Introducción a los
Modelos Gráficos Probabilistas

Francisco Javier Díez
Dpto. Inteligencia Artificial
UNED

Primera edición: octubre de 2007
Revisión: 24 de octubre de 2014

Índice general

Prefacio v

1. Fundamentos de redes bayesianas 1
Resumen . . . 1
Contexto . . . 1
Objetivos . . . 1
Requisitos previos . . . 2
Contenido . . . 2
1.1. Repaso de la teoría de la probabilidad . . . 2
1.1.1. Definiciones básicas sobre probabilidad . . . 2
1.1.2. Independencia y correlación . . . 7
1.1.3. Teorema de Bayes . . . 10
1.2. Método bayesiano ingenuo . . . 16
1.2.1. Forma racional del método bayesiano ingenuo . . . 19
1.2.2. Discusión . . . 20
1.3. Nociones sobre grafos . . . 21
1.3.1. Definiciones básicas . . . 21
1.3.2. Grafos dirigidos acíclicos . . . 24
1.4. Definición de red bayesiana . . . 26
1.4.1. Construcción de una red bayesiana . . . 26
1.4.2. Propiedad de Markov . . . 29
1.5. Grafos de dependencias e independencias probabilistas . . . 31
1.5.1. Separación en grafos dirigidos y no dirigidos . . . 31
1.5.2. Mapas de independencias . . . 32
1.5.3. Separación direccional y redes bayesianas . . . 33
1.6. Causalidad y correlación . . . 34
1.6.1. Interpretación probabilista e interpretación causal de un grafo . . . 34
1.6.2. Diferencia entre causalidad y correlación . . . 34
Bibliografía recomendada . . . 36
Actividades . . . 37

2. Inferencia en redes bayesianas 39
Resumen . . . 39
Contexto . . . 39

Objetivos . . . 39
Requisitos previos . . . 39
Contenido . . . 40
2.1. Planteamiento del problema . . . 40
2.1.1. Diagnóstico probabilista . . . 40
2.1.2. Método de fuerza bruta . . . 40
2.2. Métodos exactos . . . 42
2.2.1. Eliminación de variables . . . 42
2.2.2. Agrupamiento . . . 46
2.2.3. Variantes del método de agrupamiento . . . 59
2.2.4. Inversión de arcos . . . 67
2.3. Métodos aproximados . . . 77
2.3.1. Fundamento de los métodos estocásticos . . . 77
2.3.2. Muestreo lógico . . . 77
2.3.3. Ponderación por verosimilitud . . . 78
2.3.4. Otros métodos . . . 78
2.3.5. Complejidad computacional de los métodos estocásticos . . . 78
Bibliografía recomendada . . . 78
Actividades . . . 78

3. Construcción de redes bayesianas **81**
Resumen . . . 81
Contexto . . . 81
Objetivos . . . 81
Requisitos previos . . . 81
Contenido . . . 82
3.1. Construcción de redes causales con conocimiento experto . . . 82
3.1.1. Necesidad de la construcción manual (en algunos casos) . . . 82
3.1.2. Fase cualitativa: estructura de la red . . . 83
3.1.3. Fase cuantitativa: obtención de las probabilidades condicionales . . . 85
3.1.4. Resumen . . . 88
3.2. Modelos canónicos . . . 89
3.2.1. Modelos deterministas . . . 90
3.2.2. Modelos IIC . . . 92
3.2.3. Modelos OR/MAX . . . 93
3.2.4. Uso de los modelos canónicos en la construcción de redes bayesianas . . . 97
3.3. Aprendizaje automático a partir de bases de datos . . . 99
3.3.1. Planteamiento del problema . . . 99
3.3.2. Cuestiones generales sobre aprendizaje . . . 100
3.3.3. Aprendizaje paramétrico . . . 104
3.3.4. Aprendizaje estructural a partir de relaciones de independencia . . . 106
3.3.5. Aprendizaje estructural mediante búsqueda heurística . . . 108
3.3.6. Otras cuestiones . . . 110

Bibliografía recomendada . . . 111
Actividades . . . 111

4. Análisis de decisiones **113**
Resumen . . . 113
Contexto . . . 113
Objetivos . . . 114
Requisitos previos . . . 114
Contenido . . . 114
4.1. Fundamentos de la teoría de la decisión . . . 114
4.2. Diagramas de influencia y árboles de decisión . . . 114
4.2.1. Definición de diagrama de influencia . . . 114
4.2.2. Políticas y utilidades esperadas . . . 117
4.3. Evaluación de diagramas de influencia . . . 117
4.3.1. Expansión y evaluación de un árbol de decisión . . . 118
4.3.2. Eliminación de variables . . . 122
4.3.3. Inversión de arcos . . . 125
4.4. Construcción de diagramas de influencia . . . 131
4.5. Análisis de sensibilidad . . . 131
Bibliografía recomendada . . . 131
Actividades . . . 131

5. Aplicaciones **133**
Resumen . . . 133
Contexto . . . 133
Objetivos . . . 133
Requisitos previos . . . 133
Contenido . . . 134
5.1. Aplicaciones en medicina . . . 134
5.2. Aplicaciones en ingeniería . . . 134
5.3. Aplicaciones en informática . . . 134
5.3.1. Informática educativa . . . 134
5.3.2. Interfaces inteligentes . . . 134
5.3.3. Seguridad informática . . . 135
5.4. Aplicaciones en visión artificial . . . 135
5.5. Aplicaciones financieras y comerciales . . . 135
5.6. Otras aplicaciones . . . 135
Bibliografía recomendada . . . 135
Actividades . . . 136

Referencias **142**

Prefacio

Estos apuntes se han redactado como texto básico para varias asignaturas de la la Escuela Técnica Superior de Ingeniería Informática de la Universidad Nacional de Educación a Distancia (UNED): *Técnicas Avanzadas de Razonamiento* (quinto curso de Ingeniería Informática, ahora en extinción), *Modelos Probabilistas y Análisis de Decisiones* (cuerto curso del Grado en Ingeniería en Tecnologías de la Información), *Métodos Probabilistas en Inteligencia Artificial* (Máster en Inteligencia Artificial Avanzada) y *Análisis de Decisiones en Medicina* (Máster en Física Médica).

Algunos apartados que no están completamente desarrollados hacen referencia al libro de Castillo, Gutiérrez y Hadi [7], una obra excelente que, a pesar de que fue publicada hace bastante tiempo, sigue siendo útil para estudiantes e investigadores. También hacemos referencia a un informe técnico orientado a la medicina [22]. Ambas referencias están disponibles en Internet.

Damos las gracias a los alumnos que nos han comunicado las erratas detectadas, especialmente a José Luis Sanz Yubero, Pilar Herrera Plaza, Fernando Giner Martínez, Jon Haitz Legarreta Gorroño y Agustín Uruburu, así como al Prof. Manuel Luque, de la UNED. Todos los comentarios y sugerencias que recibamos nos serán muy útiles para preparar la próxima versión de estos apuntes.

Francisco Javier Díez Vegas
UNED, Madrid, octubre de 2014

Índice de figuras

1.1. La razón de probabilidad $RP(X)$ como función de la probabilidad $P(+x)$. . . 14
1.2. Valor predictivo positivo (prevalencia=0'1). 16
1.3. Valor predictivo negativo (prevalencia=0'1). 16
1.4. Representación del método probabilista ingenuo mediante un grafo de independencia. 18
1.5. El piloto luminoso (L) y la temperatura (T) son signos de avería (D). 19
1.6. Un grafo dirigido. 22
1.7. Caminos cerrados: tres ciclos (fila superior) y dos bucles (fila inferior). 24
1.8. La correlación entre número de cigüeñas y número de nacimientos **no** implica causalidad. 35
1.9. La correlación entre el consumo de teracola y la aparición de manchas en la piel **no** implica causalidad. 36

2.1. Red bayesiana de seis nodos. 42
2.2. Red bayesiana de siete nodos. 47
2.3. Un posible árbol de grupos para la red de la figura 2.2. Junto a cada grupo hemos escrito sus potenciales asociados. 47
2.4. Árbol de grupos para la red de la figura 2.2. Es el mismo árbol de la figura 2.3, pero aquí hemos dibujado los separadores de cada par de nodos vecinos y hemos omitido la dirección de los enlaces. 53
2.5. Construcción de un árbol de grupos para la red bayesiana de la figura 2.2. A medida que vamos eliminando nodos del grafo de dependencias va creciendo el árbol de grupos. 57
2.6. Grafo de dependencias para la red de la figura 2.2 y la evidencia $\mathbf{e} = \{+b, \neg g\}$. 60
2.7. Un posible árbol de grupos para la red de la figura 2.2, específico para la evidencia $\mathbf{e} = \{+b, \neg g\}$. 61
2.8. Grafo triangulado correspondiente al grafo de dependencias que aparece en la figura 2.5.*a*. 62
2.9. Red de cinco nodos y cinco enlaces. 68
2.10. Red resultante de invertir el enlace $X \to Y$ en la red anterior. 69
2.11. Red resultante de podar los nodos F y H en la red de la figura 2.2. La probabilidad $P(a|+g)$ obtenida a partir de esta nueva red es la misma que obtendríamos a partir de la red original. 73
2.12. Red bayesiana en que el nodo A tiene tres hijos. Si queremos convertir A en sumidero, primero tenemos que invertir el enlace $A \to D$, luego $A \to C$ y finalmente $A \to B$. 74

2.13. Inversión de arcos para el cálculo de $P(a|+g)$. Partiendo de la red de la figura 2.11, invertimos el enlace $C \to G$, lo cual obliga a añadir los enlaces $A \to G$ y $D \to C$. Luego eliminamos C, invertimos el enlace $D \to G$, eliminamos D, invertimos $B \to G$ y eliminamos B. Así llegamos a una red en que sólo aparecen A y G. 76

3.1. Causas de la isquemia y de la hipertensión arterial (HTA), según cierto libro de cardiología. 85
3.2. Reinterpretación de las causas de isquemia y de HTA. 86
3.3. Estructura interna de los modelos IIC. Las variables Z_i se introducen para explicar la construcción del modelo, pero no forman parte del modelo, en el cual sólo intervienen Y y las X_i's. 92
3.4. Estructura interna de los modelos IIC residuales. La variable auxiliar Z_L representa las causas que no están explícitas en el modelo. 93
3.5. Modelo MAX causal en Elvira. *a*) Parámetros canónicos. *b*) Tabla de probabilidad condicional (TPC). 98

4.1. DI con dos nodos de decisión (rectángulos), dos nodos de azar (óvalos) y tres nodos de utilidad (hexágonos). Observe que hay un camino dirigido, $T \to Y \to D \to U_1 \to U_0$, que incluye todas las decisiones y el nodo de utilidad global, U_0. 115
4.2. Evaluación del DI de la figura 4.1 mediante el método de inversión de arcos. . 130

Índice de tablas

3.1. Algunas de las funciones más comunes utilizadas para construir modelos canónicos. En el caso de variables booleanas, pos($\mathbf{x}$) indica el número de variables dentro de la configuración $\mathbf{x}$ que toman el valor "verdadero", "presente" o "positivo". 90
3.2. Tablas de probabilidad condicional (TPC) para algunos de los modelos deterministas inducidos por las funciones lógicas de la tabla 3.1. 91
3.3. Parámetros del modelo OR "con ruido" para el enlace $X_i \rightarrow Y$. 94
3.4. TPC para un modelo OR "con ruido". Los padres son X_1 y X_2. 95
3.5. TPC para el modelo OR residual con dos padres. 96

Capítulo 1

Fundamentos de redes bayesianas

Resumen

En este capítulo vamos a estudiar los conceptos fundamentales sobre redes bayesianas. Como preparación para su estudio, empezaremos repasando los aspectos más elementales de la teoría de la probabilidad, explicaremos el método bayesiano ingenuo (que es un antecesor de las redes bayesianas) y repasaremos también los conceptos básicos de la teoría de grafos. Por fin, en la sección 1.4 daremos la definición de red bayesiana. Con el fin de entender mejor las propiedades de estas redes, estudiaremos luego los grafos de dependencias e independencias, y veremos que el grafo de una red bayesiana es un grafo de independencia. Por último, discutiremos la diferencia entre correlación y causalidad y veremos cómo un grafo dirigido puede tener tanto una interpretación causal como una interpretación probabilista, que pueden ser muy diferentes, aunque en el caso de las redes bayesianas suelen estar íntimamente relacionadas.

Contexto

Este capítulo ofrece, por un lado, un repaso de los conceptos sobre probabilidad y sobre grafos que el alumno ha aprendido en Educación Secundaria y en algunas de las asignaturas de la carrera de informática, y por otro, presenta dos métodos desarrollados en el campo de la inteligencia artificial con el fin de resolver problemas del mundo real: el método probabilista ingenuo y las redes bayesianas.

Objetivos

El objetivo de este capítulo es llegar a entender los fundamentos de las redes bayesianas y aprender a manejar algún programa de ordenador para modelos gráficos probabilistas, como Elvira u OpenMarkov.[1]

[1] Véase `www.ia.uned.es/~elvira` y `www.openmarkov.org`.

Requisitos previos

Aunque se supone que el alumno ya posee conocimientos sobre probabilidad y sobre grafos, hacemos un breve repaso para afianzar los conceptos que vamos a utilizar a lo largo de esta asignatura.

Contenido

1.1. Repaso de la teoría de la probabilidad

1.1.1. 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 esta asignatura vamos a tratar solamente con variables discretas, podemos simplificar considerablemente la exposición tomando como punto de partida el concepto de variable aleatoria.

Definición 1.1 (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" (véanse también las definiciones 1.33 y 1.34). 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 $\mathbf{X}$.[2] La configuración $\mathbf{x} = (x_1, \ldots, x_n)$ representa la configuración de $\mathbf{X}$ en 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 estos apuntes vamos a suponer que todas las variables son discretas. Nuestro punto de partida para la definición de probabilidad será el siguiente:

Definición 1.2 (Probabilidad conjunta) Dado un conjunto de variables discretas $\mathbf{X} = \{X_1, \ldots, X_n\}$, definimos la *probabilidad conjunta* como una aplicación que a cada configuración $\mathbf{x} = (x_1, \ldots, x_n)$ le asigna un número real no negativo de modo que

$$\sum_{\mathbf{x}} P(\mathbf{x}) = \sum_{x_1} \cdots \sum_{x_n} P(x_1, \ldots, x_n) = 1 \tag{1.1}$$

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.

Definición 1.3 (Probabilidad marginal) Dada una distribución de probabilidad conjunta $P(x_1, \ldots, x_n)$, la *probabilidad marginal* para un subconjunto de variables $\mathbf{X'} = \{X'_1, \ldots, X'_{n'}\} \subset \mathbf{X}$ viene dada por

$$P(\mathbf{x'}) = P(x'_1, \ldots, x'_{n'}) = \sum_{x_i \,|\, X_i \notin \mathbf{X'}} P(x_1, \ldots, x_n) \tag{1.2}$$

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

$$P(x_i) = \sum_{x_j \,|\, X_j \neq X_i} P(x_1, \ldots, x_n) \tag{1.3}$$

Proposición 1.4 Dada una distribución de probabilidad conjunta para $\mathbf{X}$, toda distribución de probabilidad marginal obtenida a partir de ella para un subconjunto $\mathbf{X'} \subset \mathbf{X}$ es a su vez una distribución conjunta para $\mathbf{X'}$.

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

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

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

$$\sum_{\mathbf{x'}} P(\mathbf{x'}) = \sum_{x_1} \cdots \sum_{x_n} P(x_1, \ldots, x_n) = 1 \tag{1.4}$$

[2] En este libro daremos por supuesto que todos los conjuntos son disjuntos, es decir, que no tienen elementos repetidos.

con lo que concluye la demostración. □

Corolario 1.5 Si a partir de una distribución de probabilidad conjunta calculamos la distribución marginal para una variable cualquiera X_i, la suma de los valores de esta distribución ha de ser la unidad:

$$\sum_{x_i} P(x_i) = 1 \tag{1.5}$$

Ejemplo 1.6 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. (1.2). 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. □

Definición 1.7 (Probabilidad condicional) Dados dos subconjuntos disjuntos de variables, $\mathbf{X} = \{X_1, \ldots, X_n\}$ e $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$, y una configuración $\mathbf{x}$ de $\mathbf{X}$ tal que $P(\mathbf{x}) > 0$, la *probabilidad condicional* de $\mathbf{y}$ dado $\mathbf{x}$, $P(\mathbf{y}|\mathbf{x})$, se define como

$$P(\mathbf{y} \,|\, \mathbf{x}) = \frac{P(\mathbf{x}, \mathbf{y})}{P(\mathbf{x})} \tag{1.6}$$

□

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

Ejemplo 1.8 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 1.9 Dados dos subconjuntos disjuntos de variables, $\mathbf{X}$ e $\mathbf{Y}$, y una configuración $\mathbf{x}$ tal que $P(\mathbf{x}) > 0$, se cumple que

$$\forall \mathbf{x}, \quad \sum_{\mathbf{y}} P(\mathbf{y} \,|\, \mathbf{x}) = 1 \tag{1.7}$$

Demostración. Aplicando las definiciones anteriores,

$$\sum_{\mathbf{y}} P(\mathbf{y} \,|\, \mathbf{x}) = \sum_{\mathbf{y}} \frac{P(\mathbf{x}, \mathbf{y})}{P(\mathbf{x})} = \frac{1}{P(\mathbf{x})} \sum_{\mathbf{y}} P(\mathbf{x}, \mathbf{y}) = \frac{1}{P(\mathbf{x})} P(\mathbf{x}) = 1 \tag{1.8}$$

□

Observe que esta proposición es el equivalente de la ecuación (1.4) para probabilidades condicionales.

Ejercicio 1.10 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 1.6 (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 1.11 (Teorema de la probabilidad total) Dados dos subconjuntos disjuntos de variables, $\mathbf{X}$ e $\mathbf{Y}$, se cumple que

$$P(\mathbf{y}) = \sum_{\mathbf{x} \,|\, P(\mathbf{x})>0} P(\mathbf{y} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \tag{1.9}$$

Demostración. Por la definición de probabilidad marginal,

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

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

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

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

$$P(\mathbf{y} \,|\, \mathbf{z}) = \sum_{\mathbf{x} \,|\, P(\mathbf{x} \,|\, \mathbf{z})>0} P(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z}) \cdot P(\mathbf{x} \,|\, \mathbf{z}) \tag{1.10}$$

Demostración. Por la definición de probabilidad condicional,

$$P(\mathbf{y} \,|\, \mathbf{z}) = \frac{P(\mathbf{y}, \mathbf{z})}{P(\mathbf{z})} = \frac{1}{P(\mathbf{z})} \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{y}, \mathbf{z})$$

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

$$P(\mathbf{z}) > 0 \Longrightarrow \{P(\mathbf{x} \,|\, \mathbf{z}) = 0 \Leftrightarrow P(\mathbf{x}, \mathbf{z}) = 0\}$$

Por tanto

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

Ejemplo 1.13 (Continuación del ejemplo 1.6) 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,

$$\begin{aligned} P(x_2^v) &= \sum_{x_1} P(x_2^v|x_1) \cdot P(x_1) \\ &= 0'49630 \cdot 0'270 + 0'49194 \cdot 0'496 + 0'48718 \cdot 0'243 \\ &= 0'134 + 0'244 + 0'114 = 0'492 \end{aligned}$$

que es el valor que ya conocíamos. □

Finalmente, enunciamos una proposición que se deduce fácilmente de la definición de probabilidad condicional, y que nos va a ser de gran utilidad.

Proposición 1.14 (Regla de la cadena) Dado un conjunto de variables $\mathbf{X}$ y una partición $\{\mathbf{X}_1, \ldots, \mathbf{X}_k\}$ de $\mathbf{X}$, se cumple que

$$P(\mathbf{x}) = \prod_{i=1}^{k} P(\mathbf{x}_i \,|\, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_k) \tag{1.11}$$

Demostración. Por la definición de probabilidad condicional

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

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

Ejemplo 1.15 Sea $\mathbf{X} = \{A, B, C, D, E\}$. 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

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

1.1.2. Independencia y correlación

Independencia y correlación (sin condicionamiento)

Definición 1.16 (Valores independientes) Dos valores x e y de dos variables X e Y, respectivamente, son independientes sii $P(x, y) = P(x) \cdot P(y)$.

Definición 1.17 (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 1.18 (Continuación del ejemplo 1.6) 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**.

Definición 1.19 (Variables independientes) Dos variables X e Y son independientes sii todos los pares de valores x e y son independientes, es decir, sii

$$\forall x, \forall y, \quad P(x, y) = P(x) \cdot P(y) \tag{1.12}$$

Definición 1.20 (Variables correlacionadas) Dos variables X e Y están correlacionadas sii no son independientes, es decir, sii

$$\exists x, \exists y, \quad P(x, y) \neq P(x) \cdot P(y) \tag{1.13}$$

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 estos apuntes, 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 $\mathbf{X}$ e $\mathbf{Y}$, y de dos valores x e y a dos configuraciones $\mathbf{x}$ e $\mathbf{y}$.

Independencia condicional

Definición 1.21 (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 $P(x, y|z) = P(x|z) \cdot P(y|z)$.

Definición 1.22 (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

$$\forall x, \forall y, \forall z, \quad P(z) > 0 \implies P(x, y \,|\, z) = P(x \,|\, z) \cdot P(y \,|\, z) \tag{1.14}$$

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

Proposición 1.23 Sea un conjunto de variables $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$ condicionalmente independientes dada la configuración $\mathbf{x}$ de $\mathbf{X}$, es decir, $P(\mathbf{y}|\mathbf{x}) = \prod_{j=1}^{m} P(y_j|\mathbf{x})$. Para todo subconjunto $\mathbf{Y}'$ de $\mathbf{Y}$ se cumple que:

$$\forall \mathbf{y}', \; P(\mathbf{y}'|\mathbf{x}) = \prod_{j|Y_j \in \mathbf{Y}'} P(y_j|\mathbf{x}) \tag{1.15}$$

Demostración. Por la definición de probabilidad marginal,

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

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 1.24 Sean $\mathbf{Y} = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$ e $\mathbf{Y}' = \{Y_1, Y_4\}$. Supongamos que se cumple la condición (1.15), que para este ejemplo es

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

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

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

El resultado de calcular los sumatorios da la unidad, pues

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

Proposición 1.25 Sea un conjunto de variables $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$ condicionalmente independientes dada la configuración $\mathbf{x}$ de $\mathbf{X}$. Sean $\mathbf{Y}'$ e $\mathbf{Y}''$ dos subconjuntos disjuntos de $\mathbf{Y}$. Para todo par de configuraciones $\mathbf{y}'$ e $\mathbf{y}''$ se cumple que

$$P(\mathbf{y}', \mathbf{y}'' | \mathbf{x}) = P(\mathbf{y}' | \mathbf{x}) \cdot P(\mathbf{y}'' | \mathbf{x}) \tag{1.16}$$

$$P(\mathbf{y}' | \mathbf{x}, \mathbf{y}'') = P(\mathbf{y}' | \mathbf{x}) \tag{1.17}$$

Demostración. Por la proposición anterior,

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

con lo que se demuestra la primera ecuación, y de ella se deduce la segunda:

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

Ejemplo 1.26 Sea de nuevo el conjunto $\mathbf{Y} = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$ de variables condicionalmente independientes dada $\mathbf{x}$. Sean $\mathbf{Y}' = \{Y_1, Y_4\}$ e $\mathbf{Y}'' = \{Y_2\}$. La proposición que acabamos de demostrar nos dice que

$$\begin{aligned} P(y_1, y_2, y_4 \,|\, \mathbf{x}) &= P(y_1, y_4 \,|\, \mathbf{x}) \cdot P(y_2 \,|\, \mathbf{x}) \\ P(y_1, y_4 \,|\, \mathbf{x}, y_2) &= P(y_1, y_4 \,|\, \mathbf{x}) \end{aligned}$$

□

Estas dos proposiciones nos serán muy útiles más adelante.

1.1.3. Teorema de Bayes

Enunciado y demostración

La forma clásica del teorema de Bayes es la siguiente:

Teorema 1.27 (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

$$P(x \mid y) = \frac{P(x) \cdot P(y \mid x)}{\sum_{x'} P(x') \cdot P(y \mid x')} \tag{1.18}$$

Este teorema se puede generalizar así:

Teorema 1.28 (Teorema de Bayes generalizado) Dadas dos configuraciones $\mathbf{x}$ e $\mathbf{y}$ de dos subconjuntos de variables $\mathbf{X}$ e $\mathbf{Y}$, respectivamente, tales que $P(\mathbf{x}) > 0$ y $P(\mathbf{y}) > 0$, se cumple que

$$P(\mathbf{x} \mid \mathbf{y}) = \frac{P(\mathbf{x}) \cdot P(\mathbf{y} \mid \mathbf{x})}{\sum_{\mathbf{x}' \mid P(\mathbf{x}')>0} P(\mathbf{x}') \cdot P(\mathbf{y} \mid \mathbf{x}')} \tag{1.19}$$

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

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

Basta aplicar el teorema de la probabilidad total (proposición 1.11) para completar la demostración. □

Proposición 1.29 Dados tres subconjuntos disjuntos $\mathbf{X}$, $\mathbf{Y}$ y $\mathbf{Z}$, si $P(\mathbf{y}, \mathbf{z}) > 0$, se cumple que

$$P(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = P(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) \cdot P(\mathbf{y} \mid \mathbf{z}) \tag{1.20}$$

Demostración. Veamos primero que

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

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

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

como queríamos demostrar. □

Proposición 1.30 (Teorema de Bayes con condicionamiento) Dadas tres configuraciones $\mathbf{x}$, $\mathbf{y}$ y $\mathbf{z}$ de tres conjuntos de variables $\mathbf{X}$, $\mathbf{Y}$ y $\mathbf{Z}$, respectivamente, tales que $P(\mathbf{x}, \mathbf{z}) > 0$ y $P(\mathbf{y}, \mathbf{z}) > 0$, se cumple que

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

Demostración. Por la definición de probabilidad condicional,

$$P(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = \frac{P(\mathbf{x}, \mathbf{y}, \mathbf{z})}{P(\mathbf{y}, \mathbf{z})} = \frac{P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) \cdot P(\mathbf{x}, \mathbf{z})}{P(\mathbf{y}, \mathbf{z})} \tag{1.21}$$

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

$$P(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = \frac{P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) \cdot P(\mathbf{x}, \mathbf{z}) \,/\, P(\mathbf{z})}{P(\mathbf{y}, \mathbf{z}) \,/\, P(\mathbf{z})} = \frac{P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) \cdot P(\mathbf{x} \mid \mathbf{z})}{P(\mathbf{y} \mid \mathbf{z})} \tag{1.22}$$

Basta ahora aplicar la ecuación (1.10) para concluir la demostración. □

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:

Definición 1.31 (Hallazgo) Es la determinación del valor de una variable, $H = h$, a partir de un dato (una observación, una medida, etc.).

Definición 1.32 (Evidencia) Es el conjunto de todos los hallazgos disponibles en un determinado momento o situación: $\mathbf{e} = \{H_1 = h_1, \ldots, H_r = h_r\}$.

Definición 1.33 (Probabilidad a priori) Es la probabilidad de una variable o un conjunto de variables cuando no hay ningún hallazgo.

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

Definición 1.34 (Probabilidad a posteriori) Es la probabilidad de una variable o un conjunto de variables dada la evidencia $\mathbf{e}$: $P(\mathbf{x} \mid \mathbf{e})$.

Ejemplo 1.35 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:

$P(y \mid x)$	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,

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

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 $P(x \,|\, y)$, puede ser más cómodo aplicar la forma normalizada del teorema de Bayes, que es la siguiente:

$$P(x \,|\, y) = \alpha \cdot P(x) \cdot P(y \,|\, x) \tag{1.23}$$

En esta expresión,

$$\alpha \equiv \left[\sum_{x'} P(x') \cdot P(y \,|\, x') \right]^{-1} = [P(y)]^{-1} \tag{1.24}$$

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

Ejemplo 1.36 (Continuación del ejemplo 1.35) Para calcular la probabilidad a posteriori de cada universidad (es decir, la probabilidad sabiendo que es un profesor de economía) aplicamos la ecuación (1.23):

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

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

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

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

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

Observe que la probabilidad a posteriori $P(x \mid y)$ 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

$$\frac{P(x^i \mid y)}{P(x^j \mid y)} = \frac{\alpha \cdot P(x^i) \cdot P(y \mid x^i)}{\alpha \cdot P(x^j) \cdot P(y \mid x^j)} = \frac{P(x^i)}{P(x^j)} \cdot \frac{P(y \mid x^i)}{P(y \mid x^j)} \tag{1.25}$$

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 1.37 En el ejemplo 1.35 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

$$RP_{pre}(X) \equiv \frac{P(+x)}{P(\neg x)} = \frac{P(+x)}{1 - P(+x)} \tag{1.26}$$

- Razón de probabilidad de X a posteriori

$$RP_{post}(X) \equiv \frac{P(+x \mid y)}{P(\neg x \mid y)} = \frac{P(+x \mid y)}{1 - P(+x \mid y)} \tag{1.27}$$

- Razón de verosimilitud para X dado y

$$RV_X(y) \equiv \frac{P(y \mid +x)}{P(y \mid \neg x)} \tag{1.28}$$

A partir de la ecuación (1.27) podemos hallar $P(+x \mid y)$:

$$P(+x \mid y) = \frac{RP_{post}(X)}{1 + RP_{post}(X)} \tag{1.29}$$

La figura 1.1 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 $P(+x) \to 1$, $RP(X) \to \infty$.

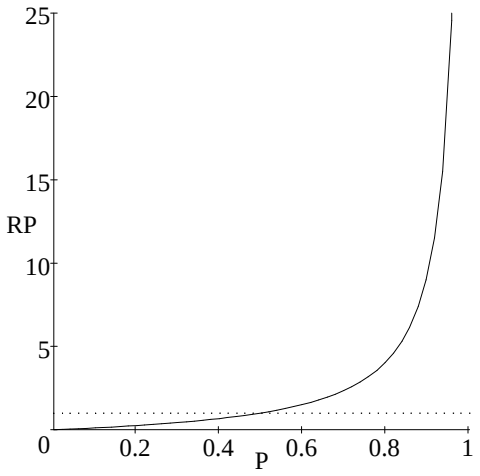


Figura 1.1: La razón de probabilidad $RP(X)$ como función de la probabilidad $P(+x)$.

Con las definiciones anteriores, la ecuación (1.25) puede expresarse como

$$RP_{post}(X) = RP_{pre}(X) \cdot RV_X(y) \tag{1.30}$$

y una vez conocida $RP_{post}(X)$ se obtiene $P(+x \mid y)$ a partir de la ecuación (1.29).

Ejemplo 1.38 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

(1.31)

y nos piden que calculemos mediante la ecuación (1.30) 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,[3]

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

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

$$P(+x \mid y^a) = \frac{RP_{post}(X)}{1 + RP_{post}(X)} = \frac{0'0059}{1 + 0'0059} = 0'0058 \tag{1.32}$$

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 a 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:

$$VPP = \frac{Sens \times Prev}{Sens \times Prev + (1 - Espec) \times (1 - Prev)} \tag{1.33}$$

$$VPN = \frac{Espec \times (1 - Prev)}{(1 - Sens) \times Prev + Espec \times (1 - Prev)} \tag{1.34}$$

En la figura 1.2 se observa que el valor predictivo positivo aumenta considerablemente al aumentar la especificidad y muy levemente al aumentar la sensibilidad; de hecho, $VPP = 1$ sólo si la especificidad vale 1; por tanto, para confirmar la presencia de una enfermedad deberemos buscar pruebas muy específicas. Análogamente, en la figura 1.3 se observa que el valor predictivo negativo aumenta al aumentar la sensibilidad, por lo que para descartar una enfermedad con certeza deberemos buscar síntomas o signos muy sensibles.

[3]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.

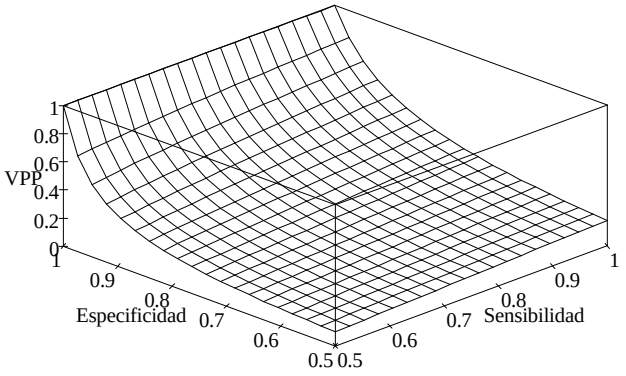


Figura 1.2: Valor predictivo positivo (prevalencia=0'1).

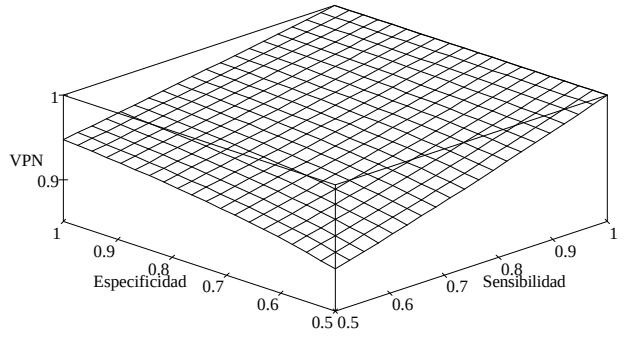


Figura 1.3: Valor predictivo negativo (prevalencia=0'1).

1.2. Método bayesiano ingenuo

Acabamos de explicar cómo puede aplicarse 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.) y por ello tenemos que ser capaces de aplicar el teorema de Bayes en 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. (1.19)) nos dice entonces

que

$$P(d_1, \ldots, d_n \,|\, h_1, \ldots, h_m) = \frac{P(d_1, \ldots, d_n) \cdot P(h_1, \ldots, h_m \,|\, d_1, \ldots, d_n)}{\sum_{d'_1, \ldots, d'_n} P(d'_1, \ldots, d'_n) \cdot P(h_1, \ldots, h_m \,|\, d'_1, \ldots, d'_n)} \tag{1.35}$$

Sin embargo, esta expresión es imposible de aplicar por la enorme cantidad de información que requiere: necesitaríamos conocer todas las probabilidades a priori $P(\mathbf{d})$ y todas las probabilidades condicionales $P(\mathbf{h} \,|\, \mathbf{d})$. 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.[4] 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 n diagnósticos posibles), de modo que la probabilidad de un diagnóstico cualquiera d viene dada por

$$P(d \,|\, h_1, \ldots, h_m) = \frac{P(d) \cdot P(h_1, \ldots, h_m \,|\, d)}{\sum_{d'} P(d') \cdot P(h_1, \ldots, h_m \,|\, d')} \tag{1.36}$$

Este modelo simplificado requiere n probabilidades a priori $P(d)$ y, si las variables H_j son binarias, $2^m \cdot n$ probabilidades condicionales $P(\mathbf{h} \,|\, d)$, 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:

$$\forall d,\ P(h_1, \ldots, h_m \,|\, d) = P(h_1 \,|\, d) \cdot \ldots \cdot P(h_m \,|\, d) \tag{1.37}$$

de modo que la probabilidad resultante para cada diagnóstico d es

$$P(d \,|\, h_1, \ldots, h_m) = \frac{P(d) \cdot P(h_1 \,|\, d) \cdot \ldots \cdot P(h_m \,|\, d)}{\sum_{d'} P(d') \cdot P(h_1 \,|\, d') \cdot \ldots \cdot P(h_m \,|\, d')} \tag{1.38}$$

o, en forma normalizada

$$P(d \,|\, h_1, \ldots, h_m) = \alpha \cdot P(d) \cdot P(h_1 \,|\, d) \cdot \ldots \cdot P(h_m \,|\, d) \tag{1.39}$$

Observe que esta expresión es una generalización de la ecuación (1.23).

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

Este modelo simplificado requiere n probabilidades a priori $P(d)$ y, si las variables H_j son binarias, $2\,m \cdot n$ probabilidades condicionales $P(h_j \mid d)$, 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.

De acuerdo con la representación gráfica de independencia probabilista, que estudiaremos en la sección 1.5, el modelo bayesiano ingenuo se corresponde con el grafo de la figura 1.4.

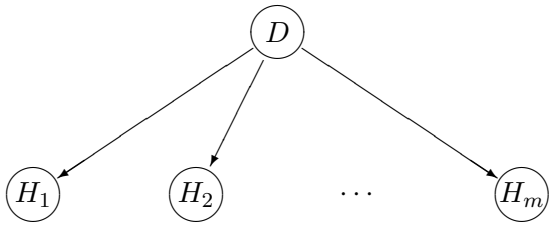


Figura 1.4: Representación del método probabilista ingenuo mediante un grafo de independencia.

Ejemplo 1.39 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 bayesiano ingenuo**.** 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 bayesiano ingenuo (fig. 1.5), que en su forma normalizada nos dice que

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

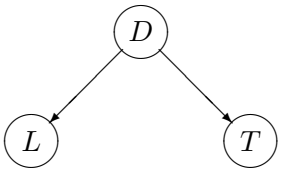


Figura 1.5: El piloto luminoso (L) y la temperatura (T) son signos de avería (D).

Concretamente, para la pregunta planteada en el problema

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

y, sustituyendo los valores numéricos,

$$\begin{cases} P(d^e \mid +l, t^r) = \alpha \cdot 0'001 \cdot 0'95 \cdot 0'01 = 0'0000095\alpha = 0'70423 \\ P(d^m \mid +l, t^r) = \alpha \cdot 0'00001 \cdot 0'99 \cdot 0'40 = 0'00000396\alpha = 0'29355 \\ P(d^n \mid +l, t^r) = \alpha \cdot 0'99899 \cdot 0'000001 \cdot 0'03 = 0'00000002997\alpha = 0'00222 \end{cases}$$

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%).

1.2.1. Forma racional del método bayesiano ingenuo

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

$$\frac{P(d \mid h_1, \ldots, h_m)}{P(d' \mid h_1, \ldots, h_m)} = \frac{P(d)}{P(d')} \cdot \frac{P(h_1 \mid d)}{P(h_1 \mid d')} \cdot \ldots \cdot \frac{P(h_m \mid d)}{P(h_m \mid d')} \tag{1.40}$$

En el problema anterior (ejemplo 1.39), si sólo quisiéramos saber si es más probable que la avería sea eléctrica o mecánica, tendríamos

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

Esto nos permite comprobar que el hallazgo $+l$ casi no influye en el diagnóstico, pues el valor de $P(+l \mid d^e)/P(+l \mid d^m)$ 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 (1.26), (1.27) y (1.28) para calcular la razón de probabilidad de D dada la evidencia $\{h_1, \ldots, h_n\}$:

$$RP_{post}(D) = RP_{pre}(D) \cdot RV_D(h_1) \cdot \ldots \cdot RV_D(h_m) \tag{1.41}$$

Esta expresión es una generalización de la (1.30) para el caso de múltiples hallazgos. A partir de $RP_{post}(D)$ se obtiene fácilmente la probabilidad posteriori mediante la ecuación (1.29).

Finalmente, conviene señalar que en el método bayesiano ingenuo (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 (1.41) 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.

1.2.2. 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 [81], los de Gorry y Barnett [30, 31] y el programa creado por de Dombal y sus colaboradores para el diagnóstico del dolor abdominal agudo [16]. Aunque estos programas dieron resultados satisfactorios, el método bayesiano ingenuo fue duramente criticado, por los motivos siguientes:

1. La *hipótesis de diagnósticos exclusivos y exhaustivos* es pocas veces aplicable en casos reales [74]. En la mayor parte de los problemas de diagnóstico médico, por ejemplo, pueden darse dos enfermedades simultáneamente, con lo que el método ingenuo 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 ingenuo, es muy cuestionable [78]. 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.)

   Estas dos hipótesis son las que hacen que el método lleve el nombre de "ingenuo", pues se considera una *ingenuidad* pensar que tales hipótesis se cumplen en el mundo real.

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 bayesiano ingenuo 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 y la lógica difusa. Sin embargo, en la década de los 80, con el desarrollo de las redes bayesianas, resurgió el interés por los métodos probabilistas en inteligencia artificial, un interés que no ha dejado de aumentar desde entonces.

1.3. Nociones sobre grafos

1.3.1. Definiciones básicas

Un grafo está formado por un conjunto de nodos y enlaces. Cualquier objeto puede ser un nodo en un grafo. El conjunto de nodos de un grafo suele denotarse por $\mathcal{N}$. Los enlaces se definen así:

Definición 1.40 (Enlace) Dado un conjunto de nodos $\mathcal{N}$, un enlace es una terna de $\mathcal{N} \times \mathcal{N} \times \{\text{dirigido, no-dirigido}\}$; es decir, una terna de la forma (X, Y, d) donde X e Y son nodos de $\mathcal{N}$ y d indica si el enlace es dirigido o no. Un enlace $(X, Y,$ dirigido) puede denotarse también como $X \to Y$; en la representación gráfica se dibuja como una flecha desde X hasta Y. Un enlace $(X, Y,$ no-dirigido) puede denotarse también como $X - Y$ y en la representación gráfica se dibuja como una línea entre X e Y.

El término castellano "enlace" corresponde al inglés "link", y en el contexto de los modelos gráficos probabilistas es sinónimo de "arco" (en inglés, "arc"). En el resto de este libro vamos a suponer siempre que los dos nodos que forman un enlace son diferentes.

Definición 1.41 (Grafo) Es un par $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ donde $\mathcal{N}$ es un conjunto de nodos y $\mathcal{A}$ un conjunto de enlaces definidos sobre los nodos de $\mathcal{N}$.

Definición 1.42 (Grafo dirigido) Es aquél cuyos enlaces son todos dirigidos.

Definición 1.43 (Grafo no dirigido) Es aquél cuyos enlaces son todos no dirigidos.

Definición 1.44 (Grafo mixto o híbrido) Es aquél que tiene tanto enlaces dirigidos como no dirigidos.

Más adelante veremos que las redes bayesianas, los diagramas de influencia y los árboles de decisión se basan en grafos dirigidos, mientras que las redes de Markov se basan en grafos no dirigidos.

Definición 1.45 (Camino) Un *camino* es un conjunto ordenado de enlaces, complementado con una ordenación de los nodos de cada enlace tal que el segundo nodo de cada enlace coincide con el primero del siguiente. Dado un enlace dirigido de un camino, si el orden de los nodos en el camino es el mismo que en el enlace, se dice que el camino atraviesa el enlace *hacia delante*, en caso contrario, decimos que lo atraviesa hacia atrás.

Ejemplo 1.46 Uno de los caminos del grafo de la figura 1.6 es $A \rightarrow D \leftarrow B$, que está definido por los enlaces $A \rightarrow D$ y $B \rightarrow D$ más la ordenación $\{\{A, D\}, \{D, B\}\}$. Este camino atraviesa el enlace $A \rightarrow D$ hacia delante y el enlace $B \rightarrow D$ hacia atrás.

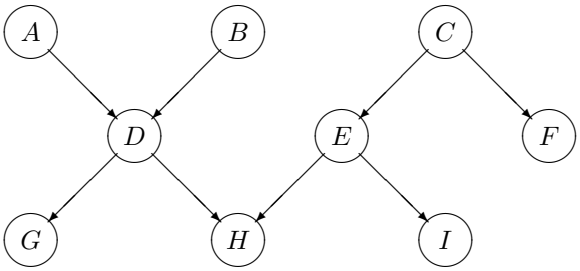


Figura 1.6: Un grafo dirigido.

Definición 1.47 (Camino dirigido) Un camino es *dirigido* cuando todos sus enlaces son dirigidos y los atraviesa todos hacia delante.

Ejercicio 1.48 ¿Cuántos caminos dirigidos contiene el grafo de la figura 1.6? Sugerencia: cuente primero los caminos de un solo enlace y luego los de dos. Observe que este grafo no contiene ningún camino dirigido de longitud mayor que dos.

Definición 1.49 (Padre) X es un *padre* de Y si y sólo si existe un arco $X \rightarrow Y$.

Los padres de X se representan como $Pa(X)$. Por semejanza con el convenio utilizado para variables y sus valores, $pa(X)$ representará la configuración de $Pa(X)$ formada al asignar un valor a cada padre de X.

Definición 1.50 (Hijo) Y es un *hijo* de X si y sólo si existe un arco $X \rightarrow Y$.

Definición 1.51 (Hermano) X es un *hermano* de Y si y sólo si existe un arco $X - Y$.

Definición 1.52 (Antepasado) X es un *antepasado* de Y si y sólo si existe (al menos) un camino dirigido desde X hasta Y.

Definición 1.53 (Descendiente) Y es un *descendiente* de X si y sólo si X es un antepasado de Y.

Observe que si X es padre de Y entonces existe un enlace $X \to Y$, lo cual implica que X es antepasado de Y.

Ejemplo 1.54 En el grafo de la figura 1.6 los padres de D son A y B: $Pa(D) = \{A, B\}$, y son sus únicos antepasados. El nodo A es antepasado de G porque existe un camino $A \to D \to G$. Los hijos de D son G y H, y son sus únicos descendientes. Los descendientes de B son D, G y H. Los antepasados de H son A, B, C, D y E.

Definición 1.55 (Camino abierto / cerrado) Si el segundo nodo del último enlace de un camino es el mismo que el primer nodo del primer enlace, se dice que el camino es *cerrado*. Si no, se dice que es *abierto*.

Ejemplo 1.56 $X \to Y \to Z$ es un camino abierto, mientras que $X \to Y \to Z \to X$ es un camino cerrado.

En un camino cerrado, si movemos el primer enlace al final y desplazamos el i-ésimo enlace a la posición anterior, el nuevo camino se considera igual al primero. Por ejemplo, los tres caminos $X \to Y \to Z \to X$, $Y \to Z \to X \to Y$, y $Z \to X \to Y \to Z$, se consideran iguales, es decir, se consideran como tres representaciones equivalentes del mismo camino.

Definición 1.57 (Ciclo) Un camino cerrado que atraviesa todos sus enlaces dirigidos hacia delante es un *ciclo*.

Definición 1.58 (Bucle) Un *bucle* es todo camino cerrado que no es un ciclo.

La figura 1.7 muestra la diferencia entre ciclos y bucles. Cada uno de los tres grafos superiores contiene un ciclo, es decir, un camino cerrado que atraviesa los tres nodos cruzando sus enlaces dirigidos (en caso de que los tenga) hacia delante. En cambio, cada grafo de la fila inferior contiene dos caminos cerrados (uno en el sentido *A*-*B*-*C* y otro en sentido contrario) pero cada uno de ellos atraviesa al menos uno de sus enlaces dirigidos hacia atrás; por tanto esos dos grafos contienen bucles pero no ciclos.

Aunque esta diferencia pueda parecer sutil, en realidad es muy importante. Imagine, por ejemplo, que estamos haciendo un razonamiento lógico y dibujamos un grafo dirigido en que cada nodo representa una proposición y un enlace desde la proposición p hasta la proposición q indica que q se deduce de p. Los nodos que no tienen padres se consideran como axiomas. Si este grafo es acíclico, todas las proposiciones son ciertas siempre que las premisas sean verdaderas. En cambio, si tuviéramos un ciclo el razonamiento no sería válido. Por ejemplo, si q se deduce de p y p se deduce de q, entonces el razonamiento no tiene validez.

Una propiedad importante, que veremos enseguida (véase la definición 1.68 y las proposiciones que la siguen) es que todo *grafo dirigido acíclico* (GDA; es decir, un grafo en que todos los enlaces son dirigidos y no contiene ciclos) induce una relación de orden parcial estricta

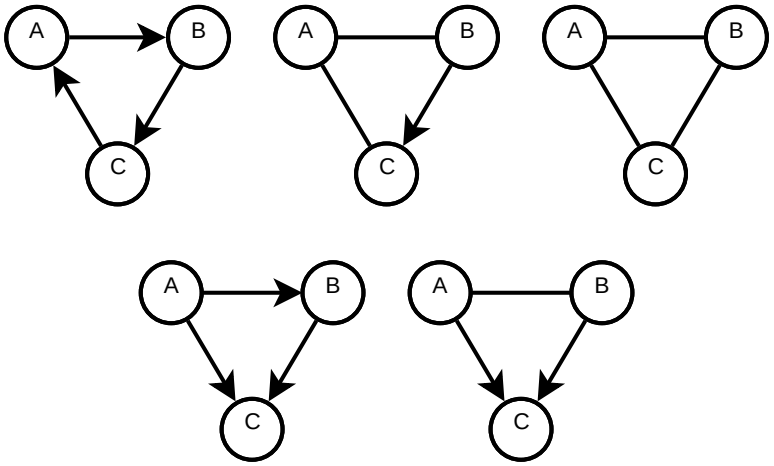


Figura 1.7: Caminos cerrados: tres ciclos (fila superior) y dos bucles (fila inferior).

entre los nodos,[5] denominada *ordenación ancestral*, mientras que en un grafo con ciclos no se puede establecer una ordenación de este tipo, porque se violaría la propiedad transitiva.

En cuanto al tema que nos ocupa, la diferencia entre ciclo y bucle es fundamental, porque la mayor parte de los modelos gráficos probabilistas, como las redes bayesianas, los diagramas de influencia y varios tipos de modelos de decisión de Markov, se definen sobre grafos dirigidos *acíclicos*. Si se definieran sobre grafos con ciclos, muchas de las propiedades matemáticas de estos modelos desaparecerían.

Por eso vamos a estudiar los GDA's en la sección siguiente con mayor detalle. Pero antes vamos a introducir tres definiciones que nos serán útiles más adelante.

Definición 1.59 (Grafo conexo) Un grafo es *conexo* si entre dos cualesquiera de sus nodos hay al menos un camino.

Definición 1.60 (Grafo simplemente conexo) Un grafo es *simplemente conexo* si entre cada par de nodos existe un único camino.

Definición 1.61 (Grafo múltiplemente conexo) Un grafo es *múltiplemente conexo* si contiene al menos un par de nodos entre los cuales hay más de un camino.

Ejemplo 1.62 El grafo de la figura 1.6 es simplemente conexo. Los cinco grafos de la figura 1.7 son múltiplemente conexos.

Ejercicio 1.63 Demuestre que si en un grafo simplemente conexo se elimina un enlace, deja de ser conexo.

1.3.2. Grafos dirigidos acíclicos

En el contexto de los GDA's, suelen establecerse las siguientes definiciones.

[5] Una *relación de orden parcial estricta* (por ejemplo, "mayor que") tiene las propiedades anti-reflexiva, anti-simétrica y transitiva. En cambio, una *relación de orden parcial no estricta* (por ejemplo, "mayor o igual que") tiene las propiedades reflexiva, anti-simétrica y transitiva.

Definición 1.64 (Familia) Es el conjunto formado por X y los padres de X: $Fam(X) = \{X\} \cup Pa(X)$.

Definición 1.65 (Nodo terminal) Es el nodo que no tiene hijos.

Definición 1.66 (Poliárbol) Un *poliárbol* es un *grafo dirigido simplemente conexo*, es decir, un grafo dirigido que no contiene ciclos ni bucles.

