{Y}' = \{Y_1, Y_4\}$. Supongamos que se cumple la condición (2.18), que para este ejemplo es

$$P(y_1, y_2, y_3, y_4, y_5 \,|\, \bar{x}) = P(y_1 \,|\, \bar{x}) \cdot P(y_2 \,|\, \bar{x}) \cdot P(y_3 \,|\, \bar{x}) \cdot P(y_4 \,|\, \bar{x}) \cdot P(y_5 \,|\, \bar{x})$$

El cálculo de $P(y_1, y_4 \,|\, \bar{x})$ se realiza así

$$\begin{aligned} P(y_1, y_4 \,|\, \bar{x}) &= \sum_{y_2} \sum_{y_3} \sum_{y_5} P(y_1, y_2, y_3, y_4, y_5 \,|\, \bar{x}) \\ &= P(y_1 \,|\, \bar{x}) \cdot P(y_4 \,|\, \bar{x}) \cdot \left[\sum_{y_2} \sum_{y_3} \sum_{y_5} P(y_2 \,|\, \bar{x}) \cdot P(y_3 \,|\, \bar{x}) \cdot P(y_5 \,|\, \bar{x}) \right] \end{aligned}$$

El resultado de calcular los sumatorios da la unidad, pues

$$\begin{aligned} &\sum_{y_2} \sum_{y_3} \sum_{y_5} P(y_2 \,|\, \bar{x}) \cdot P(y_3 \,|\, \bar{x}) \cdot P(y_5 \,|\, \bar{x}) = \\ &\qquad = \sum_{y_2} \left[P(y_2 \,|\, \bar{x}) \cdot \sum_{y_3} \sum_{y_5} P(y_3 \,|\, \bar{x}) \cdot P(y_5 \,|\, \bar{x}) \right] \\ &\qquad = \left[\sum_{y_2} P(y_2 \,|\, \bar{x}) \right] \cdot \left[\sum_{y_3} \sum_{y_5} P(y_3 \,|\, \bar{x}) \cdot P(y_5 \,|\, \bar{x}) \right] \\ &\qquad = \sum_{y_3} \sum_{y_5} P(y_3 \,|\, \bar{x}) \cdot P(y_5 \,|\, \bar{x}) = \sum_{y_3} \left[P(y_3 \,|\, \bar{x}) \cdot \sum_{y_5} P(y_5 \,|\, \bar{x}) \right] \\ &\qquad = \left[\sum_{y_3} P(y_3 \,|\, \bar{x}) \right] \cdot \left[\sum_{y_5} P(y_5 \,|\, \bar{x}) \right] = 1 \end{aligned}$$

Proposición 2.15 Sea un conjunto de variables $\bar{Y} = \{Y_1, \ldots, Y_m\}$ y una tupla $\bar{x}$ de $\bar{X}$ que separa el conjunto $\bar{Y}$, de modo que $P(\bar{y}|\bar{x}) = \prod_{j=1}^{m} P(y_j|\bar{x})$. Dados dos tuplas cualesquiera y' e y'' de dos subconjuntos $\bar{Y}'$ e $\bar{Y}''$ disjuntos de $\bar{Y}$, se cumple que

$$P(\bar{y}', \bar{y}''|\bar{x}) = P(\bar{y}'|\bar{x}) \cdot P(\bar{y}''|\bar{x}) \tag{2.19}$$

$$P(\bar{y}'|\bar{x}, \bar{y}'') = P(\bar{y}'|\bar{x}) \tag{2.20}$$

Demostración. Por la proposición anterior,

$$\begin{aligned} P(\bar{y}', \bar{y}'' \,|\, \bar{x}) &= \prod_{j \,|\, Y_j \in (\bar{Y}' \cup \bar{Y}'')} P(y_j \,|\, \bar{x}) = \left[\prod_{j \,|\, Y_j \in \bar{Y}'} P(y_j \,|\, \bar{x}) \right] \cdot \left[\prod_{j \,|\, Y_j \in \bar{Y}''} P(y_j \,|\, \bar{x}) \right] \\ &= P(\bar{y}' \,|\, \bar{x}) \cdot P(\bar{y}'' \,|\, \bar{x}) \end{aligned}$$

y de aquí se deduce que

$$P(\bar{y}' \,|\, \bar{x}, \bar{y}'') = \frac{P(\bar{x}, \bar{y}', \bar{y}'')}{P(\bar{x}, \bar{y}'')} = \frac{P(\bar{y}', \bar{y}'' \,|\, \bar{x})}{P(\bar{y}'' \,|\, \bar{x})} = \frac{P(\bar{y}' \,|\, \bar{x}) \cdot P(\bar{y}'' \,|\, \bar{x})}{P(\bar{y}'' \,|\, \bar{x})} = P(\bar{y}' \,|\, \bar{x})$$

Ejemplo 2.16 Sea de nuevo el conjunto $\bar{Y} = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$, que cumple la condición (2.18). $\bar{Y}' = \{Y_1, Y_4\}$ e $\bar{Y}'' = \{Y_2\}$. La proposición que acabamos de demostrar nos dice que

$$P(y_1, y_2, y_4 \,|\, \bar{x}) = P(y_1, y_4 \,|\, \bar{x}) \cdot P(y_2 \,|\, \bar{x})$$
$$P(y_1, y_4 \,|\, \bar{x}, y_2) = P(y_1, y_4 \,|\, \bar{x})$$

□

Estas dos proposiciones nos serán muy útiles en la sección 2.4 y en el próximo capítulo.

2.2.3 Representación gráfica de dependencias e independencias

En los modelos gráficos, cada nodo representa una variable, y los enlaces o ausencia de enlaces representan las correlaciones entre ellos. Por ejemplo, la figura 2.1 indica que el sexo y el color de ojos son variables independientes. Cuando una variable ejerce influencia causal sobre otra, se traza una flecha entre ambas, como muestra la figura 2.2, la cual representa el hecho de que la edad de una persona influye en los ingresos que percibe.

Figura 2.1: Dos variables independientes.

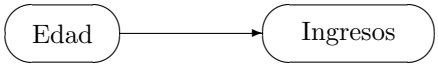

Figura 2.2: Dependencia causal entre dos variables.

En el caso de considerar tres variables correlacionadas entre sí, podemos encontrar distintos tipos de estructuras. Por ejemplo, la figura 2.3 indica que

- la edad influye en la estatura
- la edad influye en los ingresos
- hay correlación (a priori) entre la estatura y los ingresos
- cuando se conoce la edad, desaparece la correlación entre estatura e ingresos.

En cambio, la figura 2.4 indica que la edad influye en la estatura y la estatura influye en el número que calza la persona; además, la ausencia de un enlace entre la edad y el número de calzado indica que ambas variables son independientes cuando se conoce la estatura; es decir, este modelo afirma que dos personas que midan 1'75 tienen la misma probabilidad de calzar cada número, independientemente de cuál sea su edad.

Otro modelo posible es el que relaciona la edad, el sexo y la estatura (fig. 2.5). Observemos que en él sucede, en cierto sentido, lo contrario que en los modelos anteriores. Vimos

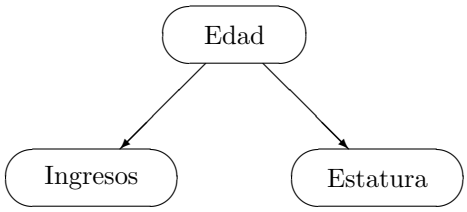

Figura 2.3: Dependencia causal entre un nodo padre y dos hijos.

Figura 2.4: Dependencia causal de tres variables en cadena.

que la estatura y los ingresos son dos variables correlacionadas a priori (fig. 2.3), pero la correlación desaparecía al conocer la edad. Del mismo modo, la edad y el número de calzado eran variables correlacionadas a priori (fig. 2.4), pero la correlación desaparecía al conocer la estatura. Sin embargo, en este último ejemplo se da la situación inversa: la edad y el sexo son variables independientes (estamos suponiendo que la esperanza de vida es igual para hombres y mujeres), pero ambas pasan a estar correlacionadas cuando se conoce la estatura.[1] Esta asimetría se conoce con el nombre de **separación direccional** y constituye la piedra angular de las redes bayesianas, como veremos detenidamente a lo largo del próximo capítulo.

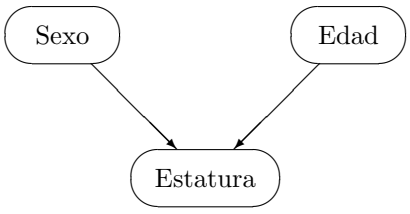

Figura 2.5: Dependencia causal entre dos padres y un hijo.

Para concluir esta sección, veamos en la figura 2.6 un ejemplo en que aparece un bucle. Este modelo nos dice que el coche que compra una persona depende de sus ingresos (obviamente) pero también de su edad: una persona joven tiende a comprar coches deportivos, con una estética moderna, mientras que una persona de mayor edad generalmente da más importancia a la comodidad y a la seguridad.

2.2.4 Diferencia entre causalidad y correlación

Se ha discutido mucho sobre si la correlación matemática representa solamente correlación o si en algunos casos representa causalidad; en realidad, lo que se discute es la esencia misma de la causalidad. Aunque ésta es una cuestión principalmente filosófica, recientemente han surgido

[1] Se puede ver la correlación con el siguiente razonamiento: un varón que mide 1'55 tiene menor probabilidad de ser adulto que una mujer que tenga esa misma estatura. Es decir, para una cierta estatura, surgen correlaciones entre el sexo y la edad.

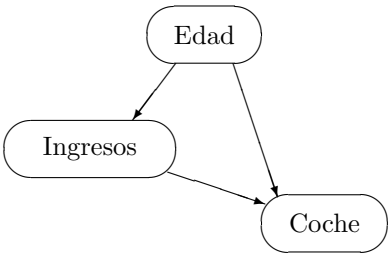

Figura 2.6: Diagrama causal en forma de bucle.

argumentos matemáticos en favor de la causalidad como algo diferente de la mera correlación. Nuestra opinión es que la causalidad existe y es distinta de la correlación. De hecho, vamos a mostrar en esta sección que causalidad implica correlación pero no a la inversa.

Por ejemplo, un estudio llevado a cabo en Inglaterra demostró que había una fuerte correlación entre el número de cigüeñas de cada localidad y el número de nacimiento de niños. ¿Podría utilizarse este hallazgo para afirmar que son las cigüeñas las que traen los niños? ¿O es acaso la presencia de niños lo que atrae a las cigüeñas?

Naturalmente, no hace falta buscar hipótesis tan extrañas para explicar tal correlación, pues existe una alternativa mucho más razonable: el número de habitantes de una localidad influye en el número de iglesias (en general, cuantos más habitantes, más iglesias), con lo que las cigüeñas tienen más campanarios donde poner sus nidos. Por otro lado, hay una correlación natural entre el número de habitantes y el número de nacimientos. Gráficamente lo representaríamos mediante la figura 2.7. Este gráfico nos dice que el número de nacimientos está correlacionado tanto con el número de cigüeñas como con el número de iglesias, pero es condicionalmente independiente de ambos dado el número de habitantes.

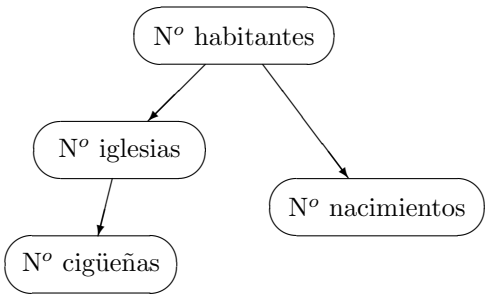

Figura 2.7: La correlación entre número de cigüeñas y número de nacimientos **no** implica causalidad.

Por poner otro ejemplo, ideado por Ross Shachter, supongamos que se ha comprobado que existe una correlación significativa entre el consumo de teracola (una bebida imaginaria) y la aparición de manchas en la piel. ¿Significa eso que las autoridades sanitarias deben prohibir la venta de esa bebida? Es posible; pero consideremos una explicación alternativa, representada por el diagrama de la figura 2.8, el cual afirma que la verdadera causa de las manchas en la piel es el contagio en la piscina. La correlación observada se explica, según este modelo, porque un aumento de la temperatura provoca, por un lado, que la gente vaya más a la piscina y, por otro, que beba más refrescos. Además, son las personas con más ingresos económicos las que pueden permitirse ir con mayor frecuencia a la piscina y, a la vez, comprar

más bebidas. Estas dos razones hacen que el consumo de teracola esté correlacionado con el hecho de ir a la piscina y, en consecuencia, correlacionado con la aparición de manchas en la piel. La correlación desaparecería si el supuesto estudio epidemiológico hubiera considerado la temperatura ambiental y los ingresos de la persona, pues el consumo de teracola y la aparición de manchas en la piel son condicionalmente independientes dadas la temperatura y los ingresos.

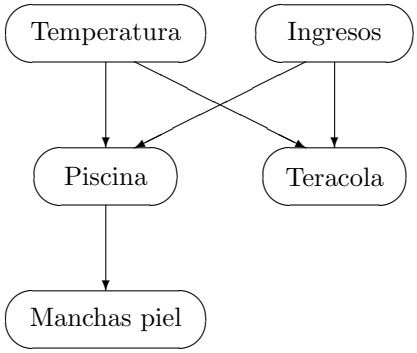

Figura 2.8: La correlación entre el consumo de teracola y la aparición de manchas en la piel **no** implica causalidad.

Con este par de ejemplos hemos intentado mostrar que correlación y causalidad son conceptos muy distintos (la causalidad implica correlación, pero la correlación no implica causalidad) y —lo que es mucho más importante en medicina, psicología, sociología, etc.— que hay que tener mucho cuidado al interpretar los resultados de un estudio epidemiológico, una estadística o una encuesta para evitar sacar conclusiones erróneas.

Por ejemplo, en 1996 publicaba cierto periódico la noticia de que un estudio llevado a cabo en Estados Unidos había demostrado que las personas que comen más tomates tienen menos riesgo de padecer cáncer (un hecho experimental) y de ahí deducía que conviene comer más tomate para reducir el riesgo de cáncer, una conclusión que podría parecer evidente, pero que es muy cuestionable a la luz de los ejemplos que acabamos de mostrar.

2.3 Teorema de Bayes

2.3.1 Enunciado y demostración

La forma clásica del teorema de Bayes es la siguiente:

Teorema 2.17 (Teorema de Bayes) Dadas dos variables X e Y, tales que $P(x) > 0$ para todo x y $P(y) > 0$ para todo y, se cumple

$$P(x \,|\, y) = \frac{P(x) \cdot P(y \,|\, x)}{\sum\limits_{x'} P(x') \cdot P(y \,|\, x')} \tag{2.21}$$

Este teorema se puede generalizar así:

Teorema 2.18 (Teorema de Bayes generalizado) Dadas dos n-tuplas $\bar{x}$ e $\bar{y}$ de dos subconjuntos de variables $\bar{X}$ e $\bar{Y}$, respectivamente, tales que $P(\bar{x}) > 0$ y $P(\bar{y}) > 0$, se cumple

que

$$P(\bar{x} \mid \bar{y}) = \frac{P(\bar{x}) \cdot P(\bar{y} \mid \bar{x})}{\sum_{\bar{x}' \mid P(\bar{x}')>0} P(\bar{x}') \cdot P(\bar{y} \mid \bar{x}')} \tag{2.22}$$

Demostración. Por la definición de probabilidad condicional, $P(\bar{x}, \bar{y}) = P(\bar{x}) \cdot P(\bar{y} \mid \bar{x})$, y por tanto,

$$P(\bar{x} \mid \bar{y}) = \frac{P(\bar{x}, \bar{y})}{P(\bar{y})} = \frac{P(\bar{x}) \cdot P(\bar{y} \mid \bar{x})}{P(\bar{y})}$$

Basta aplicar el teorema de la probabilidad total (proposición 2.7) para completar la demostración. □

Proposición 2.19 Dados tres subconjuntos disjuntos $\bar{X}$, $\bar{Y}$ y $\bar{Z}$, si $P(\bar{y}, \bar{z}) > 0$, se cumple que

$$P(\bar{x}, \bar{y} \mid \bar{z}) = P(\bar{x} \mid \bar{y}, \bar{z}) \cdot P(\bar{y} \mid \bar{z})$$

Demostración. Veamos primero que

$$P(\bar{z}) = \sum_{\bar{y}} P(\bar{y}, \bar{z}) > 0$$

Teniendo en cuenta que —por la definición de probabilidad condicional— $P(\bar{x}, \bar{y}, \bar{z}) = P(\bar{x} \mid \bar{y}, \bar{z}) \cdot P(\bar{y}, \bar{z})$, llegamos a

$$P(\bar{x}, \bar{y} \mid \bar{z}) = \frac{P(\bar{x}, \bar{y}, \bar{z})}{P(\bar{z})} = \frac{P(\bar{x} \mid \bar{y}, \bar{z}) \cdot P(\bar{y}, \bar{z})}{P(\bar{z})} = P(\bar{x} \mid \bar{y}, \bar{z}) \cdot P(\bar{y} \mid \bar{z})$$

como queríamos demostrar. □

Proposición 2.20 (Teorema de Bayes con condicionamiento) Dadas tres tuplas $\bar{x}$, $\bar{y}$ y $\bar{z}$ de tres conjuntos de variables $\bar{X}$, $\bar{Y}$ y $\bar{Z}$, respectivamente, tales que $P(\bar{x}, \bar{z}) > 0$ y $P(\bar{y}, \bar{z}) > 0$, se cumple que

$$P(\bar{x} \mid \bar{y}, \bar{z}) = \frac{P(\bar{x} \mid \bar{z}) \cdot P(\bar{y} \mid \bar{x}, \bar{z})}{\sum_{\bar{x}' \mid P(\bar{x}' \mid \bar{z})>0} P(\bar{y} \mid \bar{x}', \bar{z}) \cdot P(\bar{x}' \mid \bar{z})}$$

Demostración. Por la definición de probabilidad condicional,

$$P(\bar{x} \mid \bar{y}, \bar{z}) = \frac{P(\bar{x}, \bar{y}, \bar{z})}{P(\bar{y}, \bar{z})} = \frac{P(\bar{y} \mid \bar{x}, \bar{z}) \cdot P(\bar{x}, \bar{z})}{P(\bar{y}, \bar{z})} \tag{2.23}$$

Por otro lado, $P(\bar{x}, \bar{z}) > 0$ implica que $P(\bar{z}) > 0$, por lo que podemos escribir

$$P(\bar{x} \mid \bar{y}, \bar{z}) = \frac{P(\bar{y} \mid \bar{x}, \bar{z}) \cdot P(\bar{x}, \bar{z}) \,/\, P(\bar{z})}{P(\bar{y}, \bar{z}) \,/\, P(\bar{z})} = \frac{P(\bar{y} \mid \bar{x}, \bar{z}) \cdot P(\bar{x} \mid \bar{z})}{P(\bar{y} \mid \bar{z})} \tag{2.24}$$

Basta ahora aplicar la ecuación (2.10) para concluir la demostración. □

2.3.2 Aplicación del teorema de Bayes

En la práctica, el teorema de Bayes se utiliza para conocer la probabilidad a posteriori de cierta variable de interés dado un conjunto de hallazgos. Las definiciones formales son las siguientes:

▷ **Hallazgo.** Es la determinación del valor de una variable, $H = h$, a partir de un dato (una observación, una medida, etc.).

▷ **Evidencia.** Es el conjunto de todos los hallazgos disponibles en un determinado momento o situación: $\mathbf{e} = \{H_1 = h_1, \ldots, H_r = h_r\}$.

▷ **Probabilidad a priori.** Es la probabilidad de una variable o subconjunto de variables cuando no hay ningún hallazgo.

La probabilidad a priori de $\bar{X}$ coincide, por tanto, con la probabilidad marginal $P(\bar{x})$.

▷ **Probabilidad a posteriori.** Es la probabilidad de una variable o subconjunto de variables dada la evidencia $\mathbf{e}$. La representaremos mediante P^*:

$$P^*(\bar{x}) \equiv P(\bar{x} \,|\, \mathbf{e}) \tag{2.25}$$

Ejemplo 2.21 En un congreso científico regional participan 50 representantes de tres universidades: 23 de la primera, 18 de la segunda y 9 de la tercera. En la primera universidad, el 30% de los profesores se dedica a las ciencias, el 40% a la ingeniería, el 25% a las humanidades y el 5% restante a la economía. En la segunda, las proporciones son 25%, 35%, 30% y 10%, respectivamente, y en la tercera son 20%, 50%, 10%, 20%. A la salida del congreso nos encontramos con un profesor. ¿Cuál es la probabilidad de que sea de la tercera universidad? Y si nos enteramos de que su especialidad es la economía, ¿cuál es la probabilidad?

Solución. Si representamos mediante X la variable "universidad" y mediante Y la especialidad, la probabilidad a priori para cada una de las universidades es: $P(x^1) = 23/50 = 0'46$; $P(x^2) = 18/50 = 0'36$; $P(x^3) = 9/50 = 0'18$. Por tanto, la probabilidad de que el profesor pertenezca a la tercera universidad es "18%".

Para responder a la segunda pregunta, aplicamos el teorema de Bayes, teniendo en cuenta que la probabilidad de que un profesor de la universidad x sea de la especialidad y viene dada por la siguiente tabla:

$P(y \,\|\, x)$	x^1	x^2	x^3
y^c	0'30	0'25	0'20
y^i	0'40	0'35	0'50
y^h	0'25	0'30	0'10
y^e	0'05	0'10	0'20

Por tanto,

$$P^*(x^3) = P(x^3 \,|\, y^e) = \frac{P(x^3) \cdot P(y^e \,|\, x^3)}{\sum\limits_x P(x) \cdot P(y^e \,|\, x)} = \frac{0'18 \cdot 0'20}{0'46 \cdot 0'05 + 0'36 \cdot 0'10 + 0'18 \cdot 0'20} = 0'379$$

Es decir, la probabilidad de que un profesor de economía asistente al congreso pertenezca a la tercera universidad es el 37'9%. Observe que en este caso la evidencia era $\{Y = y^e\}$

(un solo hallazgo) y el "diagnóstico" buscado era la universidad a la que pertenece el profesor, representada por la variable X. Hemos escrito "diagnóstico" entre comillas porque estamos utilizando el término en sentido muy amplio, ya que aquí no hay ninguna anomalía, ni enfermedad, ni avería que diagnosticar. Propiamente, éste es un *problema de clasificación bayesiana*: se trata de averiguar la clase —en este ejemplo, la universidad— a la que pertenece cierto individuo. En realidad, los problemas de diagnóstico son sólo un caso particular de los problemas de clasificación. □

Forma normalizada del teorema de Bayes En el ejemplo anterior podríamos haber calculado la probabilidad para cada una de las tres universidades, una por una. Sin embargo, si necesitamos conocer las tres probabilidades $P(x \mid y)$, puede ser más cómodo aplicar la forma normalizada del teorema de Bayes, que es la siguiente:

$$P(x \mid y) = \alpha \cdot P(x) \cdot P(y \mid x) \tag{2.26}$$

En esta expresión,

$$\alpha \equiv \left[\sum_{x'} P(x') \cdot P(y \mid x') \right]^{-1} = [P(y)]^{-1} \tag{2.27}$$

pero en realidad no necesitamos preocuparnos de su significado, ya que podemos calcularla por normalización, como muestra el siguiente ejemplo.

Ejemplo 2.22 (Continuación del ejemplo 2.21) Para calcular la probabilidad a posteriori de cada universidad (es decir, la probabilidad sabiendo que es un profesor de economía) aplicamos la ecuación (2.26):

$$\begin{cases} P^*(x^1) = P(x^1 \mid y^e) = \alpha \cdot P(x^1) \cdot P(y^e \mid x^1) = \alpha \cdot 0'46 \cdot 0'05 = 0'023\alpha \\ P^*(x^2) = P(x^2 \mid y^e) = \alpha \cdot P(x^2) \cdot P(y^e \mid x^2) = \alpha \cdot 0'36 \cdot 0'10 = 0'036\alpha \\ P^*(x^3) = P(x^3 \mid y^e) = \alpha \cdot P(x^3) \cdot P(y^e \mid x^3) = \alpha \cdot 0'18 \cdot 0'20 = 0'036\alpha \end{cases}$$

Recordando que las probabilidades han de sumar la unidad, tenemos que

$$P(x^1 \mid y^e) + P(x^2 \mid y^e) + P(x^3 \mid y^e) = 0'023\alpha + 0'036\alpha + 0'036\alpha = 0'095\alpha = 1$$

de donde se deduce que $\alpha = 0'095^{-1} = 10'526$ y, por tanto,

$$\begin{cases} P^*(x^1) = 0'242 \\ P^*(x^2) = 0'379 \\ P^*(x^3) = 0'379 \end{cases}$$

Observe que la probabilidad a posteriori $P^*(x) = P(x \mid y)$ depende de dos factores: de la *probabilidad a priori* de que el profesor pertenezca a la universidad, $P(x)$, y de la proporción de profesores de la especialidad en cuestión que hay en cada universidad, $P(y \mid x)$. A este segundo factor se le conoce como *verosimilitud* (en inglés, "*likelihood*"). En el ejemplo que acabamos de considerar, $P(x^2 \mid y^e) = P(x^3 \mid y^e)$, pues, por un lado, la probabilidad a priori de la segunda universidad es el doble que el de la tercera, pero, por otro, la verosimilitud de que un profesor de economía pertenezca a la tercera es el doble que para la segunda (porque en la segunda hay un 10% de profesores de economía mientras que en la tercera hay un 20%) de modo que lo uno compensa lo otro. Vamos a insistir sobre la ponderación de probabilidad a priori y verosimilitud en el próximo apartado.

Forma racional del teorema de Bayes

Supongamos que queremos comparar la probabilidad a posteriori de dos diagnósticos, x^i y x^j En este caso, tenemos que

$$\frac{P(x^i \mid y)}{P(x^j \mid y)} = \frac{\alpha \cdot P(x^i) \cdot P(y \mid x^i)}{\alpha \cdot P(x^j) \cdot P(y \mid x^j)} = \frac{P(x^i)}{P(x^j)} \cdot \frac{P(y \mid x^i)}{P(y \mid x^j)} \tag{2.28}$$

El término $P(x^i)/P(x^j)$ se conoce como *razón de probabilidad* (en inglés, "*odds ratio*"), mientras que $P(y \mid x^i)/P(y \mid x^j)$ se denomina *razón de verosimilitud* ("*likelihood ratio*").

Ejemplo 2.23 En el ejemplo 2.21 se observa que

$$\frac{P(x^1 \mid y)}{P(x^2 \mid y)} = \frac{P(x^1)}{P(x^2)} \cdot \frac{P(y^e \mid x^1)}{P(y^e \mid x^2)} = \frac{0'46}{0'36} \cdot \frac{0'05}{0'10} = 1'278 \cdot 0'5 = 0'639$$

En efecto, $0'242/0'379 = 0'639$. Del mismo modo

$$\frac{P(x^2 \mid y)}{P(x^3 \mid y)} = \frac{P(x^2)}{P(x^3)} \cdot \frac{P(y^e \mid x^2)}{P(y^e \mid x^3)} = \frac{0'36}{0'18} \cdot \frac{0'10}{0'20} = 2 \cdot \frac{1}{2} = 1$$

Tal como decíamos en el apartado anterior, la probabilidad a posteriori es la misma para ambos valores, pues la razón de probabilidades a priori favorece a x^2 frente a x^3, mientras que la razón de verosimilitud favorece a x^3 frente a x^2 en la misma medida, por lo que ambos efectos se compensan, dando lugar a un "empate". □

Observe que, para variables no binarias, la forma racional del teorema de Bayes permite conocer la razón de probabilidad a posteriori entre dos valores, pero no sus valores concretos. Sin embargo, en el caso de una variable binaria, podemos calcular la probabilidad de cada valor a partir de su razón de probabilidad. Concretamente, cuando X toma los valores $+x$ y $\neg x$, suelen aplicarse las siguientes definiciones:

- Razón de probabilidad de X a priori

$$RP_{pre}(X) \equiv \frac{P(+x)}{P(\neg x)} = \frac{P(+x)}{1 - P(+x)} \tag{2.29}$$

- Razón de probabilidad de X a posteriori

$$RP_{post}(X) \equiv \frac{P(+x \mid y)}{P(\neg x \mid y)} = \frac{P(+x \mid y)}{1 - P(+x \mid y)} \tag{2.30}$$

- Razón de verosimilitud para X dado y

$$RV_X(y) \equiv \frac{P(y \mid +x)}{P(y \mid \neg x)} \tag{2.31}$$

A partir de la ecuación (2.30) podemos hallar $P(+x \mid y)$:

$$P(+x \mid y) = \frac{RP_{post}(X)}{1 + RP_{post}(X)} \tag{2.32}$$

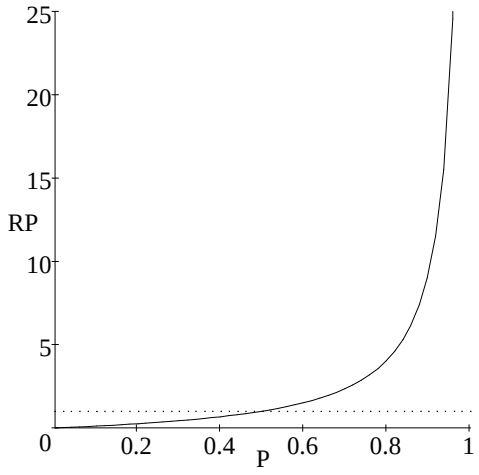

Figura 2.9: La razón de probabilidad $RP(X)$ como función de la probabilidad $P(+x)$.

La figura 2.9 representa la razón de probabilidad como función de la probabilidad. Se observa que cuando $P(+x) = 0$, $RP(X) = 0$; cuando $P(+x) < P(\neg x)$ (es decir, cuando $P(+x) < 0'5$), $RP(X) < 1$; cuando $P(+x) = P(\neg x) = 0'5$, $RP(X) = 1$; cuando $P(+x) > P(\neg x)$, $RP(X) > 1$; y, finalmente, cuando $P(+x) \to 1$, $RP(X) \to \infty$.

Con las definiciones anteriores, la ecuación (2.28) puede expresarse como

$$RP_{post}(X) = RP_{pre}(X) \cdot RV_X(y) \tag{2.33}$$

y una vez conocida $RP_{post}(X)$ se obtiene $P(+x \,|\, y)$ a partir de la ecuación (2.32).

Ejemplo 2.24 Supongamos que tenemos una enfermedad X que puede estar presente $(+x)$ o ausente $(\neg x)$, y un síntoma asociado Y que puede ser leve (y^l), moderado (y^m) o severo (y^s), aunque la mayor parte de la población no presenta el síntoma (y^a). Un estudio epidemiológico realizado con 10.000 personas ha dado la siguiente tabla:

N	$+x$	$\neg x$
y^a	50	8.500
y^l	80	1.000
y^m	100	150
y^s	70	50
Total	300	9.700

(2.34)

y nos piden que calculemos mediante la ecuación (2.33) la probabilidad de tener la enfermedad en cada caso. Para ello, debemos empezar calculando la razón de probabilidad a priori de X: $RP_{pre}(X) = 300/9.700 = 0'0309$. Si el síntoma está ausente,[2]

$$RV_X(y^a) \equiv \frac{P(y^a \mid +x)}{P(y^a \mid \neg x)} = \frac{N(+x, y^a)/N(+x)}{N(\neg x, y^a)/N(\neg x)} = \frac{0'1667}{0'8763} = 0'1902$$

[2]En realidad, estamos utilizando la tabla de frecuencias para obtener el *valor de máxima verosimilitud* de la probabilidad, pero esta es una cuestión de inferencia estadística en la que no vamos a entrar.

de modo que $RP_{post}(X) = 0'0309 \cdot 0'1902 = 0'0059$ y

$$P(+x \mid y^a) = \frac{RP_{post}(X)}{1 + RP_{post}(X)} = \frac{0'0059}{1 + 0'0059} = 0'0058 \tag{2.35}$$

Del mismo modo se calcula que $RV_X(y^l) = 2'587$, $RP_{post}(X) = 0'0800$ y $P(+x|y^l) = 0'0741$; $RV_X(y^m) = 21'5556$, $RP_{post}(X) = 0'6667$ y $P(+x|y^m) = 0'4000$; finalmente, $RV_X(y^s) = 45'2667$, $RP_{post}(X) = 1'4000$ y $P(+x|y^s) = 0'5833$. □

Sensibilidad, especificidad, prevalencia y valores predictivos En medicina, cuando tenemos una enfermedad X que puede estar presente $(+x)$ o ausente $(\neg x)$ y un hallazgo Y asociado tal enfermedad —por ejemplo, un síntoma o un signo que puede estar presente $(+y)$ o ausente $(\neg y)$, o una prueba de laboratorio que puede dar positiva $(+y)$ o negativa $(\neg y)$— es habitual emplear las siguientes definiciones:

Prevalencia	$P(+x)$
Sensibilidad	$P(+y \mid +x)$
Especificidad	$P(\neg y \mid \neg x)$
Valor predictivo positivo (VPP)	$P(+x \mid +y)$
Valor predictivo negativo (VPN)	$P(\neg x \mid \neg y)$

En este caso, el teorema de Bayes puede expresarse como:

$$VPP = \frac{Sens \times Prev}{Sens \times Prev + (1 - Espec) \times (1 - Prev)} \tag{2.36}$$

$$VPN = \frac{Espec \times (1 - Prev)}{(1 - Sens) \times Prev + Espec \times (1 - Prev)} \tag{2.37}$$

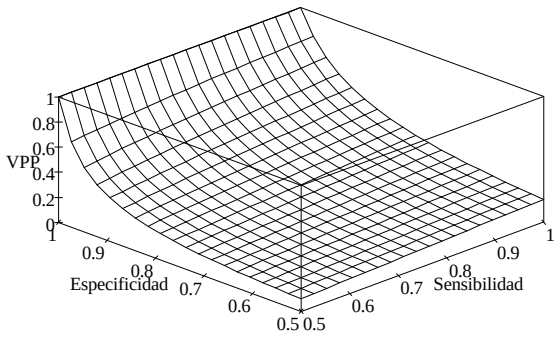

Figura 2.10: Valor predictivo positivo (prevalencia=0'1).

Se observa en estas gráficas que el valor predictivo positivo aumenta considerablemente al aumentar la especificidad; de hecho, $VPP = 1$ sólo si la especificidad vale 1; por tanto, para confirmar la presencia de una enfermedad deberemos buscar pruebas muy específicas. En cambio, el valor predictivo negativo aumenta al aumentar la sensibilidad, por lo que para descartar una enfermedad deberemos buscar síntomas o signos muy sensibles.

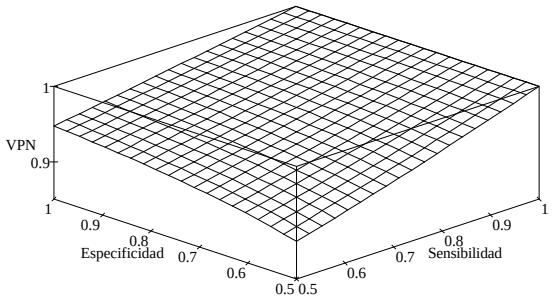

Figura 2.11: Valor predictivo negativo (prevalencia=0'1).

2.4 Método probabilista clásico

Hasta ahora hemos visto cómo aplicar el teorema de Bayes cuando tenemos una variable diagnóstico X y un hallazgo Y. Sin embargo, en los problemas del mundo real existen varios diagnósticos posibles (distintas averías, enfermedades diversas, etc.), por lo que los métodos que hemos presentado hasta ahora resultan claramente insuficientes, y debemos dar un paso hacia adelante con el fin de abordar problemas más complejos.

Una forma de intentarlo es la siguiente: supongamos que tenemos un conjunto de n enfermedades o anomalías que queremos diagnosticar; cada una de ellas vendrá representada por una variable D_i; si sólo queremos diagnosticar la presencia o ausencia de la anomalía, se tomará una variable binaria, con valores $+d_i$ y $\neg d_i$; si queremos precisar más, por ejemplo, señalando el grado de D_i, tomará varios valores d_i^k. Los m hallazgos posibles vendrán representados por las variables $H_1, \ldots, H_m$. El teorema de Bayes (ec. (2.22)) nos dice entonces que

$$P^*(d_1, \ldots, d_n) = P(d_1, \ldots, d_n \,|\, h_1, \ldots, h_m) \tag{2.38}$$

$$= \frac{P(d_1, \ldots, d_n) \cdot P(h_1, \ldots, h_m \,|\, d_1, \ldots, d_n)}{\sum_{d'_1, \ldots, d'_n} P(d'_1, \ldots, d'_n) \cdot P(h_1, \ldots, h_m \,|\, d'_1, \ldots, d'_n)} \tag{2.39}$$

Sin embargo, esta expresión es imposible de aplicar por la enorme cantidad de información que requiere: necesitaríamos conocer todas las probabilidades a priori $P(\bar{d})$ y todas las probabilidades condicionadas $P(\bar{h} \,|\, \bar{d})$. En el caso de variables binarias, habría 2^n probabilidades a priori y 2^{m+n} probabilidades condicionales, lo que significa un total de $2^{m+n} - 1$ parámetros independientes.[3] Un modelo que contenga 3 diagnósticos y 10 hallazgos posibles requiere 8.191 parámetros; para 5 diagnósticos y 20 hallazgos se necesitan $33_1554.431$ parámetros, y para 10 diagnósticos y 50 hallazgos, $1_3152.921_2504.606_1846.975$ parámetros. Obviamente, este método es inaplicable, salvo para modelos extremadamente simples.

Por ello se introduce la **hipótesis** de que los diagnósticos son exclusivos (no puede haber dos de ellos a la vez) y exhaustivos (no hay otros diagnósticos posibles). Esto permite que en vez de tener n variables D_i tengamos una sola variable, D, que toma n valores d^i (los

[3] El número de parámetros independientes es el número total de parámetros menos el número de ligaduras. En este caso, además de la ligadura $\sum_{\bar{d}} P(\bar{d}) = 1$, hay 2^n ligaduras $\sum_{\bar{h}} P(\bar{h}|\bar{d}) = 1$, una para cada $\bar{d}$, por lo que el número de parámetros independientes es $(2^n + 2^{m+n}) - (1 + 2^n) = 2^{m+n} - 1$.

n diagnósticos posibles), de modo que la probabilidad de un diagnóstico cualquiera d viene dada por

$$P^*(d) = P(d \,|\, h_1, \ldots, h_m) = \frac{P(d) \cdot P(h_1, \ldots, h_m \,|\, d)}{\sum_{d'} P(d') \cdot P(h_1, \ldots, h_m \,|\, d')} \tag{2.40}$$

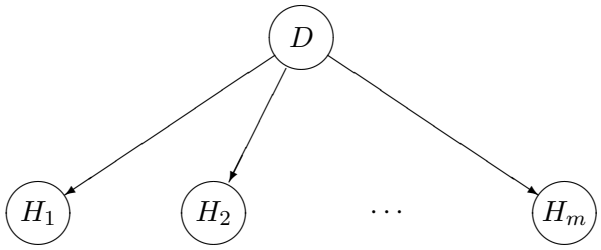

Figura 2.12: Método probabilista clásico.

Este modelo simplificado requiere n probabilidades a priori $P(d)$ y, si las variables H_j son binarias, $2^m \cdot n$ probabilidades condicionadas $P(\bar{h} \,|\, d)$, lo que significa $2^m \cdot n - 1$ parámetros independientes. Es decir, para 3 diagnósticos y 10 hallazgos harían falta 3.071 parámetros; para 5 diagnósticos y 20 hallazgos, $5_1242.879$ parámetros, y para 10 diagnósticos y 50 hallazgos, $11.258_2999.068_1426.239$. La reducción es significativa (dos órdenes de magnitud en el último caso), pero claramente insuficiente.

Por tanto, se hace necesario introducir una nueva **hipótesis**, la de *independencia condicional*: los hallazgos son condicionalmente independientes entre sí para cada diagnóstico d. En forma matemática, se expresa así:

$$P(h_1, \ldots, h_m \,|\, d) = P(h_1 \,|\, d) \cdot \ldots \cdot P(h_m \,|\, d), \qquad \forall d \tag{2.41}$$

de modo que la probabilidad resultante para cada diagnóstico d es

$$P^*(d) = \frac{P(d) \cdot P(h_1 \,|\, d) \cdot \ldots \cdot P(h_m \,|\, d)}{\sum_{d'} P(d') \cdot P(h_1 \,|\, d') \cdot \ldots \cdot P(h_m \,|\, d')} \tag{2.42}$$

o, en forma normalizada

$$P^*(d) = \alpha \cdot P(d) \cdot P(h_1 \,|\, d) \cdot \ldots \cdot P(h_m \,|\, d) \tag{2.43}$$

Observe que esta expresión es una generalización de la ecuación (2.26).

Este modelo simplificado requiere n probabilidades a priori $P(d)$ y, si las variables H_j son binarias, $2\,m \cdot n$ probabilidades condicionadas $P(h_j \,|\, d)$, lo que significa $n - 1 + m \cdot n = n \cdot (m+1) - 1$ parámetros independientes. Por tanto, para 3 diagnósticos y 10 hallazgos harían falta 32 parámetros; para 5 diagnósticos y 20 hallazgos, 104 parámetros, y para 10 diagnósticos y 50 hallazgos, 509. Con esta drástica reducción, el problema ya resulta abordable.

Ejemplo 2.25 Cierto motor puede tener una avería eléctrica (con una probabilidad de 10^{-3}) o mecánica (con una probabilidad de 10^{-5}). El hecho de que se produzca un tipo de avería no hace que se produzca una del otro tipo. Cuando hay avería eléctrica se enciende un piloto luminoso el 95% de las veces; cuando hay avería mecánica, el 99% de las veces; y cuando no hay avería, el piloto luminoso se enciende (da una falsa alarma) en un caso por millón.

Cuando no hay avería, la temperatura está elevada en el 17% de los casos y reducida en el 3%; en el resto de los casos, está en los límites de normalidad. Cuando hay avería eléctrica, está elevada en el 90% de los casos y reducida en el 1%. Cuando hay avería mecánica, está elevada en el 10% de los casos y reducida en el 40%. El funcionamiento del piloto es independiente de la temperatura. Si se enciende el piloto y la temperatura está por debajo de su valor normal, ¿cuál es el diagnóstico del motor?

Solución. Aplicamos el método probabilista clásico**.** La afirmación "el hecho de que se produzca un tipo de avería no hace que se produzca una del otro tipo" nos permite considerarlos como dos variables independientes. Sin embargo, como hemos discutido anteriormente, esto nos obligaría a considerar un modelo con muchos más parámetros de los que nos ofrece el enunciado. Por eso introducimos la hipótesis de que los diagnósticos son exclusivos, lo cual es una aproximación razonable, ya que es sumamente improbable que se den los dos tipos de avería simultáneamente: $10^{-3} \cdot 10^{-5} = 10^{-8}$. Sin embargo, estos dos diagnósticos no son exhaustivos, porque es posible que no haya avería ni eléctrica ni mecánica. Por ello, la variable diagnóstico D ha de tomar tres valores posibles: d^e (avería eléctrica), d^m (avería mecánica) y d^n (ninguna de las dos, es decir, estado de normalidad). La probabilidad a priori para D es la siguiente: $P(d^e) = 0'001$; $P(d^m) = 0'00001$; $P(d^n) = 0'99899$.

Si representamos el estado del piloto luminoso mediante la variable L, los estados posibles son $+l$ (encendido) y $\neg l$ (apagado), y la probabilidad condicional $P(l \mid d)$ viene dada por la siguiente tabla:

$P(l \mid d)$	d^e	d^m	d^n
$+l$	0'95	0'99	0'000001
$\neg l$	0'05	0'01	0'999999

La temperatura puede venir representada por una variable T, de tres valores: t^n (normal), t^e (elevada) y t^r (reducida); la tabla de probabilidad condicional es la siguiente:

$P(t \mid d)$	d^e	d^m	d^n
t^e	0'90	0'10	0'17
t^n	0'09	0'50	0'80
t^r	0'01	0'40	0'03

La afirmación del enunciado "el funcionamiento del piloto es independiente de la temperatura" podemos interpretarla como una declaración de independencia condicional entre las variables L y T para cada diagnóstico: $P(l, t \mid d) = P(l \mid d) \cdot P(t \mid d)$. Con esto, se cumplen ya las condiciones para poder aplicar el método probabilista clásico (fig. 2.13), que en su forma normalizada nos dice que

$$P^*(d) = P(d \mid l, t) = \alpha \cdot P(d) \cdot P(l \mid d) \cdot P(t \mid d)$$

Concretamente, para la pregunta planteada en el problema

$$\begin{cases} P^*(d^e) = P(d^e \mid +l, t^r) = \alpha \cdot P(d^e) \cdot P(+l \mid d^e) \cdot P(t^r \mid d^e) \\ P^*(d^m) = P(d^m \mid +l, t^r) = \alpha \cdot P(d^m) \cdot P(+l \mid d^m) \cdot P(t^r \mid d^m) \\ P^*(d^n) = P(d^n \mid +l, t^r) = \alpha \cdot P(d^n) \cdot P(+l \mid d^n) \cdot P(t^r \mid d^n) \end{cases}$$

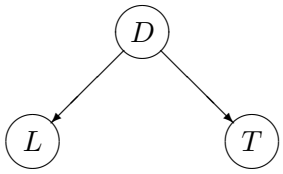

Figura 2.13: El piloto luminoso (L) y la temperatura (T) son signos de avería (D).

y, sustituyendo los valores numéricos,

$$\begin{cases} P^*(d^e) = \alpha \cdot 0'001 \cdot 0'95 \cdot 0'01 = 0'0000095\alpha = 0'70423 \\ P^*(d^m) = \alpha \cdot 0'00001 \cdot 0'99 \cdot 0'40 = 0'00000396\alpha = 0'29355 \\ P^*(d^n) = \alpha \cdot 0'99899 \cdot 0'000001 \cdot 0'03 = 0'00000002997\alpha = 0'00222 \end{cases}$$

donde el valor de α se ha calculado por normalización ($\alpha = 74.129$). En conclusión, el diagnóstico más probable es que haya avería eléctrica (70%), aunque también podría tratarse de una avería mecánica (29%). La probabilidad de que sea una falsa alarma es muy pequeña (0'22%).

2.4.1 Forma racional del método probabilista clásico

Cuando el objetivo es comparar la probabilidad relativa de dos diagnósticos, d y d', el método probabilista clásico puede expresarse en forma racional así:

$$\frac{P(d \,|\, h_1, \ldots, h_m)}{P(d' \,|\, h_1, \ldots, h_m)} = \frac{P(d)}{P(d')} \cdot \frac{P(h_1 \,|\, d)}{P(h_1 \,|\, d')} \cdot \ldots \cdot \frac{P(h_m \,|\, d)}{P(h_m \,|\, d')} \tag{2.44}$$

En el problema anterior (ejemplo 2.25), si sólo quisiéramos saber si es más probable que la avería sea eléctrica o mecánica, tendríamos

$$\begin{aligned} \frac{P(d^e \mid +l, t^r)}{P(d^m \mid +l, t^r)} &= \frac{P(d^e)}{P(d^m)} \cdot \frac{P(+l \,|\, d^e)}{P(+l \,|\, d^m)} \cdot \frac{P(t^r \,|\, d^e)}{P(t^r \,|\, d^m)} \\ &= \frac{0'001}{0'00001} \cdot \frac{0'95}{0'99} \cdot \frac{0'01}{0'40} = 100 \cdot 0'96 \cdot \frac{1}{40} = 2'40 \end{aligned}$$

Esto nos permite comprobar que el hallazgo $+l$ casi no influye en el diagnóstico, pues el valor de $P(+l \,|\, d^e)/P(+l \,|\, d^m)$ es casi la unidad; en cambio, el hallazgo t^r aporta evidencia a favor de d^m frente a d^e, pues es 40 veces más verosímil para d^m que para d^e. A pesar de eso, prevalece el diagnóstico d^e, porque su probabilidad a priori era 100 veces mayor que la de d^m.

En el caso de que D sea una variable binaria que representa la presencia ($+d$) o ausencia ($\neg d$) de una anomalía, podemos utilizar las definiciones (2.29), (2.30) y (2.31) para calcular la razón de probabilidad de D dada la evidencia $\{h_1, \ldots, h_n\}$:

$$RP_{post}(D) = RP_{pre}(D) \cdot RV_D(h_1) \cdot \ldots \cdot RV_D(h_m) \tag{2.45}$$

Esta expresión es una generalización de la (2.33) para el caso de múltiples hallazgos. A partir de $RP_{post}(D)$ se obtiene fácilmente la probabilidad posteriori mediante la ecuación (2.32).

Finalmente, conviene señalar que en el método probabilista clásico (cualquiera que sea la forma en que se exprese) sólo se han de tener en cuenta las variables-hallazgos cuyo valor se conoce; los posibles hallazgos cuyo valor no ha llegado a conocerse, deben omitirse, como si no formaran parte del modelo. Por ejemplo, si hay cuatro hallazgos posibles ($m = 4$), y en un caso particular sólo se han observado h_1 y h_4, la ecuación (2.45) queda reducida a

$$RP_{post}(D) = RP_{pre}(D) \cdot RV_D(h_1) \cdot RV_D(h_4)$$

Si más tarde se observa h_2, la nueva probabilidad a posteriori se puede calcular como

$$RP'_{post}(D) = RP_{post}(D) \cdot RV_D(h_2) = RP_{pre}(D) \cdot RV_D(h_1) \cdot RV_D(h_4) \cdot RV_D(h_2)$$

Como era de esperar, el orden en que se introducen los hallazgos no influye en el resultado final.

2.4.2 Paso de mensajes en el método probabilista clásico

Por razones que quedarán claras en el próximo capítulo, definimos unos vectores $\lambda_{H_j}(d)$ y $\lambda(d)$ como sigue:

$$\lambda_{H_j}(d) \equiv P(h_j \,|\, d) \tag{2.46}$$

$$\lambda(d) \equiv \prod_{j=1}^{m} \lambda_{H_j}(d) \tag{2.47}$$

de modo que la ecuación (2.43) puede escribirse así:

$$P^*(d) = \alpha \cdot P(d) \cdot \lambda_{H_1}(d) \cdot \ldots \cdot \lambda_{H_m}(d) \tag{2.48}$$

$$= \alpha \cdot P(d) \cdot \lambda(d) \tag{2.49}$$

Cada $\lambda_{H_j}(d)$ puede interpretarse como un mensaje que el hallazgo H_j envía al nodo-diagnóstico D, de modo que éste pueda calcular su probabilidad a posteriori $P^*(d)$ como función de su probabilidad a priori $P(d)$ y de los mensajes que recibe de cada uno de sus nodos hijos, tal como refleja la figura 2.14; el efecto de todos los mensajes $\lambda_{H_j}(d)$ se puede agrupar en un único mensaje $\lambda(d)$, tal como indica la ecuación (2.47).

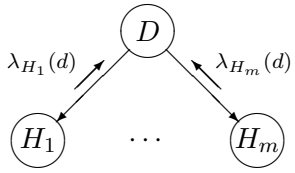

Figura 2.14: Paso de mensajes en el método probabilista clásico.

Ejemplo 2.26 Para el problema del ejemplo 2.25 tendríamos

$$\begin{aligned} P(d) &= (0'001\ 0'00001\ 0'99899) \\ \lambda_L(d) &= (0'95\ 0'99\ 0'00001) \\ \lambda_T(d) &= (0'01\ 0'40\ 0'03) \\ \lambda(d) &= (0'0095\ 0'396\ 0'0000003) \\ P(d)\cdot\lambda(d) &= (0'0000095\ 0'00000396\ 0'00000003) \\ P^*(d) &= (0'7032\ 0'2931\ 0'00221) \end{aligned}$$

□

Tan sólo conviene aclarar un pequeño detalle. En la ecuación (2.48) sólo deberían intervenir aquellas $\lambda_{H_j}(d)$ correspondientes a los hallazgos disponibles. Sin embargo, para cada nodo H_k cuyo valor se desconoce podemos considerar que $\lambda_{H_k}(d) = 1$ para todo d, con lo que H_k puede enviar también un mensaje que, por ser un vector constante, no modifica el valor de $P^*(d)$. De este modo, cada nodo H_j siempre envía un mensaje $\lambda_{H_j}(d)$, incluso cuando no se conoce qué valor toma H_j, aunque en este caso el mensaje $\lambda_{H_j}(d)$ no influye en el diagnóstico. Una vez más, las operaciones matemáticas responden a los dictados del sentido común.

En el próximo capítulo veremos que este mecanismo de paso de mensajes, convenientemente ampliado, es la base de la inferencia en redes bayesianas.

2.4.3 Discusión

El desarrollo de programas de diagnóstico basados en técnicas bayesianas comenzó en los años 60. Entre los sistemas de esa década destacan el de Warner, Toronto y Veasy para el diagnóstico de cardiopatías congénitas [63], los de Gorry y Barnett [24, 25] y el programa creado por de Dombal y sus colaboradores para el diagnóstico del dolor abdominal agudo [12]. Aunque estos programas dieron resultados satisfactorios, el método probabilista clásico fue duramente criticado, por los motivos siguientes:

1. La *hipótesis de diagnósticos exclusivos y exhaustivos* es pocas veces aplicable en casos reales [58]. En la mayor parte de los problemas de diagnóstico médico, por ejemplo, pueden darse dos enfermedades simultáneamente, con lo que el método clásico resulta totalmente inadecuado. Por otra parte, suele ser muy difícil o imposible especificar todas las causas que pueden producir un conjunto de hallazgos.

2. Igualmente, la *hipótesis de independencia condicional*, tal como se introduce en el método clásico, es muy cuestionable [59]. Normalmente, los hallazgos correspondientes a cada diagnóstico están fuertemente correlacionados, por lo que dicha hipótesis resulta inadmisible, pues lleva a sobreestimar la importancia de los hallazgos asociados entre sí. (Más adelante veremos que las redes bayesianas resuelven este problema introduciendo causas intermedias, con lo que la hipótesis de independencia condicional resulta mucho más razonable.)

