Tutorial sobre Máquinas de Vectores Soporte (SVM)

Enrique J. Carmona Suárez
ecarmona@dia.uned.es

Versión inicial: 2013 Última versión: 11 Julio 2014

Dpto. de Inteligencia Artificial, ETS de Ingeniería Informática, Universidad Nacional de Educación a Distancia (UNED), C/Juan del Rosal, 16, 28040-Madrid (Spain)

Resumen

Este tutorial presenta una introducción al mundo de las máquinas de vectores soporte (SVM, del inglés *Support Vector Machine*), aplicadas a resolver tareas tanto de clasificación como de regresión. En el primer tipo de tareas, la descripción se restringe al caso de clasificación binaria y, atendiendo al tipo de separabilidad de los ejemplos de entrada, se consideran distintas opciones. Así, en primer lugar, se aborda el caso ideal de ejemplos perfectamente separables linealmente para, seguidamente, abordar el caso más realista de ejemplos que, aunque afectados por ruido, se consideran linealmente cuasi-separables y, finalmente, se considera el caso de ejemplos no separables linealmente, donde las SVM demuestran su gran potencialidad. La descripción de las SVMs aplicadas a la tarea de regresión corre también de forma paralela, abarcando los casos tanto de regresión lineal como no lineal. Finalmente, se presentan algunas herramientas software de uso libre que implementan este tipo de paradigma de aprendizaje y con las que el usuario puede empezar a experimentar.

1. Introducción

Las máquinas de vectores soporte (SVM, del inglés *Support Vector Machines*) tienen su origen en los trabajos sobre la teoría del aprendizaje estadístico y fueron introducidas en los años 90 por Vapnik y sus colaboradores [Boser et al., 1992, Cortes & Vapnik, 1995]. Aunque originariamente las SVMs fueron pensadas para resolver problemas de clasificación binaria, actualmente se utilizan para resolver otros tipos de problemas (regresión, agrupamiento, multiclasificación). También son diversos los campos en los que han sido utilizadas con éxito, tales como visión artificial, reconocimiento de caracteres, categorización de texto e hipertexto, clasificación de proteínas, procesamiento de lenguaje natural, análisis de series temporales. De hecho, desde su introducción, han ido ganando un merecido reconocimiento gracias a sus sólidos fundamentos teóricos.

Dentro de la tarea de clasificación, las SVMs pertenecen a la categoría de los clasificadores lineales, puesto que inducen separadores lineales o hiperplanos, ya sea en el espacio original de los ejemplos de entrada, si éstos son separables o cuasi-separables (ruido), o en un espacio transformado (espacio de características), si los ejemplos no son separables linealmente en el espacio original. Como se verá más adelante, la búsqueda del hiperplano de separación en

estos espacios transformados, normalmente de muy alta dimensión, se hará de forma implícita utilizando las denominadas funciones *kernel*.

Mientras la mayoría de los métodos de aprendizaje se centran en minimizar los errores cometidos por el modelo generado a partir de los ejemplos de entrenamiento (error empírico), el sesgo inductivo asociado a las SVMs radica en la minimización del denominado *riesgo estructural*. La idea es seleccionar un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase para, de esta forma, conseguir lo que se denomina un *margen máximo* a cada lado del hiperplano. Además, a la hora de definir el hiperplano, sólo se consideran los ejemplos de entrenamiento de cada clase que caen justo en la frontera de dichos márgenes. Estos ejemplos reciben el nombre de *vectores soporte*. Desde un punto de vista práctico, el hiperplano separador de margen máximo ha demostrado tener una buena capacidad de generalización, evitando en gran medida el problema del sobreajuste a los ejemplos de entrenamiento.

Desde un punto de vista algorítmico, el problema de optimización del margen geométrico representa un problema de optimización cuadrático con restricciones lineales que puede ser resuelto mediante técnicas estándar de programación cuadrática. La propiedad de convexidad exigida para su resolución garantizan una solución única, en contraste con la no unicidad de la solución producida por una red neuronal artificial entrenada con un mismo conjunto de ejemplos.