Definición 1.67 (Árbol) Es un caso particular de poliárbol, en que cada nodo tiene un sólo padre, excepto el *nodo raíz*, que no tiene padres. Los nodos terminales de un árbol se denominan *hojas*.

El grafo de la figura 1.6 es un poliárbol, porque no contiene bucles; no es un árbol porque algunos de sus nodos (D y H) tienen más de un padre. Las nueve familias (tantas como nodos) son $\{A\}$, $\{B\}$, $\{C\}$, $\{D, A, B\}$, $\{E, C\}$, $\{F, C\}$, $\{G, D\}$, $\{H, D, E\}$ e $\{I, E\}$.

Definición 1.68 (Ordenación ancestral de un GDA) Es aquélla en que todos los antepasados de cada nodo son mayores que él.

Ejemplo 1.69 Algunas de las posibles ordenaciones ancestrales del grafo de la figura 1.6 son las siguientes:

- $A > B > C > D > E > F > G > H > I$
- $B > A > D > G > C > E > I > F > H$

Ejemplo 1.70 En la figura 1.7 se muestran dos grafos dirigidos (el que está a la izquierda en cada una de las dos filas). El primero de ellos no admite una ordenación ancestral porque contiene un ciclo. El segundo admite una única ordenación ancestral: $A > B > C$.

Vamos a demostrar ahora que todo GDA tiene al menos un orden ancestral, pero para ello necesitamos previamente la siguiente proposición.

Proposición 1.71 En todo GDA existe al menos un nodo sin padres.

Demostración. Vamos a dar una demostración algorítmica, es decir, demostraremos la existencia de ese nodo dando un procedimiento para encontrarlo.

Escogemos un nodo cualquiera y lo metemos en una lista. Procedemos recursivamente tomando el último nodo de la lista; si este nodo tiene padres, seleccionando uno cualquiera de ellos y añadiéndolo a la lista. En cada iteración hay un camino dirigido que pasa por todos los nodos de la lista, desde el último hasta el primero.

El algoritmo termina cuando el último nodo de la lista no tiene padres, y siempre termina, porque ningún nodo puede aparecer dos veces en la lista (en ese caso el grafo tendría al menos un ciclo) y el número de nodos en el grafo es finito. □

Proposición 1.72 Todo GDA tiene al menos un orden ancestral.

Demostración. Formamos una lista de nodos recursivamente sacando del grafo un nodo sin padres (es necesario borrar previamente todos los enlaces desde ese nodo hacia sus hijos) y metiéndolo en la lista. La proposición anterior nos garantiza que siempre va a haber al menos un nodo para meter en la lista. Claramente la lista está en orden ancestral, porque ningún nodo ha entrado en la lista antes que sus padres. □

Proposición 1.73 Cuando los nodos de un grafo están numerados, por ejemplo, $\{X_1, \ldots, X_n\}$, existe una permutación σ tal que $X_{\sigma(1)} > \ldots > X_{\sigma(n)}$ es una ordenación ancestral.

Demostración. La permutación se puede construir repitiendo el procedimiento de la demostración anterior: si el primer nodo metido en la lista es X_i, hacemos $\sigma(1) = i$; si el segundo nodo es X_j hacemos $\sigma(2) = j$, y así sucesivamente. □

1.4. Definición de red bayesiana

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

Definición 1.74 (Red bayesiana) Una *red bayesiana* consta de tres elementos: un conjunto de variable aleatorias, $\mathbf{X}$; un grafo dirigido acíclico (GDA) $\mathcal{G} = (\mathbf{X}, \mathcal{A})$, en que cada nodo representa una variable X_i; y una distribución de probabilidad sobre $\mathbf{X}$, $P(\mathbf{X})$, que puede ser factorizada así:

$$P(\mathbf{x}) = \prod_i P(x_i|pa(X_i)) \tag{1.42}$$

donde las $P(x_i|pa(X_i))$ son las probabilidades condicionales que se obtienen a partir de $P(\mathbf{x})$.

Observe que en esta ecuación tenemos una probabilidad condicional por cada nodo del grafo: el nodo X_i es la variable condicionada y sus padres son las variables condicionantes. Por tanto, la definición de red bayesiana establece una relación entre el grafo y la distribución de probabilidad, pues es el grafo el que determina cómo se factoriza la probabilidad. Visto de otra manera, por cada forma en que puede factorizarse una distribución de probabilidad tenemos una red bayesiana diferente, cada una con su propio GDA.

Ejemplo 1.75 El grafo de la figura 1.6 dicta la siguiente factorización de la probabilidad:

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

Ejemplo 1.76 Como ya habrá observado, el método bayesiano ingenuo es un caso particular de red bayesiana, pues el grafo de la figura 1.4 corresponde a la factorización

$$P(d, h_1, \ldots, h_m) = P(d) \cdot P(h_1 \,|\, d) \cdot \ldots \cdot P(h_m \,|\, d) \tag{1.43}$$

de donde se deducen las ecuaciones (1.37) y (1.38).

En concreto, para el ejemplo 1.39 (cf. figura 1.5), la factorización es

$$P(d, l, t) = P(d) \cdot P(l \,|\, d) \cdot P(t \,|\, d)$$

1.4.1. Construcción de una red bayesiana

En la definición de red bayesiana hemos dado por supuesto que conocíamos la probabilidad conjunta $P(\mathbf{x})$, a partir de la cual se pueden calcular todas las probabilidades marginales y condicionales. Sin embargo, el número de valores de $P(\mathbf{x})$ crece exponencialmente con el número de variables. Por eso en la práctica lo que se hace es partir de un conjunto de probabilidades condicionales, a partir de las cuales podremos calcular $P(\mathbf{x})\ldots$ si tenemos

suficientes recursos computacionales. (En el próximo capítulo estudiaremos algoritmos que en muchos problemas no triviales nos permitirán calcular las probabilidades condicionales sin tener que calcular $P(\mathbf{x})$ explícitamente.) Veamos ahora que la construcción de una red bayesiana a partir de las probabibilidades condicionales está justificada.

Teorema 1.77 (Contrucción de una red bayesiana) Sea un conjunto de variables $\mathbf{X}$, un GDA $\mathcal{G}$ en que cada nodo representa una variable de $\mathbf{X}$ y un conjunto de distribuciones de probabilidad $\{P_c(x_i|pa(X_i))\}$, es decir, una distribución por cada variable X_i y por cada configuración de sus padres en el grafo. La función $P(\mathbf{x})$, definida por

$$P(\mathbf{x}) = \prod_i P_c(x_i|pa(X_i)) \tag{1.44}$$

es una distribución de probabilidad. Se cumple además que

$$\forall i, \forall x_i, \forall pa(X_i),\ P(x_i|pa(X_i)) = P_c(x_i|pa(X_i)) \tag{1.45}$$

□

Antes de demostrar el teorema vamos a explicar su sentido. En primer lugar, la ecuación (1.44) se diferencia de la (1.42) en que las probabilidades condicionales que intervienen son las que hemos definido arbitrariamente —el subíndice c significa que son las probabilidades que hemos utilizado para construir la red— mientras que las probabilidades que aparecen en la (1.42) son las que se derivan de $P(\mathbf{x})$, es decir, $P(x_i|pa(X_i)) = P(x_i, pa(X_i))/P(pa(X_i))$, donde $P(x_i, pa(X_i))$ y $P(pa(X_i))$ son distribuciones marginales de $P(\mathbf{x})$.

La ecuación (1.45), por su parte, nos dice que las probabilidades condicionales que se deducen de $P(\mathbf{x})$ son las mismas que hemos utilizado para construir $P(\mathbf{x})$. Este resultado puede parecer una trivialidad, pero no lo es. De hecho, si aplicamos la ecuación (1.44) tomando un grafo cíclico y unas distribuciones de probabilidad P_c escogidas al azar, es casi seguro que la función $P(\mathbf{x})$ resultante no será una distribución de probabilidad ni se cumplirá la ecuación (1.45). El lector puede hacer la prueba por sí mismo con una red de dos variables y dos enlaces, $A \to B$ y $B \to A$, o bien con una red de tres variables y tres enlaces, $A \to B$, $B \to C$ y $C \to A$.

Demostración. Vamos a demostrar primero que $P(\mathbf{x})$ es una distribución de probabilidad. Es evidente que $P(\mathbf{x})$ tiene un valor no negativo para toda configuración $\mathbf{x}$. Para demostrar que sus valores suman la unidad, suponemos, sin pérdida de generalidad, que las variables de $\mathbf{X}$ están en orden ancestral,[6] y hacemos la suma de esta manera:

$$\sum_{\mathbf{x}} P(\mathbf{x}) = \sum_{x_1} \cdots \sum_{x_n} \prod_{i=1}^{n} P_c(x_i|pa(X_i))$$

Para toda i distinta de n, los padres de X_i han sido numerados antes que X_i —por ser una ordenación ancestral— y como $n > i$, X_n no puede pertenecer a $Pa(X_i)$. Por tanto $P_c(x_i|pa(X_i))$, con $i \neq n$, no depende de x_n y puede salir fuera sumatorio sobre x_n como

[6] Si las variables no estuvieran en orden ancestral, tendríamos que construir la permutación σ indicada en la proposición 1.73 y sustituir cada subíndice i que aparece en la demostración por $\sigma(i)$.

factor común. En consecuencia,

$$\sum_{\mathbf{x}} P(\mathbf{x}) = \sum_{x_1} \cdots \sum_{x_{n-1}} \left(\prod_{i=1}^{n-1} P_c(x_i|pa(X_i)) \right) \sum_{x_n} P_c(x_n|pa(X_n))$$
$$= \sum_{x_1} \cdots \sum_{x_{n-1}} \left(\prod_{i=1}^{n-1} P_c(x_i|pa(X_i)) \right)$$

porque las probabilidades condicionales de X_n suman la unidad. Observe que en esta expresión hemos eliminado la variable X_n. Es como si tuviéramos una nueva red bayesiana cuyas variables ya no son $\{X_1, \ldots, X_n\}$, sino $\{X_1, \ldots, X_{n-1}\}$.

Ahora vamos a realizar la suma sobre x_{n-1}. Como en el caso anterior, podemos sacar como factor común todas las probabilidades correspondientes a los nodos X_i con $i < n-1$, porque ninguna de ellas depende de X_{n-1}. Así obtenemos una expresión en que sólo aparecen las variables $\{X_1, \ldots, X_{n-2}\}$. Repitiendo la operación sucesivamente llegamos a

$$\sum_{\mathbf{x}} P(\mathbf{x}) = \sum_{x_1} P_c(x_1) = 1$$

Para demostrar la ecuación (1.45), vamos a calcular primero $P(x_i, anc(X_i))$, es decir, la probabilidad de una configuración cualquiera de $\{X_i\} \cup Anc(X_i)$. Definimos ahora el conjunto $\mathbf{Z}$, formado por los nodos que no son ni X_i ni sus antepasados: $\mathbf{Z} = \mathbf{X} \backslash (\{X_i\} \cup Anc(X_i))$. Si $\mathbf{Z} = \varnothing$, entonces $\{X_i\} \cup Anc(X_i) = \mathbf{X}$ y por la ecuación (1.44) tenemos:

$$P(x_i, anc(X_i)) = P_c(x_i|pa(X_i)) \prod_{j \mid X_j \in Anc(X_i)} P_c(x_j|pa(X_j)) \tag{1.46}$$

Si $\mathbf{Z} \neq \varnothing$, calculamos $P(x_i, anc(X_i))$ de este modo:

$$P(x_i, anc(X_i)) = \sum_{\mathbf{z}} P(\mathbf{x}) = \sum_{\mathbf{z}} \prod_j P_c(x_j|pa(X_j))$$

De esta ecuación vamos a eliminar una a una todas las variables de $\mathbf{Z}$, como hemos hecho anteriormente en la primera parte de esta demostración. Para ello, escogemos un nodo Z_k de $\mathbf{Z}$ que no tenga hijos.[7] Como Z_k no es padre de ningún nodo, las probabilidades condicionales de los demás nodos de la red no dependen de Z_k y podemos sacarlas como factor común de $\sum_{z_k} P_c(z_k|pa(Z_k))$, que vale 1. Así podemos ir eliminando todas las probabilidades condicionales de los nodos de $\mathbf{Z}$, hasta llegar a la ecuación 1.46, en la cual sólo aparecen las probabilidades X_i y de sus antepasados. Es como si tuviéramos una red bayesiana de la cual hubiéramos eliminado los nodos de $\mathbf{Z}$.

Por tanto, la ecuación (1.46) se cumple en todos los casos, tanto si $\mathbf{Z} = \varnothing$ como si $\mathbf{Z} \neq \varnothing$.

Definimos ahora el subconjunto $\mathbf{Y} = Anc(X_i) \backslash Pa(X_i)$, formado por los antepasados de X_i, excepto sus padres, de modo que

$$P(x_i, pa(X_i)) = \sum_{\mathbf{y}} P(x_i, \underbrace{pa(X_i), \mathbf{y}}_{anc(X_i)})$$

[7] Del mismo que en un GDA siempre hay al menos un nodo que no tiene padres, también hay al menos un nodo que no tiene hijos (la demostración es similar a la de la proposición 7). Si el único nodo sin hijos fuera X_i, todos los demás nodos serían antepasados de X_i y $\mathbf{Z}$ estaría vacío. Por tanto, cuando $\mathbf{Z}$ no está vacío debe contener algún nodo sin hijos.

Aplicamos ahora la ecuación (1.46), y teniendo en cuenta que $P_c(x_i|pa(X_i))$ no depende de $\mathbf{Y}$, porque $Pa(X_i) \cap \mathbf{Y} = \varnothing$, llegamos a

$$P(x_i, pa(X_i)) = P_c(x_i|pa(X_i)) \sum_{\mathbf{y}} \prod_{j \mid X_j \in Anc(X_i)} P_c(x_j|pa(X_j))$$

A partir de aquí calculamos $P(pa(X_i))$:

$$\begin{aligned} P(pa(X_i)) &= \sum_{x_i} P(x_i, pa(X_i)) = \left(\sum_{x_i} P_c(x_i|pa(X_i)) \right) \sum_{\mathbf{y}} \prod_{j \mid X_j \in Anc(X_i)} P_c(x_j|pa(X_j)) \\ &= \sum_{\mathbf{y}} \prod_{j \mid X_j \in Anc(X_i)} P_c(x_j|pa(X_j)) \end{aligned}$$

Finalmente,

$$P(x_i|pa(X_i)) = \frac{P(x_i, pa(X_i))}{P(pa(X_i))} = P_c(x_i|pa(X_i))$$

con lo que concluye la demostración. □

1.4.2. Propiedad de Markov

Hemos definido las redes bayesianas a partir de una distribución de probabilidad conjunta que puede ser factorizada de acuerdo con un GDA. Ahora vamos a ver otra propiedad de las redes bayesianas que también podría haberse utilizado para la definición de red bayesiana, pues como vamos a ver más adelante, es equivalente a la propiedad de factorización de la probabilidad.

Definición 1.78 (Propiedad de Markov) Una terna $(\mathbf{X}, \mathcal{G}, P)$ formada por un conjunto de variable aleatorias $\mathbf{X}$, un GDA $\mathcal{G} = (\mathbf{X}, \mathcal{A})$ en que cada nodo representa una variable X_i y una distribución de probabilidad sobre $\mathbf{X}$, $P(\mathbf{X})$, cumple la *propiedad de Markov* si y sólo si para todo nodo X, el conjunto de sus padres, $Pa(X)$, separa condicionalmente este nodo de todo subconjunto $\mathbf{Y}$ en que no haya descendientes de X. Es decir,

$$P(x|pa(X), \mathbf{y}) = P(x|pa(X)) \tag{1.47}$$

Esta definición puede resultar difícil de entender a primera vista. Para comprender mejor su significado, vamos a sacar algunas conclusiones que se deducen de ella. Las ilustraremos aplicándolas a algunos nodos del grafo que vimos en la figura 1.6:

1. Sean dos nodos X e Y sin padres. Como Y no es descendiente de X, se cumple que $P(x|pa(X), y) = P(x|pa(X))$. Ahora bien, $Pa(x) = \varnothing$, y por tanto $P(x|y) = P(x)$. Es decir, si una terna $(\mathbf{X}, \mathcal{G}, P)$ cumple la propiedad de Markov, todo par de nodos sin padres en el grafo son independientes a priori. Por ejemplo, en el grafo de la figura 1.6, los nodos (variables) A, B y C son independientes dos a dos.

2. Generalizando la propiedad anterior, podemos afirmar que todo nodo sin padres es independiente de sus no-descendientes. En la figura 1.6, A es independiente de B, C, E, F e I y de cualquier combinación de ellos. Por ejemplo, $P(a|b, e, f) = P(a)$.

3. Dos nodos que no tienen ningún antepasado común son independientes a priori. No vamos a demostrar esta propiedad para el caso general, sino sólo para los nodos D y E de la figura 1.6. Calculamos $P(d, e)$ aplicando la regla de la cadena:

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

Como E no es descendiente de D y los padres de D son A y B, por la propiedad de Markov tenemos que $P(d|a,b,e) = P(d|a,b)$. Como A y B no son descendientes de E y E no tiene padres, por la propiedad de Markov tenemos que $P(e|a,b) = P(e)$. Por tanto

$$P(d,e) = \left(\sum_a \sum_b P(d|a,b) \cdot P(a,b) \right) \cdot P(e) = \left(\sum_a \sum_b P(d,a,b) \right) \cdot P(e) = P(d) \cdot P(e)$$

lo cual demuestra que D y E son independientes a priori.

4. Si D es descendiente de A y antepasado de H, y no existe ningún otro camino desde A hasta H (véase la figura 1.6), entonces estos dos nodos quedan condicionalmente separados por D:

$$P(h|d,a) = P(h|d) \tag{1.48}$$

5. Si tanto G como H son hijos de D y no tienen ningún otro antepasado común (véase la figura 1.6), este último separa G y H, haciendo que sean condicionalmente independientes:

$$P(g|d,h) = P(g|d) \tag{1.49}$$

Ejercicio 1.79 Demostrar las propiedades 3, 4 y 5 para el caso general.

En general, la independencia (a priori o condicional) de dos nodos se pierde al conocer el valor de cualquiera de sus descendientes comunes. Volviendo a la figura 1.6, teníamos que, por la propiedad de Markov, A y E son independientes a priori: $P(a,e) = P(a) \cdot P(e)$. En cambio, si queremos demostrar $P(a,e|h) \stackrel{?}{=} P(a|h) \cdot P(e|h)$ —observe que H es descendiente tanto de A como de E— no se puede aplicar la propiedad de Markov. De hecho, en general esta igualdad no se cumple; es decir, A y E estarán correlacionados dado H, salvo en los casos excepcionales en que la distribución de probabilidad P tome unos valores numéricos concretos.

La propiedad de Markov está íntimamente relacionada con las redes bayesianas por los siguientes teoremas:

Teorema 1.80 Toda terna $(\mathbf{X}, \mathcal{G}, P)$ que cumple la propiedad de Markov constituye una red bayesiana.

Demostración. Por la regla de la cadena, cualquier distribución de probabilidad sobre $\mathbf{X} = \{X_1, \ldots, X_n\}$ puede ser factorizada así:

$$P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i|x_1, \ldots, x_{i-1})$$

Si los nodos están en orden ancestral, $\{X_1, \ldots, X_{i-1}\}$ contiene todos los padres de X_i y otros nodos que no son descendientes de X_i. Entonces, por la propiedad de Markov,

$$P(x_i|x_1, \ldots, x_{i-1}) = P(x_i|pa(X_i))$$

de modo que la distribución de probabilidad conjunta puede ser factorizada de acuerdo con la definición de red bayesiana (cf. ec. (1.42)). □

Teorema 1.81 Toda terna $(\mathbf{X},\mathcal{G}, P)$ que constituye una red bayesiana cumple la propiedad de Markov.

Para demostrar este teorema partiríamos del hecho de que la probabilidad P puede ser factorizada según el grafo $\mathcal{G}$, tomaríamos un nodo cualquiera X y un conjunto $\mathbf{Y}$ de nodos no descendientes de X y, procediendo de modo similar a la demostración del teorema 1.77, concluiríamos que se cumple la ecuación (1.47), que define la propiedad de Markov.

Uniendo estos dos teoremas recíprocos, podemos concluir que para toda terna $(\mathbf{X},\mathcal{G}, P)$ hay dos propiedades equivalentes: la factorización de la probabilidad, dada por la ecuación (1.42), y la propiedad de Markov, dada por la ecuación (1.47). Nosotros hemos definido las redes bayesianas a partir de la primera y de esta definición hemos deducido que toda red bayesiana cumple la propiedad de Markov. Alternativamente, podríamos haberlas definido a partir de la propiedad de Markov y luego haber demostrado que la probabilidad de una red bayesiana siempre puede ser factorizada de acuerdo con su GDA.

En la próxima sección vamos a presentar una tercera propiedad, equivalente a las dos anteriores, que también podría servir para definir las redes bayesianas.

1.5. Grafos de dependencias e independencias probabilistas

1.5.1. Separación en grafos dirigidos y no dirigidos

Introducimos ahora la definición de camino activo, distinguiendo dos casos: los grafos no dirigidos y los grafos dirigidos.

Definición 1.82 (Camino activo en un grafo no dirigido) Sea un grafo no dirigido $\mathcal{G}$, dos nodos A y B de $\mathcal{G}$ y un subconjunto $\mathbf{C}$ de nodos de $\mathcal{G}$ tal que ni A ni B pertenecen a $\mathbf{C}$.

- Un camino de dos nodos, es decir, A–B (un solo enlace), siempre está *activo*.
- Un camino de n nodos, es decir A–X_1–…–X_{n-2}–B está *activo* cuando ningún nodo entre A y B pertenece a $\mathbf{C}$, es decir, $\{X_1, \ldots, X_{n-2}\} \cap \mathbf{C} = \varnothing$; si algún X_i pertenece a $\mathbf{C}$ se dice que el camino está *inactivo*.

Definición 1.83 (Camino activo en un grafo dirigido) Sea un grafo dirigido $\mathcal{G}$, dos nodos A y B de $\mathcal{G}$ y un subconjunto $\mathbf{C}$ de $\mathcal{G}$ tal que ni A ni B pertenecen a $\mathbf{C}$.

- Un camino de dos nodos, es decir, $A \to B$ o $B \to A$ (un solo enlace), siempre está *activo*.
- Un camino de tres nodos puede ser de varios tipos: $A \to X \to B$, $B \to X \to A$, $A \leftarrow X \to B$ o $A \to X \leftarrow B$.

- Un camino del tipo $A \rightarrow X \rightarrow B$ o $A \leftarrow X \rightarrow B$ está *activo* si X no pertenece a $\mathbf{C}$. Cuando X pertenece a $\mathbf{C}$ se dice que el camino está inactivo porque ha sido *bloqueado* por X.

- Un camino del tipo $A \rightarrow X \leftarrow B$ está *activo* si X o alguno de sus descendientes pertenece a $\mathbf{C}$. En este caso se dice que el camino ha sido *activado* por el nodo que pertenece a $\mathbf{C}$. Cuando ni X ni sus descendientes pertenecen a $\mathbf{C}$, el camino está inactivo.

- Un camino de n nodos está activo si cada par de enlaces consecutivos forman un camino activo.

Definición 1.84 (Nodos conectados) Sea un grafo dirigido o no dirigido $\mathcal{G}$ y un subconjunto $\mathbf{C}$ de $\mathcal{G}$. Dos nodos de $\mathcal{G}$ (no pertenecientes a $\mathbf{C}$) están *conectados dado* $\mathbf{C}$ cuando existe al menos un camino activo entre ellos. En otro caso se dice que están *separados dado* $\mathbf{C}$. La notación $I_{\mathcal{G}}(A, B|\mathbf{C})$ indica que A y B están separados y por tanto $\neg I_{\mathcal{G}}(A, B|\mathbf{C})$ indica que están conectados.

Definición 1.85 (Subconjuntos conectados) Sean tres subconjuntos de nodos $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$ disjuntos dos a dos. $\mathbf{A}$ y $\mathbf{B}$ están *conectados dado* $\mathbf{C}$ cuando al menos un nodo de $\mathbf{A}$ está conectado con un nodo de $\mathbf{B}$; se expresa mediante la notación $\neg I_{\mathcal{G}}(\mathbf{A}, \mathbf{B}|\mathbf{C})$. En otro caso se dice que están *separados*: $I_{\mathcal{G}}(\mathbf{A}, \mathbf{B}|\mathbf{C})$.

Observe que estas definiciones son simétricas; es decir, si un camino de A a B está activo, también el camino inverso, de B a A, está activo. Por tanto, A está conectado con B si y sólo si B está conectado con A. Tenga en cuenta también que $\mathbf{C}$ puede ser el conjunto vacío.[8]

Seguramente el lector se preguntará por el sentido de estas definiciones. La respuesta, como el lector probablemente ha intuido, está relacionada con las propiedades de independencia probabilista en redes bayesianas; para entender el porqué de estas definiciones le recomendamos que vea el vídeo "Separación en grafos", disponible en `http://www.ia.uned.es/~fjdiez/docencia/videos-prob-dec`. Observe, sin embargo, que en estas definiciones que hemos dado sólo interviene un grafo: ninguna de ellas hace referencia a un conjunto de variables ni a una distribución de probabilidad.

Antes de concluir esta sección, debemos comentar que la separación en grafos dirigidos se denomina *separación direccional* (en inglés, *d-separation* [63, 64]), porque la dirección de los enlaces es relevante. Por el mismo motivo, la separación en grafos no dirigidos se denomina a veces *separación no direccional* (en inglés, *u-separation*, donde la *u* significa *undirected*).

1.5.2. Mapas de independencias

Definición 1.86 Dada una terna $(\mathbf{X},\mathcal{G}, P)$ formada por un conjunto de variables aleatorias $\mathbf{X}$, un grafo dirigido o no dirigido $\mathcal{G} = (\mathbf{X}, \mathcal{A})$ y una distribución de probabilidad $P(\mathbf{X})$, el grafo es un *mapa de independencias* (en inglés *I-map*) de P si para todo trío de subconjuntos $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ de $\mathbf{X}$ disjuntos dos a dos se cumple

$$I_{\mathcal{G}}(\mathbf{A}, \mathbf{B}|\mathbf{C}) \Longrightarrow P(\mathbf{a}, \mathbf{b}|\mathbf{c}) = P(\mathbf{a}|\mathbf{c}) \cdot P(\mathbf{b}|\mathbf{c}) \tag{1.50}$$

[8] La definición 1.85 no exige formalmente que $\mathbf{A}$ y $\mathbf{B}$ sean no-vacíos; simplemente cuando uno de ellos es vacío entonces están separados. Sin embargo, en la práctica esta definición sólo es útil (no trivial) cuando $\mathbf{A}$ y $\mathbf{B}$ son no-vacíos.

A veces la independencia condicional se representa mediante $I_P(\mathbf{A}, \mathbf{B}|\mathbf{C})$, donde P indica la distribución de probabilidad; por tanto, la propiedad anterior se podría reescribir así:

$$I_{\mathcal{G}}(\mathbf{A}, \mathbf{B}|\mathbf{C}) \Longrightarrow I_P(\mathbf{A}, \mathbf{B}|\mathbf{C}) \tag{1.51}$$

La definición anterior significa que, cuando un grafo es un mapa de independencias de P, si dos subconjutos de variables están separados (dado $\mathbf{C}$) en el grafo entonces son condicionalmente independientes (dado $\mathbf{C}$) en sentido probabilista. Observe que la implicación recíproca no tiene por qué ser cierta: es posible que haya alguna relación de independencia en P que no esté reflejada en $\mathcal{G}$. Por tanto, $I_{\mathcal{G}}(\mathbf{A}, \mathbf{B}|\mathbf{C})$ significa que $\mathbf{A}$ y $\mathbf{B}$ **son** condicionalmente independientes dado $\mathbf{C}$, mientras que $\neg I_{\mathcal{G}}(\mathbf{A}, \mathbf{B}|\mathbf{C})$ sólo significa que $\mathbf{A}$ y $\mathbf{B}$ **pueden estar** correlacionados dado $\mathbf{C}$.

1.5.3. Separación direccional y redes bayesianas

La relación entre separación direccional y redes bayesianas viene dada por los dos teoremas siguientes, que son recíprocos:

Teorema 1.87 El grafo de una red bayesiana constituye un mapa de independencias de la distribución de probabilidad de dicha red.

Para demostrar este teorema hay que probar que si dos variables están separadas en el grafo dado $\mathbf{C}$ (lo cual significa, entre otras cosas, que ninguno de sus descendientes pertenece a $\mathbf{C}$) entonces esas variables son condicionalmente independientes dado $\mathbf{C}$. El método sería similar al empleado en la demostración del teorema 1.77.

Teorema 1.88 Una terna ($\mathbf{X}$,$\mathcal{G}, P$) formada un conjunto de variable aleatorias $\mathbf{X}$, un GDA $\mathcal{G} = (\mathbf{X}, \mathcal{A})$ y una distribución de probabilidad $P(\mathbf{X})$, tal que $\mathcal{G}$ es un *mapa de independencias* de P, constituye una red bayesiana.

Este teorema se demostraría probando, a partir de la propiedad del mapa de independencias y la definición de caminos activos e inactivos en grafos dirigidos, que la distribución de probabilidad P puede ser factorizada según $\mathcal{G}$.

En resumen, dada una terna $(\mathbf{X}, \mathcal{G}, P)$, donde $\mathcal{G}$ es un GDA, tenemos tres propiedades equivalentes:

- la distribución P puede ser factorizada según $\mathcal{G}$ (ecuación (1.42));
- la terna cumple la propiedad de Markov (ec. (1.47));
- $\mathcal{G}$ es un mapa de independencias de P (ec. (1.51)).

Cualquiera de las tres puede servir para definir una red bayesiana, y a partir de tal definición se pueden deducir las otras dos. Nosotros hemos basado la definición de red bayesiana en la primera de ellas, y luego hemos enunciado el teorema 1.81, que permite deducir la segunda, y el teorema 1.87, que permite deducir la tercera.

1.6. 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 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. En la próxima sección vamos a ver la diferencia entre la interpretación probabilista y la interpretación causal de un grafo, y en la siguiente mostraremos algunos ejemplos que pueden ayudarnos a entender mejor la diferencia y la relación entre ambos conceptos.

1.6.1. Interpretación probabilista e interpretación causal de un grafo

En la interpretación causal de un grafo dirigido (cíclico o acíclico) un enlace $A \to B$ significa que existe un mecanismo causal que hace que la presencia de A produzca B. Si se trata de un mecanismo determinista, la presencia de A siempre producirá B, mientras que si es no determinista puede ocurrir que A no siempre produzca B, sino sólo en algunos casos.

La interpretación probabilista de un grafo dirigido generalmente se refiere a grafos acíclicos —pues no existe una interpretación generalmente aceptada para grafos con ciclos— y consiste en entender el grafo como un mapa de independencias de una distribución de probabilidad.

Por ejemplo, si tenemos un grafo con dos nodos y un enlace $A \to B$, la interpretación probabilista es que los nodos A y B están conectados, $\neg I_{\mathcal{G}}(A, B)$, y por tanto las variables A y B **pueden estar** correlacionadas en la distribución P. De hecho, en la práctica supondremos que **están** correlacionadas, pues si no, no habríamos trazado ese enlace. Observe que la interpretación probabilista del grafo $B \to A$ es la misma que la del anterior: $\neg I_{\mathcal{G}}(A, B)$. Sin embargo, la interpretación causal de ambos grafos es totalmente distinta: el primero nos dice que A es causa de B, mientras que el segundo nos dice que B es causa de A.

Considere ahora estos tres grafos: $A \to B \to C$, $A \leftarrow B \leftarrow C$, y $A \leftarrow B \to C$. Los tres son equivalentes en sentido probabilista, porque representan las mismas relaciones de conexión, que son $\neg I_{\mathcal{G}}(A, B)$, $\neg I_{\mathcal{G}}(A, B|C)$, $\neg I_{\mathcal{G}}(B, C)$, $\neg I_{\mathcal{G}}(B, C|A)$ y $\neg I_{\mathcal{G}}(A, C)$, y la misma relación de independencia, $I_{\mathcal{G}}(A, C|B)$. En cambio, la interpretación causal de cada uno de los tres grafos es totalmente diferente.

Por tanto, puede haber dos grafos que sean equivalentes en sentido probabilista, es decir, que impliquen las mismas relaciones de separación y conexión, pero nunca dos grafos distintos podrán ser equivalentes en sentido causal.

A veces en inglés se dice "*Markov equivalent*" para indicar "equivalentes en sentido probabilista".

1.6.2. Diferencia entre causalidad y correlación

En esta sección vamos a ilustrar con algunos ejemplos el conocido adagio de que "causalidad implica correlación pero correlación no implica causalidad".

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 1.8. 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.

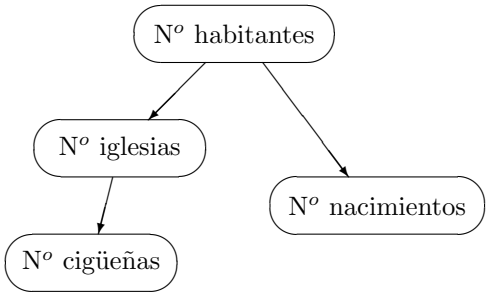


Figura 1.8: La correlación entre número de cigüeñas y número de nacimientos **no** implica causalidad.

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 1.9, 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.

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.

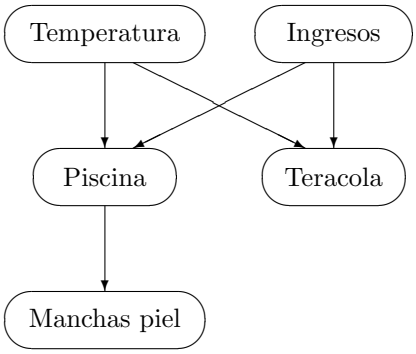


Figura 1.9: La correlación entre el consumo de teracola y la aparición de manchas en la piel **no** implica causalidad.

Bibliografía recomendada

Como dijimos en el capítulo anterior, el método bayesiano ingenuo es un tema casi cerrado 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 [66], 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 [47]. 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 [25] sobre reconocimiento de patrones y visión artificial. Una síntesis de las críticas que se plantearon al método bayesiano ingenuo desde el punto de vista de la inteligencia artificial se encuentra en [54].

La referencia fundamental sobre redes bayesianas es el famoso libro de Judea Pearl [64]. El primer libro sobre el tema escrito en español y uno de los primeros escritos en inglés fue el de Castillo, Gutiérrez y Hadi [6, 7],[9] que recomendamos especialmente como complemento de estos apuntes. En especial, recomendamos a nuestros alumnos que estudien las secciones 4.1 a 4.4, 5.2 y 5.3, 6.2, 6.4.4 a 6.6, y que intenten resolver los ejercicios correspondientes a dichas secciones.

Desde entonces se han publicado muchos libros sobre el tema. Uno de los más adecuados para nuestros alumnos, por su claridad de exposición es el de Neapolitan [57]. Otros libros muy interesantes son los de Jensen y Nielsen [40], Darwiche [15] y Koller y Friedman [44], aunque éstos son más difíciles de leer, especialmente los dos últimos.

En cuanto a la causalidad y su relación con las redes bayesianas, los dos libros más importantes son el de Spirtes, Glymour y Scheines [76] y el de Pearl [65].[10]

[9] La edición española puede obtenerse de forma gratuita en Internet en `http://personales.unican.es/gutierjm/BookCGH.html`.

[10] Judea Pearl, profesor de la Universidad de California Los Ángeles (UCLA) es uno de los investigadores contemporáneos más relevantes. Los tres libros que ha escrito se han convertido en clásicos de la inteligencia artificial. El primero de ellos, sobre búsqueda heurística, fue publicado en 1984 [62]; el segundo, sobre redes bayesianas, en 1988 [64]; y el tercero, sobre causalidad, en 2000 [65]. En marzo de 2012 se le concedió el Premio Turing de 2011, equivalente al Premio Nobel en ciencias de la computación.

Actividades

Además de realizar los ejercicios propuestos a lo largo del capítulo, conviene que el alumno empiece a utilizar algún programa informático para modelos gráficos probabilistas.[11] Recomendamos los programas Elvira y OpenMarkov, especialmente este último porque tiene la posibilidad de aprender redes bayesianas a partir de bases de datos de forma interactiva, lo cual le será útil al estudiar el capítulo 3; ambos son software libre y han sido desarrollados por la UNED, el primero de ellos en colaboración con otras universidades españolas.[12]

1. Instale el programa siguiendo las indicaciones que se dan en la página web correspondiente.

2. Lea el manual introductorio y reproduzca los ejemplos sobre redes bayesianas que se indican en él.

3. Construya una red bayesiana para el problema presentado en el ejemplo 1.35 (pág. 11). Compruebe que los resultados coinciden con los obtenidos en dicho ejemplo.

4. Ídem para el ejemplo 1.38 (pág. 14).

5. Ídem para el ejemplo 1.39 (pág. 18).

[11] En `www.cs.ubc.ca/~murphyk/Software/bnsoft.html` y en `http://www.cisiad.uned.es/congresos/POMDPs-in-OpenMarkov.pdf` puede encontrar una amplia lista de programas.

[12] Véase `www.ia.uned.es/~elvira` y `www.openmarkov.org`.

Capítulo 2

Inferencia en redes bayesianas

Resumen

En el capítulo anterior hemos visto que las probabilidades condicionales que definen una red bayesiana inducen una distribución de probabilidad conjunta, a partir de la cual se pueden calcular todas las probabilidades marginales y condicionales que nos interesen. Por ejemplo, en problemas de diagnóstico, nos va a interesar conocer la probabilidad a posteriori de cada variable o de un conjunto de variables dada la evidencia observada. Es lo que se conoce como *propagación de evidencia*.

Sin embargo, en redes medianas y grandes es imposible desarrollar explícitamente la probabilidad conjunta, por su excesivo tamaño, y por ello vamos a estudiar en este capítulo algunos algoritmos que nos permitirán calcular las probabilidades conjuntas y marginales a partir de las probabilidades condicionales que definen la red.

Vamos a estudiar dos tipos de algoritmos: exactos y aproximados. Como métodos exactos vamos a estudiar tres: la eliminación de variables, el agrupamiento y la inversión de arcos. Como métodos aproximados vamos a estudiar principalmente el muestreo lógico y la ponderación por verosimilitud.

Contexto

En el capítulo anterior hemos estudiado la teoría de las redes bayesianas y en éste vamos a estudiar los algoritmos. Por tanto, éste es continuación del anterior.

Objetivos

El objetivo de este capítulo es conocer distintos algoritmos de propagación de evidencia y la complejidad espacial (cuánta memoria de trabajo requieren) y temporal (cuánto tiempo tardan) de cada uno de ellos.

Requisitos previos

Es necesario haber comprendido bien el capítulo anterior.

Contenido

2.1. Planteamiento del problema

2.1.1. Diagnóstico probabilista

En la definición 1.74 vimos que la distribución de probabilidad conjunta de una red bayesiana puede obtenerse a partir de las probabilidades condicionales de cada nodo dados sus padres; véase también el teorema 1.77. A partir de ella podemos obtener todas las probabilidades conjuntas y marginales que nos interesen. Es lo que se conoce como *inferencia* en redes bayesianas. En particular, en problemas de diagnóstico nos interesa conocer la probabilidad a posteriori de algunas variables dada cierta evidencia.

Definición 2.1 (Hallazgo) Un *hallazgo* consiste en la asignación de un valor a una variable.

Definición 2.2 (Evidencia) Un *caso de evidencia* está formado por un conjunto de hallazgos. Suele representarse por **e**: el conjunto **E** está formado por las variables que tienen asignado un valor y la configuración **e** está formada por los valores asignados.

Por ejemplo, si la variable Fiebre toma los valores {ausente, moderada, alta} un hallazgo puede ser Fiebre = moderada. Si la variable Dolor-torácico toma los valores {ausente, presente} un hallazgo puede ser Dolor-torácico = presente. Un caso de evidencia sería {Fiebre = moderada, Dolor-torácico = presente}. Otro caso sería {Fiebre = alta, Dolor-torácico = ausente}.

Un *problema de diagnóstico* típico consiste en calcular la probabilidad a posteriori de ciertas variables de interés, $\mathbf{X}_I$, dada cierta evidencia **e**: $P(\mathbf{x}_I|\mathbf{e})$. En este caso el conjunto de variables **X** puede dividirse en tres subconjuntos disjuntos:

- variables observadas, **E**. Son las variables que forman la evidencia.
- variables de interés, $\mathbf{X}_I$. Son las variables cuya probabilidad a posteriori nos interesa conocer.
- otras variables, $\mathbf{X}_R$ (el resto). Son las demás variables que forman parte de nuestro modelo.

Por ejemplo, para el problema de diagnóstico consistente en calcular la probabilidad de que cierto paciente, que presenta fiebre moderada y dolor torácico, tenga neumonía y meningitis, la probabilidad que buscamos es P(Neumonía = presente, Meningitis = presente | Fiebre = moderada, Dolor-torácico = presente). En este caso, $\mathbf{E}$ = {Fiebre, Dolor-torácico}, $\mathbf{X}_I$ = {Neumonía, Meningitis} y $\mathbf{X}_R$ son las demás variables de la red bayesiana.

Otras veces, en vez de querer conocer la probabilidad *conjunta* de ciertas variables dada **e**, lo que nos interesa es la probabilidad *individual* de cada variable: $\{P(x_1|\mathbf{e}), P(x_2|\mathbf{e})\ldots\}$. En este caso se habla de *propagación de evidencia*, porque lo que se busca es el impacto de **e** sobre cada una de las demás variables de la red.

2.1.2. Método de fuerza bruta

Dada una red bayesiana cuya probabilidad conjunta es $P(\mathbf{x})$, una forma de calcular $P(\mathbf{x}_I|\mathbf{e})$ consiste en hallar primero las distribuciones de probabilidad marginales $P(\mathbf{x}_I, \mathbf{e})$

y $P(\mathbf{e})$, y luego aplicar la definición de probabilidad condicional:

$$P(\mathbf{x}_I, \mathbf{e}) = \sum_{\mathbf{x}_R} P(\mathbf{x}) \tag{2.1}$$

$$P(\mathbf{e}) = \sum_{\mathbf{x}_I} \sum_{\mathbf{x}_R} P(\mathbf{x}) = \sum_{\mathbf{x}_I} P(\mathbf{x}_I, \mathbf{e}) \tag{2.2}$$

$$P(\mathbf{x}_I|\mathbf{e}) = \frac{P(\mathbf{x}_I, \mathbf{e})}{P(\mathbf{e})} \tag{2.3}$$

Sin embargo, debemos recordar que el tamaño de $P(\mathbf{x})$ crece exponencialmente con el número de variables que forman la red: si la red consta de n variables binarias, $P(\mathbf{x})$ tiene 2^n valores diferentes, uno por cada configuración $\mathbf{x}$. Eso origina dos problemas: el primero es el tiempo necesario para calcular todos los valores de $P(\mathbf{x})$ a partir de las probabilidades condicionales que definen la red; el segundo es de espacio, es decir, la cantidad de memoria necesaria para almacenarlos.[1] Por ello este método, que se conoce como de fuerza bruta, sólo es aplicable en redes de pequeño tamaño (como máximo, unas 30 variables) y aún así, con un coste muy alto.

En seguida vamos a ver algoritmos más eficientes, que consiguen calcular $P(\mathbf{x}_I, \mathbf{e})$ directamente a partir de las probabilidades condicionales que definen la red bayesiana, sin tener que calcular $P(\mathbf{x})$ explícitamente, con lo cual ahorran gran cantidad de tiempo y de espacio.

Por otro lado, la propagación de evidencia puede abordarse resolviendo por separado cada uno de los problemas del tipo $P(x_i|\mathbf{e})$. Sin embargo, cada uno de estos problemas tiene muchos cálculos en común con los demás. Por tanto, es mejor aplicar algoritmos que al calcular cada probabilidad almacenen los resultados intermedios obtenidos, con el fin de reaprovecharlos posteriormente. En la sección 2.2.2 estudiaremos los métodos de agrupamiento, que son capaces de conseguir este objetivo.

Introducimos una definición que nos va a ser útil para explicar estos algoritmos.

Definición 2.3 (Potencial) Un potencial ψ definido sobre un conjunto de variables $\mathbf{X}$ es una función que asigna a cada configuración $\mathbf{x}$ un número real: $\psi : X \longrightarrow \mathbb{R}$.

Ejemplo 2.4 La probabilidad conjunta $P(\mathbf{x})$ es un potencial definido sobre $\mathbf{X}$.

Ejemplo 2.5 La probabilidad condicional $P(\mathbf{x}|\mathbf{y})$ es un potencial definido sobre $\mathbf{X} \cup \mathbf{Y}$.

Ejemplo 2.6 La suma de dos potenciales $\psi_1(\mathbf{x})$ y $\psi_2(\mathbf{y})$ da lugar a un nuevo potencial definido sobre $\mathbf{X} \cup \mathbf{Y}$. También la resta, multiplicación y división de potentiales dan lugar a nuevos potenciales.

[1] En realidad, el problema de espacio se podría solventar si en vez de almacenar en memoria todos los valores de $P(\mathbf{x})$ los generamos a medida que los necesitamos. El problema es que habría que generar cada valor tantas veces como fuéramos a utilizarlo. Es decir, evitamos el problema de la complejidad espacial a cambio de aumentar el tiempo de computación. Este fenómeno aparece con frecuencia en los algoritmos exactos de propagación de evidencia en redes bayesianas. Por ello muchos de estos algoritmos ofrecen diferentes versiones: unas reducen la complejidad temporal a costa de usar más memoria, y otras ahorran tiempo almacenando más resultados en memoria.

2.2. Métodos exactos

2.2.1. Eliminación de variables

Supongamos que, dada la red bayesiana de la figura 2.1, formada por 6 variables booleanas, tenemos la evidencia $\{H = +h\}$ y queremos calcular la probabilidad a posteriori de A. Siguiendo el método de fuerza bruta, calcularíamos $P(a, +h)$ mediante la ecuación (2.1):

$$\begin{aligned} P(a, +h) &= \sum_b \sum_d \sum_f \sum_g P(a, b, d, f, g, +h) \\ &= \sum_b \sum_d \sum_f \sum_g P(a) \cdot P(b|a) \cdot P(d|a) \cdot P(f) \cdot P(g|b, d, f) \cdot P(+h|g) \end{aligned}$$

En este caso, $\mathbf{X}_R = \{B, D, F, G\}$ tiene cuatro variables binarias, y por tanto el sumatorio da lugar a $2^4 = 16$ sumandos, cada uno de los cuales se obtiene realizando 5 multiplicaciones. Por tanto, el cálculo de $P(+a, +h)$ requiere 80 multiplicaciones y 15 sumas, y el de $P(\neg a, +h)$ otras tantas. En total, necesitamos 160 multiplicaciones y 30 sumas.

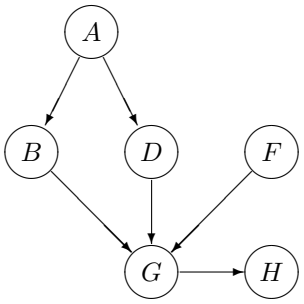


Figura 2.1: Red bayesiana de seis nodos.

Veamos cómo se puede realizar el cálculo anterior de forma más eficiente. Tomamos una de las variables de $\mathbf{X}_R$, por ejemplo, G y reunimos todos los potenciales que dependen de ella (en este ejemplo son las probabilidades condicionales $P(g|b, d, f)$ y $P(+h|g)$), los multiplicamos, sumamos sobre los valores de G y obtenemos un nuevo potencial, cuyo dominio estará formado por todas las variables que tienen algún potencial en común con G:

$$\psi_1(b, d, f) = \sum_g P(g|b, d, f) \cdot P(+h|g)$$

Por tanto,

$$P(a, +h) = \sum_b \sum_d \sum_f P(a) \cdot P(b|a) \cdot P(d|a) \cdot P(f) \cdot \psi_1(b, d, f)$$

Aplicando el mismo procedimiento vamos a eliminar ahora la variable F. En este caso, los potenciales que tenemos que multiplicar son $P(f)$ y $\psi_1(b, d, f)$:

$$\psi_2(b, d) = \sum_f P(f) \cdot \psi_1(b, d, f)$$

de modo que

$$P(a,+h) = \sum_b \sum_d P(a) \cdot P(b|a) \cdot P(d|a) \cdot \psi_2(b,d)$$

Ahora eliminamos la variable B:

$$\psi_3(a,d) = \sum_b P(b|a) \cdot \psi_2(b,d)$$
$$P(a,+h) = \sum_d P(a) \cdot P(d|a) \cdot \psi_3(a,d)$$

Por último, eliminamos D:

$$\psi_4(a) = \sum_d P(d|a) \cdot \psi_3(a,d)$$
$$P(a,+h) = P(a) \cdot \psi_4(a)$$

Observe que hemos llegado a $P(a,+h)$ sin haber calculado explícitamente los valores de la probabilidad conjunta $P(a,b,d,f,g,+h)$.

Aunque parezca que el cálculo es más complicado que con el método de fuerza bruta, en realidad hemos ahorrado un buen número de operaciones. El cálculo de $\psi_1(b,d,f)$ ha necesitado 16 multiplicaciones —una por cada configuración del tipo (b,d,f,g)— y 8 sumas, una por cada configuración (b,d,f). El cálculo de $\psi_2(b,d)$ necesita 8 multiplicaciones —una por cada configuración (b,d,f)— y 4 sumas, una por cada configuración (b,d). En estos cálculos no interviene la variable A, y por consiguiente nos van a servir tanto para $P(+a,+h)$ como para $P(\neg a,+h)$. El cálculo de $\psi_3(a,d)$ requiere igualmente 8 multiplicaciones y 4 sumas. El de $\psi_4(a,d)$, 4 multiplicaciones y 2 sumas. La última ecuación conlleva dos multiplicaciones. En total, hemos necesitado 38 multiplicaciones y 18 sumas, lo cual significa un ahorro notable frente al método de fuerza bruta, que requería 160 multiplicaciones y 30 sumas. Es decir, hemos reducido el número de operaciones en un 70%. En redes bayesianas con un número mayor de variables el ahorro puede ser mucho mayor, como veremos más adelante.

Ejercicio 2.7 Calcule cuántas operaciones serían necesarias en el ejemplo anterior si cada variable en vez de tomar sólo dos valores tomara n. Compare este resultado con el del método de fuerza bruta.

El algoritmo de eliminación de variables para el caso general sería así:

Algoritmo 2.1 (Eliminación de variables)
Entrada: probabilidades condicionales de una red bayesiana, evidencia $\mathbf{e}$
Salida: $P(\mathbf{x}_I, \mathbf{e})$
lista-de-potenciales := {probabilidades condicionales de la red reducidas según $\mathbf{e}$};
para cada variable X de $\mathbf{X}_R$
1. sacar de lista-de-potenciales todos los potenciales que dependen de X;
2. multiplicarlos y sumar sobre los valores de X;
3. añadir el potencial resultante a la lista de potenciales;

;
multiplicar todos los potenciales que quedan en lista-de-potenciales.