3. Además, sigue existiendo el problema de la *gran cantidad de parámetros* necesarios en el modelo, incluso después de introducir las dos hipótesis anteriores. Como hemos explicado ya, el modelo requiere $n\cdot(m+1)-1$ parámetros independientes, lo cual significa, por ejemplo, que para 10 diagnósticos y 50 hallazgos, se necesitan 509 parámetros; es decir,

que incluso para un problema sencillo —comparado con los que se dan en la práctica clínica diaria— la construcción del modelo es bastante complicada.

4. Por último, desde el punto de vista de la construcción de sistemas expertos, el método probabilista clásico presenta el inconveniente de que *la información no está estructurada*, lo cual complica el *mantenimiento* de la base de conocimientos, ya que ésta consiste exclusivamente en un montón de parámetros, por lo que es difícil incorporar al modelo nueva información.

Por todo ello, en la década de los 70 se buscaron métodos de diagnóstico alternativos, como fueron el modelo de factores de certeza de MYCIN (cap. 4) y la lógica difusa (cap. 5). Sin embargo, en la década de los 80, con el desarrollo de las redes bayesianas, la aplicación de los métodos probabilistas volvió a ocupar un papel destacado en el campo de la inteligencia artificial, como vamos a ver en el próximo capítulo.

2.5 Bibliografía recomendada

Como dijimos en la sección 1.2, el método probabilista clásico es un tema "muerto" desde el punto de vista de la investigación, por lo que apenas existen referencias bibliográficas recientes. Entre las pocas excepciones se encuentran el artículo de Peot [49], en que analiza las implicaciones geométricas del modelo, y los trabajos sobre construcción del modelos a partir de bases de datos mediante algoritmos de aprendizaje [37]. En cuanto a la bibliografía "clásica", se pueden consultar los artículos citados en la sección anterior, sobre aplicaciones médicas, y el famoso libro de Duda y Hart [21] sobre reconocimiento de patrones y visión artificial.

Una síntesis de las críticas que se plantearon al método probabilista clásico desde el punto de vista de la inteligencia artificial se encuentra en [38].

Capítulo 3

Redes bayesianas

En este capítulo vamos a estudiar las redes bayesianas, desde su definición formal (sec. 3.2) hasta los algoritmos de propagación, tanto para poliárboles (sec. 3.3) como para la puerta OR (sec. 3.4). Dado que estas definiciones y algoritmos son difíciles de entender al principio, hasta que el lector se ha familiarizado con ellos, hemos incluido antes un ejemplo médico, que va creciendo en grado de complejidad; su finalidad es mostrar al lector la conexión entre las propiedades formales de las redes bayesianas y el razonamiento de sentido.

3.1 Presentación intuitiva

Antes de presentar formalmente la teoría matemática de las redes bayesianas, intentaremos explicar mediante un ejemplo sencillo, tomado del campo de la medicina, el significado intuitivo de las definiciones y axiomas que luego introduciremos, para mostrar la conexión entre las redes bayesianas y el razonamiento de sentido común. En el ejemplo que vamos a discutir hemos buscado sobre todo una aproximación cualitativa, sin pretender que los factores numéricos sean exactos.

En una red bayesiana, cada nodo corresponde a una variable aleatoria, tal como la edad o el sexo de un paciente, el padecer cierta enfermedad, la presencia de un síntoma o el resultado de una prueba de laboratorio. De aquí en adelante hablaremos indistintamente de nodos y variables, y los denotaremos con letras mayúsculas, tales como X.

Ejemplo 3.1 La red bayesiana no trivial más simple que podemos imaginar consta de dos variables, que llamaremos X e Y_1, y un arco desde la primera a la segunda, como indica la figura 3.1. Por el momento, baste decir que el arco indica generalmente *influencia causal*; más adelante precisaremos el sentido de esta expresión. Utilizaremos frecuentemente el término *enlace* como sinónimo de *arco*.

Por concretar el ejemplo, podemos suponer que X representa Paludismo e Y_1 representa Gota-gruesa, la prueba más habitual para determinar la presencia de dicha enfermedad.

Cuando X es una variable binaria correspondiente a una anomalía, $+x$ indica la presencia de dicha anomalía (en nuestro ejemplo significaría "el paciente tiene paludismo") y $\neg x$ indica su ausencia ("el paciente no tiene paludismo"). Si X representa un test (por ejemplo, Gota-gruesa), $+x$ indica que el test ha dado un resultado positivo y $\neg x$ un resultado negativo.

En la práctica, la información cuantitativa de una red bayesiana viene dada por la probabilidad a priori de los nodos que no tienen padres, $P(x)$, y por la probabilidad condicional

Figura 3.1: Nodo X con un hijo Y_1.

de los nodos con padres, $P(y_1|x)$. Así, en nuestro ejemplo, se supone que conocemos

$$\begin{cases} P(+x) = 0'003 \\ P(\neg x) = 0'997 \end{cases}$$

lo cual significa que el 3 por mil de la población padece paludismo y, por tanto, la probabilidad a priori de que una persona tenga la enfermedad (es decir, la probabilidad cuando no conocemos nada más sobre esa persona) es del 0'3%. En medicina, esta probabilidad a priori se conoce como *prevalencia* de la enfermedad.

También debemos conocer $P(y|x)$, que es la probabilidad condicional del efecto dado el valor de la causa:

$$\left\{ \begin{array}{ll} P(+y_1|+x) = 0'992 & P(+y_1|\neg x) = 0'0006 \\ P(\neg y_1|+x) = 0'008 & P(\neg y_1|\neg x) = 0'9994 \end{array} \right\}$$

El significado de esta probabilidad es el siguiente: cuando hay Paludismo, el test de la Gotagruesa da positivo en el 99'2% de los casos. Este valor se conoce como *sensibilidad* del test. Cuando no hay paludismo, el test da positivo (se dice entonces que ha habido un "falso positivo") en el 0'06% de los casos. La probabilidad de que el test dé negativo cuando la enfermedad buscada está ausente —en nuestro caso es el 99'94%— se llama *especificidad*. En todos los problemas de diagnóstico, no sólo en el campo de la medicina, tratamos de encontrar las pruebas que ofrezcan el grado más alto de sensibilidad y especificidad con el menor coste posible (en términos de dinero, tiempo, riesgo, etc.).[1]

Naturalmente,

$$\begin{cases} P(+y_1|+x) + P(\neg y_1|+x) = 1 \\ P(+y_1|\neg x) + P(\neg y_1|\neg x) = 1 \end{cases}$$

o, en forma abreviada,

$$\sum_{y_1} P(y_1|x) = 1, \quad \forall x \tag{3.1}$$

Conociendo la probabilidad a priori de X y la probabilidad condicional $P(Y_1|X)$, podemos calcular la probabilidad a priori de Y_1 por el teorema de la probabilidad total (ec. (2.9)):

$$\begin{cases} P(+y_1) = P(+y_1|+x) \cdot P(+x) + P(+y_1|\neg x) \cdot P(\neg x) \\ P(\neg y_1) = P(\neg y_1|+x) \cdot P(+x) + P(\neg y_1|\neg x) \cdot P(\neg x) \end{cases}$$

[1] Recordemos que esta definición de *sensibilidad* y *especificidad* es aplicable solamente a un enlace entre variables binarias de tipo presente/ausente o positivo/negativo.

que puede escribirse en forma abreviada como

$$P(y_1) = \sum_x P(y_1|x) \cdot P(x) \tag{3.2}$$

En nuestro ejemplo,

$$\begin{cases} P(+y_1) = 0'00357 \\ P(\neg y_1) = 0'99643 \end{cases}$$

Esto significa que si hacemos el test de la gota gruesa a una persona de la que no tenemos ninguna información, hay un 0'357% de probabilidad de que dé positivo y un 99'643% de que dé negativo.

Vamos a ver ahora cómo podemos calcular la probabilidad a posteriori, es decir, la probabilidad de una variable dada la evidencia observada **e**:

$$P^*(x) \equiv P(x|\mathbf{e}) \tag{3.3}$$

a) Supongamos que la gota gruesa ha dado positivo: $\mathbf{e} = \{+y_1\}$. ¿Cuál es ahora la probabilidad de que nuestro paciente tenga paludismo? Si la prueba tuviera una fiabilidad absoluta, responderíamos que el 100%. Pero como es posible que haya habido un falso positivo, buscamos $P^*(+x)$, es decir, $P(+x|+y_1)$. Para ello, aplicamos el **teorema de Bayes**:

$$P^*(+x) = P(+x|+y_1) = \frac{P(+x) \cdot P(+y_1|+x)}{P(+y_1)} = \frac{0'003 \cdot 0'992}{0'00357} = 0'83263 \tag{3.4}$$

Es decir, de acuerdo con el resultado de la prueba, hay un 83% de probabilidad de que el paciente tenga paludismo.

También podemos calcular $P^*(\neg x)$:

$$P^*(\neg x) = P(\neg x|+y_1) = \frac{P(\neg x) \cdot P(+y_1|\neg x)}{P(+y_1)} = \frac{0'997 \cdot 0'0006}{0'00357} = 0'16737 \tag{3.5}$$

Esto significa que hay un 16'7% de probabilidad de que haya habido un falso positivo. Naturalmente, se cumple que

$$P^*(+x) + P^*(\neg x) = 1 \tag{3.6}$$

La expresión general del teorema de Bayes que hemos aplicado es

$$P^*(x) = P(x|y) = \frac{P(x) \cdot P(y|x)}{P(y)} \tag{3.7}$$

Por semejanza con el método probabilista clásico (ec. (2.46)), vamos a reescribirla como

$$P^*(x) = \alpha \cdot P(x) \cdot \lambda_{Y_1}(x) \tag{3.8}$$

donde

$$\lambda_{Y_1}(x) \equiv P(\mathbf{e}|x) = P(y_1|x) \tag{3.9}$$

$$\alpha \equiv [P(\mathbf{e})]^{-1} = [P(y_1)]^{-1} \tag{3.10}$$

Vamos a repetir ahora el cálculo anterior aplicando esta reformulación del teorema de Bayes. En primer lugar tenemos que, cuando el test da positivo,

$$\mathbf{e} = \{+y_1\} \implies \begin{cases} \lambda_{Y_1}(+x) = P(+y_1|+x) = 0'992 \\ \lambda_{Y_1}(\neg x) = P(+y_1|\neg x) = 0'0006 \end{cases} \tag{3.11}$$

Esto significa que un resultado positivo en la prueba se explica mucho mejor con la enfermedad presente que con la enfermedad ausente (en la proporción de 0'992/0'0006=1.650), lo cual concuerda, naturalmente, con el sentido común.

Por tanto,

$$\begin{cases} P^*(+x) = \alpha \cdot 0'003 \cdot 0'992 = \alpha \cdot 0'00298 \\ P^*(\neg x) = \alpha \cdot 0'997 \cdot 0'0006 = \alpha \cdot 0'000598 \end{cases}$$

Podríamos calcular α a partir de su definición (3.10), pero resulta mucho más sencillo aplicar la condición de normalización (ec. (3.6)) a la expresión anterior, con lo que se llega a

$$\alpha = [0'00298 + 0'000598]^{-1}$$

y finalmente

$$\begin{cases} P^*(+x) = 0'83263 \\ P^*(\neg x) = 0'16737 \end{cases}$$

que es el mismo resultado que habíamos obtenido anteriormente por la aplicación del teorema de Bayes en su forma clásica. Como en el capítulo anterior, la ecuación (3.8) nos dice que en la probabilidad a posteriori, $P^*(x)$, influyen dos factores: la probabilidad a priori, $P(x)$, y la verosimilitud, $\lambda_{Y_1}(x)$.

b) Y si la gota gruesa diera un resultado negativo, $\mathbf{e} = \{\neg y_1\}$, ¿cuál sería la probabilidad de que el paciente tuviera paludismo? En ese caso,

$$\mathbf{e} = \{\neg y_1\} \implies \begin{cases} \lambda_{Y_1}(+x) = P(\neg y_1|+x) = 0'008 \\ \lambda_{Y_1}(\neg x) = P(\neg y_1|\neg x) = 0'9994 \end{cases} \tag{3.12}$$

Es decir, un resultado negativo en la prueba de la gota gruesa se explica mucho mejor (en la proporción de 0'9994/0'008 = 125) cuando no hay paludismo que cuando lo hay; en otras palabras, para $\neg y_1$, el valor $\neg x$ es 125 veces más verosímil que $+x$.

Aplicando la ecuación (3.8) como en el caso anterior, obtenemos

$$\begin{cases} P^*(+x) = \alpha \cdot 0'003 \cdot 0'008 = 0'000024 \\ P^*(\neg x) = \alpha \cdot 0'997 \cdot 0'9994 = 0'999976 \end{cases}$$

donde hemos calculado α por normalización.

El resultado tan bajo para $P^*(+x)$ se explica por dos razones: por un lado, la probabilidad a priori era de sólo un 0'3%; por otro, la alta especificidad de la prueba (99'94%) es un argumento convincente para descartar la enfermedad.

De nuevo comprobamos que en la probabilidad a posteriori influyen la probabilidad a priori y la verosimilitud. □

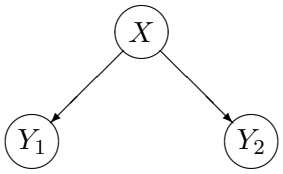

Figura 3.2: Nodo X con dos hijos.

Ejemplo 3.2 Vamos a ampliar el modelo anterior añadiendo un nuevo efecto producido por el paludismo, la Fiebre, que representaremos mediante la variable Y_2, tal como muestra la figura 3.2.

La probabilidad condicional para este segundo enlace XY_2 viene definida por

$$\left\{\begin{array}{ll} P(+y_2|+x) = 0'98 & P(+y_2|\neg x) = 0'017 \\ P(\neg y_2|+x) = 0'02 & P(\neg y_2|\neg x) = 0'983 \end{array}\right\}$$

que indica la probabilidad de que un paciente (o una persona, en general) tenga fiebre dependiendo de si tiene o no paludismo. Vemos aquí que, para el paludismo, la fiebre tiene mucha menor especificidad (98'3%) que la gota gruesa (99'94%). Así, este sencillo modelo tiene en cuenta que hay muchas otras causas que pueden producir fiebre, aunque no las incluya explícitamente.

Aplicando el teorema de la probabilidad total (ec. (2.9)) podemos calcular la probabilidad a priori de que un enfermo tenga fiebre,

$$P(+y_2) = \sum_x P(+y_2|x) \cdot P(x) = 0'01989$$

pero éste es un resultado que carece de importancia para el diagnóstico.

a) Supongamos que encontramos un paciente con fiebre, $\mathbf{e} = \{+y_2\}$, y queremos hallar la probabilidad de que tenga paludismo. En primer lugar, expresamos el teorema de Bayes en forma normalizada:

$$P^*(x) = \alpha \cdot P(x) \cdot \lambda_{Y_2}(x) \tag{3.13}$$

Ahora α vale $[P(+y_2)]^{-1}$, pero podemos prescindir de su significado y tratarla simplemente como una constante de normalización.

Para un paciente con fiebre,

$$\mathbf{e} = \{+y_2\} \implies \left\{\begin{array}{l} \lambda_{Y_2}(+x) = P(+y_2|+x) = 0'98 \\ \lambda_{Y_2}(\neg x) = P(+y_2|\neg x) = 0'017 \end{array}\right. \tag{3.14}$$

de modo que

$$\left\{\begin{array}{l} P^*(+x) = \alpha \cdot 0'003 \cdot 0'98 = 0'148 \\ P^*(\neg x) = \alpha \cdot 0'997 \cdot 0'017 = 0'852 \end{array}\right.$$

lo cual significa que hay un 14'8% de probabilidad de que el paciente tenga paludismo. Compárese con el 83'3% correspondiente a un resultado positivo de la gota gruesa (ec. (3.4)). La diferencia se debe de que esta prueba es un signo muy específico de la enfermedad, mientras que la fiebre puede estar producida por muchas otras causas.

b) Vamos a estudiar ahora el caso en que tenemos las dos observaciones y ambas indican la presencia de la enfermedad: $\mathbf{e} = \{+y_1, +y_2\}$. Al intentar calcular la probabilidad de que esa persona tenga paludismo, $P(+x|+y_1, +y_2)$ nos damos cuenta de que nos falta información, pues para aplicar el teorema de Bayes,

$$P(x|+y_1, +y_2) = \frac{P(+y_1, +y_2|\, x) \cdot P(x)}{P(+y_1, +y_2)} \tag{3.15}$$

necesitamos conocer $P(+y_1, +y_2|\, x)$ y $P(+y_1, +y_2)$.

Con la información disponible es imposible calcular estas expresiones. Por ello vamos a introducir la *hipótesis de independencia condicional*. Examinemos primero el caso en que sabemos con certeza que hay paludismo ($X = +x$). Entonces es razonable pensar que la probabilidad de que el paciente tenga o no tenga fiebre no depende de si hemos realizado el test de la gota gruesa ni del resultado que éste haya dado: la fiebre depende sólo de si hay paludismo (dando por supuesto, como parece razonable, que las demás causas de fiebre no influyen en el resultado del test). La afirmación "conociendo $X = x$, el valor de y_2 no depende del de y_1" se expresa matemáticamente como

$$P(y_2|+x, y_1) = P(y_2|+x) \tag{3.16}$$

o dicho de otro modo[2]

$$P(y_1, y_2|+x) = P(y_1|+x) \cdot P(y_2|+x) \tag{3.17}$$

Observar que estas expresiones son simétricas para Y_1 e Y_2.

Supongamos ahora que *no* hay paludismo ($X = \neg x$). La probabilidad de que el paciente presente fiebre no depende de si la gota gruesa ha dado negativo (como era de esperar) o ha dado un falso positivo por alguna extraña razón. Así tenemos

$$P(y_2|\neg x, y_1) = P(y_2|\neg x) \tag{3.18}$$

o lo que es lo mismo

$$P(y_1, y_2|\neg x) = P(y_1|\neg x) \cdot P(y_2|\neg x) \tag{3.19}$$

Uniendo las ecuaciones (3.17) y (3.19), tenemos

$$P(y_1, y_2|x) = P(y_1|x) \cdot P(y_2|x) \tag{3.20}$$

que es lo que se conoce como *independencia condicional*. Definiendo

$$\lambda(x) \equiv P(y_1, y_2|x) \tag{3.21}$$

podemos expresar dicha propiedad como

$$\lambda(x) = \lambda_{Y_1}(x) \cdot \lambda_{Y_2}(x) \tag{3.22}$$

Con esta hipótesis ya podemos calcular la probabilidad buscada. La ecuación (3.15) es equivalente a

$$P^*(x) = \alpha \cdot P(x) \cdot \lambda(x) \tag{3.23}$$

[2]Ambas expresiones son equivalentes cuando $P(+x) \neq 0$, pues

$$P(y_1, y_2|+x) = \frac{P(y_1, y_2, +x)}{P(+x)} = \frac{P(y_2|y_1, +x) \cdot P(y_1, +x)}{P(+x)} = P(y_2|+x, y_1) \cdot P(y_1|+x)$$

En nuestro ejemplo, a partir de las ecuaciones (3.11) y (3.14) tenemos

$$\mathbf{e} = \{+y_1, +y_2\} \implies \begin{cases} \lambda(+x) = 0'97216 \\ \lambda(-x) = 0'0000102 \end{cases} \tag{3.24}$$

El valor de α se calcula al normalizar, obteniendo así

$$\begin{cases} P^*(+x) = 0'99653 \\ P^*(\neg x) = 0'00347 \end{cases}$$

Naturalmente, cuando hay dos hallazgos a favor del paludismo, la probabilidad resultante (99'7%) es mucho mayor que la correspondiente a cada uno de ellos por separado (83'3% y 14'8%).

En realidad, lo que hemos hecho en este apartado no es más que aplicar el método probabilista clásico en forma normalizada (sec. 2.4); puede comprobarlo comparando las ecuaciones (3.21) y (3.22) con la (2.47) y la (2.48), respectivamente.

c) En el caso de que tengamos un hallazgo a favor y otro en contra, podemos ponderar su influencia mediante estas mismas expresiones. Por ejemplo, si hay fiebre ($+y_2$) pero hay un resultado negativo en la prueba de la gota gruesa ($\neg y_1$), las ecuaciones (3.12), (3.14) y (3.22) nos dicen que

$$\mathbf{e} = \{\neg y_1, +y_2\} \implies \begin{cases} \lambda(+x) = 0'008 \quad \cdot 0'98 \quad = 0'00784 \\ \lambda(\neg x) = 0'9994 \cdot 0'017 = 0'01699 \end{cases} \tag{3.25}$$

Vemos que hay más evidencia a favor de $\neg x$ que de $+x$ (en la proporción aproximada de 2 a 1), debido sobre todo al 0'008 correspondiente a la gota gruesa, lo cual es un reflejo de la alta sensibilidad de esta prueba (99'2%). Es decir, si hubiera paludismo, es casi seguro que lo habríamos detectado; al no haberlo detectado, tenemos una buena razón para descartarlo.

Al tener en cuenta además la probabilidad a priori de la enfermedad, nos queda finalmente

$$\begin{cases} P^*(+x) = \alpha \cdot 0'003 \cdot 0'00784 = 0'0014 \\ P^*(\neg x) = \alpha \cdot 0'997 \cdot 0'01699 = 0'9986 \end{cases}$$

Por tanto, la ponderación de la evidencia ha modificado la probabilidad desde 0'3% (valor a priori) hasta 0'14% (valor a posteriori).

De nuevo hemos aplicado el método probabilista clásico en forma normalizada.

d) Aún podemos obtener más información de este ejemplo. Imaginemos que tenemos un paciente con fiebre ($Y_2 = +y_2$) y todavía no hemos realizado la prueba de la gota gruesa. ¿Qué probabilidad hay de que ésta dé un resultado positivo o negativo? Es decir, ¿cuánto vale $P(y_1|+y_2)$?

Por la teoría elemental de la probabilidad sabemos que

$$\begin{aligned} P^*(y_1) = P(y_1|+y_2) &= \sum_x P(y_1|\, x, +y_2) \cdot P(x|+y_2) \\ &= \sum_x P(y_1|\, x, +y_2) \cdot \frac{P(x, +y_2)}{P(+y_2)} \end{aligned}$$

Aplicando la independencia condicional dada en (3.17) y definiendo[3]

$$\pi_{Y_1}(x) \equiv P(x, +y_2) = P(x) \cdot P(+y_2 | x) \tag{3.26}$$

$$\alpha \equiv [P(+y_2)]^{-1} \tag{3.27}$$

podemos reescribir la expresión anterior como

$$P^*(y_1) = \alpha \cdot \sum_x P(y_1|x) \cdot \pi_{Y_1}(x) \tag{3.28}$$

Sustituyendo los valores numéricos, tenemos que

$$\mathbf{e} = \{+y_2\} \implies \begin{cases} \pi_{Y_1}(+x) = 0'003 \cdot 0'98 &= 0'00294 \\ \pi_{Y_1}(\neg x) = 0'997 \cdot 0'017 &= 0'01695 \end{cases} \tag{3.29}$$

y finalmente

$$\begin{cases} P^*(+y_1) = 0'14715 \\ P^*(\neg y_1) = 0'85285 \end{cases} \tag{3.30}$$

Resulta muy interesante comparar la ecuación (3.28) con (3.2). Al buscar la probabilidad a priori $P(y_1)$ utilizábamos $P(x)$; ahora, al calcular $P^*(y_1)$, utilizamos $\pi_{Y_1}(x)$, que indica la probabilidad de X tras considerar la evidencia relativa a X *diferente* de Y_1.

Vemos así cómo la información que aporta el nodo Y_2 modifica la probabilidad de X, y en consecuencia también la de Y_1. El carácter simultáneamente ascendente y descendente del mecanismo de propagación es lo que nos permite utilizar la red tanto para inferencias *abductivas* (cuál es el diagnóstico que mejor explica los hallazgos) como *predictivas* (cuál es la probabilidad de obtener cierto resultado en el futuro). Un mismo nodo Y_1 puede ser fuente de información u objeto de predicción, dependiendo de cuáles sean los hallazgos disponibles y el objetivo del diagnóstico. □

Ejemplo 3.3 Consideremos una red bayesiana en que un nodo X, que representa la variable Paludismo, tiene dos padres, U_1 = País-de-origen y U_2 = Tipo-sanguíneo, dos de los factores que influyen en la probabilidad de contraer la enfermedad, tal como muestra la figura 3.3.

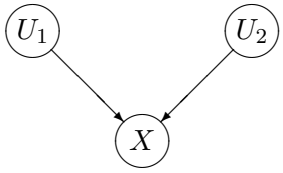

Figura 3.3: Nodo X con dos padres.

La variable U_1 podría tener muchos valores, tantos como países de origen quisiéramos considerar. Sin embargo, vamos a suponer que los agrupamos en tres zonas, de alto, medio

[3]Puede resultar extraño al lector que $\pi_{Y_1}(x)$ lleve el subíndice Y_1 a pesar de que depende del valor de la variable Y_2. El motivo es que $\pi_{Y_1}(x)$ porque recoge toda la evidencia relativa a X **diferente de** Y_1. Daremos una definición más precisa en la sección 3.3.1.

y bajo riesgo, que denotaremos por u_1^+, u_1^0 y u_1^-, respectivamente. La variable U_2 (Tipo-sanguíneo) puede tomar dos valores: $u_2^\dagger$ ó $u_2^\ddagger$.

Las probabilidades a priori para U_1 y U_2 son:

$$\begin{cases} P(u_1^+) = 0'10 \\ P(u_1^0) = 0'10 \\ P(u_1^-) = 0'80 \end{cases} \qquad \begin{cases} P(u_2^\dagger) = 0'60 \\ P(u_2^\ddagger) = 0'40 \end{cases} \tag{3.31}$$

Esto significa que la mayor parte de las personas que vamos a examinar proceden de una zona de bajo riesgo, u_1^-, y que el primer tipo sanguíneo, $u_2^\dagger$, es más frecuente que el segundo.

Las probabilidades condicionadas aparecen en la tabla 3.1. En ella vemos que, efectivamente, la zona u_1^+ es la de mayor riesgo y u_1^- la de menor. También observamos que el tipo sanguíneo $u_2^\dagger$ posee mayor inmunidad que el $u_2^\ddagger$.

$U_2 \setminus U_1$	u_1^+	u_1^0	u_1^-
$u_2^\dagger$	0'015	0'003	0'0003
$u_2^\ddagger$	0'026	0'012	0'0008

Tabla 3.1: Probabilidad de padecer paludismo, $P(+x|u_1, u_2)$.

La probabilidad de que una persona (de la que no tenemos ninguna información) padezca paludismo es

$$P(x) = \sum_{u_1,u_2} P(x|u_1, u_2) \cdot P(u_1, u_2) \tag{3.32}$$

De nuevo tenemos el problema de que no conocemos $P(u_1, u_2)$. Podemos entonces hacer la hipótesis de *independencia a priori* entre ambas variables; es decir, suponemos que los tipos sanguíneos se distribuyen por igual en las tres zonas de riesgo. Ésta es una hipótesis que habría que comprobar empíricamente. Si llegáramos a la conclusión de que existe una correlación entre ambas variables, deberíamos trazar un arco desde la una hasta la otra e introducir las correspondientes tablas de probabilidades condicionadas.

Estamos observando aquí una propiedad esencial de las RR.BB.: no sólo los arcos aportan información sobre dependencias causales, sino que también *la ausencia de un arco es una forma (implícita) de aportar información.* En nuestro caso implica que U_1 y U_2 son independientes. Matemáticamente se expresa así:

$$P(u_2|u_1) = P(u_2) \tag{3.33}$$

o bien

$$P(u_1, u_2) = P(u_1) \cdot P(u_2) \tag{3.34}$$

Con esta hipótesis podemos por fin calcular la probabilidad de X:

$$P(x) = \sum_{u_1} \sum_{u_2} P(x|u_1, u_2) \cdot P(u_1) \cdot P(u_2) \tag{3.35}$$

En nuestro caso, el valor obtenido es $P(+x) = 0'003$, que concuerda con el de los ejemplos anteriores.

a) Supongamos que nos enteramos de que la persona en cuestión procede de una zona de alto riesgo. ¿Cual es la probabilidad de que padezca la enfermedad? Una de las formas posibles de realizar el cálculo es ésta:

$$P^*(x) = P(x|u_1^+) = \frac{P(x, u_1^+)}{P(u_1^+)}$$

Si definimos

$$\pi(x) \equiv P(x, \mathbf{e}) = P(x, u_1^+) \tag{3.36}$$

$$\alpha \equiv [P(\mathbf{e})]^{-1} = [P(u_1^+)]^{-1} \tag{3.37}$$

la ecuación anterior se convierte en

$$P^*(x) = \alpha \cdot \pi(x) \tag{3.38}$$

Podemos obtener $\pi(x)$ del siguiente modo:

$$\pi(x) = \sum_{u_2} P(x, u_1^+, u_2) = \sum_{u_2} P(x|u_1^+, u_2) \cdot P(u_1^+, u_2)$$

y aplicando la independencia a priori de las causas podemos expresar la ecuación anterior como

$$\pi(x) = \sum_{u_1} \sum_{u_2} P(x|u_1, u_2) \cdot \pi_X(u_1) \cdot \pi_X(u_2) \tag{3.39}$$

que es un resultado completamente general.

En el ejemplo que estamos tratando,

$$\mathbf{e} = \{u_1^+\} \implies \left\{ \begin{array}{ll} \pi_X(u_1^+) = P(u_1^+) & \pi_X(u_2^\dagger) = P(u_2^\dagger) \\ \pi_X(u_1^0) = 0 & \pi_X(u_2^\ddagger) = P(u_2^\ddagger) \\ \pi_X(u_1^-) = 0 & \end{array} \right\} \tag{3.40}$$

y en consecuencia

$$\mathbf{e} = \{u_1^+\} \implies \left\{ \begin{array}{l} \pi(+x) = 0'00194 \\ \pi(\neg x) = 0'09806 \end{array} \right. \tag{3.41}$$

Sustituyendo este resultado en la ecuación (3.38) y normalizando (en este caso, $\alpha = 10$), hallamos que $P^*(+x) = 0'0194$; es decir, una persona originaria de una zona de alto riesgo tiene una probabilidad del 2% de padecer paludismo (frente al 0'3% general).

Las expresiones $\pi_X(u_i)$ que hemos utilizado en la deducción no son nuevas: aparecieron ya en la ecuación (3.26). Recordemos que el significado de $\pi_X(u_i)$ es que transmite a X el impacto de toda la evidencia relativa a U_i. Como no hay evidencia relativa a U_2, $\pi_X(u_2)$ coincide con la probabilidad a priori.

b) Imaginemos ahora que por alguna razón tenemos certeza absoluta de que el enfermo padece paludismo. Antes de hacer un análisis de sangre, podemos predecir qué resultado es más probable, considerando cuál de los dos tipos sanguíneos explica mejor la presencia de la enfermedad:

$$P^*(u_2) = P(u_2|+x) = \frac{P(u_2) \cdot P(+x|u_2)}{P(+x)}$$

o bien

$$P^*(u_2) = \alpha \cdot P(u_2) \cdot \lambda_X(u_2) \tag{3.42}$$

donde

$$\lambda_X(u_2) \equiv P(+x|u_2) = \sum_{u_1} P(+x|u_1, u_2) \cdot P(u_1) \tag{3.43}$$

que en nuestro ejemplo vale

$$\mathbf{e} = \{+x\} \implies \begin{cases} \lambda_X(u_2^\dagger) = 0'00204 \\ \lambda_X(u_2^\ddagger) = 0'00444 \end{cases} \tag{3.44}$$

Efectivamente, los valores de la tabla 3.1 nos han llevado a la conclusión de que el paludismo se explica mejor con el tipo sanguíneo $u_2^\ddagger$. Aplicando (3.42), obtenemos

$$\begin{cases} P^*(u_2^\dagger) = 0'408 \\ P^*(u_2^\ddagger) = 0'592 \end{cases} \tag{3.45}$$

Observamos que inicialmente el tipo $u_2^\dagger$ era el más probable (60%), pero ahora es el menos probable (40'8%) porque explica peor el paludismo.

El cálculo que hemos realizado para X y U_2 es idéntico al que hicimos en el ejemplo 3.1**.**a para Y_1 y X. Vemos de nuevo que un mismo nodo puede ser fuente de información u objeto de predicción, dependiendo de la evidencia disponible.

c) Mostraremos ahora otra de las propiedades más características de las RR.BB.: la aparición de correlaciones entre los padres de un nodo. Continuando con el caso anterior, supongamos que además de tener la certeza de que el enfermo padece paludismo sabemos que procede de un país de alto riesgo; es decir, $\mathbf{e} = \{+x, u_1^+\}$. Aplicaremos de nuevo la ecuación (3.42), aunque ahora

$$\lambda_X(u_2) \equiv P(+x, u_1^+|u_2) = P(+x|u_1^+, u_2) \cdot P(u_1^+|u_2)$$

La independencia condicional nos dice que $P(u_1^+|u_2) = P(u_1^+)$, y tenemos, por tanto,

$$\begin{aligned} \lambda_X(u_2) &= P(x|u_1^+, u_2) \cdot P(u_1^+) \\ &= \sum_{u_1} P(x|u_1, u_2) \cdot \pi_X(u_1) \end{aligned} \tag{3.46}$$

donde $\pi_X(u_1)$ es el vector que apareció en la ecuación (3.40). Al realizar los cálculos obtenemos

$$\mathbf{e} = \{+x, u_1^+\} \implies \begin{cases} \lambda_X(u_2^\dagger) = 0'0015 \\ \lambda_X(u_2^\ddagger) = 0'0026 \end{cases} \tag{3.47}$$

y de ahí

$$\begin{cases} P^*(u_2^\dagger) = 0'464 \\ P^*(u_2^\ddagger) = 0'536 \end{cases} \tag{3.48}$$

Si comparamos este resultado con el de la ecuación (3.45), observamos que la probabilidad de $u_2^\dagger$ ha aumentado del 40'8% al 46'4% como resultado de conocer la zona de origen: $U_1 = u_1^+$. Éste es el fenómeno que queríamos mostrar. *A priori*, es decir, antes de conocer el valor de x,

U_1 y U_2 eran *independientes*, por lo que la probabilidad de u_2 no variaba al conocer el valor u_1 (cf. ec. (3.33)). Sin embargo, la independencia se pierde tanto al conocer "$X = +x$" como "$X = \neg x$". Dicho de otro modo,

$$P(u_2|u_1) = P(u_2) \tag{3.49}$$

$$P(u_2|u_1, x) \neq P(u_2|x) \tag{3.50}$$

Recordando el ejemplo 3.2, vemos que allí ocurría precisamente lo contrario: las variables Y_1 e Y_2 estaban correlacionadas a priori, pero se volvían *condicionalmente independientes* al conocer el valor de X. Esta asimetría en las relaciones de independencia es un reflejo del sentido de la causalidad, es decir, de la diferencia entre causas y efectos. □

Ejemplo 3.4 Por último, consideremos el caso en que tenemos un nodo con dos causas y dos efectos (fig. 3.4). Las probabilidades condicionadas son las mismas que en los ejemplos anteriores. Por no extender demasiado esta sección, vamos a considerar solamente el caso en que tenemos un paciente que procede de una zona de alto riesgo y presenta fiebre, pero la prueba de la gota gruesa ha dado un resultado negativo. Es decir, $\mathbf{e} = \{u_1^+, \neg y_1, +y_2\}$.

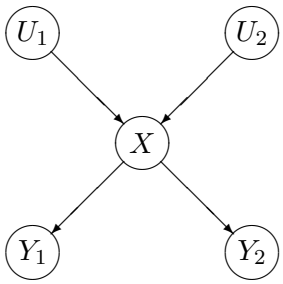

Figura 3.4: Nodo X con dos padres y dos hijos.

El teorema de Bayes nos dice que

$$P^*(x) = P(x|u_1^+, \neg y_1, +y_2) = \frac{P(x) \cdot P(u_1^+, \neg y_1, +y_2|x)}{P(u_1^+, \neg y_1, +y_2)} \tag{3.51}$$

Nuevamente necesitamos utilizar unos valores, $P(u_1^+, \neg y_1, +y_2|x)$, que no conocemos. (Si tuviéramos estos valores podríamos calcular también el denominador de la fracción.) Hemos introducido ya dos hipótesis:

1. Independencia a priori de los nodos que no tienen ningún antepasado común.
2. Independencia condicional de los dos efectos de X cuando conocemos con certeza el valor de X.

Vamos a enunciar ahora la tercera y última hipótesis, la independencia condicional (para cada valor x) entre los padres y los hijos de X:

$$P(y_1, y_2|x, u_1, u_2) = P(y_1, y_2|x) \tag{3.52}$$

o, lo que es lo mismo,

$$P(u_1, u_2, y_1, y_2|x) = P(u_1, u_2|x) \cdot P(y_1, y_2|x) \tag{3.53}$$

La interpretación de estas dos ecuaciones es clara: la probabilidad de los efectos de X depende solamente del valor que toma X, no de la combinación de factores que nos ha llevado a dicho valor. En nuestro ejemplo significa que, si hay certeza de que una persona padece paludismo, la probabilidad de que tenga fiebre y de que detectemos la enfermedad en la prueba de laboratorio no depende del país de origen ni del tipo sanguíneo. Lo mismo podemos decir de la ausencia de paludismo.[4]

De la ecuación (3.53) se deduce fácilmente, sumando sobre u_2, que

$$P(u_1, y_1, y_2|x) = P(u_1|x) \cdot P(y_1, y_2|x) \tag{3.54}$$

con lo que la ecuación (3.51) se convierte en

$$P^*(x) = \alpha \cdot \pi(x) \cdot \lambda(x) \tag{3.55}$$

Recordemos que ya habíamos definido anteriormente $\pi(x)$ y $\lambda(x)$:

$$\pi(x) \equiv P(x) \cdot P(u_1^+|x) = P(x, u_1^+) \tag{3.56}$$
$$\lambda(x) \equiv P(\neg y_1, +y_2|x) \tag{3.57}$$

y que sus valores estaban dados por (3.41) y (3.25), respectivamente. Tras unos cálculos sencillos obtenemos que $P^*(+x) = 0'0090$; es decir, con estos hallazgos, la probabilidad de que el paciente tenga paludismo es menor del 1%.

Podríamos calcular ahora la probabilidad del tipo sanguíneo en función de la evidencia, $P^*(u_2)$, pero lo vamos a omitir para no alargar más la exposición.

La ecuación (3.55) es la fórmula fundamental para el cálculo de la probabilidad en redes bayesianas. En ella aparecen dos términos importantes, $\pi(x)$ y $\lambda(x)$. El primero de ellos transmite el impacto de la evidencia correspondiente a **las causas** de X. En nuestro caso, el único hallazgo "por encima" de X era $U_1 = u_1^+$. Si no tuviéramos ninguna evidencia, $\pi(x)$ sería simplemente la probabilidad a priori $P(x)$.

El segundo, $\lambda(x)$, transmite el impacto de la evidencia correspondiente a **los efectos** de X. En el ejemplo anterior, recogía la influencia de $\neg y_1$ y $+y_2$. Si no tuviéramos ninguna evidencia, $\lambda(x)$ sería un vector constante y podríamos prescindir de él al aplicar la ecuación (3.55), sin alterar el resultado.

De las tres propiedades de independencia anteriores —ecs. (3.20), (3.34) y (3.53)— que no son más que la manifestación de la **separación direccional** (sec. 3.2.2) para esta pequeña red, se deduce que

$$P(y_1, y_2, x, u_1, u_2) = P(y_1|x) \cdot P(y_2|x) \cdot P(x|u_1, u_2) \cdot P(u_1) \cdot P(u_2) \tag{3.58}$$

Esta expresión se conoce como *factorización de la probabilidad* en una red bayesiana (cf. teorema 3.7, pág. 52). □

[4] Si por alguna razón pensáramos que esta hipótesis no es cierta, deberíamos añadir a nuestro modelo nuevos arcos con el fin de representar las influencias existentes (por ejemplo, entre el país de origen y otras causas de la fiebre) y asignarles las tablas de probabilidad oportunas.

Recapitulación

En esta sección hemos visto las propiedades más importantes de las RR.BB. En primer lugar, que la red contiene información cualitativa (la estructura del grafo) y cuantitativa (las probabilidades a priori y condicionales). Esta red constituye nuestro modelo causal y —salvo que introduzcamos algún mecanismo de aprendizaje— es invariable.

El proceso de diagnóstico consiste en introducir la evidencia disponible (asignar valores a las variables conocidas) y calcular la probabilidad a posteriori de las variables desconocidas. Se trata en realidad de un proceso de inferencia, aunque no es simbólica sino numérica.

Hemos visto además que este modelo permite tanto un razonamiento diagnóstico (cuál es la causa más probable) como predictivo (qué valor de cierta variable aparecerá con mayor probabilidad). Por otra parte, hemos comentado ya que una ventaja de las RR.BB. es que un mismo nodo puede ser fuente de información u objeto de predicción dependiendo de cuál sea la evidencia disponible, como ocurría con X o con Y_1 en los ejemplos anteriores.

Y hemos comprobado también en el ejemplo 4 que es posible realizar un cálculo incremental, modificando la probabilidad de las variables a medida que va llegando nueva evidencia, sin tener que recalcular todos los mensajes $\pi()$ y $\lambda()$.

3.2 Definición formal de red bayesiana

En la sección anterior hemos presentado de forma intuitiva qué son las redes bayesianas y cómo se propaga la evidencia, insistiendo en la importancia de las hipótesis de independencia. Ahora vamos a dar una definición matemática formal.

3.2.1 Estructura de la red. Teoría de grafos

Nuestro punto de partida consiste en un conjunto finito de *nodos* $\bar{X}$. Cada uno de ellos representa una *variable*, que puede ser discreta o continua (aunque en este texto sólo vamos a manejar variables discretas). Esta relación biunívoca entre nodos y variables nos permite emplear indistintamente ambos términos. Como vimos en el capítulo anterior, los valores de una variable deben constituir un conjunto exclusivo y exhaustivo.

Sin embargo, una diferencia importante respecto del método probabilista clásico (sec. 2.4) es que las redes bayesianas no necesitan suponer que los diagnósticos son exclusivos y exhaustivos, y por tanto no es necesario tener una variable D que represente todos los posibles diagnósticos; por ejemplo, en vez de una variable llamada D=Enfermedad, cuyos valores representasen los posibles diagnósticos correspondientes a la fiebre: neumonía, amigdalitis, paludismo, etc., en la red bayesiana tendríamos una variable Neumonía —que puede tomar dos valores (neumonía-presente y neumonía-ausente) o más de dos valores (neumonía-ausente, neumonía-leve, neumonía-moderada y neumonía-severa), dependiendo del grado de precisión que necesitemos en el diagnóstico—, otra variable Amigdalitis, Paludismo, etc. De este modo, la red bayesiana puede ofrecer dos o más diagnósticos a la vez (por ejemplo, amigdalitis-severa y neumonía-leve), lo cual era imposible con el método probabilista clásico.[5]

Introducimos a continuación algunas definiciones básicas sobre grafos:

[5]Las redes de semejanza de Heckerman [27] consituyen una notable excepción, pues en cada una de ellas hay un nodo principal, que representa los diagnósticos (supuestamente exclusivos y exhaustivos). Aunque en esto coinciden con el método probabilista clásico, se diferencian de él en que permiten que el nodo principal tenga padres y que los hijos puedan tener hijos a su vez, e incluso que haya bucles en la red.

▷ **Arco.** Es un par ordenado de nodos (X, Y).

Esta definición de arco corresponde a lo que en otros lugares se denomina *arco dirigido*. En la representación gráfica, un arco (X, Y) viene dado por una flecha desde X hasta Y, tal como muestran las figuras de los ejemplos anteriores.

▷ **Grafo dirigido.** Es un par $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ donde $\mathcal{N}$ es un conjunto de nodos y $\mathcal{A}$ un conjunto de arcos definidos sobre los nodos.

Si hubiéramos definido los arcos como pares no ordenados, tendríamos un grafo no dirigido.[6]

En el contexto de los **grafos dirigidos**, tenemos las siguientes definiciones:

▷ **Padre.** X es un *padre* de Y si y sólo si existe un arco (X, Y).

Los padres de X se representan como $pa(X)$. Por semejanza con el convenio utilizado para variables y sus valores, $pa(x)$ representará el vector formado al asignar un valor a cada nodo del conjunto $pa(X)$.

▷ **Hijo.** Y es un *hijo* de X si y sólo si existe un arco (X, Y).

▷ **Antepasado.** X es un *antepasado* de Z si y sólo si existe (al menos) un nodo Y tal que X es padre de Y e Y es antepasado de Z.

▷ **Descendiente.** Z es un *descendiente* de X si y sólo si X es un antepasado de Z.

▷ **Familia X.** Es el conjunto formado por X y los padres de X, $pa(X)$.

▷ **Nodo terminal.** Es el nodo que no tiene hijos.

Ejemplo 3.5 En la figura 3.5, los padres de D son A y B: $pa(D) = \{A, B\}$. Los hijos de D son G y H. Los antepasados de G son A, B y D. Los descendientes de A son D, G y H. Las nueve familias (tantas como nodos) son $\{A\}$, $\{B\}$, $\{C\}$, $\{D, A, B\}$, $\{E, C\}$, $\{F, C\}$, $\{G, D\}$, $\{H, D, E\}$ e $\{I, E\}$.

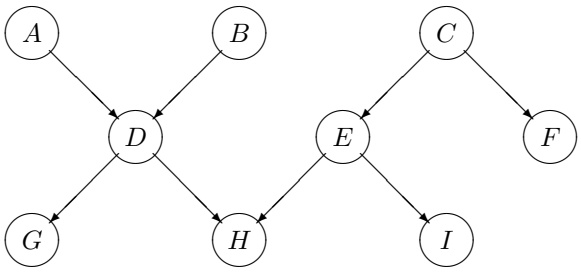

Figura 3.5: Un pequeño poliárbol.

[6] Las redes de Markov se basan en grafos no dirigidos, mientras que las redes bayesianas corresponden a grafos dirigidos.

▷ **Camino.** Un *camino* entre X_1 y X_N en una sucesión de nodos $\{X_1, \ldots, X_N\}$ pertenecientes a un grafo $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, tal que $X_i \neq X_j$ para $1 \leq i < j \leq N$ y

$$(X_i, X_{i+1}) \in \mathcal{A} \quad \text{ó} \quad (X_{i+1}, X_i) \in \mathcal{A}, \qquad \forall i, 1 \leq i < N$$

Es decir, dos nodos consecutivos de un camino —X_i y X_{i+1}— están unidos por un arco del primero al segundo o viceversa. Observe que esta definición corresponde a lo que en otros lugares se conoce como *camino abierto*.

▷ **Ciclo.** Es una sucesión de nodos $\{X_1, \ldots, X_N\}$ pertenecientes a un grafo $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, tal que *(1)* $X_i \neq X_j$ para $1 \leq i < j \leq N$, *(2)* para todo $i < N$ existe en $\mathcal{A}$ un arco (X_i, X_{i+1}), y *(3)* existe además un arco (X_N, X_1).

▷ **Bucle.** Sucesión de nodos $\{X_1, \ldots, X_N\}$ pertenecientes a un grafo $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, tal que *(1)* $X_i \neq X_j$ para $1 \leq i < j \leq N$, *(2)* para todo $i < N$ existe en $\mathcal{A}$ un arco (X_i, X_{i+1}) ó (X_{i+1}, X_i), *(3)* existe además un arco (X_N, X_1) ó (X_1, X_N) y *(4)* los arcos no forman un ciclo.

▷ **Grafo acíclico.** Es el grafo en que no hay ciclos.

Tanto el ciclo como el bucle corresponden a lo que a veces se denominan *caminos cerrados simples*. La diferencia es que en un ciclo los arcos van de cada nodo al siguiente (nunca a la inversa), mientras que la definición de bucle permite que los arcos tengan cualquiera de los dos sentidos, con la única condición de que no formen un ciclo. La distinción entre ambos es muy importante para el tema que nos ocupa, pues las redes bayesianas se definen a partir de los **grafos dirigidos acíclicos**, lo cual permite que contengan bucles pero no que contengan ciclos.

Ejemplo 3.6 En la figura 3.6.*a*, vemos que entre B y C hay dos *caminos*: $\{B, A, C\}$ y $\{B, D, C\}$, y lo mismo ocurre en 3.6.*b* y 3.6.*c*. El primero de estos tres grafos es un *ciclo*, mientras que que los dos últimos son *bucles*. Por eso estos dos últimos podrían servir para definir redes bayesianas, pero el primero no. □

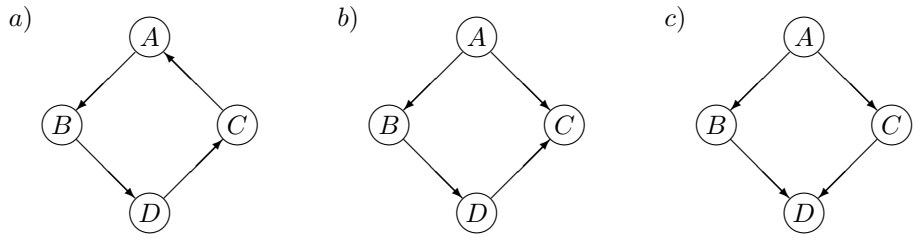

Figura 3.6: Un ciclo y dos bucles.

▷ **Grafo conexo.** Un grafo es *conexo* si entre dos cualesquiera de sus nodos hay al menos un camino.

Por tanto, un grafo no conexo es aquél que está formado por dos o más partes inconexas entre sí. Todo grafo conexo ha de pertenecer a una de las dos categorías siguientes:

▷ **Grafo simplemente conexo o poliárbol.** Un grafo es *simplemente conexo* si entre dos cualesquiera de sus nodos hay exactamente un camino.

▷ **Grafo múltiplemente conexo.** Es el que contiene ciclos o bucles.

▷ **Árbol.** Es un caso particular de poliárbol, en que cada nodo tiene un sólo padre, excepto el *nodo raíz*, que no tiene padres.

Por ejemplo, el grafo de la figura 3.5 es un poliárbol, porque no contiene bucles; no es un árbol porque algunos de sus nodos (D y H) tienen más de un padre.

3.2.2 Definición de red bayesiana

La propiedad fundamental de una red bayesiana es la *separación direccional* (llamada *d-separation* por Pearl [44, 45]), que se define así:

▷ **Separación direccional.** Dado un grafo dirigido acíclico conexo y una distribución de probabilidad sobre sus variables, se dice que hay *separación direccional* si, dado un nodo X, el conjunto de sus padres, $pa(X)$, separa condicionalmente este nodo de todo otro subconjunto $\bar{Y}$ en que no haya descendientes de X. Es decir,

$$P(x|pa(x), \bar{y}) = P(x|pa(x)) \tag{3.59}$$

Es habitual definir las redes bayesianas a partir de grafos dirigidos acíclicos (en inglés se suelen denominar *directed acyclic graph, DAG*, aunque lo correcto es decir *acyclic directed graph, ADG*). Sin embargo, nos parece importante incluir la especificación "*conexo*" por tres razones. La primera, porque muchos de los algoritmos y propiedades de las redes bayesianas sólo son correctos para grafos conexos, por lo que es mejor incluir esta característica en la definición que tener que añadirla como nota a pie de página en casos particulares. La segunda razón es que, aun en el caso de que tuviéramos un modelo con dos partes inconexas, podríamos tratarlo como dos redes bayesianas independientes. Y la tercera, porque los modelos del mundo real con que vamos a trabajar son siempre conexos; si hubiera dos partes inconexas no tendríamos uno sino dos modelos independientes.

La definición de separación direccional, aunque pueda parecer extraña a primera vista, es sencilla, y ya fue introducida en los ejemplos de la sección 3.1. En efecto, volviendo a la figura 3.4 de dicha sección (pág. 46), recordamos que, una vez conocido el valor de x, podíamos calcular la probabilidad de y_1 sin que influyeran los valores de las demás variables. Es decir, el conjunto $pa(Y_1) = \{X\}$, separa condicionalmente Y_1 de todas las demás variables de la red.

A partir de las definiciones anteriores, podemos caracterizar las redes bayesianas así:

▷ **Red bayesiana.** Es un grafo dirigido acíclico conexo más una distribución de probabilidad sobre sus variables, que cumple la propiedad de separación direccional.

El término *direccional* hace referencia a la asimetría de dicha propiedad, que se manifiesta en las siguientes propiedades de las redes bayesianas, ilustradas con el ejemplo de la figura 3.5:

1. Si A no tiene padres, entonces $P(x|pa(x)) = P(x|\varnothing) = P(x)$, y la ecuación (3.59) se traduce en $P(e|a) = P(e)$ para cada nodo E que no sea uno de los descendientes de A; en otras palabras, E es a priori independiente de A. En consecuencia, dos nodos cualesquiera D y E que no tengan ningún antepasado común son independientes a priori.

2. Si D es descendiente de A y antepasado de H, y no existe ningún otro camino desde A hasta H, entonces estos dos nodos quedan condicionalmente separados por D:

$$P(h|d, a) = P(h|d) \tag{3.60}$$

3. Si tanto G como H son hijos de D y no tienen ningún otro antepasado común, este último separa G y H, haciendo que sean condicionalmente independientes:

$$P(g|d, h) = P(g|d) \tag{3.61}$$

En general, la independencia (a priori o condicional) de dos nodos –por ejemplo, A y E— se pierde al conocer el valor de cualquiera de sus descendientes comunes —H es descendiente tanto de A como de E— pues en este caso la propiedad de separación direccional ya no es aplicable. Es muy importante que observe la relación de estas propiedades con la discusión de la sección 2.2.3.