Dado el carácter introductorio de este tutorial, los contenidos del mismo sólo abarcan una pequeña parcela del extenso campo relacionado con las máquinas vectores soporte. Por ejemplo, el problema de clasificación sólo se describirá para el caso de clases binarias. Concretamente, en la sección 2 se abordará el problema de clasificación binaria para ejemplos perfectamente separables mediante lo que se conoce como SVMs de "margen duro" (*hard margen*). Dado que en la práctica es normal que los ejemplos de entrenamiento puedan contener ruido, la sección 3 se dedica abordar el problema de clasificación binaria para ejemplos cuasi-separables linealmente, mediante lo que se denomina SVMs de "margen blando" (*soft margen*). La sección 4 culmina el problema de clasificación binaria tratando el caso de la clasificación de ejemplos no separables lienalmente mediante lo que se conoce como SVM kernelizadas. Seguidamente, la sección 5 aborda el problema de regresión mediante lo que se conoce como SVR (del inglés *Support Vector Regression machines*). En esta sección, se aborda tanto el caso de regresión lineal como el caso de regresión no lineal. En la sección 6 se describe algunos de los paquetes software de uso libre más relevante dedicados a la implementación de SVMs. Pueden ser un buen punto de comienzo para que el lector se familiarice, desde un punto de vista práctico, con este paradigma de aprendizaje. Finalmente, la sección 7 corresponde a un anexo dedicado a formular, de forma resumida, aquellos resultados de la teoría de la optimización necesarios para solucionar los diferentes problemas de optimización que surgen como consecuencia de abordar los problemas de clasificación y de regresión mediante SVMs.

2. SVM para clasificación binaria de ejemplos separables linealmente

Dado un conjunto separable de ejemplos $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, donde $\boldsymbol{x}_i \in \mathbb{R}^d$ e $y_i \in \{+1, -1\}$, se puede definir un hiperplano de separación (ver fig. 1a) como una función lineal que es capaz de separar dicho conjunto sin error:

$$D(\boldsymbol{x}) = (w_1 x_1 + ... + w_d x_d) + b =< \boldsymbol{w}, \boldsymbol{x} > +b \tag{1}$$

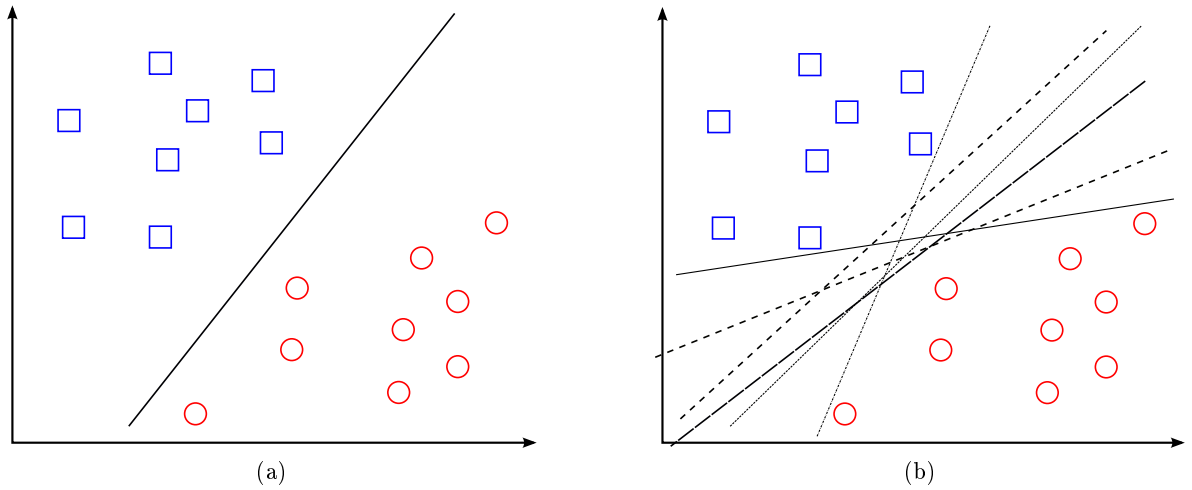


Figura 1: Hiperplanos de separación en un espacio bidimensional de un conjunto de ejemplos separables en dos clases: (a) ejemplo de hiperplano de separación (b) otros ejemplos de hiperplanos de separación, de entre los infinitos posibles.

donde $\boldsymbol{w}$ y b son coeficientes reales. El hiperplano de separación cumplirá las siguientes restricciones para todo $\boldsymbol{x}_i$ del conjunto de ejemplos:

$$\begin{array}{lll} < \boldsymbol{w}, \boldsymbol{x}_i > +b \geq 0 & \text{si} & y_i = +1 \\ < \boldsymbol{w}, \boldsymbol{x}_i > +b \leq 0 & \text{si} & y_i = -1,\ i = 1, ..., n \end{array} \tag{2}$$

o también,

$$y_i \left(< \boldsymbol{w}, \boldsymbol{x}_i > +b \right) \geq 0, \qquad i = 1, \ldots, n \tag{3}$$