Observe que al reducir las probabilidades condicionales según **e** obtenemos unos potenciales que ya no dependen de **E**. Por ejemplo, la probabilidad condicional $P(h|g)$ es un potencial que depende de G y de H; si estas variables tienen n_G y n_H valores, respectivamente, este potencial tomará $n_G \times n_H$ valores. Al reducirlo según la evidencia $\{H = +h\}$ obtenemos el potencial $P(+h|g)$, que sólo depende de G, y por tanto tendrá sólo n_G valores. El bucle del algoritmo va eliminando sucesivamente cada una de las variables de $\mathbf{X}_R$, con lo cual a la salida del bucle tenemos unos potenciales que ya no dependen de $\mathbf{E}$ ni de $\mathbf{X}_R$, sino sólo de $\mathbf{X}_I$. Al multiplicarlos obtenemos el potencial buscado, $P(\mathbf{x}_I, \mathbf{e})$, y a partir de él calculamos $P(\mathbf{x}_I|\mathbf{e})$ mediante las ecuaciones (2.2) y (2.3).

Importancia del orden de eliminación

Considere el grafo de la figura 1.4 (pág. 18). Suponga que queremos calcular $P(h_m|h_1)$ para un valor de H_1 conocido —por ejemplo, $+h_1$— y para todos los valores de H_m. El método de fuerza bruta calcularía primero $P(+h_1, h_m)$, aplicando directamente la definición de probabilidad marginal y la factorización de la probabilidad:

$$\begin{aligned} P(+h_1, h_m) &= \sum_d \sum_{h_2} \dots \sum_{h_{m-1}} P(d, +h_1, h_2, \dots, h_m) \\ &= \sum_d \sum_{h_2} \dots \sum_{h_{m-1}} P(d) \cdot P(+h_1|d) \cdot P(h_2|d) \cdot \ldots \cdot P(h_m|d) \end{aligned}$$

Si todas las variables son binarias el método necesitará $2^m \times m$ multiplicaciones y $2^m - 1$ sumas. Claramente la complejidad del método crece exponencialmente con el número de hallazgos posibles.

Vamos a aplicar ahora el método de eliminación de variables siguiendo este orden: H_2, H_3, ... , H_{m-1}, D:

$$\psi_1(d) = \sum_{h_2} P(h_2|d)$$

$$P(+h_1, h_m) = \sum_d \sum_{h_3} \dots \sum_{h_{m-1}} P(d) \cdot P(+h_1|d) \cdot \psi_1(d) \cdot P(h_3|d) \cdot \ldots \cdot P(h_m|d)$$

y así sucesivamente hasta llegar a

$$\psi_{m-2}(d) = \sum_{h_{m-1}} P(h_{m-1}|d)$$

$$P(+h_1, h_m) = \sum_d P(d) \cdot P(+h_1|d) \cdot \psi_1(d) \cdot \ldots \cdot \psi_{m-2}(d) \cdot P(h_m|d)$$

Finalmente eliminamos D directamente de la ecuación anterior.

La eliminación de cada variable H_i necesita dos sumas, una para $+d$ y otra para $\neg d$. La última ecuación necesita m multiplicaciones para cada valor de H_m. En total son $(m-2) \times 2$ sumas y $2 \times m$ multiplicaciones. ¡Una ganancia impresionante frente al método de fuerza bruta, pues hemos pasado de complejidad exponencial en m a complejidad lineal!

El mismo ahorro puede conseguirse con cualquier ordenación de las H_i's, siempre que la última variable eliminada sea D.

Supongamos ahora que un inexperto aplicara el método de eliminación de variables, pero eliminando primero la variable D:

$$\psi_1(h_2, \ldots, h_m) = \sum_d P(d) \cdot P(+h_1|d) \cdot P(h_2|d) \cdot \ldots \cdot P(h_m|d)$$
$$P(+h_1, h_m) = \sum_{h_2} \ldots \sum_{h_m - 1} \psi_1(h_2, \ldots, h_m)$$

La primera de estas dos ecuaciones necesita $2^{m-1} \times m$ multiplicaciones. La segunda, $2^{m-2} - 1$ sumas. ¡De nuevo tenemos complejidad exponencial!

Complejidad del problema

En el ejemplo anterior hemos visto que el coste computacional de la eliminación de variables puede depender drásticamente del orden de eliminación. Para ello conviene eliminar las variables intentando evitar que se formen potenciales grandes. Uno podría pensar que escogiendo un buen orden siempre tendremos una complejidad reducida. Sin embargo, eso no es cierto: hay redes bayesianas que, por la estructura de su grafo, tienen un coste computacional muy elevado, cualquiera que sea el orden de eliminación; es decir, cualquiera que sea el orden escogido, siempre se van a formar potenciales grandes. Piense en el caso extremo de una red bayesiana basada en un grafo completo (es decir, un grafo en que cada nodo está conectado con todos los demás): cualquier orden de eliminación va a dar un potencial que incluya todas las variables, exactamente igual que si aplicamos el método de fuerza bruta.

De hecho, Greg Cooper [12] demostró que el problema de la inferencia en redes bayesianas es NP-completo. Como Vd. ya sabe, los problemas NP-completos forman una familia tal que cada uno se puede transformar en otro en un tiempo polinómico,[2] y la solución del segundo se puede transformar en tiempo polinómico en una solución para el primero. Eso significa que si tenemos una red bayesiana como las que Cooper utilizó en su artículo y una pregunta del tipo $P(\mathbf{x}_I|\mathbf{e})$, podemos transformar este problema en un problema de la clase SAT, o 3SAT, o en el problema del viajero... y una vez resuelto este problema podemos transformar su solución en la respuesta a la pregunta $P(\mathbf{x}_I|\mathbf{e})$.

Aparentemente esto es una buena noticia: si alguien encuentra un algoritmo de complejidad polinómica para un problema NP-completo —cualquiera de ellos—, entonces podemos transformar nuestra red bayesiana en un problema de ese tipo y luego convertir la solución de ese problema en una solución para nuestra red, todo ello en tiempo polinómico. Sin embargo, nadie ha conseguido encontrar un algoritmo polinómico para ningún problema NP-completo; de hecho, todos los expertos en teoría de la computación están convencidos de que tal solución no existe, aunque ninguno haya conseguido demostrarlo todavía. (Éste el problema abierto más importante de la teoría de la computación. El que consiga resolverlo se hará famoso.) Si esta conjetura es cierta, eso significa que nunca se podrá encontrar un algoritmo capaz de resolver cualquier red bayesiana en tiempo polinómico.

[2] "En tiempo polinómico" significa que si el tamaño del problema, codificado según algún criterio, es n entonces existe un valor numérico a tal que el tiempo necesario para resolverlo es menor que a^n, y el valor a sirve para cualquier problema, es decir, para todas las redes bayesianas.

2.2.2. Agrupamiento

El método de agrupamiento realiza los cálculos sobre una estructura gráfica denominada *árbol de grupos*, que definimos así:

Definición 2.8 (Árbol de grupos) Sea $\mathbf{X}$ un conjunto de variables y $\{\psi_i\}$ un conjunto de potenciales tales que cada $\psi_i(\mathbf{x}_i)$ está definido sobre un subconjunto $\mathbf{X}_i \subseteq \mathbf{X}$ y para cada variable hay al menos un potencial definido sobre ella, es decir, $\bigcup_i \mathbf{X}_i = \mathbf{X}$. Un *árbol de grupos* asociado a este conjunto de potenciales es un árbol dirigido finito que cumple las siguientes propiedades:

1. Cada nodo del árbol representa un grupo, formado por un subconjunto de variables de $\mathbf{X}$. Denotaremos por $\mathbf{C}_i$ tanto el nodo del árbol como el subconjunto de variables que representa.

2. Si una variable de $\mathbf{X}$ pertenece a dos grupos distintos del árbol, entonces pertenece también a todos los grupos que se encuentran en el camino que hay entre ambos en el árbol. Es lo que se conoce como *propiedad del árbol de uniones*.

3. Cada potencial ψ_i está asociado a un grupo que contiene todas sus variables. El conjunto de potenciales asociados a $\mathbf{C}_i$ lo denotaremos por $Pot(\mathbf{C}_i)$.

En particular, vamos a estar interesados en los árboles de grupos en que $\mathbf{X}$ son las variables de una red bayesiana $(\mathbf{X},\mathcal{G},P)$ y los potenciales son las probabilidades condicionales que factorizan P; es decir, para una red de n nodos vamos a tener n potenciales, cada uno de ellos definido sobre una familia: $\psi_i(fam(X_i)) = \psi_i(x_i, pa(X_i)) = P(x_i|pa(X_i))$. Observe que la tercera condición de la definición de árbol de grupos implica que para cada familia tiene que haber al menos un grupo que contenga todas sus variables: si no se cumpliera esta condición para alguna familia $Fam(X_i)$, entonces el potencial $P(x_i|pa(X_i))$ no podría estar asociado a ningún grupo.

Ejemplo 2.9 Un árbol de grupos para la red bayesiana de la figura 2.2 puede ser el que se muestra en la figura 2.3. Es inmediato comprobar que se cumple la primera condición. También se cumple la segunda: por ejemplo, la variable C pertenece a $\mathbf{C}_1$ y a $\mathbf{C}_4$, y por tanto debe pertenecer también a todos los grupos que se encuentran en el camino que hay entre ambos, es decir, a $\mathbf{C}_2$ y a $\mathbf{C}_3$; del mismo modo, la variable D pertenece a $\mathbf{C}_2$ y a $\mathbf{C}_5$, y por tanto debe pertenecer también a $\mathbf{C}_3$. Por último, la tercera condición se comprueba fácilmente. Observe que en este ejemplo sólo hay una posibilidad de asignar los potenciales porque cada familia está contenida en un único grupo. □

De la definición anterior se deduce que para toda red bayesiana existe al menos un árbol de grupos: el formado por un solo grupo, que contiene todas las variables de la red. Sin embargo, los árboles más eficientes de cara a la inferencia son aquéllos que tienen los grupos más pequeños. La razón, como veremos en seguida con más detalle, es que el coste computacional asociado a cada grupo crece exponencialmente con el número de variables que contiene, mientras que el coste total crece linealmente con el número de grupos. Por eso es mejor tener muchos grupos pequeños, como en la figura 2.3, que un pequeño número de grupos grandes. También en este caso es aplicable el principio de "divide y vencerás".

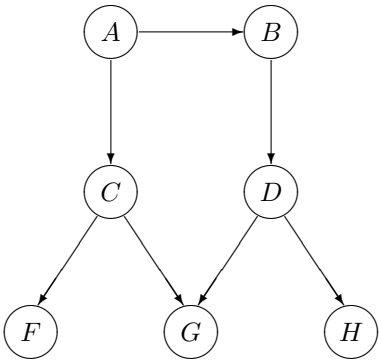


Figura 2.2: Red bayesiana de siete nodos.

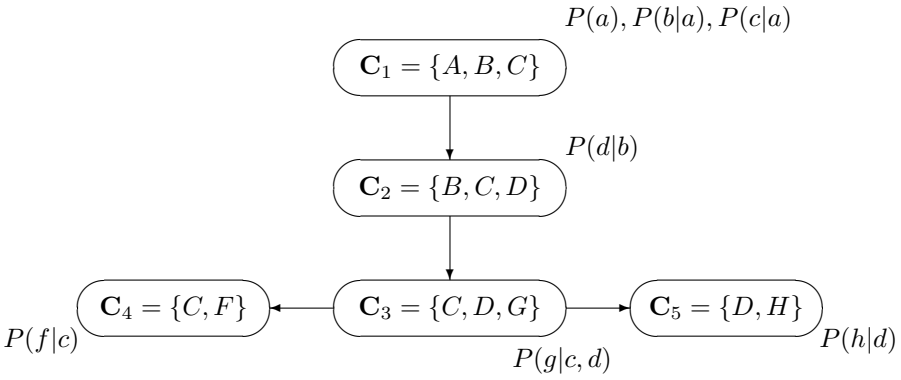


Figura 2.3: Un posible árbol de grupos para la red de la figura 2.2. Junto a cada grupo hemos escrito sus potenciales asociados.

Computación de la probabilidad en un árbol de grupos

Primera versión: un solo grupo hace todo el trabajo La primera versión del algoritmo funciona así:

1. cada grupo se encarga de pedir a sus vecinos sus potenciales asociados,
2. los proyecta sobre la evidencia,
3. los multiplica para obtener $P(\mathbf{x}_I, \mathbf{x}_R, \mathbf{e})$ para todas las configuraciones $(\mathbf{x}_I, \mathbf{x}_R)$,
4. suma sobre las variables de X_R para obtener $P(\mathbf{x}_I, \mathbf{e})$,
5. calcula $P(\mathbf{e})$ sumando sobre las configuraciones $\mathbf{x}_I$ y
6. calcula $P(\mathbf{x}_I|\mathbf{e})$ dividiendo $P(\mathbf{x}_I, \mathbf{e})$ por $P(\mathbf{e})$.

El paso 1 puede implementarse mediante paso de mensajes, como vamos a ver en el siguiente ejemplo:

Ejemplo 2.10 Supongamos que tenemos la red de la figura 2.2 y queremos calcular cualquier probabilidad conjunta, marginal o condicional a partir del árbol de la figura 2.3. Encargamos el cálculo a $\mathbf{C}_1$. Como este nodo sólo tiene algunos de los potenciales, necesita pedir los demás a los grupos que los tienen, empezando por sus vecinos, los cuales los pedirán a sus demás vecinos, y así sucesivamente siguiendo la estructura del árbol.

El proceso se inicia cuando $\mathbf{C}_1$ dice a $\mathbf{C}_2$: "recoge los potenciales que quedan de tu lado y envíamelos". A su vez, $\mathbf{C}_2$ dice a $\mathbf{C}_3$: "recoge los potenciales que quedan de tu lado y envíamelos", y éste transmite el mismo mensaje a sus vecinos, $\mathbf{C}_4$ y $\mathbf{C}_5$. El grupo $\mathbf{C}_4$, como no tiene otros vecinos a los que transmitir la petición, puede responder inmediatamente, devolviendo el único potencial que tiene $P(f|c)$; es decir, el mensaje M_{43} que $\mathbf{C}_4$ devuelve a $\mathbf{C}_3$ es el potencial $P(f|c)$. Del mismo modo, como $\mathbf{C}_5$ no tiene otros vecinos, el mensaje que devuelve a $\mathbf{C}_3$ está formado por el único potencial de $\mathbf{C}_5$: $M_{53} = \{P(h|d)\}$.

Cuando $\mathbf{C}_3$ ha reunido los mensajes de sus vecinos ya puede devolver a $\mathbf{C}_2$ el mensaje que éste le había pedido. Este mensaje contendrá los potenciales que $\mathbf{C}_3$ ha recogido de sus vecinos más los potenciales que $\mathbf{C}_3$ tenía: $M_{32} = Pot(\mathbf{C}_3) \cup M_{43} \cup M_{53} = \{P(g|c,d), P(f|c), P(h|d)\}$.

Del mismo modo, el mensaje que $\mathbf{C}_2$ devuelve a $\mathbf{C}_1$ contendrá los potenciales que $\mathbf{C}_2$ ha recibido de sus vecinos —el único vecino de $\mathbf{C}_2$, aparte de $\mathbf{C}_1$, que es quien le ha solicitado el mensaje, es $\mathbf{C}_3$— más los potenciales que $\mathbf{C}_2$ tenía: $M_{21} = Pot(\mathbf{C}_2) \cup M_{32} = \{P(d|b), P(g|c,d), P(f|c), P(h|d)\}$.

Finalmente, $\mathbf{C}_1$ agrega a los 3 potenciales que tenía asociados los 4 que ha recibido de sus vecinos y así consigue reunir los 7 que formaban parte de la red bayesiana. A partir de ellos puede calcular la probabilidad conjunta de la red y cualquier otra probabilidad marginal o condicional que deseemos. □

El nodo encargado de recopilar todos los potenciales de la red, solicitando mensajes a sus vecinos, se denomina *pivote*. En el ejemplo que acabamos de considerar el nodo pivote ha sido $\mathbf{C}_1$, que es el nodo raíz, y por eso todos los mensajes circulaban desde los hijos hacia su padre. Sin embargo, esto no tiene por qué ser así: cualquier nodo puede ser el pivote y al solicitar mensajes a sus vecinos no distinguirá entre padres e hijos. La ecuación general para calcular el mensaje que un grupo $\mathbf{C}_j$ envía a su vecino $\mathbf{C}_i$ es:

$$M_{ji} = Pot(\mathbf{C}_j) \cup \bigcup_k M_{kj} \tag{2.4}$$

donde k varía para cubrir todos los subíndices de los vecinos de $\mathbf{C}_j$, incluidos padres e hijos, excepto $\mathbf{C}_i$.

Ejercicio 2.11 Aplique de nuevo el algoritmo al árbol de la figura 2.3, como en el ejemplo anterior, pero tomando como nodo pivote $\mathbf{C}_3$. Es importante que señale claramente cuáles son los mensajes que se propagan.

Segunda versión: cada grupo hace una parte del trabajo En la sección anterior hemos visto que el nodo pivote se encargaba de recopilar todos los potenciales y de calcular las probabilidades buscadas. Vamos a ver ahora cómo es posible repartir el trabajo de modo que cada grupo realiza una parte de la computación y envía a sus vecinos sólo la información necesaria.

Ejemplo 2.12 Consideremos de nuevo la red de la figura 2.2 y el árbol de la figura 2.3 y supongamos que queremos calcular la probabilidad de B dada la evidencia $\mathbf{e} = \{+f, \neg h\}$,

es decir, $P(b|+f,\neg h)$. Para ello vamos a calcular previamente $P(a,b,c,+f,\neg h)$, que es la probabilidad de las variables de $\mathbf{C}_1$ y de la evidencia, y de ahí obtendremos $P(b,+f,\neg h)$ por marginalización:

$$P(b,+f,\neg h) = \sum_a \sum_c P(a,b,c,+f,\neg h)$$

La probabilidad $P(a,b,c,+f,\neg h)$ se calcula a partir de la factorización de la probabilidad reordenando los sumatorios para aislar la parte que no depende de los potenciales de $\mathbf{C}_1$; a esta parte la vamos a llamar M_{21}:

$$\begin{aligned} P(a,b,c,+f,\neg h) &= P(a)\cdot P(b|a)\cdot P(c|a)\cdot \sum_d \sum_g P(d|b)\cdot P(+f|c)\cdot P(g|c,d)\cdot P(\neg h|d) \\ &= \phi_1^0(a,b,c)\cdot M_{21}(b,c) \end{aligned}$$

donde:

$$\begin{aligned} \phi_1^0(a,b,c) &= P(a)\cdot P(b|a)\cdot P(c|a) \\ M_{21}(b,c) &= \sum_d \sum_g P(d|b)\cdot P(+f|c)\cdot P(g|c,d)\cdot P(\neg h|d) \end{aligned}$$

El potencial ϕ_1^0 es el producto de los potenciales asignados a $\mathbf{C}_1$, proyectados sobre la evidencia, mientras que M_{21} recoge todos los potenciales asignados a los demás grupos, también proyectados sobre $\mathbf{e}$. En la definición de M_{21} hemos sumado sobre D y G; es decir, sobre todas las variables de $\mathbf{X}_R$ que no forman parte de $\mathbf{C}_1$.[3] Eso hace que M_{21} sólo dependa de las variables de $\mathbf{C}_1$, porque la suma ha eliminado las demás: $dom(M_{21}) \subseteq \mathbf{C}_1$.[4]

La expresión $M_{21}(b,c)$ puede entenderse como un mensaje que $\mathbf{C}_2$ envía a $\mathbf{C}_1$ y puede calcularse con la información que queda del lado de $\mathbf{C}_2$ si quitáramos de la red el enlace entre $\mathbf{C}_1$ y $\mathbf{C}_2$:

$$\begin{aligned} M_{21}(b,c) &= \sum_d P(d|b)\cdot \sum_g P(+f|c)\cdot P(g|c,d)\cdot P(\neg h|d) \\ &= \sum_d \phi_2^0(b,d)\cdot M_{32}(c,d) \end{aligned}$$

[3] Al definir M_{21} no podemos sumar sobre B ni sobre C porque eso nos llevaría a un resultado erróneo. Tenga en cuenta que

$$P(b,+f,\neg h) = \sum_a \sum_c \phi_1^0(a,b,c)\cdot \underbrace{\sum_d \sum_g P(d|b)\cdot P(+f|c)\cdot P(g|c,d)\cdot P(\neg h|d)}_{M_{21}(b,c)}$$

en general es distinto de

$$\sum_a \left(\sum_c \phi_1^0(a,b,c)\right)\cdot \underbrace{\sum_{b'} \sum_{c'} \sum_d \sum_g P(d|b')\cdot P(+f|c')\cdot P(g|c',d)\cdot P(\neg h|d)}_{[M_{21}\text{ erróneo}]}$$

La diferencia entre estas dos ecuaciones es que en la primera calculamos $M_{21}(b,c)$ para unos valores de B y de C concretos, mientras que en la segunda calculamos un potencial M_{21} erróneo, que no depende de valores concretos de B y de C porque lo hemos calculado sumando sobre **todos** los valores de estas dos variables.

[4] Aquí se observa una diferencia respecto de la primera versión del algoritmo: en la primera, cada mensaje M_{21} era un conjunto de potenciales, mientras que ahora M_{21} es un potencial que depende de un subconjunto de las variables de $\mathbf{C}_1$.

donde

$$\phi_2^0(b,d) = P(d|b)$$
$$M_{32}(c,d) = \sum_g P(+f|c) \cdot P(g|c,d) \cdot P(\neg h|d)$$

Es decir, el mensaje que $\mathbf{C}_2$ envía a $\mathbf{C}_1$ se calcula a partir de los potenciales asociados a $\mathbf{C}_2$ y de los mensajes que $\mathbf{C}_2$ recibe de sus vecinos, excepto $\mathbf{C}_1$. En M_{32} hemos reunido todos los potenciales que no están asociados a $\mathbf{C}_2$ y hemos sumado sobre todas las variables que no pertenecen a $\mathbf{C}_2$, con lo cual garantizamos que $dom(M_{21}) \subseteq \mathbf{C}_2$. Uniendo este resultado al anterior, concluimos que $dom(M_{21}) \subseteq \mathbf{C}_1 \cap \mathbf{C}_2$. Éste es un caso particular de la propiedad general $dom(M_{ij}) \subseteq \mathbf{C}_i \cap \mathbf{C}_j$, que vamos a demostrar más adelante.

Reordenando los potenciales que componen M_{32} tenemos que

$$M_{32}(c,d) = \sum_g \underbrace{P(g|c,d)}_{\phi_3^0(c,d,g)} \cdot \underbrace{P(+f|c)}_{M_{43}(c)} \cdot \underbrace{P(\neg h|d)}_{M_{53}(d)}$$

Como ocurría antes, el mensaje que $\mathbf{C}_3$ envía a $\mathbf{C}_2$ se calcula a partir de los potenciales asociados a $\mathbf{C}_3$ y de los mensajes que $\mathbf{C}_3$ recibe de sus vecinos, excepto $\mathbf{C}_2$. También aquí se comprueba que $dom(M_{32}) = \{C, D\} \subseteq \mathbf{C}_2 \cap \mathbf{C}_3$, $dom(M_{43}) = \{C\} \subseteq \mathbf{C}_3 \cap \mathbf{C}_4$ y $dom(M_{53}) = \{D\} \subseteq \mathbf{C}_3 \cap \mathbf{C}_5$.

El mensaje M_{43} es el resultado del potencial inicial de $\mathbf{C}_4$, $\phi_4^0(c,f) = P(f|c)$, proyectado sobre la evidencia, $\mathbf{e} = \{+f, \neg h\}$; si $\mathbf{C}_4$ tuviera otros vecinos habría que multiplicar $\phi_4^0(c,+f)$ por los mensajes que ellos le enviaran. Como no tiene más vecinos, el algoritmo termina en esta rama del árbol.

Igualmente, M_{53} es el resultado del potencial inicial de $\mathbf{C}_5$, $\phi_5^0(d,h) = P(h|d)$, proyectado sobre la evidencia, $\mathbf{e} = \{+f, \neg h\}$. Como $\mathbf{C}_5$ no tiene más vecinos, no hace falta multiplicar por otros mensajes y el algoritmo también termina en esta rama del árbol.

Ejercicio 2.13 Con la misma red y el mismo árbol que en el ejemplo anterior, calcule $P(d|+f, \neg h)$ utilizando $\mathbf{C}_2$ como nodo pivote. ¿Cuáles de los mensajes coinciden con los de dicho ejemplo?

Como hemos visto en estos ejemplos, el algoritmo de agrupamiento permite calcular la probabilidad a posteriori $P(\mathbf{x}_I|\mathbf{e})$ siempre que en el árbol haya un grupo $\mathbf{C}_p$ (donde p significa "pivote") que contenga todas las variables de interés: $\mathbf{X}_I \subseteq \mathbf{C}_p$. En pseudocódigo puede expresarse así:

Algoritmo 2.2 (Propagación de evidencia en un árbol de grupos)
Entrada: árbol de potenciales, variables de interés $\mathbf{X}_I$, evidencia $\mathbf{e}$
Salida: $P(\mathbf{x}_I|\mathbf{e})$
$\mathbf{C}_p$:= grupo que contenga todas las variables de $\mathbf{X}_I$;
producto := 1;
para cada potencial ψ_j asociado a $\mathbf{C}_p$, es decir, $\psi_j \in Pot(\mathbf{C}_p)$,
  calcular ψ'_j como la proyección de ψ_j según la evidencia $\mathbf{e}$;
  producto := producto $\times$ ψ'_j;
;
para cada grupo $\mathbf{C}_i$ vecino de $\mathbf{C}_p$,
  calcular el mensaje M_{ip} según la ecuación (2.5) [algoritmo 2.3];

producto := producto × M_{pi};
;
// ahora producto = $P(\mathbf{c}'_p, \mathbf{e})$, donde $\mathbf{C}'_p = \mathbf{C}_p \backslash \mathbf{E}$;
$\mathbf{R} := \mathbf{C}_p \backslash (\mathbf{X}_I \cup \mathbf{E})$;
$P(\mathbf{x}_I, \mathbf{e}) := \sum_{\mathbf{r}}$ producto;
$P(\mathbf{e}) := \sum_{\mathbf{x}_I} P(\mathbf{x}_I, \mathbf{e})$;
$P(\mathbf{x}_I|\mathbf{e}) := P(\mathbf{x}_I, \mathbf{e})/P(\mathbf{e})$;
devuelve $P(\mathbf{x}_I|\mathbf{e})$.

La explicación del algoritmo es la siguiente: primero buscamos un grupo $\mathbf{C}_p$ que contenga todas las variables de interés, y será nuestro nodo pivote. En el ejemplo anterior, era $\mathbf{C}_1$, pues $\mathbf{X}_I = \{B\} \subseteq \mathbf{C}_1$. Luego multiplicamos todos sus potenciales asociados, proyectados sobre la evidencia —en el ejemplo eran $P(a)$, $P(b|a)$ y $P(c|a)$— y los mensajes que recibe de sus vecinos —$M_{21}(b, c)$—. El resultado de estas multiplicaciones nos da la probabilidad conjunta de las variables de $\mathbf{C}_p$ y de $\mathbf{e}$. En caso de que $\mathbf{C}_p$ no contenga ninguna variable observada, esta probabilidad es $P(\mathbf{c}_p, \mathbf{e})$; en nuestro ejemplo $P(\mathbf{c}_p, \mathbf{e}) = P(a, b, c, +f, \neg h)$. Sin embargo, es posible que $\mathbf{C}_p \cap \mathbf{E} \neq \varnothing$; por eso hemos introducido la definición $\mathbf{C}'_p = \mathbf{C}_p \backslash \mathbf{E}$, para poder afirmar que el producto calculado es igual a $P(\mathbf{c}'_p, \mathbf{e})$.

La probabilidad $P(\mathbf{x}_I, \mathbf{e})$ se calcula a partir de $P(\mathbf{c}'_p, \mathbf{e})$ sumando sobre las configuraciones de $\mathbf{R}$, que es el conjunto formado por las variables de $\mathbf{C}_p$ que no son ni de interés ni observadas; en nuestro ejemplo, $\mathbf{R} = \{A, C\}$ y $P(\mathbf{x}_I, \mathbf{e}) = P(b, +f, \neg h) = \sum_a \sum_c P(a, b, c, +f, \neg h)$. Finalmente, $P(b, +f, \neg h)$ se calcula como $P(b, +f, \neg h)/P(+f, \neg h)$.

El término *propagación de evidencia* se refiere al hecho de que los mensajes llevan implícita información sobre las variables observadas, de modo que el impacto de cada uno de los hallazgos se transmite a todos los grupos del árbol.

La ecuación para el cómputo de los mensajes es la siguiente:

$$M_{ji} = \sum_{\mathbf{r}_{ji}} \phi_j^0 \cdot \prod_{k \in K} M_{kj} \tag{2.5}$$

donde M_{ji} es el mensaje que $\mathbf{C}_j$ envía $\mathbf{C}_i$, ϕ_j^0 es el producto de todos los potenciales asociados a $\mathbf{C}_j$ (previamente proyectados sobre $\mathbf{e}$), K contiene los subíndices de los vecinos de $\mathbf{C}_j$ distintos de $\mathbf{C}_i$, es decir, $K = \{k \in \mathbb{N} \mid \mathbf{C}_k$ es padre o hijo de $\mathbf{C}_j$ en el árbol de unión $\wedge\ k \neq i\}$, y $\mathbf{R}_{ji}$ contiene las variables que pertenecen a $\mathbf{C}_j$ y no a $\mathbf{C}_i$: $\mathbf{R}_{ji} := \mathbf{C}_j \backslash \mathbf{C}_i$. En forma de algoritmo podemos expresarlo así:

Algoritmo 2.3 (Mensajes del árbol de grupos)
Entrada: nodos $\mathbf{C}_i$ y $\mathbf{C}_j$, vecinos en un árbol de potenciales
Salida: mensaje M_{ji}, que $\mathbf{C}_j$ envía a $\mathbf{C}_i$
producto := producto de los potenciales asociados a $\mathbf{C}_j$;
para cada vecino $\mathbf{C}_k$ de $\mathbf{C}_j$ distinto de $\mathbf{C}_i$,
  producto := producto × mensaje M_{kj};
;
$\mathbf{R}_{ji} := \mathbf{C}_j \backslash \mathbf{C}_i$;
$M_{ji} := \sum_{\mathbf{r}_{ji}}$ producto;
devuelve M_{ji}.

Es decir, el mensaje que $\mathbf{C}_j$ envía a $\mathbf{C}_i$ se calcula multiplicando los potenciales asociados a $\mathbf{C}_j$ y los mensajes que $\mathbf{C}_j$ recibe de sus demás vecinos, y luego se eliminan por suma las variables que no pertenecen a $\mathbf{C}_i$.

Como se trata de un algoritmo recursivo, el primer paso para garantizar que es correcto consiste en demostrar que termina en un número finito de pasos. Para demostrarlo hay que tener en cuenta, en primer lugar, que como los árboles no tienen bucles, es imposible que un nodo que ha solicitado un mensaje reciba posteriormente la petición de un mensaje; por tanto las peticiones de mensajes nunca llegan dos veces a un mismo nodo. Por otro lado, dado que el número de grupos en el árbol es finito, tarde o temprano tendrá que llegar la petición a un nodo sin vecinos, el cual devolverá el mensaje solicitado sin realizar nuevas peticiones.

Falta demostrar que el valor que devuelve es la probabilidad a posteriori $P(\mathbf{x}_I|\mathbf{e})$. Si el algoritmo 2.3 devolviera una lista de potenciales, sin realizar sumas ni multiplicaciones, tendríamos la primera versión del método de agrupamiento, la cual es correcta porque se limita a recopilar los potenciales. En realidad, cuando el algoritmo 2.3 multiplica potenciales y elimina una variable ello no distorsiona el resultado final, porque al calcular el mensaje M_{ji} sólo se eliminan aquellas variables que no van a aparecer en cálculos posteriores: si una variable pertenece a $\mathbf{C}_i$ no se elimina, porque $\mathbf{R}_{ji} = \mathbf{C}_j \backslash \mathbf{C}_i$; tampoco puede haber una variable que pertenezca a $\mathbf{C}_j$ y a otros grupos conectados con $\mathbf{C}_j$ a través de $\mathbf{C}_i$, porque entonces estaría también en $\mathbf{C}_i$. Ésta es la razón por la que la definición de árbol de grupos incluye la *propiedad del árbol de uniones*: para evitar que al computar los mensajes una variable sea eliminada antes de tiempo.

Vamos a examinar ahora cuál es el dominio de M_{ji}. Hemos visto ya que este mensaje sólo depende de variables de $\mathbf{C}_i$ porque las demás se han eliminado ($\mathbf{R}_{ji} = \mathbf{C}_j \backslash \mathbf{C}_i$), lo cual asegura que todos los mensajes que recibe un grupo dependen exclusivamente de las variables que contiene: $dom(M_{ji}) \subseteq \mathbf{C}_i$. Por otro lado, todos los potenciales asociados a $\mathbf{C}_j$ y todos los mensajes que recibe $\mathbf{C}_j$ dependen sólo de las variables de $\mathbf{C}_j$, lo cual implica que $dom(M_{ji}) \subseteq \mathbf{C}_j$. A la intersección de dos grupos vecinos se le denomina *separador*, y se representa por $\mathbf{S}_{ij}$, de modo que $\mathbf{S}_{ij} = \mathbf{S}_{ji} = \mathbf{C}_i \cap \mathbf{C}_j$ y $dom(M_{ji}) \subseteq \mathbf{S}_{ij}$. En la figura 2.4 hemos representado el mismo árbol de grupos que en la 2.3, pero dibujando el separador de cada par de nodos vecinos. El motivo por el que no hemos señalado la dirección de los enlaces es triple: (1) cualquier nodo del árbol puede ser pivote, y por eso no importa cuál es el nodo raíz, (2) cada nodo solicita mensajes a sus vecinos sin importar si son padres o hijos y (3) los mensajes circulan hacia el nodo pivote sin importar la dirección que tenga cada enlace. Más adelante veremos un cuarto motivo por el que no importa la dirección de cada enlace: si queremos calcular la probabilidad $P(\mathbf{c}_i, \mathbf{e})$ de cada uno de los grupos, entonces por cada enlace $\mathbf{C}_i$–$\mathbf{C}_j$ van a circular dos mensajes, M_{ij} y M_{ji}, uno en cada dirección. Todo ello hace que sea indiferente definir y construir el árbol de grupos como grafo dirigido o como no dirigido, y por eso en la literatura sobre el método de agrupamiento es habitual no dibujar la dirección de los enlaces en el árbol de grupos.

Comparación con eliminación de variables

Vamos a ver que los dos métodos estudiados hasta ahora en este capítulo son mucho más semejantes de lo que a primera vista pudiera parecer. Volviendo al ejemplo 2.12, las operaciones que realiza el método de agrupamiento sobre el árbol de la figura 2.3 son casi las mismas que si aplicáramos el método de eliminación de variables en este orden: $\{H, F, G, D, C, A\}$; la semejanza se observa mejor en la figura 2.4, que representa el mismo árbol de grupos, pero

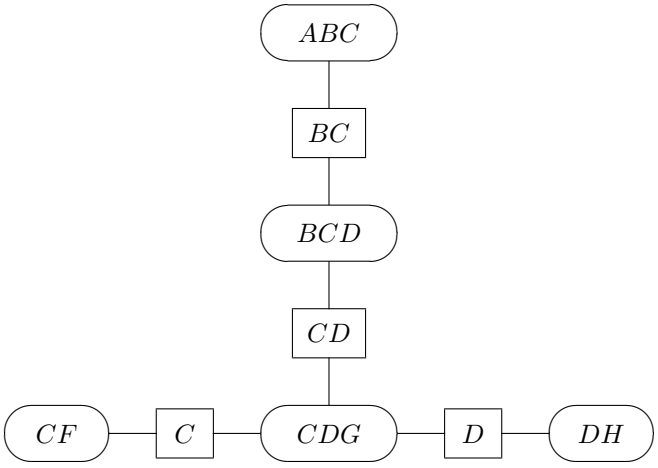


Figura 2.4: Árbol de grupos para la red de la figura 2.2. Es el mismo árbol de la figura 2.3, pero aquí hemos dibujado los separadores de cada par de nodos vecinos y hemos omitido la dirección de los enlaces.

mostrando explícitamente el separador asociado a cada enlace.

En primer lugar, cuando el hallazgo "$\neg h$" forma parte de la evidencia, tenemos que $M_{53}(d) = P(\neg h|d)$. Sin embargo, en general se calcularía así:

$$M_{53}(d) = \sum_{h} P(h|d)$$

Tanto en un caso como en otro, podemos interpretar $M_{53}(d)$ como un potencial $\psi_1(d)$ resultante de eliminar los potenciales que dependen de H (que en este caso es sólo uno).[5]

Del mismo modo, el potencial $M_{43}(c)$, que en este ejemplo es igual a $P(+f|c)$, en general sería

$$M_{43}(c) = \sum_{f} P(f|c)$$

y por tanto $M_{43}(c)$ puede interpretarse como un potencial $\psi_2(c)$ resultante de eliminar los potenciales que dependen de F.

En el método de eliminación de variables, al eliminar G obtendríamos nuevo potencial,

$$\psi_3(c, d) = \sum_{g} P(g|c, d)$$

[5]Recuerde que los subíndices de cada mensaje M indican el grupo que lo envía y el que lo recibe, mientras que el subíndice de ψ indica el orden en que se ha obtenido este potencial al eliminar variables. Por tanto, no hay ninguna relación directa entre los índices de M y el de ψ.

mientras que en el método de agrupamiento teníamos

$$\begin{aligned} M_{32}(c,d) &= \sum_g \phi_3^0(c,d,g) \cdot M_{43}(c) \cdot M_{53}(d) \\ &= M_{43}(c) \cdot M_{53}(d) \cdot \sum_g \phi_3^0(c,d,g) \\ &= \underbrace{P(+f|c)}_{\psi_2(c)} \cdot \underbrace{P(\neg h|d)}_{\psi_1(d)} \cdot \underbrace{\sum_g P(g|c,d)}_{\psi_3(c,d)} \end{aligned}$$

Al eliminar D obtenemos

$$\psi_4(b,c) = \sum_d \underbrace{P(d|b)}_{\phi_2^0(b,d)} \cdot \underbrace{\psi_1(d) \cdot \psi_2(c) \cdot \psi_3(c,d)}_{M_{32}(c,d)} = M_{21}(b,c)$$

Recordamos también que el método de agrupamiento realizaba este cálculo:

$$P(b,+f,\neg h) = \sum_a \sum_c \phi_1^0(a,b,c) \cdot M_{21}(b,c)$$

En cambio, el método de eliminación de variables realizaría éste:

$$P(b,+f,\neg h) = \sum_a P(a) \cdot P(b|a) \cdot \underbrace{\sum_c P(c|a) \cdot \psi_4(b,c)}_{\psi_5(b)}$$

Teniendo en cuenta que

$$\phi_1^0(a,b,c) = P(a) \cdot P(b|a) \cdot P(c|a)$$

y $M_{21}(b,c) = \psi_4(b,c)$, se observa que los dos métodos están realizando casi las mismas operaciones. Tan sólo hay pequeñas diferencias en el orden en que se realizan las multiplicaciones y las sumas, e incluso estas diferencias se podrían haber reducido aún más construyendo un árbol de grupos un poco diferente.

En realidad, se puede demostrar que para toda red bayesiana y todo orden de eliminación de variables existe un árbol de grupos tal que ambos métodos realizan las mismas operaciones [14]. De hecho, más adelante vamos a utilizar esta propiedad para construir un árbol de grupos a partir de un orden de eliminación de variables.

Ahora que hemos visto la semejanza entre ambos métodos, vamos a ver también sus diferencias. Supongamos que, después de calcular $P(b|+f,\neg h)$, queremos calcular también $P(a|+f,\neg h)$, $P(c|+f,\neg h)$ y $P(d|+f,\neg h)$; es decir, tenemos nuevos problemas de inferencia con la misma evidencia pero diferentes variables de interés. Las dos primeras probabilidades pueden obtenerse también a partir de $P(a,b,c,+f,\neg h)$, que es uno de los resultados intermedios obtenidos en el cálculo de $P(b|+f,\neg h)$ mediante el método de agrupamiento; por tanto podemos calcular fácilmente estas dos probabilidades sin propagar nuevos mensajes. El cálculo de $P(d|+f,\neg h)$ puede realizarse también mediante el método de agrupamiento, pero utilizando $\mathbf{C}_2 = \{B,C,D\}$ como pivote. El cálculo se realizaría así: la probabilidad conjunta de las variables de este grupo y de la evidencia es

$$P(b,c,d,+f,\neg h) = \phi_2^0(b,c,d) \cdot M_{12}(b,c) \cdot M_{32}(c,d)$$

y a partir de ahí calcularíamos $P(d, +f, \neg h)$ y $P(d|+f, \neg h)$. Observe que los valores de ϕ_2^0 y M_{32} son los mismos que habíamos obtenido al aplicar el método de agrupamiento por primera vez para calcular $P(b|+f, \neg h)$, por lo que podemos reaprovechar una parte de los cálculos anteriores. La forma general de reaprovechar los resultados intermedios es almacenar cada ϕ_i^0 en el grupo $\mathbf{C}_i$, y los mensajes M_{ij} y M_{ji} en el enlace $\mathbf{C}_i$–$\mathbf{C}_j$. De este modo, el árbol de grupos no sólo indica cómo hay que combinar los potenciales, sino que también puede utilizarse como una *caché*, que permite reducir el tiempo de computación (complejidad temporal) a costa de aumentar la cantidad de memoria requerida (complejidad espacial).

En cambio, en el método de eliminación de variables no es fácil establecer un sistema de almacenamiento (una *caché*) que permita reaprovechar los resultados intermedios, y por ello habría que repetir los cálculos desde cero para cada probabilidad condicional que queramos conocer.

Otra diferencia es que cada árbol de grupos sólo sirve para calcular algunas probabilidades a posteriori y no otras. Como hemos indicado al presentar el algoritmo 2.2, para calcular la probabilidad a posteriori $P(\mathbf{x}_I|\mathbf{e})$ hace falta que en el árbol haya un grupo $\mathbf{C}_p$ que contenga todas las variables de interés: $\mathbf{X}_I \subseteq \mathbf{C}_p$. Por ejemplo, a partir del árbol de la figura 2.3 no podemos calcular $P(a, d|+f, \neg h)$ porque no hay ningún grupo que contenga a la vez A y D.[6] En cambio, el método de eliminación de variables no tiene esta restricción.

De aquí se deduce una regla: cuando tenemos que calcular la probabilidad a posteriori de una sola variable, $P(x|\mathbf{e})$, o de un conjunto de variables, $P(\mathbf{x}_I|\mathbf{e})$, es mejor utilizar el método de eliminación de variables, porque ahorra tiempo y espacio, pues no necesita construir el árbol ni almacenar resultados intermedios; en cambio, cuando tenemos que calcular la probabilidad a posteriori de muchas variables, es mejor aplicar el método de agrupamiento... siempre que nuestro ordenador tenga suficiente memoria. Si no tiene suficiente memoria, deberemos aplicar repetidamente el método de eliminación de variables, aunque sea mucho más lento por la necesidad de repetir los mismos cálculos una y otra vez.

Construcción del árbol (bosque) de grupos

Existen varios métodos para la construcción de un árbol de grupos para una red bayesiana. Un método trivial consistiría en construir un árbol de un solo grupo que incluyera todas las variables. Sin embargo, la inferencia sobre ese árbol sería muy ineficiente, pues el tamaño del potencial asociado a su único grupo crecería exponencialmente con el número de variables de la red. De nuevo estaríamos aplicando el método de fuerza bruta en vez de aprovechar las independencias indicadas por el grafo de la red.

En esta sección vamos a explicar un método mucho más eficiente. Aunque nosotros lo vamos a aplicar especialmente a redes bayesianas, en realidad sirve para cualquier conjunto de potenciales. El primer paso de este método consiste en construir el grafo de dependencias.

Definición 2.14 (Grafo de dependencias) El *grafo de dependencias* para un conjunto de variables $\mathbf{X}$ y un conjunto de potenciales $\{\psi_i\}$, tal que cada $\psi_i(\mathbf{x}_i)$ está definido sobre un subconjunto $\mathbf{X}_i \subseteq \mathbf{X}$,. es un grafo no dirigido tal que entre dos nodos X_i y X_j hay un enlace si y sólo si existe un potencial que incluye ambas variables en su dominio.

Según esta definición, cada potencial de n variables induce $\frac{n\,(n-1)}{2}$ enlaces en el grafo, aunque es posible que un mismo enlace sea inducido por más de un potencial.

[6] Más adelante explicaremos cómo construir un árbol de grupos destinado a calcular la probabilidad a posteriori $P(\mathbf{x}_I|\mathbf{e})$.

Ejemplo 2.15 El grafo de dependencias de $\{\psi_1(a,b,c), \psi_2(c,d)\}$ tendrá cuatro enlaces: A–B, A–C, B–C y C–D.

La probabilidad condicional $P(x|u_1,\ldots,u_n)$ induce n enlaces del tipo X–U_i y $\frac{(n+1)\,n}{2}$ enlaces del tipo U_i–U_j. Por tanto, el grafo de dependencias de una red bayesiana puede construirse a partir del grafo de la red sustituyendo cada enlace dirigido $U_i \to X$ por el enlace no dirigido U_i–X y trazando un enlace entre cada par de nodos que tengan un hijo en común (salvo que ya existiera tal enlace, naturalmente).[7]

Ejemplo 2.16 El grafo de dependencias para la red bayesiana de la figura 2.2 es el que aparece en la figura 2.5.*a*. Como el nodo G tenía dos padres, que no estaban "casados", ha sido necesario trazar un enlace C–D. Los demás nodos tienen menos de dos padres y por eso no es necesario añadir más enlaces.

Un algoritmo para la construcción del árbol de grupos puede ser el siguiente. En realidad, puede ocurrir que este algoritmo no devuelva un solo árbol, sino varios, y por eso es más correcto hablar de *bosque de grupos*.

Algoritmo 2.4 (Construcción del bosque de grupos)
Entrada: conjunto de potenciales definidos sobre $\mathbf{X}$
Salida: bosque de grupos, lista de grupos raíz
construir el grafo de dependencias;
$i :=$ número de variables en $\mathbf{X}$;
grupos-huérfanos := lista vacía;
grupos-raíz := lista vacía;
mientras el grafo contenga nodos,
  seleccionar un nodo para eliminar, X;
  $\mathbf{S}_i :=$ {vecinos de X en el grafo}; *// separador de* $\mathbf{C}_i$
  $\mathbf{C}_i := \{X\} \cup \mathbf{S}_i$;
  para cada grupo $\mathbf{C}_j$ de la lista grupos-huérfanos,
    si $\mathbf{S}_j \subseteq \mathbf{C}_i$ entonces {
      trazar un enlace $\mathbf{C}_i \to \mathbf{C}_j$ en el bosque de grupos;
      sacar $\mathbf{C}_j$ de grupos-huérfanos;
    }
  si $\mathbf{S}_i = \varnothing$ entonces {
    añadir $\mathbf{C}_i$ a grupos-raíz;
  } en otro caso {
    casar los nodos de $\mathbf{S}_i$ que no estuvieran casados;[8]
    añadir $\mathbf{C}_i$ a grupos-huérfanos;
  }
  eliminar el nodo X del grafo de dependencias;
  $i := i - 1$;
;
asignar cada potencial a un grupo que contenga todas sus variables.

[7] Esta operación de "casar" los padres se denomina *moralización*, porque se considera "inmoral" que dos nodos tengan un hijo en común sin estar "casados".

[8] Es decir, trazar un enlace X_k–X_l entre cada par de nodos de $\mathbf{S}_i$ si dicho enlace no existía.

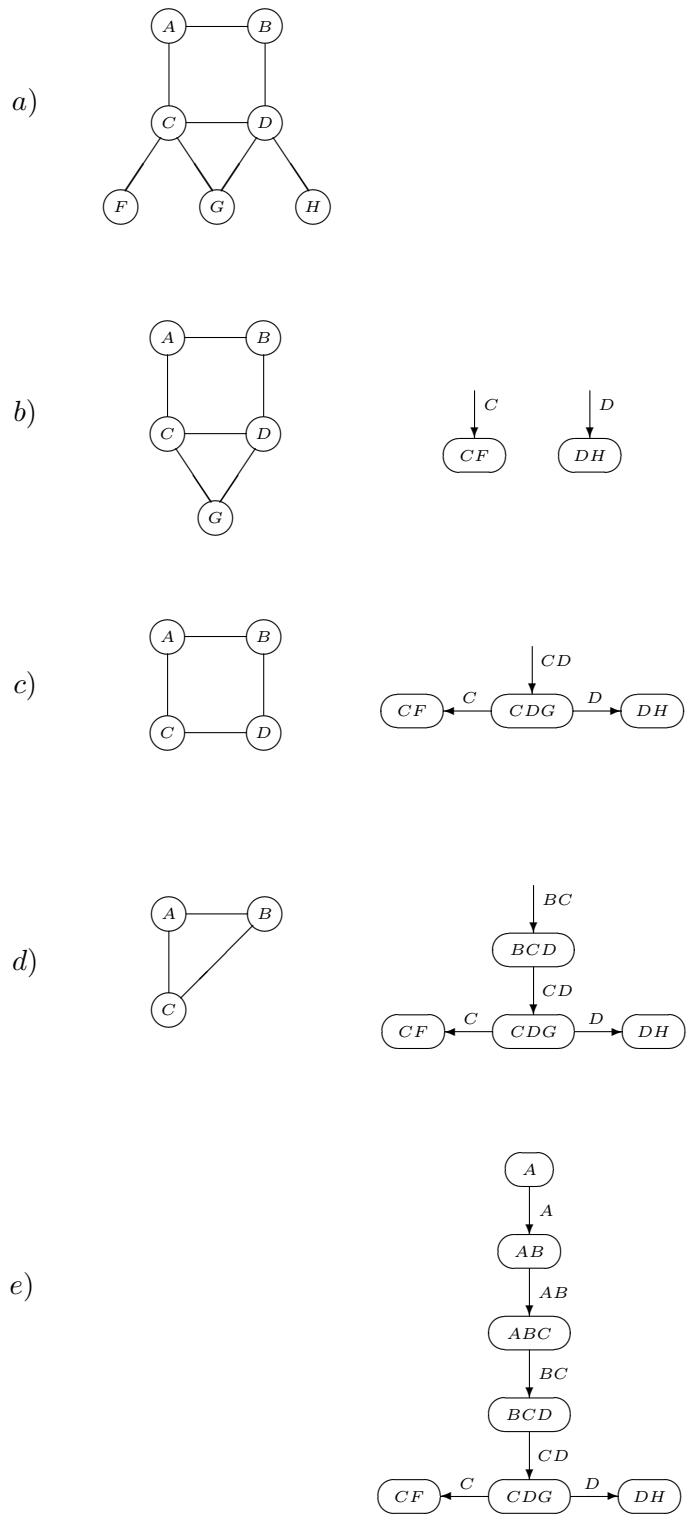


Figura 2.5: Construcción de un árbol de grupos para la red bayesiana de la figura 2.2. A medida que vamos eliminando nodos del grafo de dependencias va creciendo el árbol de grupos.