3.2.3 Factorización de la probabilidad

En la definición de red bayesiana, hemos partido de una distribución de probabilidad conjunta para las variables, $P(\bar{x})$. Aparentemente, en el caso de variables binarias, harían falta $2^N - 1$ parámetros. (Serían 2^N si no existiera la ligadura (2.1).) Sin embargo, las condiciones de independencia dadas por la separación direccional imponen nuevas restricciones, que reducen los grados de libertad del sistema. De hecho, una de las propiedades más importantes de una red bayesiana es que su distribución de probabilidad puede expresarse mediante el producto de las distribuciones condicionadas de cada nodo dados sus padres, tal como nos dice el siguiente teorema. (Recordemos que, para un nodo X sin padres, $pa(X) = \varnothing$ y, por tanto, $P(x|pa(x)) = P(x)$; es decir, la probabilidad condicionada de un nodo sin padres es simplemente la probabilidad a priori.)

Teorema 3.7 (Factorización de la probabilidad) Dada una red bayesiana, su distribución de probabilidad puede expresarse como

$$P(x_1, \ldots, x_n) = \prod_i P(x_i|pa(x_i)) \tag{3.62}$$

Demostración. Es fácil construir una ordenación de las variables en que los padres de cada nodo aparezcan siempre después de él. Supongamos, sin pérdida de generalidad, que la ordenación $\{X_1, \ldots, X_n\}$ cumple dicha propiedad. Por la proposición 2.10 (ec. (2.14)), podemos escribir

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|x_{i+1}, \ldots, x_n)$$

Ahora bien, por la forma en que hemos escogido la ordenación, el conjunto $\{X_{i+1}, \ldots, X_n\}$ incluye todos los padres de X y, en consecuencia, la separación direccional (ec. (3.59)) nos dice que

$$P(x_i|x_{i+1}, \ldots, x_n) = P(x_i|pa(x_i))$$

con lo que concluimos la demostración. □

Ejemplo 3.8 Para la red bayesiana de la figura 3.4 (pág. 46), la factorización de la probabilidad viene dada por la ecuación (3.58).

Ejemplo 3.9 Para el grafo de la figura 3.5 (pág. 49), la factorización de la probabilidad viene dada por

$$\begin{aligned} P(a,b,c,d,e,f,g,h,i) = P(a) \cdot P(b) \cdot P(c) \cdot P(d|a,b) \cdot P(e|c) \\ \cdot P(f|c) \cdot P(g|d) \cdot P(h|d,e) \cdot P(i|e) \end{aligned}$$

Podemos comprobar que cada uno de estos factores corresponde a una de las familias enumeradas en el ejemplo 3.5.

La importancia de este teorema es que nos permite describir una red bayesiana a partir de la probabilidad condicionada de cada nodo, en vez de dar la distribución de probabilidad conjunta, que requeriría un número de parámetros exponencial en el número de nodos y plantearía el grave problema de verificar la propiedad de separación direccional; sin embargo, el número de parámetros requerido para dar las probabilidades condicionadas es proporcional al número de nodos (suponiendo que el número de padres y el número de valores posibles están acotados para cada variable).

Podríamos haber definido las propiedades de independencia en términos de caminos activados o bloqueados, al estilo de Pearl, Geiger y Verma [45, págs. 317-318], [47], seguido también por Charniak [7]. En cambio, la presentación que hemos escogido se parece más a la propuesta por Neapolitan [41, cap. 5].

3.2.4 Semántica de las redes bayesianas

Hemos definido ya las redes bayesianas desde un punto de vista matemático formal. La cuestión que nos planteamos ahora es su semántica, es decir, ¿qué interpretación se le puede dar a una red bayesiana? ¿Cómo se corresponde nuestro modelo con el mundo real? ¿Por qué podemos hablar de causas y efectos en una R.B.?

Esta cuestión está ya parcialmente respondida en la sección 3.1, que fue introducida antes de la definición formal de R.B. precisamente para mostrar que los conceptos y axiomas introducidos no pareciesen arbitrarios, sino que responden a las propiedades de la causalidad, según nuestra concepción intuitiva del mundo real.

Es importante señalar que la estructura de la red, por sí misma, aporta gran cantidad de información cualitativa. En efecto, un arco XY indica, ya antes de conocer el valor concreto de la probabilidad condicional, que hay una correlación entre ambas variables: el valor que toma X influye sobre la probabilidad de Y, y viceversa. Es lo que llamamos *influencia causal directa*. Tal es la relación que existe, por ejemplo, entre el país de origen y el paludismo, o entre el paludismo y la fiebre. Profundizando un poco más, observamos que la existencia de un camino entre dos variables X e Y, con variables intermedias $\bar{Z}$, indica que hay una *influencia causal indirecta* entre ambas.

Tal como hemos discutido en la presentación intuitiva de las RR.BB., cuando nuestro sentido común, basado en la experiencia, nos dice que la influencia de una variable X sobre uno de sus efectos Y_1 (por ejemplo, del paludismo sobre la prueba de la gota gruesa) no depende

de cuáles han sido las causas o mecanismos que han producido X, ni depende tampoco de si X a dado lugar a otros efectos, entonces la red contendrá un arco desde X hasta Y_1, y no habrá ningún arco que conecte Y_1 con las demás variables. Por tanto, *la ausencia de arcos* es también una forma de expresar información. El hecho de que Y_1 depende solamente de su causa, X, se traduce matemáticamente diciendo que, conocido el valor de X, la probabilidad de Y_1 es independiente de los valores que toman esas otras variables, o dicho de otro modo, X separa Y_1 de dichas variables. Empezamos a ver aquí la relación entre el concepto de padre y el de causa, entre el de hijo y el de efecto, entre el de arco y el de influencia causal directa, entre el de independencia en los mecanismos causales y el de independencia probabilista.

En este punto es donde se manifiesta la importancia del sentido de los arcos y su relación con la idea de causalidad. Volviendo al ejemplo del paludismo, el hecho de que las variables País-de-origen y Tipo-sanguíneo no tengan ningún padre común significa que son a priori independientes, es decir, que el país no influye en el tipo sanguíneo y viceversa, de modo que, si no hay más evidencia, no podemos obtener ninguna información sobre el país de origen a partir del tipo sanguíneo, ni viceversa. Sin embargo, el hecho de que ambas variables tengan un hijo común significa que, una vez conocido el valor de ese nodo, surgen correlaciones entre los padres.[7] Podemos decir, usando la terminología de Pearl [44], que el camino entre U_1 y U_2 permanece *bloqueado* hasta que sea *activado* por la llegada de información sobre X o sobre alguno de sus descendientes.

Para el caso de los efectos de una variable ocurre precisamente lo contrario: todo médico sabe que hay correlación entre la fiebre y el test de la gota gruesa. Sin embargo, tal como discutimos en la sección 3.1, la correlación desaparece cuando averiguamos si el paciente tiene o no tiene paludismo. Es decir, el camino entre Y_1 e Y_2 está *activado* en principio, y *se bloquea* sólo al conocer el valor de X. De esta asimetría entre padres e hijos, reflejo de la asimetría que existe en el mundo real entre causas y efectos, procede el nombre de separación *direccional*.

Por tanto, hay dos formas de justificar los enlaces que introducimos u omitimos al construir nuestra red. La primera es de naturaleza teórica: formamos un modelo causal a partir de la experiencia de un especialista y trazamos los arcos correspondientes al modelo; la relación que hemos discutido entre los mecanismos causales y las propiedades matemáticas de independencia nos permite fundamentar nuestro modelo. El otro camino para justificar la red consiste en realizar una comprobación empírica a partir de un conjunto suficientemente amplio de casos, utilizando las herramientas estadísticas que se emplean habitualmente para detectar correlaciones.

Hay otro punto relativo a la semántica de las redes bayesianas, que vamos a mencionar sólo brevemente, pues aún está muy discutido. Nos referimos al debate entre los que defienden que las redes probabilistas pueden expresar causalidad y los que sostienen que éstas sólo expresan correlaciones entre variables. En realidad, no se trata de un debate limitado al campo de las redes bayesianas, sino que la existencia de la causalidad es una cuestión que se han planteado matemáticos y filósofos por lo menos desde el siglo XVII, a partir de las teorías de Hume. Para no entrar en esta cuestión citaremos solamente tres trabajos, los de Pearl y Verma [48, 46] y el de Druzdzel y Simon [19], que muestran cómo recientemente han surgido argumentos matemáticos para defender la interpretación causal frente a la meramente correlacional.

En resumen, lo que hemos intentado mostrar en esta sección es que la información cualitativa que expresa la estructura de una R.B. es más importante aún que la información

[7] La correlación que aparece entre las causas se aprecia mucho más claramente en el caso de la puerta OR (sec. 3.4).

cuantitativa, como lo demuestra el hecho de que se han construido *redes cualitativas* [64, 65], capaces de razonar a partir de las propiedades de independencia de las redes bayesianas, incluso en ausencia de valores numéricos. Por este motivo, Neapolitan [41] ha sugerido en nombre de *redes de independencia* (*independence networks*) como el más adecuado para las RR.BB.

Podríamos sintetizar todo lo dicho anteriormente repitiendo lo que Laplace afirmó en la introducción de su famoso libro *Théorie Analytique des Probabilités*:[8]

> La teoría de la probabilidad no es, en el fondo, más que el sentido común reducido al cálculo.

3.3 Propagación de evidencia en poliárboles

Vamos a estudiar ahora un algoritmo eficiente para calcular la probabilidad en una red bayesiana sin bucles. En realidad, dada una R.B., a partir de las probabilidades condicionales podríamos calcular la probabilidad conjunta según el teorema 3.7, y luego aplicar las ecuaciones (2.2) y (2.6) para calcular las probabilidades marginales y a posteriori, respectivamente. Sin embargo este método tendría complejidad exponencial incluso en el caso de poliárboles. Además, al añadir nueva evidencia tendríamos que repetir casi todos los cálculos. Por esta razón conviene encontrar algoritmos mucho más eficientes.

El algoritmo para poliárboles que presentamos en esta sección, basado en el paso de mensajes π y λ, fue desarrollado por Kim [34] a partir del que Pearl había propuesto para árboles [43]. Sin embargo, la principal limitación del algoritmo de Kim y Pearl es que no permite tratar los bucles que aparecen inevitablemente al desarrollar modelos del mundo real, por lo que en sí mismo resulta de muy poca utilidad y los constructores de RR.BB. recurren a otros que, aun perdiendo las ventajas de éste, son aplicables a todo tipo de RR.BB. Sin embargo, aquí lo vamos a estudiar con detalle por dos razones. Primera, por su sencillez y elegancia, que nos permitirán comprender mejor las propiedades de las RR.BB. Y segunda, porque el algoritmo de condicionamiento local [15], aplicable también a redes múltiplemente conexas, es una extensión de éste, que se basa en los mismos conceptos y definiciones.

Para comprender mejor el desarrollo matemático que vamos a realizar, puede ser útil al lector repasar la sección 3.1, en que aparecen sencillos ejemplos numéricos que explican por qué se introducen las definiciones de π y λ, y cómo se propaga la evidencia.

3.3.1 Definiciones básicas

Una de las propiedades fundamentales de un poliárbol es que hay un único camino entre cada par de nodos. En consecuencia, la influencia de cada hallazgo se propaga hasta un nodo X bien a través de los padres o a través de los hijos de éste, por lo que para cada nodo X podemos hacer una partición de la evidencia (recordamos que la evidencia es el conjunto de hallazgos) en dos subconjuntos, tales que

$$\mathbf{e} = \mathbf{e}_X^+ \cup \mathbf{e}_X^- \tag{3.63}$$

$$\mathbf{e}_X^+ \cap \mathbf{e}_X^- = \varnothing \tag{3.64}$$

[8] Cita tomada de Druzdzel [18].

donde $\mathbf{e}_X^+$ representa la evidencia "por encima de X" y $\mathbf{e}_X^-$ "por debajo de X" en el sentido antes mencionado.

De forma similar, la eliminación de un enlace XY divide a la red —y por tanto también la evidencia— en dos partes, una que queda "por encima" del enlace y otra que queda "por debajo". Las llamaremos $\mathbf{e}_{XY}^+$ y $\mathbf{e}_{XY}^-$, respectivamente. Al igual que en el caso anterior, se cumple que

$$\mathbf{e} = \mathbf{e}_{XY}^+ \cup \mathbf{e}_{XY}^- \tag{3.65}$$

$$\mathbf{e}_{XY}^+ \cap \mathbf{e}_{XY}^- = \varnothing \tag{3.66}$$

Ejemplo 3.10 En la figura 3.5 (pág. 49), si tuviéramos $\mathbf{e} = \{+f, +g, \neg i\}$, entonces $\mathbf{e}_E^+ = \{+f\}$ y $\mathbf{e}_E^- = \{+g, \neg i\}$. Del mismo modo, $\mathbf{e}_H^+ = \{+f, +g, \neg i\}$ y $\mathbf{e}_H^- = \varnothing$. La eliminación del enlace EH dividiría la red en dos partes, y tendríamos $\mathbf{e}_{EH}^+ = \{+f, \neg i\}$ y $\mathbf{e}_{EH}^- = \{+g\}$. □

Basándonos en la partición de la evidencia, podemos establecer las siguientes definiciones (cf. fig. 3.7):

$$\pi(x) \equiv P(x, \mathbf{e}_X^+) \tag{3.67}$$

$$\lambda(x) \equiv P(\mathbf{e}_X^- \,|\, x) \tag{3.68}$$

$$\pi_X(u_i) \equiv P(u_i, \mathbf{e}_{U_i X}^+) \tag{3.69}$$

$$\lambda_{Y_j}(x) \equiv P(\mathbf{e}_{XY_j}^- \,|\, x) \tag{3.70}$$

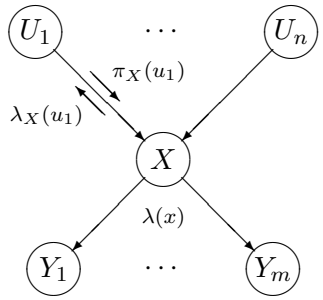

Figura 3.7: Propagación de evidencia mediante intercambio de mensajes.

El sentido de estas definiciones es el siguiente:

- $\pi(x)$ indica qué valor de X es más probable según la evidencia relacionada con las causas de X (es decir, según la evidencia "por encima" de X).
- $\lambda(x)$ indica qué valor de X explica mejor los hallazgos correspondientes a los efectos de X (la evidencia "por debajo" de X).
- $\pi_X(u)$ indica qué valor de U es más probable según la evidencia "por encima" del enlace UX.
- $\lambda_{Y_j}(x)$ indica qué valor X explica mejor la evidencia "por debajo" del enlace XY.

Para entender mejor estas explicaciones, conviene volver a los ejemplos de la sección 3.1.

Antes de concluir esta sección, señalemos que las definiciones anteriores, aunque tomadas del libro de Pearl [45], han sido modificadas de acuerdo con la propuesta de Peot y Shachter [50], con el fin de permitir un tratamiento coherente de los bucles mediante el algoritmo de condicionamiento local [15].

3.3.2 Computación de los mensajes

Recordemos una vez más que nuestro objetivo es calcular la probabilidad a posteriori de cada nodo, definida en la ecuación (2.25). A partir, de ahí,

$$\begin{aligned} P^*(x) = P(x|\,\mathbf{e}) &= \alpha\, P(x, \mathbf{e}_X^+, \mathbf{e}_X^-) \\ &= \alpha\, P(x, \mathbf{e}_X^+)\, P(\mathbf{e}_X^-|\, x, \mathbf{e}_X^+) \end{aligned}$$

donde hemos definido

$$\alpha \equiv [P(\mathbf{e})]^{-1} \tag{3.71}$$

Ahora bien, por la separación direccional sabemos que $P(\mathbf{e}_X^-|\, x, \mathbf{e}_X^+) = P(\mathbf{e}_X^-|\, x)$, de modo que, aplicando las definiciones anteriores llegamos a

$$P^*(x) = \alpha\, \pi(x)\, \lambda(x) \tag{3.72}$$

Necesitamos, por tanto, calcular los tres factores que aparecen en esta expresión. Empecemos con $\pi(x)$. Según su definición,

$$\pi(x) = P(x, \mathbf{e}_X^+) = \sum_{\bar{u}} P(x|\,\bar{u}) P(\bar{u}, \mathbf{e}_X^+)$$

Como las causas de X no tienen ningún antepasado común por estar en un poliárbol (red simplemente conexa), todas ellas y las ramas correspondientes son independientes mientras no consideremos la evidencia relativa a X o a sus descendientes:

$$\begin{aligned} P(\bar{u}, \mathbf{e}_X^+) &= P(u_1, \mathbf{e}_{U_1X}^+, \ldots, u_n, \mathbf{e}_{U_nX}^+) \\ &= \prod_{i=1}^{n} P(u_i, \mathbf{e}_{U_iX}^+) = \prod_{i=1}^{n} \pi_X(u_i) \end{aligned} \tag{3.73}$$

Por tanto,

$$\pi(x) = \sum_{\bar{u}} P(x|\,\bar{u}) \prod_{i=1}^{n} \pi_X(u_i)\,. \tag{3.74}$$

El paso siguiente consiste en calcular $\pi_X(u_i)$ o, lo que es lo mismo, $\pi_{Y_j}(x)$, puesto que en una R.B. todos los nodos son equivalentes; es sólo una cuestión de notación. La evidencia que está por encima del enlace XY_j, $\mathbf{e}_{XY_j}^+$, podemos descomponerla en varios subconjuntos: la que está por encima de X y la que está por debajo de cada enlace XY_k para los demás efectos Y_k de X (fig. 3.7). Sabemos además que X separa $\mathbf{e}_X^+$ de $\mathbf{e}_{XY_k}^-$, y separa también los

subconjuntos $\mathbf{e}^-_{XY_k}$ entre sí. Con estas consideraciones, obtenemos

$$\begin{aligned}\pi_{Y_j}(x) &= P(x, \mathbf{e}^+_{XY_j}) = P(x, \mathbf{e}^+_X, \mathbf{e}^-_{XY_k}\ k{\neq}j) \\ &= P(x, \mathbf{e}^+_X) \prod_{k\neq j} P(\mathbf{e}^-_{XY_k} | \, x) \\ &= \pi(x) \prod_{k\neq j} \lambda_{Y_k}(x) \end{aligned} \tag{3.75}$$

Para calcular esta expresión, es necesario hallar $\lambda_{Y_k}(x)$ —o $\lambda_{Y_j}(x)$, pues el resultado obtenido será válido para todos los efectos de X—. Representaremos mediante $\bar{V}$ el conjunto de causas de Y_j (o del efecto considerado) distintas de X, tal como muestra la figura 3.8. Por simplificar la notación, escribiremos $\mathbf{e}^+_{\bar{V}Y_j} = \mathbf{e}^+_{V_1Y} \cup \ldots \cup \mathbf{e}^+_{V_pY}$, con lo que nos queda $\mathbf{e}^-_{XY_j} = \mathbf{e}^-_Y \cup \mathbf{e}^+_{\bar{V}Y}$.

Recordemos que Y_j separa $\mathbf{e}^-_{Y_j}$ del resto de la red que está por encima de Y_j, e igualmente los padres de Y_j separan Y_j de $\mathbf{e}^+_{\bar{V}Y_j}$. Aplicando repetidamente la proposición 2.19, resulta

$$\begin{aligned}\lambda_{Y_j}(x) &= P(\mathbf{e}^-_{XY_j} | \, x) \\ &= \sum_{y_j} \sum_{\bar{v}} P(\mathbf{e}^-_{Y_j}, y_j, \mathbf{e}^+_{\bar{V}Y_j}, \bar{v} | \, x) \\ &= \sum_{y_j} \sum_{\bar{v}} P(\mathbf{e}^-_{Y_j} | \, y_j)\ P(y_j | \, \bar{v}, x)\ P(\mathbf{e}^+_{\bar{V}Y_j}, \bar{v} | \, x)\end{aligned}$$

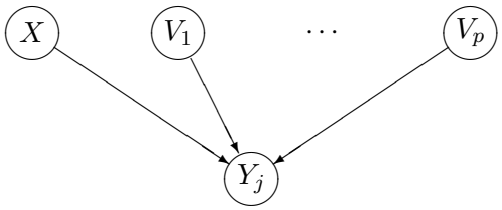

Figura 3.8: Padres de Y_j.

Puesto que las causas de Y_j son independientes a priori, podemos razonar como en la ecuación (3.73) para llegar a

$$P(\bar{v}, \mathbf{e}^+_{\bar{V}Y_j} | \, x) = P(\bar{v}, \mathbf{e}^+_{\bar{V}Y_j}) = \prod_{l=1}^{p} P(v_l, \mathbf{e}^+_{V_lY_j}) = \prod_{l=1}^{p} \pi_{Y_j}(v_l)$$

y, en consecuencia,

$$\lambda_{Y_j}(x) = \sum_{y_j} \left[\lambda(y_j) \sum_{\bar{v}} P(y_j | \, x, \bar{v}) \prod_{l=1}^{p} \pi_{Y_j}(v_l) \right] \tag{3.76}$$

Finalmente, hay que calcular $\lambda(x)$, lo cual resulta bastante sencillo:

$$\begin{aligned}\lambda(x) &= P(\mathbf{e}^-_{XY_1}, \ldots, \mathbf{e}^-_{XY_m} | \, x) \\ &= \prod_{j=1}^{m} P(\mathbf{e}^-_{XY_j} | \, x) = \prod_{j=1}^{m} \lambda_{Y_j}(x)\end{aligned} \tag{3.77}$$

Para completar el algoritmo, falta hallar la constante α que aparece en (3.72). Realizar el cálculo a partir de la definición 3.71 resultaría muy complicado en general. Sin embargo, sabemos que

$$\sum_x P^*(x) = \alpha \sum_x \pi(x)\,\lambda(x) = 1 \tag{3.78}$$

con lo que podemos obtener α como

$$\alpha = \left[\sum_x \pi(x)\,\lambda(x)\right]^{-1} \tag{3.79}$$

En la práctica, calcularemos $\pi(x)$ y $\lambda(x)$ para cada nodo y normalizaremos su producto de acuerdo con la ecuación (3.78).

Observe que por cada enlace $X \to Y$ circulan dos mensajes, $\pi_Y(x)$ de X a Y, y $\lambda_Y(x)$, de Y a X, pero ambos mensajes son vectores correspondientes a la variable X (por tanto, la dimensión del vector es $|X|$, el número de valores de X), mientras que la variable Y sólo aparece como subíndice en los dos: en $\pi_Y(x)$ indica el nodo que recibe el mensaje, mientras que en $\lambda_Y(x)$ indica el que lo envía.

3.3.3 Comentarios

Las fórmulas que acabamos de deducir son recursivas: $\pi(x)$ se calcula a partir de $\pi_X(u_i)$; $\pi_{Y_j}(x)$ a partir de $\pi(x)$ y de $\lambda_{Y_k}(x)$, etc. Necesitamos por tanto una condición de terminación para que el algoritmo esté completo. Por otro lado, necesitamos explicar cómo introducir en este esquema la evidencia observada. Resolveremos ambos problemas del siguiente modo:

Para un nodo U sin padres, $\mathbf{e}^+_U = \varnothing$, por lo que $\pi(u) = P(u)$, que es uno de los parámetros que definen la red. En este caso el problema de terminación ya lo teníamos resuelto.

Para un nodo terminal Y (nodo sin hijos), hace falta conocer $\lambda(y)$. Si no hay ninguna información sobre este nodo, asignamos el mismo número para cada valor y; por ejemplo, $\lambda(y) = 1$ para todo y. Vemos en la ecuación (3.72) que un vector $\lambda(x)$ constante no modifica el valor de $P^*(x)$. También vemos, a partir de la ecuación (3.76), que para un vector constante $\lambda(y) = 1$ podemos alterar el orden de los sumatorios y llegar a

$$\begin{aligned}\lambda_{Y_j}(x) &= c\sum_{\bar v}\left[\prod_{l=1}^{p}\pi_{Y_j}(v_l)\sum_{y_j}P(y_j|\,x,\bar v)\right]\\ &= c\sum_{\bar v}\left[\prod_{l=1}^{p}\pi_{Y_j}(v_l)\right] = c\sum_{\bar v}P(\bar v,\mathbf{e}^+_{\bar V Y_j}) = c\,P(\mathbf{e}^+_{\bar V Y_j}),\quad \forall x\end{aligned} \tag{3.80}$$

que es de nuevo un vector constante —es decir, independiente de x— y no transmite ninguna información, pues según las ecuaciones (3.72) y (3.76), un vector λ constante no influye en el resultado final.

Si hay un nodo terminal Y de valor conocido y_0 (es decir, la afirmación "$Y = y_0$" es parte de la evidencia), asignamos a $\lambda(y_0)$ un número positivo cualquiera y 0 a los demás valores de Y. Por ejemplo,

$$\begin{cases}\lambda(y_0) = 1\\ \lambda(y) \;\; = 0 \quad \text{para } y \neq y_0\end{cases}$$

lo cual implica, según (3.72),

$$\begin{cases} P(y_0) = 1 \\ P(y) \ = 0 \quad \text{para } y \neq y_0 \end{cases}$$

Vemos que, efectivamente, la probabilidad se ajusta a la afirmación de partida, "$Y = y_0$"; además sólo el valor y_0 cuenta en el sumatorio de la ecuación (3.76), por lo que podemos concluir que esta asignación de $\lambda(y)$ para nodos terminales es coherente, y así queda completo el algoritmo de propagación de evidencia en poliárboles.

En la sección siguiente vamos a mostrar un ejemplo completo de propagación de evidencia en una red bayesiana.

Ejemplo 3.11 Volvamos de nuevo a la red de la figura 3.5 (pág. 49). Recordemos que, además de tener la estructura de la red, conocemos las probabilidades a priori de los nodos sin padres: $P(a)$, $P(b)$ y $P(c)$, y las probabilidades condicionales: $P(d|\, a, b)$, $P(e|\, c)$, etc.

Supongamos que $\mathbf{e} = \{+f, +g, \neg i\}$. La asignación de landas para los nodos terminales será:

$$\begin{cases} \lambda(+f) = 1 \\ \lambda(\neg f) \ = 0 \end{cases} \quad \begin{cases} \lambda(+g) = 1 \\ \lambda(\neg g) \ = 0 \end{cases} \quad \begin{cases} \lambda(+h) = 1 \\ \lambda(\neg h) \ = 1 \end{cases} \quad \begin{cases} \lambda(+i) = 0 \\ \lambda(\neg i) \ = 1 \end{cases}$$

Queremos calcular $P^*(e)$, y por eso escogemos el nodo E como *pivote*, en el sentido de que se va a encargar de solicitar información a todos sus vecinos. Es posible que luego otros nodos soliciten los mensajes que les faltan, con el fin de computar su propia probabilidad, aunque también es posible que el nodo pivote E, una vez que ha recibido todos sus mensajes "decida" computar y enviar los mensajes de vuelta para sus vecinos, con el fin de que éstos hagan lo mismo con sus demás vecinos, y así sucesivamente hasta alcanzar todos los nodos terminales del poliárbol.

Con este esquema en mente, empezamos buscando $\pi(e)$:

$$\begin{aligned} \pi(e) &= \sum_c P(e|\, c)\, \pi_E(c) \\ \pi_E(c) &= \pi(c)\, \lambda_F(c) = P(c)\, \lambda_F(c) \\ \lambda_F(c) &= \sum_f \lambda(f)\, P(f|\, c) = P(+f|\, c) \end{aligned}$$

Así concluimos el cálculo en esta rama del árbol. Continuamos con otras ramas:

$$\begin{aligned} \lambda(e) &= \lambda_I(e)\, \lambda_H(e) \\ \lambda_I(e) &= \sum_i \lambda(i)\, P(i|\, e) = P(\neg i|\, e) \\ \lambda_H(e) &= \sum_h \lambda(h) \sum_d P(h|\, d, e)\, \pi_H(d) \end{aligned}$$

Deberíamos calcular ahora $\pi_H(d)$. Sin embargo, podemos saber ya que $\lambda_H(e)$ va a ser un vector constante porque $\lambda(h)$ también lo es. Podemos demostrarlo mediante el argumento numérico de la ecuación (3.80). Otra forma de razonarlo es a partir de las propiedades de

independencia condicional: cuando el valor de H no se conoce, D y E son independientes; recordando además que D separa G de E, tenemos

$$\begin{aligned}\lambda_H(e) &= P(\mathbf{e}^-_{EH}|\, e) = P(+g|\, e) \\ &= \sum_d P(+g|\, d, e)\, P(d|\, e) = \sum_d P(+g|\, d)\, P(d) = P(+g)\end{aligned}$$

que es un vector constante (no depende de e).

Por fin, nos queda $\lambda(e) = \lambda_I(e)$ y basta normalizar el producto $\pi(e) \cdot \lambda(e)$ para conocer $P^*(x)$. Del mismo modo podemos calcular la probabilidad a posteriori de cualquier otra variable, aprovechando —naturalmente— los resultados ya obtenidos. □

3.3.4 Implementación distribuida

El algoritmo que hemos presentado se presta inmediatamente a una implementación recursiva, según hemos comentado anteriormente. Vamos a ver ahora cómo podemos diseñar un algoritmo distribuido a partir de las mismas expresiones. (Veremos también que este método puede llevarnos a una implementación iterativa, que presenta la ventaja de que requiere mucha menos memoria de cálculo que la implementación recursiva.)

En la implementación distribuida, cada procesador corresponderá a un nodo y, por tanto, a una variable. La información que debe almacenar puede ser estática o dinámica. Aquí, "estática" significa "independiente de la evidencia observada", tal como la estructura de la red y las probabilidades condicionales. En caso de que los nodos no sean procesadores físicos reales sino que estén simulados mediante un programa de ordenador, lo primero que cada nodo necesita conocer son sus causas y sus efectos; esto puede realizarse fácilmente definiendo dos listas con punteros hacia los nodos correspondientes, las cuales codifican la topología de la red. A continuación hay que introducir la información numérica estática, a saber, las probabilidades a priori y condicionales.

La información dinámica consiste, en primer lugar, en los valores de λ para los nodos terminales, tal como hemos explicado anteriormente, y en los mensajes π y λ correspondientes a la propagación de evidencia. La propiedad más importante que se deriva de los axiomas de independencia descritos en el capítulo anterior es que, en poliárboles, cada enlace descompone la red en dos partes cuya única interacción se transmite a través de dicho enlace, y los mensajes intercambiados están desacoplados, en el sentido de que $\pi_Y(x)$ puede calcularse independientemente de $\lambda_Y(x)$, y viceversa. En la figura 3.9, que muestra los cálculos realizados en el nodo X, esta propiedad aparece como la ausencia de bucles en el flujo de información. Se puede comprobar también, observando las fórmulas de la sección 3.3.2, que toda la información requerida por un nodo para computar sus mensajes se encuentra almacenada localmente.

Un nodo X está en disposición de enviar un mensaje a su vecino W cuando y sólo cuando ha recibido ya los mensajes procedentes de todos sus demás vecinos. Un nodo X con n causas y m efectos que ha recibido q mensajes se encuentra en uno de tres estados posibles:

1. $q \leq n + m - 2$. Esto significa que X está esperando al menos dos mensajes, por lo que todavía no puede calcular ninguno de los que debe enviar.

2. $q = n + m - 1$. En este caso, X ha recibido un mensaje de cada vecino excepto de uno, que llamaremos W. Por eso X puede calcular ya el mensaje que debe enviar a W (aunque todavía no puede calcular ningún otro mensaje).

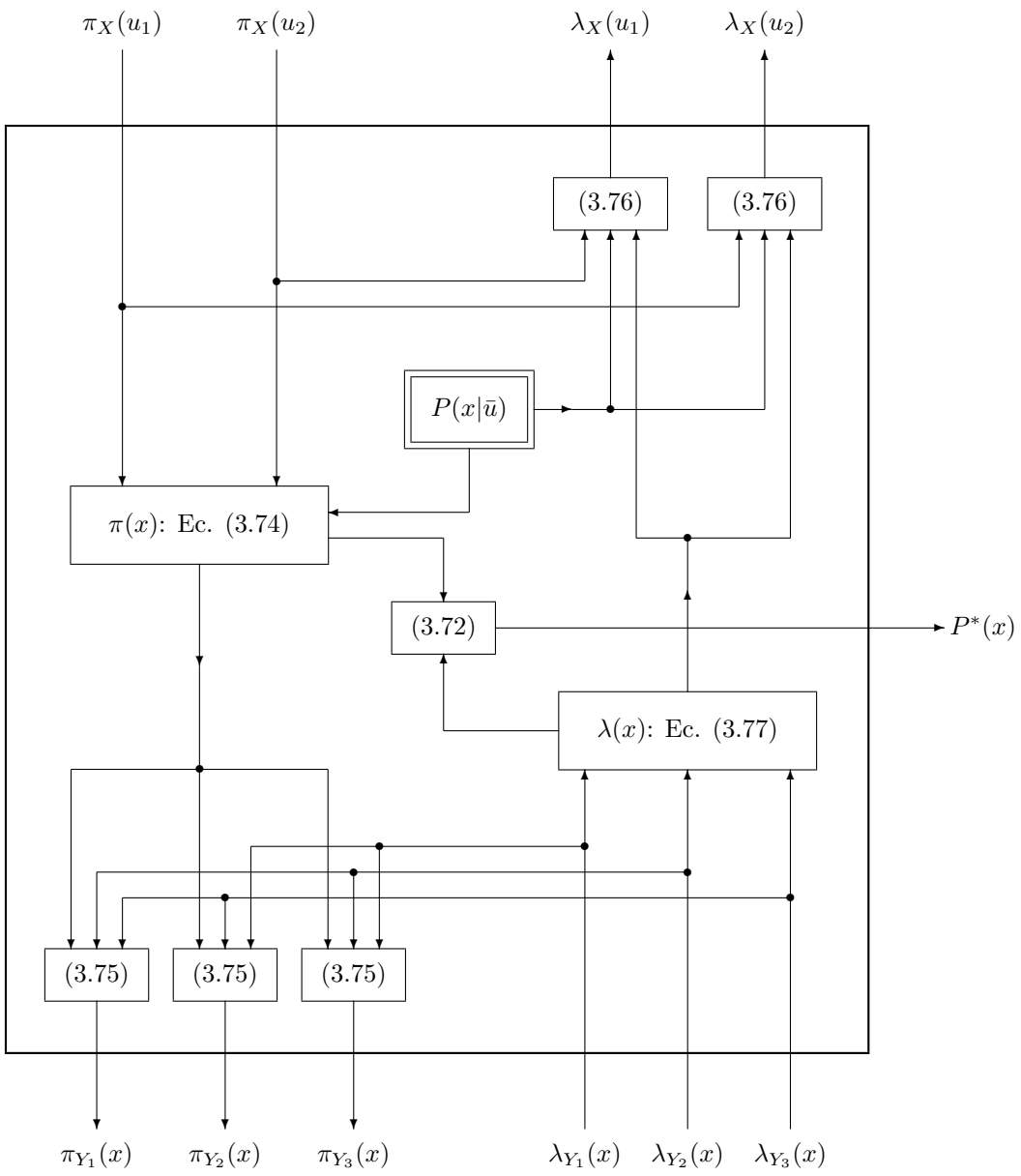

Figura 3.9: Computaciones realizadas en el nodo X.

3. $q = n + m$. Cuando X ha recibido todos los mensajes que estaba esperando, puede calcular por fin los que le faltaban por enviar.

Al principio, $q = 0$ para todos los nodos, pues aún no ha circulado ningún mensaje; por tanto, todos los nodos con un solo vecino (n+m=1) se encuentran en el estado 2; los demás se encuentran todavía en el estado 1. Es posible demostrar que siempre hay algún nodo dispuesto a enviar un mensaje, por lo que el proceso no se interrumpe nunca hasta que el algoritmo se ha completado. En vez de realizar la demostración, que es sencilla conceptualmente pero engorrosa, volvamos una vez más a la figura 3.5.

Antes de que empiece la propagación, todos los nodos que tienen un solo vecino (A, B, F, G y I) se hallan en estado 2, y los demás en estado 1. Cuando aquéllos envían sus mensajes respectivos, C y D pasan al estado 2, y lo mismo ocurre en el paso siguiente con E y H. Cuando los mensajes $\pi_H(e)$ y $\lambda_H(e)$ llegan a su destino, estos dos últimos nodos pasan al estado 3, de modo que pueden enviar ya a sus vecinos los mensajes que faltaban, y en dos pasos más queda concluido el proceso.

La discusión anterior es interesante para demostrar que no es necesario tener un mecanismo global de control, por lo que el modelo puede implementarse como una *red asíncrona* en que el número de mensajes recibido determina qué mensajes puede calcular y enviar cada nodo.

Si el algoritmo se implementa secuencialmente y la computación necesaria en cada nodo está acotada (limitando el número de padres y valores), el tiempo de computación es proporcional al número de nodos. En este caso resulta más eficiente realizar la propagación de evidencia en dos fases: recolección de mensajes hacia el nodo pivote y distribución desde él, como propusieron Jensen, Olesen y Andersen [32] para árboles de cliques.

En cambio, si hay un procesador por cada nodo, el tiempo de computación es proporcional a la longitud máxima que exista dentro de la red. La versión que hemos presentado aquí, basada en tres estados diferentes para cada nodo, se diferencia ligeramente de la de Pearl [45] en que evita computar y enviar mensajes prematuros carentes de sentido. La distinción no tiene importancia si disponemos de un procesador físico (*hardware*) por cada nodo. Pero si los procesadores conceptuales (los nodos) están simulados por un número menor de procesadores reales, el despilfarro computacional de enviar mensajes inútiles puede resultar muy caro en términos de eficiencia. En este último caso, en que los nodos hacen cola para acceder a un número limitado de procesadores físicos, encontramos el problema típico de la programación distribuida, a saber, cuál de los mensajes debe computarse primero con el fin de lograr la máxima eficiencia.

Ejemplo 3.12 Sea una red bayesiana dada por el grafo de la figura 3.10 y por las siguientes tablas de probabilidad (suponemos que todas las variables son binarias, de modo que $P(\neg a) = 1 - P(+a)$, etc.):

$$P(+a) = 0'3 \qquad P(+b) = 0'1$$
$$\left\{ \begin{array}{cc} P(+c|+a,+b) = 0'9 & P(+c|+a,\neg b) = 0'2 \\ P(+c|\neg a,+b) = 0'3 & P(+c|\neg a,\neg b) = 0'1 \end{array} \right\}$$
$$\big\{P(+d|+b) = 0'8 \qquad P(+d|\neg b) = 0\big\}$$

Dada la evidencia $\mathbf{e} = \{+a, \neg d\}$, calcular todos los mensajes π y λ que intervienen y la probabilidad a posteriori de cada variable.

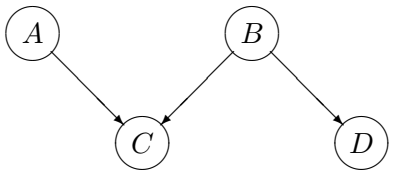

Figura 3.10: Computación distribuida de los mensajes π y λ.

Solución. Empezamos por asignar la evidencia observada. Así, el hallazgo $+a$ implica que $\pi(a) = (1\ 0)$ —este vector significa que $\pi(+a) = 1$ y $\pi(\neg a) = 0$— mientras que el hallazgo $\neg d$ se traduce en $\lambda(d) = (0\ 1)$. Como B no tiene evidencia ni directa ni procedente de sus padres, $\pi(b) = P(b) = (0'1\ 0'9)$; es decir, le asignamos su probabilidad a priori. Al nodo C le asignamos un vector constante, $\lambda(c) = (1\ 1)$, porque no tiene evidencia asociada directamente ni procedente de sus hijos.

Ahora hay que empezar a propagar la evidencia, de acuerdo con las ecuaciones (3.74) a (3.77). Vemos que A está esperando un solo mensaje, $\lambda_C(a)$, y D está esperando un solo mensaje, $\pi_D(b)$, de modo que están ya en condiciones de empezar a enviar mensajes. En cambio B y C están esperando dos mensajes cada uno, por lo que todavía no pueden enviar ninguno.

El mensaje que envía A es $\pi_C(a) = \pi(a) = (1\ 0)$, porque A no tiene otros hijos. El que envía D es $\lambda_D(b) = \sum_d \lambda(d) \cdot P(d|b) = 0 \cdot P(d|b) + 1 \cdot P(\neg d|b) = P(\neg d|b) = 1 - P(+d|b) = (0'2\ 1)$. Ahora tanto a B como a C sólo les falta recibir un mensaje, por lo que ya pueden empezar a enviar algunos mensajes.

El nodo B envía el mensaje $\pi_C(b) = \pi(b) \cdot \lambda_D(b) = (0'1\ 0'9) \cdot (0'2\ 1) = (0'02\ 0'9)$, mientras que el nodo C envía el mensaje $\lambda_C(b) = \sum_c \left[\lambda(c) \cdot \sum_a P(c|a,b) \cdot \pi_C(a)\right]$. Este mensaje se calcula así:

$$\begin{aligned}
\lambda_C(+b) &= \lambda(+c) \cdot [P(+c|+a,+b) \cdot \pi_C(+a) + P(+c|\neg a,+b) \cdot \pi_C(\neg a)] \\
&\quad + \lambda(\neg c) \cdot [P(\neg c|+a,+b) \cdot \pi_C(+a) + P(\neg c|\neg a,+b) \cdot \pi_C(\neg a)] \\
&= [0'9 \cdot 1 + 0'3 \cdot 0] + [0'1 \cdot 1 + 0'7 \cdot 0] = 1 \\
\lambda_C(\neg b) &= \lambda(+c) \cdot [P(+c|+a,\neg b) \cdot \pi_C(+a) + P(+c|\neg a,\neg b) \cdot \pi_C(\neg a)] \\
&\quad + \lambda(\neg c) \cdot [P(\neg c|+a,\neg b) \cdot \pi_C(+a) + P(\neg c|\neg a,\neg b) \cdot \pi_C(\neg a)] \\
&= [0'1 \cdot 1 + 0'7 \cdot 0] + [0'9 \cdot 1 + 0'3 \cdot 0] = 1
\end{aligned}$$

Por tanto, $\lambda_C(b) = (1\ 1)$; es decir, se trata de un vector constante, que no aporta información. En realidad, el hecho de que $\lambda(c)$ es un vector constante nos permite calcular el mensaje $\lambda_C(b)$ de forma más sencilla que como acabamos de hacerlo:

$$\begin{aligned}
\lambda_C(b) &= \sum_c \left[1 \cdot \sum_a P(c|a,b) \cdot \pi_C(a)\right] \\
&= \sum_a \left[\sum_c P(c|a,b)\right] \cdot \pi_C(a) = \sum_a \pi_C(a)
\end{aligned}$$

Esto explica por qué $\lambda_C(+b) = \lambda_C(\neg b)$, es decir, el mensaje $\lambda_C(b)$ es un vector constante, que no va a afectar al cálculo de la probabilidad a posteriori de B, pues el valor concreto que tome este vector "se pierde" al aplicar la normalización.

Siguiendo con la propagación de mensajes tenemos que $\lambda_C(a) = \sum_c [\lambda(c) \cdot \sum_b P(c|a,b) \cdot \pi_C(b)] = \sum_b [\sum_c P(c|a,b)] \cdot \pi_C(b) = \sum_b \pi_C(b) = (0'92\ 0'92)$, que es también un vector constante, lo cual demuestra que la evidencia $\neg d$ no se propaga hasta A.

El último mensaje que se propaga entre nodos es $\pi_D(b) = \pi(b) \cdot \lambda_C(b) = \pi(b) = P(b) = (0'1\ 0'9)$. Nótese que el orden en que hemos calculado los mensajes es el siguiente: $\pi_C(a)$, $\lambda_D(b)$, $\pi_C(b)$, $\lambda_C(b)$, $\lambda_C(a)$ y $\pi_D(b)$.

Por cierto, observe que $\lambda_C(b) = (1\ 1)$ ha conducido a $\pi_D(b) = \pi(b) = P(b)$, que a su vez implica que $\pi(d) = \sum_b P(d|b) \cdot \pi_D(b) = \sum_b P(d|b) \cdot P(b) = P(d)$; es decir, $\pi(d)$ coincide con la probabilidad a priori de D, lo cual demuestra que la evidencia $+a$ no se ha propagado hasta D. Visto de forma más general, $\lambda(c) = (1\ 1)$ implica que $\lambda_C(b)$ y $\lambda_C(a)$ son vectores constantes que no propagan evidencia, y esto significa que cuando no hay información sobre C ni por debajo de C el camino A–C–B está desactivado, de modo que ni la evidencia $+a$ se propaga hasta B y D ni la evidencia $\neg d$ se propaga hasta A.

Finalmente, vamos a calcular los vectores $\pi()$ y $\lambda()$ que nos faltan, con el fin de poder aplicar la ecuación (3.72) a cada nodo y calcular así su probabilidad a posteriori. Para el nodo A, $\lambda(a) = \lambda_C(a) = (1\ 1)$, porque sólo tiene un hijo, C, que no aporta ninguna evidencia; por tanto, $P^*(a) = \alpha\, \pi(a)\, \lambda(a) = \alpha\, \pi(a) = (1\ 0)$; es decir, $P^*(+a) = 1$, como debe ser, pues $+a$ forma parte de la evidencia. Para B, $\lambda(b) = \lambda_C(\neg b)\, \lambda_D(\neg b) = (1\ 1) \cdot (0'2\ 1) = (0'2\ 1)$ y $P^*(b) = \alpha\, \pi(b)\, \lambda(b) = \alpha \cdot (0'1\ 0'9) \cdot (0'2\ 1) = (0'022\ 0'978)$. Para C, $\pi(c) = \sum_a \sum_b P(c|a,b)\, \pi_C(a)\, \pi_C(b) = (0'198\ 0'722)$ y $P^*(c) = \alpha \cdot (0'198\ 0'722) \cdot (1\ 1) = (0'215\ 0'785)$. Por último, para D, $\pi(d) = \sum_b P(d|b) \cdot \pi_D(b) = (0'08\ 0'92)$ y $P^*(d) = \alpha\, \pi(d)\, \lambda(d) = \alpha \cdot (0'08\ 0'92) \cdot (0\ 1) = (0\ 1)$, lo cual también era de esperar, porque $\neg d$ forma parte de la evidencia. □

Insistimos una vez más en que en este ejemplo la computación se ha realizado de forma distribuida: en vez de tener un nodo pivote que se encargue de centralizar la recogida y distribución de la información, como ocurría en el ejemplo 3.11, aquí cada nodo "sabe" en todo momento qué mensajes ha recibido y, por tanto, cuáles puede enviar.

Ejercicio 3.13 Repetir los cálculos del ejemplo anterior para la evidencia $\mathbf{e} = \{+c, \neg d\}$.

Solución. Los mensajes y las probabilidades son (en el orden en que se calculan): $\lambda(c) = (1\ 0)$; $\lambda(d) = (0\ 1)$; $\pi(a) = (0'3\ 0'7)$; $\pi(b) = (0'1\ 0'9)$; $\pi_C(a) = (0'3\ 0'7)$; $\lambda_D(b) = (0'2\ 1)$; $\lambda_C(b) = (0'48\ 0'13)$; $\pi_C(b) = (0'02\ 0'9)$; $\pi(c) = (0'1266\ 0'7934)$; $\lambda(b) = (0'096\ 0'13)$; $\lambda_C(a) = (0'198\ 0'096)$; $\pi_D(b) = (0'048\ 0'117)$; $\lambda(a) = (0'198\ 0'096)$; $\pi(d) = (0'0384\ 0'1266)$; $P(a) = (0'4692\ 0'5308)$; $P(b) = (0'0758\ 0'9242)$; $P(c) = (1\ 0)$; $P(d) = (0\ 1)$.

3.4 La puerta OR/MAX

3.4.1 La puerta OR binaria

Hemos visto que, en el caso general, la probabilidad condicional viene dada por una tabla, tal como la que aparece en la página 3.1. Por tanto, el número de parámetros requerido para una familia crece exponencialmente con el número de padres. Esto conlleva varios inconvenientes. El más grave es la obtención de dichos parámetros: si obtenemos los resultados a partir de

una base de datos, necesitamos gran cantidad de casos para que los parámetros obtenidos sean fiables; si la ausencia de una base de datos nos obliga a recurrir a la estimación subjetiva de un experto humano, resultará muy complicado para él responder a tantísimas preguntas correspondientes a una casuística compleja: "¿Cuál es la probabilidad de que el paciente presente fiebre dado que tenga paludismo, neumonía y apendicitis y no tenga amigdalitis, ni meningitis, ni etc., etc.?"

El segundo problema que plantea el modelo general, una vez obtenidos los parámetros, es la cantidad de espacio de almacenamiento que requiere cuando el número de padres es grande (por ejemplo, para un nodo binario con 10 padres binarios, la tabla de probabilidad condicional tiene $2^{1+10} = 2.048$ parámetros). Y por último, otro grave inconveniente es que el tiempo de computación para la propagación de evidencia crece también exponencialmente con el número de padres de la familia considerada.

Por estas razones, es conveniente buscar modelos simplificados de interacción causal que simplifiquen la construcción de RR.BB. y la computación de la probabilidad. Pearl [45] los llama modelos *canónicos* porque son aplicables a numerosos campos, no son soluciones *ad hoc* para resolver un problema concreto de un dominio particular. Los más famosos entre ellos son las puertas OR y MAX probabilistas, que suponen una generalización de los correspondientes modelos deterministas.

En la puerta OR probabilista (*noisy OR-gate* [45, sec. 4.3.2]) se supone que cada causa U_i actúa para producir el efecto X, pero existe un *inhibidor* I_i que bloquea la influencia; es como si U_i estuviera inactiva. Por tanto, el parámetro fundamental es la probabilidad de que actúe el inhibidor (q_i) o bien su parámetro complementario, $c_i = 1 - q_i$, la probabilidad de que la causa U_i actuando en ausencia de otras causas llegue a producir X:

$$P(+x|+u_i, \neg u_{j\,[j\neq i]}) = c_i = 1 - q_i$$

Tenemos así la probabilidad de X en el caso de que haya una única causa presente y las demás están ausentes. Para hallar la probabilidad de X en el caso de que haya más de una causa presente, se introduce la hipótesis de que X sólo está ausente cuando todas las causas están ausentes o cuando para cada causa U_i que está presente ha actuado el correspondiente inhibidor I_i. Se supone que no sólo las causas sino también los inhibidores actúan independientemente, lo cual implica la independencia en sentido probabilista. En consecuencia,

$$P(\neg x|\bar{u}) = \prod_{i\in T_U} q_i$$

donde T_U indica el subconjunto de las causas de X que están presentes ($T_U \subset \bar{U}$).

A partir de aquí podemos construir la tabla $P(x|\bar{u})$ y aplicar el algoritmo de propagación general desarrollado en la sección anterior. Pero así habríamos resuelto uno sólo de los inconvenientes anteriores (el de la obtención de los parámetros), pues ya vimos que la complejidad de este algoritmo crecía exponencialmente con el número de padres. Por fortuna, existen expresiones para la puerta OR que llevan a un tiempo de propagación proporcional al tamaño de la familia. Dichas expresiones se encuentran en [45]. Nosotros, en vez de deducirlas aquí, las presentaremos como un caso particular de las correspondientes a la puerta MAX que vamos a estudiar a continuación.

3.4.2 Definición de la puerta MAX

Existe una generalización de la puerta OR binaria, que fue propuesta por Max Henrion [30] como modelo para la obtención del conocimiento; el nombre de "puerta MAX", su formulación

matemática y los algoritmos de propagación que discutimos a continuación fueron publicados por primera vez en [13].

Para llegar a una formulación matemática del modelo es necesario introducir previamente el siguiente concepto:

Definición 3.14 (Variable graduada) Es la variable X que puede estar ausente o presente con g_X grados de intensidad. Tiene por tanto g_X+1 valores posibles, a los que asignaremos enteros tales que $X=0$ significa "ausencia de X" y los números sucesivos indican grados de mayor intensidad.

Ejemplo 3.15 Supongamos que la variable X =Neumonía puede estar ausente o presente con tres grados de intensidad ($g_X=3$): leve, moderada o severa. Entonces $X=0$ significa "el paciente no tiene neumonía", $X=1$ significa "el paciente tiene neumonía leve", etc. □

Observe que el concepto de *graduada* no es sinónimo de *multivaluada*. De hecho, son dos conceptos independientes: por un lado, no todas las variables multivaluadas representan distintos grados de intensidad, y por otro lado, hay variables binarias graduadas, como son las que hemos visto hasta ahora de tipo presente/ausente o positivo/negativo, cuyos valores representábamos por $+x$ y $\neg x$. (La definición de variable graduada nos dice que a $\neg x$ le corresponde el valor 0 y a $+x$ el valor 1.) Más aún, *las variables que intervienen en la puerta OR binaria son siempre variables graduadas*, pues no tiene sentido plantear dicho modelo para variables no graduadas, tales como el sexo.

El modelo de interacción de las puertas OR/MAX es bastante general; aquí lo vamos a definir en el contexto de las RR.BB., aunque sería aplicable a otros métodos de tratamiento de la incertidumbre. Por simplificar la escritura, llamaremos $\tilde{U}_i$ al conjunto de todas las causas de X excluida U_i:

$$\tilde{U}_i \equiv \bar{U} \setminus U_i$$

Definición 3.16 (Puerta OR/MAX) En una red bayesiana, dada una variable graduada X con n padres $U_1,\ldots,U_n$ (también variables graduadas), decimos que interaccionan mediante una *puerta MAX* cuando se cumplen las dos condiciones siguientes:

$$1.\quad P(X=0|\bar{U}=0)=1 \tag{3.81}$$

$$2.\quad P(X\leq x|\bar{u}) = \prod_i P(X\leq x|U_i=u_i,\tilde{U}_i=0) \tag{3.82}$$

Si X y las U_i son todas binarias, se dice que interactúan mediante una *puerta OR*.