o de forma más compacta

$$y_i D(\boldsymbol{x}_i) \geq 0, \qquad i = 1, \ldots, n \tag{4}$$

Tal y como se puede deducir fácilmente de la fig. 1b, el hiperplano que permite separar los ejemplos no es único, es decir, existen infinitos hiperplanos separables, representados por todos aquellos hiperplanos que son capaces de cumplir las restricciones impuestas por cualquiera de las expresiones equivalentes (3-4). Surge entonces la pregunta sobre si es posible establecer algún criterio adicional que permita definir un hiperplano de separación óptimo. Para ello, primero, se define el concepto de *margen* de un hiperplano de separación, denotado por τ, como la mínima distancia entre dicho hiperplano y el ejemplos más cercano de cualquiera de las dos clases (ver fig. 2a). A partir de esta definición, un hiperplano de separación se denominará *óptimo* si su margen es de tamaño máximo (fig. 2b).

Una propiedad inmediata de la definición de hiperplano de separación óptimo es que éste equidista del ejemplo más cercano de cada clase. La demostración de esta propiedad se puede hacer fácilmente por reducción al absurdo. Supongamos que la distancia del hiperplano óptimo al ejemplo más cercano de la clase $+1$ fuese menor que la correspondiente al ejemplo más cercano de la clase -1. Esto significaría que se puede alejar el hiperplano del ejemplo de la clase $+1$ una distancia tal que la distancia del hiperplano a dicho ejemplo sea mayor que antes y, a su vez, siga siendo menor que la distancia al ejemplo más cercano de la clase -1. Se llega así al absurdo de poder aumentar el tamaño del margen cuando, de partida, habíamos supuesto que éste era máximo (hiperplano óptimo). Se aplica un razonamiento similar en el

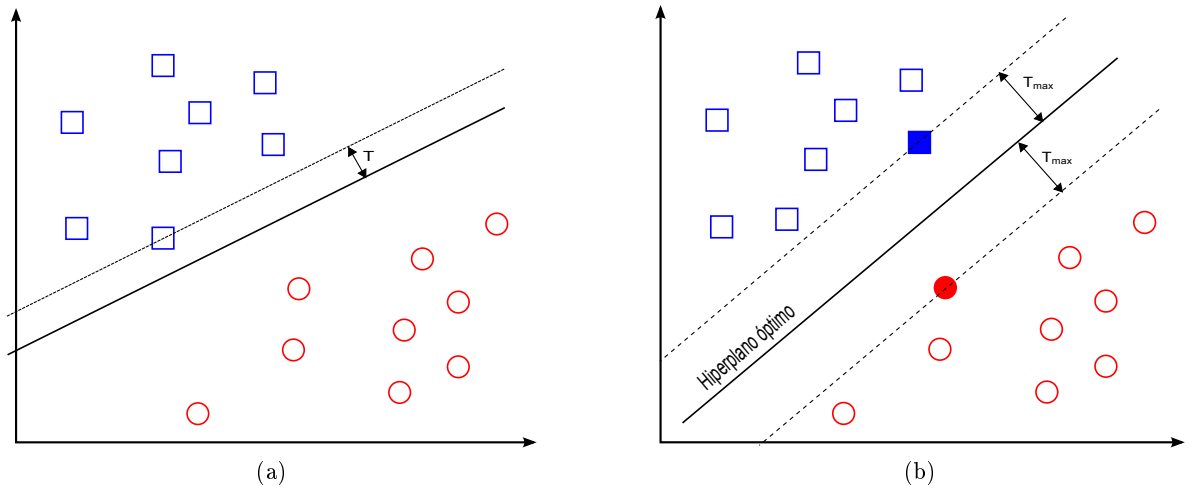


Figura 2: Margen de un hiperplano de separación: (a) hiperplano de separación no-óptimo y su margen asociado (no máximo) (b) hiperplano de separación óptimo y su margen asociado (máximo).

caso de suponer que la distancia del hiperplano óptimo al ejemplo más cercano de la clase -1 fuese menor que la correspondiente al ejemplo más cercano de la clase $+1$.