Ejemplo 2.17 La figura 2.5 ilustra la aplicación del algoritmo a la red bayesiana de la figura 2.2. Primero construimos el grafo de dependencias, que tiene 7 nodos, tantos como la red bayesiana (véase la figura 2.5.*a*). El árbol de grupos aún está vacío.

Seleccionamos H como nodo para eliminar, lo cual nos da el grupo $\mathbf{C}_7 = \{D, H\}$, cuyo separador es $\mathbf{S}_7 = \{D\}$. A continuación eliminamos el nodo F, lo cual nos da el grupo $\mathbf{C}_6 = \{C, F\}$, cuyo separador es $\mathbf{S}_6 = \{C\}$. En ese momento quedan cinco nodos en el grafo de dependencias y el bosque tiene dos grupos, ambos huérfanos (véase la figura 2.5.*b*).

Al eliminar G se forma el grupo $\mathbf{C}_5 = \{C, D, G\}$, cuyo separador es $\mathbf{S}_5 = \{C, D\}$. Dado que C y D ya están casados, no hace falta añadir ningún enlace. Como $\mathbf{S}_6 \subseteq \mathbf{C}_5$ y $\mathbf{S}_7 \subseteq \mathbf{C}_5$, este grupo se convierte en padre de $\mathbf{C}_6$ y $\mathbf{C}_7$, que salen de la lista de grupos huérfanos, y entra en ella $\mathbf{C}_5$ (figura 2.5.*c*).

Al eliminar D tenemos $\mathbf{C}_4 = \{B, C, D\}$ y $\mathbf{S}_4 = \{B, C\}$. Como B y C no están casados, hay que añadir un enlace entre ellos (figura 2.5.*d*). El hecho de que $\mathbf{S}_5 \subseteq \mathbf{C}_4$ permite trazar el enlace $\mathbf{C}_4 \rightarrow \mathbf{C}_5$. Por eso $\mathbf{C}_5$ sale de la lista de grupos huérfanos y entra $\mathbf{C}_4$.

Continuamos eliminando las variables C, B y A, lo cual nos lleva a $\mathbf{C}_3 = \{A, B, C\}$, $\mathbf{S}_3 = \mathbf{C}_2 = \{A, B\}$, $\mathbf{S}_2 = \mathbf{C}_1 = \{A\}$ y $\mathbf{S}_1 = \varnothing$ (figura 2.5.*e*). En este caso el bosque tiene un sólo árbol, cuya raíz es $\mathbf{C}_1 = \{A\}$. □

Ejercicio 2.18 Construya un bosque de grupos para el grafo de dependencias formado por siete variables y los siguientes enlaces: *A*–*B*, *B*–*C*, *D*–*E*, *D*–*F*, *D*–*G*, y *E*–*F*. Solución: el bosque tendrá dos árboles, uno de tres grupos y otro de cuatro.

Vamos a hacer algunas observaciones sobre este algoritmo.

1. El bosque contiene tantos grupos como nodos había en el grafo de dependencias, es decir, tantos como variables había en el conjunto $\mathbf{X}$ sobre el que están definidos los potenciales, pues cada grupo se forma al eliminar una variable.

2. Si el grafo de dependencias de un conjunto de potenciales es conexo, como ocurre en el ejemplo anterior, entonces el algoritmo devolverá un bosque de un solo árbol. (Recordemos que, según la definición 1.59, un grafo es conexo si entre cada par de nodos existe al menos un camino.) Si no lo es, el algoritmo devolverá un árbol por cada uno de los componentes conexos del grafo. En todos los casos, cada árbol tendrá tantos grupos como nodos había en el componente conexo (del grafo de dependencias) que ha dado lugar a ese árbol.

3. Al eliminar un nodo X y formar el grupo $\mathbf{C}_i$, el separador $\mathbf{S}_i$ contiene la intersección de $\mathbf{C}_i$ con los nodos que aún quedan en el grafo de dependencias. También se cumple que $\mathbf{C}_i = \{X\} \cup \mathbf{S}_i$. Por tanto, si $\mathbf{S}_i$ está vacío eso significa que $\mathbf{C}_i$ sólo contiene un nodo y que este nodo no tenía vecinos en el grafo de dependencias, es decir, X era el último nodo de uno de los componentes conexos del grafo. El grupo $\mathbf{C}_i = \{X\}$ se constituirá en la raíz de un árbol; ese árbol contendrá todas (y solamente) las variables conectadas con X en el grafo de dependencias.

4. Todo subconjunto completamente conexo en el grafo de dependencias (es decir, un conjunto tal que entre cada par de nodos existe un enlace) estará contenido en algún grupo del árbol. La razón es que en cuanto se elimine un nodo de ese conjunto se formará un grupo con dicho nodo y todos sus vecinos, el cual incluirá todos los nodos del subconjunto (y quizá algunos nodos más). Por ejemplo, $\{C, G\}$ y $\{C, D, G\}$ son

completamente conexos en el grafo de dependencias de la figura 2.5.*a*; al eliminar el nodo G se añade al árbol un grupo que contiene todas sus variables, como puede verse en la figura 2.5.*e*.

5. Si el potencial ψ_i está definido sobre $\mathbf{X}_i$ entonces los nodos de $\mathbf{X}_i$ forman un subconjunto completamente conexo y por eso habrá un grupo que contenga todas las variables de $\mathbf{X}_i$. Eso es lo que garantiza que siempre será posible asignar ψ_i a algún grupo del árbol. Así se satisface la tercera condición de la definición de árbol de grupos (cf. pág. 46).

6. Al formar el grupo $\mathbf{C}_i$ casamos todos los nodos de $\mathbf{S}_i$. El propósito de esta operación es asegurar que $\mathbf{S}_i$ sea un conjunto completamente conexo en el grafo de dependencias resultante, lo cual nos asegura que si $\mathbf{S}_i$ no está vacío tarde o temprano se va a formar un grupo que contenga todas sus variables y, en consecuencia, adoptará a $\mathbf{C}_i$ como hijo. Eso garantiza que cuando el algoritmo termine no habrá ningún nodo huérfano; véase de nuevo la figura 2.5.*e*.

7. Para demostrar que el algoritmo 2.4 es correcto sólo falta demostrar que cada uno de los árboles formados cumple la propiedad del árbol de uniones introducida en la definición 2.8 (pág. 46). Este resutado queda garantizado por la siguiente proposición.

Proposición 2.19 Cada uno de los árboles formados por el algoritmo 2.4 cumple la propiedad del árbol de uniones, es decir, la condición 2 de la definición 2.8.

Demostración. Tenemos que demostrar que si una variable X aparece en dos grupos cualesquiera del árbol, $\mathbf{C}_i$ y $\mathbf{C}_j$, entonces aparece también en todos los grupos que se encuentran en el camino que hay entre ellos. Examinemos primero el caso en que $\mathbf{C}_i$ sea antepasado de $\mathbf{C}_j$ (piense, por ejemplo, en la variable D y en los grupos $\{B, C, D\}$ y $\{D, H\}$ de la fig. 2.5). En este caso sabemos que $\mathbf{C}_j$, el descendiente, se formó antes que $\mathbf{C}_i$. El hecho de que $X \in \mathbf{C}_i$ implica que X aún estaba en el grafo de dependencia cuando se formó $\mathbf{C}_j$. Por tanto $X \in \mathbf{S}_j$ y también pertenece al padre de $\mathbf{C}_j$ en el grafo. Aplicando el mismo razonamiento repetidamente llegamos a la conclusión de que X pertenece a todos los grupos que están en el camino que asciende desde $\mathbf{C}_j$ hasta $\mathbf{C}_i$.

Si $\mathbf{C}_j$ es antepasado de $\mathbf{C}_i$ la demostración es similar.

Tan sólo falta examinar el caso en que ambos grupos tengan antepasados comunes. Sea $\mathbf{C}_k$ el grupo formado al eliminar la variable X. Obviamente todos los grupos que contienen a X se han formado antes que $\mathbf{C}_k$. Por eso si un grupo $\mathbf{C}_l$ distinto de $\mathbf{C}_k$ contiene a X entonces X pertenece también a su separador, $\mathbf{S}_l$, y a su padre en el árbol, y así sucesivamente, de modo que en la cadena de antepasados de $\mathbf{C}_l$ tarde o temprano aparecerá $\mathbf{C}_k$; todos los grupos que se encuentran en el camino que asciende desde $\mathbf{C}_l$ hasta $\mathbf{C}_k$ contienen la variable X.

Sea $\mathbf{C}_m$ el nodo más alto del camino que hay entre $\mathbf{C}_i$ y $\mathbf{C}_j$. Como $\mathbf{C}_k$ es antepasado de $\mathbf{C}_i$ y de $\mathbf{C}_j$, el grupo $\mathbf{C}_m$ ha de ser necesariamente $\mathbf{C}_k$ o uno de sus descendientes. Por tanto, todos los grupos que se encuentran entre $\mathbf{C}_i$ y $\mathbf{C}_m$, ambos inclusive, contienen a X, y todos los que se encuentran entre $\mathbf{C}_j$ y $\mathbf{C}_m$ también, con lo cual queda demostrada la proposición.

2.2.3. Variantes del método de agrupamiento

Bosque de grupos específico para la evidencia

Según la exposición que hemos seguido en este capítulo, el algoritmo 2.4 construye el árbol de grupos, independientemente de cuál sea la evidencia, y luego es el algoritmo 2.2 el que se

encarga de proyectar los potenciales según la evidencia disponible. Una forma más eficiente de realizar los cálculos consiste en proyectar los potenciales primero y luego construir el árbol de grupos (naturalmente, el algoritmo 2.2 ya no necesitaría proyectar los potenciales sobre la evidencia).

Eso tiene el inconveniente de que hay que construir un grafo diferente para cada caso de evidencia, es decir, para cada conjunto de hallazgos, pero en general ese coste computacional se ve sobradamente compensado por el hecho de tener un grafo de dependencias con menos nodos y menos enlaces, como puede verse en el siguiente ejemplo, lo cual conduce a un árbol de grupos en que la inferencia es mucho más eficiente en tiempo y en consumo de memoria.[9]

Ejemplo 2.20 Dada la red bayesiana de la figura 2.2, vamos a construir un árbol de grupos específico para la evidencia $\mathbf{e} = \{+b, \neg g\}$. Los potenciales que vamos a asignar al árbol son: $P(a)$, $P(+b|a)$, $P(c|a)$, $P(d|+b, c)$, $P(f|c)$, $P(\neg g|c, d)$ y $P(h|d)$. Observe que $P(+b|a)$ es un potencial que sólo depende de la variable A, y $P(\neg g|c, d)$ sólo depende de las variables C y D. En primer lugar, construimos el grafo de dependencias que aparece en la figura 2.6. Este grafo contiene un enlace entre dos variables si y sólo si existe un potencial que dependa de ambas. Un posible árbol de grupos para este grafo es el que se muestra en la figura 2.7. Observe que este árbol tiene tres grupos de dos nodos cada uno, mientras que el de la figura 2.3 tiene tres grupos de tres variables y dos de dos.

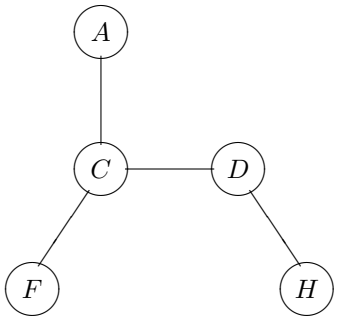


Figura 2.6: Grafo de dependencias para la red de la figura 2.2 y la evidencia $\mathbf{e} = \{+b, \neg g\}$.

Ejercicio 2.21 Construya un bosque de grupos para la red bayesiana de la figura 2.2 y la evidencia $\mathbf{e} = \{+a, +d\}$. ¿Cuántos árboles tiene este bosque?

Ejercicio 2.22 Repita el ejercicio anterior para la evidencia $\mathbf{e} = \{+b, +g\}$.

Para concluir esta sección, vamos a mencionar brevemente la importancia del orden de eliminación de nodos, que es muy semejante a la importancia del orden en el método de eliminación de variables (cf. pág. 44), lo cual no es sorprendente, dada la estrecha relación que existe entre ambos métodos, como ya hemos comentado. De hecho, incluso podemos utilizar los mismos ejemplos.

[9] Un ahorro adicional, en general bastante significativo, puede lograrse si antes de proyectar los potenciales podamos de la red los nodos sumideros, de los cuales hablaremos en la sección 2.2.4.

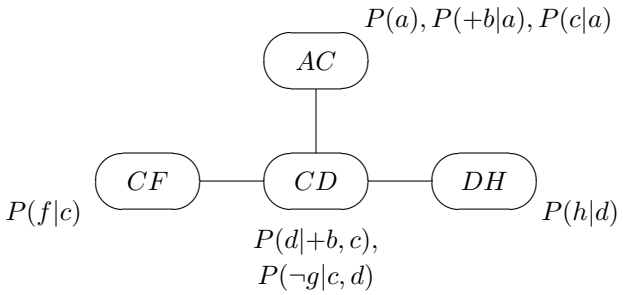


Figura 2.7: Un posible árbol de grupos para la red de la figura 2.2, específico para la evidencia $\mathbf{e} = \{+b, \neg g\}$.

Volvamos de nuevo al grafo de la figura 1.4 (pág. 18). Si el primer nodo eliminado del grafo de dependencias es D, entonces el primer grupo formado contendrá todas las variables: $\mathbf{C}_{m+1} = \{D, H_1, \ldots, H_m\}$; el siguiente, contendrá todas menos D: $\mathbf{C}_m = \{H_1, \ldots, H_m\}$, etc. El mensaje que $\mathbf{C}_{m+1}$ envía a $\mathbf{C}_m$ depende de m variables, el de $\mathbf{C}_m$ a $\mathbf{C}_{m-1}$ de $m-1$ variables, y así sucesivamente. Este método sólo representa una ventaja frente al de fuerza bruta en que puede reaprovechar cálculos si hay que calcular varias probabilidades a posteriori, pero a costa de emplear mucha más memoria. En cambio, si las variables se eliminan en el orden $\{H_1, \ldots, H_m, D\}$ cada grupo sólo contendrá dos variables, con lo que la computación resultará mucho más eficiente: recordemos que la complejidad computacional de este método crece exponencialmente con el tamaño de los grupos.

Como regla general, conviene eliminar primero aquellos nodos que no añaden enlaces al grafo de dependencias. En el ejemplo anterior, el orden $\{H_1, \ldots, H_m, D\}$ no añade ningún enlace, mientras que si se elimina primero D hay que añadir un enlace entre cada par de variables $\{H_i, H_j\}$. Por desgracia, encontrar el orden óptimo de eliminación es un problema NP-completo, pero afortunadamente existen heurísticas que con un coste computacional pequeño suelen dar buenos órdenes de eliminación. Desde luego, la primera condición que debe cumplir una heurística es no añadir enlaces salvo que sea estrictamente necesario, y por eso prácticamente todas las heurísticas propuestas empiezan por eliminar los nodos que tienen todos sus vecinos casados. Una de las heurísticas más sencillas consiste en eliminar primero el nodo que tenga que añadir menos enlaces. Más adelante mencionaremos otras formas de obtener un orden de eliminación.

Esta sección está destinada a alumnos avanzados que quieran estudiar las principales referencias bibliográficas sobre el método de agrupamiento para ampliar los conocimientos presentados en este capítulo. Por eso vamos a explicar aquí las diferencias existentes entre la versión que nosotros hemos presentado y las que se encuentran habitualmente en la literatura. Los alumnos que no vayan a consultar otras fuentes bibliográficas pueden omitir esta sección o, mejor aún, leerla sólo superficialmente.

Grafos triangulados

Uno de los conceptos que aparecen con más frecuencia en la literatura sobre este tema es el de grafo triangulado y triangulación, que se definen así.

Definición 2.23 (Grafo triangulado) Un grafo no dirigido está *triangulado* si para todo ciclo de longitud mayor que tres existe al menos un enlace que une dos nodos no consecutivos

en el ciclo.

Dicho en forma negativa, un grafo es no triangulado si contiene al menos un ciclo de longitud mayor que tres tal que no existe ningún enlace entre un par de nodos no consecutivos de ese ciclo.

Definición 2.24 (Triangulación) La *triangulación* de un grafo consiste en añadir enlaces que lo conviertan en triangulado (si no lo era ya).

Ejemplo 2.25 En la figura 2.5.*a* tenemos un ciclo de longitud cinco, *A*–*B*–*D*–*G*–*C*–*A*; como hay un enlace *C*–*D* entre nodos no consecutivos, este ciclo no impide que el grafo sea triangulado. En cambio, el ciclo *A*–*B*–*D*–*C*–*A* hace que el grafo no sea triangulado, pues es de longitud cuatro y no hay en el grafo ningún enlace que una dos nodos no consecutivos de ese ciclo. Existen dos formas de triangular ese grafo: añadir el enlace *A*–*D* y añadir el enlace *B*–*C*, pues cualquiera de los unirá dos nodos no consecutivos del ciclo. Observe que en el proceso de construcción del árbol de grupos el algoritmo 2.4 (fig. 2.5.*d*) añade el enlace *B*–*C*; si añadiéramos este enlace al grafo de la figura 2.5.*a* obtendríamos el grafo triangulado que aparece en la figura 2.8. □

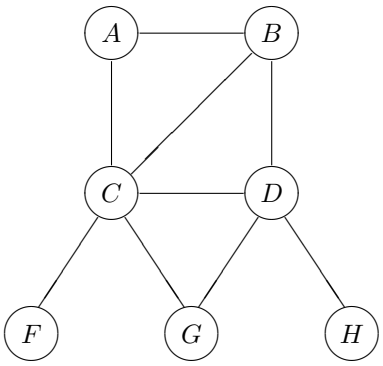


Figura 2.8: Grafo triangulado correspondiente al grafo de dependencias que aparece en la figura 2.5.*a*.

El resultado anterior se puede generalizar, dando lugar a un algoritmo de triangulación de grafos basado en un orden de eliminación de nodos. En realidad, es el mismo algoritmo 2.4, pero prescindiendo de la construcción del árbol de grupos y de las operaciones con potenciales.

Algoritmo 2.5 (Triangulación por eliminación de nodos)
Entrada: grafo no dirigido
Salida: grafo triangulado
hacer una copia del grafo no dirigido,
mientras la copia del grafo contenga nodos,
  seleccionar un nodo para eliminar, X;
  $\mathbf{S}_i := \{$vecinos de X en la copia del grafo$\}$;
  si $\mathbf{S}_i \neq \varnothing$ entonces {
    casar los nodos de $\mathbf{S}_i$ que no estuvieran casados, recordando
      qué enlaces han sido añadidos;

}
eliminar el nodo X de la copia del grafo;
añadir al grafo los enlaces añadidos al eliminar los nodos.

La demostración de que el grafo producido por este algoritmo está triangulado es muy sencilla: si el grafo original tenía un ciclo de longitud n, con $n > 3$, y no había ningún enlace para ningún par de nodos adyacentes, la eliminación de uno de los nodos del ciclo (en la copia del grafo) añadirá un enlace entre sus dos vecinos (tanto en la copia como en el original), que no eran adyacentes en el ciclo, con lo cual ese ciclo ya no puede impedir que el grafo sea triangulado. Ciertamente, el nuevo enlace da lugar a un ciclo de longitud $n-1$ que no estaba en el grafo original, pero también para él se añadirá un enlace entre nodos no adyacentes cuando se elimine alguno de sus nodos, y así sucesivamente.

Imaginemos ahora que a un grafo cualquiera le aplicamos el algoritmo 2.5, obteniendo así un grafo triangulado. En este segundo grafo el primer nodo que hemos eliminado tiene todos sus vecinos casados, y podemos eliminarlo sin tener que añadir enlaces. Este resultado se puede generalizar a cualquier grafo triangulado (no sólo a los que se han triangulado por eliminación de nodos):

Proposición 2.26 En un grafo triangulado siempre hay al menos un nodo que tiene todos sus vecinos casados.

Al eliminar este nodo, el grafo resultante también será triangulado, por lo que podemos aplicar la operación repetidamente hasta eliminar todos los nodos sin haber añadido ningún enlace.

En un grafo no triangulado, como el de la figura 2.5.*a*, puede haber nodos que tengan todos sus vecinos casados, pero si eliminamos estos nodos, tarde o temprano llegaremos a un grafo en que no habrá ningún nodo que tenga todos sus vecinos casados (fig. 2.5.*c*). Para poder seguir eliminado nodos tendremos que añadir enlaces que contribuyan a la triangulación de este grafo.

Conglomerados de un grafo triangulado

En este apartado vamos a introducir el concepto de árbol de conglomerados y a ver su relación con el de grafo triangulado.

Definición 2.27 (Conglomerado) En un grafo no dirigido, un *conglomerado* es todo subconjunto de nodos completamente conexo maximal.

Definición 2.28 (Bosque de conglomerados) Dado un grafo no dirigido, el *bosque de conglomerados* está formado por un conjunto de árboles tales que cada nodo (de un árbol) representa un conglomerado (del grafo original) y cada árbol cumple la propiedad del árbol de uniones.

Ejemplo 2.29 Los conglomerados del grafo de la figura 2.8 son $\{A, B, C\}$, $\{B, C, D\}$, $\{C, F\}$, $\{C, D, G\}$ y $\{D, H\}$. Los conjuntos $\{A\}$ y $\{A, B\}$ son completamente conexos pero no son conglomerados porque no son maximales, dado que también $\{A, B, C\}$ es completamente conexo. Un posible árbol de conglomerados para este grafo es el de la figura 2.3 (pág. 47). Observe que el árbol de esta figura es muy similar al de la fig. 2.5.*e*; la diferencia principal es que en el primero sólo aparecen los conglomerados.

Ejercicio 2.30 Construya un bosque de conglomerados para el grafo del ejercicio 2.18. Solución: se puede construir un bosque formado por dos árboles, de dos conglomerados cada uno. Compare este resultado con el bosque de grupos que obtuvo en el ejercicio 2.18.

La diferencia principal entre árbol de grupos y árbol de conglomerados es que el primero se refiere a un conjunto de potenciales y el segundo a un árbol no dirigido; sin embargo, si nos fijamos en el árbol de dependencias asociado al conjunto de potenciales, esta diferencia desaparece. Por otro lado, también es posible asignar potenciales a los nodos de un árbol de conglomerados, como veremos más adelante. Por tanto, la diferencia principal entre ambos es que en un árbol de grupos —como el de la figura 2.5.*e*— cada nodo representa un conjunto de nodos (del grafo de dependencias) que no tiene por qué ser completamente conexo ni maximal,[10] mientras que en un árbol de conglomerados —como el de la figura 2.3 o el que ha obtenido en el ejercicio 2.30— cada nodo representa un conglomerado, es decir, un subconjunto completamente conexo maximal. Por eso puede decirse que un árbol de conglomerados es un caso particular de árbol de grupos.

Enunciamos ahora una proposición que no vamos a demostrar, pero que nos va a ser muy útil como fundamento de un algoritmo que construye un bosque de conglomerados.

Proposición 2.31 Los conglomerados de un grafo triangulado pueden organizarse en un bosque de conglomerados. Cada componente conexo del grafo triangulado da lugar a un árbol.

Corolario 2.32 Los conglomerados de un grafo triangulado conexo pueden organizarse en un árbol de conglomerados.

Ejemplo 2.33 En ejemplo anterior hemos visto que los conglomerados del grafo de la figura 2.8, que está triangulado, pueden organizarse en un árbol de conglomerados. En cambio, el grafo de la figura 2.5.*c* no está triangulado, y sus conglomerados, que son $\{A, B\}$, $\{A, C\}$, $\{B, D\}$ y $\{C, D\}$, no pueden organizarse en un árbol de grupos. Por ejemplo, si probamos con el árbol $\{A, C\} \rightarrow \{A, B\} \rightarrow \{B, D\} \rightarrow \{C, D\}$ vemos que no cumple la propiedad del árbol de uniones porque la variable C aparece en el raíz y en el nodo hoja pero no aparece en los grupos intermedios.

La proposición anterior puede demostrarse mediante el siguiente algoritmo; para que el algoritmo sea correcto, es necesario que el nodo X seleccionado en cada iteración tenga todos sus vecinos casados, lo cual es siempre posible, por la proposición 2.26.

Algoritmo 2.6 (Construcción del bosque de conglomerados)
Entrada: grafo triangulado
Salida: bosque de conglomerados, lista de conglomerados raíz
construir el grafo de dependencias;
conglomerados-huérfanos := lista vacía;
conglomerados-raíz := lista vacía;

[10] Sin embargo, el algoritmo 2.4 sólo forma un grupo cuando todos los nodos están casados, aunque para ello tenga que añadir algunos enlaces. Estos enlaces son los mismos que se necesitan para triangular el grafo. Por eso, cada grupo del árbol de grupos representa un conjunto completamente conexo del grafo de dependencias, aunque no todos los grupos son maximales. Por ejemplo, en el árbol de la figura 2.5.*e* los grupos $\{A\}$ y $\{A, B\}$ no son maximales porque están contenidos en $\{A, B, C\}$, que sí es completamente conexo maximal, es decir, representa un conglomerado del grafo triangulado de la figura 2.8.

$i := 1$;
mientras el grafo contenga nodos {
  seleccionar un nodo para eliminar, X;
  $\mathbf{C}_i := \{X$ y sus vecinos en el grafo$\}$;
  $\mathbf{S}_i := \{$vecinos de $\mathbf{C}_i$ que tienen vecinos fuera de $\mathbf{C}_i\}$; *// separador de* $\mathbf{C}_i$
  para cada conglomerado $\mathbf{C}_j$ de la lista conglomerados-huérfanos,
    si $\mathbf{S}_j \subseteq \mathbf{C}_i$ entonces {
      trazar un enlace $\mathbf{C}_i \rightarrow \mathbf{C}_j$ en el bosque de conglomerados;
      sacar $\mathbf{C}_j$ de conglomerados-huérfanos;
    }
  si $\mathbf{S}_i = \varnothing$ entonces {
    añadir $\mathbf{C}_i$ a conglomerados-raíz;
  } en otro caso {
    añadir $\mathbf{C}_i$ a conglomerados-huérfanos;
  }
  eliminar del grafo no dirigido los nodos de $\mathbf{C}_i \backslash \mathbf{S}_i$;
  $i := i + 1$;
}.

Ejercicio 2.34 Aplique este algoritmo al grafo de la figura 2.8 y compruebe que se obtiene el mismo árbol de grupos de la figura 2.3, en el cual cada grupo es un conglomerado. (La diferente numeración de los grupos es irrelevante.)

Ejercicio 2.35 Aplique este algoritmo al grafo del ejercicio 2.18. ¿Obtiene el mismo resultado que en el ejercicio 2.30?

Las diferencias de este algoritmo frente al 2.4 son las siguientes:

1. En cada iteración el algoritmo 2.4 elimina un solo nodo, mientras que éste puede eliminar varios nodos. Por ejemplo, si tenemos un conjunto de nodos completamente conexo $\{A, B, C, D\}$ y sólo A y B tienen otros vecinos, al eliminar D eliminaremos también C.

2. Este algoritmo toma como entrada un grafo triangulado y por tanto siempre puede eliminar algún nodo sin necesidad de añadir enlaces. Sin embargo, podríamos modificar este algoritmo para que acepte grafos no triangulados: basta añadir la línea de código "casar los nodos de $\mathbf{S}_i$ que no estuvieran casados" antes de la línea "añadir $\mathbf{C}_i$ a grupos-huérfanos", como en el algoritmo 2.4.

3. El algoritmo 2.4 toma como entrada un conjunto de potenciales y al final asigna cada potencial a un grupo del árbol (mejor dicho, a un grupo de algún árbol del bosque), mientras que en el 2.6 no se hace ninguna referencia a potenciales. Sin embargo, también esta diferencia puede soslayarse modificando el algoritmo 2.6 para que trabaje sobre el grafo de dependencias asociado a un conjunto de potenciales y al final asigne los potenciales a los conglomerados.

4. El orden de numeración de los grupos o conglomerados es diferente.

Como hemos visto, de estas cuatro diferencias, sólo la primera es relevante. Las demás pueden eliminarse modificando ligeramente el segundo de los algoritmos.

Vamos a demostrar ahora que el algoritmo 2.6 es correcto, es decir, que da como resultado un árbol de conglomerados. Este resultado se deduce de las dos proposiciones siguientes:

Proposición 2.36 Los árboles formados por el algoritmo 2.6 cumplen la propiedad del árbol de uniones.

La demostración es análoga a la de la proposición 2.19.

Proposición 2.37 Los conjuntos $\{\mathbf{C}_i\}$ formados por el algoritmo 2.6 son conglomerados.

Demostración. Cada conjunto $\mathbf{C}_i$ es completamente conexo porque hemos escogido X con la condición de que todos sus vecinos estuvieran casados.

Vamos a demostrar que son maximales, es decir, que ninguno está incluido en otro. Examinemos ahora dos nodos del bosque tales que $\mathbf{C}_i$ es padre de $\mathbf{C}_j$. Se cumple que $\mathbf{C}_j \not\subset \mathbf{C}_i$ porque $\mathbf{C}_j$ contiene al menos el nodo que al ser eliminado condujo a la formación de $\mathbf{C}_j$; naturalmente, este nodo, ya eliminado, no puede pertenecer a $\mathbf{C}_i$, que se ha formado más tarde. Por otro lado, si el grupo $\mathbf{C}_i$ se ha formado al eliminar un nodo que no pertenece a $\mathbf{S}_j$ entonces $\mathbf{C}_i \not\subset \mathbf{C}_j$, y si $\mathbf{C}_i$ se forma al eliminar un nodo X perteneciente a $\mathbf{S}_j$ entonces $\mathbf{C}_i$ contendrá también el vecino de X que no pertenece a $\mathbf{C}_j$ (véase la definición de $\mathbf{S}_j$), lo cual implica que $\mathbf{C}_i \not\subset \mathbf{C}_j$.

Por tanto, si dos nodos del árbol, $\mathbf{C}_i$ y $\mathbf{C}_j$, son adyacentes, ninguno puede estar contenido en el otro. Esta propiedad se cumple también para nodos no adyacentes, pues por la propiedad del árbol de uniones, si uno estuviera incluido en el otro también estaría incluido en todos los nodos que se encuentran en el camino entre ambos, incluido el nodo adyacente al primero, lo cual es imposible. □

Probablemente el lector se esté haciendo esta pregunta: "entonces, ¿qué es mejor: un árbol de grupos o un árbol de conglomerados?" En realidad, dado que el segundo es un caso particular del primero, como ya hemos explicado, la pregunta anterior puede formularse así: "¿merece la pena tener en un árbol de grupos algún grupo que sea subconjunto de otro?". A primera vista la respuesta es negativa, pues si $\mathbf{C}_j \subset \mathbf{C}_i$ entonces cualquier probabilidad a posteriori que podamos calcular a partir del primero podemos calcularla también a partir del segundo, y por tanto $\mathbf{C}_j$ es innecesario. Por eso la mayor parte de la investigación sobre el método de agrupamiento se ha centrado en el estudio de los árboles de conglomerados, siguiendo la propuesta inicial de Lauritzen y Spiegelhalter [50]. Sin embargo, los grupos que son un subconjunto propio de otros pueden contruibuir a almacenar resultados intermedios, con lo cual se consigue ahorrar tiempo de computación a costa de consumir más memoria, tal como demostraron experimentalmente Lepar y Shenoy [51]; para conseguir este ahorro conviene organizar los grupos en forma de árbol binario, es decir, un árbol en que cada grupo tenga sólo dos hijos [73].

Algoritmos para determinar el orden de eliminación de los nodos

Una forma de triangular un grafo es ir eliminando sus nodos en cierto orden, según el algoritmo 2.5. Otra forma de hacerlo es mediante el *algoritmo de máxima cardinalidad*, que va numerando los nodos y añadiendo enlaces sin eliminar los nodos. Ahora bien, si la numeración de los nodos resultante de este algoritmo es $\{X_1, \ldots, X_n\}$, entonces los enlaces añadidos son los mismos que si hubiéramos aplicado el algoritmo 2.5 eliminando los nodos en el orden inverso, $\{X_1, \ldots, X_n\}$. Eso implica que cualquier triangulación obtenida mediante máxima

cardinalidad podría haberse obtenido también por eliminación de nodos. Sin embargo, lo recíproco no es cierto: hay ciertas triangulaciones resultantes de la eliminación de nodos que no pueden obtenerse mediante el algoritmo de máxima cardinalidad.

Otro inconveniente del método de máxima cardinalidad es que no aplica ninguna heurística: cuando hay varios nodos que tienen el mismo número de vecinos numerados, en vez de aplicar algún criterio para seleccionar el mejor candidato entre ellos, se escoge un nodo arbitrariamente, lo cual hace que los resultados sean casi siempre peores que si se hubiera empleado una heurística de numeración de nodos.

A pesar de estas desventajas, hemos mencionado aquí el algoritmo de máxima cardinalidad porque el primer artículo en que se propuso el método de agrupamiento para inferencia en redes bayesianas [50] utilizaba dicho algoritmo para triangular el grafo de dependencias (que en aquel artículo se denominaba "grafo moral"), y desde entonces otros libros de texto han seguido exponiéndolo así.

En cambio, nosotros hemos explicado cómo triangular un grafo mediante eliminación de variables. Más aún, hemos propuesto un algoritmo capaz de formar un árbol de grupos o un árbol de conglomerados a la vez que triangula el grafo. Como heurística de eliminación, hemos mencionado la del "mínimo número de enlaces añadidos", que consiste en eliminar primero el nodo que necesite añadir menos enlaces para casar todos sus vecinos. Naturalmente, eso implica eliminar primero los nodos que tienen todos sus vecinos casados, de modo que si un grafo ya está triangulado esta heurística no añade ningún enlace nuevo (cf. proposición 2.26).

Sin embargo, existen otras heutísticas más eficientes, como la de Cano y Moral [3], que con un coste computacional pequeño consiguen órdenes de eliminación cercanos al óptimo.

2.2.4. Inversión de arcos

Vamos a explicar en esta sección el método de inversión de arcos, que, aunque inicialmente fue propuesto para diagramas de influencia, también es aplicable a redes bayesianas. Primero veremos cómo invertir un arco[11] y luego cómo eliminar nodos sumideros.

Cómo invertir un arco

Primero nos vamos a centrar en el ejemplo más sencillo, que consiste en una red de sólo dos nodos, luego estudiaremos un caso con cinco nodos y finalmente el caso más general.

Ejemplo 2.38 Sea una red bayesiana formada por dos variables, X e Y, y un enlace $X \to Y$. Las probabilidades que definen esta red son $P(x)$ y $P(y|x)$. La probabilidad conjunta resultante es

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

Ahora queremos construir una red bayesiana equivalente a la anterior, es decir, que contenga las mismas variables, X e Y, y que represente la misma probabilidad conjunta $P(x, y)$ pero que tenga un enlace $Y \to X$ en vez de $X \to Y$. Las probabilidades que definen esta nueva red son $P(y)$ y $P(x|y)$, que aún no las conocemos, pero que podemos calcular así:

$$P(y) = \sum_x P(x, y) = \sum_x P(x) \cdot P(y|x) \tag{2.7}$$

[11] En esta sección los términos *arco* y *enlace* son sinónimos. Utilizaremos más el primero por ser el que aparece en la literatura que trata este método.

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

Observe que esta última ecuación es precisamente el teorema de Bayes.

En resumen, en este ejemplo la inversión del arco $X \to Y$ ha constido en que, a partir de una red bayesiana cuyas probabilidades condicionales son $P(x)$ y $P(y|x)$, hemos construido una red bayesiana equivalente cuyas probabilidades son $P(y)$ y $P(x|y)$. Estas últimas las hemos obtenido a partir de las primeras mediante la aplicación del teorema de Bayes.

Ejemplo 2.39 Sea una red bayesiana formada por cinco variables, A, B, C, X e Y, y los siguientes enlaces: $A \to X$, $B \to X$, $B \to Y$, $C \to Y$ y $X \to Y$, tal como se muestra en la figura 2.9. Las probabilidades que definen esta red son $P(a)$, $P(b)$, $P(c)$, $P(x|a,b)$ y $P(y|b,c,x)$. Nos interesan especialmente las dos últimas. La probabilidad conjunta es

$$P(a,b,c,x,y) = P(a) \cdot P(b) \cdot P(c) \cdot \underbrace{P(x|a,b) \cdot P(y|x,b,c)} \tag{2.9}$$

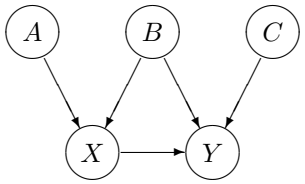


Figura 2.9: Red de cinco nodos y cinco enlaces.

Proposición 2.40 En la red bayesiana anterior,

$$P(x,y|a,b,c) = P(x|a,b) \cdot P(y|x,b,c) \tag{2.10}$$

Demostración. La regla de la cadena nos dice que

$$P(x,y|a,b,c) = P(y|a,b,c,x) \cdot P(x|a,b,c) \tag{2.11}$$

La propiedad de Markov afirma que "en una red bayesiana todo nodo es independiente de sus no-descendientes dados sus padres". En la red original A no es descendiente de Y y $Pa(Y) = \{B, C, X\}$; por tanto,

$$P(y|a,b,c,x) = P(y|b,c,x) \tag{2.12}$$

Del mismo modo, C no es descendiente de X y $Pa(X) = \{A, B\}$; por tanto,

$$P(x|a,b,c) = P(x|a,b) \tag{2.13}$$

Uniendo estos tres resultados queda demostrada la proposición. □

Ahora vamos a aplicar de nuevo la regla de la cadena, como en la ecuación (2.11), pero intercambiando los papeles de X e Y:

$$P(x,y|a,b,c) = P(x|a,b,c,y) \cdot P(y|a,b,c) \tag{2.14}$$

Tanto $P(x|a,b,c,y)$ como $P(y|a,b,c)$ pueden calcularse a partir de $P(x,y|a,b,c)$, que a su vez puede calcularse según la ecuación (2.10):

$$P(y|a,b,c) = \sum_x P(x,y|a,b,c) = \sum_x P(x|a,b) \cdot P(y|b,c,x) \tag{2.15}$$

$$P(x|a,b,c,y) = \frac{P(x,y|a,b,c)}{P(y|a,b,c)} = \frac{P(x|a,b) \cdot P(y|b,c,x)}{\sum_{x'} P(x'|a,b) \cdot P(y|b,c,x')} \tag{2.16}$$

Esta última ecuación es una aplicación del teorema de Bayes con condicionamiento.

A partir de estos resultados podemos construir una nueva red bayesiana formada por las mismas variables que la original, A, B, C, X e Y, y siete enlaces (véase la figura 2.10): cuatro de ellos son los mismos que en la red original: $A \to X$, $B \to X$, $B \to Y$ y $C \to Y$; el quinto, $Y \to X$, es el resultado de invertir el enlace $X \to Y$; y los dos últimos, $A \to Y$ y $C \to X$, se deben a que en la nueva red los nodos X e Y comparten sus padres: el nodo A, que en la red original sólo era padre de X, en la nueva red va a ser también padre de Y, y el nodo C, que sólo era padre de Y, ahora también va a ser padre de X. El nodo B, que era padre de ambos en la red original, sigue siéndolo en la nueva red.

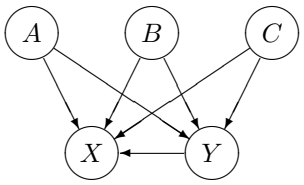


Figura 2.10: Red resultante de invertir el enlace $X \to Y$ en la red anterior.

Las probabilidades condicionales de la nueva red son $P(a)$, $P(b)$, $P(c)$, $P(y|a,b,c)$ y $P(x|a,b,c,y)$, y dan lugar a la siguiente probabilidad conjunta.

$$P'(a,b,c,x,y) = P(a) \cdot P(b) \cdot P(c) \cdot \underbrace{P(y|a,b,c) \cdot P(x|a,b,c,y)} \tag{2.17}$$

De las ecuaciones (2.10) y (2.14) se deduce que

$$P(x|a,b) \cdot P(y|b,c,x) = P(y|a,b,c) \cdot P(x|a,b,c,y) \tag{2.18}$$

y por tanto las dos redes representan la misma probabilidad conjunta: $P'(a,b,c,x,y) = P(a,b,c,x,y)$.

Caso general Vamos a estudiar finalmente el caso general, que es semejante al anterior, con dos diferencias: la primera, que en vez de tener tres nodos, A, B y C, vamos a tener tres conjuntos de nodos $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$, y la segunda, que ahora los nodos de $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$ pueden tener antepasados y descendientes, e incluso puede haber enlaces entre los nodos de $\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}$, y tanto X como Y pueden tener otros descendientes. El problema se plantea así:

Sean dos nodos X e Y que forman parte de una red bayesiana, tales que existe un enlace $X \to Y$ y no existe ningún otro camino dirigido desde X hasta Y. Sean $\mathbf{A}$ el conjunto de

nodos (de azar y/o de decisión) que son padres de X y no de Y, $\mathbf{C}$ el conjunto de nodos que son padres de Y (excepto X) pero no de X, y $\mathbf{B}$ el conjunto de los padres comunes; es decir,

$$\mathbf{A} = Pa(X) \backslash Pa(Y) \tag{2.19}$$
$$\mathbf{B} = Pa(X) \cap Pa(Y) \tag{2.20}$$
$$\mathbf{C} = [Pa(Y) \backslash \{X\}] \backslash Pa(X) = Pa(Y) \backslash [\{X\} \cup Pa(X)] \tag{2.21}$$

Las probabilidades condicionales de estos nodos son $P(x|\mathbf{a}, \mathbf{b})$ y $P(y|\mathbf{b}, \mathbf{c}, x)$.

Proposición 2.41 Para la red anterior se cumple que

$$P(x, y|\mathbf{a}, \mathbf{b}, \mathbf{c}) = P(x|\mathbf{a}, \mathbf{b}) \cdot P(y|x, \mathbf{b}, \mathbf{c}) \tag{2.22}$$

Demostración. La demostración es análoga a la de la proposición 2.40. Por la regla de la cadena,

$$P(x, y|\mathbf{a}, \mathbf{b}, \mathbf{c}) = P(y|\mathbf{a}, \mathbf{b}, \mathbf{c}, x) \cdot P(x|\mathbf{a}, \mathbf{b}, \mathbf{c}) \tag{2.23}$$

Sabemos también que ningún nodo de $\mathbf{A}$ es descendiente de Y —porque en ese caso la red tendría un ciclo— y que $Pa(Y) = \mathbf{B} \cup \mathbf{C} \cup \{X\}$; por tanto, la propiedad de Markov nos dice que

$$P(y|\mathbf{a}, \mathbf{b}, \mathbf{c}, x) = P(y|\mathbf{b}, \mathbf{c}, x) \tag{2.24}$$

También sabemos que ningún nodo de $\mathbf{C}$ puede ser descendiente de X —porque en ese caso habría otro camino desde X hasta Y, en contra de la condición impuesta al definir esta red— y que $Pa(X) = \mathbf{A} \cup \mathbf{B}$; por tanto, aplicando de nuevo la propiedad de Markov llegamos a

$$P(x|\mathbf{a}, \mathbf{b}, \mathbf{c}) = P(x|\mathbf{a}, \mathbf{b}) \tag{2.25}$$

Uniendo estos tres resultados queda demostrada la proposición. □

A partir de esta red bayesiana podemos construir una nueva red equivalente en la cual el enlace $X \to Y$ ha sido sustituido por $Y \to X$ y ambos nodos comparten sus padres, es decir, se añade un enlace desde cada nodo de $\mathbf{A}$ hasta Y y desde cada nodo de $\mathbf{C}$ hasta X. Las probabilidades de estos nodos en la nueva red pueden calcularse a partir de $P(x, y|\mathbf{a}, \mathbf{b}, \mathbf{c})$:

$$P(y|\mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_x P(x, y|\mathbf{a}, \mathbf{b}, \mathbf{c}) \tag{2.26}$$

$$P(x|\mathbf{a}, \mathbf{b}, \mathbf{c}, y) = \frac{P(x, y|\mathbf{a}, \mathbf{b}, \mathbf{c})}{P(y|\mathbf{a}, \mathbf{b}, \mathbf{c})} \tag{2.27}$$

La equivalencia de ambas redes se justifica por la siguiente ecuación:

$$P(x|\mathbf{a}, \mathbf{b}) \cdot P(y|\mathbf{b}, \mathbf{c}, x) = P(x, y|\mathbf{a}, \mathbf{b}, \mathbf{c}) = P(x|\mathbf{a}, \mathbf{b}, \mathbf{c}, y) \cdot P(y|\mathbf{a}, \mathbf{b}, c) \tag{2.28}$$

En resumen, la inversión de un arco consta de cinco pasos:

1. Invertir el arco en el grafo.
2. Compartir padres.

3. Calcular $P(x, y|\mathbf{a}, \mathbf{b}, \mathbf{c})$ mediante la ecuación (2.22).

4. Calcular $P(y|\mathbf{a}, \mathbf{b}, \mathbf{c})$ mediante la ecuación (2.26) y asignar esta probabilidad al nodo Y.

5. Calcular $P(x|\mathbf{a}, \mathbf{b}, \mathbf{c}, y)$ mediante la ecuación (2.27) y asignar esta probabilidad al nodo X.

Para recordar más fácilmente estos pasos, observe su semejanza con lo que hicimos en el ejemplo 2.38: partiendo de $P(x)$ y $P(y|x)$, que son las probabilidades de la red original, calculamos la probabilidad $P(x, y)$ y de ella obtenemos primero $P(y)$ y luego $P(x|y)$. La única diferencia es que ahora los nodos de $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$ (todos los padres de X e Y) actúan como variables condicionantes.

Hemos visto ya que la condición de que la red original no contenga ningún otro camino dirigido desde X hasta Y era necesaria para demostrar la proposición 2.41. Ahora tenemos una razón más que justifica dicha condición: si hubiera otro camino dirigido desde X hasta Y, ese camino, junto con el enlace $Y \to X$ de la nueva red, formaría un ciclo, y por tanto el nuevo modelo ya no sería una red bayesiana. Por eso, antes de invertir un arco debemos comprobar siempre que no existe otro camino dirigido entre ambos nodos.

Poda de sumideros

La siguiente definición hace referencia a los subconjuntos $\mathbf{E}$ (evidencia), $\mathbf{X}_I$ (variables de interés) y $\mathbf{X}_R$ (resto de las variables) definidos en la página 40.

Definición 2.42 (Sumidero) Sea una red bayesiana $\mathcal{B} = (\mathbf{X}, \mathcal{G}, P(\mathbf{X}))$ y dos subconjuntos de variables de ella, $\mathbf{E}$ y $\mathbf{X}_I$. Un nodo S es un *sumidero* si y sólo si $S \in \mathbf{X}_R = \mathbf{X} \backslash (\mathbf{E} \cup \mathbf{X}_I)$ y no tiene hijos.

Ejemplo 2.43 Dada la red de la figura 2.2 y los subconjuntos $\mathbf{X}_I = \{A, B\}$ y $\mathbf{E} = \{D, F\}$, los sumideros son G y H.

Definición 2.44 (Poda de un nodo sin hijos) La poda de un nodo sin hijos S de una red bayesiana $\mathcal{B} = (\mathbf{X}, \mathcal{G}, P(\mathbf{X}))$ consiste en construir una nueva red $\mathcal{B}' = (\mathbf{X}', \mathcal{G}', P'(\mathbf{X}'))$ tal que $\mathbf{X}' = \mathbf{X} \backslash \{S\}$, el grafo $\mathcal{G}'$ se obtiene a partir de $\mathcal{G}$ eliminado el nodo S y todos los enlaces que llegaban a S, y la probabilidad conjunta de $\mathcal{B}'$ es

$$P'(\mathbf{x}') = \sum_s P(\mathbf{x}) \tag{2.29}$$

Para demostrar que la definición es consistente, tenemos que comprobar que $\mathcal{B}'$ es efectivamente una red bayesiana (cf. def. 1.74, pág. 26). Está claro que $\mathcal{G}'$ es un grafo dirigido acíclido y que cada uno de sus nodos representa una variable de $\mathbf{X}'$. Por otro lado,

$$P'(\mathbf{x}') = \sum_s P(\mathbf{x}) = \prod_{i|X_i \neq S} P(x_i|pa(X_i)) \underbrace{\sum_s P(s|pa(S))}_{1} = \prod_{i|X_i \neq S} P(x_i|pa(X_i)) \tag{2.30}$$

y además

$$P(x_i|pa(X_i)) = P'(x_i|pa(X_i)) \tag{2.31}$$

(la demostración de esta ecuación queda como ejercicio para el lector), de modo que se cumple la ecuación (1.42) y por tanto $\mathcal{B}'$ es una red bayesiana. La ecuación (2.31) es importante no sólo como paso intermedio para llegar a demostrar la ecuación (1.42), sino también porque nos dice que las probabilidades condicionales que definen la red podada $\mathcal{B}'$ son las mismas que las que teníamos para $\mathcal{B}$ (excepto la del nodo podado, naturalmente).

Proposición 2.45 Sea una red bayesiana $\mathcal{B}$, dos subconjuntos de variables $\mathbf{X}_I$ y $\mathbf{E}$, y un sumidero S. Sea $\mathcal{B}'$ la red resultante de podar S. La probabilidad $P'(\mathbf{x}_I, \mathbf{e})$, obtenida a partir de $\mathcal{B}'$, es la misma que $P(\mathbf{x}_I, \mathbf{e})$, obtenida a partir de $\mathcal{B}$.

Demostración. Tenemos que $\mathbf{X}_R = \mathbf{X}\backslash(\mathbf{X}_I \cup \mathbf{E})$, $\mathbf{X}' = \mathbf{X}\backslash\{S\}$ y $\mathbf{X}'_R = \mathbf{X}'\backslash(\mathbf{X}_I \cup \mathbf{E}) = \mathbf{X}_R\backslash\{S\}$. Por la ecuación (2.30),

$$P'(\mathbf{x}_I, \mathbf{e}) = \sum_{\mathbf{x}'_R} P'(\mathbf{x}) = \sum_{\mathbf{x}'_R} \prod_{i|X_i \neq S} P(x_i|pa(X_i))$$

Por otro lado,

$$\begin{aligned} P(\mathbf{x}_I, \mathbf{e}) &= \sum_{\mathbf{x}_R} P(\mathbf{x}) = \sum_{\mathbf{x}_R} \prod_i P(x_i|pa(X_i)) \\ &= \sum_{\mathbf{x}'_R} \sum_s P(s|pa(S)) \prod_{i|X_i \neq S} P(x_i|pa(X_i)) \\ &= \sum_{\mathbf{x}'_R} \prod_{i|X_i \neq S} P(x_i|pa(X_i)) \underbrace{\sum_s P(s|pa(S))}_{1} = P'(\mathbf{x}_I, \mathbf{e}) \end{aligned}$$

con lo que queda demostrada la proposición.