Podemos utilizar la notación x^0 ≡"$X=0$" para expresar ambas condiciones en forma abreviada como

$$1.\quad P(x^0|\bar{u}^0)=1 \tag{3.83}$$

$$2.\quad P(X\leq x|\bar{u}) = \prod_i P(X\leq x|u_i,\tilde{u}_i^0) \tag{3.84}$$

Intentaremos ahora explicar el significado de esta definición. La primera condición es fácil de interpretar: significa que, si todas las causas que pueden producir X están ausentes, entonces tenemos la seguridad de que también X estará ausente. Más adelante relajaremos esta restricción.

La segunda condición (ec. (3.82)) nos dice que $X \leq x$ sólo cuando ninguna de las causas U_i (actuando como si las demás causas estuvieran ausentes) ha elevado X a un grado superior a x. Dicho con otras palabras, el grado que alcanza X es el *máximo* de los grados producidos por las causas actuando independientemente; ésta es la razón por la que se denomina "puerta MAX". Al igual que en la puerta OR binaria, el resultado es el máximo de los valores de las entradas; esta coincidencia era de esperar, pues el modelo graduado es tan sólo una generalización del caso binario.

La importancia de esta definición es que permite calcular todos los valores de $P(x|\bar{u})$ a partir de un reducido número de parámetros. Para la familia X, serán las probabilidades X condicionadas a que una sola de las causas esté presente:

$$c_{X=x}^{U_i=u_i} \equiv P(X=x|\, U_i=u_i, \tilde{U}_i = 0) \tag{3.85}$$

que podemos escribir en forma abreviada como

$$c_x^{u_i} \equiv P(x|u_i, \tilde{u}_i^0) \tag{3.86}$$

En principio, el número de parámetros para el enlace U_iX es $(g_{U_i}+1)\cdot(g_X+1)$. Sin embargo, la suma de las probabilidades debe ser la unidad y, en consecuencia,

$$c_{x^0}^{u_i} = 1 - \sum_{x=1}^{g_X} c_x^{u_i}. \tag{3.87}$$

Por otra parte, la primera condición de la definición de la puerta OR (ec. (3.81)) es equivalente a decir que

$$c_x^{u_i^0} = \begin{cases} 1 & \text{para } x = 0 \\ 0 & \text{para } x \neq 0\,. \end{cases} \tag{3.88}$$

Por tanto, sólo se necesitan $g_{u_i} \cdot g_X$ parámetros para este enlace.

Ejemplo 3.17 Supongamos que tenemos una porción de red representada por la figura 3.11, y que cada una de las tres variables puede tomar los siguientes valores:

U_1 = Neumonía $\rightarrow$ {ausente, leve, moderada, severa}
U_2 = Paludismo $\rightarrow$ {ausente, presente}
X = Fiebre $\rightarrow$ {ausente, leve, elevada}

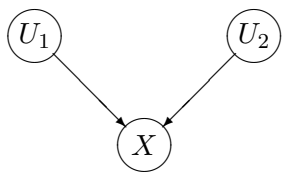

Figura 3.11: Ejemplo de puerta MAX.

Vemos que $g_{U_1} = 3$, $g_{U_2} = 1$ y $g_X = 2$. En el modelo general, para esta familia necesitaríamos una tabla con 16 parámetros (hay 4×2=8 combianaciones posibles de u_1 y u_2; para cada una de ellas deberíamos dar tres valores $P(x|u_1, u_2)$, pero como la suma de los tres debe ser la

unidad, sólo hace falta dar dos, de modo que necesitamos en total 8×2=16 parámetros). Sin embargo, para la puerta MAX, indicaremos los valores de $c_x^{u_1}$ y $c_x^{u_2}$, tal como muestran las tablas 3.2 y 3.3. En ellas se observa que el número de parámetros se ha reducido a la mitad. Si hubiéramos tenido más causas en vez de sólo dos, el ahorro habría sido mayor. A partir de estas dos pequeñas tablas, aplicando los axiomas anteriores podemos construir la tabla $P(x|u_1, u_2)$ completa necesaria para aplicar el algoritmo general. Sin embargo, existe una solución mucho más eficiente, que evita tener que calcular dicha tabla, como mostraremos en la próxima sección. □

$X \setminus U_1$	leve	moderada	severa
leve	0'50	0'40	0'20
elevada	0'20	0'50	0'80

Tabla 3.2: Parámetros $c_x^{u_1}$.

$X \setminus U_2$	presente
leve	0'20
elevada	0'75

Tabla 3.3: Parámetros $c_x^{u_2}$.

Causas no explícitas

Si en el modelo general tomamos $P(+x|\neg u_1, \neg u_2) > 0$, esto significa que X puede estar presente incluso cuando U_1 y U_2 están ausentes. Sin embargo, en la puerta OR/MAX, la propiedad (3.81) nos dice que cuando todas las causas de X están ausentes, sabemos con certeza que X estará ausente.

La propiedad en sí es razonable, pero existe el problema de que en la práctica es imposible considerar explícitamente todas las causas, pues éstas pueden ser muy numerosas e incluso muchas de ellas serán desconocidas; esto se ve especialmente claro en el caso de la medicina. La cuestión es importante aunque, afortunadamente, tiene una solución muy sencilla: para cada nodo X incluiremos un nodo X^* que agrupe todas las causas que no aparezcan explícitamente en el modelo.

Podemos suponer que el valor de este nodo siempre es presente ($\pi(+x^*) = 1$) y que su eficacia para producir X viene dada por los parámetros $c_x^{x^*}$ (de forma abreviada, c_x^*); en caso de que X sea una variable binaria, basta conocer un solo número, c_{+x}^*, pues $c_{\neg x}^*$ será siempre 0.

Al desarrollar el algoritmo de propagación para la puerta OR/MAX, veremos que el impacto de cada causa U_i se traduce en una $Q_{U_i}(x)$, y todas éstas se combinan de acuerdo con la ecuación (3.93). Por tanto, podemos aplicar, la propiedad asociativa del producto y agrupar varias causas en una sola sin violar los principios axiomáticos de las redes bayesianas. Lo que queremos decir es que está matemáticamente justificado incluir las causas no explícitas en un solo nodo y asignar al enlace correspondiente unos valores c_x^* que combinan los parámetros de todas ellas como si se tratara de una sola causa.

3.4.3 Algoritmo de propagación

Hemos resuelto ya los dos primeros problemas que presentaba el modelo general, pues ya no necesitamos obtener ni almacenar un número exponencial de parámetros por familia. Veamos a continuación cómo podemos resolver el tercero, es decir, cómo podemos realizar eficientemente la propagación de evidencia. Empezamos introduciendo la siguiente definición:

$$Q(x) \equiv P(X \leq x, \mathbf{e}_X^+) \tag{3.89}$$

Es fácil obtener $Q(x)$ a partir de $\pi(x)$

$$Q(x) = \sum_{x'=0}^{x} \pi(x') \tag{3.90}$$

y viceversa

$$\pi(x) = \begin{cases} Q(x) - Q(x-1) & \text{para } x \neq 0 \\ Q(0) & \text{para } x = 0 \end{cases} \tag{3.91}$$

Queremos encontrar ahora un algoritmo eficiente para calcular $Q(x)$. Aplicando las ecuaciones (3.83) y (3.73), podemos escribir

$$\begin{aligned} Q(x) &= \sum_{\bar{u}} P(X \leq x|\bar{u}) \quad P(\bar{u}, \mathbf{e}_X^+) \\ &= \sum_{\bar{u}} \left[\prod_i P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i|\mathbf{e}_{U_iX}^+) \right] \end{aligned}$$

En esta expresión podemos invertir el orden del productorio y de los sumatorios. En efecto,

$$\begin{aligned} &\sum_{u_1} \left[\prod_{i=1}^{n} P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i, \mathbf{e}_{U_iX}^+) \right] \\ &= \left[\sum_{u_1} P(X \leq x|u_1, \tilde{u}_1^0) \quad P(u_1, \mathbf{e}_{U_1X}^+) \right] \cdot \prod_{i=2}^{n} P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i, \mathbf{e}_{U_iX}^+) \\ &= P(X \leq x, \mathbf{e}_{U_1X}^+, \tilde{u}_1^0) \prod_{i=2}^{n} P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i|\mathbf{e}_{U_iX}^+) \\ &= Q_{U_1} \cdot \prod_{i=2}^{n} P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i|\mathbf{e}_{U_iX}^+) \end{aligned}$$

donde hemos introducido la definición

$$Q_{U_i}(x) \equiv P(X \leq x, \mathbf{e}_{U_iX}^+, \tilde{u}_i^0) \tag{3.92}$$

que es la probabilidad de $X = x$ considerando toda la evidencia por encima del enlace U_iX, en caso que todas las demás causas de X estuvieran ausentes.

Sustituyendo este resultado en la expresión de $Q(x)$ tenemos

$$Q(x) = Q_{U_1}(x) \cdot \sum_{u_2,\ldots,u_n} \left[\prod_{i=2}^{n} P(X \leq x|u_i, \tilde{u}_i^0) \quad P(u_i|\mathbf{e}_{U_iX}^+) \right]$$

y repitiendo la misma operación n veces llegamos a

$$Q(x) = \prod_i Q_{U_i}(x) \tag{3.93}$$

Lo que necesitamos ahora es una fórmula sencilla para calcular $Q_{U_i}(x)$. Para ello, definimos un nuevo conjunto de parámetros $C_x^{u_i}$:

$$C_x^{u_i} \equiv P(X \leq x | u_i, \tilde{u}_i^0) \tag{3.94}$$

que podemos calcular a partir de las $c_x^{u_i}$, según las ecuaciones (3.86) y (3.87):

$$\begin{aligned} C_x^{u_i} &= \sum_{x'=0}^{x} c_x^{u_i} = c_{x^0}^{u_i} + \sum_{x'=1}^{x} c_x^{u_i} = 1 - \sum_{x'=1}^{g_X} c_x^{u_i} + \sum_{x'=1}^{x} c_x^{u_i} \\ &= 1 - \sum_{x'=x+1}^{g_x} c_x^{u_i} \end{aligned} \tag{3.95}$$

Estos nuevos parámetros pueden ser almacenados junto con la descripción de la red (para ahorrar tiempo de computación) o calculados cuando se los necesita, aunque también es posible definir la red a partir de las $C_x^{u_i}$ en lugar de las $c_x^{u_i}$.

Desde aquí, el cálculo de $Q_{U_i}(x)$ es inmediato:

$$\begin{aligned} Q_{U_i}(x) &= \sum_{u_i} P(X \leq x | u_i, \tilde{u}_1^0) \quad P(u_i, \mathbf{e}^+_{U_i X}) \\ &= \sum_{u_i} C_x^{u_i} \quad \pi_X(u_i) \qquad (3.96) \\ &= \sum_{u_i} \pi_X(u_i) \left[1 - \sum_{x'=x+1}^{g_X} c_x^{u_i} \right] \qquad (3.97) \end{aligned}$$

En el caso de que tengamos además unas c_x^* correspondientes a las causas no explícitas en el modelo, podemos manejarlas como si se tratara de una causa similar a las demás y calcular la respectiva $Q^*(X)$, que deberá incluirse en el productorio de la ecuación (3.93). El tratamiento de las causas no explícitas es, por tanto, muy sencillo.

Hemos resuelto ya la primera parte del problema: cómo calcular $\pi(x)$ para la familia X en tiempo proporcional al número de padres. También el cálculo de $\lambda(x)$ y el de $\pi_X(u_i)$ o $\lambda_{Y_j}(x)$ están resueltos, pues podemos aplicar las ecuaciones (3.77) y (3.75), ya que en ellas no aparece $P(x|\bar{u})$ y por tanto no varían al pasar del caso general a la puerta OR/MAX. Lo que nos falta por resolver es cómo calcular $\lambda_Y(u_i)$, es decir, el mensaje que X envía a cada uno de sus padres.

La ecuación (3.76) puede escribirse para la familia X como

$$\lambda_X(u_i) = \sum_x \left[\lambda(x) \sum_{\tilde{u}_i} P(x|\bar{u}) \prod_{j \neq i} \pi_X(u_j) \right] \tag{3.98}$$

Observe que, dentro de esta expresión, el valor de u_i en $\bar{u} = (u_1, \ldots, u_n)$ está fijo (depende de qué $\lambda_X(u_i)$ estamos calculando), mientras que el valor de las demás variables u_j va cambiando según indica el sumatorio.

Ahora bien, una forma de fijar el valor u'_i para la variable U_i es asignarle un vector $\pi(u_i)$ definido así:

$$[\pi(u'_i)]_{U_i=u_i} = \begin{cases} 1 & \text{para } u_i = u'_i \\ 0 & \text{para } u_i \neq u'_i \end{cases} \tag{3.99}$$

puesto que entonces, según (3.72) y (3.75),

$$[P(u_i)]_{U_i=u'_i} = [\pi_X(u_i)]_{U_i=u'_i} = \begin{cases} 1 & \text{para } u_i = u'_i \\ 0 & \text{para } u_i \neq u'_i \end{cases}$$

y también

$$[P(x|\bar{u})]_{U_i=u'_i} = \sum_{u_i} P(x|\bar{u})\,[\pi(u_i)]_{U_i=u'_i}$$

Sustituyendo este resultado en la ecuación (3.98), tenemos

$$\lambda_X(u'_i) = \sum_x \left[\lambda(x) \sum_{\bar{u}} P(x|\bar{u}) \prod_j \pi_X(u_j) \right]_{U_i=u'_i}$$

o bien

$$\lambda_X(u_i) = \sum_x \lambda(x)\,[\pi(x)]_{U_i=u_i} \tag{3.100}$$

Aquí, $[\pi(x)]_{U_i=u_i}$ debe calcularse como hicimos anteriormente, es decir, con las ecuaciones (3.91) y (3.93), aunque ahora la ecuación (3.96) se simplifica para convertirse en

$$[Q_{U_i}(x)]_{U_i=u_i} = C_x^{u_i} \tag{3.101}$$

de acuerdo con el valor de $[\pi_X(u_i)]_{U_i=u'_i}$ indicado anteriormente.

Dicho con otras palabras, el algoritmo de la puerta OR/MAX puede expresarse así: para calcular $\pi(x)$, transformamos cada mensaje $\pi_X(u_i)$ en $Q_{U_i}(x)$, y los multiplicamos todos para obtener $Q(x)$, a partir del cual es muy sencillo obtener $\pi(x)$.

Cuando queremos calcular $\lambda_X(u_i)$ seguimos un procedimiento similar: para las causas U_j distintas de U_i tomamos las mismas $Q_{U_j}(x)$ que antes; para U_i, en cambio, tomamos la $Q_{U_i}(x)$ correspondiente al valor u_i según la ecuación (3.101), y repetimos —para cada valor de U_i— el mismo proceso que en el cálculo de $\pi(x)$.

3.4.4 Implementación distribuida

La implementación distribuida de la puerta OR/MAX es muy similar a la del caso general. Sin embargo, ahora los parámetros c_x^u (las tablas 3.2 y 3.3, por ejemplo) no son una característica del nodo X ni de esta familia en conjunto, sino que están asociados a cada enlace UX. Fíjese en la figura 3.12 y en la tabla 3.4 y observe dónde se almacenan las c_x^u. El nodo X debe saber solamente qué tipo de interacción debe aplicar: caso general o puerta OR/MAX; en el segundo caso, los parámetros se encontrarán almacenados en los enlaces correspondientes (salvo los parámetros c_x^*, correspondientes a las causas no explícitas, que se almacenarán en el propio nodo, para no tener que añadir un nodo que represente *OTRAS-CAUSAS-DE-X*). Comparando la figura 3.12 con la relativa al caso general (fig. 3.9, pág. 62), observamos que ahora los enlaces no son canales de información pasivos, sino procesadores activos que transforman cada mensaje $\pi_X(u)$ en $Q_U(x)$ y generan además $\lambda_X(u)$, liberando así al nodo X de algunas computaciones.

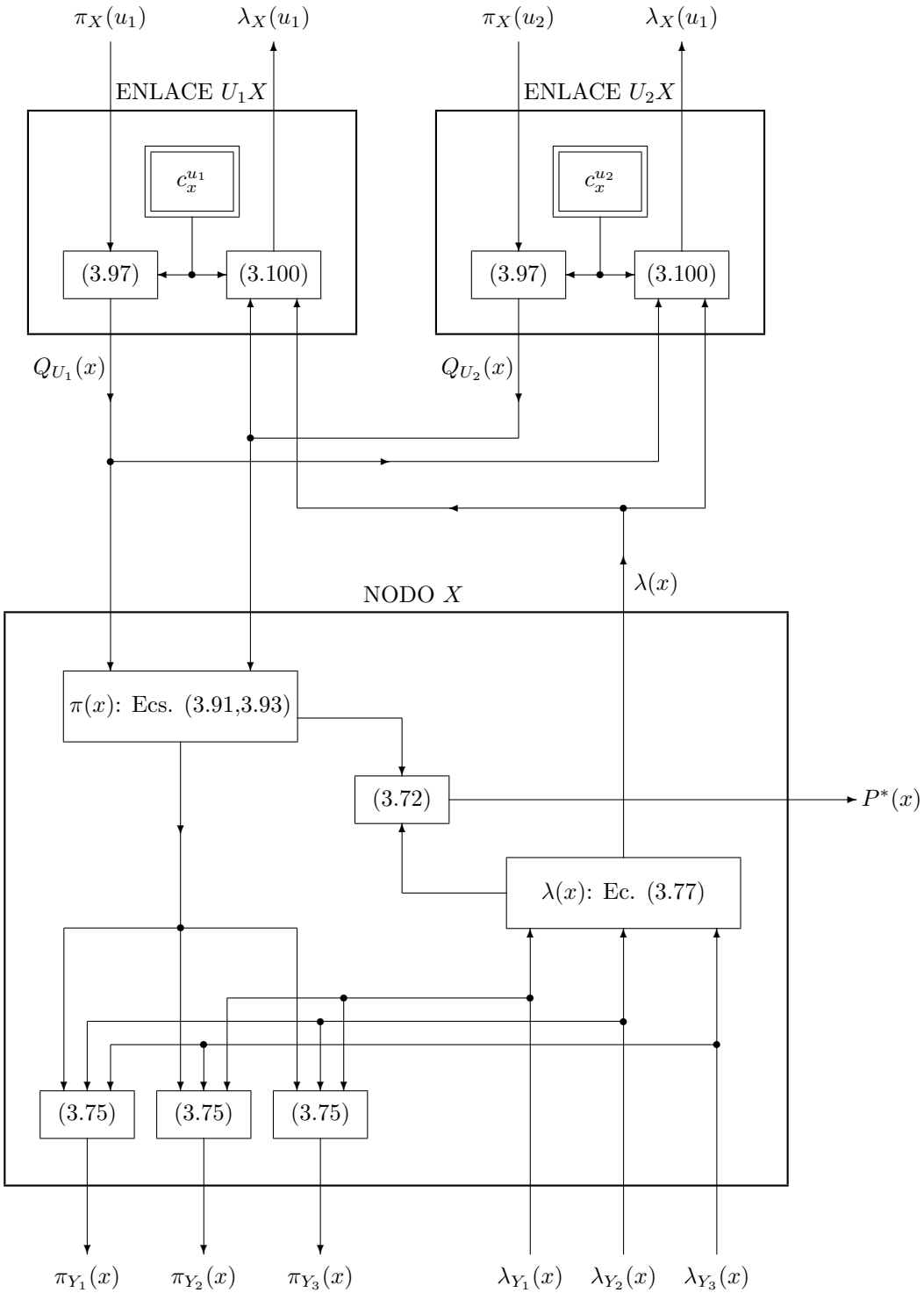

Figura 3.12: Computaciones realizadas en la puerta OR.

		Modelo general	Puerta OR
Nodo X	Almacena	$P(x\|u)$	c_x^*
	Recibe	$\pi_X(u_i)$, $\lambda_{Y_j}(x)$	$Q_{U_i}(X)$, $\lambda_{Y_j}(x)$
	Envía	$\lambda_X(u_i)$, $\pi_{Y_j}(x)$, $P^*(x)$	$\lambda(x)$, $\pi_{Y_j}(x)$, $P^*(x)$
Enlace U_iX	Almacena		$c_x^{u_i}$
	Recibe		$\pi_X(u_i)$, $\lambda(x)$, $Q_{U_{i'}}(x)$
	Envía		$Q_{U_i}(x)$, $\lambda_X(u_i)$

Tabla 3.4: Caso general y puerta OR.

3.4.5 Semántica

Sería interesante desarrollar el ejemplo 3.17 para ver cómo funciona la propagación de evidencia en la puerta OR/MAX y comprobar que los resultados obtenidos matemáticamente coinciden con lo previsible mediante nuestro sentido común. Lamentablemente, desarrollar con detalle un solo ejemplo con los múltiples casos posibles nos ocuparía mucho más espacio de lo que podemos permitirnos. Por otra parte, puede encontrarse un ejemplo bien explicado en [45, sec. 4.3.2]; aunque allí se describe solamente la puerta OR, el tratamiento resulta muy similar al que pudiéramos realizar para la puerta MAX.

Lo que vamos a discutir en esta sección es la relación entre el modelo general y la puerta OR/MAX. En la sección 3.2.4 hablamos de la semántica de las redes bayesianas, refiriéndonos al modelo general. Allí vimos la relación entre los axiomas de independencia y los mecanismos causales percibidos intuitivamente. De igual modo, estudiar la semántica de la puerta OR/MAX consiste en establecer una relación entre los axiomas de la definición 3.16 y nuestros conceptos de causalidad. Esta cuestión fue abordada parcialmente al introducir dicha definición. En efecto, allí se mostró que la ecuación (3.81) significa que el efecto X está ausente cuando todas las causas que lo producen están ausentes, lo cual concuerda naturalmente con el sentido común, y la ecuación (3.82) significa que el grado que alcanza el efecto X es el máximo de los que producirían sus causas actuando independientemente.

Al igual que discutimos al hablar de la semántica de las redes bayesianas en general, podemos afirmar aquí que hay dos formas posibles de justificar la utilización de la puerta OR/MAX como modelo simplificado al construir nuestra red. La primera —la teórica— consiste en crear en nuestra mente un modelo de cómo actúan las causas a la hora de producir el efecto considerado. Por ejemplo, si suponemos que las diferentes causas de una enfermedad actúan independientemente, en el sentido de que, tal como dice la definición de puerta MAX, el grado más probable de la enfermedad es el máximo de los que producirían las causas, entonces estamos en condiciones de aplicar nuestro modelo simplificado; si no es así, debemos recurrir al modelo general.

La segunda forma de justificar nuestro modelo consiste en realizar estudios empíricos sobre un amplio número de casos y ver hasta qué punto la puerta OR/MAX puede considerarse como aproximación satisfactoria, y en esto pueden utilizarse dos criterios. Uno de ellos, el

más estricto, consistiría en exigir que los resultados estadísticos para la familia X se ajustaran a los predichos por la expresión 3.82; el otro criterio, más flexible, se conformaría con que la red en su conjunto ofreciera diagnósticos acertados, dentro de ciertos límites, aunque las predicciones para la familia X no fueran completamente correctas.[9]

Por último, al hablar de la semántica debemos insistir en la diferencia que existe entre el modelo general y la puerta OR/MAX. Si volvemos al ejemplo 3.3, comprobamos que hay una interacción entre el país de origen y el tipo sanguíneo como *factores condicionantes* de la probabilidad de contraer paludismo. Sin embargo, en el ejemplo 3.17, tenemos dos *causas*, neumonía y paludismo, cada una de las cuales por sí misma es capaz de producir fiebre, interactuando mediante una puerta MAX. Por tanto, podemos afirmar que la puerta OR/MAX refleja mejor el concepto intuitivo de causalidad que utilizamos en nuestra vida cotidiana. Así, cuando decimos que "A es una **causa** de C" entendemos que "A produce o puede producir C". Nótese el contraste con el primer ejemplo, referido al caso general: los nodos País-de-origen y Tipo-sanguíneo son los padres del nodo Paludismo, pero nadie diría "el país de origen produce paludismo" y menos aún "el tipo de sangre produce paludismo", sino "el país de origen y el tipo de sangre son dos **factores condicionantes** que influyen en la probabilidad de contraer paludismo".

De esta diferencia se deduce una ventaja más de la puerta OR/MAX frente al modelo general, además de las que habíamos mencionado anteriormente: a la hora de generar *explicaciones lingüísticas*, si los nodos de la familia X interactúan mediante una puerta OR/MAX podemos decir "la causa que [con mayor probabilidad] ha producido X es U_i", "la presencia de U_i explica por qué se ha producido X, y por tanto ya no es necesario sospechar la presencia de U_j" o "al descartar U_i por dichas razones, aumenta nuestra sospecha de que la causa más probable de X es U_j". En el modelo general, no es posible —al menos no es fácil— generar este tipo de explicaciones a partir de una tabla de probabilidades.

3.5 Bibliografía recomendada

Dado que las redes bayesianas son un tema de gran actualidad, la bibliografía relevante es extensa y crece día a día. Entre los libros publicados destaca el de Judea Pearl [45], que es la obra de referencia principal. Otro libro que recomendamos encarecidamente es el editado por J. A. Gámez y J. M. Puerta, *Sistemas Expertos Probabilísticos* [22]; sus ventajas principales son: que cubre casi todos los aspectos de las redes bayesianas (algoritmos de propagación, aprendizaje automático, modelos temporales, aplicaciones médicas e industriales...), que está escrito con fines didácticos, que ha aparecido muy recientemente, por lo que contiene los resultados y las referencias bibliográficas más actuales, y, aunque ésta es una ventaja de orden secundario, que está escrito en castellano. Otros libros didácticos, aunque todos ellos con aportaciones originales, son el de Neapolitan [41], el de Castillo, Gutiérrez y Hadi [6],[10] el de Cowell y otros [10] y el de Jensen [31]. También la tesis doctoral de F. J. Díez Vegas [14], disponible en Internet, puede servir de introducción al tema. La aplicación de las redes bayesianas y los diagramas de influencia a la medicina está descrita en [33] y [16].

[9] Estos dos criterios pueden aplicarse también a sistemas expertos no bayesianos. Por ejemplo, en la evaluación de MYCIN [66] se tomó el segundo de ellos y se consideró que la realización del programa había sido un éxito. Sin embargo, con el criterio más estricto, se habría cuestionado la validez del programa, pues éste contiene numerosas inconsistencias, como explicamos en la sección 4.4.

[10] En `http://correo.unican.es/~gutierjm/BookCGH.html` puede obtenerse este libro de forma gratuita.

Javier Díez y Marek Druzdzel están escribiendo actualmente un artículo que explica detalladamente la puerta OR/MAX y otros modelos canónicos, como las puertas AND, MIN, XOR, etc. El lector interesado en la aplicación de redes bayesianas a problemas del mundo real podrá encontrar dicho artículo dentro de unos meses en la página `http://www.dia.uned.es/~fjdiez/public.html`.

Quien tenga acceso a la WWW puede encontrar numerosos enlaces de interés a partir de la página `http://www.ia.uned.es/~fjdiez/bayes/rbayes.html`. En particular, recomendamos al alumno que trabaje con alguno de los programas gratuitos para redes bayesianas que se indican en ella, especialmente con el programa Elvira, desarrollado conjuntamente por varias universidades españolas, que puede obtenerse en Internet en `http://www.ia.http://www.ia.uned.es/~elvira`.

Capítulo 4

Modelo de factores de certeza de MYCIN

El modelo de factores de certeza surgió ante la necesidad de que MYCIN, un sistema experto basado en reglas (sec. 4.1), fuese capaz de representar y razonar con la incertidumbre expresada por los médicos que participaban en el proyecto. En este capítulo explicamos cómo se definieron los factores de certeza (sec. 4.2) y cómo se combinan cuando se encadenan las reglas (sec. 4.3), y analizamos los problemas que plantea el modelo desde el punto de vista matemático (sec. 4.4).

4.1 El sistema experto MYCIN

4.1.1 Características principales

Para muchos, el programa DENDRAL, desarrollado en la Universidad de Stanford a partir de 1965 [2, 4] es el primer sistema experto, pues posee muchas de las características básicas de los sistemas expertos: dominio reducido (su objetivo era determinar la estructura de moléculas orgánicas mediante espectrometría de masas), separación entre conocimiento e inferencia (introdujo el uso de reglas para representar el conocimiento), razonamiento simbólico (utiliza conceptos como cetona y aldehído), formación de hipótesis y búsqueda heurística (una búsqueda exhaustiva sería imposible). Gracias a estas características alcanzó una eficacia superior a la de cualquier experto humano.

Sin embargo, otros consideran que el primer sistema experto fue MYCIN, también desarrollado en la Universidad de Stanford en la década de los 70 [3] como sistema de consulta para el diagnóstico y tratamiento de enfermedades infecciosas (recomendación de la terapia antimicrobiana más adecuada en cada caso). Es sin duda el sistema experto más famoso, el que ha dado lugar a más proyectos derivados y el que ha marcado el paradigma de todos los sistemas expertos de la actualidad. Entre las aportaciones de MYCIN destacan las siguientes:[1]

- separación entre *base de conocimientos*, que en su caso estaba formada por unas 400 reglas, *base de afirmaciones*, que es donde se almacenan temporalmente las conclusiones

[1] Puede encontrarse una explicación más completa y más detallada en cualquier libro de inteligencia artificial y sistemas expertos, p.ej. [40, cap. 6].

obtenidas, y *motor de inferencia*, que es la parte del programa encargada de combinar los datos y las reglas con el fin de obtener nuevas conclusiones, tal como muestra la figura 4.1 (en DENDRAL, las reglas estaban codificadas junto con el resto del programa);

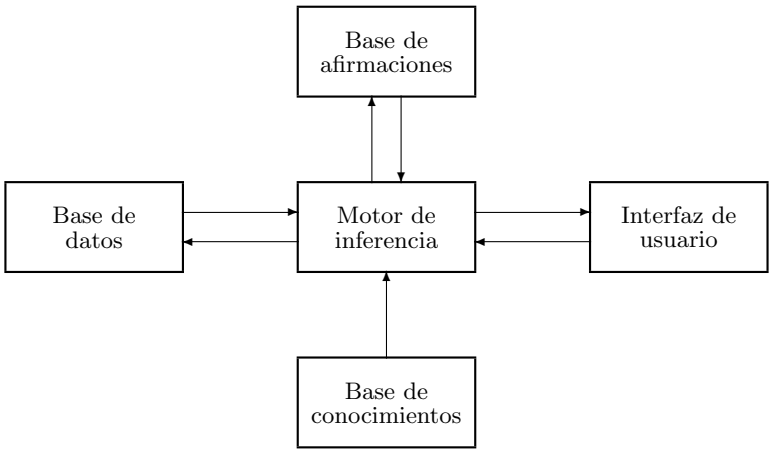

Figura 4.1: Estructura típica de un sistema basado en reglas.

- *encadenamiento hacia atrás* o basado en objetivos (DENDRAL utilizaba encadenamiento hacia delante);
- *diálogo (relativamente) flexible*: el usuario no sólo podía introducir información cuando el sistema lo solicitaba —como en los sistemas anteriores a MYCIN— sino también en cualquier otro momento;
- capacidad de *explicación del razonamiento*, mediante la traza de las reglas encadenadas (la capacidad de explicación es esencial, especialmente en un sistema experto médico);
- *tratamiento de la incertidumbre* mediante **factores de certeza**, que es el tema al que dedicamos este capítulo.

Con MYCIN se consolidó la utilización de reglas como método general para desarrollar sistemas expertos, aunque la posibilidad de programar la solución de un problema mediante reglas se conoce desde Post [51]; Newell y Simon la introdujeron en el campo de la inteligencia artificial en 1972 [42].

El procedimiento de inferencia en MYCIN consiste en buscar una regla que nos permita confirmar la hipótesis buscada. Por ejemplo, para saber si un organismo es bacteroide escogemos la regla siguiente [61, pág. 71]:

```
($AND (SAME CNTXT GRAM GRAMNEG)
      (SAME CNTXT MORPH ROD)
      (SAME CNTXT AIR ANAEROBIC))
(CONCLUDE CNTXT IDENTITY BACTEROIDES TALLY .6)
```

que significa[2]

SI el organismo es gram-negativo
Y tiene forma de bastón
Y es anaerobio
ENTONCES el organismo es bacteroide (con una certeza de 0'6).

Para poder ejecutar la regla y concluir que se trata de un organismo bacteroide, hace falta confirmar cada una de sus premisas. El sistema trata primero de ver si es gram-negativo buscando en la base de afirmaciones dicha aseveración; si no la encuentra, intentará aplicar otra regla de la que pueda deducir esta afirmación (y así sucesivamente, en lo que se conoce como *encadenamiento hacia atrás* de las reglas); si todas las reglas fallan, preguntará al usuario si el organismo es gram-negativo. Una vez confirmada la primera premisa, el sistema tratará de confirmar la segunda, y si se confirman todas, concluirá que es bacteroide; si no, examinará otras reglas. El encadenamiento hacia atrás corresponde a una inferencia *basada en hipótesis*, pues son éstas el objetivo que guía la búsqueda. En cambio, el *encadenamiento hacia delante* —el que empleaba DENDRAL— corresponde a una inferencia *basada en datos*, pues un dato confirmado puede disparar una o más reglas, lo cual puede provocar que se disparen otras, y así sucesivamente.

4.1.2 Motivación del modelo de factores de certeza

Como vimos en la sección 2.4.3, en los años 60 se empezó a utilizar el teorema de Bayes para la creación de sistemas expertos de diagnóstico médico. Sin embargo, este modelo fue criticado porque exigía hipótesis inverosímiles (exclusividad de los diagnósticos e independencia condicional), porque requería un número elevado de parámetros, generalmente difíciles de obtener, y porque no permitía estructurar la información (la "base de conocimientos" consistía en un montón de parámetros), con la consiguiente dificultad de refinar el modelo a medida que se obtuviera nueva información sobre las relaciones entre hallazgos y diagnósticos.

Por otro lado, los creadores de MYCIN, animados por el éxito alcanzado por DENDRAL, estaban intentando construir un programa basado en reglas, y necesitaban un modelo de tratamiento de la incertidumbre capaz de adaptarse a la modularidad de las reglas, es decir, capaz de asociar cierta información a cada regla y combinar localmente esa información a medida que se encadenasen las reglas. Claramente, el método bayesiano clásico (el único modelo probabilista disponible en aquel momento) estaba muy lejos de satisfacer tales condiciones. Por estas razones el tratamiento riguroso de la probabilidad resultaba inviable dentro de MYCIN.

Hubo además otra razón que llevó a buscar un modelo alternativo a la teoría de la probabilidad. Se trataba de la relación entre creencia a favor y creencia en contra. En principio esto puede parecer trivial, ya que la creencia en contra es lo opuesto de la creencia a favor. De hecho, la teoría de probabilidad nos dice que, definiendo las siguientes variables,

- E_1 = El organismo es gram-negativo
- E_2 = El organismo tiene forma de bastón

[2]El término `CNTXT` corresponde a una *variable* que se asociará al organismo correspondiente en cada caso. El uso de variables dentro de las reglas es una de las características que sitúa las reglas más cerca de la lógica de predicados que de la lógica proposicional (cf. cap. 5).

- E_3 = El organismo es anaerobio
- H = El organismo es bacteroide[3]

la regla mostrada en la sección anterior puede expresarse en términos de probabilidad a posteriori como

$$P(H|E_1 \wedge E_2 \wedge E_3) = 0'7 \tag{4.1}$$

lo cual implica que

$$P(\neg H|E_1 \wedge E_2 \wedge E_3) = 1 - 0'7 = 0'3 \tag{4.2}$$

Sin embargo, los médicos que colaboraban en el proyecto MYCIN no estaban de acuerdo con el hecho de que la primera igualdad implique la segunda, pues aunque $E_1 \wedge E_2 \wedge E_3$ aporte evidencia a favor de H eso no significa –según ellos— que aporte igualmente evidencia en contra de $\neg H$. La razón es que $P(H|E)$ procede de una relación de causa-efecto; en cambio, es posible que no exista ninguna relación causa-efecto entre E y $\neg H$, como sugiere la ecuación $P(\neg H|E) = 1 - P(H|E)$. Por eso, en el proyecto MYCIN se apreció la conveniencia de contar con una teoría que pudiera considerar por separado la evidencia a favor de una hipótesis (*confirmation*) y la evidencia en contra (*disconfirmation*).

Por todos estos motivos, Edward Shortliffe, uno de los investigadores principales del proyecto, empezó a desarrollar un método alternativo a la probabilidad, especialmente adaptado a las reglas, que fuera fácilmente computable y que considerase por separado la evidencia a favor y la evidencia en contra de cada hipótesis. Para ello se inspiró en la teoría de la confirmación de Carnap [5, pág. 19], quien distinguía dos tipos de probabilidad [5, pág. 19]:

- **Probabilidad-1**: "Es el *grado de confirmación* de una hipótesis H a partir de una aseveración de evidencia E; por ejemplo, un informe observacional. Es un concepto *lógico* semántico. Una afirmación sobre este concepto no se basa en la observación de hechos, sino en un análisis lógico" [énfasis añadido]. Es decir, se trata de una relación entre los *conceptos* E y H.
- **Probabilidad-2**: "Es la frecuencia relativa (a la larga) de una propiedad de sucesos o cosas respecto de otra cosa. Un aserto sobre este concepto es fáctico, empírico." Por tanto, es un concepto ligado a la frecuencia de eventos reproducibles.

De aquí se deduce que, para Carnap, la confirmación se basa en una implicación lógica. Sin embargo, los investigadores de MYCIN la interpretaron con más flexibilidad. Por ejemplo, así como la observación de un cuervo negro *confirmaría* (en el sentido de que *daría credibilidad* a) la hipótesis de que "todos los cuervos son negros", por el principio de inducción, Shortliffe y Buchanan [57] consideran que el hecho de que un organismo sea gram-positivo *confirma* la hipótesis de que es un estreptococo, aunque la conclusión esté basada en conocimiento empírico y no en un análisis lógico.

Por otro lado, Carnap [5] distinguía también tres formas de confirmación:

1. **clasificatoria:** "la evidencia E confirma la hipótesis H";

[3]Nótese que en el capítulo 2 H significaba **hallazgo**, es decir una *variable* cuyos valores eran $+h$ y $\neg h$, mientras que en éste va a significar siempre una **hipótesis**, de modo que la *proposición* correspondiente a la hipótesis se representará mediante $\neg H$.

2. **comparativa:** "E_1 confirma H en mayor medida que E_2 confirma H" o "E confirma más H_1que H_2";

3. **cuantitativa:** "E confirma H en grado x".

En MYCIN se utilizó una aproximación cuantitativa, aunque el objetivo último era comparativo: se trataba de que dos o tres identidades de organismos alcanzaran una confirmación mucho más fuerte que el resto, con lo cual las primeras constituirían el diagnóstico y recibirían el tratamiento terapéutico indicado. Por tanto, no importaba conocer la certeza absoluta correspondiente a cada hipótesis, sino saber si la certeza de unas pocas hipótesis era mucho mayor que la de las demás [57].

Esta observación es importante por la siguiente razón. En principio, si los factores de certeza miden la **variación** (aumento o disminución) de la credibilidad, para determinar la certeza con que se cree una hipótesis, habría que tomar la credibilidad inicial y agregar o descontar el efecto de la evidencia recibida. [Véase, por ejemplo, la ecuación (2.26), en que para calcular la probabilidad a posteriori se tiene en cuenta la probabilidad a priori (credibilidad inicial) y la verosimilitud (grado de confirmación de la hipótesis en función de la evidencia).] Sin embargo, el método de MYCIN prescinde de la credibilidad inicial y clasifica las hipótesis solamente en función del grado de confirmación aportado por la evidencia, apoyándose en el argumento expresado en el párrafo anterior. Volveremos sobre este punto en la sección 4.4.

4.2 Definición de los factores de certeza

4.2.1 Factor de certeza de cada regla

En MYCIN, el factor de certeza (FC) de cada regla "Si E entonces H" se definió como **grado de confirmación**, más concretamente como la diferencia entre la creencia a favor (*measure of belief*, MB) y la creencia en contra (*measure of disbelief*, MD):

$$MB(H,E) = \begin{cases} \dfrac{P(H|E) - P(H)}{1 - P(H)} & \text{si } P(H|E) \geq P(H) \\ 0 & \text{si } P(H|E) < P(H) \end{cases} \tag{4.3}$$

$$MD(H,E) = \begin{cases} 0 & \text{si } P(H|E) \geq P(H) \\ \dfrac{P(H) - P(H|E)}{P(H)} & \text{si } P(H|E) < P(H) \end{cases} \tag{4.4}$$

$$FC(H,E) = MB(H,E) - MD(H,E) = \tag{4.5}$$

$$= \begin{cases} \dfrac{P(H|E) - P(H)}{1 - P(H)} & \text{si } P(H|E) \geq P(H) \\ \dfrac{P(H|E) - P(H)}{P(H)} & \text{si } P(H|E) < P(H) \end{cases} \tag{4.6}$$

Observe que MB es una medida proporcional del **aumento** de la credibilidad (no de la credibilidad absoluta); más exactamente, es una medida de la disminución proporcional de la falta de credibilidad, $1 - P(H)$. Del mismo modo, MD es una medida proporcional del **aumento** de la creencia en contra de H; más exactamente, es una medida de la disminución proporcional de $P(H)$.

Los factores de certeza cumplen las siguientes propiedades:

1. Intervalos:

$$0 \leq MB \leq 1 \tag{4.7}$$
$$0 \leq MD \leq 1 \tag{4.8}$$
$$-1 \leq FC \leq 1 \tag{4.9}$$

   Los valores positivos de FC corresponden a un aumento en la creencia en una hipótesis, mientras que los valores negativos corresponden a una disminución en la creencia. Un FC positivo indica que la evidencia confirma (total o parcialmente) la hipótesis ya que $MB > MD$. Un FC negativo significa que la evidencia descarta (total o parcialmente) la hipótesis, ya que $MB < MD$.

2. Factor de certeza de la negación de una hipótesis:

$$MD(\neg H, E) = MB(H, E) \tag{4.10}$$
$$MB(\neg H, E) = MD(H, E) \tag{4.11}$$
$$FC(\neg H, E) = -FC(H, E) \tag{4.12}$$

3. Confirmación total (la evidencia confirma con seguridad absoluta la hipótesis):

$$P(H|E) = 1 \Longrightarrow \begin{cases} MB(H, E) = 1 \\ MD(H, E) = 0 \\ FC(H, E) = 1 \end{cases} \tag{4.13}$$

4. Exclusión total (la evidencia descarta la hipótesis con absoluta certeza):

$$P(H|E) = 0 \Longrightarrow \begin{cases} MB(H, E) = 0 \\ MD(H, E) = 1 \\ FC(H, E) = -1 \end{cases} \tag{4.14}$$

5. Falta de evidencia:

$$P(H|E) = P(H) \Longrightarrow \begin{cases} MB(H, E) = 0 \\ MD(H, E) = 0 \\ FC(H, E) = 0 \end{cases} \tag{4.15}$$

6. Límite probabilista:

$$P(H) \ll 1 \wedge P(H) \ll P(H|E) \Longrightarrow FC(H, E) \approx P(H|E) \tag{4.16}$$

En el modelo de factores de certeza no hay ligaduras entre los valores de $MB(H, E)$ y $MD(H, E)$; en efecto, por la propiedad 2 se cumple que

$$MB(H, E) + MB(\neg H, E) = MB(H, E) + MD(H, E) \tag{4.17}$$

pero no es necesario que $MB(H, E) + MB(\neg H, E)$ valga 0 ni valga 1 ni ningún otro valor predeterminado. [Nótese el contraste con la teoría de la probabilidad, en que se exige que $P(H|E) + P(\neg H|E) = 1$.]

4.2.2 Factor de certeza de cada valor

El modelo de MYCIN no sólo asigna un factor de certeza a cada regla de la base de conocimientos (cf. fig. 4.1), sino también a cada terna objeto-atributo-valor de la base de afirmaciones.[4] Algunas de las cuádruplas resultantes podrían ser éstas:

objeto	**atributo**	**valor**	**certeza**
paciente	nombre	Sisebuto-Gómez	$1'0$
organismo-1	forma	bastón	$0'8$
organismo-1	identidad	estafilococo	$0'2$
organismo-1	identidad	estreptococo	$-0'3$
organismo-2	forma	bastón	$-1'0$

Esto significa que tenemos certeza absoluta de que el nombre del paciente es Sisebuto Gómez, que hay fuerte evidencia que indica que el organismo-1 tiene forma de bastón, evidencia leve de que es un estafilococo y evidencia leve en contra de que sea un estreptococo; igualmente, existe certeza de que el organismo-2 no tiene forma de bastón.

4.3 Propagación de la evidencia en una red de inferencia

Cuando tenemos una regla "Si E entonces H", la confirmación E conlleva la confirmación de H, como vamos a ver en seguida. Cuando hay una red de reglas, la evidencia se propaga mediante la aplicación repetida de dos esquemas de simples: combinación convergente y combinación secuencial.

4.3.1 Modus ponens incierto

Supongamos que tenemos una regla "Si E entonces H, con $FC(H,E)$" y de algún modo concluimos E con certeza $FC(E)$. En este caso, podemos concluir H con una certeza $FC(H)$, que es función de $FC(E)$ y $FC(H,E)$. El mecanismo de inferencia se denomina *modus ponens*, y se puede representar así:

$$\frac{\begin{array}{l}\text{Si } E \text{ entonces } H, \text{ con } FC(H,E)\\ E, \text{ con } FC(E)\end{array}}{H, \text{ con } FC(H) = f_{mp}(FC(E), FC(H,E))}$$

donde

$$f_{mp}(x,y) = \begin{cases} x \cdot y & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases} \tag{4.18}$$

Es decir, la regla sólo se dispara cuando $FC(E) > 0$.

Por ejemplo, a partir de la regla "Si llueve entonces hace frío, con $FC = 0'6$" y la confirmación de "Llueve", con $FC(\text{"Llueve"}) = 0'8$, podemos concluir que $FC(\text{"Hace frío"}) = 0'48$.

[4] El hecho de que cada terna de la base de afirmaciones tiene un factor de certeza asociado es un hecho que muy pocas veces se menciona de forma explícita en los textos que describen el modelo de MYCIN (incluidas las referencias originales). Nosotros, en cambio, queremos resaltar esta realidad porque, a nuestro juicio, hace que se entienda mejor el método de inferencia en sí y la forma en que fue implementado.

Observe que $f_{mp}(1,1) = 1$, es decir,

$$[FC(E) = 1 \wedge FC(H,E) = 1] \Longrightarrow FC(H) = 1 \tag{4.19}$$

lo cual significa que el modus ponens clásico (sec. 5.1.1) es un caso particular del modus ponens de MYCIN.

Ejercicio 4.1 Dibuje la gráfica tridimensional de $f_{mp}(x,y) : [-1,1] \times [-1,1] \to [-1,1]$ o, lo que es lo mismo, la gráfica de $FC(H)$ en función de $FC(E)$ y $FC(H,E)$. □

En la práctica, puede ocurrir que una ligera evidencia a favor de E haga disparar una regla que, a la larga, apenas va a aumentar la certeza de H. Esto origina dos problemas: el primero es de eficiencia, pues el sistema "pierde tiempo" en cálculos poco significativos, y el segundo, más serio, es que, en reglas compuestas, puede llevar al sistema a plantear al usuario preguntas que a la larga van a resultar irrelevantes, con lo que el usuario desconfiará de la capacidad del sistema (y en un sistema destinado a la medicina es esencial contar con la confianza del usuario, es decir, el médico, porque de otro modo rechazará sistemáticamente el consejo que el sistema le ofrezca). Por estos dos motivos los creadores de MYCIN establecieron un umbral de 0'2, de modo que

$$f_{mp}(x,y) = \begin{cases} x \cdot y & \text{si } x > 0'2 \\ 0 & \text{si } x \leq 0'2 \end{cases} \tag{4.20}$$

4.3.2 Combinación de reglas convergentes

Supongamos que tenemos dos reglas, "Si E_1 entonces H, con $FC(H,E_1)$" y "Si E_2 entonces H, con $FC(H,E_2)$"; es decir, se trata de dos hallazgos distintos, E_1 y E_2, que apoyan la misma hipótesis H, tal como indica la figura 4.2. Es lo que se denomina a veces combinación de reglas en paralelo, o más propiamente, combinación de reglas convergentes.

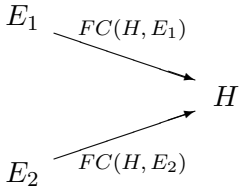

Figura 4.2: Combinación de reglas convergentes.

Si sabemos que E_1 y que E_2, podemos concluir H con una certeza

$$FC(H, E_1 \wedge E_2) = f_{conv}(FC(H,E_1), FC(H,E_2)) \tag{4.21}$$

como si las dos reglas anteriores se combinaran en una sola:

Si E_1 entonces H, con $FC(H,E_1)$
Si E_2 entonces H, con $FC(H,E_2)$

Si $E_1 \wedge E_2$ entonces H, con $FC(H, E_1 \wedge E_2)$

Las propiedades deseables para la combinación convergente de reglas son las siguientes:

1. Simétrica:
$$FC(H, E_1 \wedge E_2) = FC(H, E_2 \wedge E_1) \tag{4.22}$$

2. Asociativa:
$$FC(H, E_1 \wedge (E_2 \wedge E_3)) = FC(H, (E_1 \wedge E_2) \wedge E_3) \tag{4.23}$$

3. Monótona:
$$FC(H, E_2) > 0 \Longrightarrow FC(H, E_1 \wedge E_2) > FC(H, E_1) \tag{4.24}$$

4. Cancelación de evidencia contradictoria:
$$FC(H, E_1) = -FC(H, E_2) \neq 1 \Longrightarrow FC(H, E_1 \wedge E_2) = 0 \tag{4.25}$$

5. Evidencia nula:
$$FC(H, E_1) = 0 \Longrightarrow FC(H, E_1 \wedge E_2) = FC(H, E_2) \tag{4.26}$$

6. Predominio de la confirmación total:
$$FC(H, E_1) = 1 \Longrightarrow [FC(H, E_2) \neq -1 \Longrightarrow FC(H, E_1 \wedge E_2) = 1] \tag{4.27}$$

7. Predominio de la negación total:
$$FC(H, E_1) = -1 \Longrightarrow [FC(H, E_2) \neq 1 \Longrightarrow FC(H, E_1 \wedge E_2) = -1] \tag{4.28}$$

8. Contradicción:
$$[FC(H, E_1) = 1 \wedge FC(H, E_2) = -1] \Longrightarrow FC(H, E_1 \wedge E_2) \text{ no está definido} \tag{4.29}$$

Ejercicio 4.2 Trate de explicar intuitivamente cada una de estas propiedades. Por ejemplo, la primera significa que el orden en que se introduce la evidencia no afecta a la certeza de la conclusión. □

Función de combinación original

Inicialmente se utilizó la siguiente función de combinación para MYCIN:

$$f_{conv}(x, y) = \begin{cases} x + y(1 - x) & \text{si } x > 0 \text{ e } y > 0 \\ x + y(1 + x) & \text{si } x < 0 \text{ e } y < 0 \\ x + y & \text{si } -1 < x \cdot y \leq 0 \\ \text{no definida} & \text{si } x \cdot y = -1 \end{cases} \tag{4.30}$$

En esta definición, el primer caso corresponde a dos hallazgos positivos, es decir, la situación en que tanto E_1 como E_2 aportan evidencia a favor de H; el segundo caso corresponde a dos hallazgos negativos, y los casos tercero y cuarto a evidencia contradictoria.

Ejercicio 4.3 Dibuje la gráfica tridimensional de $f_{conv}(x, y) : [-1, 1] \times [-1, 1] \rightarrow [-1, 1]$ o, lo que es lo mismo, la gráfica de $FC(H, E_1 \wedge E_2)$ en función de $FC(H, E_1)$ y $FC(H, E_2)$.

Ejercicio 4.4 Demuestre que esta función cumple las 5 primeras de las ocho propiedades deseables que acabamos de enunciar, pero no cumple las tres últimas.□

El hecho de que la función de combinación utilizada originalmente en MYCIN no satisface el predominio de la confirmación total (propiedad 6) plantea un problema en la práctica, y es que aunque un hallazgo E_1 confirme con certeza absoluta la hipótesis H, es posible que un hallazgo dudoso E_2 anule la certeza aportada por E_1:

$$\frac{\begin{array}{l}\text{Si } E_1 \text{ entonces } H, \text{ con } FC(H,E_1)=1\\ \text{Si } E_2 \text{ entonces } H, \text{ con } FC(H,E_2)=-0'8\end{array}}{\begin{array}{l}\text{Si } E_1 \wedge E_2 \text{ entonces } H,\\ \qquad \text{con } FC(H,E_1\wedge E_2)=0'2\end{array}} \tag{4.31}$$

Análogamente, es posible también que un solo hallazgo negativo anule el efecto de un número cualquiera de hallazgos positivos mucho más fiables:

$$\frac{\begin{array}{l}\text{Si } E_1 \text{ entonces } H, \text{ con } FC(H,E_1)=0'99\\ \text{Si } E_2 \text{ entonces } H, \text{ con } FC(H,E_2)=0'99\\ \text{Si } E_3 \text{ entonces } H, \text{ con } FC(H,E_3)=0'999\\ \text{Si } E_4 \text{ entonces } H, \text{ con } FC(H,E_4)=-0'8\end{array}}{\begin{array}{l}\text{Si } E_1 \wedge E_2 \wedge E_3 \wedge E_4 \text{ entonces } H,\\ \qquad \text{con } FC(H,E_1\wedge E_2\wedge E_3\wedge E_4)=0'199999\end{array}} \tag{4.32}$$

Es decir, a pesar de tener tres reglas que confirman H con casi total seguridad —pues $FC(H,E_1\wedge E_2\wedge E_3)=0'9999999$— el factor de certeza resultante es menor que 0'2 (el umbral señalado en el apartado anterior), de modo que la cuarta regla, a pesar de ser incierta, es capaz de contrarrestar la evidencia aportada por las tres anteriores.

Para resolver este problema, se introdujo una nueva función de combinación, que es la siguiente.