Por geometría, se sabe que la distancia entre un hiperplano de separación $D(x)$ y un ejemplo x' viene dada por

$$\frac{|D(\boldsymbol{x}')|}{\|\boldsymbol{w}\|} \tag{5}$$

siendo $|\cdot|$ el operador valor absoluto, $\|\cdot\|$ el operador norma de un vector y $\boldsymbol{w}$ el vector que, junto con el parámetro b, define el hiperplano $D(x)$ y que, además, tiene la propiedad de ser perpendicular al hiperplano considerado. Haciendo uso de la expresiones (4) y (5), todos los ejemplos de entrenamiento cumplirán que:

$$\frac{y_i D(\boldsymbol{x}_i)}{\|\boldsymbol{w}\|} \geq \tau, \qquad i = 1, \ldots, n \tag{6}$$

De la expresión anterior, se deduce que encontrar el hiperplano óptimo es equivalente a encontrar el valor de $\boldsymbol{w}$ que maximiza el margen. Sin embargo, existen infinitas soluciones que difieren solo en la escala de $\boldsymbol{w}$. Así, por ejemplo, todas las funciones lineales $\lambda\,(< \boldsymbol{w}, \boldsymbol{x} > +b)$, con $\lambda \in \mathbb{R}$, representan el mismo hiperplano. Para limitar, por tanto, el número de soluciones a una sola, y teniendo en cuenta que (6) se puede expresar también como

$$y_i D(\boldsymbol{x}_i) \geq \tau \|\boldsymbol{w}\|, \qquad i = 1, \ldots, n \tag{7}$$

la escala del producto de τ y la norma de $\boldsymbol{w}$ se fija, de forma arbitraria, a la unidad, es decir

$$\tau \|\boldsymbol{w}\| = 1 \tag{8}$$

Llegando a la conclusión final de que aumentar el margen es equivalente a disminuir la norma de $\boldsymbol{w}$, ya que la expresión anterior se puede expresar como

$$\tau = \frac{1}{\|\boldsymbol{w}\|} \tag{9}$$

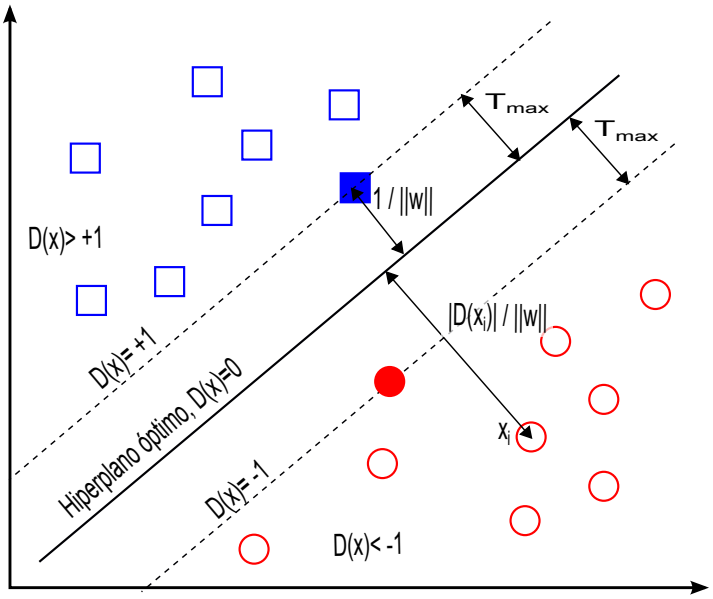


Figura 3: La distancia de cualquier ejemplo,$\boldsymbol{x}_i$, al hiperplano de separación óptimo viene dada por $|D(\boldsymbol{x}_i)| \,/\, \|\boldsymbol{w}\|$. En particular, si dicho ejemplo pertenece al conjunto de vectores soporte (identificados por siluetas sólidas), la distancia a dicho hiperplano será siempre $1/\|\boldsymbol{w}\|$. Además, los vectores soporte aplicados a la función de decisión siempre cumplen que $|D(\boldsymbol{x})| = 1$.

Por tanto, de acuerdo a su definición, un hiperplano de separación óptimo (ver fig. 3) será aquel que posee un margen máximo y, por tanto, un valor mínimo de $\|\boldsymbol{w}\|$ y, además, está sujeto a la restricción dada por (7), junto con el criterio expresado por (8), es decir,

$$y_i D(\boldsymbol{x}_i) \geq 1, \qquad i = 1, \ldots, n \tag{10}$$

o lo que es lo mismo

$$y_i \left(< \boldsymbol{w}, \boldsymbol{x}_i > + b\right) \geq 1, \qquad i = 1, \ldots, n \tag{11}$$

El concepto de margen máximo está relacionado directamente con la capacidad de generalización del hiperplano de separación, de tal forma que, a mayor margen, mayor distancia de separación existirá entre las dos clases. Los ejemplos que están situados a ambos lados del hiperplano óptimo y que definen el margen o, lo que es lo mismo, aquellos para los que la restricción (11) es una igualdad, reciben el nombre de *vectores soporte* (ver fig. 3). Puesto que estos ejemplos son los más cercanos al hiperplano de separación, serán los más difíciles de clasificar y, por tanto, deberían ser los únicos ejemplos a considerar a la hora de construir dicho hiperplano. De hecho, se demostrará más adelante, en esta misma sección, que el hiperplano de separación óptimo se define sólo a partir de estos vectores.