Ejemplo 2.46 Dada la red de la figura 2.2, supongamos que queremos calcular $P(a|+g)$. Para ello calculamos primero $P(a, +g)$ para todas las configuraciones de A. Como F y H son sumideros, podemos calcular esta probabilidad en la red resultante de podar F. En esta red, H sigue siendo un sumidero, y por eso también podemos podarlo, llegando así a la red que se muestra en la figura 2.11. □

Este resultado no es específico del método de inversión de arcos, sino que puede aplicarse a cualquier red —dada cierta evidencia y cierto conjunto de variables de interés— antes de aplicar cualquier algoritmo de inferencia. Volviendo al ejemplo anterior, el cálculo de $P(a, +g)$, y por tanto el de $P(a|+g)$, va a dar el mismo resultado en la red original y en la red podada, cualquiera que sea el algoritmo que apliquemos: fuerza bruta, eliminación de variables, agrupamiento, inversión de arcos, etc. En todos los casos el cálculo va a ser más eficiente en la red podada.

Hay que tener en cuenta, por último, que la eliminación de un sumidero puede hacer que otros nodos se conviertan en sumideros. Por ejemplo, si queremos calcular $P(a|\neg b)$ en la red de la figura 2.2, los sumideros son inicialmente F, G y H. Tras eliminarlos, C y D se convierten en sumideros y también pueden ser eliminados, de modo que el cálculo de $P(a|\neg b)$ puede realizarse sobre una red que sólo tenga los nodos A y B.

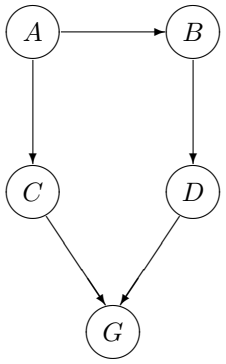


Figura 2.11: Red resultante de podar los nodos F y H en la red de la figura 2.2. La probabilidad $P(a|+g)$ obtenida a partir de esta nueva red es la misma que obtendríamos a partir de la red original.

Método de inversión de arcos

El método de inversión de arcos consiste en combinar los dos resultados anteriores: cuando queremos eliminar un nodo, primero lo convertimos en sumidero y luego lo podamos. Para demostrar que el método siempre es aplicable, necesitamos la siguiente proposición:

Proposición 2.47 En toda red bayesiana, para todo nodo X que tenga al menos un hijo existe un nodo Y, hijo de X, tal que no existe ningún otro camino dirigido desde X hasta Y (es decir, aparte del enlace $X \to Y$).

Demostración. Aplicamos el siguiente procedimiento, que nos va dar una lista ordenada de hijos de X. (Es posible que algunos de los hijos de X no figuren en esta lista.) Escogemos arbitrariamente un hijo de X, que va a ser el primer nodo de la lista, Y_1. Si existe otro hijo de X, que llamaremos Y_2, tal que hay un camino dirigido desde Y_1 hasta Y_2 entonces añadimos Y_2 a la lista, y así sucesivamente. Desde cada nodo de esta lista parte un camino dirigido que pasa por todos los nodos que aparecen antes que él en la lista. Eso implica que ningún nodo puede aparecer dos veces, pues entonces el grafo contendría un ciclo. Como el número de hijos de X es finito, llega un momento en que no se pueden añadir más nodos a la lista. El último de ellos cumple la condición requerida en el enunciado: es el nodo Y. □

Ejemplo 2.48 Vamos a aplicar el método de la demostración anterior a la red de la figura 2.12, con $X = A$. Escogemos arbitrariamente uno de los hijos de A, que va a ser $Y_1 = B$. Como existe otro camino dirigido desde A hasta B, que es $A \to C \to B$, tomamos $Y_2 = C$. También ahora existe otro camino desde A hasta C, que es $A \to D \to C$. Por tanto, $Y_3 = D$. Así tenemos una lista ordenada $\{B, C, D\}$ de hijos de A, tal que desde cada nodo parte un camino dirigido que pasa por todos los nodos que aparecen antes que él en la lista. El último nodo de la lista cumple la condición indicada en la proposición anterior.

Corolario 2.49 En toda red bayesiana, para todo nodo X que tenga al menos un hijo existe un nodo Y, hijo de X, tal que el enlace $X \to Y$ se puede invertir.

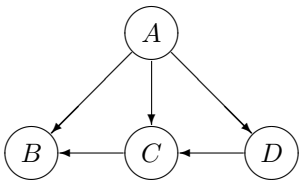


Figura 2.12: Red bayesiana en que el nodo A tiene tres hijos. Si queremos convertir A en sumidero, primero tenemos que invertir el enlace $A \to D$, luego $A \to C$ y finalmente $A \to B$.

Observe que la demostración de la proposición anterior nos proporciona un algoritmo para buscar entre los arcos que salen de X uno que pueda ser invertido. Así, en el ejemplo anterior, si queremos invertir uno de los enlaces que salen de A, podemos intentarlo con $A \to B$. Sin embargo, esto no es posible, porque hay otro camino dirigido desde A hasta B, que pasa por C. Por tanto, intentamos invertir el enlace $A \to C$, lo cual tampoco es posible, porque existe otro camino dirigido desde A hasta C, que pasa por D. Como no hay otro enlace dirigido desde A hasta D, podemos invertir este enlace, transformándolo en $D \to A$.

Cuando hemos invertido un enlace que parte de un nodo, si este nodo aún tiene más hijos podemos invertir otro enlace, y así sucesivamente, hasta que X no tenga más hijos, tal como afirma el siguiente corolario.

Corolario 2.50 Dada una red bayesiana, todo nodo con n hijos se puede transformar en un nodo sin hijos mediante n inversiones de arcos.

En el ejemplo anterior, el nodo A, que tiene tres hijos, puede convertirse en un nodo sin hijos mediante tres inversiones de arcos: $A \to D$, $A \to C$ y $A \to B$, que han de realizarse necesariamente en este orden.

Uniendo todos estos resultados, ya podemos explicar cómo se aplica este método para resolver el problema del diagnóstico probabilista, es decir, para calcular cualquier probabilidad $P(\mathbf{x}_I|\mathbf{e})$. Los pasos son los siguientes. Primero, escogemos un nodo S que no sea variable de interés ni pertenezca a la evidencia: $S \in \mathbf{X}_R = \mathbf{X}\backslash(\mathbf{X}_I \cup \mathbf{E})$. Si S tiene hijos, buscamos un enlace saliente de S que pueda ser invertido, y así recursivamente hasta que S sea un sumidero y pueda ser podado. Repetimos la operación para todos los nodos de $\mathbf{X}_R$ hasta llegar a una red que sólo contenga los nodos de $\mathbf{X}_I$ y de $\mathbf{E}$. La probabilidad $P(\mathbf{x}_I, \mathbf{e})$ es la probabilidad conjunta de esta red, que se puede calcular mediante la factorización de la probabilidad. Con esto, el problema está resuelto.

Ejemplo 2.51 Supongamos que queremos calcular $P(a|{+}g)$ para la red de la figura 2.2 mediante el método de inversión de arcos. Nos interesa tener una red en que sólo aparezcan las variables A y G, por lo que tenemos que eliminar B, C, D, F y H. Primero podamos los sumideros, que son F y H, con lo cual llegamos a la red de la figura 2.11. Como ya no quedan más sumideros, tenemos que empezar a invertir arcos hasta que algún nodo se convierta en sumidero. Dentro de $\mathbf{X}_R = \{B, C, D\}$, nos fijamos en el nodo C, que sólo tiene un hijo, G. Comprobamos que no existe ningún otro camino dirigido de C a G. Por tanto, invertimos el enlace $C \to G$, lo cual hace que A pase a ser padre de G y D a ser padre de C (véase la

figura 2.13.a). Las nuevas probabilidades de estos dos nodos se calculan así:

$$P(c, g|a, d) = P(c|a) \cdot P(g|c, d)$$
$$P(g|a, d) = \sum_c P(c, g|a, d)$$
$$P(c|g, a, d) = \frac{P(c, g|a, d)}{P(g|a, d)}$$

En realidad no haría falta calcular $P(c|g, a, d)$, porque en la nueva red el nodo C va a ser eliminado (fig. 2.13.b).

Ahora vamos a eliminar el nodo D, que tiene un solo hijo, G. Comprobamos que no existe ningún otro camino dirigido de D a G. Por tanto, invertimos el enlace $D \to G$, lo cual hace que A pase a ser padre de D y B a ser padre de G (fig. 2.13.c). Las nuevas probabilidades de estos dos nodos se calculan así:

$$P(d, g|a, b) = P(d|b) \cdot P(g|a, d)$$
$$P(g|a, b) = \sum_d P(d, g|a, b)$$
$$P(d|g, a, b) = \frac{P(d, g|a, b)}{P(g|a, b)}$$

Como en el caso anterior, no haría falta calcular $P(d|g, a, b)$ porque D va a ser eliminado inmediatamente (fig. 2.13.d).

Por último eliminamos B, para lo cual es necesario invertir el enlace $B \to G$ (fig. 2.13.e). Se comprueba fácilmente que no existe ningún otro camino dirigido de B a G. En este caso no hace falta añadir enlaces, porque ningún nodo tiene un padre que no sea padre del otro. Las nuevas probabilidades se calculan así:

$$P(b, g|a) = P(b|a) \cdot P(g|a, b)$$
$$P(g|a) = \sum_b P(b, g|a)$$
$$P(b|a, g) = \frac{P(b, g|a)}{P(g|a)}$$

Ahora B es un sumidero y puede ser eliminado, con lo cual llegamos a una red que sólo tiene la variable de interés, A, y la variable observada, G (fig. 2.13.f). Las probabilidades conjuntas $P(+a, +g)$ y $P(\neg a, +g)$ se obtienen de la factorización de la probabilidad de esta red:

$$P(a, +g) = P(a) \cdot P(+g|a)$$

y a partir de ellas calculamos la probabilidad buscada, $P(a|+g)$.

Importancia del orden de eliminación de nodos

Como en los métodos anteriores, el orden de eliminación es muy importante, pues aunque el resultado final va a ser el mismo, el coste computacional (en tiempo y en memoria requerida) puede ser muy diferente.

Por poner un ejemplo, consideremos de nuevo el grafo de la figura 1.4 (pág. 18) y supongamos que queremos calcular $P(h_m|h_1)$ para un valor de H_1 conocido —por ejemplo, $+h_1$—

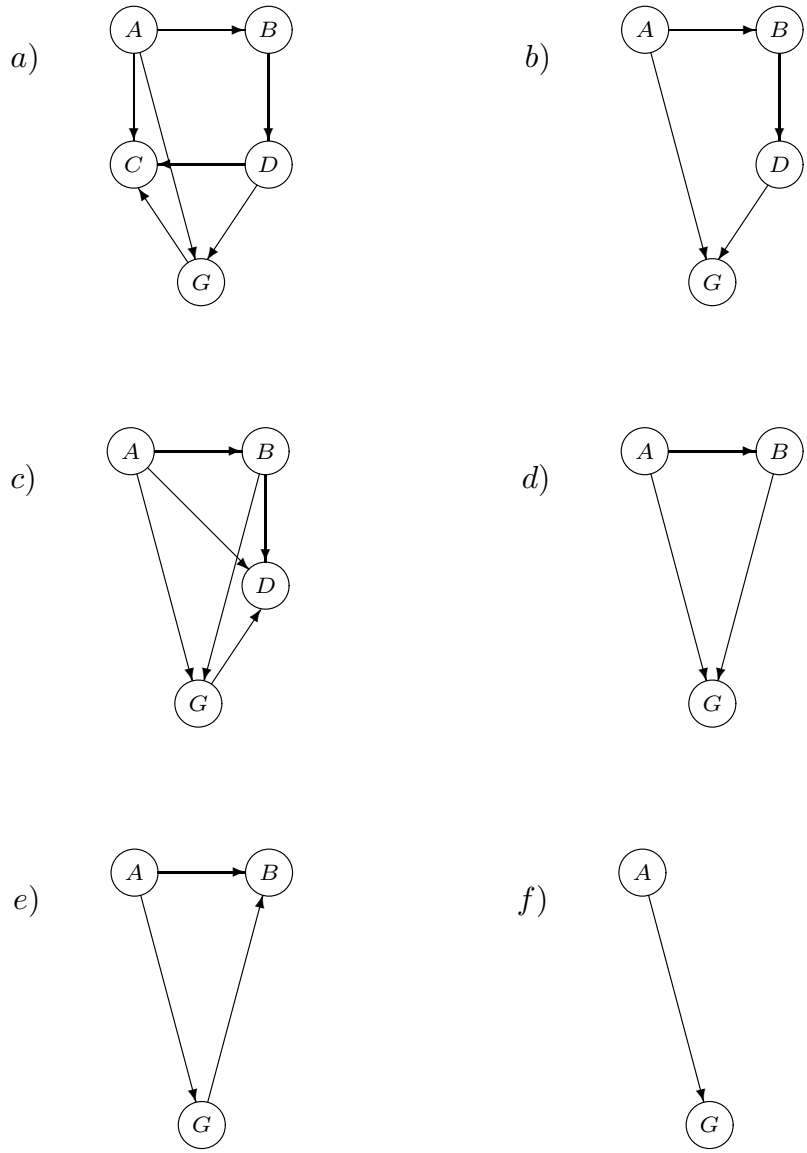


Figura 2.13: Inversión de arcos para el cálculo de $P(a|+g)$. Partiendo de la red de la figura 2.11, invertimos el enlace $C \rightarrow G$, lo cual obliga a añadir los enlaces $A \rightarrow G$ y $D \rightarrow C$. Luego eliminamos C, invertimos el enlace $D \rightarrow G$, eliminamos D, invertimos $B \rightarrow G$ y eliminamos B. Así llegamos a una red en que sólo aparecen A y G.

y para todos los valores de H_m. Si eliminamos primero los sumideros, nos queda una red que sólo contiene tres nodos, D, H_1 y H_m, y dos enlaces, $D \to H_1$ y $D \to H_m$. Para convertir D en sumidero invertimos primero uno de los enlaces, por ejemplo, $D \to H_1$, y luego el otro, $D \to H_m$, lo que obliga a añadir el enlace $H_1 \to H_m$. En la primera inversión sólo intervienen dos variables y en la segunda tres, y por tanto el coste computacional es pequeño.

En cambio, si queremos convertir D en sumidero antes de eliminar los nodos H_2 a H_{m-1}, el resultado es que habrá un enlace entre cada par de nodos (H_i, H_j). En consecuencia, habrá un nodo H_k que tenga de padres a todos los demás H's y su tabla de probabilidad condicional tendrá un tamaño del orden $\exp(m)$. Por otro lado, en cada inversión de arcos intervendrán cada vez más variables, hasta que en la última intervendrán todas ellas, $m+1$, lo cual implica un coste computacional enorme..

Queda claro que conviene eliminar primero los sumideros, porque el coste computacional es casi nulo, pero a partir de ahí ya no es evidente qué nodos conviene eliminar primero. Más aún, una vez elegido el nodo a eliminar, si éste tiene varios hijos puede haber varios enlaces como candidatos a ser invertidos, y no es fácil determinar cuál debe ser invertido primero. De nuevo, el problema de determinar el orden óptimo de inversiones y eliminaciones para el método de inversión de arcos es un problema NP-completo, por lo que se hace necesario utilizar reglas heurísticas.

2.3. Métodos aproximados

Existen básicamente dos tipos de métodos aproximados para el cálculo de la probabilidad en redes bayesianas: estocásticos y deterministas. Los *métodos estocásticos* consisten en realizar N simulaciones con una distribución de probabilidad extraída de la red bayesiana y, en función de los resultados obtenidos, estimar la probabilidad buscada, que en general es de la forma $P(\mathbf{x}_I|\mathbf{e})$. Los *métodos aproximados deterministas* generalmente consisten en modificar la red bayesiana (por ejemplo, eliminando los enlaces menos relevantes) para aplicar después un método exacto de propagación de evidencia.[12] En este capítulo sólo vamos a estudiar los métodos estocásticos, dando especial importancia a dos de ellos: el muestreo lógico y la ponderación por verosimilitud.

2.3.1. Fundamento de los métodos estocásticos

Leer la sección 9.1 y estudiar las secciones 9.2 y 9.3 del libro de Castillo et al. [7].[13]

2.3.2. Muestreo lógico

Estudiar la sección 9.4 del libro de Castillo et al. [7], titulada "El método de aceptación-rechazo".

[12] Hay un método aproximado determinista, el *muestreo sistemático*, que no sigue esta regla (realizar una computación exacta sobre una red aproximada a la original) sino que, a pesar de ser determinista, se parece más a los métodos de simulación estocástica. Lo mencionaremos en la sección 2.3.4; aparece descrito con detalle en la sección 9.9 del libro de Castillo et al. [7],

[13] Recordamos que este libro puede obtenerse de forma gratuita en Internet en `http://personales.unican.es/gutierjm/BookCGH.html`.

2.3.3. Ponderación por verosimilitud

Estudiar la sección 9.6 del libro de Castillo et al. [7], titulada "El método de la función de verosimilitud pesante".

2.3.4. Otros métodos

Leer las secciones 9.5, 9.7, 9.8 y 9.9 del libro de Castillo et al. [7]. Estas secciones no hace falta estudiarlas en profundidad.

2.3.5. Complejidad computacional de los métodos estocásticos

Leer la sección 9.11 del libro de Castillo et al. [7].

Bibliografía recomendada

La bibliografía sobre inferencia en redes bayesianas es abundantísima. Todos los libros mencionados en la bibliografía del capítulo anterior (página 36) tratan el tema; a nuestro juicio, el que lo explica con más extensión y detalle es el de Darwiche [15].

Yendo a las referencias históricas, los primeros algoritmos para redes bayesianas, que surgieron en la década de los 80, aparecen descritos en el famoso libro de Judeal Pearl [64]. El método de eliminación de variables es tan simple que resulta difícil determinar quién es su autor; una variante de este método, conocida como *bucket elimination* (*eliminación "por cubos" o "por calderos"*), propuesta por Rina Dechter [17], se ha hecho popular en los últimos años. El primer método de agrupamiento fue propuesto por Lauritzen y Spiegelhalter [50] en 1988, y posteriormente mejorado por Jensen et al. [41, 38] y por Shenoy [73].

La inversión de arcos fue propuesta por Olmsted [60] como método para la evaluación de diagramas de influencia, aunque el algoritmo fue completado por Shachter [72].

La bibliografía sobre métodos aproximados puede encontrarse en el libro de Castillo et al. [7]. Sin embargo, este libro no menciona el método estocástico que en la actualidad está dando los mejores resultados: el muestreo por importancia. Este método, bien conocido en estadística, fue aplicado por primera vez a las redes bayesianas por Shachter y Peot [71]. Algunas versiones más eficientes han sido propuestas por investigadores españoles [4, 34, 55] y por el grupo de la Universidad de Pittsburgh [8, 83].

Actividades

El alumno puede escoger algunas de éstas:

1. Resolver los ejercicios propuestos en este capítulo.
2. Asignar valores numéricos a las tablas de probabilidad de las redes bayesianas utilizadas en los ejemplos y realizar los cálculos numéricos. Por ejemplo, para la red de la figura 2.1 las tablas podrían ser las siguientes:

	$P(a)$
$+a$	0'02
$\neg a$	0'98

$P(b\|a)$	$+a$	$\neg a$
$+b$	0'7	0'1
$\neg b$	0'3	0'9

etc.

3. Introducir esas redes Elvira o en OpenMarkov (véanse las actividades del capítulo 1) y comprobar que los resultados de la inferencia coinciden con los obtenidos en la actividad anterior.

4. Resolver algunos de los ejercicios propuestos en los capítulos 8 y 9 de [7]. Comprobar que los resultados coinciden con los que ofrecen Elvira u OpenMarkov.

5. Analizar el código fuente del programa Elvira o de OpenMarkov[14] y tratar de entender cómo están implementados los algoritmos de inferencia exactos y aproximados.

6. Leer los artículos más recientes sobre muestreo por importancia: [8, 55, 83], en especial este último.

[14] Disponible en en `http://leo.ugr.es/~elvira/Bayelvira/bayelvira2.tar.gz` o en `www.openmarkov.org/developers.html`, respectivamente.

Capítulo 3

Construcción de redes bayesianas

Resumen

Hay dos formas de construir una red bayesiana, que podríamos denominar informalmente como "automática" y "manual". La primera consiste en tomar una base de datos y aplicarle alguno de los algoritmos que vamos a estudiar en la sección 3.3. Este proceso se conoce también como *aprendizaje de redes bayesianas*. El otro método consiste en construir primero, con la ayuda de un experto, un grafo causal, al que se le añaden posteriormente las probabilidades condicionales.

Contexto

Este capítulo se basa en los conceptos introducidos en el capítulo 1. Los conceptos expuestos en la sección 3.1, que explica la construcción de redes bayesianas causales con conocimiento experto, también serán útiles para la construcción de diagramas de influencia (sec. 4.4).

Por otro lado, la construcción de redes bayesianas a partir de bases de datos es una de las técnicas más utilizadas para la minería de datos y, por tanto, la sección 3.3 enlaza con otras asignaturas del Máster en Inteligencia Artificial Avanzada, tales como *Minería de datos*, *Minería de la web* y *Descubrimiento de información en textos*.

Objetivos

Los tres objetivos de este capítulo son:

1. que el alumno aprenda a construir redes bayesianas con la ayuda de expertos humanos;
2. que conozca los principales algoritmos de aprendizaje de redes bayesianas a partir de bases de datos;
3. que sepa construir redes bayesianas a partir de bases de datos utilizando alguna herramienta informática como Elvira u OpenMarkov.

Requisitos previos

Es necesario conocer bien los conceptos expuestos en el capítulo 1.

Contenido

3.1. Construcción de redes causales con conocimiento experto

3.1.1. Necesidad de la construcción manual (en algunos casos)

Como hemos mencionado en la introducción de este capítulo, es posible construir de forma automática una red bayesiana aplicando algún algoritmo de aprendizaje a una base de datos en que todas las variables que nos interesan estén representadas y que contenga un número de casos suficientemente grande. Se trata por tanto de un método rápido y barato, pues lleva poco tiempo y no requiere la colaboración de expertos. En cambio, la construcción "manual" de redes bayesianas requiere la colaboración de expertos que conozcan bien el problema que queremos modelar, y aun con la ayuda de éstos es una tarea compleja y que consume mucho tiempo: como vamos a ver a continuación, la construcción del grafo causal es difícil en muchos casos, y la obtención de las probabilidades numéricas es casi siempre aún más difícil.

Sin embargo, a pesar de que hay muchísimos algoritmos de aprendizaje disponibles en la actualidad y de que este método presenta muchas ventajas, la mayor parte de las redes bayesianas que se están utilizando en la actualidad se han construido de forma manual ¿A qué se debe esto? La razón principal es que para muchos problemas reales no existen bases de datos, o si existen, no son suficientemente grandes ni detalladas. Por ejemplo, en el caso de la medicina, prácticamente todas las bases de datos disponibles en la actualidad son incompletas, en el sentido de que solamente recogen algunas de las *variables observadas* (que son, a su vez, un subconjunto de las variables observables) y el *diagnóstico final* que ha realizado el médico, pero no suelen incluir todas las variables intermedias necesarias para que se cumpla la propiedad de separación condicional (cf. sec. 1.5.1).

Por otro, la existencia de "huecos" en las bases de datos (son los llamados "valores ausentes" o "*missing values*") dificultan la aplicación de los algoritmos de aprendizaje. La hipótesis de que los valores ausentes faltan al azar (es decir, que la probabilidad de que haya un "hueco" es independiente del valor que debería figurar en él) rara vez se cumple en la práctica: cuando un valor no ha sido registrado en la base de datos, no es por azar, sino por alguna razón. Eso hace que el aprendizaje de redes bayesianas a partir de bases de datos con muchos valores ausentes pueda dar resultados absurdos como consecuencia de haber aplicado una hipótesis totalmente errónea.

Otra razón es que los métodos de aprendizaje muchas veces producen grafos no causales. Por ejemplo, tal como dijimos en la sección 1.6.1, los tres grafos $A \to B \to C$, $A \leftarrow B \leftarrow C$, y $A \leftarrow B \to C$ son equivalentes en sentido probabilista y por tanto la información contenida en una base de datos no permite distinguir entre un grafo y otro. Ahora bien, desde el punto de vista causal los tres son completamente diferentes. Si el verdadero grafo causal para cierto problema fuera el primero y el método de aprendizaje devolviera el segundo o el tercero, al experto —especialmente si está acostumbrado a interpretar los grafos en sentido causal— le parecería que el modelo es erróneo, pues tanto en el segundo como en el tercero el enlace $B \to A$ parece querer indicar que B es causa de A. Ciertamente, habría que explicar al experto que el modelo obtenido no debe interpretarse de modo causal sino sólo en sentido probabilista, pero aun así no estaría satisfecho: los expertos humanos siempre prefieren trabajar con modelos causales. En cambio, las redes bayesianas construidas manualmente se basan en grafos causales, lo cual hace que sean más fácilmente aceptadas por los expertos que las van a utilizar.

Por todo ello en la práctica muchas veces es necesario construir las redes bayesianas de forma manual, que es el método que vamos a estudiar en esta sección.

Como hemos comentado, este método requiere la ayuda de un experto humano que conozca a fondo el problema que queremos modelar; con su ayuda, se construye primero un grafo causal (fase cualitativa), al que se le añaden posteriormente las probabilidades condicionales (fase cuantitativa). En nuestra exposición de este proceso, haremos referencia frecuentemente al caso de la medicina, por dos razones: primero, porque es el campo que mejor conocemos, dado que la mayor parte de las redes bayesianas que hemos construido han sido para diferentes problemas médicos (véanse las páginas `www.cisiad.uned.es`); el segundo, porque en medicina se encuentran prácticamente todas las dificultades que pueden surgir en otros campos, de modo que al tratar este campo el lector estará preparado para abordar cualquier otro, que en general va a ser más sencillo.

3.1.2. Fase cualitativa: estructura de la red

El primer paso de la construcción de una red consiste en seleccionar las variables relevantes para el problema que queremos resolver, tanto las **anomalías** que queremos diagnosticar como los **datos observables**. Por ejemplo, en el caso de la cardiología las anomalías pueden ser: hipertensión en la aurícula derecha, dilatación del ventrículo derecho, hipertrofia del ventrículo izquierdo, mixoma en la aurícula izquierda, calcificación de la válvula pulmonar, disminución de la fracción de eyección, fibrosamiento pericárdico, etc. En el caso del diagnóstico de una planta industrial, en cambio, deberemos incluir una variable por cada una de los posibles fallos que pueden presentarse: fallo de un generador, fallo de una válvula, etc.

En el caso de la medicina, los datos observables incluyen, en primer lugar, los datos personales del enfermo. Algunos de los datos administrativos, tales como el nombre, dirección, número de historia clínica, etc., no influyen en el diagnóstico, y por tanto no se representan en la red, mientras que otros, como el sexo, la edad o el país de origen, han de representarse explícitamente por la importancia que tienen para el diagnóstico. También hay que incluir en la red los antecedentes (enfermedades que ha sufrido, medicación que ha recibido, intervenciones que se le han practicado, factores hereditarios, factores de riesgo) y los síntomas y signos relevantes para el tipo de enfermedades que queremos diagnosticar; volviendo al ejemplo de la cardiología, la disnea, la ortopnea, los labios cianóticos, el dolor precordial, los edemas... constituyen una pequeña muestra de los numerosos síntomas y signos que intervienen en el diagnóstico cardiológico. A la hora de determinar cuáles son los síntomas y signos relevantes, puede ser útil pensar en cada una de las partes del organismo (cabeza, tronco y extremidades) o en los cuatro apartados clásicos en que se divide la exploración física: inspección, auscultación, palpación y percusión. Por último, hay que tener en cuenta las pruebas complementarias, también llamadas pruebas de laboratorio: analítica, radiografía, electrocardiografía, técnicas de ultrasonidos, resonancia magnética, gammagrafía, etc., etc. En algunas de estas pruebas, como la ECG, hay que decidir si la red bayesiana va a considerar datos primarios (tales como la amplitud QRS, la duración del intervalo PR, la elevación del segmento ST...) o datos interpretados (hipertrofia ventricular izquierda, signos de isquemia, bloqueo auriculoventricular, etc.). También es necesario decidir en algunos casos si se va a trabajar con datos cuantitativos (fracción de eyección = 60%) o con datos cualitativos (fracción de eyección levemente disminuida).

En el caso del diagnóstico de un proceso industrial, los datos observables corresponden generalmente a los sensores disponibles: temperatura, presión, caudal, etc..

Una cuestión importante es determinar cuántos valores va a tomar cada variable; es uno de los aspectos de lo que podríamos denominar **el problema de la granularidad**. En principio, cuanto más detallado sea el modelo, mayor precisión podrá tener en el diagnóstico. Sin embargo, construir modelos más detallados significa, por un lado, que hay que obtener más probabilidades numéricas (con lo cual se pierde fiabilidad en los datos obtenidos) y, por otro, que la computación va a llevar más tiempo y la interpretación de resultados va a ser más difícil.

Por ejemplo, podemos definir la hipertensión del ventrículo izquierdo como una variable binaria que toma dos valores: presente o ausente. Sin embargo, en general conviene precisar más, dando cuatro o cinco valores: hipertensión ausente, leve, moderada, severa, muy severa. Si quisiéramos construir un modelo más detallado, podríamos distinguir entre presión telesistólica y presión telediastólica; este segundo modelo podría "razonar" con mucha más precisión que el anterior, pero es también mucho más difícil de construir y de evaluar.

Por poner otro ejemplo, al determinar cuántos valores toma la variable Dolor, en principio convendría considerar, además de su intensidad, duración y localización, todos los matices como irradiación, recurrencia, factores agravantes, factores atenuantes, etc. Sin embargo, incluir todos estos aspectos nos llevaría a un modelo sumamente complicado, para el que no podríamos obtener todas las probabilidades numéricas, por lo que en la práctica conviene seleccionar solamente aquellas características más relevantes para las enfermedades que queremos diagnosticar.

En consecuencia, hay que buscar un punto de equilibrio: un modelo que no incluya el grado de detalle necesario no va a ser fiable, pero un modelo en que intentemos incluir demasiados detalles tampoco va a ser fiable por la falta de precisión en los parámetros probabilistas.

Después, hay que trazar los **enlaces causales** entre las variables. En la mayor parte de los casos es bastante sencillo determinar cuáles deben ser los enlaces, aunque en otros la situación es complicada. Por ejemplo, según cierto libro de cardiología, los factores que influyen en hipertensión arterial son la edad, el sexo, el tabaquismo, la raza, la herencia, la obesidad, el estrés y la ingestión de sodio. El mismo libro, al hablar de la isquemia, señala entre los factores de riesgo la edad, el sexo, el tabaquismo, la hipertensión arterial, la hipercolesterolemia y la diabetes. Con esta información podríamos construir la red causal que aparece en la figura 3.1.

Sin embargo, sabemos que el tabaquismo depende del sexo, de la edad, de la raza y del estrés; la obesidad depende de la herencia y del tipo de alimentación; los factores de riesgo hereditarios dependen de la raza. Teniendo en cuenta estas consideraciones, podemos construir el grafo de la figura 3.2, en el que hemos añadido algunos enlaces para representar tales influencias (relaciones causales) y hemos suprimido otros. Tenga en cuenta que en una red bayesiana no sólo son importantes los enlaces que hemos trazado, sino también los que hemos omitido. Podemos decir que **cada enlace ausente representa una hipótesis de independencia causal**. Por ejemplo, el grafo de la figura 3.2 afirma que el ser varón no aumenta por sí mismo la probabilidad de infarto, sino que el hecho de que los varones tengan más infartos que las mujeres se debe a que éstos fuman más y tienen más estrés. Ésta es una hipótesis que habría que comprobar mediante un estudio epidemiológico. También afirma que el tabaquismo no aumenta el riesgo de isquemia directa sino indirectamente, mediante el aumento de la hipertensión arterial (otra afirmación que habría que comprobar), y que la ingesta de sodio y la hipercolesterolemia son independientes de la edad, el sexo y la raza (unas afirmaciones más que dudosas).

Por eso conviene recalcar una vez más que en una red bayesiana tanto la presencia como

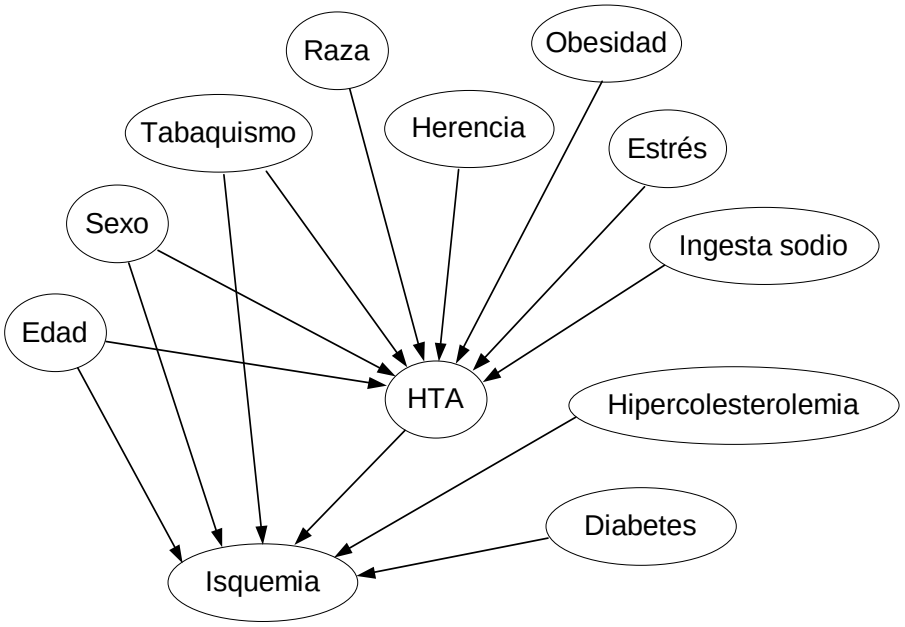


Figura 3.1: Causas de la isquemia y de la hipertensión arterial (HTA), según cierto libro de cardiología.

la ausencia de un enlace representa una afirmación sobre la relación entre dos variables. En consecuencia, sería deseable realizar estudios estadísticos exhaustivos no sólo para obtener las probabilidades numéricas que definen la red, sino también para comprobar si las afirmaciones de dependencia e independencia condicional contenidas en la red se corresponden con la realidad.

3.1.3. Fase cuantitativa: obtención de las probabilidades condicionales

Una vez que hemos determinado cuáles son las variables, los valores que toma cada una de ellas y los enlaces causales, queda la labor más difícil, que es construir las tablas de probabilidad. En el caso de la medicina, las fuentes de información numérica son las cuatro que vamos a analizar a continuación.

Estudios epidemiológicos

Lo ideal sería que todas las probabilidades condicionales que forman parte de la red se obtuvieran de estudios epidemiológicos. Tiene la ventaja de que es el método más directo y más sencillo. Sin embargo, en la práctica resulta muy costoso, en tiempo y en dinero, realizar un estudio diseñado específicamente para obtener todas las probabilidades necesarias. Por eso en la construcción de redes bayesianas casi nunca se utiliza este método.

La literatura médica

Buscar las probabilidades condicionales en libros y revistas tiene la ventaja de que es mucho más barato que realizar estudios epidemiológicos y mucho más fiable que obtenerlas

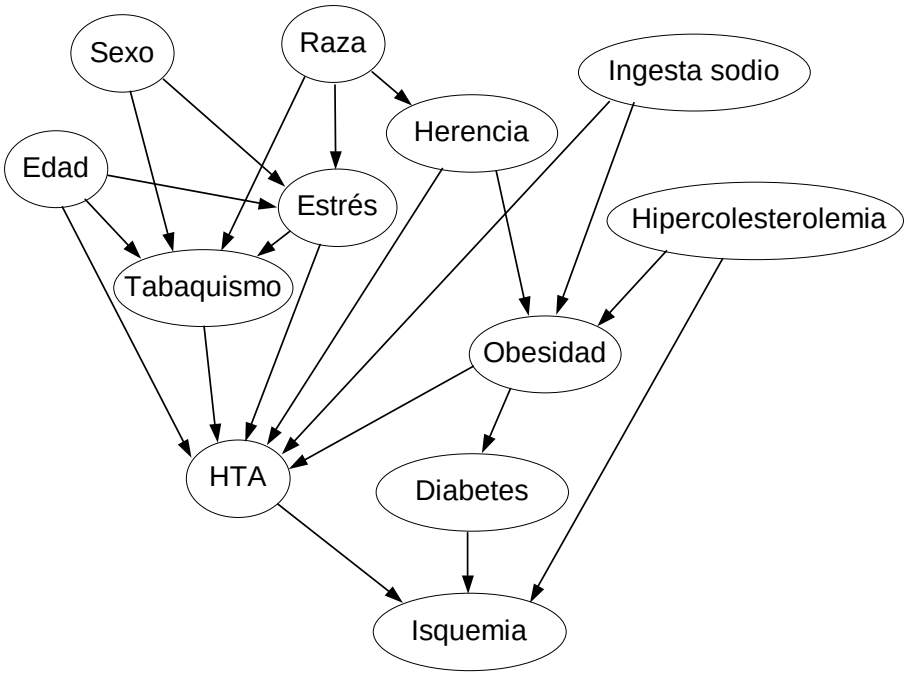


Figura 3.2: Reinterpretación de las causas de isquemia y de HTA.

mediante estimaciones subjetivas. El aspecto negativo es que resulta muy difícil encontrar en las publicaciones científicas las probabilidades condicionales necesarias para construir una red bayesiana, examinemos como ejemplo el siguiente fragmento, extraído de un libro especializado:

> El tumor primario *más común* en el corazón *adulto* es el mixoma y el *75%* de ellos se localiza en la aurícula izquierda, *habitualmente* en mujeres. [Cursiva añadida]

En esta breve cita aparecen dos términos difusos, *adulto* y *habitualmente*. Esto nos plantea varios interrogantes: ¿Desde qué edad se considera a una persona como *adulta*? ¿Distingue entre adultos y ancianos o los engloba a todos en el mismo grupo? ¿Qué frecuencia debemos entender por "*habitualmente*"? Hay estudios psicológicos que podrían ofrecer una cierta ayuda a la hora de convertir las expresiones cualitativas en probabilidades numéricas, pero las variaciones en las asignaciones son tan amplias que en la práctica resultan de poca ayuda; véase, por ejemplo, [20, sec. 3.2]. Por otro lado, además de los términos difusos, aparece una proposición comparativa, "*más común*", que no indica cuáles son las prevalencias de los tumores, ni siquiera la proporción relativa entre ellas.

La existencia de un número concreto, "*el 75%*", tampoco es de gran ayuda. ¿Se trata de un número exacto medido experimentalmente o es una estimación subjetiva? Aun suponiendo que sea un resultado experimental, tampoco podemos introducir este dato directamente en nuestra red bayesiana, porque no expresa una probabilidad condicional entre causas y efectos. Es más, aunque se tratara de la probabilidad o frecuencia de una causa dado el efecto, como suele aparecer en la bibliografía médica ("el 70% de las estenosis mitrales son de origen reumático"), el problema sería el mismo: necesitamos conocer la probabilidad del efecto

dadas sus causas —o en el caso del modelo OR— la probabilidad de que la causa produzca el efecto: "La fiebre reumática produce estenosis mitral en el 18% de los casos"; éste es un dato que aparece alguna vez en la literatura, pero mucho menos frecuentemente de lo que desearíamos.

Sin embargo, el "75%" mencionado no es una información que debamos despreciar. Con una red bayesiana construida mediante estimaciones subjetivas es posible calcular la probabilidad de que el tumor sea un mixoma y la probabilidad de que se localice en la aurícula izquierda; el resultado debe coincidir con el que habíamos encontrado en el libro. Si no es así, tendremos que revisar los parámetros de la red (las probabilidades condicionales) que han influido en ese desajuste, hasta llegar a una concordancia entre el modelo y los datos disponibles.

Por último, la disparidad entre distintos investigadores puede ser sorprendente. Así, hay quien afirma que el prolapso mitral se da en el 20% de las mujeres jóvenes, mientras que otros sitúan esta proporción en el 1 o el 2%. La diferencia se debe a la técnica (ecocardiografía de modo M frente a eco bidimensional) y al criterio empleados. A la hora de construir una red bayesiana es preciso determinar claramente en cada caso cuál es el criterio que se va a seguir, para evitar que se produzcan ambigüedades y confusiones.

Bases de datos

Extraer las probabilidades a partir de una base de datos tiene la ventaja de ser un método rápido y barato. El problema es que a veces las bases de datos tienen muchos "huecos" (en inglés, "*missing values*"), como ya hemos comentado. Es habitual en los estudios suponer que los valores faltan al azar ("*random missing values*"), lo cual no se corresponde con la realidad: cuando un dato falta es por alguna razón, como ya hemos comentado. Otro de los problemas es que las bases de datos suelen estar sesgadas, pues por ejemplo la "prevalencia" de una enfermedad en un hospital no coincide en absoluto con la prevalencia en la población general [24].

Por cierto, cuando dibujamos primero el grafo de la red y luego obtenemos todas las probabilidades de una base de datos (así se hizo, por ejemplo, en el sistema Hepar II, para diagnóstico de enfermedades hepáticas [61]), la construcción de la red no es completamente "manual" ni "automática", sino un caso híbrido, pues el grafo se construye con la ayuda de un experto mientras que las probabilidades pueden extraerse de la base de datos mediante un algoritmo.

Estimaciones subjetivas

En muchas ocasiones no es posible encontrar en la literatura científica la información que se necesita y tampoco es posible (por limitaciones de tiempo y de recursos materiales) llevar a cabo la investigación correspondiente. Por ello frecuentemente es necesario recurrir a la estimación subjetiva de expertos humanos. En este caso, es importante que quienes evalúan las probabilidades sean especialistas muy veteranos, que hayan examinado un número significativo de pacientes, para que las estimaciones sean fiables.

En este sentido, es importante también tener muy en cuenta los estudios psicológicos que indican que la forma de plantear las preguntas puede llevar a resultados muy distintos; véanse, por ejemplo, los estudios ya clásicos recogidos en la obra de Kahneman et al. [42] y otros más recientes que se mencionan en [67], que son casi todos ellos importantes en relación con este tema, en particular el de Fishhoff [26], titulado "*Debiasing*", por los consejos que da sobre

cómo evitar los sesgos que suelen producirse en la estimación subjetiva de la probabilidad. También es muy interesante el libro de Morgan y Henrion [56]. (Un resumen de estos trabajos se encuentra en [20, sec. 3.1]).

Naturalmente, a la hora de construir una red bayesiana es posible contar con probabilidades condicionales procedentes de las cuatro fuentes mencionadas: estudios epidemiológicos, literatura médica, bases de datos y estimaciones subjetivas. Sin embargo, a la hora de unir todas estas probabilidades en un único modelo hay que ser muy cuidadoso, porque en algunas ocasiones la combinación directa de datos numéricos procedentes de orígenes diversos puede dar lugar a inconsistencias del modelo, que se traducen en diagnósticos y actuaciones terapéuticas incorrectas [24].

También debemos señalar que, a pesar de la distinción que hemos hecho entre las dos fases de la construcción manual de una red bayesiana, en la práctica suele tratarse de un proceso iterativo, de modo que muchas veces mientras se están obteniendo las probabilidades condicionales se hacen algunos retoques en la estructura de la red; véase, por ejemplo, el artículo [45], que describe la construcción de Prostanet, una red bayesiana para el diagnóstico de cáncer de próstata y otras enfermedades urológicas.

A pesar de estas contribuciones, la obtención del conocimiento necesario para construir un sistema experto sigue siendo hoy en día más un arte que una técnica, de modo que los resultados obtenidos dependen en gran medida de las dotes personales de quien lo realiza. En inteligencia artificial existen numerosos libros y artículos sobre el tema, e incluso una revista, el *Journal of Knowledge Acquisition*, aunque la mayor parte de los trabajos suponen explícita o implícitamente que la representación del conocimiento se va a realizar mediante reglas, por lo que hay muy poca bibliografía específica sobre la obtención de probabilidades condicionales para sistemas expertos bayesianos.

3.1.4. Resumen

Como síntesis de lo expuesto en esta sección, enumeramos a continuación los pasos necesarios para construir una red bayesiana "a mano":

1. Seleccionar las **variables** que intervienen. A cada variable le corresponde un nodo en la red.

2. Trazar los **enlaces** *en sentido causal*; es decir, considerar para cada enlace cuál es la *causa* y cuál el *efecto*. Por ejemplo, ¿es la enfermedad o anomalía la causa de que aparezca el síntoma/signo, o viceversa? Hay que recordar que, por definición, una red bayesiana puede tener bucles pero no puede tener ciclos.

3. Una vez construido el grafo, está determinada cuál es la forma de la **factorización de la probabilidad**, es decir, la forma de las tablas de probabilidad.[1]

4. Determinar qué **valores** (cuántos y cuáles) toma cada variable. Si la variable es continua, habrá que discretizarla, porque los algoritmos que hemos estudiado en este curso sólo admiten variables discretas.

[1] Recordamos que por cada nodo Y cuyos padres son $\{X_1, \ldots, X_n\}$ ha de haber una (y sólo una) tabla de probabilidad $P(y|x_1, \ldots, x_n)$; para un nodo sin padres, la probabilidad condicional es simplemente la probabilidad a priori. A pesar de que esto es algo muy elemental, a veces hay estudiantes que lo olvidan y construyen una tabla de probabilidad por cada padre: $P(y|x_1)$, $P(y|x_2)$, etc., lo cual es un error grave.

5. Ver en cuáles de las familias que forman la red es posible aplicar alguno de los **modelos canónicos** de los que vamos a definir en la próxima sección. Los que más se usan son el modelo OR y el modelo MAX. Esto simplifica notablemente la asignación de probabilidades, pero hay que tener en cuenta que este modelo simplificado sólo es aplicable cuando se cumplen las condiciones señaladas en dicha sección.

6. Asignar **probabilidades numéricas**, como hemos indicado en la sección 3.1.3). Si la familia interactúa mediante un modelo IIC, hay que asignar las probabilidades de cada enlace, más la probabilidad residual c_L. En otro caso, no queda más remedio que indicar explícitamente la tabla de probabilidad condicional para cada familia, lo cual puede ser bastante complejo cuando el número de padres es elevado; por eso, a la hora de construir el grafo de la red, conviene utilizar algún "truco" (por ejemplo, introducir variables intermedias adicionales) con el fin de que cada nodo no tenga más de dos o tres padres (a partir de tres o cuatro padres el problema se complica notablemente si no podemos aplicar ningún modelo canónico).

Este proceso de construcción de redes se simplifica muchísimo con la ayuda de algunas de las herramientas informáticas existentes en la actualidad, tales como OpenMarkov, que permiten dibujar grafos en pantalla, añadir y eliminar nodos y enlaces fácilmente, y asignar valores a las variables; además generan automáticamente las tablas de probabilidad condicional que el usuario debe rellenar; y, lo que es aún más importante, permiten introducir hallazgos y comprobar cómo se modifica la probabilidad de las variables de interés, lo cual es muy útil para corregir las probabilidades condicionales asignadas si los resultados del diagnóstico no coinciden con lo que era de esperar [46].

Por eso recomendamos encarecidamente a los alumnos que practiquen con OpenMarkov (o con otra herramienta similar), introduciendo los ejemplos del texto y otros que al alumno se le ocurran, pues eso le ayudará a comprender mucho mejor todo lo que estamos estudiando en esta asignatura.

3.2. Modelos canónicos

En una red bayesiana, para un nodo Y cuyos padres son $\mathbf{X} = \{X_1, \ldots, X_n\}$, la tabla de probabilidad condicional (TPC) consta de un número de parámetros que crece exponencialmente con n. En concreto, si la variable Y puede tomar m valores, la distribución de probabilidad $P(y|\mathbf{x})$ tiene m parámetros, de los cuales $m-1$ son independientes (porque su suma ha de ser la unidad), y si cada padre puede tomar m valores, habrá m^n configuraciones de $\mathbf{X}$, de modo que el número total de parámetros es m^{n+1}, de los cuales $(m-1) \times m^n$ son independendientes. Por ejemplo, para un nodo binario con 5 padres binarios necesitamos 32 parámetros independientes. Si tenemos una familia con 4 padres y cada nodo toma 3 valores, necesitamos 162 parámetros.

En la práctica, construir una TPC con más de 3 o 4 padres es muy difícil, casi imposible, no sólo por el elevado número de parámetros que se necesita, sino también porque cada uno de ellos es muy difícil de estimar. Por ejemplo, para estimar la probabilidad $P(y|\mathbf{x})$ un experto debe recordar los casos en que ha visto la configuración $\mathbf{x}$ y decir aproximadamente en cuántos de ellos la variable Y tomaba el valor y. El problema es que si la configuración $\mathbf{x}$ es muy poco frecuente, la estimación no va a ser fiable. Por ejemplo, la estimación de P(fiebre|paludismo, neumonía, bronquitis) puede ser muy difícil porque es casi seguro que ningún médico ha visto nunca un paciente que padezca simultáneamente las tres enfermedades. Igualmente, en una

base de datos tampoco habrá ningún paciente con esa configuración de enfermedades, ni es previsible que encontremos un número suficiente de pacientes en esa situación al hacer un estudio epidemiológico, ni es esperable encontrar ese dato en la literatura médica.

En estas situaciones pueden ser útiles los modelos canónicos, denominados así porque son como los bloques elementales a partir de los cuales se pueden construir modelos probabilistas más sofisticados. En este tema vamos a estudiar los dos tipos de modelos canónicos más importantes: los modelos deterministas y los basados en la hipótesis de *independencia de influencia causal* (IIC). Los primeros no necesitan ningún parámetro. Los segundos, que se basan en los primeros, requieren un número de parámetros proporcional al número de padres, en vez de ser exponencial, y además los parámetros suelen ser más fáciles de estimar que en el caso de una TPC general.

3.2.1. Modelos deterministas

El tipo de modelo canónico más sencillo es aquél en que el valor de Y es una función de los valores de los padres: $y = f(x_1, \ldots, x_n)$. En este caso la tabla de probabilidad condicional (TPC) es

$$P(y|\mathbf{x}) = \begin{cases} 1 & \text{si } y = f(\mathbf{x}) \\ 0 & \text{en otro caso .} \end{cases} \tag{3.1}$$

Por tanto, la principal ventaja de estos modelos es que no necesitan ningún parámetro numérico.