Función de van Melle

Dado que la función de combinación original f_{conv} no cumplía las propiedades deseables, los diseñadores de MYCIN diseñaron una nueva, conocida como función de combinación de van Melle [62]:

$$f^{VM}_{conv}(x,y)=\begin{cases} x+y(1-x) & \text{si } x>0 \text{ e } y>0\\ x+y(1+x) & \text{si } x<0 \text{ e } y<0\\ \dfrac{x+y}{1-\min(|x|,|y|)} & \text{si } -1<x\cdot y\leq 0\\ \text{no definida} & \text{si } x\cdot y=-1\end{cases} \tag{4.33}$$

Se observa que la única diferencia respecto de f_{conv} aparece en el tercer caso, es decir, cuando tenemos un hallazgo a favor de H y otro en contra de H (salvo cuando los dos hallazgos aportan evidencia absoluta y contradictoria, que es el cuarto caso, en el cual la función no está definida).

Al aplicar esta nueva función a los ejemplos anteriores obtenemos $FC(H,E_1\wedge E_2)=1$ en vez de $0'2$ para el primero y $FC(H,E_1\wedge E_2\wedge E_3\wedge E_4)=0'9999995$ en vez de $0'199999$ para el segundo, con lo cual se evitan las inconsistencias ocasionadas por f_{conv}.

Ejercicio 4.5 Dibuje la gráfica tridimensional de $f^{VM}_{conv}(x,y) : [-1,1] \times [-1,1] \to [-1,1]$ y compárela con la de $f_{conv}(x,y)$, dada por la ecuación (4.30).

Ejercicio 4.6 Demuestre que esta función cumple las ocho propiedades deseables enunciadas anteriormente.

4.3.3 Combinación secuencial de reglas

Supongamos que tenemos dos reglas tales que el consecuente de una de ellas coincide con el antecedente de la otra: "Si A entonces B, con $FC(B,A)$" y "Si B entonces C, con $FC(C,B)$":

$$A \xrightarrow{FC(B,A)} B \xrightarrow{FC(C,B)} C$$

En estas condiciones, si de algún modo llegamos a confirmar A, podemos deducir B, y de ahí podemos deducir C; es como si las dos reglas anteriores dieran lugar a una nueva regla "Si A entonces C, con $FC(C,A)$":

$$\frac{\begin{array}{l}\text{Si } A \text{ entonces } B, \text{ con } FC(B,A)\\ \text{Si } B \text{ entonces } C, \text{ con } FC(C,B)\end{array}}{\text{Si } A \text{ entonces } C, \text{ con } FC(C,A)}$$

donde

$$FC(C,A) = \begin{cases} FC(B,A) \cdot F(C,B) & \text{si } FC(B,A) \geq 0 \\ 0 & \text{si } FC(B,A) < 0 \end{cases} \tag{4.34}$$

Ejercicio 4.7 Dibuje la gráfica tridimensional de $FC(C,A)$ en función de $FC(B,A)$ y $FC(C,B)$. □

Las propiedades de la combinación secuencial de reglas son las siguientes:

1. Asociativa:
$$[A \wedge (A \to B)] \wedge (B \to C) = A \wedge [(A \to B) \wedge (B \to C)] \tag{4.35}$$
(Esta expresión ha de entenderse en el sentido de que el valor de $FC(C)$ ha de ser el mismo tanto si se calcula primero $FC(B)$ a partir de la primera regla —mediante la ec. (4.18)— y luego se aplica la segunda, como si A se aplicara directamente sobre la regla resultante de unir ambas.)

2. Propagación con certeza total:
$$FC(B,A) = 1 \Longrightarrow FC(C,A) = F(C,B) \tag{4.36}$$

3. Propagación nula:
$$FC(B,A) \leq 0 \Longrightarrow FC(C,A) = 0 \tag{4.37}$$

Ejemplo 4.8 Tenemos el siguiente conjunto de reglas:

$$\begin{array}{l}\text{Si } A \text{ entonces } C, \text{ con } FC(C,A) = 0'8\\ \text{Si } B \text{ entonces } C, \text{ con } FC(C,B) = 0'5\\ \text{Si } C \text{ entonces } D, \text{ con } FC(D,C) = 0'7\\ \text{Si } D \text{ entonces } F, \text{ con } FC(F,D) = 0'9\\ \text{Si } E \text{ entonces } F, \text{ con } FC(E,F) = -0'3\end{array}$$

(La red de inferencia resultante se muestra en la figura 4.3.) Si conocemos A con una certeza de 0'9, y B y E con certeza total, ¿cuál es el factor de certeza resultante para cada una de las demás proposiciones?

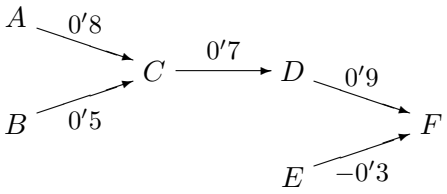

Figura 4.3: Pequeña red de inferencia.

Solución. Empezando por la izquierda, la primera regla nos permite asignar a C un factor de certeza de $FC(C) = FC(C, A) \cdot FC(A) = 0'9 \cdot 0'8 = 0'72$, de acuerdo con la ecuación (4.18). La segunda regla conduce a $FC(C) = FC(C, B) \cdot FC(B) = 0'5 \cdot 1 = 0'5$. Dado que se trata de dos reglas convergentes, combinamos ambos factores de certeza mediante la ecuación (4.30),[5] con lo que tenemos que $FC(C) = 0'72 + 0'28 \cdot 0'5 = 0'86$.

Aplicando de nuevo el modus ponens, tenemos que $FC(D) = FC(D, C) \cdot FC(C) = 0'7 \cdot 0'86 = 0'602$ y $FC(F) = FC(F, D) \cdot FC(D) = 0'9 \cdot 0'602 = 0'5418$. Ahora bien, tenemos por otro lado que $FC(F) = FC(F, E) \cdot FC(E) = (-0'3) \cdot 1 = -0'3$. Combinando estos dos últimos factores de certeza según la función de van Melle llegamos a

$$FC(F) = \frac{0'5418 - 0'3}{1 - \min(0'5418, 0'3)} = \frac{0'2418}{0'7} = 0'3454$$

Ejemplo 4.9 Mr. Holmes recibe una llamada telefónica de su vecino, el Dr. Watson, quien le dice que está oyendo sonar una alarma antirrobo en la casa de Mr. Holmes. Cuando va a salir corriendo, se acuerda de que el Dr. Watson tiene fama de ser un gran bromista, y por eso decide llamar primero a su vecina, Mrs. Gibbons, que es bastante más fiable, y ella le confirma que está sonando la alarma. ¿Con qué certeza puede creer Mr. Holmes que ha habido un (intento de) robo en su casa?

Solución. Este problema se podría plantear mediante las siguientes proposiciones:

- W: El Dr. Watson afirma que ha sonado la alarma
- G: Mrs. Gibbons afirma que ha sonado la alarma
- A: Ha sonado (realmente) la alarma
- R: Ha habido un intento de robo

y las siguientes reglas:

- R_1: Si W entonces A, $FC(A, F) = 0'5$

[5] En este caso da lo mismo aplicar la función de combinación original de MYCIN (ec. (4.30)) que la función de van Melle (ec. (4.33)) porque ambas coinciden cuando los dos argumentos son del mismo signo.

- R_2: Si G entonces A, $FC(A, G) = 0'9$
- R_3: Si A entonces R, $FC(R, A) = 0'99$

Sabemos que $FC(W) = FC(G) = 1$. A partir de ahí, combinando las dos reglas convergentes R_1 y R_2, se llega que $FC(A) = 0'9 + 0'1 \cdot 0'5 = 0'95$; es decir, del testimonio combinado del Dr. Watson y Mrs. Gibbons obtenemos certeza casi absoluta de que ha sonado la alarma. Finalmente, por la regla R_3 concluimos que $FC(R) = 0'95 \cdot 0'99 = 0'9405$, con lo que concluimos con certeza razonable que ha habido un intento de robo. □

4.3.4 Combinación de evidencia en el antecedente

En los apartados anteriores hemos visto cómo combinar reglas cuyo antecedente está formado por una cláusula simple. Sin embargo, los sistemas expertos con frecuencia necesitan representar reglas como "Si el paciente **no** es alérgico a la penicilina...", "Si el organismo es anaerobio **y** tiene forma de bastón...", "Si el organismo es un estreptococo **o** un estafilococo...", "Si la temperatura del condensador de vapor es mayor de 90º **o** la presión es mayor de 2 atmósferas...". La forma en que MYCIN trata este tipo de reglas consiste en tomar el opuesto del factor de certeza para la negación, el mínimo de los factores de certeza para la conjunción y el máximo para la disyunción:

$$FC(\neg E) = -FC(E) \tag{4.38}$$
$$FC(E_1 \wedge E_2) = \min(FC(E_1), FC(E_2)) \tag{4.39}$$
$$FC(E_1 \vee E_2) = \max(FC(E_1), FC(E_2)) \tag{4.40}$$

Es decir,

$$\frac{\begin{array}{l}\text{Si } \neg E \text{ entonces } H, \text{ con } FC = r \\ E, \text{ con } FC = x\end{array}}{H, \text{ con } FC = f_{mp}(-x, r)}$$

$$\frac{\begin{array}{l}\text{Si } E_1 \wedge E_2 \text{ entonces } H, \text{ con } FC = r \\ E_1, \text{ con } FC = x \\ E_2, \text{ con } FC = y\end{array}}{H, \text{ con } FC = f_{mp}(\min(x, y), r)}$$

$$\frac{\begin{array}{l}\text{Si } E_1 \vee E_2 \text{ entonces } H, \text{ con } FC = r \\ E_1, \text{ con } FC = x \\ E_2, \text{ con } FC = y\end{array}}{H, \text{ con } FC = f_{mp}(\max(x, y), r)}$$

En caso de expresiones anidadas, se aplican las ecuaciones anteriores sucesivamente. Por ejemplo,

$$\begin{aligned} &FC((E_1 \wedge E_2 \wedge E_3) \vee (E_4 \wedge \neg E_5)) \\ &\quad = \max(FC(E_1 \wedge E_2 \wedge E_3), FC(E_4 \wedge \neg E_5)) \\ &\quad = \max\{\min[FC(E_1), FC(E_2), FC(E_3)], \min[FC(E_4), -FC(E_5)\} \end{aligned}$$

4.4 Problemas del modelo de factores de certeza

Aunque MYCIN superó muy satisfactoriamente la evaluación, en la cual tuvo que medirse con los mejores expertos de su especialidad [66], el modelo de factores de certeza seguía teniendo serios problemas desde el punto de vista matemático. Los más importantes son éstos:

1. **Reglas equivalentes conducen a conclusiones distintas.** Las reglas "Si E_1 entonces H" y "Si E_2 entonces H" son equivalentes a "Si E_1 o E_2 entonces H". Supongamos que cada una de las dos primeras reglas tenga un factor de certeza de $0'5$; el factor de certeza de la regla combinada debe ser también $FC(H, E_1 \vee E_2) = 0'5$, para que se cumpla que $FC(H, E_1) = f_{mp}(FC(E_1 \vee E_2, E_1), FC(H, E_1 \vee E_2)) = \max(1, 0) \times 0'5 = 0'5$.

   Ahora bien, en caso de que conozcamos con certeza absoluta tanto E_1 como E_2, la combinación de las dos primeras reglas nos dice que $FC(H, E_1 \wedge E_2) = f^{VM}_{conv}(0'5, 0'5) = 0'75$, mientras que la aplicación de la regla combinada —equivalente a ellas según la lógica clásica— nos dice que $FC(H, E_1 \wedge E_2) = f_{mp}(FC(E_1 \vee E_2, E_1 \wedge E_2), FC(H, E_1 \vee E_2)) = \max(1, 1) \times 0'5 = 0'5 \neq 0'75$.

   El problema es que no hay ningún criterio para decidir cuál de estos dos resultados es más correcto, el $0'75$ o el $0'5$, porque tan correcto sería utilizar las reglas "Si E_1 entonces H" y "Si E_2 entonces H" como su equivalente "Si E_1 o E_2 entonces H".

2. **No considera correlaciones entre proposiciones.** Por ejemplo, supongamos que tenemos dos proposiciones, tales que FC("El organismo es un estreptococo") $= 0'8$ y FC("El organismo es un estafilococo") $= 0'3$. Dado que ambas son incompatibles, deberíamos tener que FC("El organismo es un estreptococo y un estafilococo") $= 0$. Sin embargo, la ecuación (4.38) nos dice que FC("estreptococo y estafilococo") $= \min(0'8, 0'3) = 0'3 \neq 0$.

3. **No considera correlaciones entre hallazgos.** Por ejemplo, supongamos que tenemos las tres reglas siguientes:

   Si llueve en Madrid, entonces llueve en toda España ($FC = 0'5$)
   Si llueve en León, entonces llueve en toda España ($FC = 0'5$)
   Si llueve en Barcelona, entonces llueve en toda España ($FC = 0'4$)

   Sabiendo que llueve en estas tres ciudades podemos concluir que FC("llueve en toda España") $= 0'82$. En cambio, si tenemos las reglas

   Si llueve en Madrid, entonces llueve en toda España ($FC = 0'5$)
   Si llueve en Móstoles, entonces llueve en toda España ($FC = 0'5$)
   Si llueve en Getafe, entonces llueve en toda España ($FC = 0'5$)
   Si llueve en Leganés, entonces llueve en toda España ($FC = 0'5$)

   y sabemos que llueve en cada una de estas cuatro localidades, podemos concluir que FC("llueve en toda España") $= 0'9375$, lo cual indica mayor certeza que en el caso anterior, 0'82, a pesar de que en este segundo caso los cuatro hallazgos están fuertemente correlacionados entre sí, por la proximidad geográfica entre las localidades, y por tanto la evidencia conjunta no es tan fuerte como en el primer caso, en que la correlación es mucho menor.

4. **Incoherencia de los valores calculados.** Supongamos que tenemos tres hipótesis, H_1, H_2 y H_3, y un hallazgo E que aporta evidencia a favor de cada una de ellas. Las probabilidades y los factores de certeza correspondientes son:

$$\begin{array}{lll} P(H_1) = 0'9 & P(H_1|E) = 0'9 & FC(H_1, E) = 0 \\ P(H_2) = 0'8 & P(H_2|E) = 0'9 & FC(H_2, E) = 0'5 \\ P(H_3) = 0'5 & P(H_3|E) = 0'9 & FC(H_3, E) = 0'8 \end{array}$$

Es decir, aunque $P(H_i|E)$ toma el mismo valor en los tres casos, al disminuir $P(H_i)$, aumenta $FC(H_i, E)$, lo cual es absurdo, pues si la verosimilitud es la misma en los tres casos, el FC debería ser mayor para la hipótesis más probable a priori; sin embargo, aquí ocurre lo contrario.

Por poner otro ejemplo, sean dos hipótesis H_1 y H_2, y un hallazgo E tales que

$$\begin{array}{lll} P(H_1) = 0'8 & P(H_1|E) = 0'9 & FC(H_1, E) = 0'5 \\ P(H_2) = 0'1 & P(H_2|E) = 0'7 & FC(H_2, E) = 0'67 \end{array}$$

Es decir, en caso de que $FC(E) = 1$ podemos concluir que $FC(H_2) = 0'67 > FC(H_1) = 0'5$, lo cual indica que tenemos más certeza en H_2 que en H_1, a pesar de que H_1 no sólo es mucho más probable a priori, sino también más verosímil.

5. **Falta de sensibilidad.** Comparemos estos dos casos:

Si $E_1 \wedge E_2 \wedge E_3$ entonces H, con $FC = 1$
E_1, con $FC = 1$
E_2, con $FC = 1$
E_3, con $FC = 0'001$

H, con $FC = 0'001$

Si $E_1 \wedge E_2 \wedge E_3$ entonces H, con $FC = 1$
E_1, con $FC = 0'001$
E_2, con $FC = 0'001$
E_3, con $FC = 0'001$

H, con $FC = 0'001$

Lo razonable sería que $FC(H)$ fuera mayor en el primer caso que en el segundo. Esta falta de sensibilidad se debe al hecho de tomar el mínimo de los factores de certeza para calcular la certeza de la conjunción (ec. (4.38)).

6. **Pseudo-independencia condicional.** Si tenemos dos reglas "Si E_1 entonces H" y

"Si E_2 entonces H" debe cumplirse que[6]

$$\frac{P(E_1 \wedge E_2|H)}{P(E_1 \wedge E_2)} = \frac{P(E_1|H)}{P(E_1)} \cdot \frac{P(E_2|H)}{P(E_2)} \tag{4.42}$$

$$\frac{P(E_1 \wedge E_2|\neg H)}{P(E_1 \wedge E_2)} = \frac{P(E_1|\neg H)}{P(E_1)} \cdot \frac{P(E_2|\neg H)}{P(E_2)} \tag{4.43}$$

Nótese que la independencia condicional exigida por el método probabilista clásico sería (cf. ec. (2.41))

$$P(E_1 \wedge E_2|H) = P(E_1|H) \cdot P(E_2|H) \tag{4.44}$$

$$P(E_1 \wedge E_2|\neg H) = P(E_1|\neg H) \cdot P(E_2|\neg H) \tag{4.45}$$

mientras que, en general, $P(E_1 \wedge E_2) \neq P(E_1) \cdot P(E_2)$. Estas dos últimas igualdades **solamente** pueden justificarse cuando hay dos mecanismos causales independientes $H \rightarrow E_1$ y $H \rightarrow E_2$. Ésta es la razón principal por la que el método probabilista clásico fue tan criticado y una de las razones que llevó a los diseñadores de MYCIN a crear un modelo alternativo.

Ahora bien, como acabamos de ver, para que la combinación de reglas convergentes en MYCIN sea válida, deben cumplirse las ecuaciones (4.42) y (4.43), que son muy similares a la denostada condición de independencia condicional, pero con el inconveniente de que no pueden justificarse ni siquiera en el caso de mecanismos causales independientes. Dicho coloquialmente, "salimos de Guatemala para meternos en Guatepeor".

4.4.1 Creencia absoluta frente a actualización de creencia

El punto 4 de los que acabamos de señalar es el que produce los resultados más contrarios al sentido común y a la teoría de la probabilidad. La causa de este comportamiento incorrecto es

[6]La demostración es la siguiente. De las ecuaciones (4.21) y (4.30) se deduce que

$$FC(H, E_1 \wedge E_2) = FC(H, E_1) + [1 - FC(H, E_1)] \cdot FC(H, E_2)$$

lo cual es equivalente a

$$1 - FC(H, E_1 \wedge E_2) = [1 - FC(H, E_1)] \cdot [1 - FC(H, E_2)] \tag{4.41}$$

Por otro lado, la ecuación (4.6) nos dice que

$$1 - FC(H, E) = 1 - \frac{P(H, E) - P(H)}{1 - P(H)} = \frac{P(\neg H|E)}{P(\neg H)}$$

y teniendo en cuenta que $P(\neg H|E) \cdot P(E) = P(E, \neg H) = P(E|\neg H) \cdot P(\neg H)$, llegamos a

$$\frac{P(E|\neg H)}{P(E)} = \frac{P(\neg H|E)}{P(\neg H)} = 1 - FC(H, E)$$

lo que sustituido en (4.41) nos dice que

$$\frac{P(E_1 \wedge E_2|\neg H)}{P(E_1 \wedge E_2)} = \frac{P(E_1|\neg H)}{P(E_1)} \cdot \frac{P(E_2|\neg H)}{P(E_2)}$$

Así se prueba la ecuación (4.43).

Esta demostración sigue siendo válida si sustituimos H por $\neg H$, y viceversa, con lo que se demuestra también la ecuación (4.42).

la confusión entre **creencia absoluta** ("la evidencia E, ¿con qué certeza confirma la hipótesis H?") y **actualización de la creencia** ("¿cuánto ha aumentado o disminuido la creencia en la hipótesis H al conocer la evidencia E?").[7]

Por un lado, el factor de certeza de cada proposición fue interpretado como creencia absoluta, de modo que, si $FC(H_1) \gg FC(H_2)$, se tomaba H_1 como conclusión, descartando H_2 (cf. sec. 4.2.2). En cambio, los factores de certeza de las reglas fueron definidos —a partir de la probabilidad— como medida de la actualización de creencia (cf. sec. 4.2.2, especialmente, ec. (4.6)), aunque en la práctica no se obtuvieron a partir de probabilidades ni objetivas ni subjetivas, sino que fueron estimados directamente por los expertos humanos (los médicos) que participaban en el proyecto. Más aún, la forma habitual de asignar un factor de certeza a la regla "Si E entonces H"no consistía en preguntar a los expertos "Dada la evidencia E, ¿en qué medida aumenta nuestra creencia en H?" (actualización de creencia), sino "Dada la evidencia E, ¿con qué certeza podemos concluir H?" (creencia absoluta). Es decir, el factor de certeza de cada regla, aunque definido como actualización de creencia, fue obtenido como creencia absoluta.

Esta confusión se ha producido en varios campos. Los filósofos de la ciencia suelen llamar *estudio de la confirmación* a la investigación de problemas relacionados con la creencia (en la sección 1.2 hemos hablado de la evolución del concepto de probabilidad: en algunas épocas, como la actual, predomina cada vez con más fuerza la interpretación subjetivista, mientras que en otras ha predominado la interpretación objetiva-frecuentista). El esfuerzo de los filósofos por describir creencia subjetiva, frecuencias y cambios en la creencia, produjo confusión acerca de la diferencia entre creencia absoluta y actualización en la creencia. Carnap utilizó el término *grado de confirmación* para resaltar la diferencia entre la interpretación subjetiva y la interpretación clásica frecuencial. Pero, en realidad, Carnap y otros emplearon la expresión *grado de confirmación* para referirse a dos conceptos muy diferentes: la probabilidad subjetiva y la actualización de la creencia. Cuando corrigieron este error, era ya demasiado tarde para evitar la confusión que se había creado.

Muchas de las paradojas históricas de teoría de la confirmación tienen su raíz en la confusión entre estos dos conceptos. Algunas de estas paradojas llevaron a la conclusión de que la probabilidad era incapaz de abordar los aspectos esenciales de la confirmación. Carnap, Barker, Harré y otros, al ver la confirmación como análoga a la probabilidad absoluta, se explayaron en lo que parecían ser aspectos intrigantes de la relación entre la confirmación de una hipótesis y la confirmación de la negación de la hipótesis.

Posteriormente, Popper y otros han señalado que la teoría de la probabilidad permite medidas de actualización de la creencia y, por tanto, es un instrumento útil para el estudio de la confirmación. Más tarde, Good demostró que varias expresiones probabilistas satisfacen una versión ligeramente modificada de los axiomas de Popper para medidas del cambio en la creencia.

4.4.2 La supuesta modularidad de las reglas

Hemos comentado ya que la ventaja más pregonada de las reglas es su modularidad, es decir la capacidad de representar el conocimiento mediante reglas independientes y de propagar la evidencia mediante computaciones locales ("locales" significa que al encadenar dos reglas no

[7]Como hemos señalado ya, en el método probabilista clásico la diferencia estaba clara: $P(H|E)$ era la credibilidad absoluta mientras que la verosimilitud $\lambda_H(E) = P(E|H)$ era el factor que indicaba la actualización de la creencia, pues $P(H|E) = \alpha \cdot P(H) \cdot \lambda_H(E)$.

debemos preocuparnos ni de las demás reglas contenidas en la base de conocimientos ni de las demás hipótesis confirmadas o descartadas hasta ese momento). En otro lugar [14, sec. 2.4.1] hemos discutido los problemas que esta modularidad origina en cuanto al mantenimiento de la base de conocimientos, debido a la falta de estructura. Vamos a señalar ahora los problemas que origina en cuanto al tratamiento de la incertidumbre. (Un tratamiento más detallado puede encontrarse en [45, sec. 1.2] y en [41, cap. 4].)

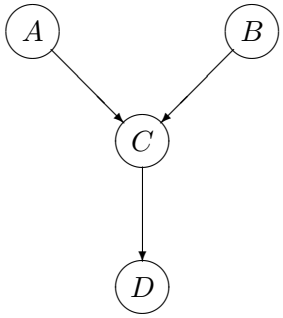

Figura 4.4: Nodo C con dos causas y un efecto.

Empezaremos con un ejemplo muy sencillo. Supongamos que una alteración C puede tener dos causas A y B, y un efecto D, como indica la figura 4.4. Al intentar construir un sistema basado en reglas que nos permita realizar inferencias sobre este modelo tan simple, encontramos varias dificultades.

La primera consiste en que las reglas no permiten establecer inferencias bidireccionales. En efecto, según este modelo de reglas, la presencia de C nos hace sospechar la presencia de A como mecanismo causante. Recíprocamente, si descubrimos A, pensaremos que probablemente ha producido C. Ahora bien, si incluimos en la base de conocimientos dos reglas, "$A \to C$" y "$C \to A$", el aumento de la credibilidad de una de ellas aumenta la credibilidad de la otra, y viceversa, dando lugar a un ciclo sin fin.

Existen dos formas de intentar solucionarlo. Una de ellas consiste en incluir sólo una de las dos reglas, con lo cual estamos limitando la capacidad de inferencia de nuestro sistema. La otra consiste en no actualizar la credibilidad de cada proposición o variable más que una sola vez (ésta es la solución que suelen adoptan las herramientas comerciales para la construcción de sistemas expertos). El inconveniente de esta segunda opción es que entonces las actualizaciones de la credibilidad no se transmiten a las reglas encadenadas. Así, en nuestro ejemplo anterior, una vez que la regla "$A \to C$" dispara la regla "$C \to D$", la credibilidad de D no se modificará aunque B aporte nueva evidencia a favor o en contra de C. El resultado es que el orden de llegada de la información puede influir en las conclusiones obtenidas.

La segunda dificultad reside en que las reglas no distinguen entre causas y efectos. En el ejemplo anterior, vemos que tanto A como B como D pueden aportar evidencia a favor de C, ¡pero de forma diferente! En efecto, al observar D aumenta la credibilidad de C y, por tanto, aumenta la credibilidad de B (y también de A) como causa posible. Supongamos que luego observamos también A. De nuevo aumenta la credibilidad de C, pero en este caso la credibilidad de B no aumentará, sino que disminuirá, pues A puede explicar la presencia de C, con lo cual disminuye nuestra sospecha sobre B (este mecanismo se denomina en inglés *explaining away*, y es típico de la puerta OR probabilista). En resumen: un aumento en la

credibilidad de C puede conducir a un aumento o a una reducción de la credibilidad de B, dependiendo de cuál sea el origen de la evidencia. Sin embargo, en un sistema basado en reglas, "$C \rightarrow B$" hará aumentar la credibilidad de B, sea cual sea el origen de la evidencia.

Por supuesto, podemos escribir reglas más sofisticadas, tales como "$(C \wedge \neg A) \rightarrow B$", pero entonces el sistema se vuelve mucho más complejo, pues los casos imaginables son casi ilimitados, por lo que en la práctica esta solución se hace inviable.

La tercera dificultad está relacionada con la anterior. Imaginemos que hemos deducido B a partir de D, pero luego observamos A, con lo cual debe disminuir la credibilidad de B. Sin embargo, no nos está permitido incluir la regla "$A \rightarrow \neg B$ ni otras similares, pues la presencia de A por sí misma no excluye B: ambas causas pueden estar presentes.

Existe, finalmente, un cuarto problema. Supongamos que tenemos dos informes médicos (o dos indicios de cualquier clase) que apuntan a un mismo diagnóstico. Ahora bien, si descubrimos que el segundo de ellos se basó en el diagnóstico del primero, la fiabilidad conjunta de ambos informes es menor que si los dos médicos hubieran llegado a la misma conclusión independientemente (recuérdese el ejemplo del punto 3 de la sección 4.4, pág. 90). La incapacidad de tratar fuentes de evidencia correlacionadas es otra de las limitaciones de los sistemas basados en reglas.

Heckerman y Henrion [28, 29] llegaron a la conclusión de que el origen de los problemas mencionados consiste en aplicar al razonamiento con incertidumbre un método, el encadenamiento de reglas, que sólo es válido en el campo de la lógica, donde todas las proposiciones son ciertas o falsas, pero nunca dudosas. En efecto, en el campo de la lógica sí existe la *modularidad semántica*, la cual significa que podemos deducir una conclusión a partir de unas premisas

—independientemente de cómo fueron deducidas dichas premisas, e

—independientemente de que existan otras proposiciones o reglas.

Estas dos propiedades se denominan desacoplo (*detachment*) y localidad (*locality*), respectivamente [45, pág. 5]. Sin embargo, en el razonamiento incierto no se cumplen dichas propiedades, tal como hemos discutido en los ejemplos anteriores. De aquí se deduce que **solamente es correcto utilizar reglas en dominios deterministas**, pues éstas son incapaces de tratar las correlaciones que surgen a causa de la incertidumbre.

4.4.3 ¿Por qué MYCIN funcionaba tan bien?

En vista de las graves incoherencias del modelo de factores de certeza, resulta sorprendente el gran éxito que alcanzó el sistema experto MYCIN, que en las pruebas de evaluación realizadas demostró que sus diagnósticos y recomendaciones terapéuticas eran al menos tan buenos como los de los mejores expertos de la especialidad [66].

Una de las razones puede ser la poca sensibilidad de los resultados de MYCIN frente a variaciones numéricas en el factor de certeza; en efecto, Clancey y Cooper [3, págs. 218–219] demostraron en 1979 que, aunque los factores de certeza en vez de variar entre -1 y 1 sólo pudieran tomar $n+1$ valores $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ (lo que obligaba a redondear todos los FC), el factor de certeza resultante para las hipótesis principales solía variar, pero rara vez variaba la ordenación entre ellas, de modo que en la mayor parte de los casos el organismo diagnosticado y la terapia recomendada eran los mismos que cuando los factores de certeza podían tomar cualquier valor del intervalo $[-1, 1]$. Solamente cuando $n \leq 3$ se producía una degradación significativa en los resultados, lo cual demuestra la poca sensibilidad ante variaciones en los

valores de los factores de certeza.

Otra de las razones puede ser que en medicina las hipótesis suelen tener una prevalencia pequeña ($P(H) \ll 1$), de modo que cada hallazgo generalmente produce un importante incremento relativo en la probabilidad ($P(H|E) \gg P(H|E)$) lo cual implica que $FC(H,E) \approx P(H|E)$ (cf. ec. (4.16)), por lo que en realidad muchos de los factores de certeza estimados se corresponden de cerca con la probabilidad a posteriori; dicho de otro modo, pocas veces se da la situación indicada en el punto 4 de la página 91, con lo cual se evita una de las principales fuentes de incoherencia del sistema.

4.5 Bibliografía recomendada

La obra de referencia obligada es el libro de Buchanan y Shortliffe [3], *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, en que se recopilaron, actualizaron y discutieron los artículos más importantes publicados dentro del proyecto MYCIN. Este libro, todo un "clásico", tiene la rara virtud de que, a pesar de lo mucho que se ha investigado campo de los sistemas expertos desde entonces, sigue conservando prácticamente intacto todo su interés, lo cual demuestra la calidad del trabajo realizado. Especialmente, en la parte dedicada a los de factores de certeza, además de incluir una selección de artículos revisados, añade un capítulo retrospectivo sobre el modelo y una selección de los mensajes de correo electrónico en que los investigadores discutían la problemática del modelo. En estos mensajes y en el capítulo 12, de Adams [1], se puede ver cómo los propios creadores del modelo eran conscientes ya entonces de algunas de las incoherencias que hemos señalado en la sección 4.4.[8]

El lector interesado puede encontrar más información y numerosas referencias en [45, sec. 1.2], [41, cap. 4] y [14, sec. 2.4].

[8] Lo que resulta más curioso es que Buchanan y Shortliffe, insatisfechos con el modelo de factores de certeza, incluyeron en su libro un capítulo sobre la teoría de Dempster-Shafer [23], a pesar de no se había utilizado en absoluto dentro del proyecto MYCIN; el motivo es que estaban convencidos en aquel momento (1984) de que esta teoría podía constituir una forma de corregir y salvar el modelo de factores de certeza. De hecho, cuando David Heckerman inició el doctorado, su director de tesis, Edward Shortliffe, le indicó que estudiase la teoría de Dempster-Shafer con este fin (comunicación personal de D. Heckerman al autor). Sin embargo, lo que Heckerman demostró es que es imposible construir un modelo de factores de certeza coherente, salvo para casos triviales [26], y a partir de ese momento decidió utilizar redes bayesianas en su tesis doctoral [27]. Heckerman es hoy en día uno de los científicos más destacados de este campo y el director del grupo de investigación de Microsoft sobre redes bayesianas aplicadas a la informática. Shortliffe, por su parte ha dicho en varias ocasiones (por ejemplo, en [11]) que, si hoy tuviera que construir MYCIN de nuevo, utilizaría redes bayesianas.

Capítulo 5

Lógica difusa

A lo largo de este capítulo, la lógica de proposiciones (sección 5.1) nos va a servir de base para la lógica de predicados (sección 5.2), y ésta, a su vez, para la teoría de conjuntos (sección 5.3). Esto nos permitirá justificar el isomorfismo existente entre la lógica y la teoría de conjuntos. Además, vamos a establecer el paralelismo entre la lógica de proposiciones precisas, predicados unitarios precisos y conjuntos nítidos, por un lado, y las proposiciones imprecisas, predicados unitarios imprecisos y conjuntos difusos, por otro. Análogamente, mostraremos la conexión entre los predicados n-arios y las relaciones entre conjuntos (sección 5.4), lo cual nos servirá para justificar los mecanismos de inferencia difusos (sección 5.4.3).

5.1 Lógica de proposiciones

Antes de explicar la lógica difusa (en sentido restringido) vamos a empezar repasando la lógica clásica. Suponemos que el lector ya está familiarizado con ella, y por eso nos vamos a limitar a una breve presentación, pues una exposición detallada requeriría todo un libro.

El punto de partida de la **lógica proposicional** son las *proposiciones elementales*. Cada proposición elemental corresponde a una frase simple del lenguaje natural, como "está lloviendo" o "7 es impar". A partir de las proposiciones elementales se pueden formar *proposiciones compuestas* mediante la aplicación de *conectivas*, que suelen ser cinco: una conectiva unitaria —la negación ($\neg$)— y cuatro conectivas binarias: la conjunción ($\wedge$), la disyunción ($\vee$), la implicación ($\rightarrow$) y la doble implicación ($\leftrightarrow$). Por ejemplo, si p representa la proposición "7 es impar" y q representa "7 es menor que 4", la proposición $\neg p$ significa "7 **no** es impar", $p \wedge q$ significa "7 es impar **y** menor que 4", $p \vee q$ significa "7 es impar **o** menor que 4", $p \rightarrow q$ significa "7 es impar **implica** que es menor que 4" y $p \leftrightarrow q$ significa "7 es impar **si y sólo si** es menor que 4".[1]

[1] Conviene señalar, sin embargo, que la implicación matemática (representada por "$\rightarrow$") no siempre coincide con la implicación que nosotros entendemos habitualmente, y que en este texto estamos representando mediante "$\Longrightarrow$". Por ejemplo, la proposición "$(7 < 3) \rightarrow (2^2 = 4)$" se considera cierta en la lógica matemática, como veremos en seguida, mientras que la afirmación supuestamente equivalente "$(7 < 3)$ implica que $(2^2 = 4)$" es absurda. El análisis de esta cuestión es sumamente complejo, hasta el punto de que dio lugar a una nueva rama de la lógica, la *lógica modal*, con el fin de representar adecuadamente la implicación; de hecho, no hay una única lógica modal, sino distintas variantes, lo cual indica que ninguna de ellas ha conseguido dar una solución completamente satisfactoria al problema.

Nosotros, en este texto, cuando escribimos la implicación natural ($p \Longrightarrow q$) estamos indicando dos cosas: *(1)* que tanto p como q son contingentes, es decir, son ciertas en unos casos y falsas en otros, y *(2)* que siempre que p es cierta, q también lo es; más adelante, al introducir las proposiciones imprecisas, el significado exacto

A cada una de las proposiciones elementales p se le asigna un **valor de verdad** $v(p)$ dado por una **función de verdad** v. En la **lógica clásica**,

$$v(p) \in \{0, 1\} \tag{5.1}$$

de modo que $v(p) = 1$ significa que p es verdadera y $v(p) = 0$ que es falsa.

En las **lógicas n-valuadas**,

$$v(p) \in \mathcal{V}_n = \left\{0, \frac{1}{n-1}, \frac{2}{n-1}, \cdots, \frac{n-2}{n-1}, 1\right\} \tag{5.2}$$

En la **lógica difusa**,

$$v(p) \in [0, 1] \tag{5.3}$$

En cualquier caso, los valores de verdad están siempre entre 0 y 1.

Dado que cuando p es verdadera, $v(p) = 1$, **afirmar p es lo mismo que decir que $v(p) = 1$**:[2]

$$p \Longleftrightarrow v(p) = 1 \tag{5.4}$$

Por el contrario, cuando p es falsa, $v(p)$=0. Por ejemplo, $v($"7 es impar"$)$=1, $v($"7 es menor que 4"$)$=0. Los valores de $v(p)$ más próximos a 1 indican que la proposición p es más verdadera, y viceversa. Por ejemplo, $v($"una persona que mide 1'80 es alta"$) > v($"una persona que mide 1'75 es alta"$)$.

Por tanto, cada función v representa nuestro estado de conocimiento acerca del mundo. Es posible que dos personas distintas tengan creencias diferentes, o que las creencias de una persona varíen con el tiempo, o que se pueda razonar sobre mundos hipotéticos, y por ello puede haber distintas funciones de verdad v. Por ejemplo, es posible tener dos v's diferentes tales que $v_1($"Pedro es hermano de Luis"$)$=1 y $v_2($"Pedro es hermano de Luis"$)$=0.

Dada una función de verdad v, una proposición p se denomina *precisa* cuando $v(p) \in \{0, 1\}$; se denomina *imprecisa (en sentido amplio)* cuando $v(p) \in [0, 1]$ e *imprecisa (en sentido estricto)* cuando $v(p) \in (0, 1)$. Más adelante veremos que proposiciones como "Juan es alto" o "la sopa está caliente" son imprecisas, porque generalmente no son ni totalmente ciertas ni totalmente falsas.

Hay dos proposiciones especiales: la *proposición segura*, que se representa mediante $\mathbf{1}$, siempre es verdadera (en todos los mundos posibles); es decir, $\forall v$, $v(\mathbf{1}) = 1$; del mismo modo, la *proposición imposible*, que se representa mediante $\mathbf{0}$, siempre es falsa: $\forall v$, $v(\mathbf{0}) = 0$. Ambas son proposiciones precisas.

El valor de verdad de una **proposición compuesta** se obtiene a partir de los *valores de verdad* de las *proposiciones* que la componen y de las *funciones* que definen las *conectivas* que las unen, tal como indican las siguientes ecuaciones:

$$v(\neg p) = f_\neg(v(p)) \tag{5.5}$$
$$v(p \wedge q) = f_\wedge(v(p), v(q)) \tag{5.6}$$
$$v(p \vee q) = f_\vee(v(p), v(q)) \tag{5.7}$$
$$v(p \rightarrow q) = f_\rightarrow(v(p), v(q)) \tag{5.8}$$
$$v(p \leftrightarrow q) = f_\leftrightarrow(v(p), v(q)) \tag{5.9}$$

---

de $p \Longrightarrow q$ es que siempre que $v(p) = 1$ entonces también $v(q) = 1$.

Algo semejante podemos decir de la diferencia entre la doble-implicación matemática ($\leftrightarrow$) y la equivalencia lógica ($\Longleftrightarrow$).

[2] Conviene tener esto muy presente a lo largo de todo el capítulo para entender correctamente muchas de las proposiciones que vamos a enunciar.

Distintas elecciones de estas funciones f dan lugar a diferentes lógicas (por ejemplo, la lógica de Łukasiewicz se basa en unas funciones distintas a las de la lógica de Kleene, como veremos más adelante).

Equivalencia entre proposiciones

Dos proposiciones p y q son *equivalentes* cuando toman el mismo valor de verdad cualquiera que sea la asignación de valores de verdad para las proposiciones elementales

$$p \equiv q \iff \forall v,\ v(p) = v(q) \tag{5.10}$$

[El primer ejemplo de equivalencia que vamos a encontrar es la propiedad de involución de la lógica clásica, "$\neg\neg p \equiv p$", pues $v(\neg\neg p) = v(p)$ tanto si $v(p) = 0$ como si $v(p) = 1$.]

Naturalmente, para que dos proposiciones p y q distintas sean equivalentes al menos una de ellas ha de ser compuesta, pues si las dos fueran elementales podríamos escoger una función v tal que $v(p) = 1$ y $v(q) = 0$, de modo que ya no serían equivalentes.

También debemos señalar que es posible que dos proposiciones (compuestas) sean equivalentes en una lógica y no equivalentes en otra. [Por ejemplo, la equivalencia $p \to q \equiv \neg(p \wedge \neg q)$ es cierta en la lógica clásica y en la de Kleene, pero no en la de Łukasiewicz.]

Las proposiciones que son equivalentes a **1**, es decir, aquéllas cuyo valor de verdad siempre es 1, se denominan *tautologías*. [En seguida veremos que "$p \vee \neg p$" es una tautología en la lógica clásica, porque $v(p \vee \neg p) = 1$ tanto si $v(p) = 1$ como si $v(p) = 0$.] En cambio, las que son equivalentes a la proposición **0**, es decir, aquéllas cuyo valor de verdad siempre es 0, se denominan *contradicciones*. [En seguida veremos que "$p \wedge \neg p$" es una contradicción en la lógica clásica, porque $v(p \wedge \neg p) = 0$ tanto si $v(p) = 1$ como si $v(p) = 0$.] Naturalmente, aparte de **1** y **0**, sólo las proposiciones compuestas pueden ser tautologías o contradicciones, pues a las proposiciones elementales siempre se les puede asignar tanto el valor de verdad 0 como 1.

Las propiedades básicas de la equivalencia de proposiciones vienen dadas por la tabla 5.1.

Reflexiva	$p \equiv p$
Simétrica	$p \equiv q \Longrightarrow q \equiv p$
Transitiva	$(p \equiv q \wedge q \equiv r) \Longrightarrow p \equiv r$

Tabla 5.1: Propiedades de la equivalencia de proposiciones.

Por cierto, nótese que "$p \equiv q$" es una proposición precisa, pues siempre es cierta o falsa, y no admite grados de verdad intermedios.

Tipos de implicación y doble-implicación

Antes de estudiar distintos modelos lógicos, vamos a introducir los distintos tipos de implicación y doble-implicación que aparecen en la tabla 5.2, la cual ha de leerse así: una *implicación rigurosa* es aquélla que cumple que $p \to q \iff \neg p \vee q$ o, lo que es lo mismo,[3] $f_{\to}(a, b) = 1 \iff a = 0 \vee b = 1$, y así sucesivamente.

[3]Insistimos en que, aunque "$p \to q$" es generalmente una proposición imprecisa, la equivalencia "$\iff$" se aplica sólamente entre proposiciones totalmente ciertas; es decir, la afirmación $p \to q \iff \neg p \vee q$ significa que $v(p \to q) = 1$ si y sólo si $v(\neg p \vee q) = 1$.

Implicación rigurosa	$p \to q \iff \neg p \vee q$ $f_{\to}(a,b) = 1 \iff a = 0 \vee b = 1$
Implicación amplia	$p \to q \iff v(p) \leq v(q)$ $f_{\to}(a,b) = 1 \iff a \leq b$
Doble-implicación rigurosa	$p \leftrightarrow q \iff [p \wedge q] \vee [\neg p \wedge \neg q]$ $f_{\leftrightarrow}(a,b) = 1 \iff [a = b = 1 \vee a = b = 0]$
Doble-implicación amplia	$p \leftrightarrow q \iff [v(p) = v(q)]$ $f_{\leftrightarrow}(a,b) = 1 \iff a = b$

Tabla 5.2: Tipos de implicación y doble implicación.

Se demuestra fácilmente que cuando los únicos valores de verdad posibles son 0 y 1, una implicación [doble-implicación] es rigurosa si y sólo si es implicación [doble-implicación] amplia; en cambio, cuando se admiten valores de verdad distintos de 0 y 1, una implicación amplia [doble-implicación] no puede ser rigurosa, y viceversa. [Más adelante veremos que la implicación de Łukasiewicz es amplia, mientras que la de Kleene es rigurosa.] Naturalmente, es posible en principio que una implicación o doble-implicación no sea ni amplia ni rigurosa.

En caso de una implicación rigurosa,

$$f_{\to}(a,b) = 1 \Longrightarrow a = 0 \vee b = 1$$

y tanto si $a = 0$ como si $b = 1$ se cumple que $a \leq b$; en consecuencia, la propiedad

$$f_{\to}(a,b) = 1 \Longrightarrow a \leq b \tag{5.11}$$

se cumple tanto para las implicaciones rigurosas como para las implicaciones amplias. Del mismo modo, puede demostrarse que la propiedad

$$f_{\leftrightarrow}(a,b) = 1 \Longrightarrow a = b \tag{5.12}$$

se cumple tanto para las dobles-implicaciones amplias como para las rigurosas.

Tras haber introducido los conceptos básicos, vamos a estudiar a continuación algunas de las lógicas más conocidas: la lógica clásica, las lógicas multivaluadas de Łukasiewicz y de Kleene, y la lógica difusa.

5.1.1 Lógica clásica

Las funciones que definen la **lógica clásica** son las que aparecen en la tabla 5.3. (Para entender mejor su significado, le recomendamos que vuelva a mirar las ecuaciones (5.5) a (5.9).)

La tabla 5.4 muestra las principales propiedades que cumplen estas conectivas así definidas.[4] Conviene señalar que en estas tablas p, q y r representan proposiciones genéricas, es decir, pueden ser proposiciones simples o compuestas, y las propiedades se cumplen para toda p, para toda q y para toda r. Por ejemplo, la 1ª ley de Morgan debe leerse así: "$\forall p,\ \forall q,\ \neg(p \wedge q) \equiv \neg p \vee \neg q$".

[4] La propiedad del *tercio excluso* recibe este nombre porque afirma que "o p o $\neg p$, y no existe una tercera posibilidad". La denominación de *monotonía* quedará clara más adelante.

a	b	$f^C_\neg(a)$	$f^C_\wedge(a,b)$	$f^C_\vee(a,b)$	$f^C_\rightarrow(a,b)$	$f^C_\leftrightarrow(a,b)$
1	1	0	1	1	1	1
1	0	0	0	1	0	0
0	1	1	0	1	1	0
0	0	1	0	0	1	1

Tabla 5.3: Valores de verdad para las funciones que definen las conectivas clásicas.

Observe que todas las propiedades, excepto las de monotonía, son de la forma $p_1 \equiv p_2$, lo cual, según la definición de equivalencia entre proposiciones, significa que $\forall v,\ v(p_1) = v(p_2)$. Siguiendo con el ejemplo anterior, la 1ª ley de Morgan significa: "$\forall v,\ \forall p,\ \forall q,\ v(\neg(p \wedge q)) = v(\neg p \vee \neg q)$"; es decir, para toda asignación de valores de verdad, v, y para todo par de proposiciones, p y q, la función v asigna los mismos valores de verdad a la proposición $\neg(p \wedge q)$ y a la proposición $\neg p \vee \neg q$.

La forma de demostrar esta propiedad es la siguiente: el valor de verdad de $\neg(p \wedge q)$ y de $\neg p \vee \neg q$ depende sólo de los valores de verdad $v(p)$ y $v(q)$; como sólo hay cuatro formas posibles en que v puede asignar estos valores —correspondientes a las cuatro filas de la tabla 5.5—, basta comprobar que las columnas $v(\neg(p \wedge q))$ y $v(\neg p \vee \neg q)$ de esta tabla coinciden. Del mismo modo se interpretan y demuestran las demás propiedades.

En la tabla 5.4 que estamos comentando aparecen dos versiones de cada propiedad de monotonía. Cuando la implicación es una implicación amplia (vea la tabla 5.14), como ocurre en la lógica clásica, la monotonía-p y la monotonía-v son equivalentes. Sin embargo, esta última tiene la ventaja de que no depende de ninguna implicación particular, por lo que nos será más útil a la hora de definir la monotonía de la negación, la conjunción y la disyunción en lógica difusa (sec. 5.1.3). En cambio, las propiedades de monotonía-p tienen la ventaja de que no hacen referencia explícita a la asignación de valores de verdad a las proposiciones, y por eso se pueden generalizar para predicados precisos —como veremos en la sección 5.2.1 (pág. 124)— y para conjuntos nítidos (tabla 5.19, pág. 140).

El hecho de que la lógica proposicional cumple las propiedades descritas por la tabla 5.4 nos permite afirmar que es un *álgebra de Boole*. (En realidad, las propiedades de monotonía no forman parte de la definición de álgebra de Boole, pero nos interesa incluirlas en la misma tabla.) Más adelante veremos que la teoría de conjuntos clásica, con las propiedades del complementario, unión e intersección, cumple también estas propiedades, por lo que constituye igualmente un álgebra de Boole.

Observe que, según la tabla 5.3, si sabemos que $p \rightarrow q$, es decir, si sabemos que $v(p \rightarrow q)$=1, hay tres posibilidades:

$$\left\{ \begin{array}{l} v(p) = 0,\ v(q) = 0 \\ v(p) = 0,\ v(q) = 1 \\ v(p) = 1,\ v(q) = 1 \end{array} \right\} \tag{5.13}$$

y se excluye la posibilidad de que $\{v(p) = 1,\ v(q) = 0\}$. A partir de esta observación se comprueba inmediatamente que la función de implicación de la lógica clásica, $f^C_\rightarrow$, es tanto una implicación rigurosa como una implicación amplia, y se demuestran fácilmente las propiedades que aparecen en la tabla 5.6.

También se demuestra a partir de la tabla 5.3 que si $p \leftrightarrow q$ entonces $v(p) = v(q)$, lo que

Negación	Negación de **1**	$\neg\mathbf{1} \equiv \mathbf{0}$
	Negación de **0**	$\neg\mathbf{0} \equiv \mathbf{1}$
	Involución	$\neg\neg p \equiv p$
	Monotonía-p	$p \to q \Longrightarrow \neg q \to \neg p$
	Monotonía-v	$v(p) \leq v(q) \Longrightarrow v(\neg p) \geq v(\neg q)$
Conjunción	Conmutativa	$p \wedge q \equiv q \wedge p$
	Asociativa	$p \wedge (q \wedge r) \equiv (p \wedge q) \wedge r$
	Elemento neutro	$p \wedge \mathbf{1} \equiv p$
	Elemento absorbente	$p \wedge \mathbf{0} \equiv \mathbf{0}$
	Idempotencia	$p \wedge p \equiv p$
	Ley de contradicción	$p \wedge \neg p \equiv \mathbf{0}$
	Monotonía-p	$p \to q \Longrightarrow (p \wedge r) \to (q \wedge r)$
	Monotonía-v	$v(p) \leq v(q) \Longrightarrow v(p \wedge r) \leq v(q \wedge r)$
Disyunción	Conmutativa	$p \vee q \equiv q \vee p$
	Asociativa	$p \vee (q \vee r) \equiv (p \vee q) \vee r$
	Elemento neutro	$p \vee \mathbf{0} \equiv p$
	Elemento absorbente	$p \vee \mathbf{1} \equiv \mathbf{1}$
	Idempotencia	$p \vee p \equiv p$
	Tercio excluso	$p \vee \neg p \equiv \mathbf{1}$
	Monotonía-p	$p \to q \Longrightarrow (p \vee r) \to (q \vee r)$
	Monotonía-v	$v(p) \leq v(q) \Longrightarrow v(p \vee r) \leq v(q \vee r)$
Propiedades combinadas	Distributiva de la conjunción	$p \wedge (q \vee r) \equiv (p \wedge q) \vee (p \wedge r)$
	Distributiva de la disyunción	$p \vee (q \wedge r) \equiv (p \vee q) \wedge (p \vee r)$
	1ª ley de Morgan	$\neg(p \wedge q) \equiv \neg p \vee \neg q$
	2ª ley de Morgan	$\neg(p \vee q) \equiv \neg p \wedge \neg q$
	Absorción de la conjunción	$p \vee (p \wedge q) \equiv p$
	Absorción de la disyunción	$p \wedge (p \vee q) \equiv p$

Tabla 5.4: Propiedades de la lógica clásica.

$v(p)$	$v(q)$	$v(p \wedge q)$	$v(\neg(p \wedge q))$	$v(\neg p)$	$v(\neg q)$	$v(\neg p \vee \neg q)$
1	1	1	0	0	0	0
1	0	0	1	0	1	1
0	1	0	1	1	0	1
0	0	0	1	1	1	1

Tabla 5.5: Demostración de la 1ª ley de Morgan.

Implicación	Neutralidad de la verdad	$\mathbf{1} \to p \equiv p$
	Predominio de la falsedad	$\mathbf{0} \to p \equiv \mathbf{1}$
	Identidad	$p \to p \equiv \mathbf{1}$
	Intercambio	$p \to (q \to r) \equiv q \to (p \to r)$
	Contraposición	$p \to q \equiv \neg q \to \neg p$
	Monotonía-p en el 1er argumento	$p \to q \Longrightarrow (q \to r) \to (p \to r)$
	Monotonía-p en el 2º argumento	$p \to q \Longrightarrow (r \to p) \to (r \to q)$
	Monotonía-v en el 1er argumento	$v(p) \leq v(q) \Longrightarrow v(p \to r) \geq v(q \to r)$
	Monotonía-v en el 2º argumento	$v(p) \leq v(q) \Longrightarrow v(r \to p) \leq v(r \to q)$

Tabla 5.6: Propiedades de la implicación de proposiciones clásica.

demuestra que la lógica clásica se basa en una doble-implicación amplia; de hecho, cuando $p \leftrightarrow q$ sólo hay dos posibilidades:

$$\left\{ \begin{array}{l} v(p) = 0,\ v(q) = 0 \\ v(p) = 1,\ v(q) = 1 \end{array} \right\} \tag{5.14}$$

por lo que es también una doble-implicación rigurosa.