Tipo de función	Tipo de variables	Nombre	Definición
lógica	booleana	NOT	$y \Longleftrightarrow \neg x$
		OR	$y \Longleftrightarrow x_1 \vee \ldots \vee x_n$
		AND	$y \Longleftrightarrow x_1 \wedge \ldots \wedge x_n$
		XOR	$y \Longleftrightarrow \text{pos}(\mathbf{x}) = 1$
		exactamente-r	$y \Longleftrightarrow \text{pos}(\mathbf{x}) = r$
		umbral	$y \Longleftrightarrow \text{pos}(\mathbf{x}) \geq r$
algebraica	ordinal	MINUS	$y = -x$
		INV	$y = x_{\max} - x$
		MAX	$y = \max(x_1, \ldots, x_n)$
		MIN	$y = \min(x_1, \ldots, x_n)$
		ADD	$y = x_1 + \ldots + x_n$
		promedio	$y = \frac{1}{n}(x_1 + \ldots + x_n)$
		promedio discreto	$y = \left\lceil \frac{1}{n}(x_1 + \ldots + x_n) \right\rceil$
		combinación lineal	$y = a_0 + a_1 x_1 + \ldots + a_n x_n$

Tabla 3.1: Algunas de las funciones más comunes utilizadas para construir modelos canónicos. En el caso de variables booleanas, pos($\mathbf{x}$) indica el número de variables dentro de la configuración $\mathbf{x}$ que toman el valor "verdadero", "presente" o "positivo".

La tabla 3.1 muestra algunas de las funciones que se han propuesto para construir modelos canónicos. Las funciones lógicas se aplican a variables booleanas, es decir, del tipo "verdadero/falso", "presente/ausente" o "positivo/negativo". Observe que todas las funciones de la tabla 3.1 son conmutativas (excepto la combinación lineal cuando las a_i son diferentes), lo que implica que el orden de los padres en el correspondeinte modelo canónico es irrelevante.

La tabla 3.2 muestra las TPC's correspondientes a algunas de las funciones definidas en la tabla 3.1, de acuerdo con la ecuación (3.1).

Función	TPC		
NOT	$P(+y\|x)$	$+x$	$\neg x$
		0	1
OR	$P(+y\|x_1, x_2)$	$+x_1$	$\neg x_1$
	$+x_2$	1	1
	$\neg x_2$	1	0
AND	$P(+y\|x_1, x_2)$	$+x_1$	$\neg x_1$
	$+x_2$	1	0
	$\neg x_2$	0	0
XOR	$P(+y\|x_1, x_2)$	$+x_1$	$\neg x_1$
	$+x_2$	0	1
	$\neg x_2$	1	0

Tabla 3.2: Tablas de probabilidad condicional (TPC) para algunos de los modelos deterministas inducidos por las funciones lógicas de la tabla 3.1.

A veces se utiliza el término "puerta" (en inglés, *gate*) para referirse a estos modelos canónicos, por analogía con las puertas utilizadas en electrónica. Por ejemplo, en un circuito, una puerta OR es un componente que tiene dos entradas, las cuales pueden tomar los valores 0 (potencial bajo) y 1 (potencial alto); basta que cualquiera de ellas tome el valor 1 para que la salida tome también el valor 1; es decir, si la primera entrada vale 1 o la segunda vale 1, entonces la salida vale 1, y por eso recibe el nombre de puerta OR ("or" en inglés significa "o" en castellano). Análogamente, en el modelo canónico que acabamos de mostrar, basta que X_1 o X_2 (o ambos) estén presentes para que Y esté presente. Por eso al modelo canónico OR determinista a veces se le llama "puerta OR", por analogía con la electrónica, considerando a X_1 y X_2 como entradas y a Y como salida. Por la misma razón se habla a veces de "puerta AND" y "puerta XOR" para referirse a los correspondientes modelos dados en la tabla 3.1.

Estos modelos se denominan "deterministas" porque, conociendo el valor de X_1 y el de X_2, se conoce con certeza el de Y; esto contrasta con los modelos probabilistas que vamos a ver a continuación, en los que los valores de X_1 y X_2 no determinan el valor de Y, sino sólo su probabilidad.

3.2.2. Modelos IIC

Los modelos deterministas no son muy comunes en la práctica, al menos en los dominios en que se aplican las redes bayesianas, que se introducen precisamente para tratar la incertidumbre. En esta sección vamos a presentar otro tipo de modelos: los que se basan en la hipótesis de *independencia de la interacción causal* (IIC). Estudiaremos dos subclases: los modelos "con ruido" y los residuales; los segundos son un caso particular de los primeros.[2]

Modelos IIC "con ruido"

Los modelos IIC se construyen a partir de los modelos deterministas introduciendo n variables auxiliares $\{Z_1, \ldots, Z_n\}$, como se indica en la figura 3.3, de modo que Y es una función determinista de las Z_i's y el valor de cada Z_i depende de forma probabilista de X_i, según la TPC $P(z_i|x_i)$. La TPC $P(y|\mathbf{x})$ se obtiene eliminando por suma las Z_i's:

$$P(y|\mathbf{x}) = \sum_{\mathbf{z}} P(y|\mathbf{z}) \cdot P(\mathbf{z}|\mathbf{x}) \ . \tag{3.2}$$

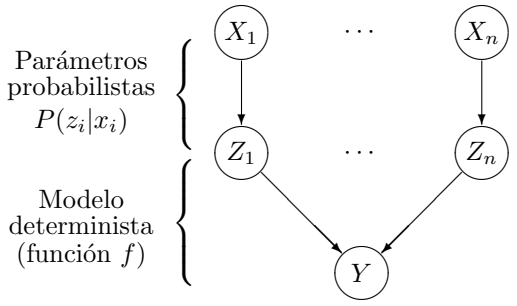


Figura 3.3: Estructura interna de los modelos IIC. Las variables Z_i se introducen para explicar la construcción del modelo, pero no forman parte del modelo, en el cual sólo intervienen Y y las X_i's.

La hipótesis de independencia de influencia causal (IIC) implica que los mecanismos causales (y los inhibidores) por los cuales las X_i's influencian el valor de Y son diferentes y no interactúan. En el grafo de la figura 3.3, esto se manifiesta en la ausencia de enlaces de la forma $X_i \to Z_j$ o $Z_i \to Z_j$ con $i \neq j$. De ahí se deduce que

$$P(\mathbf{z}|\mathbf{x}) = \prod_i P(z_i|x_i) \ . \tag{3.3}$$

Esta factorización, junto con las ecuaciones (3.1) y (3.2), lleva a

$$P(y|\mathbf{x}) = \sum_{\mathbf{z}|f(\mathbf{z})=y} \prod_i P(z_i|x_i) \ . \tag{3.4}$$

Ésta es la ecuación general de los modelos IIC.

[2]Puede que en una primera lectura la definición de los modelos IIC resulte demasiado abstracta. Los conceptos expuestos aquí se entenderán mejor cuando estudiemos dos modelos concretos: OR y MAX.

Nótese que cada parámetro $P(z_i|x_i)$ de un modelo IIC está asociado a un enlace particular, $X_i \rightarrow Y$, mientras que cada parámetro $P(y|\mathbf{x})$ de una TPC general corresponde a una cierta configuración $\mathbf{x}$ en que intervienen todos los padres de Y, y por tanto, no puede ser asociado a ningún enlace particular. Esta propiedad, que se deduce de la hipótesis de IIC, conlleva dos ventajas desde el punto de vista de la obtención de las probabilidades. La primera es una reducción significativa del número de parámetros necesarios, desde $O(\exp(n))$ en el caso de una TPC general hasta $O(n)$ en un modelo IIC. Por ejemplo, para un nodo binario con 10 padres binarios, la TPC consta de $2^{11} = 2,048$ parámetros numéricos. Al añadir un solo padre el número se duplica, convirtiéndose en $2^{12} = 4,096$ parámetros. En cambio, un modelo IIC sólo necesita 10 y 11 parámetros, respectivamente. La segunda ventaja es que los parámetros de un modelo IIC tienen una interpretación intuitiva sencilla, lo cual facilita la estimación subjetiva por parte de los expertos humanos. Es decir, que los modelos IIC no sólo necesitan muchos menos parámetros que las TPC generales, sino que los parámetros requeridos son mucho más intuitivos y fáciles de obtener.

Modelos IIC residuales

En muchos casos no es posible ni deseable incluir en un modelo todas las variables que ejercen influencia sobre Y. En este caso, si se cumplen ciertas hipótesis [23], podemos construir un modelo en que sólo algunas de las causas de Y aparecen explícitamente, mientras que la influencia de las demás pueden ser representada mediante un *parámetro residual* $P(z_L)$. Este parámetro puede interpretarse como la probabilidad a priori de una variable auxiliar (imaginaria) que representa las causas no explícitas, tal como se muestra en la figura 3.4.

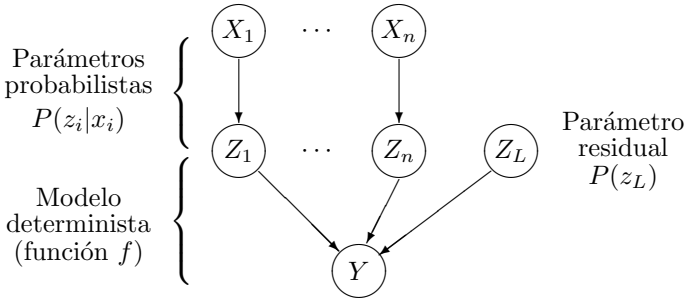


Figura 3.4: Estructura interna de los modelos IIC residuales. La variable auxiliar Z_L representa las causas que no están explícitas en el modelo.

Vamos a estudiar a continuación los dos modelos IIC más utilizados en la práctica: OR y MAX.

3.2.3. Modelos OR/MAX

En la sección 3.2.1 hemos estudidado los modelos OR y MAX deterministas. En ésta vamos a explicar cómo se pueden construir modelos IIC basados en ellos.

Modelo OR "con ruido"

La interpretación causal de este modelo es que cada X_i representa una causa de Y y cada Z_i indica si X_i ha producido Y o no. El término "con ruido" significa, en este caso, que es posible que algunas causas no produzcan el efecto cuando están presentes. En ese caso, $\neg z_i$ significa que X_i no ha producido Y, bien porque X_i estaba ausente, bien porque cierto inhibidor I_i ha impedido que X_i produjera Y. Si denotamos mediante q_i la probabilidad de que el inhibidor I_i esté activo, entonces la probabilidad de que X_i, estando presente, produzca Y es

$$c_i = P(+z_i| + x_i) = 1 - q_i \ . \tag{3.5}$$

En la práctica, $c_i > 0$, porque si c_i fuera cero, entonces X_i no podría ser una causa de Y y no lo incluiríamos entre sus padres.

Naturalmente, cuando X_i está ausente no puede producir Y, y por eso

$$P(+z_i|\neg x_i) = 0 \ . \tag{3.6}$$

$P(z_i\|x_i)$	$+x_i$	$\neg x_i$
$+z_i$	c_i	0
$\neg z_i$	$1 - c_i$	1

Tabla 3.3: Parámetros del modelo OR "con ruido" para el enlace $X_i \rightarrow Y$.

Mediante la ecuación (3.4) podemos obtener la TPC teniendo en cuenta que $f_{\text{OR}}(\mathbf{z}) = \neg y$ solamente para la configuración $(\neg z_1, \ldots, \neg z_n)$. Por tanto,

$$P(\neg y|\mathbf{x}) = \prod_{i=1}^{n} P(\neg z_i|x_i) = \prod_{i \in I+(\mathbf{x})} P(\neg z_i| + x_i) \cdot \prod_{i \in I\neg(\mathbf{x})} P(\neg z_i|\neg x_i)$$

donde $I_+(\mathbf{x})$ representa los padres de Y que toman el valor $+x_i$ (es decir, "verdadero", "presente" o "positivo") en la configuración $\mathbf{x}$ e $I_\neg(\mathbf{x})$ a los demás. Por ejemplo, si $\mathbf{x} = (+x_1, \neg x_2, +x_3)$ entonces $I_+(\mathbf{x}) = \{X_1, X_3\}$ e $I_\neg(\mathbf{x}) = \{X_2\}$. Utilizando los parámeros introducidos en las ecuaciones (3.5) y (3.6), tenemos que

$$P(\neg y|\mathbf{x}) = \prod_{i \in I_+(\mathbf{x})} q_i = \prod_{i \in I_+(\mathbf{x})} (1 - c_i) \ . \tag{3.7}$$

Cuando hay dos causas X_1 y X_2 de un efecto común Y, esta ecuación conduce a la tabla 3.4. Observe que si $c_1 = c_2 = 1$, esta tabla coincide con la TPC del modelo OR determinista (tabla 3.2). De hecho, el modelo OR determinista es un caso particular del modelo OR "con ruido" en que $c_i = 1$ para toda i.

La ecuación (3.7) implica que cuando todas las causas están ausentes, entonces también Y está ausente:

$$P(\neg y|\neg x_1, \ldots, \neg x_n) = 1 \ . \tag{3.8}$$

Del mismo modo, cuando X_i está presente y las demás causas de Y están ausentes, entonces

$$P(+y| + x_i, \neg x_{j\,(\forall j,\, j \neq i)}) = c_i \ , \tag{3.9}$$

$P(+y\|x_1, x_2)$	$+x_1$	$\neg x_1$
$+x_2$	$c_1 + (1 - c_1) \cdot c_2$	c_2
$\neg x_2$	c_1	0

Tabla 3.4: TPC para un modelo OR "con ruido". Los padres son X_1 y X_2.

lo cual es coherente con la definición de c_i como la probabilidad de que X_i cause Y (cf. ecuación (3.5)).

Ejercicio 3.1 Una máquina de producción de piezas de plástico puede averiarse debido a tres causas: la mala calidad de la materia prima utilizada, un calentamiento excesivo del torno o un exceso de tensión. Hay un sensor que detecta la mala calidad de la materia prima en el 95% de los casos, lo cual hace que aunque la materia sea defectuosa sólo provoque una avería en el 5% de los casos. Del mismo modo, un sensor de temperatura hace que la máquina generalmente se detenga a tiempo en el 93% de los casos, por lo cual un calentamiento sólo provoca averías en el 7% de los casos. Sin embargo, el diferencial instalado no funciona correctamente, y por ello la probabilidad de que un exceso de tensión provoque una avería es el 80%.

1. Construya una TPC para este problema aplicando la ecuación (3.7).
2. Construya la TPC con el programa Elvira o con OpenMarkov, introduciendo los parámetros dados en el enunciado. Compruebe que los valores calculados por el programa coinciden con los que Vd. ha calculado.
3. ¿Cuántos parámetros independientes habría necesitado para construir esta TPC si no hubiera aplicado el modelo OR?

Modelo OR residual

Como hemos dicho en la sección 3.2.2, en general es imposible en la práctica incluir todas las causas posibles de un efecto, y por ello necesitamos una versión residual del modelo OR. En este caso, la variable Z_L también es booleana, y la probabilidad residual $P(z_L)$ viene dada por el parámetro c_L,

$$\begin{cases} P(+z_L) = c_L \\ P(\neg z_L) = 1 - c_L \ . \end{cases}$$

La TPC para este modelo es [23],

$$P(\neg y|\mathbf{x}) = (1 - c_L) \cdot \prod_{i \in I_+(\mathbf{x})} (1 - c_i) \ . \tag{3.10}$$

La tabla 3.5 muestra la forma de la TPC para $n = 2$. Observe que cuando $c_L = 0$, esta tabla coincide con la del modelo OR "con ruido" (tabla 3.4).

Cuando todas las causas explícitas (los padres de Y en la red bayesiana) toman el valor "ausente", la probabilidad de que Y esté presente es

$$P(+y|\neg x_1, \ldots, \neg x_n) = c_L \ , \tag{3.11}$$

$P(+y\|x_1,x_2)$	$+x_1$	$\neg x_1$
$+x_2$	$1-(1-c_1)\cdot(1-c_2)\cdot(1-c_L)$	$c_2+(1-c_2)\cdot c_L$
$\neg x_2$	$c_1+(1-c_1)\cdot c_L$	c_L

Tabla 3.5: TPC para el modelo OR residual con dos padres.

lo que significa que Z_L puede interpretarse como una variable que indica si las causas implícitas (es decir, las que no aparecen como padres de Y en la red bayesiana) han producido Y o no.

Ejercicio 3.2 En el ejercicio 3.1 hemos considerado tres causas por las cuales una máquina de producción de piezas de plástico podía averiarse. Sin embargo, hay otros muchos factores que no hemos considerado y que pueden producir una avería con una probabilidad de 0'001.

1. Construya una TPC para este problema aplicando la ecuación (3.10). Puede aprovechar los cálculos que hizo en el ejercicio 3.1.
2. Construya la TPC con el programa Elvira o con OpenMarkov, introduciendo los parámetros dados en el enunciado. Compruebe que los valores calculados por el programa coinciden con los que Vd. ha calculado.

Modelo MAX

El modelo MAX "con ruido" surgió como una extensión del modelo OR "con ruido" para variables multivaluadas. En él, cada Z_i representa el valor de Y producido por X_i. El valor de Y resultante es el máximo de los valores individuales producidos por cada uno de los padres: $y = f_{\text{MAX}}(\mathbf{z})$. Los parámetros para el enlace $X_i \to Y$ son

$$c_{z_i}^{x_i} = P(z_i|x_i) \tag{3.12}$$

o lo que es lo mismo,

$$c_y^{x_i} = P(Z_i = y|x_i) \tag{3.13}$$

Cada $c_y^{x_i}$ representa la probabilidad de que X_i, tomando el valor x_i, eleve el valor de Y hasta y.

Existe una versión causal de este modelo en la cual cada variable tiene un *estado neutro*, denotado por $\neg y$ o $\neg x_i$, que en general representa la ausencia de anomalía. El estado neutro de Y es el mínimo de sus posibles valores; por ejemplo, los valores de Y pueden ser $\text{dom}(Y) = \{ausente, leve, moderada, severa\}$, de modo que $\neg y = ausente$. Cuando X_i está en su estado neutro no altera el valor de Y; podríamos decir que cuando la anomalía X_i está ausente no produce la anomalía Y. Por tanto,

$$c_{\neg y}^{\neg x_i} = P(Z_i = \neg y|\neg x_i) = 1\ . \tag{3.14}$$

De la ecuación (3.4) se deduce que la anomalía Y está ausente cuando cada uno de sus padres está en estado neutro (es decir, cuando todas las anomalías X_i están ausentes):

$$P(\neg y|\neg x_1, \ldots, \neg x_n) = 1\ . \tag{3.15}$$

También se deduce de la ecuación (3.4) que

$$P(y|x_i, \neg x_{j\,(\forall j,\, j\neq i)}) = P(Z_i = y|x_i) = c_y^{x_i}\ , \tag{3.16}$$

lo cual significa que el parámetro $c_y^{x_i}$ representa la probabilidad de que Y tome el valor y cuando X_i toma el valor x_i y las demás causas de Y permanecen en sus estados neutros. Es decir, $c_y^{x_i}$ representa la capacidad de X_i para elevar el estado de Y hasta el valor y independientemente de los estados de otras causas de Y (que quizá van a elevar Y hasta un valor más alto).

Existe también un modelo MAX residual, semejante al modelo OR residual, en el cual el parámetro c_y^L representa la probabilidad de que $Y = y$ cuando todas las causas de Y explícitas en el modelo están en sus estados neutros:

$$c_y^L = P(y|\neg x_1, \ldots, \neg x_n) \; . \tag{3.17}$$

Ejemplo 3.3 (Modelo MAX) Una cierta enfermedad Y puede deberse al exceso o al defecto de la sustancia X_1 en la sangre del paciente. También puede deberse a una anomalía X_2. Podemos representar este problema escogiendo tres variables, Y, X_1 y X_2, con los siguientes dominios:

$$\begin{aligned} \mathrm{dom}(Y) &= \{ausente, leve, moderada, severa\} \\ \mathrm{dom}(X_1) &= \{reducido, normal, aumentado\} \\ \mathrm{dom}(X_2) &= \{ausente, presente\} \end{aligned}$$

Sus valores neutros son $\neg y$ =*ausente*, $\neg x_1$ =*normal*, y $\neg x_2$ =*ausente*, respectivamente. La figura 3.5.*a* muestra los parámetros para este modelo en Elvira, y la figura 3.5.*b* la TPC calculada a partir de ellos. Podemos observar que cuando X_1 y X_2 están en sus estados neutros (5ª columna), la probabilidad de Y coincide con la probabilidad residual mostrada en la figura 3.5.*a*, lo cual concuerda con la ecuación (3.17).

3.2.4. Uso de los modelos canónicos en la construcción de redes bayesianas

El modelo OR se puede aplicar a una familia dentro de una red bayesiana cuando se cumplen las siguientes condiciones:

1. ¿Son booleanas todas las variables de esa familia?

   Por ejemplo, si uno de los padres es la variable *Sexo*, el modelo OR no se puede aplicar porque no es posible indentificar uno de sus valores (*varón* o *mujer*) con la causa de una anomalía y el otro con su asuencia. Si el *Sexo* figura entre los padres de una familia, generalmente actúa como un factor de riesgo o como precondición para otra variable, y en ese caso la elección más adecuada puede ser el modelo AND [23].

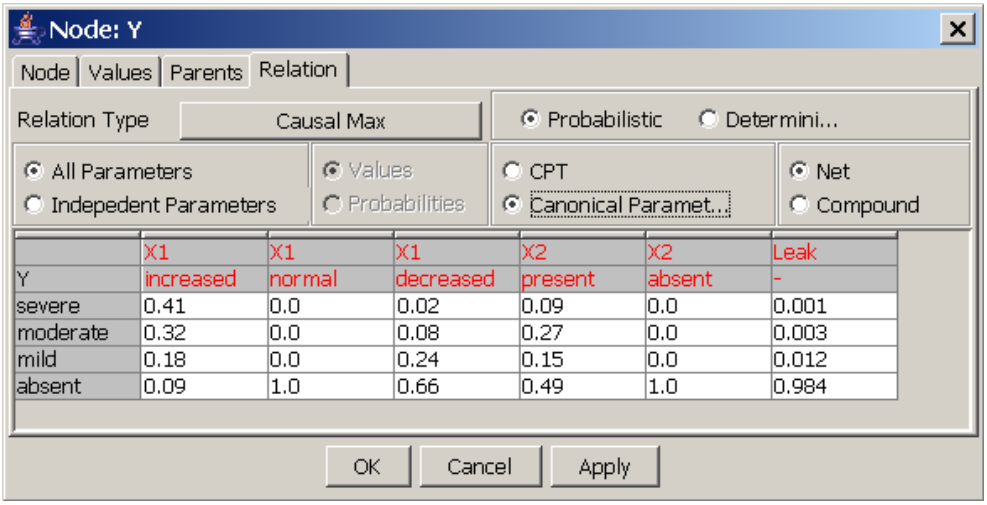
2. ¿Hay un mecanismo causal para cada padre, X_i, tal que X_i es capaz de producir Y en ausencia de las demás anomalías?

   Recordemos que en el modelo OR la variable Z_i (figura 3.3) representa el hecho de que el efecto Y ha sido producido por X_i. Cuando el efecto no puede ser producido por mecanismos causales individuales, sino que es necesaria la concurrencia de varias causas, entonces no se puede aplicar el modelo OR.

3. ¿Cómo son los mecanismos causales?

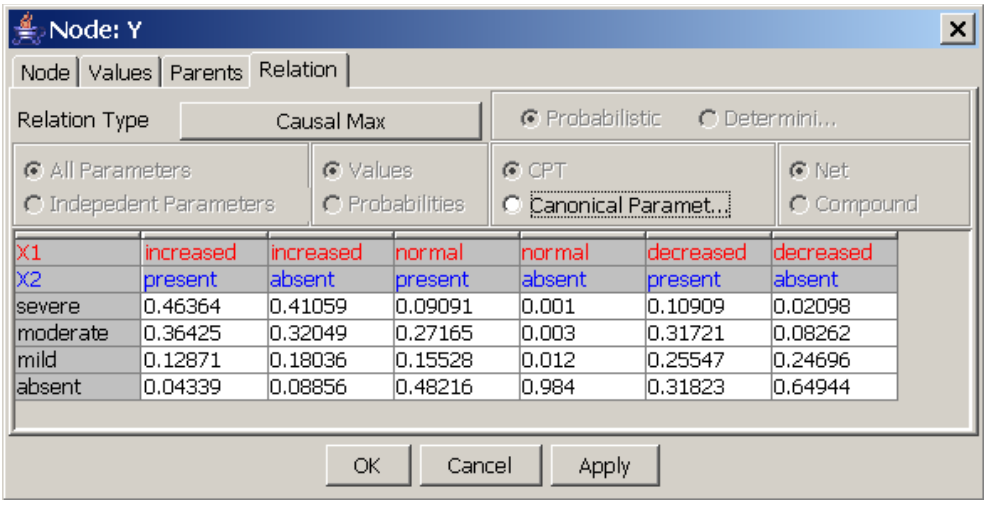
   Si algunos de los mecanismos son no-deterministas, entonces hay que aplicar el modelo OR "con ruido" en lugar del modelo determinista. Si hay otras causas no explícitas en

a)



	X1	X1	X1	X2	X2	Leak
Y	increased	normal	decreased	present	absent	-
severe	0.41	0.0	0.02	0.09	0.0	0.001
moderate	0.32	0.0	0.08	0.27	0.0	0.003
mild	0.18	0.0	0.24	0.15	0.0	0.012
absent	0.09	1.0	0.66	0.49	1.0	0.984

b)



X1	increased	increased	normal	normal	decreased	decreased
X2	present	absent	present	absent	present	absent
severe	0.46364	0.41059	0.09091	0.001	0.10909	0.02098
moderate	0.36425	0.32049	0.27165	0.003	0.31721	0.08262
mild	0.12871	0.18036	0.15528	0.012	0.25547	0.24696
absent	0.04339	0.08856	0.48216	0.984	0.31823	0.64944

Figura 3.5: Modelo MAX causal en Elvira. *a*) Parámetros canónicos. *b*) Tabla de probabilidad condicional (TPC).

el modelo capaces de producir Y, entonces hay que aplicar el modelo OR residual en vez del modelo "con ruido".

4. ¿Son independientes los mecanismos causales?

   En general ésta es la condición más difícil de comprobar, pues en muchos casos nuestro conocimiento del problema es limitado y no podemos estar seguros de que los mecanismos causales y sus inhibidores no interactúan. En la práctica, cuando no hay interacciones conocidas suponemos que se cumple esta condición, y así podemos aplicar el modelo OR "con ruido" o residual. Si no se cumple esta condición, tendremos que utilizar otro modelo que no sea del tipo IIC; algunos de los modelos propuestos en la literatura son el modelo OR "con ruido" recursivo (en inglés, *recursive noisy OR*, o RNOR), que también tiene una versión que incluye inhibidores (*inhibited RNOR*) [23].

Los criterios para la aplicación del modelo MAX son similares.

Existen otros modelos canónicos de tipo IIC basado en las funciones AND, MIN, XOR, umbral, etc. (véase la tabla 3.1), así como otros tipos de modelos, entre los cuales se encuentran los *modelos canónicos simples*, que requieren menos parámetros que los de tipo IIC. La explicación de estos modelos, los criterios para utilizarlos en la construcción de redes bayesianas y algunos consejos para la obtención de los parámetros numéricos pueden encontrarse en el artículo ya citado, [23].

3.3. Aprendizaje automático a partir de bases de datos

Con la generalización del uso de los ordenadores, que se produjo sobre todo en la década de los 80, la disponibilidad de bases de datos no ha dejado de crecer exponencialmente. Ello hace que exista un interés cada vez mayor en extraer conocimiento de ellas, es decir, en construir modelos matemáticos que permitan predecir el futuro y tomar las mejores decisiones. Es lo que se conoce como *minería de datos*.

En esta sección vamos a estudiar la construcción de modelos gráficos probabilistas, concretamete redes bayesianas, a partir de bases de datos. Éste es un campo de gran actualidad. Podríamos decir que al menos un tercio de las publicaciones que se producen cada año sobre redes bayesianas están dedicadas a esta cuestión. Dado que se trata de un problema matemáticamente complejo y que la literatura existente es amplísima, nos vamos a limitar a explicar algunos conceptos básicos y a dar referencias para que el lector interesado pueda consultar referencias especializadas.

3.3.1. Planteamiento del problema

De forma general, podemos decir que el problema del aprendizaje consiste en construir, a partir de un conjunto de datos, el modelo que mejor represente la realidad, o mejor dicho, una porción del mundo real en la cual estamos interesados. Como en el caso de la construcción manual de redes bayesianas, el aprendizaje de este tipo de modelos tiene dos aspectos: el aprendizaje paramétrico y el aprendizaje estructural. En algún caso puede ocurrir que tengamos ya el grafo de la red, construido con la ayuda de un experto, y estemos interesados solamente en la obtención de las probabilidades condicionales; sería un caso de aprendizaje paramétrico. Sin embargo, es más habitual que deseemos construir a partir de los datos la red completa, es decir, tanto el grafo como los parámetros de la red.

Hay dos métodos principales para construir la red. El primero consiste en realizar, a partir de las frecuencias observadas en la base de datos, una estimación de la distribución de probabilidad que rige el mundo real; las relaciones de dependencia e independencia probabilista de dicha distribución indican cuál debe ser la estructura del grafo. Es decir, se trata de buscar un grafo que sea mapa de independencias de la distribución de probabilidad (véase la sec. 1.5.2). Naturalmente, puede haber más de una solución, pues como vimos en la sección 1.6.1, existen grafos equivalentes en sentido probabilista, es decir, grafos que representan las mismas relaciones de independencias y de (posibles) dependencias.

El otro método de aprendizaje estructural consiste en realizar una búsqueda heurística utilizando alguna medida de calidad: en general, se parte de una red sin enlaces y se van añadiendo enlaces uno a uno hasta que la red representa adecuadamente la distribución de probabilidad obtenida de la base de datos. También en este método hay que tener en cuenta la existencia de grafos equivalentes en sentido probabilista.

Antes de ver con detalle cada uno de estos métodos (el aprendizaje paramétrico en la sec. 3.3.3, el estructural basado en relaciones de independencia en la 3.3.4 y el estructural mediante búsqueda heurística en la 3.3.5), vamos a estudiar primero algunas cuestiones generales sobre aprendizaje que nos ayudarán a entender mejor el aprendizaje de redes bayesianas.

3.3.2. Cuestiones generales sobre aprendizaje

El problema del sobreajuste

Hemos dicho antes que el problema del aprendizaje consiste en construir el modelo que mejor represente una porción del mundo real en la cual estamos interesados. Sin embargo, en la práctica no conocemos toda la realidad, sino sólo un conjunto de datos. Por ejemplo, si queremos construir un modelo para el diagnóstico de cáncer de hígado, no conocemos la realidad completa (todos los posibles pacientes), sino sólo los casos recogidos en cierta base de datos. Podríamos pensar entonces que el objetivo es construir el modelo que más se ajusta a los datos disponibles. Sin embargo, se comprueba en muchos casos que el modelo que mejor se ajusta *a los datos* no es necesariamente el que mejor se ajusta *a la realidad*.

Este fenómeno se denomina *sobreajuste* (en inglés, *overfitting*) y tiende a ocurrir cualquiera que sea el tipo de modelo que queremos aprender (ya sea un árbol de clasificación, una red neuronal, un conjunto de reglas difusas...). Vamos a ver tres ejemplos para el caso de las redes bayesianas, dos de aprendizaje paramétrico y uno de aprendizaje estructural.

Ejemplo 3.4 Volvamos al ejemplo 1.39, concretamente a la tabla de $P(t|d)$ que aparece en la página 19. Supongamos que tenemos una base de datos de unos 150.000 casos, de los cuales aproximadamente 150 corresponderán a avería eléctrica, pues $P(d^e) = 0'001$. La probabilidad de tener temperatura reducida en caso de avería eléctrica, $P(t^r|d^e)$, es 0'01. Por tanto, es posible que en la base de datos no haya ningún caso de avería eléctrica con temperatura reducida. (La probabilidad de que ocurra esto es $0'99^{150} = 0'22$.) En ese caso, el modelo que mejor se ajusta a los datos dirá que $P(t^r|d^e) = 0$. Según este modelo, cuando hay temperatura reducida es imposible que exista avería eléctrica, por muchos otros hallazgos que pudieran confirmar ese tipo de avería. El problema es que ha habido un sobreajuste del modelo a los datos.

Ejemplo 3.5 Supongamos que queremos construir un modelo para el diagnóstico diferencial de tres enfermedades, E_a, E_b y E_c, a partir de una serie de hallazgos. Entre ellos se encuentra

un síntoma S tal que $P(+s|e_a) = 0'40$, $P(+s|e_b) = 0'05$ y $P(+s|e_c) = 0'01$. Si la base de datos contiene 150 pacientes con la enfermedad E_c es posible que ninguno de ellos presente el síntoma S. (La probabilidad de que ocurra esto es $0'99^{150} = 0'22$.) Al realizar el aprendizaje de la red bayesiana, el modelo que más se ajusta a los datos dirá que $P(\neg s|e_c) = 1$ y $P(+s|e_c) = 0$, y según él, la probabilidad de E_c dada la presencia del síntoma es 0; es decir, el hallazgo $+s$ sería capaz de anular toda la evidencia a favor de E_c que pudieran aportar los demás hallazgos. Esto se debe a que ha habido un sobreajuste del modelo a los datos.

Ejemplo 3.6 Sea un problema de tres variables, A, B y C, tal que B y C son condicionalmente independientes dado A, es decir, $P(b|a) \cdot P(c|a) = P(b, c|a)$. De ahí se deduce que las frecuencias observadas en la base de datos cumplirán aproximadamente la igualdad $N(+a, +b) \cdot N(+a, +c) = N(+a, +b, +c) \cdot N(+a)$. Sin embargo, es muy improbable que esta igualdad se cumpla exactamente; es decir, lo más probable es que aparezca una pequeña correlación accidental entre B y C (dado A) en la base de datos, a pesar de que en el mundo real son condicionalmente independientes. El modelo que mejor se ajustaría a dicha base de datos incluiría un enlace espurio $B \to C$ o $C \to B$, debido al sobreajuste.

Supongamos ahora que tenemos una red bayesiana cuyo grafo es el de la figura 1.6 (pág. 22).

Este último ejemplo nos muestra que las correlaciones espurias existentes en la base de datos pueden llevar a construir un modelo que contenga un enlace entre cada par de nodos, es decir, un grafo completo. Eso plantea tres problemas:

1. La falta de precisión, que ya hemos comentado en el ejemplo 3.5.

2. La pérdida de información: el modelo resultante no mostraría ninguna de las relaciones de independencia existentes en el mundo real. Por ejemplo, en el grafo de la figura 1.6 (pág. 22) se observan ciertas relaciones de independencia entre las variables, mientras que si trazáramos un enlace entre cada par de variables perderíamos esa información.

3. El tamaño del modelo: una red bayesiana de n variables basada en un grafo completo contendrá una tabla de probabilidad de n variables, cuyo tamaño será el mismo que el de la probabilidad conjunta, y además otra tabla de $n-1$ variables, otra de $n-2$, etc. En este caso, construir una red bayesiana es peor que representar la tabla de probabilidad conjunta explícitamente. Como el tamaño del problema crece de forma exponencial, con los ordenadores actuales sólo podríamos construir redes bayesianas de unas 25 o 30 variables. Sin embargo, luego veremos que hoy en día existen algoritmos de aprendizaje capaces de construir modelos con cientos de variables.

La forma habitual de evitar el sobreaprendizaje a la hora de construir el grafo de la red es tratar de construir modelos sencillos, es decir, con un pequeño número de enlaces. Más adelante veremos cómo lo consigue cada uno de los métodos de aprendizaje que vamos a estudiar.

Aprendizaje probabilista

Denominamos aprendizaje probabilista a aquél que se basa en los principios de la teoría de la probabilidad, independientemente de que el modelo aprendido sea de tipo probabilista (por ejemplo, una red bayesiana) o no probabilista (por ejemplo, una red neuronal). Dentro de él vamos a estudiar dos modalidades: el de máxima verosimilitud y el bayesiano.

Aprendizaje de máxima verosimilitud La verosimilitud es una función, habitualmente representada por λ, que indica en qué medida cada hipótesis o cada modelo explica los datos observados:

$$\lambda(\text{modelo}) = P(\text{datos} \mid \text{modelo}) \tag{3.18}$$

El aprendizaje de máxima verosimilitud consiste en tomar la hipótesis o el modelo que maximiza esta función.

Ejemplo 3.7 Deseamos construir un modelo que indique la probabilidad de supervivencia para cierta enfermedad. Se trata de un modelo muy simple, pues sólo tiene un parámetro, θ, la tasa de supervivencia. Sabemos que el verdadero valor de θ ha de estar entre entre 0 y 1. Para cada paciente individual,

$$\begin{cases} P(\text{supervivencia}) = \theta \\ P(\text{fallecimiento}) = 1 - \theta \end{cases} \tag{3.19}$$

Supongamos ahora que tenemos una base de datos de n pacientes que sufren esa enfermedad, de los cuales han sobrevidido m, y a partir de ellos tratamos de estimar el valor del parámetro θ. (Como dicen Castillo et al. [7, pág. 505], "en el lenguaje de los estadísticos, el aprendizaje estadístico se llama *estimación*".) La verosimilitud para estos datos es[3]

$$\lambda(\theta) = P(\text{datos}|\theta) = \frac{n!}{m!(n-m)!}\,\theta^m\,(1-\theta)^{n-m} \tag{3.20}$$

La estimación de máxima verosimilitud para θ consiste en tomar el valor que maximiza $\lambda(\theta)$, que en este caso es $\theta = m/n$.[4]

Aprendizaje bayesiano Uno de los tipos de aprendizaje desarrollados en el campo de la inteligencia artificial, inspirado en la estadística bayesiana, es el demominado *aprendizaje bayesiano*, que consiste en asignar una probabilidad a priori a cada uno de los modelos, $P(\text{modelo})$. La probabilidad a posteriori del modelo dados los datos se define mediante el teorema de Bayes:

$$P(\text{modelo} \mid \text{datos}) = \frac{P(\text{modelo}) \cdot P(\text{datos} \mid \text{modelo})}{P(\text{datos})} \tag{3.21}$$

Observe la semejanza con el problema del diagnóstico probabilista estudiado en la sección 1.1.3. Allí se trataba de diagnosticar la enfermedad que mejor explicaba los síntomas.

[3] La probabilidad $P(m|\theta)$ viene dada por la distribución de Bernouilli. El factor θ^m corresponde a la probabilidad de que sobrevivan m pacientes, $(1-\theta)^{n-m}$ es la probabilidad de que mueran los demás, y $n!/(m!(n-m)!)$ es el número de combinaciones de n elementos tomados de m en m.

[4] Para maximizar esta función, resulta más cómodo tomar su logaritmo neperiano (como el logaritmo es una función monótona, el máximo de λ es el mismo que el de $\ln\lambda$):

$$\ln\lambda(\theta) = \ln\frac{n!}{m!(n-m)!} + m\ln\theta + (n-m)\ln(1-\theta)$$

$$\frac{d}{d\theta}\ln\lambda(\theta) = m\frac{1}{\theta} - (n-m)\frac{1}{1-\theta}$$

$$\frac{d}{d\theta}\ln\lambda(\theta) = 0 \Longleftrightarrow \theta = \frac{m}{n}$$

Se puede comprobar además que la derivada segunda de $\ln\lambda(\theta)$ es negativa en todo el intervalo (0,1), y por tanto $\lambda(\theta)$ tiene un máximo en $\theta = m/n$.

Aquí tratamos de "diagnosticar" el modelo que mejor explica los datos observados. En ambos casos, el "diagnóstico" se basa en una probabilidad a priori, que representa el conocimiento que teníamos antes de observar las observaciones, y en una verosimilitud que indica en qué medida cada una de nuestras hipótesis (en este caso, cada modelo) explica los datos observados.

En la ecuación anterior el denominador es una constante, en el sentido de que es la misma para todos los modelos. Por tanto, si no queremos conocer la probabilidad absoluta, sino sólo cuál de los modelos tiene mayor probabilidad que los demás, podemos quedarnos con una versión simplificada de ella:

$$P(\text{modelo} \mid \text{datos}) \propto P(\text{modelo}) \cdot P(\text{datos} \mid \text{modelo}) \tag{3.22}$$

El aprendizaje de máxima verosimilitud es un caso particular del aprendizaje bayesiano, pues cuando la probabilidad a priori es constante, entonces la probabilidad a posteriori es proporcional a la verosimilitud,

$$P(\text{modelo}) = \text{constante} \Longrightarrow P(\text{modelo} \mid \text{datos}) \propto P(\text{datos} \mid \text{modelo}) \tag{3.23}$$

y maximizar una de ellas es lo mismo que maximizar la otra.

Ejemplo 3.8 Supongamos que en el problema del ejemplo anterior la probabilidad a priori de θ es constante: $P(\theta) = c$. En este caso,

$$P(\theta|\text{datos}) \propto P(\theta) \cdot P(\text{datos}|\theta) \propto \theta^m \, (1-\theta)^{n-m} \tag{3.24}$$

El máximo de $P(\theta|m)$ es el mismo que el de $\lambda(\theta)$, es decir, $\theta = m/n$.

Ejemplo 3.9 Volviendo de nuevo al ejemplo 3.7, tomamos como probabilidad a priori $P(\theta) = c\ \theta^k(1-\theta)^l$, donde c es una constante de normalización. En este caso, la probabilidad a posteriori es:[5]

$$P(\theta|\text{datos}) \propto P(\theta) \cdot P(\text{datos}|\theta) \propto \theta^{k+m}(1-\theta)^{l+n-m} \tag{3.25}$$

El máximo de esta función corresponde al valor $\theta = (k+m)/(k+l+n)$.

En los dos ejemplos anteriores, la distribución $P(\theta)$ indica la probabilidad de un parámetro probabilista (una probabilidad), y por eso suele decirse que $P(\theta)$ es una *probabilidad de segundo orden.*

Insistimos en que lo más característico del aprendizaje bayesiano es el hecho de asignar una probabilidad a priori a cada uno de los modelos. En los dos ejemplos anteriores, el modelo venía caracterizado por el parámetro θ, y por eso hemos identificado $P(\text{modelo})$ con $P(\theta)$.

Recordamos también que el aprendizaje bayesiano **no** está especialmente relacionado con las redes bayesianas: como vamos a ver a continuación, hay métodos de aprendizaje de redes bayesianas que no tienen nada que ver con el aprendizaje bayesiano, y viceversa, puede aplicarse el aprendizaje bayesiano para construir modelos que no tienen nada que ver con las redes bayesianas; por ejemplo, una red neuronal.

[5] Quizá el lector se pregunte por qué hemos escogido esa probabilidad a priori. Observe que, de acuerdo con la ecuación (3.24), si la probabilidad a priori es constante y los datos observados nos dicen que han sobrevivido k pacientes y han muerto l, la probabilidad a posteriori de θ es proporcional a $\theta^k(1-\theta)^l$. Por tanto, la probabilidad a priori considerada en este ejemplo podría proceder de tales datos.

Supongamos ahora que tenemos una segunda base de datos con m supervivientes y $(n-m)$ fallecidos. Para estos datos, la verosimilitud es proporcional a $\theta^m(1-\theta)^{n-m}$. La probabilidad a posteriori final es $\theta^{k+m}(1-\theta)^{l+n-m}$. Este resultado es coherente con el hecho de que en total ha habido $k+m$ supervivientes y $l+n-m$ fallecidos.

3.3.3. Aprendizaje paramétrico

El aprendizaje paramétrico da por supuesto que conocemos la estructura (el grafo) de la red bayesiana y, en consecuencia, también conocemos la factorización de la probabilidad (cf. ec. (1.42)) y cuáles son las probabilidades condicionales que forman la red. En el caso de variables discretas, podemos considerar que cada probabilidad condicional $P(x_i|pa(X_i))$ es un parámetro, que llamaremos $\theta_{P(x_i|pa(X_i))}$. Sin embargo, para cada configuración $pa(X_i)$ se cumple que

$$\sum_{x_i} P(x_i|pa(X_i)) = 1 \tag{3.26}$$

Por tanto, si X_i toma n_{X_i} valores, el número de parámetros independientes para cada configuración $pa(X_i)$ es $n_{X_i} - 1$.

Notación Dado que resulta engorroso en muchos casos escribir $\theta_{P(x_i|pa(X_i))}$, en la bibliografía sobre aprendizaje es habitual utilizar la notación θ_{ijk}, cuyo significado es el siguiente:

- i representa la variable X_i. Por tanto, si la red tiene n_I variables, se cumple que $1 \leq i \leq n_I$.
- j representa la j-ésima configuración de los padres de X_i. Si X_i tiene k padres binarios, entonces hay 2^k configuraciones de $Pa(X_i)$, lo cual implica que $1 \leq j \leq 2^k$.
- k representa el valor que toma la variable X_i. El k-ésimo valor de X_i se puede escribir como x_i^k. Si esta variable puede tomar n_{X_i} valores, entonces $1 \leq k \leq n_{X_i}$.

Con esta notación, la ecuación anterior se puede reescribir como

$$\forall i, \forall j, \quad \sum_{k=1}^{n_{X_i}} \theta_{ijk} = 1 \tag{3.27}$$

También es habitual que $\boldsymbol{\theta}$ denote el conjunto de parámetros de la red (todas las probabilidades condicionales), $\boldsymbol{\theta}_i$ el subconjunto de probabilidades condicionales asociadas a la familia de X_i, y $\boldsymbol{\theta}_{ij}$ el conjunto de probabilidades condicionales correspondientes a la j-ésima configuración de $Pa(X_i)$. Por tanto, en el caso de variables binarias, tenemos:

- Para cada configuración $Pa(X_i)$, $\boldsymbol{\theta}_{ij} = \{\theta_{ijk} \mid 1 \leq k \leq n_{X_i}\}$.
- Para cada variable X_i, $\boldsymbol{\theta}_i = \bigcup_{j=1}^{2^k} \boldsymbol{\theta}_{ij}$.
- Para la red, $\boldsymbol{\theta} = \bigcup_{i=1}^{n_I} \boldsymbol{\theta}_i$.

Cuando tenemos una base de datos, N_{ijk} representa el número de casos en que la variable X_i toma el valor x_i^k y los padres de X_i toman los valores correspondientes a la j-ésima configuración.

Aprendizaje (estimación) de máxima verosimilitud

Una forma sencilla de estimar cada uno de ellos es por el método de la máxima verosimilitud: si la base de datos contiene N_{ij} casos en que los padres de X_i toman los valores correspondientes a la j-ésima configuración de $Pa(X_i)$, y en N_{ijk} de esos casos la variable X_i toma su k-ésimo valor, entonces la estimación del parámetro $\theta_{P(x_i|pa(X_i))}$ es $\hat{\theta}_{P(x_i|pa(X_i))} = N_{ijk}/N_{ij}$. De acuerdo con la notación introducida anteriormente:

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{N_{ij}} \tag{3.28}$$

Esta ecuación tiene dos problemas. El primero es que cuando $N_{ij} = 0$ (es decir, cuando no hay ningún caso en la base de datos correspondiente a esa configuración de padres de X_i) entonces también $N_{ijk} = 0$ y por tanto nos encontramos con una indeterminación. El otro problema es el sobreajuste, tal como hemos comentado en el ejemplo 3.5. Estos dos problemas se resuelven mediante la estimación bayesiana, como vamos a ver en seguida.

Aprendizaje bayesiano

Una alternativa más compleja es utilizar el aprendizaje bayesiano, lo cual implica dar una distribución de probabilidad para los parámetros, $P(\boldsymbol{\Theta})$. Observe que estamos utilizando la letra griega theta *en mayúscula* porque en el caso del aprendizaje bayesiano cada parámetro se trata como una variable aleatoria, es decir, una variable que tiene una distribución de probabilidad asociada.

La primera cuestión, por tanto, es determinar la forma de $P(\boldsymbol{\Theta})$ y la segunda obtener un algoritmo que permita estimar los valores de $\boldsymbol{\Theta}$ de forma eficiente, es decir, con un consumo de tiempo y de memoria razonables. Para poder abordar ambas cuestiones es necesario introducir hipótesis adicionales y restricciones sobre la forma de $P(\boldsymbol{\Theta})$. La mayor parte de los métodos propuestos en la literatura se basan en la hipótesis de *independencia de los parámetros*, que se expresa así:

$$P(\boldsymbol{\theta}) = \prod_i \prod_j P(\boldsymbol{\theta}_{ij}) \tag{3.29}$$

Hay autores que dividen esta hipótesis en dos partes: la primera es la *independencia global de los parámetros*, es decir que dos parámetros (las probabilidades condicionales) de dos familias diferentes son independientes entre sí,

$$P(\boldsymbol{\theta}) = \prod_i P(\boldsymbol{\theta}_i)\ ; \tag{3.30}$$

la segunda es la *independencia local de los parámetros*, es decir, que dentro de una familia los parámetros correspondientes a una configuración de los padres[6] son independientes de los correspondientes a las demás configuraciones:

$$P(\boldsymbol{\theta}_i) = \prod_j P(\boldsymbol{\theta}_{ij}) \tag{3.31}$$

Luego hay que indicar qué forma tiene la distribución de probabilidad a priori para los parámetros de una configuración, $P(\boldsymbol{\theta}_{ij})$. Lo más frecuente en la literatura es suponer que

[6] Si X_i es una variable binaria sólo se necesita un parámetro por cada configuración de los padres.

se trata de una distribución de Dirichlet. Dado el nivel de este curso, no vamos a entrar en los detalles del proceso. Tan sólo vamos a comentar que cuando no hay conocimiento a priori es razonable asignar una probabilidad uniforme (constante) a todos los valores de los parámetros. En este caso, es decir, cuando tenemos una distribución de Dirichlet uniforme, el estimador de θ_{ijk} correspondiente al valor máximo de la probabilidad a posteriori es

$$\hat{\theta}_{ijk} = \frac{N_{ijk} + 1}{N_{ij} + n_{X_i}} \tag{3.32}$$

La diferencia de esta ecuación con la (3.28) es que sumamos 1 en el numerador, lo cual obliga a sumar n_{X_i} para que las probabilidades sumen la unidad (cf. ecuaciones (3.26) y (3.27)). Esta modificación se suele denominar *corrección de Laplace*. Observe que el resultado es el mismo que si aplicáramos la estimación de máxima verosimilitud pero añadiendo a la base de datos un caso ficticio por cada configuración de la familia de X_i; es decir, para cada configuración de $Pa(X_i)$ se añaden n_{X_i} casos, uno por cada valor x_i^k de X_i.

Así se resuelven los dos problemas que presentaba el aprendizaje de máxima verosimilitud. El primero, es que aunque $N_{ij} = 0$ el cociente está definido, pues en ese caso $\hat{\theta}_{ijk} = 1/n_{X_i}$ para todo k. El segundo es que se evita el sobreajuste. Así en el ejemplo 3.5, en vez de tener $\hat{P}(\neg s|e_c) = 1$ y $\hat{P}(+s|e_c) = 0$, tendríamos $\hat{P}(\neg s|e_c) = (1+1)/(1+2) = 2/3$ y $\hat{P}(+s|e_c) = (0+1)/(1+3) = 1/3$.

Existen otras variantes del método, similares a la corrección de Laplace, que en vez de añadir un caso ficticio por cada configuración de $Pa(X_i)$ y cada valor de X_i, añaden α_{ijk} casos, donde α_{ijk} puede ser un número no entero:

$$\hat{\theta}_{ijk} = \frac{N_{ijk} + \alpha_{ijk}}{N_{ij} + \alpha_{ij}} \tag{3.33}$$

donde α_{ij}, definido como $\alpha_{ij} = \sum_k \alpha_{ijk}$, se denomina *espacio muestral equivalente*, porque el resultado es el mismo que si hubiéramos realizado la estimación de los parámetros de $\boldsymbol{\theta}_{ij}$ mediante el método de máxima verosimilitud pero añadiendo α_{ij} casos a la base de datos. Estos valores de α representan la información a priori, es decir, la probabilidad a priori de los parámetros.

En principio, los valores de las α's se podrían ajustar para reflejar las estimaciones subjetivas aportadas por los expertos. Sin embargo, en la práctica nunca (o casi nunca) hay expertos capaces de aportar información a priori, por lo que lo más habitual es aplicar la corrección de Laplace (es decir, $\alpha_{ijk} = 1$ en todos los casos) o bien tomar un valor de corrección menor, por ejemplo $\alpha_{ijk} = 0'5$, que es como una "corrección de Laplace suavizada". En la literatura sobre el tema a veces se muestran los resultados experimentales obtenidos con diferentes valores de esta corrección.

3.3.4. Aprendizaje estructural a partir de relaciones de independencia

Los primeros algoritmos de aprendizaje estructural de redes bayesianas que surgieron estaban basados en un análisis de las relaciones de dependencia e independencia presentes en la distribución de probabilidad P: el problema consiste en encontrar un grafo dirigido acíclico (GDA) que sea un mapa de independencias (I-mapa) de P. En realidad, buscamos un I-mapa minimal; es decir, si hay dos grafos que sólo se diferencian en que hay un enlace que aparece en el primero pero no en el segundo y ambos son I-mapas de P preferiremos el segundo, por

cuatro motivos: porque muestra más relaciones de independencia que el primero, porque va a necesitar menos espacio de almacenamiento (alguna de las tablas será más pequeña), porque va a ser más preciso (al necesitar menos parámetros podrá estimarlos con mayor fiabilidad, reduciendo además el riesgo de sobreajuste) y porque conducirá a una computación más eficiente. Una vez obtenido el grafo, ya se puede realizar el aprendizaje paramétrico para hallar las probabilidades condicionales y completar así la red bayesiana.

Obtención de las relaciones de dependencia e independencia En la práctica nunca conocemos P (la distribución de probabilidad del mundo real), sino un conjunto de casos obtenidos del mundo real y recogidos en una base de datos. Por eso, el primer problema que debe afrontar este método es cómo obtener las relaciones de dependencia e independencia de P a partir de la base de datos. A primera vista podríamos pensar que cuando dos variables están correlacionadas en la base de datos es porque están correlacionadas en P. Sin embargo, también es posible que se trate de una correlación espuria, es decir, accidental. Para ello se suelen aplicar los tests de independencia de la estadística clásica, es decir, se realiza un *contraste de hipótesis* para discernir estas dos posibilidades:

- **Hipótesis experimental**, H_E: Las variables están correlacionadas.
- **Hipótesis nula**, H_0: Las variables son independientes, es decir, la correlación que se observa en la base de datos se debe al azar.

Luego se aplica un test χ^2 para determinar la probabilidad de que siendo cierta la hipótesis nula, H_0, se produzca por azar una correlación tan grande como la observada entre las variables (u otra mayor). Esta probabilidad se denomina p. Si p es inferior a cierto umbral α, conocido como *nivel de significancia*, se rechaza la hipótesis nula, es decir, se concluye que realmente las variables están correlacionadas. En el problema que nos ocupa, eso lleva a incluir una relación de dependencia como dato de entrada para el aprendizaje estructural. En cambio, si $p \geq \alpha$, se incluye una relación de independencia.

La dificultad de esta búsqueda de dependencias e independencias es que el número de relaciones posibles crece super-exponencialmente con el número de variables, pues hay que examinar cada relación del tipo $I_P(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$, con la única condición de que los tres subconjuntos sean disjuntos y $\mathbf{X}$ e $\mathbf{Y}$ sean no vacíos. Para cada relación hay que realizar tantos tests de independencia como configuraciones existen para $\mathbf{X}$, $\mathbf{Y}$ y $\mathbf{Z}$; al menos, hay que hacer todos los tests necesarios hasta encontrar una combinación de configuraciones en que el test determine que hay correlación entre $\mathbf{X}$ e $\mathbf{Y}$ dado $\mathbf{Z}$, lo cual nos llevaría a incluir $\neg I_P(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ en la lista de relaciones.

Por este motivo, el aprendizaje basado en relaciones sólo puede utilizarse para problemas en que el número de variables es muy reducido.

Construcción del grafo Una vez obtenida la lista de relaciones, hay que construir el grafo. El algoritmo más conocido es el denominado PC, de Spirtes et al. [75, 76]. El algoritmo consta de dos fases. En la primera, toma como punto de partida un grafo completo no dirigido y va eliminando enlaces basándose en las relaciones de independencia. La segunda fase consiste en asignar una orientación a los enlaces del grafo no dirigido obtenido en la primera fase. Los detalles de este algoritmo, así como su justificación teórica, pueden encontrarse en las referencias citadas y en el libro de Neapolitan [58, cap. 10], que estudia además otros algoritmos de aprendizaje estructural similares.

3.3.5. Aprendizaje estructural mediante búsqueda heurística

Un método alternativo de aprendizaje estructural consiste en utilizar una métrica para determinar cuál es el mejor modelo. La primera dificultad que encontramos es que el número de modelos es infinito. Por ello descomponemos el problema en dos partes: buscar el mejor grafo y buscar los mejores parámetros para cada grafo posible. La segunda parte es el llamado aprendizaje paramétrico, que como ya hemos visto, puede resolverse en un tiempo razonable introduciendo algunas hipótesis. Por tanto, el problema se reduce a examinar todos los grafos posibles —su número es finito— y realizar un aprendizaje paramétrico para cada uno de ellos. Sin embargo, esta propuesta sólo es válida para problemas con muy pocas variables, porque el número de grafos posibles crece de forma super-exponencial con el número de variables. La solución es aplicar el concepto de *búsqueda heurística* desarrollado en el campo de la inteligencia artificial: dado que no es posible examinar todas las posibles soluciones, sino sólo una parte muy pequeña de ellas, vamos a realizar un proceso de búsqueda que consiste en generar unas pocas posibles soluciones, seleccionar la mejor (o las mejores), y a partir de ella(s) generar otras nuevas, hasta encontrar una que satisfaga ciertos criterios. Este proceso de *búsqueda en calidad*[7] necesita ser guiado por una *métrica* (en inglés, *score*) que indique la calidad de cada posible solución.

Por tanto, cada algoritmo de aprendizaje de este tipo se caracteriza por dos elementos: la métrica y el algoritmo de búsqueda. En este texto sólo vamos a dar una breve introducción al tema. Para profundizar en el tema, conviene estudiar el capítulo 11 del libro [7], donde se dan muchos más detalles, así como la bibliografía recomendada al final de este capítulo.

Métricas de calidad

Las métricas más utilizadas para determinar cuál es la mejor red bayesiana dados ciertos datos son de tres tipos principales:

Métricas bayesianas En estas métricas la calidad de una red se identifica con su probabilidad a posteriori, dada por la ecuación (3.22). La probabilidad a priori de una red es el resultado de la probabilidad de su grafo y de la probabilidad de los parámetros dado el grafo:

$$P(\text{red}) = P(\text{grafo}) \cdot P(\text{parámetros} \mid \text{grafo}) \tag{3.34}$$

Las métricas bayesianas se diferencian unas de otras en la forma de asignar estas dos distribuciones de probabilidad, $P(\text{grafo})$ y $P(\text{parámetros} \mid \text{grafo})$. Esta última es la probabilidad a priori de los parámetros dado el grafo. No vamos a hablar ahora de esta cuestión porque ya la hemos discutido en la sec. 3.3.3 (aprendizaje paramétrico bayesiano).

En cuanto a $P(\text{grafo})$, hay varias posibilidades. Por ejemplo, la métrica K2 [13], que es una de las más conocidas, supone que todos los grafos tienen la misma probabilidad a priori. Otra posibilidad sería dar mayor probabilidad a las redes que tienen menos enlaces y menos padres en cada familia, con el fin de evitar el sobreajuste.[8]

Por último, hay que especificar la probabilidad $P(\text{datos} \mid \text{red})$, para lo cual también es necesario introducir hipótesis adicionales. En particular, es habitual suponer que la base de

[7] En el campo de la inteligencia artificial se distinguen tres tipos principales de búsqueda: búsqueda en profundidad (en inglés, *depth-first search*), búsqueda en anchura (*breadth-first search*) y búsqueda en calidad (*best-first search*).

[8] En el algoritmo K2, que utiliza la métrica K2, el sobreajuste se evita de otra forma: limitando el número de padres que puede tener cada nodo.

datos está completa —es decir, no hay datos ausentes (cf. sec. 3.1.1)— y que los casos de la base de datos son condicionalmente independientes dado el modelo.

En resumen, la métrica resultante, que, como hemos dicho, identifica la calidad de la red con su probabilidad a posteriori, viene dada por

$$P(\text{red} \mid \text{datos}) \propto P(\text{red}) \cdot P(\text{datos} \mid \text{red}) \tag{3.35}$$

$$= P(\text{grafo}) \cdot P(\text{parámetros} \mid \text{grafo}) \cdot P(\text{datos} \mid \text{grafo, parámetros}) \tag{3.36}$$

En las secciones 11.3 a 11.5 de [7] están explicadas la métrica K2 (en la sec. 11.4.3, con el nombre de "medida de Cooper-Herskovits") y otras medidas bayesianas.

Métrica de longitud mínima de descripción Estas métricas se basan en la posibilidad de representar el modelo y los datos en forma codificada. La codificación del modelo ocupa más espacio cuanto más complejo sea el modelo. La representación de los datos será más breve cuanto más cerca esté el modelo de los datos. El concepto de mínima longitud de descripción (LMD; en inglés, *minimum description length*, MDL) se refiere a la descripción más breve posible, es decir, utilizando la mejor codificación.

Al identificar la calidad de un modelo con su LMD cambiada de signo se atienden dos objetivos: por un lado, se asigna mayor calidad a los modelos más simples, lo cual es un antídoto contra el sobreajuste; por otro, se valoran más los modelos que más se ajustan a los datos. Buscar las redes bayesianas de mayor calidad es lo mismo que buscar las de menor LMD.

Una descripción más detallada de este método se encuentra en [7, sec. 11.6].

Medidas de información Otra forma de medir el ajuste entre la red y los datos consiste en calcular la *información mutua*, cuyo valor numérico se determina a partir de la teoría de la información. El problema de identificar la calidad de la red con la información mutua es que los modelos más complejos permiten alcanzar valores más altos en esta métrica, lo cual lleva al sobreajuste. Para evitarlo, se añade a la métrica un término adicional, denominado *penalización*, que resta un valor más alto cuanto mayor es la complejidad del modelo; véase [7, sec. 11.7].

Algoritmos de búsqueda

Como hemos dicho ya, en la práctica es imposible examinar todos los grafos posibles y realizar un aprendizaje paramétrico para cada uno de ellos. Por eso se utilizan técnicas de búsqueda heurística.

El primer algoritmo de búsqueda propuesto en la literatura fue K2 [13]. El punto de partida es un grafo vacío. En cada iteración, el algoritmo considera todos los enlaces posibles, es decir, todos aquellos que no forman un ciclo, y añade aquél que conduce a la red bayesiana de mayor calidad.[9] El algoritmo termina cuando no se pueden añadir más enlaces (el número de padres por nodo está acotado, con el fin de evitar el sobreajuste) o cuando no es posible aumentar la calidad de la red añadiendo un enlace.

[9] En el artículo original de Cooper and Herskovits [13], la métrica utilizada era K2. Por eso la métrica y el algoritmo de búsqueda llevan el mismo nombre. Naturalmente, es posible utilizar el algoritmo de búsqueda K2 con cualquier otra métrica y, recíprocamente, utilizar la métrica K2 con cualquier otro algoritmo de búsqueda.

El principal problema de este método es que se trata de un *algoritmo voraz*, por lo que una vez añadido un enlace nunca lo borra. Eso hace que la probabilidad de quedar atrapado en un máximo local sea muy alta.

Una forma de reducir (no de eliminar) este problema consiste en dar mayor flexibililidad al algoritmo, permitiendo que en cada paso se realice una de estas operaciones:

1. Añadir un enlace (siempre que no cree un ciclo).

2. Borrar un enlace.

3. Invertir un enlace (siempre que no se cree un ciclo).

El precio que se paga para reducir la probabilidad de máximos locales es una mayor complejidad computacional, pues el número de grafos que hay que examinar es mucho mayor.

Concluimos esta sección señalando que hay métodos híbridos de aprendizaje que combinan la búsqueda heurística con la detección de relaciones de dependencia e independencia.

3.3.6. Otras cuestiones

Para terminar, vamos a mencionar cuatro cuestiones de gran importancia, pero que exceden el nivel de las asignaturas para las que se han redactado estos apuntes.

1. **Datos incompletos.** Los métodos que hemos descrito en este capítulo suponen que la base de datos es completa, es decir, que no tiene valores ausentes (cf. sec. 3.1.1). Sin embargo, en la práctica la mayor parte de las bases de datos no cumplen esta condición. Algunos de los métodos más utilizados para abordar este problema están descritos en [7, sec. 11.10] y en [58, secs. 6.5, 7.1.5 y 8.3].

2. **Aprendizaje de redes causales.** Como vimos en la sección 1.6.1, algunas redes bayesianas admiten una interpretación causal (es decir, cada enlace $X \to Y$ representa un mecanismo causal por el que el valor de X influye sobre el valor de Y, mientras que otras sólo admiten una interpretación probabilista (es decir, como un conjunto de relaciones de independencia condicional). Los métodos de aprendizaje expuestos en este capítulo sólo garantizan la interpretación probabilista. Como introducción a los métodos de aprendizaje que permiten obtener modelos causales, recomendamos el libro de Neapolitan [58, secs. 6.5, 7.1.5 y 8.3], en especial las secciones 1.5, 2.6 y 11.4.2 y todo el capítulo 10. Un tratamiento más extenso se encuentra en los libros de Spirtes et al. [76], Glymour y Cooper [29] y Pearl [65].

3. **Variables ocultas.** Habitualmente, la red bayesiana contiene las mismas variables que la base de datos a partir de la cual ha sido construida. Sin embargo, en algunos casos la distribución de probabilidad asociada a la base de datos puede contener algunas relaciones de dependencia e independencia que no pueden ser modelizadas adecuadamente mediante un GDA. Eso puede hacernos sospechar la presencia de una variable oculta (en inglés, *hidden variable* o *latent variable*) que induce tales relaciones. Algunos métodos para construir redes bayesianas que incluyen variables ocultas pueden verse en [58, sec. 8.5].

4. **Clasificadores basados en redes bayesianas.** Los métodos de aprendizaje descritos en este capítulo tratan de construir una red bayesiana cuya probabilidad conjunta se asemeje lo más posible a la distribución de probabilidad del mundo real. Por eso las métricas de calidad descritas en las secciones anteriores miden, sobre todo, el ajuste entre la red bayesiana y los datos. Sin embargo, en la mayor parte de las aplicaciones prácticas las redes bayesianas se utilizan como clasificadores; por ejemplo, para determinar qué enfermedad padece una persona, para detectar qué avería tiene una máquina, para predecir si un cliente va a devolver el préstamo solicitado, para distinguir el correo interesante del correo basura, etc. Ahora bien, la red que mejor clasifica no es necesariamente la que mejor modeliza la probabilidad. Por ello, encontrar el mejor clasificador basado en una red bayesiana es un problema diferente del aprendizaje en general. El lector interesado en el tema puede encontrar abundantes referencias en Internet introduciendo el término "Bayesian network classifiers".

Bibliografía recomendada

El mejor libro que conocemos sobre construcción de redes bayesianas con ayuda de expertos humanos es el de Ley Borrás [52]; aunque el autor se centra en los diagramas de influencia, todo lo que dice puede aplicarse a las redes bayesianas, en particular, el capítulo dedicado a la obtención de las probabilidades. En cuanto a los modelos canónicos, puede encontrar una descripción detallada, con indicaciones sobre cómo aplicarlos en la práctica, en [23].

En cuanto al aprendizaje automático de redes bayesianas, recomendamos encarecidamente el libro de Richard Neapolitan, *Learning Bayesian Networks* [58]. Es un libro muy completo (el único tema importante que no trata es el de los clasificadores bayesianos) y muy bien escrito, al alcance de todo alumno de posgrado, pues —a diferencia de la mayor parte de los artículos que se encuentran en la literatura— no da nada por supuesto. También es muy recomendable el artículo de Heckerman [32]. Por otro lado, todos los libros generales sobre redes bayesianas citados en la bibliografía del capítulo 1 (página 36) dedican uno o varios capítulos a este tema.

Actividades

1. Realice los ejercicios propuestos en la sección 3.2 de este capítulo.

2. Realice los ejercicios que aparecen al final del capítulo 11 del libro de Castillo et al. [7].

3. Utilizando el programa OpenMarkov, realice los ejercicios de aprendizaje interactivo de redes bayesianas a partir de bases de datos que se indican en el tutorial, que está disponible en `www.openmarkov.org/learning`. Estos ejercicios le ayudarán a ver cómo actúan, paso a paso, los algoritmos de aprendizaje estudiados en las secciones 3.3.4 y 3.3.5.

Capítulo 4

Análisis de decisiones

Resumen

En este capítulo vamos a estudiar principalmente dos modelos de análisis de decisiones: los árboles de decisión (AD) y los diagramas de influencia (DI). Los DIs pueden entenderse como una extensión de las redes bayesianas, pues éstas sólo tienen nodos de azar, mientras que aquéllos tienen también nodos de decisión y de utilidad.

Empezaremos por plantear los fundamentos de la teoría probabilista de la decisión (sec. 4.1), para estudiar después los dos modelos mencionados (sec. 4.2). Como veremos, la evaluación de un AD o un DI consiste en determinar la utilidad esperada y la estrategia de actuación óptima, es decir, la mejor política para cada una de las decisiones. Una forma de evaluar un DI consiste en desarrollar un AD equivalente y evaluarlo (cf. sec. 4.3.1). Además de este método, que es el primero que se propuso en la literatura, vamos a estudiar dos algoritmos más eficientes: la eliminación de variables (sec. 4.3.2) y la inversión de arcos (sec. 4.3.3); ambos son muy similares a los métodos del mismo nombre para redes bayesianas estudiados en el capítulo 2. En la sección 4.4 estudiaremos la construcción de diagramas de influencia para resolver problemas del mundo real y en la 4.5 el análisis de sensibilidad.

Contexto

Este capítulo se basa en los tres anteriores, en particular en la teoría de la probabilidad, en la factorización de la probabilidad de acuerdo con un grafo dirigido acíclico (GDA), en los algoritmos de inferencia para redes bayesianas y en la construcción manual de redes bayesianas.

Los diagramas de influencia están íntimamente relacionados con los modelos de decisión de Markov (en inglés se denominan *Markov decision processes*), que son muy utilizados en varios campos de la inteligencia artificial, principalmente en el aprendizaje con refuerzo y en planificación, especialmente en el campo de la robótica. Por tanto, este capítulo enlaza con otras asignaturas del Máster en Inteligencia Artificial Avanzada, tales como *Métodos de aprendizaje en IA*, *Robótica perceptual y autónoma* y otras.

Objetivos

El objetivo de este capítulo es que el alumno conozca los fundamentos de la teoría probabilista de la decisión, los principales modelos de representación del conocimiento para análisis de decisiones y los algoritmos para evaluarlos. También debe ser capaz de construir modelos para problemas reales utilizando algún programa de ordenador, como Elvira u OpenMarkov. Por último, debe conocer la existencia de otros modelos que no vamos a estudiar con detalle en esta asignatura, como los modelos de decisión de Markov ya mencionados, que son muy utilizados en distintos dominios; el objetivo es que, si algún día los necesita en su práctica profesional, pueda acudir a las referencias bibliográficas que damos al final de este capítulo y, con conocimientos adquiridos en esta asignatura, pueda aprender por sí mismo a aplicarlos según sus necesidades.

Requisitos previos

Es necesario haber estudiado los tres capítulos anteriores.

Contenido

4.1. Fundamentos de la teoría de la decisión

Estudiar la sección 1 de [22], especialmente los conceptos de valor esperado y utilidad esperada y la distinción entre ambos. Puede ser muy útil ver el vídeo docente 4.1, titulado "Introducción a la teoría de la decisión", que se encuentra en `http://www.ia.uned.es/~fjdiez/docencia/videos-prob-dec`.

4.2. Diagramas de influencia y árboles de decisión

Estudiar la sección 2 de [22] y ver los vídeos 4.2 y 4.3. Así se entenderán mejor las definiciones que vamos a dar a continuación.

4.2.1. Definición de diagrama de influencia

Información cualitativa: el grafo del DI y su significado

Un DI contiene tres clases de nodos: *nodos de azar* $\mathbf{V}_C$, *nodos de decisión* $\mathbf{V}_D$, y *nodos de utilidad* $\mathbf{V}_U$—véase la fig. 4.1. Los nodos de azar representan eventos que no pueden ser controlados por el decisor. Los nodos de decisión corresponden a acciones que el decisor puede controlar. Los nodos de utilidad representan las preferencias del decisor. Los nodos de utilidad no pueden ser padres de nodos de azar o de decisión.

Hay dos clases de nodos de utilidad: *nodos de utilidad ordinarios*, cuyos padres son nodos de decisión y/o de utilidad (tales como U_1 y U_2 en la fig. 4.1), y *nodos super-valor*, cuyos padres son nodos de utilidad (por ejemplo, el nodo U_0 de la fig. 4.1). Suponemos que en cada DI hay un nodo de utilidad que es el único nodo de utilidad o un descendiente de todos los demás nodos de utilidad y, por tanto, no tiene hijos; a este nodo lo llamamos U_0.

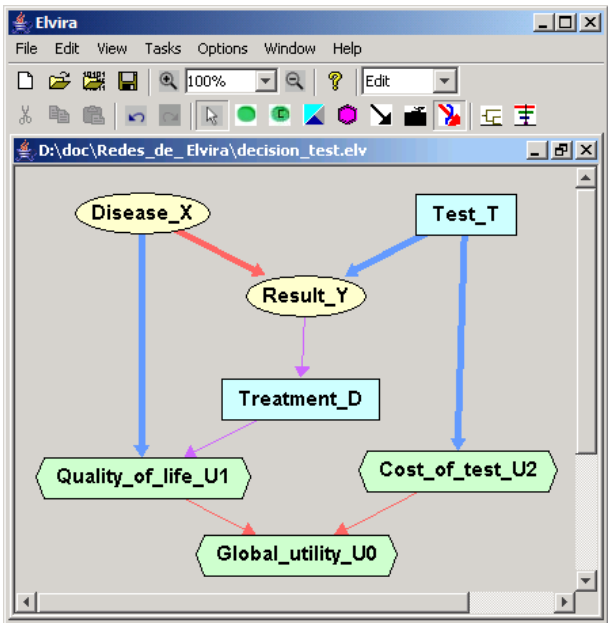


Figura 4.1: DI con dos nodos de decisión (rectángulos), dos nodos de azar (óvalos) y tres nodos de utilidad (hexágonos). Observe que hay un camino dirigido, $T \to Y \to D \to U_1 \to U_0$, que incluye todas las decisiones y el nodo de utilidad global, U_0.

Hay tres clases de enlaces en un DI, dependiento del tipo de nodo al que apuntan. Los enlaces que apuntan a nodos de azar indican dependencia probabilista, como en el caso de las redes bayesianas. Los enlaces que apuntan a nodos de decisión indican disponibilidad de información; por ejemplo, el enlace $Y \to D$ significa que el estado de Y es conocido al tomar la decisión D. Los enlaces que apuntan a nodos de utilidad indican dependencia funcional: para nodos de utilidad ordinarios, representan el dominio de la función de utilidad asociada; para un nodo super-valor, indican que la utilidad asociada es función (generalmente la suma o el producto) de las funciones de utilidad de sus padres.

Debe haber, además, un camino dirigido que incluye todos los nodos de decisión e indica el orden en que se toman las decisiones. Esto induce una partición del conjunto de variables de azar, $\mathbf{V}_C$, tal que en un DI que tenga n decisiones $\{D_0, \ldots, D_{n-1}\}$, la partición contiene $n+1$ subconjuntos $\{\mathbf{C}_0, \mathbf{C}_1, ..., \mathbf{C}_n\}$, donde $\mathbf{C}_i$ es el conjunto de variables del tipo C tales que existe un enlace $C \to D_i$ pero no existe ningún enlace $C \to D_j$ con $j < i$; es decir, $\mathbf{C}_i$ representa el conjunto de variables de azar conocidas para D_i y desconocidas para las decisiones previas. $\mathbf{C}_n$ es el conjunto de variales de las cuales no parte ningún enlace a ningún nodo de decisión, es decir, las variables cuyo valor nunca se conoce directamente. En el ejemplo anterior (fig. 4.1), $D_0 = T$, $D_1 = D$, $\mathbf{C}_0 = \varnothing$, $\mathbf{C}_1 = \{Y\}$, y $\mathbf{C}_2 = \{X\}$.

Las variables cuyo valor es conocido por el decisor al tomar la decisión D_i se denominan *predecesores informativos* de D_i y se denotan por $PredInf(D_i)$. Introducimos la *hipótesis de no-olvido* (en inglés, *no-forgetting hypothesis*), que dice que el decisor recuerda todas las observaciones previas y todas las decisiones que ha tomado. Al introducir esta hipótesis,

tenemos que

$$\begin{aligned} \mathit{PredInf}(D_i) &= \mathit{PredInf}(D_{i-1}) \cup \{D_{i-1}\} \cup \mathbf{C}_i & (4.1)\\ &= \mathbf{C}_0 \cup \{D_0\} \cup \mathbf{C}_1 \cup \ldots \cup \{D_{i-1}\} \cup \mathbf{C}_i \ . & (4.2) \end{aligned}$$

Supongamos que existe un nodo de azar X, dos decisiones D_1 y D_2, un enlace $X \to D_1$ y un camino dirigido desde D_1 hasta D_2. En este caso X es un predecesor informativo de D_1 y, por la hipótesis de no-olvido, también lo es de D_2, tanto si el grafo contiene el enlace $X \to D_2$ como si no. A este enlace se le denomina *arco de recuerdo* (en inglés, *non-forgetting link*) porque sirve para indicar explícitamente que el decisor recuerda el valor de X cuando va a tomar la decisión D_2.

Existe otro tipo de arco de recuerdo. Supongamos que un DI contiene dos decisiones, D_1 y D_2, un enlace $D_1 \to D_2$ y además otro camino dirigido desde D_1 hasta D_2. En este caso, aunque ese enlace no estuviera en la red, D_1 seguiría siendo una decisión anterior a D_2 y, por la hipótesis de no-olvido, un predecesor informativo de D_2. El enlace $D_1 \to D_2$ se denomina *de recuerdo* porque indica que al tomar la segunda decisión el decisor recuerda qué opción escogió para la primera.

En ambos casos los arcos de recuerdo son redundantes, es decir, da lo mismo que estén o no, porque no modifican la semántica del DI.

Información cuantitativa: probabilidades y utilidades

La información cuantitativa que define un DI se da al asignar a cada nodo de azar C una distribución de probabilidad $P(c|pa(C))$ para cada configuración de sus padres (como en las redes bayesianas), asignando a cada nodo de utilidad ordinario U una función $\psi_U(pa(U))$ que asocia a cada configuración de sus padres un número real, y asignando a cada nodo super-valor una función de combinación de utilidades (generalmente la suma o la multiplicación). El dominio de cada función U viene dado por sus *predecesores funcionales*, $\mathit{PredFunc}(U)$. Para un nodo de utilidad ordinario, $\mathit{PredFunc}(U) = Pa(U)$, y para un nodo super-valor $\mathit{PredFunc}(U) = \bigcup_{U' \in Pa(U)} \mathit{PredFunc}(U')$. En el ejemplo anterior, $\mathit{PredFunc}(U_1) = \{X, D\}$, $\mathit{PredFunc}(U_2) = \{T\}$, y $\mathit{PredFunc}(U_0) = \{X, D, T\}$. Para simplificar la notación supondremos que $\mathit{PredFunc}(U_0) = \mathbf{V}_C \cup \mathbf{V}_D$, lo cual incluye la posibilidad de que U_0 en realidad dependa sólo de un subconjunto de $\mathbf{V}_C \cup \mathbf{V}_D$.

Para cada configuración $\mathbf{v}_D$ de las variables de decisión $\mathbf{V}_D$ tenemos una distribución de probabilidad sobre las variables de azar $\mathbf{V}_C$:

$$P(\mathbf{v}_C : \mathbf{v}_D) = \prod_{C \in \mathbf{V}_C} P(c|pa(C)) \tag{4.3}$$

que representa la probabilidad de que la configuración $\mathbf{v}_C$ ocurra en el mundo real si el decisor escogiera siempre (ciegamente) los valores indicados por $\mathbf{v}_D$.

Ejemplo 4.1 Para el DI de la figura 4.1, tenemos que[1]

$$P(x, y : t, d) = P(x) \cdot P(y|x : t)$$

[1]En este capítulo utilizarmos el símbolo ":" para indicar que las decisiones actúan como variables condicionantes pero sin formar parte de una distribución de probabilidad conjunta. En este ejemplo, como no hay una probabilidad conjunta $P(x, y, t, d)$ de la cual pueda derivarse $P(x, y|t, d)$ según la ecuación (1.6), hemos escrito en su lugar $P(x, y : t, d)$. Por el mismo motivo hemos escrito $P(y|x : t)$ en vez de $P(y|x, t)$.

En este caso las probabilidades de las variables de azar no dependen de la decisión D porque X e Y ocurren antes que D en el tiempo.

4.2.2. Políticas y utilidades esperadas

Una *política estocástica* para una decisión D es una distribución de probabilidad definida sobre D y condicionada por el conjunto de sus predecesores informativos, $P_D(d|PredInf(D))$. Si P_D es degenerada, es decir, si consta solamente de ceros y unos, decimos que es una *política determinista*. Una política determinista puede entenderse como una función π_D que asigna a cada configuración de $PredInf(D)$ un valor de D: $\pi_D(predInf(D)) = d$.

Una *estrategia* Δ para un DI consta de un conjunto de políticas, una para cada decisión, $\{P_D|D \in \mathbf{V}_D\}$. Una estrategia Δ induce una distribución conjunta sobre $\mathbf{V}_C \cup \mathbf{V}_D$ definida por

$$\begin{aligned} P_\Delta(\mathbf{v}_C, \mathbf{v}_D) &= P(\mathbf{v}_C : \mathbf{v}_D) \prod_{D \in \mathbf{V}_D} P_D(d|PredInf(D)) \\ &= \prod_{C \in \mathbf{V}_C} P(c|pa(C)) \prod_{D \in \mathbf{V}_D} P_D(d|pa(D)) \end{aligned} \tag{4.4}$$

La *utilidad esperada de una estrategia* Δ se define así

$$UE(\Delta) = \sum_{\mathbf{v}_C} \sum_{\mathbf{v}_D} P_\Delta(\mathbf{v}_C, \mathbf{v}_D)\psi(\mathbf{v}_C, \mathbf{v}_D) \tag{4.5}$$

donde ψ es la utilidad asociada al nodo U_0, es decir, la utilidad global. La *estrategia óptima* es aquella que maximiza la utilidad esperada:

$$\Delta_{opt} = \arg\max_{\Delta \in \Delta^*} UE(\Delta) \tag{4.6}$$

La *utilidad esperada de un DI*, a veces llamada *máxima utilidad esperada*, es la que se obtiene al aplicar la estrategia óptima:

$$UE = UE(\Delta_{opt}) = \max_{\Delta \in \Delta^*} UE(\Delta) \tag{4.7}$$

4.3. Evaluación de diagramas de influencia

La *evaluación* de un DI consiste en encontrar la UE y la estrategia óptima. Una forma de hacerlo sería mediante la aplicación directa de la ecuación (4.7), lo cual exigiría evaluar todas las estrategias posibles. En la práctica este método es inviable, porque el número de estrategias es enorme.

Afortunadamente, se puede demostrar que

$$UE = \sum_{\mathbf{c}_0} \max_{d_0} \ldots \sum_{\mathbf{c}_{n-1}} \max_{d_{n-1}} \sum_{\mathbf{c}_n} P(\mathbf{v}_C : \mathbf{v}_D)\psi(\mathbf{v}_C, \mathbf{v}_D) \tag{4.8}$$

No incluimos la demostración en este texto porque es bastante complicada, más por la notación que por los conceptos que intervienen.

Ejemplo 4.2 La UE del DI de la figura 4.1 es

$$UE = \max_t \sum_y \max_d \sum_x P(x) \cdot P(y|x:t) \cdot \underbrace{(U_1(x,d) + U_2(t))}_{U_0(x,d,t)} \tag{4.9}$$

Ejercicio 4.3 Indicar qué forma toma la ecuación (4.8) para cada uno de los ejemplos que aparecen en [22].

Vamos a estudiar a continuación tres métodos de evaluación de diagramas de influencia. En cierto modo podemos decir que son tres formas distintas de aplicar la ecuación (4.8). El primero consiste en construir un árbol de decisión equivalente y evaluarlo. Los otros dos son una adaptación de los métodos de eliminación de variables y de inversión de arcos estudiados para redes bayesianas.

Observe que en la ecuación (4.8) cada sumatorio se aplica a un conjunto de variables, $\mathbf{C}_i$. Este sumatorio puede descomponerse en tantos sumatorios como variables haya en el conjunto $\mathbf{C}_i$. De este modo tenemos un operador de maximización por cada variable de decisión y un operador de marginalización (sumatorio) por cada variable de azar. Los operadores del mismo tipo conmutan entre sí, pero no con los del otro tipo. Es decir, si tenemos que aplicar consecutivamente dos sumatorios, podemos cambiar el orden entre ellos sin que afecte al resultado; sin embargo, no podemos intercambiar el orden de un operador de maximización y uno de marginalización, porque eso afectaría al resultado: compare el ejemplo 1 de [22], en que se calcula $\max_d \sum_x U(x,d)$, con el ejemplo 2, en que se calcula $\sum_x \max_d U(x,d)$, y vea que las utilidades esperadas son diferentes.

Esta observación es importante, porque en cada uno de los métodos que vamos a exponer a continuación, dentro de cada subconjunto $\mathbf{C}_i$ podemos ordenar las variables como queramos, pero siempre hay que respetar el orden $\mathbf{C}_0, D_0, \mathbf{C}_1, D_1, \ldots, \mathbf{C}_{n-1}, D_{n-1}, \mathbf{C}_n$. Ahora bien, aunque el orden de las variables dentro de cada $\mathbf{C}_i$ no va a afectar a la utilidad esperada ni a la política óptima resultantes, sí puede afectar al coste computacional, en términos de tiempo y memoria requeridos. Para cada uno de los métodos, buscar el orden óptimo de las variables es un problema NP. Para cada método existen reglas heurísticas y otras técnicas que ayudan a encontrar un orden casi óptimo, pero no las vamos a estudiar en este libro.

4.3.1. Expansión y evaluación de un árbol de decisión

Expansión del árbol de decisión

La computación de la ecuación (4.8) puede realizarse mediante la expansión y evaluación de un árbol de decisión. Suponemos que las variables que componen cada conjunto $\mathbf{C}_i$ son $\{C_i^1, \ldots, C_i^{m_i}\}$, donde $m_i = card(\mathbf{C}_i)$. Tomamos la variable C_0^1 como nodo raíz y dibujamos tantas ramas como valores tiene esa variable;[2] colocamos un nodo de C_0^2 en cada una de las ramas anteriores, y así sucesivamente hasta agotar las variables de $\mathbf{C}_0$. En cada una de las ramas de $C_0^{m_0}$ (la última variable $\mathbf{C}_0$) ponemos un nodo de D_0 y de cada nodo sacamos una rama por cada valor de esta variable. Continuamos expandiendo el árbol, según el orden dado por

$$\{C_0^1, \ldots, C_0^{m_0}, C_1^1, \ldots, C_{n-1}^{m_{n-1}}, C_n^1, \ldots, C_n^{m_n}\} \tag{4.10}$$

Ahora vamos a calcular la probabilidad de cada rama que parte de un nodo de azar. Denotamos por C_k la variable que ocupa el k-ésimo lugar en la ordenación anterior y definimos

[2] Si el conjunto $\mathbf{C}_0$ estuviera vacío, tomaríamos D_0 como nodo raíz.

$\check{\mathbf{C}}_k$ como el subconjunto de las $k-1$ variables anteriores a C_k, es decir, las variables de azar que aparecen a la izquierda de C_k en el árbol. Definimos también, de forma recursiva, las siguientes distribuciones de probabilidad:

$$P(\check{\mathbf{c}}_k : \mathbf{v}_D) = \sum_{c_k} P(c_k, \check{\mathbf{c}}_k : \mathbf{v}_D) \tag{4.11}$$

Se trata de una definición recursiva porque k toma valores desde m hasta 1. Para $k = m$, tenemos que $\{C_m\} \cup \check{\mathbf{C}}_m = \mathbf{V}_C$ y $P(c_k, \check{\mathbf{c}}_k : \mathbf{v}_D) = P(\mathbf{v}_C : \mathbf{v}_D)$. La probabilidad conjunta $P(\check{\mathbf{c}}_k : \mathbf{v}_D)$, definida sobre $k-1$ variables, se calcula a partir de $P(c_k, \check{\mathbf{c}}_k : \mathbf{v}_D)$, que está definida sobre k variables. A partir de estas probabilidades conjuntas definimos estas probabilidades condicionales:

$$P(c_k|\check{\mathbf{c}}_k : \mathbf{v}_D) = \frac{P(c_k, \check{\mathbf{c}}_k : \mathbf{v}_D)}{P(\check{\mathbf{c}}_k : \mathbf{v}_D)} \tag{4.12}$$

En consecuencia, la probabilidad $P(\mathbf{v}_C : \mathbf{v}_D)$ puede factorizarse de este modo:

$$P(\mathbf{v}_C : \mathbf{v}_D) = \prod_{k=1}^{m} P(c_k|\check{\mathbf{c}}_k : \mathbf{v}_D) \tag{4.13}$$

Observe que esta ecuación es simplemente la regla de la cadena, aplicada para cada configuración $\mathbf{v}_D$.

Recordando cómo hemos definido las C_k, esta ecuación también puede escribirse así:

$$P(\mathbf{v}_C : \mathbf{v}_D) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} P(c_i^j|\check{\mathbf{c}}_i^j : \mathbf{v}_D) \tag{4.14}$$

Ejemplo 4.4 En el ejemplo 4.1 hemos visto que, para el DI de la figura 4.1, $P(x, y : t, d) = P(x) \cdot P(y|x : t)$. A partir de $P(x, y : t, d)$ podemos obtener la probabilidad conjunta

$$P(y : t, d) = \sum_{x} P(x, y : t, d)$$

y la probabilidad condicional

$$P(x|y : t, d) = \frac{P(x, y : t, d)}{P(y : t, d)}$$

Proposición 4.5 La probabilidad $P(c_i^j|\check{\mathbf{c}}_i^j : \mathbf{v}_D)$ no depende de las decisiones que quedan a la derecha del nodo C_i^j en el árbol, es decir, es independiente de los valores que toman las variables $\{D_i, \ldots, D_{n-1}\}$:

$$P(c_i^j|\check{\mathbf{c}}_i^j : \mathbf{v}_D) = P(c_i^j|\check{\mathbf{c}}_i^j : d_0, \ldots, d_{i-1}) \tag{4.15}$$

Ejemplo 4.6 En el ejemplo anterior, la proposición anterior implica que $P(y : t, d) = P(y : t)$. Es fácil comprobarlo, pues $P(x, y : t, d) = P(x) \cdot P(y|x : t)$ no depende de d, y por tanto $P(y : t, d)$, que se calcula a partir de $P(x, y : t, d)$, tampoco.

Demostración. [Esta demostración no es necesaria para entender el resto de la exposición. El lector que lo desee puede omitirla.] Transformamos el DI en una red bayesiana (RB) eliminando los nodos de utilidad, eliminando los arcos de información (es decir, los enlaces cuyo destino es un nodo de decisión) y sustituyendo cada nodo de decisión D por un nodo de azar, que representa la misma variable que en el diagrama de influencia. A este nodo, que no tiene padres, le asignamos arbitrariamente una distribución de probabilidad, $P(d)$. Las variables de la RB son $\mathbf{V}_C \cup \mathbf{V}_D$ y su probabilidad conjunta es

$$P_{\mathrm{RB}}(\mathbf{v}_C, \mathbf{v}_D) = \prod_{C \in \mathbf{V}_C} P(c|pa(C)) \prod_{D \in \mathbf{V}_D} P(d)$$

Como ningún nodo de $\mathbf{V}_D$ es descendiente de ningún nodo de $\mathbf{V}_C$, tenemos que

$$P_{\mathrm{RB}}(\mathbf{v}_D) = \prod_{D \in \mathbf{V}_D} P(d)$$

$$P_{\mathrm{RB}}(\mathbf{v}_C|\mathbf{v}_D) = \prod_{C \in \mathbf{V}_C} P(c|pa(C))$$

y, de acuerdo con la ecuación (4.3),

$$P(\mathbf{v}_C : \mathbf{v}_D) = P_{\mathrm{RB}}(\mathbf{v}_C|\mathbf{v}_D)$$

En particular,

$$P(c_i^j|\check{\mathbf{c}}_i^j : \mathbf{v}_D) = P_{\mathrm{RB}}(c_i^j|\check{\mathbf{c}}_i^j, \mathbf{v}_D) = P_{\mathrm{RB}}(c_i^j|\check{\mathbf{c}}_i^j, d_0, \ldots, d_{n-1})$$

Por otro lado, el valor de C_i^j se conoce antes de tomar la decisión D_i, lo cual implica que existe un camino dirigido desde C_i^j hasta el nodo D_i y hasta cada nodo D_j con $j > i$. Por tanto, C_i^j no puede ser descendiente de $\{D_i, \ldots, D_{n-1}\}$ en el DI, y como el grafo de la RB se formó quitando algunos enlaces del DI, tampoco puede serlo en la RB. Por el mismo razonamiento, como todos los nodos de $\check{\mathbf{C}}_i^j$ pertenecen a $\mathit{PredInf}(D_i)$, $\check{\mathbf{C}}_i^j$ no puede contener ningún nodo que sea descendiente de $\{D_i, \ldots, D_n\}$ en la RB. Además, en la RB ningún nodo D_i tiene padres. Por tanto, por la propiedad de Markov tenemos que

$$P_{\mathrm{RB}}(c_i^j, \check{\mathbf{c}}_i^j, d_0, \ldots, d_{n-1}) = P_{\mathrm{RB}}(c_i^j, \check{\mathbf{c}}_i^j, d_0, \ldots, d_{i-1}) \cdot P_{\mathrm{RB}}(d_i, \ldots, d_{n-1})$$

y

$$P_{\mathrm{RB}}(\check{\mathbf{c}}_i^j, d_0, \ldots, d_{n-1}) = P_{\mathrm{RB}}(\check{\mathbf{c}}_i^j, d_0, \ldots, d_{i-1}) \cdot P_{\mathrm{RB}}(d_i, \ldots, d_{n-1})$$

de donde se deduce que

$$\begin{aligned} P_{\mathrm{RB}}(c_i^j|\check{\mathbf{c}}_i^j, d_0, \ldots, d_{n-1}) &= \frac{P_{\mathrm{RB}}(c_i^j, \check{\mathbf{c}}_i^j, d_0, \ldots, d_{n-1})}{P_{\mathrm{RB}}(\check{\mathbf{c}}_i^j, d_0, \ldots, d_{n-1})} \\ &= \frac{P_{\mathrm{RB}}(c_i^j, \check{\mathbf{c}}_i^j, d_0, \ldots, d_{i-1}) \cdot P_{\mathrm{RB}}(d_i, \ldots, d_{n-1})}{P_{\mathrm{RB}}(\check{\mathbf{c}}_i^j, d_0, \ldots, d_{i-1}) \cdot P_{\mathrm{RB}}(d_i, \ldots, d_{n-1})} \\ &= P(c_i^j|\check{\mathbf{c}}_i^j : d_0, \ldots, d_{i-1}) \end{aligned}$$

que no depende de $\{D_i, \ldots, D_n\}$, con lo cual concluye la demostración. □

Como consecuencia de esta proposición, la probabilidad $P(\mathbf{v}_C : \mathbf{v}_D)$ queda así:

$$P(\mathbf{v}_C : \mathbf{v}_D) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} P(c_i^j | \check{\mathbf{c}}_i^j : d_0, \ldots, d_{i-1}) \tag{4.16}$$

En resumen, hemos construido un árbol de decisión en que la probabilidad de cada rama que parte de un nodo C_i^j puede calcularse a partir de las probabilidades que definen el DI, y esta probabilidad sólo depende de los valores que toman las variables que se encuentran a la izquierda de ese nodo en el árbol.

Evaluación del árbol de decisión

Uniendo las ecuaciones (4.8) y (4.16), tenemos que

$$UE = \sum_{\mathbf{c}_0} \max_{d_0} \ldots \sum_{\mathbf{c}_{n-1}} \max_{d_{n-1}} \sum_{\mathbf{c}_n} \prod_{i=1}^{n} \prod_{j=1}^{m_i} P(c_i^j | \check{\mathbf{c}}_i^j : d_0, \ldots, d_{i-1}) \psi(\mathbf{v}_C, \mathbf{v}_D) \tag{4.17}$$

Vamos a ver que el método de evaluación de árboles de decisión expuesto en [22] es simplemente una forma eficiente de aplicar esta ecuación.

En primer lugar, el potencial $P(c_i^j | \check{\mathbf{c}}_i^j : d_0, \ldots, d_{i-1})$ se puede sacar como factor común de todos los sumatorios sobre $\mathbf{c}_j$ con $j > i$ y todos los operadores de maximización sobre d_j con $j \geq i$, de modo que

$$UE = \sum_{\mathbf{c}_0} \prod_{j=1}^{m_0} P(c_0^j | \check{\mathbf{c}}_0^j) \max_{d_0} \ldots \sum_{\mathbf{c}_{n-1}} \max_{d_{n-1}} \sum_{\mathbf{c}_n} \prod_{j=1}^{m_n} P(c_n^j | \check{\mathbf{c}}_n^j : d_0, \ldots, d_{n-1}) \psi(\mathbf{v}_C, \mathbf{v}_D) \tag{4.18}$$

En segundo lugar, cada sumatorio sobre $\mathbf{C}_i$ puede descomponerse en m_i sumatorios, uno por cada variable C_i^j, y cada potencial $P(c_i^j | \check{\mathbf{c}}_i^j : d_0, \ldots, d_{i-1})$ puede sacarse como factor común a los sumatorios sobre $C_i^{j'}$ en que $j' > j$. Por tanto,

$$\begin{aligned} UE = & \sum_{C_0^1} P(c_0^1) \cdots \sum_{C_0^{m_0}} P(c_0^{m_0} | \check{\mathbf{c}}_0^{m_0}) \max_{d_0} \\ & \sum_{C_1^1} P(c_1^1 | \check{\mathbf{c}}_1^1 : d_0) \cdots \sum_{C_1^{m_1}} P(c_1^{m_1} | \check{\mathbf{c}}_1^{m_1} : d_0) \max_{d_1} \\ & \ldots \\ & \sum_{C_{n-1}^1} P(c_{n-1}^1 | \check{\mathbf{c}}_{n-1}^1 : d_0, \ldots, d_{n-2}) \cdots \sum_{C_{n-1}^{m_{n-1}}} P(c_{n-1}^{m_{n-1}} | \check{\mathbf{c}}_{n-1}^{m_{n-1}} : d_0, \ldots, d_{n-2}) \max_{d_{n-1}} \\ & \sum_{C_n^1} P(c_n^1 | \check{\mathbf{c}}_n^1 : d_0, \ldots, d_{n-1}) \cdots \sum_{C_n^{m_n}} P(c_n^{m_n} | \check{\mathbf{c}}_n^{m_n} : d_0, \ldots, d_{n-1}) \psi(\mathbf{v}_C, \mathbf{v}_D) \end{aligned} \tag{4.19}$$

Es decir, la utilidad asociada a cada nodo de la variable $C_n^{m_n}$ se calcula como la media de las utilidades de sus ramas, ponderadas según la probabilidad de cada rama, $P(c_n^{m_n} | \check{\mathbf{c}}_n^{m_n} : d_0, \ldots, d_{n-1})$. Al evaluar el nodo $C_n^{m_n-1}$, las utilidades que acabamos de obtener se van a ponderar por las probabilidades $P(c_n^{m_n-1} | \check{\mathbf{c}}_n^{m_n-1} : d_0, \ldots, d_{n-1})$, y así sucesivamente para todas las variables de $\mathbf{C}_n$. Luego nos encontramos con la decisión que está más a la derecha, D_{n-1}; su utilidad se calcula maximizando las utilidades de sus ramas. Así continúa la

evaluación del árbol, de izquierda a derecha, según las operaciones indicadas por la ecuación anterior.

De esta forma queda justificado formalmente el método de evaluación de árboles de decisión.

4.3.2. Eliminación de variables

El método de eliminación de variables es el más sencillo de entender. Vamos a explicar dos versiones: sin división de potenciales y con división. En ambos casos partimos de una lista de potenciales, de los cuales m son de probabilidad (uno por cada nodo de azar) y uno de utilidad. Hay dos formas de eliminar una variable: por marginalización (es decir, aplicando un sumatorio) o por maximización. En ambos casos, se sacan de la lista todos los potenciales que dependen de dicha variable.

Eliminación sin división de potenciales

A partir de aquí las dos versiones difieren. La primera de ellas, que no distingue entre los potenciales de probabilidad y de utilidad, multiplica todos los potenciales extraídos, aplica la marginalización o la maximización, según corresponda, y mete en la lista el potencial resultante. Así continúa hasta eliminar todas las variables, con lo cual obtiene un número real, que es la utilidad esperada.