Modus ponens y modus tollens en la lógica clásica

De la tabla 5.3 se deduce que, cuando $v(p)$=1, la única posibilidad de que $v(p \to q)$=1 es que $v(q)$=1. Por tanto, cuando sabemos que $p \to q$ y p son ciertas podemos deducir que q también es cierta:

$$(p \to q) \wedge p \Longrightarrow q \tag{5.15}$$

Este silogismo se denomina *modus ponens*, y se suele representar así:

$$\begin{array}{l} p \to q \\ p \\ \hline q \end{array}$$

Del mismo modo se comprueba que cuando $v(q)$=0, la única posibilidad de que $v(p \to q)$=1 es que $v(p)$=0. Por tanto, cuando sabemos que $p \to q$ es cierta y q es falsa podemos deducir que p también es falsa:

$$(p \to q) \wedge \neg q \Longrightarrow \neg p \tag{5.16}$$

Este silogismo se denomina *modus tollens*, y se suele representar así:

$$\begin{array}{l} p \rightarrow q \\ \neg q \\ \hline \neg p \end{array}$$

Más adelante veremos cómo se puede generalizar en el caso de las lógicas multivaluadas, de modo que, cuando $v(p \rightarrow q) \approx 1$, se cumple que si $v(p) \approx 1$ entonces $v(q) \approx 1$ (modus ponens aproximado) y si $v(q) \approx 0$ entonces $v(p) \approx 0$ (modus tollens aproximado).

5.1.2 Lógicas multivaluadas

Desde hace muchos siglos, varios filósofos han señalado que el hecho de tener que considerar toda proposición como verdadera o falsa es una limitación de la lógica clásica, puesto que hay afirmaciones indeterminadas. Para el propio Aristóteles, las proposiciones referidas al futuro no son verdaderas ni falsas, sino que pueden acabar siendo tanto lo uno como lo otro. También en la interpretación de Copenhage de la mecánica cuántica es posible que alguna proposición, tal como "el espín de cierto electrón es $-\frac{1}{2}$", no sea ni verdadera ni falsa, sino indeterminada, hasta que se haga un experimento para medirlo.

Por ello, teniendo en cuenta que en la lógica matemática clásica el valor de verdad de cada proposición es siempre 0 o 1, parece razonable asignar a las proposiciones indeterminadas valores intermedios. Así surgieron las lógicas trivaluadas (en las que el valor de verdad de una proposición indeterminada es $\frac{1}{2}$), que en seguida fueron generalizadas para dar lugar a las lógicas n-valuadas. Hemos dicho "lógicas trivaluadas", en plural, porque existen varias; a continuación vamos a estudiar dos de las más conocidas: la de Łukasiewicz y la de Kleene.

Lógica multivaluada de Łukasiewicz

La lógica trivaluada de Łukasiewicz viene dada por las funciones que se definen en la tabla 5.7. Observe que se trata de una extensión de la lógica clásica (bivaluada), en el sentido de que cuando los valores de verdad son 0 y 1, los valores de cada f son los mismos que en la lógica clásica; esto se puede comprobar comparando la tabla 5.7 con la 5.3.

a	b	$f^L_\neg(a)$	$f^L_\wedge(a,b)$	$f^L_\vee(a,b)$	$f^L_\rightarrow(a,b)$	$f^L_\leftrightarrow(a,b)$
1	1	0	1	1	1	1
1	$\frac{1}{2}$	0	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$
1	0	0	0	1	0	0
$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	1	1	$\frac{1}{2}$
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	1
$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
0	1	1	0	1	1	0
0	$\frac{1}{2}$	1	0	$\frac{1}{2}$	1	$\frac{1}{2}$
0	0	1	0	0	1	1

Tabla 5.7: Funciones para la lógica trivaluada de Łukasiewicz.

Del mismo modo, una *lógica n-valuada de Łukasiewicz* es aquélla en que $v(p) \in \mathcal{V}_n$ (ec. (5.2)) y las funciones que definen las conectivas son las siguientes:

$$
\begin{aligned}
f^L_{\neg}(a) &= 1 - a \\
f^L_{\wedge}(a,b) &= \min(a,b) && (5.17)\\
f^L_{\vee}(a,b) &= \max(a,b) && (5.18)\\
f^L_{\rightarrow}(a,b) &= \min(1, 1-a+b) && (5.19)\\
&= 1 - \max(0, a-b) && (5.20)\\
&= \begin{cases} 1 & \text{si } a \leq b \\ 1-(a-b) & \text{si } a > b \end{cases} && (5.21)\\
f^L_{\leftrightarrow}(a,b) &= 1 - |a-b| && (5.22)
\end{aligned}
$$

(Hemos expresado $f^L_{\rightarrow}$ de tres formas equivalentes con el fin de poder utilizar en cada caso la que más nos convenga).

Observe que esta definición funcional evita tener que construir una tabla diferente para cada n. Observe también que cuando n=2, tenemos la lógica clásica, mientras que cuando n=3 tenemos la lógica trivaluada definida por la tabla 5.7.

Se puede demostrar también que la lógica multivaluada de Łukasiewicz cumple **todas las propiedades de la lógica clásica** que aparecen en la tabla 5.4, **excepto la ley de contradicción y el tercio excluso**. En efecto, estas dos propiedades sólo se cumplen si los valores de verdad son 0 o 1, pues si $0 < v(p) < 1$, entonces $v(p \wedge \neg p) = \max(v(p), 1-v(p)) > 0$ y $v(p \vee \neg p) < 1$, de modo que ni $p \wedge \neg p \equiv \mathbf{0}$ ni $p \vee \neg p \equiv \mathbf{1}$.

Se comprueba además fácilmente que la lógica de Łukasiewicz utiliza una implicación amplia y una doble-implicación amplia.

Modus ponens y modus tollens en la lógica de Łukasiewicz

De la ecuación (5.21) se deduce que

$$v(p \rightarrow q) = 1 \Longrightarrow v(q) \geq v(p) \tag{5.23}$$

En particular, cuando $v(p \rightarrow q) = 1$, si $v(p) = 1$ entonces $v(q) = 1$ (modus ponens clásico) y si $v(q) = 0$ entonces $v(p) = 0$ (modus tollens clásico).

También de la ecuación (5.21) se deduce que

$$
\begin{aligned}
v(p \rightarrow q) < 1 &\Longrightarrow v(p \rightarrow q) = 1 - (v(p) - v(q)) \\
&\Longrightarrow v(q) = v(p) - [1 - v(p \rightarrow q)] = v(p \rightarrow q) - [1 - v(p)] && (5.24)
\end{aligned}
$$

lo cual nos dice que si $v(p \rightarrow q) \approx 1$ y $v(p) \approx 1$ entonces $v(q) \approx 1$ (modus ponens aproximado), aunque cuanta menos certeza tengamos sobre $p \rightarrow q$ y p, menos certeza tendremos sobre q. Por ejemplo, si $v(p \rightarrow q) = 0'95$ y $v(p) = 0'8$ entonces $v(q) = 0'75$.[5]

Uniendo las ecuaciones (5.23) y (5.24) podemos concluir que

$$v(q) \geq v(p) - [1 - v(p \rightarrow q)] = v(p) + v(p \rightarrow q) - 1 \tag{5.25}$$

[5]Se da, sin embargo, la siguiente paradoja; supongamos que $v(p) = 0'001$; si $v(p \rightarrow q) = 1$ entonces no hay ninguna restricción para $v(q)$; en cambio, si $v(p \rightarrow q) = 0'999$ entonces $v(q) = 0$. Es decir, que una disminución (aunque sea insignificante) en nuestra certeza sobre $p \rightarrow q$ nos lleva de no saber nada sobre q a descartar q con toda certeza. A nuestro juicio no hay ninguna razón de sentido común que justifique este comportamiento.

También tenemos que

$$v(p \to q) < 1 \Longrightarrow v(p) = v(q) + [1 - v(p \to q)] \tag{5.26}$$

lo cual nos dice que si $v(p \to q) \approx 1$ y $v(q) \approx 0$ entonces $v(p) \approx 0$ (modus tollens aproximado). Por ejemplo, si $v(p \to q) = 0'95$ y $v(q) = 0'1$ entonces $v(p) = 0'15$.[6]

Lógica multivaluada de Kleene

La lógica trivaluada de Kleene viene dada por las funciones de la tabla 5.8. Observe que es casi idéntica a la de Łukasiewicz, pues sólo difiere de ella en los valores de $f_{\to}\left(\frac{1}{2}, \frac{1}{2}\right)$ y $f_{\leftrightarrow}\left(\frac{1}{2}, \frac{1}{2}\right)$. También ésta es una generalización de la lógica clásica, como se puede comprobar comparando la tabla 5.8 con la 5.3.

a	b	$f^K_{\neg}(a)$	$f^K_{\wedge}(a,b)$	$f^K_{\vee}(a,b)$	$f^K_{\to}(a,b)$	$f^K_{\leftrightarrow}(a,b)$
1	1	0	1	1	1	1
1	$\frac{1}{2}$	0	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$
1	0	0	0	1	0	0
$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	1	1	$\frac{1}{2}$
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
0	1	1	0	1	1	0
0	$\frac{1}{2}$	1	0	$\frac{1}{2}$	1	$\frac{1}{2}$
0	0	1	0	0	1	1

Tabla 5.8: Funciones para la lógica trivaluada de Kleene.

Del mismo modo, una *lógica n-valuada de Kleene* es aquélla en que $v(p) \in \mathcal{V}_n$ (ec. (5.2)) y las funciones que definen las conectivas son las siguientes:

$$f^K_{\neg}(a) = 1 - a \tag{5.27}$$

$$f^K_{\wedge}(a,b) = \min(a,b) \tag{5.28}$$

$$f^K_{\vee}(a,b) = \max(a,b) \tag{5.29}$$

$$f^K_{\to}(a,b) = \max(1-a,b) \tag{5.30}$$

$$f^K_{\leftrightarrow}(a,b) = \min(\max(1-a,b), \max(1-b,a)) \tag{5.31}$$

Es decir, coincide con la de Łukasiewicz en la definición de la negación, la conjunción y la disyunción, pero difiere en la implicación y la doble implicación.

Como en el caso anterior, cuando n=2, tenemos la lógica clásica, mientras que cuando n=3 tenemos la lógica trivaluada de Kleene (tabla 5.8).

Se puede demostrar también que la lógica multivaluada de Kleene cumple **todas las propiedades de la lógica clásica** que aparecen en la tabla 5.4, **excepto la ley de contradicción y el tercio excluso**, pues si $0 < v(p) < 1$, entonces $v(p \wedge \neg p) > 0$ y $v(p \vee \neg p) < 1$.

[6] Tenemos de nuevo una paradoja análoga a la anterior. Supongamos que $v(q) = 0'999$; si $v(p \to q) = 1$ entonces $v(p)$ puede tomar cualquier valor entre 0 y 0'999; en cambio, si $v(p \to q) = 0'999$ entonces $v(p) = 1$. Es decir, una disminución (aunque sea insignificante) en nuestra certeza sobre $p \to q$ nos lleva de no saber prácticamente nada sobre p a confirmar p con toda certeza.

Se comprueba además que la lógica de Kleene utiliza una implicación rigurosa y una doble-implicación rigurosa, en contraste con la de Łukasiewicz, que se basa en una implicación amplia y una doble-implicación amplia.

Modus ponens y modus tollens para la lógica de Kleene

De la ecuación (5.30) se deduce que

$$v(p \to q) = \max(1 - v(p), v(q)) \geq 1 - v(p)$$

y por tanto

$$v(p \to q) + v(p) \geq 1$$

Se puede deducir también que

$$v(p \to q) + v(p) = 1 \iff v(p \to q) = 1 - v(p) \iff v(q) \leq 1 - v(p) \tag{5.32}$$
$$v(p \to q) + v(p) > 1 \iff v(p \to q) > 1 - v(p) \iff v(p \to q) = v(q) \geq 1 - v(p) \tag{5.33}$$

Esta última expresión nos dice que si $v(p \to q) = 1$ y $v(p) = 1$ entonces $v(q) = v(p \to q) = 1$ (modus ponens clásico) y si $v(p \to q) \approx 1$ y $v(p) \approx 1$ entonces $v(q) = v(p \to q) \approx 1$ (modus ponens aproximado); por ejemplo, si $v(p \to q) = 0'95$ y $v(p) = 0'8$ entonces $v(q) = 0'95$.[7]

De la ecuación (5.30) se deduce también que

$$v(p \to q) > v(q) \Longrightarrow v(p \to q) = 1 - v(p)$$

Por tanto, si $v(p \to q) = 1$ y $v(q) = 0$ entonces $v(p) = 1 - v(p \to q) = 0$ (modus tollens clásico) y si $v(p \to q) \approx 1$ y $v(q) \approx 0$ entonces $v(p) = 1 - v(p \to q) \approx 0$ (modus tollens aproximado); por ejemplo, si $v(p \to q) = 0'95$ y $v(q) = 0$ entonces $v(p) = 0'05$.[8]

Otras lógicas multivaluadas

Existen otras lógicas trivaluadas, tales como la de Bochvar, la de Heyting, la de Reichenbach... (cf. [35, sec. 8.2]), algunas de las cuales ni siquiera utilizan las funciones mínimo y máximo para la conjunción y la disyunción. Dado el carácter introductorio de este texto, no las vamos a tratar aquí; nos limitamos a señalar que todas ellas son extensiones de la lógica clásica (bivaluada), en el sentido de que cuando los valores de verdad son 0 y 1, los valores de cada f son los mismos que en la lógica clásica.

[7] De nuevo tenemos paradojas análogas a las que se daban para la implicación de Łukasiewicz. En efecto, supongamos que $v(p \to q) = 1$; si $v(p) = 0$ entonces $v(q)$ puede tomar cualquier valor, mientras que si $v(p) = 0'001$ entonces $v(q) = v(p \to q) = 1$. Es decir, el hecho de saber que p no es totalmente falso (aunque el valor de $v(p)$ pueda ser tan próximo a 0 como queramos) nos lleva a confirmar q con certeza total. Tampoco en este caso hay, a nuestro juicio, ninguna razón de sentido común que justifique este comportamiento.

[8] En este caso la paradoja es la siguiente: supongamos que $v(p \to q) = 1$; si $v(q) = 1$ entonces $v(p)$ puede tomar cualquier valor, porque $\max(1 - v(p), v(q))$ va a ser siempre 1; en cambio, si $v(q) = 0'999$ entonces $v(p)$ tiene que ser 0 para $\max(1 - v(p), v(q)) = 1$. Es decir, el hecho de saber que q no es totalmente cierto (aunque el valor de $v(q)$ pueda ser tan próximo a 1 como queramos) nos lleva a descartar p con certeza total.

5.1.3 Lógica difusa

Acabamos de estudiar las lógicas n-valuadas, que son aquéllas que toman valores de verdad en $\mathcal{V}_n$ (ec. (5.2)). En la de Łukasiewicz y en la de Kleene, las funciones f están definidas a partir de la suma, la resta, el máximo, el mínimo y el valor absoluto, lo cual garantiza que si a y b pertenecen a $\mathcal{V}_n$ también $f_\neg(a)$, $f_\wedge(a,b)$, $f_\vee(a,b)$, $f_\to(a,b)$ y $f_\leftrightarrow(a,b)$ pertenecen. Observe que en una lógica n-valuada no se podría tener, por ejemplo, una función de conjunción como $f_\wedge(a,b) = a \cdot b$, pues ello implicaría que $f_\wedge(\frac{1}{n}, \frac{1}{n}) = \frac{1}{n^2} \notin \mathcal{V}_n$. Ésta es una de las razones por las que en la lógica difusa se permite que $v(p)$ tome cualquier valor del intervalo $[0,1]$ (ec. (5.3)), con lo cual existe mayor libertad a la hora de definir las funciones lógicas. En este sentido, la lógica difusa es una generalización de las lógicas n-valuadas, del mismo modo que éstas generalizaban la lógica clásica.

Naturalmente, las lógicas n-valuadas que hemos estudiado en la sección anterior no exigen que los valores de verdad sean números racionales ni que el conjunto de valores sea finito; de hecho, basta que el valor de verdad asignado a las proposiciones elementales pertenezca al intervalo [0,1] para que los valores de verdad de las proposiciones compuestas pertenezcan también al intervalo [0,1]. En consecuencia, tanto la lógica de Łukasiewicz como la de Kleene son casos particulares de la lógica difusa, pero pueden existir otras muchas, vamos a estudiar a continuación las propiedades que han de cumplir cada una de ellas.

Una de las condiciones fundamentales, denominada *límite clásico*, es que toda lógica difusa ha de ser una extensión de la lógica clásica, en el sentido de que cuando los valores de verdad de dos proposiciones está en $\{0,1\}$, los valores de verdad de las proposiciones resultantes al aplicar las conectivas sean los mismos que en la lógica clásica. Por ejemplo, la propiedad del tercio excluso ($p \vee \neg p \equiv \mathbf{1}$) de la lógica clásica es equivalente a decir que $f^C_\vee(0,1) = 1$; por tanto, toda lógica difusa debe cumplir que $f_\vee(0,1) = f^C_\vee(0,1) = 1$, aunque en general $f_\vee(a,b) \neq 1$ cuando a o b sean distintos de 0 y 1.

Además de mantener el límite clásico, vamos a examinar qué otras propiedades de la lógica clásica es posible y conveniente mantener en la lógica difusa; para ello, vamos a estudiar las condiciones que deben cumplir las funciones f, pues cada una de las propiedades expresadas en la tabla 5.4 se traduce en una propiedad que deben cumplir las f's.

Funciones de negación

Las propiedades deseables para una función de negación en lógica difusa aparecen en la tabla 5.9, y proceden de las propiedades de la lógica clásica (tabla 5.4).

Negación	Límite clásico	$f_\neg(1) = 0,\ f_\neg(0) = 1$
	Involución	$f_\neg(f_\neg(a)) = a$
	Monotonía-v	$a \leq b \Longrightarrow f_\neg(a) \geq f_\neg(b)$

Tabla 5.9: Propiedades definitorias de la función negación.

Otra propiedad deseable de $f_\neg$ es la **continuidad**, pues no es razonable que un cambio infinitesimal en a produzca un cambio brusco en $f_\neg(a)$; por tanto,

$$\lim_{\delta \to 0} f_\neg(a+\delta) = f_\neg(a)$$

Sin embargo, a diferencia de lo que hacen otros textos sobre lógica difusa, no consideramos necesario incluir la continuidad entre las propiedades axiomáticas de $f_\neg$, pues viene garantizada

por la siguiente proposición:

Proposición 5.1 Toda función de negación $f_\neg$ es estrictamente decreciente, invertible, simétrica (respecto del eje $x = y$) y continua.

Demostración. Dado que es decreciente por definición (cf. tabla 5.9), basta tener en cuenta que

$$f_\neg(a) = f_\neg(b) \Longrightarrow f_\neg(f_\neg(a)) = f_\neg(f_\neg(b)) \Longrightarrow a = b$$

para demostrar que es estrictamente decreciente:

$$a < b \Longrightarrow f_\neg(a) > f_\neg(b)$$

La función inversa es ella misma: $f_\neg^{-1} = f_\neg$, pues $f_\neg^{-1}(f_\neg(a)) = f_\neg(f_\neg(a)) = a$. Por otro lado, de la propiedad de involución se deduce que $f_\neg(x) = y \Longrightarrow f_\neg(y) = x$, lo que prueba la simetría.

Intuitivamente vemos que $f_\neg$ ha de ser continua, porque es una función decreciente y si tuviera un escalón entonces existiría un y que no sería imagen de ningún x, lo cual es absurdo, porque para todo y siempre podemos tomar $x = f_\neg(y)$, de modo que $f_\neg(x) = y$. Sin embargo, la demostración rigurosa, que ofrecemos a continuación, es bastante más compleja; el lector que lo desee puede omitir su estudio, porque se trata más de una cuestión de cálculo infinitesimal que de razonamiento aproximado.

Por definición, una función del intervalo $[0, 1]$ es continua si y sólo si cumple la siguiente propiedad:

$$\forall a, 0 \leq a \leq 1, \forall \varepsilon, \varepsilon > 0, \exists \delta \mid [\delta > 0] \wedge \underbrace{[\forall x, 0 \leq x \leq 1, |x - a| < \delta \Longrightarrow |f_\neg(x) - f_\neg(a)| \leq \varepsilon]}_{p_{a,\varepsilon,\delta}}$$

Una vez dados a y ε, tales que $0 \leq a \leq 1$ y $\varepsilon > 0$, vamos a demostrar que la δ definida por las siguientes ecuaciones

$$\begin{aligned} x_1 &= f_\neg(\min(f_\neg(a) + \varepsilon, 1)) \\ x_2 &= f_\neg(\max(f_\neg(a) - \varepsilon, 0)) \\ \delta &= \begin{cases} x_2 & \text{si } a = 0 \\ 1 - x_1 & \text{si } a = 1 \\ \min(x_1 - a, a_2 - x) & \text{si } 0 < a < 1 \end{cases} \end{aligned}$$

cumple $\delta > 0$ y la proposición $p_{a,\varepsilon,\delta}$. En primer lugar, observamos que

$$\begin{aligned} a = 0 &\Longrightarrow x_1 = 0 \qquad a > 1 \Longrightarrow 0 \leq x_1 < a \\ a = 1 &\Longrightarrow x_2 = 1 \qquad a < 1 \Longrightarrow a < x_2 \leq 1 \end{aligned}$$

lo cual implica que $\delta > 0$ en todos los casos.

Por otro lado, de las desigualdades $0 \leq x \leq 1$ y $|x - a| < \delta$ se deduce que

$$\begin{aligned} &a = 0 \Longrightarrow x = |x - 0| < \delta = x_2 \Longrightarrow x_1 = 0 \leq x < x_2 \\ &a = 1 \Longrightarrow 1 - x = |x - 1| < \delta = 1 - x_1 \Longrightarrow x_1 < x \leq 1 = x_2 \\ &0 < a < 1 \Longrightarrow = \left\{ \begin{array}{l} x - a < \delta \leq x_2 - a \\ a - x < \delta \leq a - x_1 \end{array} \right\} \Longrightarrow x_1 < x < x_2 \end{aligned}$$

Es decir, en todos los casos se cumple que

$$x_1 \leq x \leq x_2$$

y por tanto

$$\min(f_\neg(a) + \varepsilon, 1) = f_\neg(x_1) \geq f_\neg(x) \geq f_\neg(x_2) = \max(f_\neg(a) - \varepsilon, 0)$$

$$f_\neg(a) + \varepsilon \geq f_\neg(x) \geq f_\neg(a) - \varepsilon$$

$$|f_\neg(x) - f_\neg(a)| \leq \varepsilon$$

como queríamos demostrar. □

Proposición 5.2 Para toda función de negación $f_\neg$ existe un *punto de equilibrio* a_e, $0 < a_e < 1$, tal que $f_\neg(a_e) = a_e$.

Demostración. Definimos una nueva función $g(a) = f_\neg(a) - a$. Dado que g es continua, $g(0) = 1 > 0$ y $g(1) = -1 < 0$, por el teorema de Bolzano debe existir un a_e, $0 < a_e < 1$, tal que $g(a_e) = f_\neg(a_e) - a_e = 0$. □

La función $f_\neg$ más conocida y utilizada es la que en lógica difusa se conoce como *función de negación estándar*:

$$f_\neg(a) = 1 - a \tag{5.34}$$

El punto de equilibrio para esta función es $a_e = 0'5$.

Para cada valor del parámetro β tal que $\beta \in (-1, \infty)$, la función

$$f_\neg(a) = \frac{1 - a}{1 + \beta a} \tag{5.35}$$

es también una función de negación; el conjunto de todas estas funciones (para distintos valores de β) se denomina *familia de negaciones de Sugeno*. La función de negación estándar es un miembro particular de la familia de Sugeno, correspondiente al caso en que $\beta = 0$.

Del mismo modo, la *familia de negaciones de Yager* viene dada por

$$f_\neg(a) = (1 - a^\beta)^{\frac{1}{\beta}} \tag{5.36}$$

donde $\beta \in (0, \infty)$. Cuando $\beta = 1$, tenemos la negación estándar.

Ejercicio 5.3 (opcional) Demuestre que cada miembro de estas dos familias es una función de negación, y calcule su punto de equilibrio. □

Existen otras propiedades matemáticas, tales como las relativas al punto de equilibrio o a los generadores de funciones de negación, que no vamos a tratar en este texto porque, a nuestro juicio no aportan nada desde el punto de vista de su aplicación a la inteligencia artificial. El lector interesado puede encontrar éste y otros temas en la bibliografía recomendada al final del capítulo.

Funciones de conjunción: normas triangulares

Acabamos de ver que es posible construir funciones de negación difusas que mantengan todas las propiedades de la negación clásica. Sin embargo, en seguida vamos a ver que es imposible definir funciones de conjunción y de disyunción difusas que conserven todas las propiedades de la lógica clásica. Por ello vamos a exigir a estas funciones que mantengan las propiedades más importantes, aunque sea al precio de perder otras menos importantes o menos deseables.

En primer lugar, definimos las *normas triangulares* (a veces llamadas *t-normas* o simplemente *normas*) como aquellas funciones que cumplen las propiedades indicadas en la tabla 5.10 (compárelas con las de la conjunción clásica, tabla 5.4, pág. 102).

Conmutativa	$f_\wedge(a,b) = f_\wedge(b,a)$
Asociativa	$f_\wedge(a, f_\wedge(b,c)) = f_\wedge(f_\wedge(a,b), c)$
Elemento neutro	$f_\wedge(a,1) = a$
Monotonía-v	$a \leq b \Longrightarrow f_\wedge(a,c) \leq f_\wedge(b,c)$

Tabla 5.10: Propiedades de las normas triangulares.

Proposición 5.4 Toda norma triangular cumple el límite clásico de la conjunción.

Demostración. Por la propiedad de elemento neutro, $f_\wedge(0,1) = 0$ y $f_\wedge(1,1) = 1$. Por la conmutativa, $f_\wedge(1,0) = 0$. Por la monotonía, $f_\wedge(0,0) \leq f_\wedge(1,0)$, lo que implica que $f_\wedge(0,0) = 0$. □

Proposición 5.5 Para toda norma triangular hay un elemento absorbente, que es el cero:

$$\forall a, \quad f_\wedge(a,0) = 0 \tag{5.37}$$

Demostración. Para todo valor a tenemos, por un lado, que $f_\wedge(a,0) \leq f_\wedge(1,0) = f_\wedge(0,1) = 0$, y por otro, que $f_\wedge(a,0) \geq 0$, de donde se deduce que $f_\wedge(a,0) = 0$. □

En cambio, hay normas triangulares que no cumplen la propiedad de idempotencia

$$\forall a, \quad f_\wedge(a,a) = a \tag{5.38}$$

o la ley de contradicción

$$\forall a, \quad f_\wedge(a, f_\neg(a)) = 0 \tag{5.39}$$

para ninguna función de negación $f_\neg$, como veremos más adelante.

Proposición 5.6 Para toda norma triangular se cumple que

$$f_\wedge(a,b) = 1 \Longleftrightarrow a = b = 1 \tag{5.40}$$

Demostración. (Por reducción al absurdo) Si uno de los dos argumentos, por ejemplo a, fuera menor que 1, tendríamos que

$$f_\wedge(a,b) \leq f_\wedge(a,1) = a < 1$$

lo cual es una contradicción. □

La *conjunción estándar* (es decir, la que se usa en la lógica difusa estándar) viene dada por la función *mínimo*,

$$f_\wedge(a,b) = \min(a,b) \tag{5.41}$$

que es una norma triangular.

Otras normas triangulares son el *producto algebraico*

$$f_\wedge(a,b) = a \cdot b \tag{5.42}$$

la *diferencia acotada*

$$f_\wedge(a,b) = \max(0, a+b-1) \tag{5.43}$$

y la *conjunción drástica*

$$f_\wedge(a,b) = f_\wedge^{\min}(a,b) = \begin{cases} b & \text{si } a = 1 \\ a & \text{si } b = 1 \\ 0 & \text{en los demás casos} \end{cases} \tag{5.44}$$

Entre las cuatro existe una relación de orden:

$$\forall a, \forall b, \quad f_\wedge^{\min}(a,b) \le a \cdot b \le \max(0, a+b-1) \le \min(a,b) \tag{5.45}$$

Ejercicio 5.7 Dibuje estas cuatro funciones de $[0,1] \times [0,1] \to [0,1]$.

Ejercicio 5.8 Demuestre que cada una de ellas es una norma y que se cumple la desigualdad anterior.

Observe que todas estas normas, excepto la conjunción drástica, son continuas. Como dijimos al hablar de la negación, conviene que $f_\wedge$ sea continua para evitar que un cambio infinitesimal en a o en b provoque un cambio brusco en $f_\wedge(a,b)$, por lo que en la práctica sólo se utilizan normas continuas.[9]

Existen también varias familias de normas; por ejemplo, la *familia de Yager* viene dada por

$$f_\wedge(a,b) = Y_\wedge^\beta(a,b) = 1 - \min\left\{1, [(1-a)^\beta + (1-b)^\beta]^{\frac{1}{\beta}}\right\} \tag{5.46}$$

donde $\beta \in (0,\infty)$. Aunque β no puede valer nunca 0 ni infinito, se cumple que

$$\lim_{\beta\to 0} Y_\wedge^\beta(a,b) = \min(a,b) \tag{5.47}$$

$$\lim_{\beta\to \infty} Y_\wedge^\beta(a,b) = f_\wedge^{\min}(a,b) \tag{5.48}$$

y que

$$\forall \beta, \ f_\wedge^{\min}(a,b) \le Y_\wedge^\beta(a,b) \le \min(a,b) \tag{5.49}$$

En realidad, esta última expresión es un caso particular de una afirmación más general.

Proposición 5.9 Para toda norma triangular $f_\wedge$ se cumple que

$$f_\wedge^{\min}(a,b) \le f_\wedge(a,b) \le \min(a,b) \tag{5.50}$$

[9] Algunos autores incluyen la continuidad entre las propiedades axiomáticas de las normas, y por eso no consideran la conjunción drástica como una norma.

Ejercicio 5.10 Demostrar esta proposición.

Ésta es la razón por la que denotamos por $f_\wedge^{\min}$ a la conjunción drástica, pues es la norma mínima. Del mismo modo podríamos escribir $f_\wedge^{\max} = \min$, pues la función mínimo es la norma máxima.

Proposición 5.11 La función mínimo es la única norma idempotente.

Demostración. Sea una norma idempotente $f_\wedge$. Para dos valores a y b tal que $a \leq b$,

$$a = f_\wedge(a, a) \leq f_\wedge(a, b) \leq f_\wedge(a, 1) = a$$

de donde se deduce que $f_\wedge(a, b) = a$. Del mismo modo se demuestra que, cuando $b \leq a$, $f_\wedge(a, b) = b$, con lo que concluye la demostración. □

Teniendo en cuenta que la función mínimo no cumple la ley de contradicción —pues tal como afirma la proposición 5.2, para cualquier $f_\neg$ existe un punto de equilibrio a_e, $0 < a_e < 1$, de modo que $f_\wedge(a_e, f_\neg(a_e)) = \min(a_e, a_e) = a_e \neq 0$— se deduce que

Corolario 5.12 Ninguna norma que cumpla la ley de contradicción puede ser idempotente.

Corolario 5.13 No hay ninguna norma que cumpla todas las propiedades de la conjunción clásica.

Ante la imposibilidad de satisfacer todas las propiedades simultáneamente, la idempotencia ($p \wedge p \equiv p$) parece más importante que la ley de contradicción, la cual afirma que es imposible que una proposición y su negación sean ciertas a la vez, ni siquiera parcialmente ciertas, pues $p \wedge \neg p \equiv \mathbf{0}$ implica que $v(p \wedge \neg p) = f_\wedge(v(p), v(\neg p)) = 0$. Los defensores de la lógica difusa sostienen que ésta es una restricción innecesaria y poco realista, porque en el caso de conceptos difusos es posible que tanto una afirmación como su contraria tengan algo de verdad. Más aún, hay autores que, invocando la filosofía oriental del Ying y el Yang, afirman que todo lo que existe tiene también algo de su contrario: todo hombre tiene algún rasgo femenino en su carácter, y viceversa; todo lo tibio es caliente en cierto modo y frío en cierto modo, etc. Nuria, la niña a la que está dedicado este libro, se dio cuenta bastante pronto de que no era totalmente pequeña ni totalmente mayor, y por eso cuando tenía dos años y pocos meses solía decir: "soy uno poco pequeña y uno poco mayó". Por eso en lógica difusa el hecho de que algunas normas triangulares incumplan la ley de contradicción no se considera como un inconveniente, sino como una auténtica ventaja. Dado el carácter introductorio de este texto, no vamos a entrar a discutir tales afirmaciones. Lo mencionamos solamente porque ésta es una de las razones por las que la lógica difusa estándar toma la función mínimo para definir la conjunción (y la intersección de conjuntos, como veremos más adelante), dado que esta función preserva la idempotencia y elude a la vez la ley de contradicción.

Ejercicio 5.14 Indique cuáles de las normas triangulares mostradas en esta sección cumplen la ley de contradicción (para la negación estándar) y cuáles no.

Funciones de disyunción: conormas triangulares

Análogamente, las *conormas triangulares* (a veces llamadas *t-conormas* o simplemente *conormas*) son, por definición, las funciones que cumplen las propiedades de la tabla 5.11 (compárelas con las de la disyunción clásica, tabla 5.4, pág. 102).

Conmutativa	$f_\vee(a,b) = f_\vee(b,a)$
Asociativa	$f_\vee(a, f_\vee(b,c)) = f_\vee(f_\vee(a,b),c)$
Elemento neutro	$f_\vee(a,0) = a$
Monotonía-v	$a \leq b \Longrightarrow f_\vee(a,c) \leq f_\vee(b,c)$

Tabla 5.11: Propiedades de las conormas triangulares.

Proposición 5.15 Toda conorma triangular cumple el límite clásico de la disyunción.

Proposición 5.16 Para toda conorma triangular hay un elemento absorbente, que es el uno:

$$\forall a, \quad f_\vee(a,1) = 1 \tag{5.51}$$

(Las demostraciones son análogas a las de las proposiciones 5.4 y 5.5.)

En cambio, hay conormas que no cumplen la propiedad de idempotencia

$$\forall a, \quad f_\vee(a,a) = a \tag{5.52}$$

o la ley del tercio excluso

$$\forall a, \quad f_\vee(a, f_\neg(a)) = 1 \tag{5.53}$$

para ninguna función de negación $f_\neg$, como veremos más adelante.

Proposición 5.17 Para toda conorma triangular se cumple que

$$f_\vee(a,b) = 0 \Longleftrightarrow a = b = 0 \tag{5.54}$$

La demostración es análoga a la de la proposición 5.6.

La *disyunción estándar* (es decir, la que se usa en la lógica difusa estándar) viene dada por la función *máximo*:

$$f_\vee(a,b) = \max(a,b) \tag{5.55}$$

que es una conorma triangular.

Otras conormas triangulares son la *suma algebraica*

$$f_\vee(a,b) = a + b - a \cdot b \tag{5.56}$$

la *suma acotada*

$$f_\vee(a,b) = \min(1, a+b) \tag{5.57}$$

y la *disyunción drástica*

$$f_\vee(a,b) = f_\vee^{\max}(a,b) = \begin{cases} b & \text{si } a = 0 \\ a & \text{si } b = 0 \\ 1 & \text{en los demás casos} \end{cases} \tag{5.58}$$

Entre las cuatro existe una relación de orden:

$$f_\vee^{\max}(a,b) \geq a + b - a \cdot b \geq \min(1, a+b) \geq \max(a,b) \tag{5.59}$$

Observe que todas ellas, excepto la disyunción drástica, son continuas. Por las razones expuestas en las secciones anteriores, en la práctica sólo se utilizan conormas continuas.

Ejercicio 5.18 Dibuje estas cuatro funciones de $[0,1] \times [0,1] \to [0,1]$. Compare las gráficas con las del ejercicio 5.7.

Ejercicio 5.19 Demuestre que cada una de ellas es una conorma y que se cumple la desigualdad anterior.

La *familia de conormas de Yager* viene dada por

$$f_\vee(a,b) = Y_\vee^\beta(a,b) = \min\left[1, (a^\beta + b^\beta)^{\frac{1}{\beta}}\right] \tag{5.60}$$

donde $\beta \in (0,\infty)$. Aunque β no puede valer nunca 0 ni infinito, se cumple que

$$\lim_{\beta\to 0} Y_\vee^\beta(a,b) = \max(a,b) \tag{5.61}$$

$$\lim_{\beta\to \infty} Y_\vee^\beta(a,b) = f_\vee^{\max}(a,b) \tag{5.62}$$

Proposición 5.20 Para toda conorma triangular $f_\vee$ se cumple que

$$f_\vee^{\max}(a,b) \geq f_\vee(a,b) \geq \max(a,b) \tag{5.63}$$

Ésta es la razón por la que denotamos por $f_\vee^{\max}$ a la disyunción drástica, pues es la conorma máxima. Del mismo modo podríamos escribir $f_\vee^{\min} = \max$, pues la función máximo es la conorma mínima.

Proposición 5.21 La función máximo es la única conorma idempotente.

Teniendo en cuenta que la función máximo no cumple la ley del tercio excluso, pues para un a tal que $0 < a < 1$, $\max(a, 1-a) \neq 0$, se deduce que

Corolario 5.22 Ninguna conorma que cumpla la ley del tercio excluso puede ser idempotente.

Corolario 5.23 No hay ninguna conorma que cumpla todas las propiedades de la disyunción clásica.

Las razones por las que la lógica difusa estándar elige la función máximo para definir la disyunción [y la unión de conjuntos] son las mismas que las que justifican la elección de la función mínimo para definir la conjunción [y la intersección] (cf. sección anterior).

Uno de los temas que no hemos mencionado al hablar de las normas y conormas es la posibilidad de definirlas a partir de funciones generatrices. El motivo es que, a nuestro juicio, éste es solamente un aspecto matemático que no aporta nada desde el punto de vista del razonamiento en inteligencia artificial. En cualquier caso, el lector interesado puede encontrar la información en [35, cap. 3] y [60, cap. 2]

Normas y conormas conjugadas

Hemos analizado las condiciones que deben cumplir la negación, la conjunción y la disyunción por separado para conservar las propiedades más importantes de la lógica clásica. Sin

embargo, si exigimos que se mantengan también las propiedades combinadas que aparecen en la tabla 5.4, surgen nuevas condiciones. Por ejemplo, la 1ª ley de Morgan nos dice que

$$f_\neg(f_\wedge(a,b)) = f_\vee(f_\neg(a), f_\neg(b)) \tag{5.64}$$

Esta ecuación es válida para todos los valores de a y b, incluidos $f_\neg(a)$ y $f_\neg(b)$:

$$f_\neg(f_\wedge(f_\neg(a), f_\neg(b))) = f_\vee(f_\neg(f_\neg(a)), f_\neg(f_\neg(b))) = f_\vee(a,b)$$

de donde se deduce que

$$f_\vee(a,b) = f_\neg(f_\wedge(f_\neg(a), f_\neg(b))) \tag{5.65}$$

y, aplicando $f_\neg$ a cada miembro de esta ecuación, obtenemos

$$f_\neg(f_\vee(a,b)) = f_\wedge(f_\neg(a), f_\neg(b)) \tag{5.66}$$

que es la 2ª ley de Morgan. Por tanto, del hecho de que la función de negación $f_\neg$ es idempotente se deduce la siguiente proposición:

Proposición 5.24 La primera ley de Morgan se cumple si y sólo si se cumple la segunda.

Observe que la ecuación (5.64) es equivalente a

$$f_\wedge(a,b) = f_\neg(f_\vee(f_\neg(a), f_\neg(b))) \tag{5.67}$$

Por tanto, podemos utilizar esta expresión para definir $f_\wedge$ a partir de $f_\neg$ y $f_\vee$, del mismo modo que se podría haber utilizado la (5.65) para definir $f_\vee$ en función de $f_\neg$ y $f_\wedge$.

Proposición 5.25 Si $f_\neg$ es una función de negación difusa y $f_\wedge$ es una norma triangular, la función $f_\vee$ definida por la ecuación (5.65) es una conorma. (Se dice que $f_\vee$ es la *conorma dual* o *conjugada* de $f_\wedge$ respecto de $f_\neg$.)

Proposición 5.26 Si $f_\neg$ es una función de negación difusa y $f_\vee$ es una conorma triangular, la función $f_\wedge$ definida por la ecuación (5.67) es una norma. (Se dice que $f_\wedge$ es la *dual* o *conjugada* de $f_\vee$ respecto de $f_\neg$.)

Ejercicio 5.27 Demostrar estas tres proposiciones.

De la discusión anterior a la proposición 5.24 se deduce la siguiente:

Proposición 5.28 La función $f_\wedge$ es la norma dual de $f_\vee$ si y sólo si $f_\vee$ es la conorma dual de $f_\wedge$. (Suele decirse que $f_\wedge$ y $f_\vee$ son *conjugadas*.)

La tabla 5.12 recoge en su primera columna varias normas y en la segunda las correspondientes normas conjugadas respecto de la negación estándar.

Proposición 5.29 Si $f_\wedge$ y $f_\vee$ son conjugadas entre sí, y $f'_\wedge$ y $f'_\vee$ son conjugadas entre sí,

$$f_\wedge(a,b) \leq f'_\wedge(a,b) \Longleftrightarrow f_\vee(a,b) \geq f'_\vee(a,b) \tag{5.68}$$

Por tanto, observando la tabla 5.12 se entiende la relación entre las ecuaciones (5.45) y (5.63).

Norma	Conorma
$\min(a,b)$	$\max(a,b)$
$a \cdot b$	$a+b-a\cdot b$
$\max(0,a+b-1)$	$\min(1,a+b)$
$f_{\wedge}^{\min}(a,b)$	$f_{\vee}^{\max}(a,b)$
$Y_{\wedge}^{\beta}(a,b)$	$Y_{\vee}^{\beta}(a,b)$

Tabla 5.12: Normas y conormas conjugadas.

Funciones de implicación

En lógica difusa es habitual definir la función de implicación $f_{\rightarrow}$ a partir de las funciones $f_{\neg}$, $f_{\wedge}$ y $f_{\vee}$, aplicando alguna de las propiedades de la lógica clásica. Concretamente, la propiedad clásica

$$p \rightarrow q \equiv \neg p \vee q \tag{5.69}$$

se traduce en

$$f_{\rightarrow}(a,b) = f_{\vee}(f_{\neg}(a),b) \tag{5.70}$$

Cada $f_{\rightarrow}$ obtenida al escoger una función de negación $f_{\neg}$ y una conorma $f_{\vee}$ para la expresión anterior se denomina *implicación* S (porque S es el símbolo elegido habitualmente para representar una conorma). Por ejemplo, si tomamos la negación estándar (ec. (5.34)) y la suma acotada (ec. (5.57)), respectivamente, obtenemos la función de **implicación de Łukasiewicz** (ec. (5.19)), que es, por tanto, una implicación S.

En cambio, si tomamos la negación estándar y la disyunción estándar obtenemos la **implicación de Kleene** (ec. (5.30)), a veces llamada de Kleene-Dienes, que es también una implicación S. Tomando la negación estándar y la suma algebraica (ec. (5.56)) obtenemos la **implicación de Reichenbach**:

$$f_{\rightarrow}^{R}(a,b) = 1 - a + ab \tag{5.71}$$

que es, por cierto, una implicación rigurosa.

Otra propiedad de la lógica clásica en la cual puede basarse la definición de implicación es

$$p \rightarrow q \equiv \neg p \vee (p \wedge q) \tag{5.72}$$

la cual se traduce en[10]

$$f_{\rightarrow}(a,b) = f_{\vee}(f_{\neg}(a), f_{\wedge}(a,b)) \tag{5.73}$$

Las funciones generadas a partir de esta expresión utilizando distintas $f_{\neg}$, $f_{\wedge}$ y $f_{\vee}$ se denominan *implicaciones QL*. Por ejemplo, utilizando las funciones de negación, conjunción y disyunción estándares se obtiene la **implicación de Zadeh**:

$$f_{\rightarrow}^{Z}(a,b) = \max(1-a, \min(a,b)) \tag{5.74}$$

[10]Nótese que en la lógica clásica la propiedad $\neg p \vee (p \wedge q)$ es equivalente a $\neg p \vee q$, por lo que es indiferente que definamos $p \rightarrow q$ a partir de una o de otra. Sin embargo, en lógica difusa estas dos propiedades dan lugar a funciones de implicación distintas.

que es una implicación rigurosa. Con la negación estándar, la suma acotada y la diferencia acotada, respectivamente, se obtiene la **implicación de Kleene**.[11]

La ventaja de definir la implicación a partir de una de estas dos propiedades es que la propiedad del límite clásico queda automáticamente garantizada.

También se cumple en la lógica clásica la propiedad siguiente:

$$[p \wedge (p \to q)] \to q \tag{5.75}$$

de modo que

$$f_\wedge(a, \underbrace{f_\to(a,b)}_{c}) \leq b \tag{5.76}$$

En ^{o}, $f_\to(a,b)$ es el máximo de los valores de c que cumplen la propiedad anterior:

$$\begin{array}{ll} f_\wedge(0,c) \leq 0 \Longrightarrow c = 0 \vee c = 1 & f_\to(0,0) = 1 \\ f_\wedge(0,c) \leq 1 \Longrightarrow c = 0 \vee c = 1 & f_\to(0,1) = 1 \\ f_\wedge(1,c) \leq 0 \Longrightarrow c = 0 & f_\to(1,0) = 0 \\ f_\wedge(1,c) \leq 1 \Longrightarrow c = 0 \vee c = 1 & f_\to(1,1) = 1 \end{array} \tag{5.77}$$

es decir

$$f_\to(a,b) = \max\{c \in \{0,1\} \mid f_\wedge(a,c) \leq b\} \tag{5.78}$$

Esta propiedad se puede generalizar para la lógica difusa:

$$f_\to(a,b) = \sup\{c \in [0,1] \mid f_\wedge(a,c) \leq b\} \tag{5.79}$$

Ejercicio 5.30 Demostrar que para toda norma $f_\wedge$ la implicación resultante de esta definición cumple el límite clásico. □

Las funciones de implicación que se obtienen al introducir distintas normas triangulares $f_\wedge$ en esta expresión se denominan *implicaciones R*. Por ejemplo, al tomar la diferencia acotada tenemos la **implicación de Łukasiewicz**,[12] que es, por tanto, una implicación R. En cambio, con la conjunción estándar obtenemos la **implicación de Gödel**:

$$f^G_\to(a,b) = \sup\{c \in [0,1] \mid \min(a,c) \leq b\} = \begin{cases} 1 & \text{si } a \leq b \\ b & \text{si } a > b \end{cases} \tag{5.80}$$

que es una implicación amplia; etc.

[11] En efecto, la implicación resultante es $f_\to(a,b) = \min(1, 1-a+\max(0, a+b-1))$. Si $1-a \geq b$, entonces $\max(0, a+b-1) = 0$ y $f_\to(a,b) = 1-a$; si $1-a < b$, entonces $\max(0, a+b-1) = a+b-1$ y $f_\to(a,b) = b$; por tanto, $f_\to(a,b) = \max(1-a, b) = f^K_\to(a,b)$.

[12] La demostración es la siguiente: teniendo en cuenta que $b \geq 0$,

$$\max(0, a+c-1) \leq b \Longleftrightarrow [(0 \leq b) \wedge (a+c-1 \leq b)] \Longleftrightarrow a+c-1 \leq b$$

de donde se deduce que

$$\sup\{c \in [0,1] \mid \max(0, a+c-1) \leq b\} = \sup\{c \in [0,1] \mid c \leq 1-a+b\} = \min(1, 1-a+b) = f^L_\to(a,b)$$

Proposición 5.31 Toda función de implicación de cualquiera de las tres familias (S, QL o R), cumple la propiedad de "neutralidad de la verdad":

$$\forall b,\ f_{\to}(1,b) = b \tag{5.81}$$

Proposición 5.32 Toda función de implicación S o R es monótona en el primer argumento:

$$\forall a_1, \forall a_2,\ a_1 \leq a_2 \Longrightarrow f_{\to}(a_1,b) \geq f_{\to}(a_2,b) \tag{5.82}$$

Ejercicio 5.33 Demostrar estas dos proposiciones.

Como hemos visto, cada una de estas tres familias procede de una propiedad de la lógica clásica. Otras propiedades podrían dar lugar a otras familias, aunque en la literatura sólo se han estudiado estas tres. Por cierto, hemos comprobado ya que estas familias no son exclusivas, porque algunas funciones de implicación que conocemos pertenecen a dos de ellas; tampoco son exhaustivas, porque la **implicación de Gaines-Rescher**:

$$f_{\to}^{GR}(a,b) = \begin{cases} 1 & \text{si } a \leq b \\ 0 & \text{si } a > b \end{cases} \tag{5.83}$$

por ejemplo, no pertenece a ninguna de ellas.

La tabla 5.14 muestra estas funciones de implicación, junto con las propiedades que cumplen y el año en fueron publicadas por primera vez. Las propiedades que cumplen son las siguientes: LC es el **límite clásico**, C indica que la función es **continua**, y las letras A y R indican si se trata de una implicación **amplia** o **rigurosa** (podría ocurrir que alguna de estas funciones no fuera ni amplia ni rigurosa, pero no es el caso); los números 1 a 7 corresponden a las siete propiedades de la tabla 5.13, que a su vez proceden de las propiedades de la implicación clásica (tabla 5.6, pág. 103). Observe que la única función de implicación que no cumple la propiedad de "neutralidad de la verdad" es la de Gaines-Rescher, que no pertenece a ninguna de las familias S, QL o R, como exige la proposición 5.31 (ec. (5.81)), y que las funciones que no cumplen la monotonía en el primer argumento —la de Zadeh y la de Gaines-Rescher— no pueden pertenecer a S ni a R, por la proposición 5.32 (ec. (5.82)).

1.	Neutralidad de la verdad	$f_{\to}(1,a) = a$
2.	Predominio de la falsedad	$f_{\to}(0,a) = 1$
3.	Identidad	$f_{\to}(a,a) = 1$
4.	Intercambio	$f_{\to}(a, f_{\to}(b,c)) = f_{\to}(b, f_{\to}(a,c))$
5.	Contraposición	$f_{\to}(a,b) = f_{\to}(1-b, 1-a)$
6.	Monotonía-v en el 1$^{\text{er}}$ argumento	$a \leq b \Longrightarrow f_{\to}(a,c) \geq f_{\to}(b,c)$
7.	Monotonía-v en el 2$^{\text{o}}$ argumento	$a \leq b \Longrightarrow f_{\to}(c,a) \leq f_{\to}(c,b)$

Tabla 5.13: Algunas propiedades que cumplen ciertas funciones de implicación.

Al igual que hicimos en la sección 5.1.2, podríamos intentar ahora obtener el valor de $v(q)$ —o, al menos, alguna restricción para $v(q)$— a partir de los valores de $v(p \to q)$ y $v(p)$, con lo que tendríamos un modus ponens difuso, y análogamente, un modus tollens difuso. Sin embargo, no es éste el método que se sigue habitualmente en la lógica difusa, sino que se llega a estos silogismos mediante la composición de relaciones difusas, tal como veremos en la sección 5.4.3.

Autor	$f_{\rightarrow}(a,b)$	**Tipo**	**Propiedades**	**Año**
Łukasiewicz	$\min(1, 1-a+b)$	S, R	LC, C, A, 1, 2, 3, 4, 5, 6, 7	1920
Kleene	$\max(1-a, b)$	S, QL	LC, C, R, 1, 2, –, 4, 5, 6, 7	1938
Reichenbach	$1-a+ab$	S	LC, C, R, 1, 2, –, 4, 5, 6, 7	1935
Zadeh	$\max(1-a, \min(a,b))$	QL	LC, C, R, 1, 2, –, –, –, –, 7	1973
Gödel	$\begin{cases} 1 & \text{si } a \leq b \\ b & \text{si } a > b \end{cases}$	R	LC, C, A, 1, 2, 3, 4, –, 6, 7	1976
Gaines-Rescher	$\begin{cases} 1 & \text{si } a \leq b \\ 0 & \text{si } a > b \end{cases}$		LC, –, A, –, 2, 3, 4, 5, 6, 7	1969

Tabla 5.14: Algunas de las funciones de implicación más conocidas.

Modificadores difusos para proposiciones

Dado un conjunto de proposiciones, podemos aplicar a cada una de ellas uno o varios modificadores difusos, correspondientes a expresiones lingüísticas, para dar lugar a proposiciones difusas como "p es muy cierta" o, de forma equivalente, "es muy cierto que p". El valor de verdad de la proposición resultante se calcula aplicando cierta función matemática al valor de la proposición original p. Concretamente, algunos de los modificadores difusos más habituales y sus correspondientes funciones son:

$$v(\text{“}p \text{ es muy cierto”}) = v(p)^2$$

$$v(\text{“}p \text{ es bastante cierto”}) = \begin{cases} 2v(p)^2 & \text{si } v(p) \leq 0'5 \\ 1-2[1-v(p)]^2 & \text{si } v(p) > 0'5 \end{cases}$$

$$v(\text{“}p \text{ es más o menos cierto”}) = v(p)^{\frac{1}{2}}$$

$$v(\text{“}p \text{ es falso”}) = 1 - v(p)$$

$$v(\text{“}p \text{ es muy falso”}) = [1-v(p)]^2$$

En la sección 5.2.1 veremos que existen modificadores equivalentes para los predicados difusos.

5.2 Lógica de predicados

5.2.1 Predicados unitarios

Dado un conjunto X no vacío, denominado *conjunto universal*, cada *predicado* unitario P puede definirse como una función que asigna a cada elemento x de X la proposición $P(x) =$ "x es P". Por ejemplo, el predicado "Mayor que 3" asigna al elemento 5 la proposición "5 es mayor que 3"; en la notación habitual, "Mayor-que-3(5)" = "5 es mayor que 3". Igualmente, dado el predicado "Hermano de Juan" y el elemento Antonio, "Hermano-de-Juan(Antonio)" = "Antonio es hermano de Juan".