Ejemplo 4.7 Hemos visto en el ejemplo 4.2 que la utilidad asociada al DI de la figura 4.1 es

$$UE = \max_t \sum_y \max_d \sum_x P(x) \cdot P(y|x:t) \cdot U_0(x,d,t)$$

Vamos a eliminar primero la variable X. Sacamos de la lista los potenciales que dependen de X, que en este caso son todos. Al multiplicarlos y marginalizar sobre X, obtenemos un nuevo potencial,

$$\psi(y,t,d) = \sum_x P(x) \cdot P(y|x:t) \cdot U_0(x,d,t)$$

que metemos en la lista. Nuestro problema se ha transformado en el siguiente:

$$UE = \max_t \sum_y \max_d \psi(y,t,d)$$

Eliminamos ahora la variable D. Sacamos de la lista todos los potenciales que dependen de D, que en este caso es uno solo, y maximizamos sobre D, con lo cual obtenemos un nuevo potencial,

$$\psi(y,t) = \max_d \psi(y,t,d)$$

que volvemos a meter en la lista. Al maximizar hemos obtenido la política óptima para D:

$$\pi_D^{opt}(y,t) = \arg\max_d \psi(y,t,d)$$

Ahora el problema es

$$UE = \max_t \sum_y \psi(y,t)$$

Para eliminar Y sacamos de la lista el potencial $\psi(y,t)$ y marginalizamos sobre Y, con lo cual obtenemos el potencial

$$\psi(t) = \sum_y \psi(y,t)$$

El problema se reduce a

$$UE = \max_t \psi(t)$$

Por último, eliminamos T por maximización, con lo cual obtenemos la utilidad esperada y la política óptima para T:

$$\pi_T^{opt}() = \arg\max_t \psi(t)$$

Eliminación con división de potenciales

La eliminación de variables con división de potenciales, en cambio, distingue entre los potenciales de probabilidad y el de utilidad. En cada paso del algoritmo, los potenciales de probabilidad representan la distribución $P(\mathbf{c} : \mathbf{d})$, donde $\mathbf{C}$ y $\mathbf{D}$ son los conjuntos de variables de azar y de decisión, respectivamente, que aún no han sido eliminadas. La utilidad viene dada por $U(\mathbf{c}, \mathbf{d})$. Cuando vamos a eliminar la variable de azar C, el algoritmo factoriza esta probabilidad en dos elementos: $P(c|\check{\mathbf{c}} : \mathbf{d})$ y $P(\check{\mathbf{c}} : \mathbf{d})$, donde $\check{\mathbf{C}} = \mathbf{C}\backslash\{C\}$. Estas probabilidades se pueden calcular así:

$$P(\check{\mathbf{c}} : \mathbf{d}) = \sum_c P(c, \check{\mathbf{c}} : \mathbf{d}) \tag{4.20}$$

$$P(c|\check{\mathbf{c}} : \mathbf{d}) = \frac{P(c, \check{\mathbf{c}} : \mathbf{d})}{P(\check{\mathbf{c}} : \mathbf{d})} \tag{4.21}$$

Esto nos permite calcular una nueva utilidad,

$$U(\check{\mathbf{c}}, \mathbf{d}) = \sum_c P(c|\check{\mathbf{c}} : \mathbf{d}) \cdot U(c, \check{\mathbf{c}}, \mathbf{d}) \tag{4.22}$$

Es fácil comprobar que se cumple la siguiente igualdad:

$$\underbrace{\sum_c P(c, \check{\mathbf{c}} : \mathbf{d}) \cdot U(c, \check{\mathbf{c}}, \mathbf{d})}_{\psi(\check{\mathbf{c}}, \mathbf{d})} = P(\check{\mathbf{c}} : \mathbf{d}) \underbrace{\sum_c P(c|\check{\mathbf{c}} : \mathbf{d}) \cdot U(c, \check{\mathbf{c}}, \mathbf{d})}_{U(\check{\mathbf{c}}, \mathbf{d})} \tag{4.23}$$

El miembro de la izquierda corresponde a la eliminación de variables sin divisiones, que multiplica los potenciales de probabilidad y de utilidad directamente, y así obtiene un potencial ψ que no distingue entre probabilidad y utilidad. En cambio, el miembro de la derecha corresponde a la versión con divisiones, que primero obtiene la probabilidad condicional $P(c|\check{\mathbf{c}} : \mathbf{d})$, para poder calcular la utilidad $U(\check{\mathbf{c}}, \mathbf{d})$, que se va a conservar aparte de la probabilidad $P(\check{\mathbf{c}} : \mathbf{d})$.

La pregunta que se hará el lector es: "Si el resultado es el mismo (es decir, ambas versiones van a dar la misma utilidad esperada y las mismas políticas), ¿para qué realizar una división de potenciales que es innecesaria y tiene un coste computacional no despreciable?" La razón es que a la hora de **explicar** los resultados al usuario, conviene conservar la distinción entre la probabilidad y la utilidad, pues esto nos permitirá mostrar al usuario cuál es la utilidad esperada para cada una de las opciones de cierta decisión.

La conclusión es inmediata: si sólo queremos calcular la utilidad esperada y/o la estrategia óptima, es mejor aplicar la versión sin divisiones, porque es más eficiente. En cambio, si queremos conocer la utilidad asociada a cada escenario y cada posible elección del decisor, debemos aplicar la versión con divisiones, aunque su coste computacional sea mayor.

En realidad, ninguna de las dos versiones trabaja explícitamente con $P(c, \check{\mathbf{c}} : \mathbf{d})$, sino con una factorización de la misma y, como hemos visto varias veces en este libro, al realizar los cálculos conviene conservar la factorización tanto como sea posible. Por eso, al aplicar la ecuación (4.20) lo que se hace es sacar de la lista sólo los potenciales de probabilidad que dependen de C. Primero los multiplicamos, con lo cual obtenemos un nuevo potencial, $\psi_1(c, \check{\mathbf{c}} : \mathbf{d})$. Sea $\psi_2(\check{\mathbf{c}} : \mathbf{d})$ el producto de los potenciales de probabilidad que no dependen de C; naturalmente, en vez de calcular ψ_2 explícitamente, conservamos sus factores en la lista. Tenemos por tanto que

$$P(\check{\mathbf{c}} : \mathbf{d}) = \psi_2(\check{\mathbf{c}} : \mathbf{d}) \sum_c \psi_1(c, \check{\mathbf{c}} : \mathbf{d}) = \psi_2(\check{\mathbf{c}} : \mathbf{d}) \cdot \psi_1(\check{\mathbf{c}} : \mathbf{d}) \tag{4.24}$$

$$P(c|\check{\mathbf{c}} : \mathbf{d}) = \frac{\psi_1(c, \check{\mathbf{c}} : \mathbf{d}) \cdot \psi_2(\check{\mathbf{c}} : \mathbf{d})}{\psi_1(\check{\mathbf{c}} : \mathbf{d}) \cdot \psi_2(\check{\mathbf{c}} : \mathbf{d})} = \frac{\psi_1(c, \check{\mathbf{c}} : \mathbf{d})}{\psi_1(\check{\mathbf{c}} : \mathbf{d})} \tag{4.25}$$

El algoritmo para eliminar una variable de azar C queda así:

1. Sacamos de la lista de potenciales de probabilidad todos los que dependen de C y los multiplicamos, para obtener $\psi_1(c, \check{\mathbf{c}} : \mathbf{d})$.
2. Calculamos $\psi_1(\check{\mathbf{c}} : \mathbf{d})$ y lo añadimos a la lista.
3. Calculamos $P(c|\check{\mathbf{c}} : \mathbf{d})$ y lo utilizamos para calcular $U(\check{\mathbf{c}}, \mathbf{d})$, según la ecuación (4.22). Ésta va a ser la nueva utilidad.

La eliminación de una variable de decisión D es más sencilla. En principio, es posible que D aparezca en los potenciales de probabilidad, pero en ese caso se trata de una variable redundante, es decir, la probabilidad no depende de la variable de decisión que vamos a eliminar. Por ejemplo, al eliminar la variable D_2 podríamos encontrar que los potenciales de probabilidad que —al menos aparentemente— dependen de ella son $\psi_1(x, y, d_1, d_2)$ y $\psi_2(y, z, d_2)$; se puede demostrar que su producto no depende de D_2, es decir, para todo par de valores d_2^i y d_2^j se cumple que $\psi_1(x, y, d_1, d_2^i) \cdot \psi_2(y, z, d_2^i) = \psi_1(x, y, d_1, d_2^j) \cdot \psi_2(y, z, d_2^j)$. Por tanto, podemos sustituir ψ_1 y ψ_2 por un nuevo potencial de probabilidad, ψ_3, definido así:

$$\psi_3(x, y, z, d_1) = \psi_1(x, y, d_1, d_2^i) \cdot \psi_2(y, z, d_2^i)$$

donde d_2^i es un valor de cualquiera de D_2, que podemos escoger arbitrariamente porque la elección no afecta a ψ_3.

El algoritmo para eliminar una variable de decisión D queda así:

- Sacamos de la lista de potenciales de probabilidad todos los que dependen de D, los multiplicamos y los proyectamos sobre un valor cualquiera de D, arbitrariamente escogido. El potencial resultante lo metemos en la lista de potenciales de probabilidad.
- La nueva utilidad será $U' = \max_d U$.

De este modo, la variable D queda eliminada tanto de los potenciales de probabilidad como del potencial de utilidad.

Ejemplo 4.8 Vamos a resolver de nuevo el mismo problema que en el ejemplo anterior, pero ahora con división de potenciales. Recordamos que la utilidad esperada era

$$UE = \max_t \sum_y \max_d \sum_x P(x) \cdot P(y|x:t) \cdot U_0(x,d,t)$$

A diferencia de la versión anterior, para eliminar la variable X sólo sacamos de la lista los potenciales de probabilidad que depende de X; al multiplicarlos obtenemos $\psi_1(x, y:t) = P(x) \cdot P(y|x:t)$; por marginalización —es decir, sumando sobre X— obtenemos $\psi_1(y:t)$, que en este caso es igual a $P(y:t)$; por división obtenemos $P(x|y:t)$. Por tanto, nuestro problema se ha transformado en

$$\begin{aligned} UE &= \max_t \sum_y \max_d P(y:t) \sum_x P(x|y:t) \cdot U_0(x,d,t) \\ &= \max_t \sum_y \max_d P(y:t) \cdot U(y,d,t) \end{aligned}$$

Cada valor de $U(y,d,t)$ indica la utilidad esperada cuando el decisor toma las decisiones $T = t$ y $D = d$ y la variable Y toma el valor y.[3]

Para eliminar D hay que maximizar el potencial de utilidad:

$$\begin{aligned} UE &= \max_t \sum_y P(y:t) \max_d U(y,d,t) \\ &= \max_t \sum_y P(y:t) \cdot U(y,t) \end{aligned}$$

con lo cual obtenemos de paso la política óptima para D.

Para eliminar Y, tomamos todos los potenciales que dependen de Y, que en este caso es uno solo, y por tanto no hace falta multiplicar potenciales de probabilidad, ni marginalizar, ni dividir, sino que directamente multiplicamos ese potencial por el de utilidad, y luego sumamos sobre los valores de Y, con lo cual obtenemos $U(t)$.

Por último eliminamos T por maximización, con lo cual obtenemos la utilidad esperada:

$$UE = \max_t U(t)$$

Como se ve, la eliminación de variables con división de potenciales es muy semejante a la versión sin divisiones. La ventaja es que siempre tenemos la distinción entre utilidad y probabilidad, y el precio que se paga es el coste computacional de hacer las divisiones.

4.3.3. Inversión de arcos

El tercer método que vamos a estudiar para la evaluación de DIs es la inversión de arcos, que es similar al método del mismo nombre para redes bayesianas. En realidad, este método fue propuesto en el contexto de los DIs [60, 72] y nosotros lo hemos adaptado a las redes bayesianas, porque creemos que así es más fácil de entender. De hecho, un enlace $X \to Y$ sólo puede invertirse si X e Y son nodos de azar (aunque sus padres, en caso de que los

[3] Observe que en el ejemplo anterior, en que no hacíamos divisiones, obteníamos un potencial $\psi(y,d,t)$ que no era una utilidad, sino el producto de la probabilidad $P(y,d,t)$ y la utilidad $U(y,d,t)$, dos potenciales que en esta nueva versión se mantienen por separado.

tengan, pueden ser tanto nodos de decisión como de azar); nunca se invierte un enlace en que uno de los nodos sea de decisión o de utilidad.

En primer lugar debemos observar que en la ecuación (4.8), que procede de la (4.5), ψ representa la utilidad asociada al nodo U_0, es decir, la utilidad global. Si el DI tiene un solo nodo de utilidad, ψ es la tabla de utilidad para ese nodo. En cambio, si el DI tiene nodos super-valor, tenemos que calcular ψ a partir de las tablas de utilidad de los nodos de utilidad ordinarios y de las funciones de combinación asociadas a los nodos super-valor. Esto es lo mismo que transformar el DI en uno equivalente que sólo tiene un nodo de utilidad, cuyos padres vienen dados por la unión de los conjuntos de padres de todos los nodos de utilidad ordinarios.

Ejemplo 4.9 Para el DI de la figura 4.1 esta transformación consistiría en eliminar los nodos de utilidad ordinarios, U_1 y U_2, y hacer que X, D y T sean padres de U_0. La tabla de utilidad para U_0 en el nuevo DI es $U_0(x, d, t) = U_1(x, d) + U_2(t)$. □

La evaluación de un DI consiste en aplicar la siguiente ecuación,

$$UE = \sum_{\mathbf{c}_0} \max_{d_0} \ldots \sum_{\mathbf{c}_{n-1}} \max_{d_{n-1}} \sum_{\mathbf{c}_n} \prod_{C \in \mathbf{V}_C} P(c|pa(C))\psi(pa(U)) \tag{4.26}$$

que procede de unir las ecuaciones (4.3) y (4.8). Primero eliminamos una a una las variables de $\mathbf{C}_n$, luego D_{n-1}, y así sucesivamente, hasta eliminar todas las de $\mathbf{C}_0$. Por tanto, este método es muy similar al de eliminación de variables en DIs. La diferencia es que el de eliminación de variables trabaja sobre los potenciales, prescindiendo del grafo, mientras que el de inversión de arcos trabaja sobre el DI. Cada vez que elimina un nodo obtiene un nuevo DI, hasta llegar a un DI que no tiene nodos de azar ni de decisión: sólo tiene un nodo, que será de utilidad. Como este nodo no tiene padres, su tabla de utilidad sólo tendrá un valor, que será la utilidad esperada. Naturalmente, cada vez que elimina un nodo de decisión obtiene la política óptima para esa decisión.

Vamos a explicar primero cómo se elimina un nodo de azar y cómo se elimina un nodo de decisión, y luego indicaremos cómo seleccionar en cada iteración del algoritmo el nodo que debe ser eliminado.

Eliminación de un nodo de azar

Cuando vamos a eliminar un nodo de azar C, podemos encontrar tres situaciones:

1. C no tiene hijos; en este caso se dice que C es un sumidero (cf. def. 2.42).
2. C tiene un único hijo, que es U.
3. C tiene al menos uno o varios hijos de azar y posiblemente también tiene como hijo a U. (C no puede tener como hijo un nodo de decisión porque entonces C sería un predecesor informativo de D y por tanto no se podría eliminar C antes que D.)

En la primera situación, la variable C aparece solamente en el potencial $P(c|pa(C))$. No aparece ni en ψ ni en otras probabilidades condicionales. Por tanto, podemos sacar factor común a los demás potenciales, de modo que nos queda $\sum_c P(c|pa(C))$, que siempre vale 1. Es decir, podemos eliminar $\sum_c$ y $P(c|pa(C))$ del miembro derecho de la ecuación (4.26) sin

alterar la utilidad esperada. En el DI esta operación es equivalente a eliminar el nodo C (y su tabla de probabilidad) directamente.

En la segunda situación, el único potencial de probabilidad que depende de C es la tabla de probabilidad de C, y por tanto el cálculo que debemos hacer es $\sum_c P(c|pa(C))\psi(c, \check{\mathbf{c}})$, donde $\check{\mathbf{C}}$ representa los demás padres del nodo de utilidad: $\check{\mathbf{C}} = Pa(U)\backslash\{C\}$. El resultado es un nuevo potencial de utilidad,

$$U(\check{\mathbf{c}}, pa(C)) = \sum_c P(c|pa(C))\psi(c, \check{\mathbf{c}}) \tag{4.27}$$

En el DI esta operación equivale a eliminar el nodo C y trazar enlaces desde los nodos que eran padres de C hasta U, porque la nueva tabla de utilidad de U va a ser $U(\check{\mathbf{c}}, pa(C))$.

En la tercera situación, C aparece en varios potenciales de probabilidad.[4] Lo que se hace en este caso es invertir todos los enlaces necesarios hasta que C no tenga hijos (primera situación) o tenga a U como único hijo (segunda situación), lo cual siempre es posible por el corolario que vamos a presentar a continuación, y estas dos situaciones ya sabemos cómo resolverlas.

Definición 4.10 (Inversión de un arco en un DI) Sean X e Y dos nodos de azar de un DI tales que existe un enlace $X \to Y$. La inversión de este enlace se realiza sustituyéndolo por el enlace $Y \to X$, trazando enlaces desde los padres de X hasta Y y desde los padres de Y (excepto X) hasta X, y sustituyendo las probabilidades condicionales de estos nodos por las que se definen en las ecuaciones (2.22), (2.26) y (2.27).

Proposición 4.11 En todo DI, para todo nodo de azar X que tenga al menos un hijo de azar existe un nodo Y, hijo de X, tal que no existe ningún camino dirigido desde X hasta Y. Al invertir el enlace $X \to Y$ se obtiene un nuevo DI cuya utilidad esperada y cuya estrategia óptima son las mismas que en el DI original.

La demostración es prácticamente igual a la de la proposición 2.47, pues da lo mismo que los padres X e Y sean nodos de azar o de decisión.

Corolario 4.12 En un DI, todo nodo de azar con n hijos de azar y sin hijos que sean nodos de decisión se puede transformar mediante n inversiones de enlaces en un nodo sin hijos de azar ni de decisión.

En resumen, si C es un sumidero, lo eliminamos del DI sin más. Si su único hijo es U, lo eliminamos por marginalización, como indica la ecuación (4.27). Y si C tiene hijos que sean nodos de azar, invertimos tantos enlaces como sea necesario hasta llegar a una de las dos situaciones anteriores.

[4] Si estuviéramos en el método de eliminación de variables con división de potenciales, bastaría multiplicar esos potenciales para obtener una probabilidad condicional y una probabilidad conjunta, dadas por las ecuaciones (4.20) y (4.21); pero aquí no es tan sencillo, porque $P(\check{\mathbf{c}} : \mathbf{d})$ no es una probabilidad condicional de una sola variable, sino una probabilidad conjunta, la cual no puede formar parte de un DI. Por eso tenemos que buscar otra solución.

Eliminación de un nodo de decisión

Cuando vamos a eliminar un nodo de decisión D, podemos encontrar dos situaciones:

1. D no tiene hijos (también en este caso se dice que D es un sumidero);
2. el único hijo de D es U.

No puede darse la situación de que D tenga como hijo algún nodo de decisión o de azar porque los únicos nodos que se eliminan después que D en el grafo son sus predecesores informativos, es decir, sus antepasados en el grafo. Los demás nodos, incluyendo naturalmente los hijos de D y todos sus demás descendientes, se eliminan antes que D.

Por tanto, D nunca puede tener como hijo un nodo de azar, lo cual implica que ninguna probabilidad condicional depende de la variable D. Pensando de nuevo en la ecuación (4.26), pero teniendo en cuenta que el operador que aparece más a la derecha es $\max_d$, sacamos las probabilidades condicionales como factor común.

Si D no tiene hijos, entonces no pertenece a $Pa(U)$, por tanto ψ no depende de D y, en consecuencia, ψ no se modifica al maximizar sobre D. En el DI esta operación es equivalente a eliminar el nodo sumidero D, sin más.[5]

Si D es padre de U (su único hijo), definimos $\check{\mathbf{D}}$ como el conjunto de los demás padres de U, de modo que $U(pa(U)) = U(d, \check{\mathbf{d}})$. Al maximizar sobre D obtenemos una nueva utilidad,

$$U(\check{\mathbf{d}}) = \max_d U(d, \check{\mathbf{d}}) \tag{4.28}$$

En el DI esta operación consiste en eliminar el nodo D (y todos los enlaces que entraban o salían de él, entre los cuales estaba $D \to U$), de modo que en el nuevo DI los padres de U van a ser $\check{\mathbf{D}}$, es decir, todos los padres que tenía en el DI original, excepto D. La utilidad de U en el nuevo DI es $U(\check{\mathbf{d}})$.

En este caso no hace falta que U herede los padres que D tenía en el DI original, porque $U(\check{\mathbf{d}})$ no depende de ellos (salvo que formen parte de $\check{\mathbf{D}}$, naturalmente; pero en ese caso no hace falta añadir enlaces porque ya existen).

Algoritmo de inversión de arcos

El algoritmo de inversión de arcos actúa iterativamente: en cada iteración elimina un nodo de azar o de decisión, como ya hemos explicado. Para completar el algoritmo sólo nos falta

[5] Observe que si un nodo D_k no tiene hijos, entonces ni las probabilidades condicionales ni la utilidad dependen de D, y por tanto

$$\begin{aligned} UE &= \sum_{\mathbf{c}_0} \max_{d_0} \ldots \sum_{\mathbf{c}_k} \max_{d_k} \sum_{\mathbf{c}_{k+1}} \cdots \sum_{\mathbf{c}_{n-1}} \max_{d_{n-1}} \sum_{\mathbf{c}_n} \prod_{C \in \mathbf{V}_C} P(c|pa(C))\psi(pa(U)) \\ &= \sum_{\mathbf{c}_0} \max_{d_0} \ldots \sum_{\mathbf{c}_k} \sum_{\mathbf{c}_{k+1}} \cdots \sum_{\mathbf{c}_{n-1}} \max_{d_{n-1}} \sum_{\mathbf{c}_n} \prod_{C \in \mathbf{V}_C} P(c|pa(C))\psi(pa(U)) \end{aligned}$$

lo cual implica que en el DI el nodo sumidero D_k se puede eliminar en cualquier momento, no es necesario eliminar primero las variables $\mathbf{C}_{k+1}, D_{k+1}, \ldots, \mathbf{C}_n$.

Sin embargo, en la práctica nunca vamos a encontrar esta situación, porque no tiene sentido incluir en el DI un nodo de decisión que no afecta ni a las probabilidades de otras variables ni a la utilidad. Tampoco puede haber nodos de decisión sumidero como resultado de aplicar el algoritmo de inversión de arcos (aunque sí es habitual convertir nodos de azar en sumideros, con el fin de poder eliminarlos).

indicar cómo se selecciona en cada iteración el nodo que se va a eliminar. Distinguimos dos situaciones, recordando que $\mathbf{C}_n$ es el conjunto de nodos de azar que no tienen entre sus hijos ningún nodo de decisión.

1. Si $\mathbf{C}_n$ contiene nodos (es decir, si hay nodos de azar que no son padres de ningún nodo de decisión), ésos son los primeros que deben ser eliminados. Podemos escoger arbitrariamente cualquiera de ellos para ser eliminado, aunque por motivos de eficiencia conviene eliminar primero los nodos de azar sumideros y los que tienen a U como único hijo, con el fin de realizar el menor número posible de inversiones de arcos.

2. Si $\mathbf{C}_n$ está vacío, es porque todos los nodos de azar tienen como hijo algún nodo de decisión. Por definición, en un DI siempre hay un camino dirigido que contiene todos los nodos de decisión (cf. sec. 4.2.1); en esta segunda situación ello implica que el último nodo de ese camino es un nodo de decisión que es descendiente de todos los demás nodos de azar y de decisión del DI. Éste es el nodo que debe ser eliminado en la próxima iteración del algoritmo.

Al eliminar la decisión D_i, los nodos de $\mathbf{C}_i$ ya no tienen entre sus hijos ningún nodo de decisión; si $\mathbf{C}_i$ no está vacío, volvemos de nuevo a la primera situación; si lo está, eliminamos D_{i-1}, y así sucesivamente. El algoritmo termina porque el número de nodos en el DI es finito.

Ejemplo 4.13 Para evaluar el DI de la figura 4.1 mediante el método de inversión de arcos, lo primero que debemos hacer es transformarlo para que tenga un solo nodo de utilidad, como hemos explicado en el ejemplo 4.9; el resultado se muestra en la figura 4.2.*a*. Vemos que X es la única variable de azar que no tiene como hijo ningún nodo de decisión. Por tanto, ésta es la primera que vamos a eliminar. Pero no podemos hacerlo directamente, porque tiene un hijo, Y. Por eso invertimos el enlace $X \to Y$, lo cual implica que deben compartir los mismos padres; es decir, tenemos que trazar un enlace $T \to X$. El grafo del nuevo DI es el que se muestra en la figura 4.2.*b*. Las probabilidades de estos X e Y en el nuevo diagrama se calculan como indican las ecuaciones (2.22), (2.26) y (2.27):

$$\begin{aligned} P(x, y : t) &= P(x) \cdot P(y|x : t) \\ P(y : t) &= \sum_x P(x, y : t) \\ P(x|y : t) &= \frac{P(x, y : t)}{P(y : t)} \end{aligned}$$

Ahora el único hijo de X es U, y por tanto podemos eliminar X. Los padres de X en el grafo de la figura 4.2.*b* pasan a ser padres de U en el de la fig. 4.2.*c*. La nueva tabla de utilidad es

$$U(y, d, t) = \sum_x P(x|y : t) \cdot U(x, d, t)$$

Como ya no hay más nodos de azar no observables, eliminamos un nodo de decisión. El nodo de decisión que es hijo de todos los demás nodos de decisión y de azar es D. Al eliminarlo obtenemos el DI que se muestra en la figura 4.2.*d*, cuya tabla de utilidad es

$$U(y, t) = \max_d U(y, d, t)$$

Al maximizar sobre D obtenemos la política óptima para esta decisión.

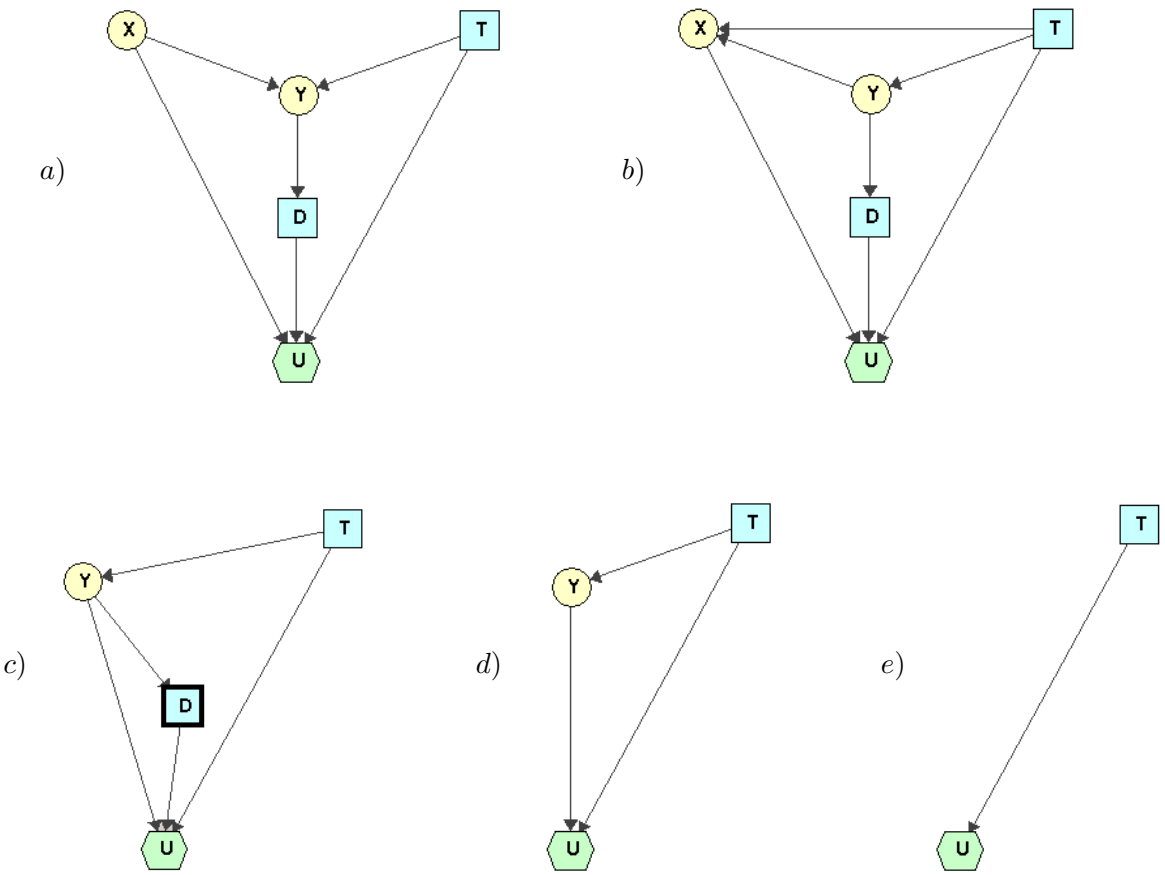


Figura 4.2: Evaluación del DI de la figura 4.1 mediante el método de inversión de arcos.

En este DI, Y es el único nodo de azar que no tiene como hijo ningún nodo de decisión. Por eso eliminamos Y, con lo cual obtenemos el DI de la figura 4.2.e, cuya tabla de utilidad es

$$U(t) = \sum_{y} P(y:t) \cdot U(y,t)$$

Finalmente, al eliminar el único nodo que queda obtenemos un DI que sólo tiene el nodo U. Su utilidad coincide con la utilidad esperada del DI original:

$$UE = \max_{t} U(t)$$

Al maximizar sobre T obtenemos la política óptima para esta decisión.

Como habrá observado el lector, los tres métodos de evaluación de DIs estudiados en esta sección (en realidad son cuatro, porque el de eliminación de variables tiene dos versiones) realizan operaciones muy similares. En concreto, el método de expansión-evaluación de árboles de decisión y el de inversión de arcos, si eliminan las variables en el mismo orden, realizan exactamente los mismos cálculos, aunque este último, en vez de tener las probabilidades y las utilidades dispersas en las ramas del árbol, las almacena de forma compacta en los nodos del DI, lo cual resulta más eficiente desde el punto de vista computacional.

4.4. Construcción de diagramas de influencia

Leer la sección 3 de [22] y ver el vídeo 4.4 que se encuentra en `http://www.ia.uned.es/~fjdiez/docencia/videos-prob-dec`.

4.5. Análisis de sensibilidad

El alumno que lo desee, puede estudiar la sección 4 de [22].

Bibliografía recomendada

En este capítulo hemos planteado brevemente los fundamentos de la teoría bayesiana de la decisión. Una presentación formal puede encontrarse en el trabajo original de von Neumann y Morgenstern [80] y en el libro de Raiffa [69].

En cuanto a los diagramas de influencia, el libro clásico es el de Howard y Matheson [37]. A nuestro juicio, que coincide con el de muchos expertos en la materia, el libro que mejor explica los modelos de análisis de decisiones tanto árboles de decisión como diagramas de influencia, con un extenso tratamiento del análisis de sensibilidad, es el de Clemen y Reilly [9]; una de las mejores cualidades de este libro es la cantidad de ejemplos tomados de la vida real. El libro de Ríos, Bielza y Mateos [70] sobre teoría de la decisión explica los árboles de decisión, los diagramas de influencia, la teoría de la decisión multicriterio, el análisis de sensibilidad, etc. Los aspectos computacionales de la evaluación de los diagramas de influencia pueden encontrarse en estos dos libros y también en los citados en la bibliografía del capítulo 1 (página 36).

En cuanto a la construcción de diagramas influencia, recomendamos encarecidamente dos libros de Ley Borrás [52, 53]. El primero de ellos explica en detalle la obtención de las probababilidades, como ya dijimos en el capítulo 3. El segundo aborda el *análisis de decisiones integral*, que incluye entre sus fases la construcción de un modelo (por ejemplo, un árbol de decisión o un diagrama de influencia), pero es un proceso mucho más amplio y complejo.

En cuanto a los procesos de decisión de Markov (PDM), que pueden verse como una generalización de los diagramas de influencia, el libro más famoso es el de Puterman [68], aunque por su nivel matemático resulta muy difícil de leer. El de Sutton y Barto [77] trata muy bien los PDMs aplicados al aprendizaje en inteligencia artificial. El de Ghallab et al. [28, cap. 16] trata los PDM y los PDMPO aplicados a la planificación en inteligencia artificial. Estos dos últimos libros son muy recomendables para los alumnos que quieran aplicar los modelos de análisis de decisiones en alguno de estos campos.

Actividades

Realizar los ejercicios que aparecen al final de [22], sin mirar las soluciones. Una vez realizados, consultar las soluciones para comprobar que los ha resuelto correctamente.

Capítulo 5

Aplicaciones

Resumen

En este capítulo vamos a dar referencias de algunas de las numerosas aplicaciones de los modelos gráficos probabilistas (MGPs) que se han desarrollado en los campos más variados: a partir de 1990 el crecimiento ha sido exponencial y uno no deja de sorprenderse del número y variedad de aplicaciones que surgen cada día.

Contexto

Las referencias citadas permitirán al lector apreciar la importancia de los modelos y los métodos estudiados en esta asignatura.

Objetivos

El objetivo, aparte de motivar al alumno, es que conozca algunas de las aplicaciones existentes y que pueda buscar por sí mismo otras aplicaciones en el campo que más le interese, con el fin de que pueda desarrollar sus propios modelos si algún día lo necesita en su práctica profesional. Por tanto, no es necesario que el alumno lea detenidamente todos los artículos ni que llegue a comprender todos los detalles de cómo se ha construido cada aplicación. Cada alumno dedicará más tiempo a unas aplicaciones o a otras según sus gustos y sus intereses profesionales.

Requisitos previos

Para entender mejor este capítulo conviene haber estudiado los anteriores. Sin embargo, también sería posible leer este capítulo antes de estudiar el resto, como motivación para el estudio de la asignatura.

Contenido

5.1. Aplicaciones en medicina

La medicina es el campo donde más aplicaciones de MGPs se han construido.

Leer el artículo [21] (medicina.pdf)*, que ofrece una visión de conjunto sobre los MGPs en medicina.*

Leer el artículo [45] (prostanet.pdf)*, que explica la construcción de una red bayesiana para cáncer de próstata y otras enfermedades urológicas. Lo hemos seleccionado porque explica el proceso de construcción del modelo, mediante refinamientos sucesivos de una red causal, mientras que casi todos los artículos que se publican sólo describen la red final, no el proceso que se ha seguido hasta llegar a ella.*

Por otra parte, el artículo de revisión de los modelos canónicos [23], que ya hemos citado anteriormente, incluye en la sección 8.2 una docena de referencias sobre aplicaciones de los modelos canónicos probabilistas en medicina.

5.2. Aplicaciones en ingeniería

Leer el capítulo 12 del libro de Castillo et al. [7].

Otras aplicaciones interesantes son la desarrollada por Volkswagen para predecir la demanda de componentes para la fabricación del VW Golf [27] (volkswagen.pdf) y la desarrollada por Boeing para el diagnóstico y mantenimiento de sus aviones comerciales [43] (boeing.pdf).

5.3. Aplicaciones en informática

5.3.1. Informática educativa

Este es otro campo en que cada vez se utilizan más los modelos gráficos probabilistas. Un par de artículos interesantes sobre modelado del estudiante en informática educativa son los de Conati et al. [10] (student-modeling-Conati.pdf) y Zapata [84] (student-models-Zapata.pdf).

Vomlel [79] (adaptive-testing.pdf) ha utilizado redes bayesianas para construir un programa de ordenador que selecciona las preguntas de un test en función de lo que el alumno ha respondido hasta ese momento.

5.3.2. Interfaces inteligentes

El sistema operativo Microsoft Windows 95 incluía un MGP para el diagnóstico de problemas de impresión.[1] El artículo de Heckerman et al. [33] (troubleshooting.pdf) describe éste y otros modelos de diagnóstico y reparación de averías. También Hewlett-Packard ha desarrollado un modelo para diagnóstico de impresoras, denominado SACSO [39] (sacso.pdf).

Por su parte, Microsoft está investigando activamente el uso de MGPs para hacer interfaces que respondan mejor a las necesidades y expectativas de los usuarios: es el denominado Proyecto Lumière [35] (lumiere.pdf). En las páginas web `http://research.microsoft.`

[1] En la página web `http://www.cs.huji.ac.il/~galel/Repository` se encuentran varias redes bayesianas y diagramas de influencia para distintos dominios, incluida esta red.

com/~horvitz/lum.htm y http://research.microsoft.com/~horvitz/lumiere.HTM puede obtener información adicional, principalmente documentos en PDF y algunos vídeos de demostración.

5.3.3. Seguridad informática

Hoy en día la mayor parte de los filtros de correo basura se basan en el método bayesiano ingenuo, que hemos estudiado en la sección 1.2; véase el artículo "*Bayesian spam filtering*" en la Wikipedia (http://en.wikipedia.org/wiki/Bayesian_spam_filtering). Sobre el filtrado de correo basura en Microsoft Outlook pueden verse las transparencias que se encuentran en el archivo mail-filtering.pdf.

También se están utilizando los MGPs para vigilar las redes informáticas y evitar que los intrusos puedan causar daños; véase, por ejemplo, [1] (cybercrime.pdf).

5.4. Aplicaciones en visión artificial

Para este apartado hemos seleccionado dos artículos: el primero [18] (visual-tracking.pdf) describe un trabajo desarrollado por la división de investigación que Intel (el fabricante de microprocesadores) tiene en China; el segundo [19] (visual-activity-recognition.pdf) es un trabajo hecho en el Instituto Técnológico y de Estudios Superiores de Monterrey, en Cuernavaca, México.

5.5. Aplicaciones financieras y comerciales

En este apartado recomendamos el libro *Probabilistic Methods for Financial and Marketing Informatics*, de Neapolitan [59]. En la primera parte (caps. 1 a 6) ofrece una excelente revisión de las redes bayesianas y los modelos gráficos probabilistas, en la segunda (caps. 7 a 10) analiza diferentes aplicaciones en el campo de las finanzas y en la tercera (caps. 11 y 12) describe varias aplicaciones en mercadotecnia (*marketing*).

5.6. Otras aplicaciones

Por último, vamos a señalar algunas aplicaciones desarrolladas en otras áreas, para que el lector se haga una idea de la variedad de campos en que se están aplicando los MGPs.

Entre las aplicaciones para sistemas de información geográfica (en inglés, *geographic information systems*, GIS), hemos seleccionado el de Laskey et al. [49] (GIS-Laskey.pdf)

Como curiosidad, hemos escogido una aplicación a la meteorología [5] (meteorologia.pdf), desarrollada en Cantabria, España, y una sobre música [82] (harmony.pdf), que utiliza los procesos de decisión de Markov para armonizar (poner acordes de acompañamiento) a una partitura.

Bibliografía recomendada

El lector interesado puede encontrar bastantes aplicaciones más en los libros de Clemen y Reilly[9], Ley Borrás [53] y Neapolitan [59]. Recordamos también que, como dijimos en

la sección de bibliografía recomendada del capítulo 4, el libro de Sutton y Barto [77] trata muy bien los PDMs aplicados al aprendizaje en inteligencia artificial y el de Ghallab et al. [28, caps. 16 y 17] trata los PDM y los PDMPO aplicados a la planificación en inteligencia artificial.

Actividades

Hoy en día los MGPs se están aplicando en casi todos los campos. Por ello sugerimos como actividad que el estudiante realice una búsqueda en Internet sobre el tema que más le interese. Por ejemplo, quien esté interesado en el ajedrez puede realizar una búsqueda con los términos "Bayesian networks" y "chess", o bien "influence diagrams" y "chess", o bien "Markov", "decision" y "chess". Se sorprenderá al ver cuántas referencias encuentra.

Referencias

[1] N. Abouzakhar, A. Gani, G. Manson, M. Abuitbel y D. King. Bayesian learning networks approach to cybercrime detection. En: *Proceedings of the 2003 PostGraduate Networking Conference (PGNET-2003)*, Liverpool, UK, 2003.

[2] J. M. Agosta (ed.). *Proceedings of the First UAI Bayesian Modelling Applications Workshop (BMAW'03)*, Acapulco, Mexico, 2003.

[3] A. Cano y S. Moral. Heuristic algorithms for the triangulation of graphs. En: B. Bouchon-Meunie, R. R. Yager y I. A. Zadeh (eds.), *Advances in Intelligent Computing (IPMU-94)*, págs. 98–107. Springer-Verlag, Berlin, 1995.

[4] J. E. Cano, L. D. Hernández y S. Moral. Importance sampling algorithms for the propagation of probabilities in belief networks. *International Journal of Approximate Reasoning*, **15**:77–92, 1996.

[5] R. Cano, C. Sordo y J. M. Gutiérrez. Applications of Bayesian networks in meteorology. En: J. A. Gámez, S. Moral y A. Salmerón (eds.), *Advances in Bayesian Networks*, volume 146 de *Studies in Fuzziness and Soft Computing*, págs. 309–327, Berlin, 2004. Springer.

[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. Castillo, J. M. Gutiérrez y A. S. Hadi. *Sistemas Expertos y Modelos de Redes Probabilísticas*. Academia de Ingeniería, Madrid, 1997.

[8] J. Cheng y M. J. Druzdzel. AIS-BN: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. *Journal of Artificial Intelligence Research*, **13**:155–188, 2000.

[9] R. T. Clemen y T. A. Reilly. *Making Hard Decisions*. Duxbury, Pacific Grove, CA, 2001.

[10] C. Conati, A. S. Gertner, K. VanLehn y M. J. Druzdzel. On-line student modeling for coached problem solving using Bayesian networks. En: *Proceedings of the Sixth International Conference on User Modeling (UM'97)*, págs. 231–242. Springer, Vienna, Austria, Chia Laguna, Italy, 1997.

[11] G. Cooper y S. Moral (eds.). *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98)*, San Francisco, CA, 1998. Morgan Kauffmann.

[12] G. F. Cooper. A method for using belief networks as influence diagrams. En: R. D. Shachter, T. Levitt, L. N. Kanal y J. F. Lemmer (eds.), *Uncertainty in Artificial Intelligence 4 (UAI'88)*, págs. 55–63, Amsterdam, The Netherlands, 1988. Elsevier Science Publishers.

[13] G. F. Cooper y E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**:309–347, 1992.

[14] F. G. Cozman. Generalizing variable elimination in Bayesian networks. En: *Proceedings of the IBERAMIA/SBIA 2000 Workshops (Workshop on Probabilistic Reasoning in Artificial Intelligence)*, págs. 27–32, São Paulo, Brazil, 2000.

[15] A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, New York, 2009.

[16] 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.

[17] R. Dechter. Bucket elimination: A unifying framework for probabilistic inference. En: Horvitz y Jensen [36], págs. 211–219.

[18] Q. Diao, J. Lu, W. Hu, Y. Zhang y G. Bradski. DBN models and a prediction method for visual tracking. En: Agosta [2].

[19] R. Díaz de León y L. E. Sucar. A general Bayesian network model for visual activity recognition. En: Agosta [2].

[20] F. J. Díez. *Sistema Experto Bayesiano para Ecocardiografía*. Tesis doctoral, Dpto. Informática y Automática, UNED, Madrid, 1994.

[21] 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.

[22] F. J. Díez. Teoría probabilista de la decisión en medicina. Informe Técnico CISIAD-07-01, UNED, Madrid, 2007.

[23] F. J. Díez y M. J. Druzdzel. Canonical probabilistic models for knowledge engineering. Technical Report CISIAD-06-01, UNED, Madrid, Spain, 2006.

[24] M. J. Druzdzel y F. J. Díez. Combining knowledge from different sources in probabilistic models. *Journal of Machine Learning Research*, **4**:295–316, 2003.

[25] R. O. Duda y P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[26] B. Fishhoff. Debiasing. En: D. Kahneman, P. Slovic y A. Tversky (eds.), *Judgement under Uncertainty: Heuristics and Biases*, cap. 31, págs. 249–267. Cambridge University Press, Cambridge, UK, 1982.

[27] J. Gebhardt, H. Detmer y A. L. Madsen. Predicting parts demand in the automotive industry—an application of probabilistic graphical models. En: Agosta [2].

[28] M. Ghallab, D. Nau y P. Traverso. *Automated Planning : Theory and Practice*. Morgan Kaufmann, San Francisco, CA, 2004.

[29] C. Glymour y G. F. Cooper. *Computation, Causation and Discovery*. The MIT Press, Cambridge, Massachusetts, 1999.

[30] G. A. Gorry. Computer-assisted clinical decision making. *Methods of Information in Medicine*, **12**:45–51, 1973.

[31] G. A. Gorry y G. O. Barnett. Experience with a model of sequential diagnosis. *Computers and Biomedical Research*, **1**:490–507, 1968.

[32] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmon, WA, 1995.

[33] D. Heckerman, J. S. Breese y K. Rommelse. Decision-theoretic troubleshooting. *Communications of the ACM*, **38**:49–57, 1995.

[34] L. D. Hernández, S. Moral y A. Salmerón. A Monte Carlo algorithm for probabilistic propagation in belief networks based on importance sampling and stratified simulation techniques. *International Journal of Approximate Reasoning*, **18**:53–91, 1998.

[35] E. Horvitz, J. Breese, D. Heckerman, D. Hovel y K. Rommelse. The Lumière project: Bayesian user modeling for inferring the goals and needs of software users. En: Cooper y Moral [11], págs. 256–265.

[36] E. Horvitz y F. V. Jensen (eds.). *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI'96)*, San Francisco, CA, 1996. Morgan Kauffmann.

[37] R. A. Howard y J. E. Matheson. *Readings on the Principles and Applications of Decision Analysis*. Strategic Decisions Group, Menlo Park, CA, 1984.

[38] F. V. Jensen, B. Chamberlain, T. Nordahl y F. Jensen. Analysis in HUGIN of data conflict. En: B. D'Ambrossio, P. Smets y P. Bonissone (eds.), *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence (UAI'91)*, págs. 519–528, San Mateo, CA, 1991. Morgan Kauffmann.

[39] F. V. Jensen, U. Kjærulff, B. Kristiansen, H. Langseth, C. Skaanning, J. Vomlel y M. Vomlelová. The SACSO methodology for troubleshooting complex systems. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **15**:321–333, 2001.

[40] F. V. Jensen y T. D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York, segunda edición, 2007.

[41] 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.

[42] D. Kahneman, P. Slovic y A. Tversky (eds.). *Judgement under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK, 1982.

[43] O. Kipersztok y G. Provan. A framework for diagnostic inference of commercial aircraft systems. En: Agosta [2].

[44] D. Koller y N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, Cambridge, MA, 2009.

[45] C. Lacave y F. J. Díez. Knowledge acquisition in Prostanet, a Bayesian network for diagnosing prostate cancer. *Lecture Notes in Computer Science*, **2774**:1345–1350, 2003.

[46] C. Lacave, A. Oniśko y F. J. Díez. Use of Elvira's explanation facilities for debugging probabilistic expert systems. *Knowledge-Based Systems*, **19**:730–738, 2006.

[47] 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.

[48] K. B. Laskey, S. M. Mahoney y J. Goldsmith (eds.). *Proceedings of the Fifth UAI Bayesian Modelling Applications Workshop (BMAW'07)*, Vancouver, British Columbia, Canada, 2007.

[49] K. B. Laskey, E. J. Wright y P. C. G. da Costa. Envisioning uncertainty in geospatial information. En: Laskey et al. [48].

[50] S. L. Lauritzen y D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, **50**:157–224, 1988.

[51] V. Lepar y P. P. Shenoy. A comparison of Lauritzen-Spiegelhalter, HUGIN, and Shenoy-Shafer architectures for computing marginals of probability distributions. En: Cooper y Moral [11], págs. 328–337.

[52] R. Ley Borrás. *Análisys de Incertidumbre y Riesgo para la Toma de Decisiones*. Comunidad Morelos, Orizaba, Ver., México, 2001.

[53] R. Ley Borrás. *Análisis de Decisiones Integral*. Consultoría en Decisiones, Orizaba, Ver., México, 2009.

[54] 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.

[55] S. Moral y A. Salmerón. Dynamic importance sampling in Bayesian networks based on probability trees. *International Journal of Approximate Reasoning*, **38**:245–261, 2005.

[56] M. G. Morgan y M. Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge, UK, 1990.

[57] R. E. Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. Wiley-Interscience, New York, 1990.

[58] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, Upper Saddle River, NJ, 2004.

[59] R. E. Neapolitan y X. Jiang. *Probabilistic Methods for Financial and Marketing Informatics*. Morgan Kaufmann, San Francisco, CA, 2007.

[60] S. M. Olmsted. *On Representing and Solving Decision Problems.* Tesis doctoral, Dept. Engineering-Economic Systems, Stanford University, CA, 1983.

[61] A. Oniśko, M. J. Druzdzel y H. Wasyluk. Extension of the Hepar II model to multiple-disorder diagnosis. En: M. Kłopotek, M. Michalewicz y S. T. Wierzchoń (eds.), *Intelligent Information Systems*, págs. 303–313. Springer-Verlag, Heidelberg, 2000.

[62] J. Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving.* Addison-Wesley, Boston, MA, 1984.

[63] J. Pearl. Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, **29**:241–288, 1986.

[64] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[65] J. Pearl. *Causality. Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, UK, 2000.

[66] M. A. Peot. Geometric implications of the Naive Bayes assumption. En: Horvitz y Jensen [36], págs. 414–419.

[67] S. Plous. *The Psychology of Judgment and Decision Making.* McGraw-Hill, New York, 1993.

[68] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, New York, 1994.

[69] H. Raiffa. *Decision Analysis. Introductory Lectures on Choices under Uncertainty.* Addison-Wesley, Reading, MA, 1968.

[70] S. Ríos, C. Bielza y A. Mateos. *Fundamentos de los Sistemas de Ayuda a la Decisión.* Ra-Ma, Madrid, 2002.

[71] R. Shachter y M. Peot. Simulation approaches to general probabilistic inference on belief networks. En: P. Bonissone, M. Henrion, L. N. Kanal y J. F. Lemmer (eds.), *Uncertainty in Artificial Intelligence 6 (UAI'90)*, págs. 221–231, Amsterdam, The Netherlands, 1990. Elsevier Science Publishers.

[72] R. D. Shachter. Evaluating influence diagrams. *Operations Research*, **34**:871–882, 1986.

[73] P. P. Shenoy. Binary join trees for computing marginals in the Shenoy-Shafer architecture. *International Journal of Approximate Reasoning*, **17**:1–25, 1997.

[74] 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.

[75] P. Spirtes y C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, **9**:62–72, 1991.

[76] P. Spirtes, C. Glymour y R. Scheines. *Causation, Prediction and Search.* The MIT Press, Cambridge, Massachusetts, segunda edición, 2000.

[77] R. S. Sutton y A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[78] P. Szolovits y S. G. Pauker. Categorical and probabilistic reasoning in medicine. *Artificial Intelligence*, **11**:115–144, 1978.

[79] J. Vomlel. Building adaptive tests using Bayesian networks. *Kybernetica*, **40**:333–348, 2004.

[80] J. von Neumann y O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

[81] 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.

[82] L. Yi y J. Goldsmith. Automatic generation of four-part harmony. En: Laskey et al. [48].

[83] C. Yuan y M. J. Druzdzel. Importance sampling algorithms for Bayesian networks: Principles and performance. *Mathematical and Computer Modeling*, **43**:1189–1207, 2005.

[84] D. Zapata-Rivera. Indirectly visible Bayesian student models. En: Laskey et al. [48].