Con el fin de evitar sutilezas matemáticas, vamos a suponer generalmente que X es un conjunto finito, como ocurre en todos los problemas prácticos. Por ejemplo, si X es un conjunto de personas, siempre es finito, aunque incluyéramos en él todas las personas que han

existido y existirán en los próximos milenios. Incluso las medidas supuestamente continuas se realizan siempre en la práctica sobre conjuntos finitos.[13]

Al predicado $P^{\mathbf{1}}$ que asigna a todo elemento de X la proposición segura (cf. pág. 98), $\forall x,\ P^{\mathbf{1}}(x) = \mathbf{1}$, le denominamos *predicado seguro*, pues $\forall x, \forall v,\ v(P^{\mathbf{1}}(x)) = 1$. Análogamente, al predicado $P^{\mathbf{0}}$, tal que $\forall x,\ P^{\mathbf{0}}(x) = \mathbf{0}$, le denominamos *predicado imposible*, pues $\forall x, \forall v,\ v(P^{\mathbf{0}}(x)) = 0$.

Un *predicado preciso* P es aquél que siempre origina proposiciones $P(x)$ precisas, mientras que un *predicado impreciso* es aquél que genera proposiciones imprecisas, al menos para ciertos valores x de X. Por ejemplo, "Mayor que 3" es un predicado preciso, mientras que "Aproximadamente igual a cero" es impreciso, ya que la proposición "Aproximadamente-igual-a-cero(0'25)", es decir, "0'25 es aproximadamente igual a 0", no es totalmente verdadera ni totalmente falsa.

Los conceptos que se utilizan en matemáticas corresponden generalmente a predicados precisos; por ejemplo, en la aritmética tenemos conceptos como par, impar, primo, mayor que, menor que, etc., que son claramente verdaderos o falsos. Sin embargo, también existen otros muchos —como grande, pequeño, mucho mayor que, mucho menor que, aproximadamente igual a...— que dan lugar a proposiciones imprecisas. Por ejemplo, la proposición "0'01 es aproximadamente igual a 0" es más cierta que "0'1 es aproximadamente igual a 0", pero ésta tampoco es totalmente falsa.

En la vida cotidiana, podemos encontrar predicados precisos, como mayor-de-edad, soltero, casado, hijo-único, tiene-permiso-de-conducir, etc., pero son muchos más los predicados imprecisos: joven, viejo, alto, bajo, gordo, delgado, rico, pobre, inteligente, habla-inglés, sabe-informática, etc., etc. Incluso en la ciencia abundan los predicados imprecisos, especialmente en el campo de la medicina, donde encontramos numerosas expresiones difusas, como sano, enfermo, edad avanzada, presión alta, dolor agudo, fatiga leve, tumor grande, síntoma evidente, técnica sensible, diagnóstico complejo, pronóstico grave, terapia arriesgada, alta mortalidad, etc., etc.

Como veremos más adelante, los predicados precisos dan lugar a conjuntos nítidos (los conjuntos clásicos), en los que un elemento o pertenece completamente al conjunto o no pertenece en absoluto, mientras que los predicados imprecisos dan lugar a conjuntos difusos, en los que el grado de pertenencia varía dentro del intervalo $[0, 1]$.

Cuantificadores y predicados compuestos

El *cuantificador universal* $\forall$ se define así:

$$\forall x, P(x) \equiv \bigwedge_{x \in X} P(x) \tag{5.84}$$

[13] Problamente el lector estará pensando: "¿Y qué pasa con las escalas continuas, como la estatura, por ejemplo? Lo habitual es considerar que la la estatura se mide sobre la escala continua de los números reales positivos ($X = \mathbb{R}^+$)". Nuestra respuesta es la siguiente: en contra de lo que se suele afirmar, insistimos en que la estatura, como cualquier otra magnitud, se mide sobre un conjunto X finito. De hecho, la estatura de una persona no suele ser superior a 2'50 m., y se suele medir con una precisión de centímetros. Para asegurarnos que no nos quedamos cortos en los límites, supongamos que tenemos el conjunto X de todos los números racionales de 0 a 10 con 5 decimales o menos; ciertamente es un conjunto finito más que suficiente para medir la estatura de las personas (en metros). Del mismo modo, cualquier otra escala de las consideradas continuas puede representarse —por tomar límites generosos— mediante el conjunto de los números decimales de -10^{500} a 10^{500} con un máximo de 200 decimales, que es un conjunto enorme (2^{700}), pero finito.

Por tanto, sostenemos que en la práctica no es una limitación el suponer que X es finito.

y el *cuantificador existencial* $\exists$ así:

$$\exists x, P(x) \equiv \bigvee_{x \in X} P(x) \tag{5.85}$$

Por tanto,

$$v(\forall x, P(x)) = \underset{x \in X}{f_\wedge} v(P(x)) \tag{5.86}$$

$$v(\exists x, P(x)) = \underset{x \in X}{f_\vee} v(P(x)) \tag{5.87}$$

donde $f_\wedge$ y $f_\vee$ son las funciones de conjunción y disyunción generalizadas para n argumentos, que se definen recursivamente a partir de las respectivas funciones binarias:[14]

$$f_\wedge(a) = a \quad f_\wedge(a_1, \ldots, a_n) = f_\wedge(f_\wedge(a_1, \ldots, a_{n-1}), a_n) \tag{5.88}$$

$$f_\vee(a) = a \quad f_\vee(a_1, \ldots, a_n) = f_\vee(f_\vee(a_1, \ldots, a_{n-1}), a_n) \tag{5.89}$$

Ejemplo 5.34 Dado el conjunto universal $X = \{x_1, x_2, x_3\}$ entonces

$$\forall x, P(x) \equiv P(x_1) \wedge P(x_2) \wedge P(x_3)$$

y por tanto,

$$v(\forall x, P(x)) = f_\wedge(v(P(x_1)), v(P(x_2)), v(P(x_3)))$$

es decir, la proposición $\forall x, P(x)$ es cierta si y sólo si cada una de las tres proposiciones $P(x_i)$ es cierta. Análogamente,

$$\exists x, P(x) \equiv P(x_1) \vee P(x_2) \vee P(x_3)$$

$$v(\exists x, P(x)) = f_\vee(v(P(x_1)), v(P(x_2)), v(P(x_3)))$$

de modo que $\exists x, P(x)$ si y sólo si alguna de las proposiciones $P(x_i)$ es cierta. □

Nótese que, si P es un predicado preciso, las proposiciones "$\forall x, P(x)$" y "$\exists x, P(x)$" son precisas, mientras si P es difuso las proposiciones $\forall x, P(x)$ y $\exists x, P(x)$ son difusas.[15]

Proposición 5.35 Las funciones de conjunción y disyunción generalizadas cumplen que

$$f_\wedge(a_1, \ldots, a_n) = 1 \iff \forall i,\ a_i = 1 \tag{5.90}$$

$$f_\vee(a_1, \ldots, a_n) = 0 \iff \forall i,\ a_i = 0 \tag{5.91}$$

Demostración. Por la definición de $f_\wedge$ y el hecho de que $f_\wedge(1, 1) = 1$, es fácil probar por inducción completa que

$$\forall i,\ a_i = 1 \Rightarrow f_\wedge(a_1, \ldots, a_n) = 1$$

La implicación recíproca se demuestra aplicando la proposición 5.17 repetidamente,

$$f_\wedge(a_1, \ldots, a_n) = f_\wedge(f_\wedge(a_1, \ldots, a_{n-1}), a_n) = 1 \Longrightarrow a_n = f_\wedge(a_1, \ldots, a_{n-1}) = 1$$

hasta llegar a demostrar que $a_n = a_{n-1} = \ldots = a_1 = 1$.

[14] Cuando X es finito, la generación de $f_\wedge(a, b) = \min(a, b)$ sigue siendo la función min; en cambio, si X es infinito, la generalización de la función mínimo es la función ínfimo, del mismo modo que la generalización de la función máximo es la función supremo.

[15] Existen también cuantificadores difusos —como "muchos", "casi todos", "casi ninguno", "unos 15", "muchos más de 100"— que no vamos a tratar aquí. El lector interesado puede consultar [35, sec. 8.4].

Corolario 5.36 La proposición $\forall x, P(x)$ es cierta si y sólo si el predicado $P(x)$ es cierto para todo x:

$$\forall x, P(x) \Longleftrightarrow v(\forall x, P(x)) = 1 \Longleftrightarrow \forall x, v(P(x)) = 1 \tag{5.92}$$

Corolario 5.37 La proposición $\exists x, P(x)$ es falsa si y sólo si el predicado $P(x)$ es falso para todo x:

$$\neg(\exists x, P(x)) \Longleftrightarrow v(\exists x, P(x)) = 0 \Longleftrightarrow \forall x, v(P(x)) = 0 \tag{5.93}$$

La negación, conjunción y disyunción de predicados se definen a partir de las correspondientes conectivas para proposiciones (nótese que en cada una de estas definiciones hay un "$\forall x$" implícito, y que el signo "$=$" indica igualdad de proposiciones):

$$[\neg P](x) = \neg[P(x)] \tag{5.94}$$
$$[P \wedge Q](x) = P(x) \wedge Q(x) \tag{5.95}$$
$$[P \vee Q](x) = P(x) \vee Q(x) \tag{5.96}$$

Proposición 5.38 Para todo conjunto X finito (no vacío) se cumple que

$$\neg\left(\forall x, P(x)\right) \equiv \exists x, \neg P(x) \tag{5.97}$$
$$\neg\left(\exists x, P(x)\right) \equiv \forall x, \neg P(x) \tag{5.98}$$

Demostración. Dado un conjunto de proposiciones, la 1ª ley de Morgan generalizada para n proposiciones (donde n es un entero positivo), $\neg(p_1 \wedge \ldots \wedge p_n) = \neg p_1 \vee \ldots \vee \neg p_n$, se prueba por inducción completa sobre n. Para $n = 1$ es trivial. Cuando, $n = 2$ es la ley de Morgan ordinaria (cf. tabla 5.4); en la sección 5.1.1 se explicó cómo demostrarla. Si la ley generalizada se cumple para $n - 1$, entonces

$$\begin{aligned}\neg(p_1 \wedge \ldots \wedge p_n) &= \neg[(p_1 \wedge \ldots \wedge p_{n-1}) \wedge p_n] = \neg(p_1 \wedge \ldots \wedge p_{n-1}) \vee \neg p_n \\ &= (\neg p_1 \vee \ldots \vee \neg p_{n-1}) \vee \neg p_n = \neg p_1 \vee \ldots \vee \neg p_n\end{aligned}$$

Con este resultado, la equivalencia (5.97) se demuestra simplemente aplicando la definición de $\forall$ y $\exists$. La demostración de (5.98) es similar. Nótese que esta demostración por inducción completa es correcta porque X es finito; para un conjunto infinito no sería válida. □

Comparación de predicados

Definición 5.39 (Equivalencia de predicados) Dos predicados P y Q son *equivalentes* (se representa mediante "$P \equiv Q$") cuando las respectivas proposiciones asignadas a cada x son equivalentes:

$$P \equiv Q \Longleftrightarrow [\forall x,\ P(x) \equiv Q(x)] \Longleftrightarrow [\forall x,\ \forall v,\ v(P(x)) = v(Q(x))] \tag{5.99}$$

Por ejemplo, para todo P y todo Q, tenemos que $P \wedge Q \equiv Q \wedge P$. En realidad, todas las propiedades de la tabla 5.4 siguen siendo válidas si se sustituye cada proposición (representada por una letra minúscula) por un predicado (letra mayúscula).

Del mismo modo que podemos aplicar las conectivas de negación, conjunción y disyunción a unos predicados con el fin de obtener nuevos predicados, podemos aplicar las conectivas de

implicación y doble-implicación a dos predicados, aunque con la diferencia muy importante de que en este caso no se obtienen nuevos predicados sino proposiciones:

$$P \to Q \equiv [\forall x,\ P(x) \to Q(x)] \tag{5.100}$$
$$P \leftrightarrow Q \equiv [\forall x,\ P(x) \leftrightarrow Q(x)] \tag{5.101}$$

Insistiendo en el punto anterior, recomendamos comparar estas dos últimas definiciones con la (5.95) y la (5.96) para entender por qué $P \to Q$ y $P \leftrightarrow Q$ son proposiciones, mientras que $\neg P$, $P \wedge Q$ y $P \vee Q$ son predicados.

Cuando P y Q son predicados precisos, entonces $P \to Q$ y $P \leftrightarrow Q$ son proposiciones precisas (por la propiedad del límite clásico), mientras que cuando P y Q son predicados imprecisos las proposiciones resultantes son imprecisas. Nótese la diferencia con la equivalencia de predicados (ec. (5.99)), pues $P \equiv Q$ es siempre una proposición precisa, aunque P y Q sean imprecisos. Otra diferencia muy importante es que la verdad o falsedad de $P \equiv Q$ no depende de la función de verdad v (cf. ec. (5.99)), mientras que el grado de verdad de $P \to Q$ y $P \leftrightarrow Q$ sí depende de v. La relación entre $P \equiv Q$ y $P \leftrightarrow Q$ es la siguiente:

Proposición 5.40 Cuando "$\leftrightarrow$" es una equivalencia amplia,

$$P \equiv Q \iff \forall v,\ P \leftrightarrow Q \tag{5.102}$$

Demostración. Por la definición de doble-implicación amplia (tabla 5.2),

$$P \equiv Q \iff \forall x,\ \forall v,\ v(P(x)) = v(Q(x)) \iff \forall v,\ \forall x,\ P(x) \leftrightarrow Q(x) \iff \forall v,\ P \leftrightarrow Q$$

□

Propiedades de la lógica de predicados

Naturalmente, las propiedades de una lógica de predicados dependen de la lógica de proposiciones en que se basa. Por ejemplo, para la lógica de proposiciones clásica tenemos la lógica de predicados clásica; si en la tabla 5.4 sustituimos cada proposición p por un predicado P, la proposición $\mathbf{1}$ por $P^{\mathbf{1}}$ y $\mathbf{0}$ por $P^{\mathbf{0}}$, todas las propiedades —excepto las de monotonía-v, que no tendrían sentido— siguen siendo válidas. En cambio, ninguna lógica multivaluada ni lógica difusa puede satisfacer todas esas propiedades, por los motivos discutidos en la sección 5.1.3.

Ciñiéndonos igualmente al ámbito de la lógica clásica, a partir de las propiedades de la implicación de proposiciones (tabla 5.6, pág. 103) se deducen las propiedades de la implicación de predicados que aparecen en la tabla 5.15. Por ejemplo, la primera de ellas se demuestra así:

$$P^{\mathbf{1}} \to P \equiv \big[\forall x,\ P^{\mathbf{1}}(x) \to P(x)\big] \equiv [\forall x,\ \mathbf{1} \to P(x)] \equiv [\forall x,\ P(x)] \equiv \bigwedge_{x \in X} P(x)$$

Por cierto, $P^{\mathbf{0}} \to P$ y $P \to P$ son tautologías, porque son equivalentes a la proposición segura.

Modificadores difusos para predicados

Del mismo modo en que se aplican modificadores a proposiciones (sec. 5.1.3), es posible también aplicar modificadores lingüísticos a los predicados. Por ejemplo, al predicado "Grande"

Implicación	Neutralidad del predicado seguro	$P^{\mathbf{1}} \to P \equiv \bigwedge_{x \in X} P(x)$
	Predominio del predicado imposible	$P^{\mathbf{0}} \to P \equiv \mathbf{1}$
	Identidad	$P \to P \equiv \mathbf{1}$
	Intercambio	$P \to (Q \to R) \equiv Q \to (P \to R)$
	Contraposición	$P \to Q \equiv \neg Q \to \neg P$
	Monotonía-p en el 1$^{\text{er}}$ argumento	$P \to Q \Longrightarrow (Q \to R) \to (P \to R)$
	Monotonía-p en el 2$^{\text{o}}$ argumento	$P \to Q \Longrightarrow (R \to P) \to (R \to q)$

Tabla 5.15: Propiedades de la implicación de predicados.

podemos aplicarle el modificador "muy" con el fin de obtener el predicado "Muy grande". A cada modificador mod le corresponde una función matemática tal que

$$v(\text{“[mod } P](x)\text{”}) = f_{\text{mod}}(v(P(x))) \tag{5.103}$$

Algunos de los modificadores difusos más habituales y sus correspondientes funciones son:

$$v(\text{“Muy-}P(x)\text{”}) = v(P(x))^2$$

$$v(\text{“Bastante-}P(x)\text{”}) = \begin{cases} 2v(P(x))^2 & \text{si } v(P(x)) \leq 0'5 \\ 1 - 2[1 - v(P(x))]^2 & \text{si } v(P(x)) > 0'5 \end{cases}$$

$$v(\text{“Más-o-menos-}P(x)\text{”}) = v(P(x))^{\frac{1}{2}}$$

$$v(\text{“No-}P(x)\text{”}) = 1 - v(P(x))$$

$$v(\text{“No-muy-}P(x)\text{”}) = [1 - v(P(x))]^2$$

5.2.2 Modus ponens para predicados

Hemos visto ya en las secciones 5.1.1 y 5.1.2 cómo se puede aplicar el modus ponens entre proposiciones. En ésta vamos a mostrar cómo aplicarlo en la lógica de predicados. El ejemplo más típico es el siguiente: de la regla "todos los hombres son mortales" y la afirmación "Sócrates es hombre" se puede deducir que "Sócrates es mortal". Formalmente se representa así: si P es el predicado "Hombre" y Q el predicado "Mortal", la regla anterior puede expresarse mediante la proposición $P \to Q$, que, tal como fue definida en la ecuación (5.100), es equivalente a $\forall x,\ P(x) \to Q(x)$. La afirmación "Sócrates es hombre" se puede representar mediante P(Sócrates), y la conclusión a la que queremos llegar es Q(Sócrates) = "Sócrates es mortal".

En este caso, como los predicados que intervienen son precisos (las proposiciones resultantes son totalmente verdaderas o totalmente falsas), podemos abordarlo desde la lógica clásica. Formalmente se representa así:

$$\frac{\begin{array}{l} P \to Q \\ P(x_0) \end{array}}{Q(x_0)}$$

o, si se prefiere,

$$\frac{\begin{array}{l}\forall x,\ P(x) \to Q(x)\\ P(x_0)\end{array}}{Q(x_0)}$$

La justificación del razonamiento es la siguiente. Según el corolario 5.36, cuando la proposición $P \to Q$ es cierta, ha de ser cierta también cada una de las proposiciones $P(x) \to Q(x)$; en particular, $P(x_0) \to Q(x_0)$ ha de ser cierta, y teniendo en cuenta que tanto $P(x_0)$ como $Q(x_0)$ son proposiciones, podemos aplicar el modus ponens que vimos en la sección 5.1.1:

$$\frac{\begin{array}{l}P(x_0) \to Q(x_0)\\ P(x_0)\end{array}}{Q(x_0)}$$

En caso de que P y Q sean predicados imprecisos, debemos recurrir a alguna lógica multivaluada o difusa. En este caso, tenemos que

$$v(P \to Q) = v(\forall x,\ P(x) \to Q(x)) = \underset{x\in X}{f_\wedge}\, v(P(x) \to Q(x))$$

y, por la ecuación (5.50), llegamos a[16]

$$\underset{x\in X}{f_\wedge}\, v(P(x) \to Q(x)) \leq \min_{x\in X} v(P(x) \to Q(x))$$

de donde se deduce que

$$v(P(x_0) \to Q(x_0)) \geq \min_{x\in X} v(P(x) \to Q(x)) = v(P \to Q) \tag{5.104}$$

En caso de la lógica de Łukasiewicz, uniendo esta última ecuación a la (5.23) y a la (5.24) tenemos que

$$\begin{aligned} v(Q(x_0)) &\geq v(P(x_0)) + v(P(x_0) \to Q(x_0)) - 1 \geq v(P(x_0)) + v(P \to Q) - 1 && (5.105)\\ &= v(P(x_0)) - [1 - v(P \to Q)] && (5.106)\end{aligned}$$

De nuevo se comprueba que "$v(P \to Q) \approx 1$ y $v(P(x_0)) \approx 1$ implica que $v(Q(x_0)) \approx 1$".

Por ejemplo, supongamos que tenemos la regla "Toda curva cerrada es peligrosa" con un grado de verdad de 0'8 y la afirmación "La curva es cerrada" con un grado de verdad de 0'7. De aquí podemos deducir que el grado de verdad de la afirmación "La curva es peligrosa" es, al menos, $0'7 - (1 - 0'8) = 0'5$.

Del mismo modo se podría estudiar el modus tollens para la lógica de Łukasiewicz, y ambos silogismos para la lógica de Kleene, pero no nos vamos a alargar más en esta exposición. Tampoco nos vamos a detener a analizar el modus ponens con reglas como "$P(x) \to Q(y)$", porque esta regla no tiene cuantificadores y, por tanto, basta aplicar el modus ponens para proposiciones, tal como fue expuesto en la sección 5.1.1.

Nótese que, en la exposición que acabamos de hacer del modus ponens, el predicado P que aparece en la regla $P \to Q$ ha de ser el mismo que el de la afirmación $P(x_0)$. En caso de tener una afirmación $P'(x_0)$ no podríamos deducir nada, aunque el predicado P' fuera casi igual a P. En la sección 5.4.3 veremos cómo realizar inferencia imprecisa en este caso: es lo que se denomina *modus ponens difuso*.

[16] Si X fuera infinito deberíamos tomar el ínfimo en vez del mínimo, pero la demostración sería igualmente válida.

5.3 Teoría de conjuntos

5.3.1 Conjuntos y predicados

Supongamos que tenemos un predicado preciso P que toma valores dentro de cierto conjunto finito X. Para ciertos elementos de X, la proposición $P(x)$ será cierta —es decir, $v(P(x)) = 1$—, mientras que para otros será falsa: $v(P(x)) = 0$. Por tanto, todo predicado define un subconjunto A de X, formado por aquéllos elementos de X que cumplen la condición impuesta por el predicado:

$$A = \{x \in X \mid P(x)\} \Longleftrightarrow x \in A \equiv P(x) \tag{5.107}$$

Ejemplo 5.41 Sea el conjunto universal $X_{500} = \{x \in \mathbb{Z} \mid 0 < x < 10^{500}\}$ y el predicado "Múltiplo de 7". Tenemos que v("Múltiplo-de-7(14)")=1, mientras que v("Múltiplo-de-7(9)")=0. Es decir, $x \in A \equiv$ "Múltiplo-de-7(x)", de modo que $A = \{0, 7, 14, 21, \ldots\}$. □

Recíprocamente, podemos definir el predicado P_A, "pertenece a A", que asigna a cada elemento x la proposición "$x \in A$":

$$P_A(x) = x \in A \tag{5.108}$$

Uniendo esta observación a la anterior, concluimos que **todo predicado induce un subconjunto y todo subconjunto induce un predicado**.

Se cumple además que

$$A = \{x \in X \mid P(x)\} \Longleftrightarrow x \in A \equiv P(x) \Longleftrightarrow P_A \equiv P \tag{5.109}$$

Cuando P es un predicado impreciso, las tres ecuaciones anteriores son válidas, pero entonces la proposición $x \in A$ puede tomar valores distintos de 0 y 1. Es decir, puede haber algún elemento x que no pertenezca totalmente a A, aunque tampoco sea totalmente cierto que no pertenezca. Se dice en este caso que el conjunto A, inducido por un *predicado impreciso*, es un *conjunto difuso*, en contraposición a los conjuntos inducidos por *predicados precisos*, que se denominan *conjuntos nítidos*.[17] Por ejemplo, dado el predicado impreciso "mucho mayor que 10", el conjunto de los números mucho mayores que 10 es un conjunto difuso. ¿Pertenece el número 60 a este conjunto? La respuesta no es ni un sí rotundo ni un no rotundo. Ciertamente, el número 60 pertenece más que el 40 y menos que el 100, de modo que $0 < v(40 \in A) < v(60 \in A) < v(100) < 1$. Volveremos sobre este punto al hablar del grado de pertenencia.

Proposición 5.42 Dos predicados son equivalentes si y sólo si definen el mismo subconjunto.

Demostración. Para dos predicados cualesquiera P y Q,

$$P \equiv Q \Longrightarrow P(x) \equiv Q(x) \Longrightarrow \{x \in X \mid P(x)\} = \{x \in X \mid Q(x)\} \tag{5.110}$$

Demostramos ahora la implicación recíproca. Sea A el conjunto definido por P; de las ecuaciones (5.107) y (5.108) se deduce que $P \equiv P_A$. Si Q define el mismo conjunto A, entonces $Q \equiv P_A$, y por tanto $P \equiv Q$. □

[17] Cuando se dice "conjunto difuso" puede entenderse en sentido restringido, como "conjunto no nítido", o en sentido amplio, de modo que los conjuntos nítidos son un caso particular de los conjuntos difusos. Del contexto se deducirá en cada caso el sentido utilizado.

Esta proposición es muy importante, pues nos asegura que para cada propiedad de la doble-implicación entre predicados existe una propiedad de igualdad entre conjuntos, y viceversa, como veremos en las próximas secciones.

Proposición 5.43 El predicado seguro define el conjunto total, mientras que el predicado imposible define el conjunto vacío:

$$P_X \equiv P^{\mathbf{1}} \tag{5.111}$$

$$P_\varnothing \equiv P^{\mathbf{0}} \tag{5.112}$$

5.3.2 Funciones características

Función característica de un conjunto nítido

Dada una función $\mu : X \to \{0, 1\}$, algunos elementos de X tomarán el valor 1 mientras que otros tomarán el valor 0. Los primeros forman un subconjunto A:

$$A = \{x \in X \mid \mu(x) = 1\} \tag{5.113}$$

Ejemplo 5.44 Sea el conjunto universal X_{500} definido en el ejemplo 5.41 y la función μ

$$\mu(x) = x \bmod 2$$

que asigna 1 a los números impares y 0 a los números pares. Por tanto

$$A = \{x \in X \mid x \bmod 2 = 1\} = \{1, 3, 5, 7, \ldots\}$$

□

Recíprocamente, todo subconjunto nítido A define una función $\mu_A : X \to \{0, 1\}$, denominada *función característica*, que toma el valor 1 para los elementos que pertenecen a A y 0 para los que no pertenecen:

$$\mu_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases} \tag{5.114}$$

es decir,

$$\mu_A(x) = v(x \in A) \tag{5.115}$$

El valor $\mu_A(x)$ se denomina *pertenencia* del elemento x al conjunto A. Obviamente,

$$\mu_X(x) = 1, \ \forall x \tag{5.116}$$

$$\mu_\varnothing(x) = 0, \ \forall x \tag{5.117}$$

Es decir, **toda función** $\mu : X \to \{0, 1\}$ **define un subconjunto (nítido)** y **todo subconjunto** A **induce una función** $\mu_A : X \to \{0, 1\}$.

Uniendo esta afirmación a los resultados de la sección anterior, podemos concluir que

Proposición 5.45 Existe una relación biunívoca entre

- los subconjuntos nítidos de X,
- los predicados precisos definidos sobre X y

- las funciones $\mu : X \to \{0, 1\}$. □

De las ecuaciones (5.107) y (5.115) se deduce inmediatamente que

Proposición 5.46 Para todo predicado P,

$$\boxed{A = \{x \in X \mid P(x)\} \Longleftrightarrow x \in A \equiv P(x) \Longleftrightarrow P_A \equiv P \Longleftrightarrow \mu_A(x) = v(P(x))} \tag{5.118}$$

□

Dicho de otro modo: si A viene definido por un predicado, la función característica μ_A se obtiene a partir de los valores de verdad correspondientes. Tenemos, por tanto, cuatro formas equivalentes de caracterizar la relación entre un predicado y el conjunto inducido por él. Este resultado es la piedra angular en que basaremos todos los desarrollos del resto del capítulo.

Ejemplo 5.47 Dado de nuevo el conjunto universal X_{500}, queremos definir el subconjunto A de los números menores que 10. En este caso, el predicado es P="Menor que 10", que asigna a cada x la proposición "x es menor que 10". Por tanto,

$$A = \{x \in X \mid P(x)\} = \{x \in X \mid x < 10\}$$

La función característica correspondiente es

$$\mu_A(x) = v(x < 10)$$

de modo que

$$\mu_A(x) = \begin{cases} 1 & \text{si } x < 10 \\ 0 & \text{si } x \geq 10 \end{cases}$$

Ejercicio 5.48 Dibuje las funciones $\mu(x)$ y $\mu_A(x)$ que aparecen en los dos ejemplos anteriores (para $x \leq 15$).

Función característica de un conjunto difuso

Acabamos de ver que existe un isomorfismo entre los subconjuntos nítidos, los predicados precisos y las funciones $\mu : X \to \{0, 1\}$, tal como indica la expresión (5.118). De modo análogo, si partimos de un subconjunto difuso A y aplicamos la ecuación (5.115) define una función $\mu_A : X \to [0, 1]$. Recíprocamente, toda función $\mu : X \to [0, 1]$ indica en qué medida cada elemento de X pertenece a A,

$$v(x \in A) = \mu(x) \tag{5.119}$$

que es tanto como definir A. Por tanto, concluimos que para cada función $\mu : X \to [0, 1]$ existe un subconjunto difuso de X, y viceversa. Uniendo este resultado a los de la sección 5.3.1 podemos concluir que

Proposición 5.49 Existe una relación biunívoca entre

- los subconjuntos difusos de X,
- los predicados imprecisos definidos sobre X y

- las funciones $\mu : X \to [0,1]$. □

Por las mismas razones, la expresión (5.118) sigue siendo válida si en vez de tener conjuntos nítidos, predicados precisos y funciones $\mu : X \to \{0,1\}$, tenemos conjuntos difusos, predicados imprecisos y funciones $\mu : X \to [0,1]$.

Por tanto, una forma posible de definir un conjunto difuso es dar su función característica. También en la teoría de conjuntos clásica era posible definir los conjuntos (nítidos) mediante funciones características, pero ésta era una posibilidad que no se utilizaba en la práctica, pues casi siempre resulta más fácil enunciar el predicado que define un conjunto —por ejemplo, $A = \{x \in X \mid \text{Múltiplo-de-3}(x)\}$— o enumerar los elementos que lo componen —en este caso, $A = \{0, 3, 6, 9, \ldots\}$— que dar la función característica. Sin embargo, la forma más eficiente de definir un conjunto estrictamente difuso consiste en dar su función característica, especialmente cuando X es un conjunto numérico y $\mu_A(x)$ puede expresarse mediante una función algebraica.

Ejemplo 5.50 Sea el conjunto universal formado por los números reales ($X = \mathbb{R}$); el conjunto A de los números próximos a 0 puede definirse así

$$\mu_A(x) = \frac{1}{1 + \beta x^2}$$

donde β es un parámetro que puede ajustarse según convenga (figura 5.1). □

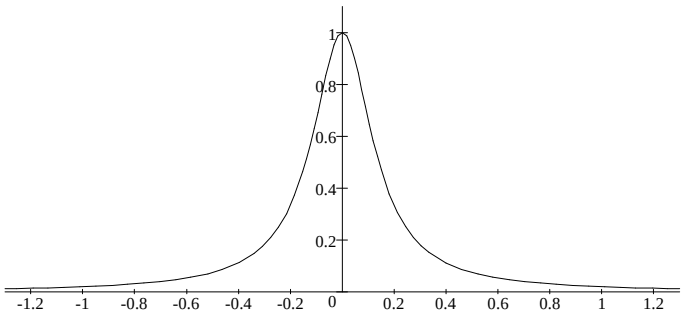

Figura 5.1: Función característica del conjunto A de números próximos a 0 (β=50).

Sin embargo, cuando X no es un conjunto numérico o no hemos encontrado una función numérica que exprese adecuadamente μ_A, la definición funcional de A es imposible. Lo que se hace en este caso es definir A indicando explícitamente el grado de pertenencia de cada x:

$$A = \mu_A(x_1)|x_1 + \ldots + \mu_A(x_n)|x_n \tag{5.120}$$

Naturalmente, esta forma de definir A sólo es posible cuando X es un conjunto finito, o cuando $\mu_A(x)$ solamente es mayor que 0 para un subconjunto finito de X (que se suele denominar *soporte de* A, y es un conjunto nítido).

Ejemplo 5.51 Sea de nuevo el conjunto universal X_{500}; el conjunto A de los números próximos a 0 puede definirse mediante

$$A = 1|0 + 0'7|1 + 0'4|2 + 0'05|3 \tag{5.121}$$

y se entiende que $\mu_A(x) = 0$ para $x \geq 4$. □

En otros casos es posible utilizar una **escala numérica**. Por ejemplo, sea un conjunto de varones $X = \{\text{Antonio, Juan, Luis, Roberto}\ldots\}$, y queremos determinar cuál es el conjunto A de las personas altas. Nuestro sentido común nos dice que el grado de pertenencia a A depende de la estatura: si Juan y Antonio miden lo mismo, $\mu_A(\text{Juan}) = \mu_A(\text{Antonio})$. Por tanto, dada una escala Y para medir la estatura (por ejemplo, en centímetros) y la función $f_Y(x) : X \to Y$ que a cada persona le asigna su estatura, podemos definir una función $\mu'_A(y) : Y \to [0, 1]$, de modo que

$$\mu_A(x) = \mu'_A(f_Y(x)) \tag{5.122}$$

es decir, $\mu_A = \mu'_A \circ f_Y$. En nuestro ejemplo, la función μ'_A podría ser

$$\mu'_A(y) = \frac{1}{1 + e^{-0'45(y-175)}} \tag{5.123}$$

de modo que si $f_Y(\text{Luis})$=180, $\mu_A(\text{Luis})$=$\mu'_A(180)$=0'88, o dicho de otro modo, el *grado de pertenencia* de Luis al conjunto de personas altas es 0'88; si $f_Y(\text{Roberto})$=190, $\mu_A(\text{Roberto})$ =$\mu'_A(190)$=0'98; si $f_Y(\text{Antonio})$=170, $\mu_A(\text{Antonio})$=$\mu'_A(170)$=0'12.

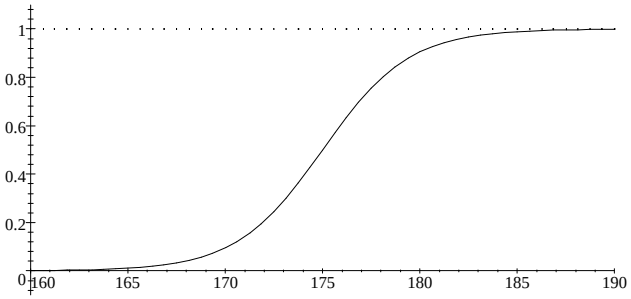

Figura 5.2: Función $\mu'_A(y)$: grado de pertenencia al conjunto A de personas altas, en función de la estatura en centímetros, y.

Observe que

$$\int_{-\infty}^{\infty} \mu'_A(y) = \int_{-\infty}^{\infty} \frac{1}{1 + e^{-0'4(x-175)}} = \infty$$

de modo que μ'_A no está normalizada en el sentido de la teoría de la probabilidad (para ello, la integral anterior debería valer 1 en vez de infinito). De hecho, en lógica difusa se utiliza un concepto diferente de normalización:

Definición 5.52 (Conjunto normalizado) Un conjunto A está normalizado si y sólo si

$$\max_{x \in X} \mu_A(x) = 1 \tag{5.124}$$

En estos dos ejemplos hemos construido $\mu_A(x)$ y $\mu'_A(y)$ escogiendo dos de las funciones que más se utilizan para este fin (la primera, en forma de campana y, la segunda, una sigmoide), ajustando los parámetros "a ojo", es decir, por el método de ensayo y error, hasta conseguir

que el resultado se ajustara a nuestra estimación intuitiva de μ. Naturalmente, existen métodos más rigurosos para la construcción de funciones características, tanto con la ayuda de uno o varios expertos como a partir de una base de datos, aunque en la mayor parte de las aplicaciones las funciones suelen construirse mediante estimaciones subjetivas de un experto humano, que en muchos casos es el propio diseñador del sistema. Por eso no vamos a entrar en la descripción de tales métodos, la mayoría de los cuales tienen más un interés académico que real; el lector interesado puede encontrarlos en [35, cap. 11] y [60, sec. 5.1].

5.3.3 Igualdad de conjuntos

La igualdad de dos conjuntos se define mediante un criterio que indique si son iguales o no, es decir, que nos indique el grado de verdad de la proposición $A = B$.

Definición 5.53 (Igualdad de A y B: $A = B$) Dados dos conjuntos A Y B,

$$A = B \equiv P_A \leftrightarrow P_B \equiv (\forall x,\ x \in A \leftrightarrow x \in B) \tag{5.125}$$

donde "$\leftrightarrow$" es una doble-implicación amplia.[18]

De esta definición se deduce que

$$\begin{aligned} v(A = B) = v(\forall x,\ x \in A \leftrightarrow x \in B) &= \underset{x \in X}{f_\wedge}\, v\,(x \in A \leftrightarrow x \in B) \\ &= \underset{x \in X}{f_\wedge}\, f_\leftrightarrow(v(x \in A), v(x \in B)) \qquad (5.126) \\ &= \underset{x \in X}{f_\wedge}\, f_\leftrightarrow(\mu_A(x), \mu_B(x)) \qquad (5.127) \end{aligned}$$

Aplicando la proposición 5.35 a las dos últimas igualdades se obtienen respectivamente las siguientes proposiciones:

Proposición 5.54 Dos conjuntos son estrictamente iguales si y sólo si cada x de X pertenece con el mismo grado a A y a B:

$$A = B \Longleftrightarrow [\forall x,\ v(x \in A) = v(x \in B)] \tag{5.128}$$

Proposición 5.55 Dos conjuntos son estrictamente iguales si y sólo si sus funciones características son iguales:

$$A = B \Longleftrightarrow [\forall x,\ \mu_A(x) = \mu_B(x)] \tag{5.129}$$

Ejemplo 5.56 Para el conjunto universal X_{500}, los conjuntos A y B definidos por las siguientes funciones,

$$\begin{aligned} \mu_A(x) &= x \bmod 2 \\ \mu_B(x) &= \begin{cases} 0 & \text{si } x = 0 \\ 1 - \mu_B(x-1) & \text{si } x > 0 \end{cases} \end{aligned}$$

son el mismo subconjunto ($A = B$), porque ambas funciones asignan el mismo valor a cada x. Por cierto, se trata del subconjunto nítido de los números impares. □

[18] En el ejemplo 5.59 veremos que el motivo de exigir una doble-implicación amplia (en vez de permitir una rigurosa) es mantener la propiedad reflexiva de la igualdad.

Nótese que la propiedad (5.129) se deduce de la definición (5.125), pero no a la inversa, porque la propiedad (5.129) no dice cuál es el valor de verdad de $A = B$ cuando $\mu_A(x) \neq \mu_B(x)$ para algún x. En cambio, la definición (5.125) da lugar a la ecuación (5.127), que permite calcular $v(A = B)$ en todos los casos.

En la *teoría de conjuntos difusos estándar* (a veces llamada impropiamente *lógica difusa estándar*), la igualdad entre conjuntos $A = B$ se define como una propiedad precisa, de modo que dos conjuntos son iguales si y sólo si sus funciones características son exactamente iguales

$$v(A = B) = \begin{cases} 1 & \text{si } \forall x,\ \mu_A(x) = \mu_B(x) \\ 0 & \text{en los demás casos} \end{cases} \tag{5.130}$$

(Observe la semejanza de esta definición con la ec. (5.129)). Por tanto, la proposición $A = B$ sólo puede ser totalmente cierta o totalmente falsa, sin posibilidad de grados intermedios. En cambio, nuestra definición hace que $A = B$ sea una proposición imprecisa para conjuntos difusos. Volveremos sobre este punto en el ejemplo 5.59.

Del mismo modo que dos funciones $\mu : [0, 1]$ iguales definen el mismo subconjunto, dos predicados equivalentes definen también el mismo subconjunto:

Proposición 5.57 Dados dos predicados P y Q, y dos conjuntos $A = \{x \in X \mid P(x)\}$ y $B = \{x \in X \mid Q(x)\}$,

$$P \equiv Q \Longrightarrow A = B \tag{5.131}$$

Demostración.

$$P \equiv Q \Longleftrightarrow \forall v, \forall x,\ v(P(x)) = v(Q(x)) \Longleftrightarrow \forall v, \forall x,\ v(x \in A) = v(x \in B) \Longleftrightarrow \forall v,\ A = B$$

□

La propia demostración nos dice que la proposición inversa no es cierta, porque la verdad de $A = B$ depende de cada v concreto, mientras que la equivalencia de predicados es independiente de v (cf. definición. (5.99)). En cambio, la "equivalencia" $P \leftrightarrow Q$ depende de v,[19] y ello nos permite enunciar la siguiente

Proposición 5.58 Dados dos predicados P y Q, y dos conjuntos $A = \{x \in X \mid P(x)\}$ y $B = \{x \in X \mid Q(x)\}$,

$$P \leftrightarrow Q \equiv A = B \tag{5.132}$$

Demostración. Por la ecuación (5.101),

$$v(P \leftrightarrow Q) = v(\forall x,\ P(x) \leftrightarrow Q(x)) = v(\forall x,\ x \in A \leftrightarrow x \in B) = v(A = B)$$

□

La diferencia entre la definición 5.125 y esta proposición es que los predicados que aparecen en la definición son "Pertenece a A" y "Pertenece a B", mientras que los que aparecen en la proposición son dos predicados genéricos, como "Par" y "Múltiplo de 2". Lo que nos dice esta proposición es que predicados equivalentes definen el mismo conjunto. En este ejemplo, la

[19] Recordemos que otra diferencia importante es que $P \equiv Q$ es siempre una proposición precisa, aunque los predicados sean imprecisos, mientras que la proposición $P \leftrightarrow Q$ es imprecisa.

equivalencia entre los predicados "Par" y "Múltiplo de 2" hace que el conjunto de los números pares sea el mismo que el de los múltiplos de 2.

Las propiedades básicas de la igualdad clásica vienen dadas por la tabla 5.16; naturalmente, estas propiedades se pueden demostrar tanto a partir de la ecuación (5.125) como a partir de la 5.129.

Reflexiva	$A = A$
Simétrica	$A = B \Longrightarrow B = A$
Transitiva	$(A = B \wedge B = C) \Longrightarrow A = C$

Tabla 5.16: Propiedades de la igualdad de conjuntos.

Igualdad de conjuntos nítidos

La lógica clásica se basa en una doble-implicación amplia (cf. pág. 103); por tanto, en la teoría de conjuntos clásica dos conjuntos son iguales si sólo si sus funciones características son iguales. De hecho, para los conjuntos clásicos, $A = B$ es una proposición precisa. Cuando $A = B$, sólo existen dos posibilidades para cada x:

$$A = B \Longleftrightarrow \forall x,\ [\mu_A(x) = \mu_B(x) = 0 \vee \mu_A(x) = \mu_B(x) = 1]$$

Igualdad de conjuntos difusos

Como hemos dicho ya, los conjuntos difusos se apoyan en la lógica difusa, del mismo modo que los conjuntos nítidos se apoyan en la lógica clásica. Sin embargo, no existe una única lógica difusa, sino distintas modalidades, que vienen dadas por las diferentes posibilidades de elegir las funciones de negación, conjunción, disyunción, implicación y doble-implicación.

Ejemplo 5.59 Sea $X = \{x_1, x_2, x_3, x_4\}$ y los tres conjuntos

$$\begin{aligned} A &= 0'2|x_1 + 0'3|x_2 + 0'7|x_3 + 0'96|x_4 \\ B &= 0'2|x_1 + 0'4|x_2 + 0'7|x_3 + 0'95|x_4 \\ C &= 0'9|x_1 + 0'8|x_2 + 0'1|x_3 + 0'05|x_4 \end{aligned}$$

Se observa inmediatamente que, con la definición de igualdad estándar, $v(A{=}B) = v(A{=}C) = 0$, a pesar de que A y B son casi iguales. □

Sin embargo, esto va en contra del espíritu de la lógica difusa, que se enorgullece de admitir grados de verdad (distintos tonos de gris, como se suele decir) donde la lógica clásica sólo ve blanco o negro.

Por eso nos parece más adecuado tomar como definición de igualdad la expresión 5.125, la cual permite que el valor de verdad de $A = B$ pueda ser, en principio, cualquier número entre 0 y 1. Una peculiaridad de tal definición es que este valor depende de la elección de $f_\wedge$ y $f_\leftrightarrow$.

Así, en el ejemplo anterior, si tomamos la norma estándar y la doble-implicación de Łukasiewicz,

$$\begin{aligned} v(A = A) &= \min_x (1 - |\mu_A(x) - \mu_A(x)|) = 1 \\ v(A = B) &= \min_x (1 - |\mu_A(x) - \mu_B(x)|) = 0'9 \\ v(A = C) &= \min_x (1 - |\mu_A(x) - \mu_B(x)|) = 0'09 \end{aligned}$$

mientras que, con la norma del producto y la doble-implicación de Łukasiewicz,

$$\begin{aligned} v(A = A) &= \mathrm{prod}_x (1 - |\mu_A(x) - \mu_A(x)|) = 1 \\ v(A = B) &= \mathrm{prod}_x (1 - |\mu_A(x) - \mu_B(x)|) = 0'891 \\ v(A = C) &= \mathrm{prod}_x (1 - |\mu_A(x) - \mu_B(x)|) = 0'0054 \end{aligned}$$

Por cierto, si tomamos la norma estándar y la doble-implicación de Kleene

$$v(A = A) = \min_x \{\min[\max(1 - \mu_A(x), \mu_A(x)), \max(1 - \mu_A(x), \mu_A(x))]\} = 0'6 < 1$$

Ésta es la razón por la que en la definición de igualdad de conjuntos se exige una doble-implicación amplia, con el fin de que se cumpla la propiedad reflexiva, $v(A = A) = 1$, también para conjuntos difusos.

5.3.4 Inclusión de conjuntos

La inclusión de un conjunto en otro se define de modo análogo a la igualdad: dando un criterio para hallar el grado de verdad de la proposición $A \subseteq B$.

Definición 5.60 (Inclusión de A en B: $A \subseteq B$) Dados dos conjuntos A y B,

$$A \subseteq B \equiv P_A \to P_B \equiv [\forall x,\ x \in A \to x \in B] \tag{5.133}$$

donde "$\to$" es una implicación amplia.

De esta definición se deduce que

$$\begin{aligned} v(A \subseteq B) = v(\forall x,\ x \in A \to x \in B) &= v\left(\bigwedge_{x \in X} f_{\to}(\mu_A(x), \mu_B(x))\right) \\ &= \underset{x \in X}{f_{\wedge}} f_{\to}(v(x \in A), v(x \in B)) \qquad (5.134) \\ &= \underset{x \in X}{f_{\wedge}} f_{\to}(\mu_A(x), \mu_B(x)) \qquad (5.135) \end{aligned}$$

Por analogía con la sección anterior, podemos enunciar las siguientes proposiciones (omitimos las demostraciones, porque son muy similares):

Proposición 5.61 El conjunto A está estrictamente incluido en B si y sólo si cada x de X pertenece con a A con igual o menor grado que a B:

$$A \subseteq B \iff [\forall x,\ v(x \in A) \leq v(x \in B)] \tag{5.136}$$

Proposición 5.62 El conjunto A está estrictamente incluido en B si y sólo si la función característica del primero es menor que la del segundo para todo x:

$$A \subseteq B \Longleftrightarrow [\forall x,\ \mu_A(x) \leq \mu_B(x)] \tag{5.137}$$

Proposición 5.63 Dados dos predicados P y Q, y dos conjuntos $A = \{x \in X \mid P(x)\}$ y $B = \{x \in X \mid Q(x)\}$,

$$P \rightarrow Q \Longleftrightarrow A \subseteq B \tag{5.138}$$

La diferencia entre la definición 5.60 (ec. (5.133)) y esta proposición es que los predicados que aparecen en la definición son "Pertenece a A" y "Pertenece a B", mientras que los que aparecen en la proposición son dos predicados genéricos.

Ejemplo 5.64 Todos los enteros no negativos múltiplos de cuatro son divisibles por 2. De la implicación "Múltiplo de 4"→"Par" se deduce la inclusión $\{x \in X \mid \text{Múltiplo-de-4}(x)\} \subseteq \{x \in X \mid \text{Par}(x)\}$,.tal como afirma esta última proposición. En efecto, $\{0, 4, 8, 12, \ldots\} \subseteq \{0, 2, 4, 6, 8, \ldots\}$. □

Como en la sección anterior, debemos señalar que la propiedad (5.137) se deduce de la definición (5.133), pero no a la inversa.

En la *teoría de conjuntos difusos estándar*, la inclusión de un conjunto en otro, $A \subseteq B$, se define como una propiedad precisa, de modo que dos conjuntos son iguales si y sólo si sus funciones características son exactamente iguales

$$v(A = B) = \begin{cases} 1 & \text{si } \forall x,\ \mu_A(x) \leq \mu_B(x) \\ 0 & \text{en los demás casos} \end{cases} \tag{5.139}$$

Por tanto, la proposición $A \subseteq B$ sólo puede ser totalmente cierta o totalmente falsa, sin posibilidad de grados intermedios. En cambio, la definición (5.133) hace que $A \subseteq B$ sea una proposición imprecisa cuando los conjuntos son difusos.

Las propiedades principales de la inclusión se recogen en la tabla 5.17.

Reflexiva	$A \subseteq A$
Antisimétrica	$(A \subseteq B \wedge B \subseteq A) \Longrightarrow A = B$
Transitiva	$(A \subseteq B \wedge B \subseteq C) \Longrightarrow A \subseteq C$

Tabla 5.17: Propiedades de la inclusión entre conjuntos.

Inclusión para conjuntos nítidos

En la lógica clásica, $A \subseteq B$ es siempre una proposición precisa. Cuando $A \subseteq B$ existen tres posibilidades para cada x:

$$\left\{ \begin{array}{l} \mu_A(x) = 0,\ \mu_B(x) = 0 \\ \mu_A(x) = 0,\ \mu_B(x) = 1 \\ \mu_A(x) = 1,\ \mu_B(x) = 1 \end{array} \right\} \tag{5.140}$$

y se excluye la posibilidad de que $\{\mu_A(x) = 1,\ \mu_B(x) = 0\}$, porque no cumple la condición $\mu_A(x) \leq \mu_B(x)$.

Inclusión para conjuntos difusos

En la lógica difusa estándar la inclusión entre conjuntos se define mediante la ecuación (5.139), lo cual conlleva el inconveniente de que la afirmación $A \subseteq B$ sólo puede ser totalmente cierta o totalmente falsa. Para los conjuntos del ejemplo 5.59, la definición estándar de inclusión conduce a $v(A \subseteq B) = v(C \subseteq B) = 0$, a pesar de que A casi está incluido en B, pues bastaría que $\mu_A(x_4)$ valiera 0'95 en vez de 0'96 para que $v(A \subseteq B) = 1$; incluso cuando $\mu_A(x_4) = 0'95$, parece evidente que $v(A \subseteq B)$ debería ser mayor que $v(C \subseteq B)$.

Por eso resulta más razonable tomar como definición de igualdad la expresión 5.133, la cual permite que el valor de verdad de $A \subseteq B$ pueda ser, en principio, cualquier número entre 0 y 1. Naturalmente, el valor de $v()$ de la elección de $f_\wedge$ y $f_\rightarrow$.

Ejemplo 5.65 Para los conjuntos del ejemplo 5.59, tomando la conjunción estándar y la implicación de Łukasiewicz,

$$\begin{aligned}
v(A \subseteq A) &= \min_x (\min(1, 1 - \mu_A(x) + \mu_A(x))) = 1 \\
v(A \subseteq B) &= \min_x (\min(1, 1 - \mu_A(x) + \mu_B(x))) = 0'99 \\
v(B \subseteq A) &= \min_x (\min(1, 1 - \mu_B(x) + \mu_A(x))) = 0'9 \\
v(A \subseteq C) &= \min_x (\min(1, 1 - \mu_A(x) + \mu_C(x))) = 0'09 \\
v(C \subseteq A) &= \min_x (\min(1, 1 - \mu_A(x) + \mu_C(x))) = 0'3
\end{aligned}$$

Si hubiéramos aplicado la implicación de Kleene tendríamos, según la ecuación (5.30), que $v(A \subseteq A) = \min_x \max(1 - a, a) = 0'7$, mientras que lo deseable sería que sería que $v(A \subseteq A) = 1$ para todo conjunto A. Por esta razón se exige en la definición de inclusión de conjuntos una implicación amplia, con el fin de preservar la propiedad reflexiva.

5.3.5 Composición de conjuntos: complementario, unión e intersección

Teniendo en cuenta la proposición 5.46, podemos definir el conjunto $\bar{A}$ (**complementario** de A) de cuatro modos equivalentes:

- $\bar{A} = \{x \in X \mid \neg(x \in A)\}$ (5.141)
- $x \in \bar{A} \equiv \neg(x \in A)$ (5.142)
- $P_{\bar{A}} = \neg P_A$ (5.143)
- $\mu_{\bar{A}}(x) = f_\neg(\mu_A(x))$ (5.144)

El conjunto $A \cap B$ (**intersección** de A y B) puede definirse también de cuatro modos equivalentes:

- $A \cap B = \{x \in X \mid x \in A \wedge x \in B\}$ (5.145)
- $x \in A \cap B \Longleftrightarrow x \in A \wedge x \in B$ (5.146)
- $P_{A \cap B} = P_A \wedge P_B$ (5.147)
- $\mu_{A \cap B}(x) = f_\wedge(\mu_A(x), \mu_B(x))$ (5.148)

y análogamente $A \cup B$ (**unión** de A y B):

- $A \cup B = \{x \in X \mid x \in A \vee x \in B\}$ (5.149)
- $x \in A \cup B \Longleftrightarrow x \in A \vee x \in B$ (5.150)
- $P_{A \cup B} = P_A \vee P_B$ (5.151)
- $\mu_{A \cup B}(x) = f_{\vee}(\mu_A(x), \mu_B(x))$ (5.152)

La diferencia de conjuntos se define a partir de las operaciones anteriores:

$$A \setminus B = A \cap \bar{B} \tag{5.153}$$

Composición de conjuntos clásicos

En el caso clásico, los valores de $\mu_{\bar{A}}(x)$, $\mu_{A \cap B}(x)$ y $\mu_{A \cup B}(x)$ en función de $\mu_A(x)$ y $\mu_B(x)$ vienen dados por la tabla 5.18, que se deduce inmediatamente de la tabla 5.3 aplicando las ecuaciones (5.144), (5.148) y (5.152).

$\mu_A(x)$	$\mu_B(x)$	$\mu_{\bar{A}}(x)$	$\mu_{A \cap B}(x)$	$\mu_{A \cup B}(x)$
1	1	0	1	1
1	0	0	0	1
0	1	1	0	1
0	0	1	0	0

Tabla 5.18: Complementario, unión e intersección de conjuntos clásicos.

Ejemplo 5.66 Sea de nuevo el conjunto universal X_{500} y $A = \{x \mid x < 10\}$, de modo que

$$\mu_A(x) = \begin{cases} 1 & \text{si } x < 10 \\ 0 & \text{si } x \geq 10 \end{cases}$$

La función característica de $\bar{A}$ puede obtenerse a partir de la ecuación (5.144) y la función $f_{\neg}$ definida en la tabla 5.3:

$$\mu_{\bar{A}}(x) = \begin{cases} 0 & \text{si } x < 10 \\ 1 & \text{si } x \geq 10 \end{cases}$$

También se puede calcular $\bar{A}$ por la ecuación (5.5):

$$\bar{A} = \{x \mid \neg(x < 10)\} = \{x \mid x \geq 10\}$$

Si $B = \{x \in X \mid x \text{ es impar}\}$, su función característica es $\mu_B(x) = (x + 1) \bmod 2$. Utilizando las mismas ecuaciones que para $\bar{A}$,

$$\mu_{\bar{B}}(x) = 1 - ((x + 1) \bmod 2) = x \bmod 2$$

$$\bar{B} = \{x \mid x \text{ no es impar}\}$$

La intersección y la unión de estos conjuntos vienen dadas por

$$\mu_{A \cap B}(x) = \begin{cases} 1 & \text{si 1, 3, 5, 7 y 9} \\ 0 & \text{si 0, 2, 4, 6, 8, 10, 11, 12, 13 } \ldots \end{cases}$$

$$\mu_{A\cup B}(x) = \begin{cases} 1 & \text{si} \quad 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 13, 15, 17 \ldots \\ 0 & \text{si} \quad 0{,}10, 12, 14, 16, 18 \ldots \end{cases}$$

□

El complementario, la intersección y la unión de conjuntos clásicos (nítidos) cumplen las propiedades que aparecen en la tabla 5.19. Estas propiedades hacen que la teoría clásica de conjuntos constituya un *álgebra de Boole*.[20] Cada una de estas propiedades puede demostrarse al menos de tres modos:

- por las definiciones (5.141), (5.145) y (5.149),
- por las propiedades de los predicados (tabla 5.4),
- por las propiedades de las funciones características (tabla 5.3).

Por ejemplo, la propiedad del elemento neutro de la intersección puede demostrarse así:

- por la ecuación (5.145)

$$A \cap X = \{x \in X \mid x \in A \wedge x \in X\} = \{x \in X \mid x \in A\} = A$$

- dado que los predicados que definen A, X y $A \cap X$ son, respectivamente P_A, $P_X \equiv P^{\mathbf{1}}$ y $P_{A\cap X}$, aplicando la ecuación (5.147) y la propiedad de predicados $P_A \wedge P^{\mathbf{1}}{=}P_A$ (que se deduce de la propiedad de proposiciones $p \wedge \mathbf{1} {=} p$), se demuestra que $P_{A\cap X} = P_A$;
- de la ecuación (5.148) se deduce que $f_\wedge(a, 1) = a$ tanto si a vale 1 como si vale 0; por tanto, teniendo en cuenta que $\mu_X(x) = v(x \in X) = 1$, obtenemos que $\mu_{A\cap X}(x) = f_\wedge(\mu_A(x), \mu_X(x)) = f_\wedge(\mu_A(x), 1) = \mu_A(x)$.

Y así pueden demostrarse una por una todas las propiedades.

Composición de conjuntos difusos

Ejemplo 5.67 Dados los conjuntos del ejemplo 5.59, aplicando las ecuaciones (5.144), (5.148) y (5.152) y las funciones de la lógica difusa estándar, obtenemos que para x_1,

$$\begin{aligned} \mu_{\bar{A}}(x_1) &= f_\neg(\mu_A(x_1)) = 1 - 0'2 = 0'8 \\ \mu_{B\cap C}(x_1) &= f_\wedge(\mu_B(x_1), \mu_C(x_1)) = \min(0'2, 0'9) = 0'2 \\ \mu_{B\cup C}(x_1) &= f_\vee(\mu_B(x_1), \mu_C(x_1)) = \max(0'2, 0'9) = 0'9 \end{aligned}$$

Haciendo los mismos cálculos para cada elemento de X llegamos a:

$$\begin{aligned} \bar{A} &= 0'8|x_1 + 0'7|x_2 + 0'3|x_3 + 0'04|x_4 \\ B \cap C &= 0'2|x_1 + 0'4|x_2 + 0'1|x_3 + 0'05|x_4 \\ B \cup C &= 0'9|x_1 + 0'8|x_2 + 0'7|x_3 + 0'95|x_4 \end{aligned}$$

Ejercicio 5.68 Con los mismos conjuntos del ejemplo 5.59, calcular $\bar{C}$, $A \cap C$ y $A \cup C$.

[20] Las propiedades de monotonía no forman parte de la definición de álgebra de Boole, pero nos interesa incluirlas todas en la misma tabla.

Complementario	Involución	$\overline{\overline{A}} = A$
	Complementario de X	$\bar{X} = \varnothing$
	Complementario de $\varnothing$	$\varnothing = X$
	Monotonía	$A \subseteq B \Longrightarrow \bar{B} \subseteq \bar{A}$
Intersección	Conmutativa	$A \cap B = B \cap A$
	Asociativa	$A \cap (B \cap C) = (A \cap B) \cap C$
	Elemento neutro	$A \cap X = A$
	Elemento absorbente	$A \cap \varnothing = \varnothing$
	Idempotencia	$A \cap A = A$
	Ley de contradicción	$A \cap \bar{A} = \varnothing$
	Monotonía	$A \subseteq B \Longrightarrow A \cap C \subseteq B \cap C$
Unión	Conmutativa	$A \cup B = B \cup A$
	Asociativa	$A \cup (B \cup C) = (A \cup B) \cup C$
	Elemento neutro	$A \cup \varnothing = A$
	Elemento absorbente	$A \cup X = X$
	Idempotencia	$A \cup A = A$
	Tercio excluso	$A \cup \bar{A} = X$
	Monotonía	$A \subseteq B \Longrightarrow A \cup C \subseteq B \cup C$
Propiedades combinadas	Distributiva de la intersección	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
	Distributiva de la unión	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
	1ª ley de Morgan	$\overline{A \cap B} = \bar{A} \cup \bar{B}$
	2ª ley de Morgan	$\overline{A \cup B} = \bar{A} \cap \bar{B}$
	Absorción de la intersección	$A \cup (A \cap B) = A$
	Absorción de la unión	$A \cap (A \cup B) = A$

Tabla 5.19: Propiedades de la teoría de conjuntos clásica.

5.3.6 Recapitulación

En las presentaciones de la lógica difusa es habitual definir el complementario, la intersección y la unión de conjuntos a partir de las funciones complementario (c), intersección (i) y unión (u) que intervienen en las siguientes ecuaciones:

$$\begin{aligned} \mu_{\bar{A}}(x) &= c(\mu_A(x)) \\ \mu_{A\cap B}(x) &= i(\mu_A(x), \mu_B(x)) \\ \mu_{A\cup B}(x) &= u(\mu_A(x), \mu_B(x)) \end{aligned}$$

para estudiar luego las propiedades que deben cumplir c, i y u. Nosotros, en cambio, hemos intentado mostrar que, si partimos de una lógica (es decir, de una definición de $f_\neg$, $f_\wedge$ y $f_\wedge$), la función complementario coincide con la función negación ($c = f_\neg$), la intersección con la conjunción ($i = f_\wedge$) y la unión con la disyunción ($u = f_\wedge$), estableciendo así el isomorfismo entre la lógica y la teoría de conjuntos. Esto nos ha permitido explicar por qué, si la lógica clásica forma un álgebra de Boole, también la teoría de conjuntos clásica ha de ser un álgebra

de Boole, pues cada propiedad de la primera da lugar a una propiedad de la segunda, y viceversa.

De aquí se deduce también, a partir de los corolarios 5.13 y 5.23, que no es posible construir una teoría de conjuntos difusos que cumpla todas las propiedades de la teoría de conjuntos clásica.

Por otro lado, el hecho de definir la igualdad de conjuntos a partir de la doble-equivalencia (mejor dicho, de cualquier doble-equivalencia amplia) y la inclusión a partir de cualquier implicación amplia nos ha llevado a que $A = B$ y $A \subseteq B$ sean proposiciones imprecisas cuando A y B son conjuntos difusos, en contra de lo que es habitual en la teoría de conjuntos difusos estándar.

5.4 Relaciones e inferencia

En la sección 5.2 hemos estudiado los predicados unitarios, que son aquéllos que asignan a cada elemento x de X una proposición. Ahora bien, puede darse el caso de que el conjunto X venga dado por el producto cartesiano de varios conjuntos: $\hat{X} = X_1 \times \ldots \times X_n$,[21] de modo que cada elemento x de X será una n-tupla: $x = (x_1, \ldots, x_n)$. Además de las propiedades que hemos visto para los predicados unitarios y para los subconjuntos (complementario, unión, intersección, etc., las cuales siguen siendo válidas en este caso), existen otras propiedades específicas, que vamos a estudiar en esta sección.

5.4.1 Predicados n-arios y relaciones

Cuando un predicado P está definido sobre un conjunto $\hat{X} = X_1 \times \ldots \times X_n$, se dice que es un predicado n-ario. En este caso, $P(\hat{x}) = P(x_1, \ldots, x_n)$. Cuando n=1, tenemos predicados unitarios, que son los que hemos estudiado en la sección 5.2.1; para n=2, predicados binarios; para n=3, predicados ternarios, etc. Al igual que ocurría con los predicados unitarios, todo predicado n-ario puede ser preciso o difuso.

Ejemplo 5.69 Dado el conjunto universal $\hat{X} = \mathbb{R}^2$, "Mayor que" es un predicado binario preciso, porque dado un par de números reales (a, b), la proposición "Mayor-que(a, b)"= "a es mayor que b" es siempre verdadera o falsa. En cambio, el predicado binario "Mucho mayor que" es impreciso, pues no se puede definir exactamente para cuáles de los pares (a, b) la proposición "Mucho-mayor-que(a, b)" es totalmente cierta y para cuáles es totalmente falsa.

Ejemplo 5.70 Si C es un conjunto de ciudades, "Cerca-de" es un predicado binario impreciso definido sobre $\hat{X} = C^2$. Como ocurría en la sección 5.3.2 —cf. (ec. (5.122)— el valor de verdad de la proposición "Cerca-de(a, b)" puede venir dado por una función de la distancia entre ambas ciudades:

$$v(\text{Cerca-de}(a, b)) = \mu'_{\text{Cerca-de}}(\text{distancia}(a, b))$$

Del mismo modo, el predicado ternario "Entre", tal que "Entre(a, b, c)"="a está entre b y c", podría venir dado por una función del ángulo que forman sus argumentos:

$$v(\text{Entre}(a, b, c)) = \mu'_{\text{Entre}}(\angle bac)$$

[21] El acento circunflejo sobre $\hat{X}$ indica solamente que este conjunto es el resultado de un producto cartesiano. Nótese que en la definición de este producto no se exige que los conjuntos sean distintos entre sí. Por ejemplo, podríamos tener $\hat{X} = X \times X$ o $\hat{X} = X \times Y \times Z \times Y$.

Así tendríamos que v(Entre(Zaragoza, Madrid, Barcelona))=1 porque el ángulo Madrid-Zaragoza-Barcelona vale aproximadamente 180º, mientras que v(Entre(Sevilla, Madrid, Barcelona))=0 porque el ángulo que forman es muy pequeño.

Del mismo modo que un predicado unitario aplicado a los elementos de X define un subconjunto de X, todo **predicado n-ario** P aplicado a los elementos de $\hat{X} = X_1 \times \ldots \times X_n$ define un subconjunto R, $R \subseteq \hat{X}$, que en este caso recibe el nombre de **relación n-aria**:[22]

$$R = \{\hat{x} \in \hat{X} \mid P(\hat{x})\} \Longleftrightarrow \hat{x} \in R \equiv P(\hat{x}) \tag{5.154}$$

Por eso hablaremos indistintamente del *predicado* "Cerca de" o de la *relación* "Cerca de", según nos convenga en cada caso.

Por otro lado, teniendo en cuenta que R es un subconjunto de $\hat{X}$, a cada elemento $\hat{x} = (x_1, \ldots, x_n)$ de $\hat{X}$ podemos asignarle un grado de pertenencia a R, de acuerdo con la ecuación (5.115):

$$\mu_R(\hat{x}) = v(\hat{x} \in R) = v(P(\hat{x})) \tag{5.155}$$

Naturalmente, los **predicados precisos** definen **relaciones clásicas** (también llamadas **relaciones nítidas**), que son aquéllas en las $\mu_R(\hat{x})$ vale necesariamente 0 o 1, mientras que los **predicados imprecisos** definen **relaciones difusas**, en las que $\mu_R(\hat{x})$ puede tomar cualquier valor del intervalo $[0, 1]$.

Ejemplo 5.71 La relación R definida por el predicado "Mayor que":

$$R = \{(a, b) \in \mathbb{R}^2 \mid a > b\} \Longleftrightarrow (a, b) \in R \equiv a > b$$

es una relación nítida, pues de acuerdo con la ecuación (5.154) el valor de verdad de la proposición "$(a, b) \in R$", que es el mismo que el de la proposición "$a > b$", siempre va a ser 0 o 1. Es decir, todo par (a, b) o pertenece completamente a R —como es el caso del par $(4, 3)$— o no pertenece en absoluto a R, como ocurre con los pares $(2, 5)$ y $(7, 7)$. □

Relación de identidad

Dado un conjunto X, la **relación de identidad** I_X es una relación binaria $I_X \subseteq X \times X$, que se define así:

$$I_X = \{(x, x) \mid x \in X\} \tag{5.156}$$

o, lo que es lo mismo,

$$\mu_{I_X}(x, x') = \begin{cases} 1 & \text{si } x = x' \\ 0 & \text{si } x \neq x' \end{cases} \tag{5.157}$$

[22] Nótese que esta ecuación no es más que un caso particular de la ec. (5.107). Cuando n=1, tenemos un subconjunto (como los que hemos estudiado en la sec. 5.3); por eso no se habla nunca de relaciones unitarias, sino de subconjuntos, y cuando decimos "relaciones n-arias" suponemos generalmente que $n \geq 2$.

Relación recíproca de una relación binaria

Dada una relación binaria $R \subseteq X \times Y$, la **relación recíproca** $R^{-1} \subseteq Y \times X$ puede definirse de tres modos equivalentes:

$$\bullet \quad R^{-1} = \{(y,x) \mid (x,y) \in R\} \tag{5.158}$$

$$\bullet \quad (y,x) \in R^{-1} \equiv (x,y) \in R \tag{5.159}$$

$$\bullet \quad \mu_{R^{-1}}(y,x) = \mu_R(x,y) \tag{5.160}$$

Es decir, y está relacionado con x (mediante R^{-1}) si y sólo si x está relacionado con y (mediante R).

Observe que en general $R \circ R^{-1} \neq I_X$ y $R^{-1} \circ R \neq I_Y$, como demuestra el siguiente contraejemplo. Sean $X = \{x^1, x^2\}$, $Y = \{y^1, y^2\}$ y $R = \{(x^1, y^1), (x^1, y^2)\}$. Se tiene entonces que

$$\begin{aligned} R^{-1} &= \{(y^1, x^1), (y^2, x^1)\} \\ I_X &= \{(x^1, x^1), (x^2, x^2)\} \\ I_Y &= \{(y^1, y^1), (y^2, y^2)\} \\ R \circ R^{-1} &= \{(x^1, x^1)\} \neq I_X \\ R^{-1} \circ R &= \{(y^1, y^1), (y^1, y^2), (y^2, y^1), (y^1, y^2)\} \neq I_Y \end{aligned}$$

Por eso nos ha parecido más adecuado llamar a R^{-1} *relación recíproca*, a pesar de que otros autores la denominan *relación inversa*.

5.4.2 Composición de relaciones

Dada una relación $(m+1)$-aria $R_1 \subseteq \hat{X} = X_1 \times \ldots \times X_m \times Y$ y una relación $(n+1)$-aria $R_2 \subseteq \hat{Z} = Y \times Z_1 \times \ldots \times Z_n$, podemos definir la relación $(m+n)$-aria $R_1 \circ R_2 \subseteq X_1 \times \ldots \times X_m \times Z_1 \times \ldots \times Z_n$ de tres modos equivalentes:

$$\bullet R_1 \circ R_2 = \{(x_1, \ldots, x_m, z_1, \ldots, z_n) \mid \exists y,\ \hat{x} \in R_1 \wedge \hat{z} \in R_2\} \tag{5.161}$$

$$\bullet (x_1, \ldots, x_m, z_1, \ldots, z_n) \in R_1 \circ R_2 \equiv \exists y,\ \hat{x} \in R_1 \wedge \hat{z} \in R_2 \tag{5.162}$$

$$\bullet \mu_{R_1 \circ R_2}(x_1, \ldots, x_m, z_1, \ldots, z_n) = \underset{y \in Y}{f_\vee} f_\wedge(\mu_{R_1}(\hat{x}), \mu_{R_2}(\hat{z})) \tag{5.163}$$

Naturalmente, si R_1 y R_2 son relaciones nítidas, $R_1 \circ R_2$ también lo será; en este caso, las funciones $f_\wedge$ y $f_\vee$ son las correspondientes a la lógica clásica. En cambio, si R_1 y R_2 son relaciones difusas, su composición nos da una nueva relación difusa, pero en este caso la relación $R_1 \circ R_2$ resultante dependerá de la norma y conorma escogidas para $f_\wedge$ y $f_\vee$. (Veremos dos ejemplos en seguida.)

Nótese que la composición de relaciones no es conmutativa. De hecho, la composición $R_2 \circ R_1$ sólo es posible si la variable Z_n (la última de R_2) es la misma que X_1 (la primera de R_1). Aun así, $R_2 \circ R_1 \subseteq Y \times Z_1 \times \ldots \times Z_{n-1} \times X_2 \times \ldots \times X_m \times Y$, por lo que generalmente no puede ser igual a $R_1 \circ R_2 \subseteq X_1 \times \ldots \times X_m \times Z_1 \times \ldots \times Z_n$.

Sin embargo, sí se cumplen las propiedades asociativa

$$(R_1 \circ R_2) \circ R_3 = R_1 \circ (R_2 \circ R_3) \tag{5.164}$$

y de elemento neutro

$$I_{X_1} \circ R = R \circ I_{Z_n} = R \tag{5.165}$$

Se comprueba también fácilmente que

$$(R_1 \circ R_2)^{-1} = R_2^{-1} \circ R_1^{-1} \tag{5.166}$$

Hay dos casos que nos interesan especialmente por su relación con la inferencia: la composición de un subconjunto (relación unitaria) con una relación binaria, y la composición de dos relaciones binarias. Los estudiamos a continuación.

Composición de dos relaciones binarias

Dadas dos relaciones binarias $R_{XY} \subseteq X \times Y$ y $R_{YZ} \subseteq Y \times Z$, la relación compuesta $R_{XZ} = R_{XY} \circ R_{YZ} \subseteq X \times Z$ viene dada por

$$R_{XZ} = R_{XY} \circ R_{YZ} = \{(x,z) \mid \exists y,\ (x,y) \in R_{XY} \wedge (y,z) \in R_{YZ}\} \tag{5.167}$$

o, lo que es lo mismo,

$$\mu_{R_{XZ}}(x,z) = \mu_{R_{XZ} \circ R_{YZ}}(x,z) = \underset{y \in Y}{f_\vee}\, f_\wedge(\mu_{R_{XY}}(x,y), \mu_{R_{YZ}}(y,z)) \tag{5.168}$$

La primera de estas dos ecuaciones (que no son más una reescritura de la (5.161) y la (5.163), respectivamente), nos dice que el valor x está relacionado con z si y sólo si existe un y que sirve de "puente" entre ambos. La segunda nos dice lo mismo pero de otra forma: dado un x y un z, examinamos todos los "caminos" x–y–z (uno para cada y) y nos quedamos con el que establece la relación más fuerte entre x y z. Por eso la primera ecuación es más adecuada para relaciones nítidas, mientras que la segunda es más adecuada para relaciones difusas.

Ejemplo 5.72 Sea X el conjunto de los continentes, Y el de los países y Z el de los idiomas. Tenemos que $(x,y) \in R_{XY}$ si y sólo si el país y tiene (al menos parte de) su territorio en el continente x; por ejemplo (Europa, España) $\in R_{XY}$ y (África, España) $\in R_{XY}$, mientras que (Asia, España) $\notin R_{XY}$. Del mismo modo, $(y,z) \in R_{YZ}$ si y sólo si el idioma z es oficial en el país y; así, (España, gallego) $\in R_{YZ}$, mientras que (Italia, esperanto) $\notin R_{YZ}$. La ecuación (5.167) nos dice que $(x,z) \in R_{XZ}$ —es decir, que el idioma z es oficial en el continente x— si y sólo si existe un país y del continente x que tiene z como idioma oficial. □

Ejemplo 5.73 Sean los conjuntos $X = \{x^1, x^2, x^3, x^4\}$, $Y = \{y^1, y^2\}$ y $Z = \{z^1, z^2, z^3\}$, y las relaciones R_{XY} y R_{YZ} dadas por las siguientes matrices (el elemento de la i-ésima fila y la j-ésima columna de la matriz de R_{XY} es el valor de $\mu_R(x^i, y^j)$):

$$R_{XY} = \begin{pmatrix} 0'3 & 0'5 \\ 0 & 0'7 \\ 0'2 & 0'4 \\ 0'6 & 0'3 \end{pmatrix} \qquad R_{YZ} = \begin{pmatrix} 1 & 0'9 & 0'7 \\ 0 & 0'2 & 0'4 \end{pmatrix}$$

Tomando la t-norma min para la conjunción y la conorma max para la disyunción, la ecuación (5.163) se traduce en

$$\mu_{R_{XZ}}(x,z) = \max_{y \in Y} \min(\mu_{R_{XY}}(x,y), \mu_{R_{YZ}}(y,z)) \tag{5.169}$$

Por ejemplo,

$$\mu_{R_{XZ}}(x^1, z^1) = \max[\min(\mu_{R_{XY}}(x^1, y^1), \mu_{R_{YZ}}(y^1, z^1)), \min(\mu_{R_{XY}}(x^1, y^2), \mu_{R_{YZ}}(y^2, z^1))]$$
$$= \max[\min(0'3, 1), \min(0'5, 0)] = \max(0'3, 0) = 0'3$$

(Nótese la semejanza con la multiplicación de matrices en espacios vectoriales.) Del mismo modo se calculan los demás valores, con lo que se llega a

$$R_{XZ} = R_{XY} \circ R_{YZ} = \begin{pmatrix} 0'3 & 0'5 \\ 0 & 0'7 \\ 0'2 & 0'4 \\ 0'6 & 0'3 \end{pmatrix} \circ \begin{pmatrix} 1 & 0'9 & 0'7 \\ 0 & 0'2 & 0'4 \end{pmatrix} = \begin{pmatrix} 0'3 & 0'3 & 0'4 \\ 0 & 0'2 & 0'4 \\ 0'2 & 0'2 & 0'4 \\ 0'6 & 0'6 & 0'6 \end{pmatrix}$$

Esta forma de composición se denomina max-min. En cambio, si tomamos la norma prod para la conjunción y la conorma max para la disyunción (composición max-prod), que se suele representar mediante $\odot$) tenemos

$$\mu_{R_{XZ}}(x, z) = \max_{y \in Y} \mu_{R_{XY}}(x, y) \cdot \mu_{R_{YZ}}(y, z) \tag{5.170}$$

y por tanto

$$R'_{XZ} = R_{XY} \odot R_{YZ} = \begin{pmatrix} 0'3 & 0'5 \\ 0 & 0'7 \\ 0'2 & 0'4 \\ 0'6 & 0'3 \end{pmatrix} \odot \begin{pmatrix} 1 & 0'9 & 0'7 \\ 0 & 0'2 & 0'4 \end{pmatrix} = \begin{pmatrix} 0'3 & 0'27 & 0'21 \\ 0 & 0'14 & 0'28 \\ 0'2 & 0'18 & 0'16 \\ 0'6 & 0'54 & 0'42 \end{pmatrix}$$

□

Composición de un conjunto con una relación binaria

Dado un subconjunto $A \subseteq X$, y una relación binaria $R_{XY} \subseteq X \times Y$, la composición de ambos nos da un conjunto B de Y:

$$B = A \circ R_{XY} = \{y \mid \exists x,\ x \in A \wedge (x, y) \in R_{XY}\} \tag{5.171}$$

o, lo que es lo mismo,

$$\mu_B(y) = \mu_{A \circ R_{XY}}(y) = \underset{x \in X}{f_\vee} f_\wedge(\mu_A(x), \mu_{R_{XY}}(x, y)) \tag{5.172}$$

(Estas dos ecuaciones son tan sólo una reescritura de la (5.161) y la (5.163), respectivamente.)

Por tanto, la relación R_{XY} equivale a una función f_{XY} que a cada conjunto A de X le asigna un conjunto B de Y,

$$f_{XY}(A) = A \circ R_{XY} = B \tag{5.173}$$

de modo que, partiendo de información sobre el dominio X, obtenemos información sobre el dominio Y.

Ejemplo 5.74 (Continuación del ejemplo 5.72) Recordemos que X era el conjunto de continentes, Y el de países y Z el de idiomas. Dados los conjuntos

$$A = \{\text{Oceanía}\} \subseteq X$$
$$B = \{\text{Nueva Zelanda}\} \subseteq Y$$

y las relaciones definidas en el ejemplo 5.72, tenemos

$$A \circ R_{XY} = \{\text{Australia, Nueva Zelanda, Papúa Nueva Guinea, Fiji}\} \subseteq Y$$
$$B \circ R_{YZ} = \{\text{inglés, maorí}\} \subseteq Z$$
$$A \circ R_{YZ} = \{\text{inglés, maorí, pidgin, fijano, hindi}\} \subseteq Z$$

Estos resultados nos dicen que, si sabemos que una persona es originaria de Oceanía, podemos deducir que procede de Australia, Nueva Zelanda, Papúa Nueva Guinea o Fiji, y que sus idiomas oficiales están dentro del conjunto {inglés, maorí, pidgin, fijano, hindi}.

Por tanto, las relaciones R_{XY} y R_{XZ} nos permiten obtener información sobre los países (Y) y sobre los idiomas (Z), respectivamente, a partir de información sobre los continentes (X). Comprobamos así cómo se puede utilizar una relación R_{XY} para realizar **inferencias** sobre Y a partir de información sobre X, del mismo modo que R_{YZ} permite obtener información sobre los idiomas a partir de información sobre los países. Observe también que la relación recíproca R_{XY}^{-1} nos permite obtener para cada país el continente o continentes en que tiene su territorio. Por ejemplo, $\{\text{España}\} \circ R_{XY}^{-1} = \{\text{Europa, África}\}$. Del mismo modo, $\{\text{inglés}\} \circ R_{YZ}^{-1}$ nos da el conjunto de los países que tiene el inglés como lengua oficial.

Ejemplo 5.75 (Continuación del ejemplo 5.73) Dado el conjunto difuso $A = 0'8|x^1 + 0'2|x^2 + 0'1|x^4$, tenemos que

$$A \circ R_{XY} = \begin{pmatrix} 0'8 & 0'2 & 0 & 0'1 \end{pmatrix} \circ \begin{pmatrix} 0'3 & 0'5 \\ 0 & 0'7 \\ 0'2 & 0'4 \\ 0'6 & 0'3 \end{pmatrix} = \begin{pmatrix} 0'3 & 0'5 \end{pmatrix}$$

Es decir, $f_{XY}(A) = A \circ R_{XY} = 0'3|y^1 + 0'5|y^2$. Tenemos también que

$$\begin{aligned} A \odot R'_{XZ} &= A \odot (R_{XY} \odot R_{YZ}) \\ &= \begin{pmatrix} 0'8 & 0'2 & 0 & 0'1 \end{pmatrix} \odot \begin{pmatrix} 0'3 & 0'27 & 0'21 \\ 0 & 0'14 & 0'28 \\ 0'2 & 0'18 & 0'16 \\ 0'6 & 0'36 & 0'42 \end{pmatrix} = \begin{pmatrix} 0'24 & 0'216 & 0'128 \end{pmatrix} \end{aligned}$$

de modo que $A \odot R'_{XZ} = 0'24|z^1 + 0'216|z^2 + 0'128|z^3$. Observe que $A \odot R'_{XZ} = A \odot (R_{XY} \odot R_{YZ}) = (A \odot R_{XY}) \odot R_{YZ}$. Es decir, si entendemos cada relación R como una regla de inferencia, da lo mismo componer primero R_{XY} con R_{YZ} para obtener la regla de inferencia R'_{XZ} que componer A con R_{XY} para obtener un subconjunto $A \odot R_{XY}$ que luego se combina con la regla R_{YZ}.

5.4.3 Modus ponens difuso

En la sección 5.2.2 mencionamos que nos interesaría poder realizar una inferencia como

$$\frac{\begin{array}{l}P \to Q\\ P'(x_0)\end{array}}{Q'(x_0)}$$

de modo que cuando P' esté próximo a P la conclusión Q' esté próxima a Q. Por ejemplo, dada la regla "si la curva es cerrada es peligrosa" y la afirmación "la curva es muy cerrada" nos interesaría poder deducir que "la curva es muy peligrosa", o al menos que "la curva es peligrosa". Con la misma regla y la afirmación "la curva es bastante cerrada" nos gustaría poder deducir que "la curva es bastante peligrosa" o alguna conclusión similar.

También desearíamos poder aplicar el modus ponens para el caso siguiente:

$$\frac{\begin{array}{l}P(x) \to Q(y)\\ P'(x)\end{array}}{Q'(y)}$$

donde $x \in X$ e $y \in Y$ (tanto si X e Y son conjuntos distintos como si son el mismo conjunto).

Una forma de abordar este problema consiste en considerar el isomorfismo entre predicados y conjuntos (que incluye como caso particular el isomorfismo entre predicados n-arios y relaciones). Si A, B, A' y B' son los conjuntos inducidos por P, Q, P' y Q' respectivamente, el silogismo anterior puede expresarse como[23]

$$\frac{\begin{array}{l}x \in A \to y \in B\\ x \in A'\end{array}}{y \in B'}$$

Por tanto, nuestro problema inicial se ha reducido a convertir la regla "$P(x) \to Q(y)$" (o, lo que es lo mismo, la regla "$x \in A \to y \in B$") en una relación $R_{A\to B} \subseteq X \times Y$, pues entonces podremos aplicar la composición de relaciones, de modo que

$$B' = A' \circ R_{A\to B} \tag{5.175}$$

Naturalmente, $R_{A\to B}$ ha de venir dada por A y B, y para ello suele utilizarse la implicación, de la que ya hemos hablado ampliamente en la sección 5.1:

$$(x, y) \in R_{A\to B} \equiv (x \in A) \to (y \in B) \tag{5.176}$$

lo cual equivale a

$$\mu_{A\to B}(x, y) = f_{\to}(\mu_A(x), \mu_B(y)) \tag{5.177}$$

[23]Para algunas de las propiedades que vamos a discutir más adelante, conviene exigir la condición de que el predicado P sea completamente cierto para algún x; esta condición puede expresarse —al menos— de tres formas equivalentes:

$$\exists x, P(x) = 1 \iff \exists x, x \in A \iff \max_{x\in X} \mu_A(x) = 1 \tag{5.174}$$

En lógica difusa se dice que el conjunto A está *normalizado* (cf. definición 5.52).

(Hemos escrito $\mu_{A\to B}$ en vez de $\mu_{R_{A\to B}}$ para simplificar la notación.) En consecuencia, la ecuación (5.168) se convierte en

$$\mu_{B'}(y) = \underset{x\in X}{f_\vee}\ f_\wedge(\mu_{A'}(x),\, \mu_{A\to B}(x,y)) \tag{5.178}$$

$$= \underset{x\in X}{f_\vee}\ f_\wedge(\mu_{A'}(x),\, f_\to(\mu_A(x), \mu_B(y))) \tag{5.179}$$

Ejemplo 5.76 Sean los conjuntos $A = 0|x^1 + 0'4|x^2 + 0'8|x^3 + 1|x^4$ y $B = 0'2|y^1 + 0'7|y^2 + 1|y^3$. Para cada una de las funciones de implicación que aparecen en la tabla 5.14 (Łukasiewicz, Kleene, Reichenbach, Zadeh, Gödel y Gaines-Rescher), la ecuación (5.177) da lugar a las siguientes relaciones de implicación:

$$R^L_{A\to B} = \begin{pmatrix} 1 & 0'8 & 0'4 & 0'2 \\ 1 & 1 & 0'9 & 0'7 \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad R^K_{A\to B} = \begin{pmatrix} 1 & 0'6 & 0'2 & 0'2 \\ 1 & 0'7 & 0'7 & 0'7 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

$$R^R_{A\to B} = \begin{pmatrix} 1 & 0'68 & 0'36 & 0'2 \\ 1 & 0'88 & 0'76 & 0'7 \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad R^Z_{A\to B} = \begin{pmatrix} 1 & 0'6 & 0'2 & 0'2 \\ 1 & 0'6 & 0'7 & 0'7 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

$$R^G_{A\to B} = \begin{pmatrix} 1 & 0'2 & 0'2 & 0'2 \\ 1 & 1 & 0'7 & 0'7 \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad R^{GR}_{A\to B} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Sea ahora el conjunto $A' = 0|x^1 + 0'3|x^2 + 0'6|x^3 + 1|x^4 \subset A$. El conjunto $B' = A' \circ R_{A\to B}$ dependerá de las funciones de conjunción, disyunción e implicación seleccionadas, tal como indican la ecuación (5.178) o la (5.179). Por ejemplo tomando la función min para la conjunción y max para la disyunción,[24] tenemos que

$$\mu_B(y) = \max_{x\in X} \min(\mu_A(x), \mu_{A\to B}(x,y))$$

lo que da lugar a

$$B'^L = 0'4|y^1 + 0'7|y^2 + 1|y^3 \qquad B'^K = 0'3|y^1 + 0'7|y^2 + 1|y^3$$

$$B'^R = 0'36|y^1 + 0'7|y^2 + 1|y^3 \qquad B'^Z = 0'3|y^1 + 0'7|y^2 + 1|y^3$$

$$B'^G = 0'2|y^1 + 0'7|y^2 + 1|y^3 \qquad B'^{GR} = 0|y^1 + 0'3|y^2 + 1|y^3$$

□

Observe que en este ejemplo, a pesar de que $A' \subset A$, la única relación de implicación que cumple la inclusión $B' \subseteq B$ es la de Gaines-Rescher. Naturalmente, en vez de aplicar la composición max-min podríamos haber escogido otro par conorma-norma (no necesariamente conjugadas) para cada función de implicación, con el fin de que se cumpliera que

$$A' \subseteq A \implies B' = (A' \circ R_{A\to B}) \subseteq B \tag{5.180}$$

[24] Una vez más, si X fuera un conjunto infinito deberíamos tomar el supremo en vez del máximo.

Esta propiedad puede deducirse —por la monotonía de la función de implicación— a partir de otra propiedad más general:

$$A' = A \implies B' = (A' \circ R_{A\to B}) = B$$

la cual se expresa de forma más simple como

$$A \circ R_{A\to B} = B \tag{5.181}$$

Proposición 5.77 En la lógica clásica se cumple la propiedad (5.181) para todo conjunto A no vacío.

Demostración. Sea $B' = A \circ R_{A\to B}$. Por las ecuaciones (5.171) y (5.176) tenemos que

$$B' = \{y \mid \exists x,\ x \in A \wedge (x,y) \in R_{A\to B}\} = \{y \mid \exists x,\ x \in A \wedge [(x \in A) \to (y \in B)]\}$$

Por un lado,

$$y \in B' \Longrightarrow \exists x,\ x \in A \wedge [(x \in A) \to (y \in B)] \Longrightarrow y \in B$$

lo cual prueba que $B' \subseteq B$. Por otro lado, si $y \in B$, la condición $[(x \in A) \to (y \in B)]$ se cumple para todo $x \in X$ (tanto si $x \in A$ como si $x \notin A$); como A no es un conjunto vacío, existe al menos un x_0 tal que $x_0 \in A$ (con valor de verdad 1), se cumple que $x_0 \in A \wedge [(x_0 \in A) \to (y \in B)]$, y por tanto $y \in B'$, lo cual prueba que $B \subseteq B'$. □

Proposición 5.78 Una implicación (difusa) $f_\to$ tipo R cumple la propiedad (5.181) para todo conjunto A normalizado, siempre que la conorma $f_\vee$ de la función de composición (ec. (5.179)) sea max y $f_\wedge$ sea la norma que generó $f_\to$ (ec. (5.79)).

Demostración. Puesto que A está normalizado, existe un $x_0 \in X$ tal que $\mu_A(x_0) = 1$; por tanto,

$$\mu_{B'}(y) \geq f_\wedge(\mu_A(x_0), f_\to(\mu_A(x_0), \mu_B(y))) = f_\to(1, \mu_B(y))$$

Como la función $f_\to$ es de tipo R, la ecuación (5.81) nos dice que

$$\mu_{B'}(y) \geq \mu_B(y)$$

Por otro lado, de la definición de $f_\to$ (ec. (5.79)) se deduce también que $f_\wedge(a, f_\to(a,b)) \leq b$, lo cual implica que

$$\forall x, \forall y,\ f_\wedge(\mu_A(x), f_\to(\mu_A(x), \mu_B(y))) \leq \mu_B(y)$$

y por tanto,

$$\mu_{B'}(y) = \max_{x\in X} f_\wedge(\mu_A(x), f_\to(\mu_A(x), \mu_B(y))) \leq \mu_B(y)$$

Uniendo estos dos resultados se demuestra que $\forall y$, $\mu_{B'}(y) = \mu_B(y)$, y por tanto $B' = B$. □

Esta proposición es importante por el motivo siguiente: la igualdad (5.181) es una de las propiedades deseables para el modus ponens difuso. Ahora bien, como hemos visto en el ejemplo anterior, dicha propiedad no se cumple en general, a no ser que escojamos de forma cuidadosa y coordinada las tres funciones —$f_\wedge$, $f_\vee$ y $f_\to$— que intervienen en la ecuación (5.179), lo cual no es tarea fácil. La proposición que acabamos de demostrar nos ofrece un medio de realizar dicha elección: tomamos una norma cualquiera, la implicación R generada por ella (ec. (5.79)) y la conorma max; de este modo tenemos garantizado que se cumple la propiedad buscada.

Otra forma de hacer que se cumpla esta propiedad consiste en construir una tabla que indique cuál es el conjunto $B' = A \circ R_{A\to B}$ resultante cuando se toman distintas funciones $f_\wedge$, $f_\vee$ y $f_\to$; ello nos permite escoger una combinación tal que $B' = B$. En el libro de Klir y Yuan [35, sec. 11.3] se puede encontrar una tabla de este tipo para distintas funciones de implicación y distintas normas, cuando la conorma es max/sup.

También incluye dicho libro otra tabla que indica qué combinaciones de funciones cumplen la propiedad

$$\bar{B} \circ R^{-1}_{A\to B} = \bar{A} \tag{5.182}$$

correspondiente al modus tollens, y una tercera tabla que indica las combinaciones que cumplen la propiedad

$$R_{A\to B} \circ R_{B\to C} = R_{A\to C} \tag{5.183}$$

denominada "silogismo hipotético", pues permite agrupar las reglas $A \to B$ y $B \to C$ en una nueva regla $A \to C$, de modo que para todo A' se cumpla que $(A' \circ R_{A\to B}) \circ R_{B\to C} = A' \circ (R_{A\to B} \circ R_{B\to C}) = A' \circ R_{A\to C}$.

Sin embargo, como reconocen los propios autores, aún no hay criterios generales que permitan escoger adecuadamente las funciones que intervienen en el modus ponens difuso a partir de las propiedades que se desean cumplir, por lo que éste es uno de los temas de investigación más importantes que quedan abiertos en el campo de la lógica difusa.

5.5 Bibliografía recomendada

Como dijimos en la sección 1.2, el número de artículos, libros, revistas y congresos dedicados a la lógica difusa es de varios millares. Dado el carácter introductorio de este texto, vamos a recomendar solamente cuatro libros, que pueden servir como punto de partida para una búsqueda bibliográfica más extensa. El primero es el de Klir y Yuan [35], que, además de ser muy claro y bien organizado, es realmente exhaustivo, pues cubre todos los aspectos de la lógica difusa e incluye más de 1.800 referencias. El segundo es el de Timothy Ross [54], que, como el anterior, cubre todos los aspectos importantes de la lógica difusa; la diferencia es que el de Klir y Yuan expone primero la teoría y luego dedica un capítulo a cada uno de los campos de aplicación, mientras que el segundo intercala la exposición teórica y los ejemplos (aunque sólo incluye aplicaciones en ingeniería). El tercer libro que recomendamos es el de Trillas, Alsina y Terricabras [60], cuya mejor cualidad, a nuestro juicio, es que ofrece una interesante discusión de las implicaciones filosóficas de la lógica difusa (sazonada, por cierto, con citas de Antonio Machado); a nuestro juicio, resulta más difícil de comprender que los dos anteriores, a pesar de ser el único de los cuatro que está escrito en castellano. Por último, como obra de referencia (no como material didáctico), recomendamos el libro de Dubois, Yager y Prade [20], que recopila un gran número artículos originales sobre las diversas ramas de la lógica difusa y la teoría de la posibilidad.

Bibliografía

[1] J. B. Adams. Probabilistic reasoning and certainty factors. En: B. G. Buchanan y E. H. Shortliffe (eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, cap. 12, págs. 263–271. Addison-Wesley, Reading, MA, 1984.

[2] B. G. Buchanan y E. A. Feigenbaum. DENDRAL and Meta-DENDRAL: Their applications dimension. *Artificial Intelligence*, **11**:5–24, 1978.

[3] B. G. Buchanan y E. H. Shortliffe (eds.). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA, 1984.

[4] B. G. Buchanan, G. Sutherland y E. A. Feigenbaum. Heuristic DENDRAL: A program for generating explanatory hypotheses in organic chemistry. En: B. Meltzer y D. Michie (eds.), *Machine Intelligence 4*. Edinburgh University Press, Edinburgh, 1969.

[5] R. Carnap. The two concepts of probability. En: *Logical Foundations of Probability*, págs. 19–51. University of Chicago Press, Chicago, 1950.

[6] E. Castillo, J. M. Gutiérrez y A. S. Hadi. *Expert Systems and Probabilistic Network Models*. Springer-Verlag, New York, 1997. Versión española: *Sistemas Expertos y Modelos de Redes Probabilísticas*, Academia de Ingeniería, Madrid, 1997.

[7] E. Charniak. Bayesian Networks without tears. *AI Magazine*, **12**:50–63, 1991.

[8] P. Cohen y M. Grinberg. A framework for heuristic reasoning about uncertainty. En: *Proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI–83)*, págs. 355–357, Karlsruhe, Germany, 1983.

[9] P. Cohen y M. Grinberg. A theory of heuristic reasoning about uncertainty. *AI Magazine*, **4**:17–23, 1983.

[10] R. G. Cowell, A. P. Dawid, S. L. Lauritzen y D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York, 1999.

[11] R. Davis, B. G. Buchanan y E. H. Shortliffe. Retrospective on "Production rules as a representation for a knowledge-based consultation program". *Artificial Intelligence*, **59**:181–189, 1993.

[12] F. T. de Dombal, J. R. Leaper, J. R. Staniland, A. McCann y J. Horrocks. Computer-aided diagnosis of acute abdominal pain. *British Medical Journal*, **2**:9–13, 1972.

[13] F. J. Díez. Parameter adjustment in Bayes networks. The generalized noisy OR–gate. En: *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI'93)*, págs. 99–105, Washington D.C., 1993. Morgan Kaufmann, San Mateo, CA.

[14] F. J. Díez. *Sistema Experto Bayesiano para Ecocardiografía.* Tesis doctoral, Dpto. Informática y Automática, UNED, Madrid, Spain, 1994.

[15] F. J. Díez. Local conditioning in Bayesian networks. *Artificial Intelligence*, **87**:1–20, 1996.

[16] F. J. Díez. Aplicaciones de los modelos gráficos probabilistas en medicina. En: J. A. Gámez y J. M. Puerta (eds.), *Sistemas Expertos Probabilísticos*, págs. 239–263. Universidad de Castilla-La Mancha, Cuenca, 1998.

[17] J. Doyle. A truth maintenance system. *Artificial Intelligence*, **12**:231–272, 1979.

[18] M. J. Druzdzel. *Probabilistic Reasoning in Decision Support Systems: From Computation to Common Sense.* Tesis doctoral, Dept. Engineering and Public Policy, Carnegie Mellon University, 1993.

[19] M. J. Druzdzel y H. A. Simon. Causality in Bayesian belief networks. En: *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI'93)*, págs. 3–11, Washington D.C., 1993. Morgan Kaufmann, San Mateo, CA.

[20] D. Dubois, R. R. Yager y H. Prade. *Readings in Fuzzy Sets for Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1993.

[21] R. O. Duda y P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York, 1973.

[22] J. A. Gámez y J. M. Puerta (eds.). *Sistemas Expertos Probabilísticos.* Universidad de Castilla-La Mancha, Cuenca, 1998.

[23] J. Gordon y E. H. Shortliffe. The Dempster-Shafer theory of evidence. En: B. G. Buchanan y E. H. Shortliffe (eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, cap. 13, págs. 272–292. Addison-Wesley, Reading, MA, 1984.

[24] G. A. Gorry. Computer-assisted clinical decision making. *Methods of Information in Medicine*, **12**:45–51, 1973.

[25] G. A. Gorry y G. O. Barnett. Experience with a model of sequential diagnosis. *Computers and Biomedical Research*, **1**:490–507, 1968.

[26] D. Heckerman. Probabilistic interpretations for MYCIN's certainty factors. En: L. N. Kanal y J. F. Lemmer (eds.), *Uncertainty in Artificial Intelligence*, págs. 167–196. Elsevier Science Publishers, Amsterdam, The Netherlands, 1986.

[27] D. E. Heckerman. *Probabilistic Similarity Networks.* Tesis doctoral, Dept. Computer Science, Stanford University, STAN–CS–90–1316, 1990.

[28] D. E. Heckerman y E. J. Horvitz. On the expresiveness of rule-based systems for reasoning with uncertainty. En: *Proceedings of the 6th National Conference on AI (AAAI–87)*, págs. 121–126, Seattle, WA, 1987.

[29] D. E. Heckerman y E. J. Horvitz. The myth of modularity in rule-based systems for reasoning with uncertainty. En: J. F. Lemmer y L. N. Kanal (eds.), *Uncertainty in Artificial Intelligence 2*, págs. 23–34. Elsevier Science Publishers, Amsterdam, The Netherlands, 1988.

[30] M. Henrion. Some practical issues in constructing belief networks. En: L. N. Kanal, T. S. Levitt y J. F. Lemmer (eds.), *Uncertainty in Artificial Intelligence 3*, págs. 161–173. Elsevier Science Publishers, Amsterdam, The Netherlands, 1989.

[31] F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York, 2001.

[32] F. V. Jensen, K. G. Olesen y S. K. Andersen. An algebra of Bayesian belief universes for knowledge-based systems. *Networks*, **20**:637–660, 1990.

[33] P. Juez Martel y F. J. Díez Vegas. *Probabilidad y Estadística en Medicina. Aplicaciones en la Práctica Clínica y en la Gestión Sanitaria*. Ed. Díaz de Santos, Madrid, 1996.

[34] J. H. Kim. *CONVINCE: A conversational inference consolidation engine*. Tesis doctoral, Dept. Computer Science, University of California, Los Angeles, 1983.

[35] G. J. Klir y B. Yuan. *Fuzzy Sets and Fuzzy Logic. Theory and Applications*. Prentice Hall, Upper Saddle River, NJ, 1995.

[36] P. Krause y D. Clark. *Representing Uncertain Knowledge. An Artificial Intelligence Approach*. Intellect Books, Oxford, UK, 1993.

[37] P. Larrañaga. Aprendizaje automático de modelos gráficos II. Aplicaciones a la clasificación supervisada. En: J. A. Gámez y J. M. Puerta (eds.), *Sistemas Expertos Probabilísticos*, págs. 141–160. Universidad de Castilla-La Mancha, Cuenca, 1998.

[38] J. McCarthy y P. Hayes. Some phylosophical problems from the standpoint of Artificial Intelligence. En: B. Meltzer y D. Michie (eds.), *Machine Intelligence 4*, págs. 463–502. Edinburgh University Press, Edinburgh, 1969.

[39] W. S. McCulloch y W. H. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**:115–133, 1943.

[40] J. Mira, A. E. Delgado, J. G. Boticario y F. J. Díez. *Aspectos Básicos de la Inteligencia Artificial*. Sanz y Torres, Madrid, 1995.

[41] R. E. Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. Wiley-Interscience, New York, 1990.

[42] A. Newell y H. A. Simon. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ, 1972.

[43] J. Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. En: *Proceedings of the 2nd National Conference on AI (AAAI–82)*, págs. 133–136, Pittsburgh, Pennsylvania, 1982.

[44] J. Pearl. Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, **29**:241–288, 1986.

[45] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988. Reimpreso con correcciones en 1991.

[46] J. Pearl. From conditional oughts to qualitative decision theory. En: *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI'93)*, págs. 12–20, Washington D.C., 1993. Morgan Kaufmann, San Mateo, CA.

[47] J. Pearl, D. Geiger y T. Verma. Conditional independence and its representations. *Kybernetika*, **25**:33–44, 1989.

[48] J. Pearl y T. S. Verma. A statistical semantics for causation. *Statistics and Computing*, **2**:91–95, 1992.

[49] M. A. Peot. Geometric implications of the Naive Bayes assumption. En: *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI'94)*, págs. 414–419, Seattle, WA, 1996. Morgan Kaufmann, San Francisco, CA.

[50] M. A. Peot y R. D. Shachter. Fusion and propagation with multiple observations in belief networks. *Artificial Intelligence*, **48**:299–318, 1991.

[51] E. Post. Formal reductions of the general combinatorial problem. *American Journal of Mathematics*, **65**:197–268, 1943.

[52] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, **13**:81–132, 1980.

[53] A. Rosenblueth, N. Wiener y J. Bigelow. Behavior, purpose and teleology. *Philosophy of Science*, **10**:18–24, 1943.

[54] T. J. Ross. *Fuzzy Logic with Engineering Applications*. McGraw-Hill, New York, 1995.

[55] G. Shafer. Probability judgment in artificial intelligence and expert systems. *Statistical Science*, **2**:3–44, 1987.

[56] G. Shafer y J. Pearl. *Readings in Uncertain Reasoning*. Morgan Kaufmann, San Mateo, CA, 1990.

[57] E. H. Shortliffe y B. G. Buchanan. A model of inexact reasoning in medicine. En: B. G. Buchanan y E. H. Shortliffe (eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, cap. 11, págs. 233–262. Addison-Wesley, Reading, MA, 1984.

[58] E. H. Shortliffe, B. G. Buchanan y E. A. Feigenbaum. Knowledge engineering for medical decision making: A review of computer-based clinical decision aids. *Proceedings of the IEEE*, **67**:1207–1224, 1979.

[59] P. Szolovits y S. G. Pauker. Categorical and probabilistic reasoning in medicine. *Artificial Intelligence*, **11**:115–144, 1978.

[60] E. Trillas, C. Alsina y J. M. Terricabras. *Introducción a la Lógica Borrosa*. Ariel, Barcelona, 1995.

[61] W. van Melle. The structure of the MYCIN system. En: B. G. Buchanan y E. H. Shortliffe (eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, cap. 4, págs. 67–77. Addison-Wesley, Reading, MA, 1984.

[62] W. van Melle, E. H. Shortliffe y B. G. Buchanan. EMYCIN: A knowledge engineer's tool for constructing rule-based expert systems. En: B. G. Buchanan y E. H. Shortliffe (eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, cap. 15, págs. 302–313. Addison-Wesley, Reading, MA, 1984.

[63] H. R. Warner, A. F. Toronto y L. G. Veasy. Experience with Bayes' theorem for computer diagnosis of congenital heart disease. *Annals of the New York Academy of Sciences*, **115**:558–567, 1964.

[64] M. P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, **44**:257–303, 1990.

[65] M. P. Wellman. Graphical inference in qualitative probabilistic networks. *Networks*, **20**:687–701, 1990.

[66] V. L. Yu, L. M. Fagan, S. M. Wraith, W. J. Clancey, A. C. Scott, J. F. Hannigan, R. L. Blum, B. G. Buchanan y S. N. Cohen. An evaluation of MYCIN's advice. En: B. G. Buchanan y E. H. Shortliffe (eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, cap. 31, págs. 589–596. Addison-Wesley, Reading, MA, 1984.