La búsqueda del hiperplano óptimo para el caso separable puede ser formalizado como el problema de encontrar el valor de $\boldsymbol{w}$ y b que minimiza el funcional $f(\boldsymbol{w}) = \|\boldsymbol{w}\|$ sujeto a las restricciones (11), o de forma equivalente[1]

$$\begin{aligned} &\text{mín} \quad f(\boldsymbol{w}) = \tfrac{1}{2} \|\boldsymbol{w}\|^2 = \tfrac{1}{2} < \boldsymbol{w}, \boldsymbol{w} > \\ &\text{s.a.} \quad y_i \left(< \boldsymbol{w}, \boldsymbol{x}_i > + b\right) - 1 \geq 0, \qquad i = 1, \ldots, n \end{aligned} \tag{12}$$

[1]Obsérvese que es equivalente minimizar $f(\boldsymbol{w}) = \|\boldsymbol{w}\|$ o el funcional propuesto en (12). El proceso de minimización de este funcional equivalente, en lugar del original, permitirá simplificar la notación posterior, obteniendo expresiones más compactas.

Este problema de optimización con restricciones corresponde a un problema de programación cuadrático y es abordable mediante la *teoría de la optimización*. Dicha teoría establece que un problema de optimización, denominado primal, tiene una forma dual si la función a optimizar y las restricciones son funciones estrictamente convexas. En estas circunstancias, resolver el problema dual permite obtener la solución del problema primal.

Así, puede demostrarse que el problema de optimización dado por (12) satisface el criterio de convexidad y, por tanto, tiene un dual. En estas condiciones, y aplicando los resultados descritos en el anexo, al final de este tutorial, se pueden enumerar los siguientes pasos encaminados a resolver el problema primal:

En el primer paso, se construye un problema de optimización sin restricciones utilizando la función Lagrangiana[2]:

$$L\left(\boldsymbol{w}, b, \boldsymbol{\alpha}\right) = \frac{1}{2} < \boldsymbol{w}, \boldsymbol{w} > - \sum_{i=1}^{n} \alpha_i \left[y_i \left(< \boldsymbol{w}, \boldsymbol{x}_i > + b\right) - 1\right] \tag{13}$$

donde los $\alpha_i \geq 0$ son los denomiandos multiplicadores de Lagrange.

El segundo paso consiste en aplicar las condiciones de Karush-Kuhn-Tucker (ver anexo al final de este tutorial), también conocidas como condiciones KKT:

$$\frac{\partial L\left(\boldsymbol{w}^*, b^*, \boldsymbol{\alpha}\right)}{\partial \boldsymbol{w}} \equiv \boldsymbol{w}^* - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = 0,\ i = 1, \ldots, n \tag{14}$$

$$\frac{\partial L\left(\boldsymbol{w}^*, b^*, \boldsymbol{\alpha}\right)}{\partial b} \equiv \sum_{i=1}^{n} \alpha_i y_i = 0,\ i = 1, \ldots, n \tag{15}$$

$$\alpha_i \left[1 - y_i \left(< \boldsymbol{w}^*, \boldsymbol{x}_i > + b^*\right)\right] = 0,\ i = 1, \ldots, n \tag{16}$$

Las restricciones representadas por (14-15) corresponden al resultado de aplicar la primera condición KKT, y las expresadas en (16), al resultado de aplicar la denominada condición complementaria (segunda condición KKT). Las primeras permiten expresar los parámetros de $\boldsymbol{w}$ y b en términos de α_i:

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i,\ i = 1, \ldots, n \tag{17}$$

y, además, establecen restricciones adicionales para los coeficientes α_i:

$$\sum_{i=1}^{n} \alpha_i y_i = 0,\ i = 1, \ldots, n \tag{18}$$

Con las nuevas relaciones obtenidas, se construirá el problema dual. Así, bastara usar (17) para expresar la función Lagrangiana sólo en función de α_i. Antes de ello, se puede reescribir (13) como

$$L\left(\boldsymbol{w}, b, \boldsymbol{\alpha}\right) = \frac{1}{2} < \boldsymbol{w}, \boldsymbol{w} > - \sum_{i=1}^{n} \alpha_i y_i < \boldsymbol{w}, \boldsymbol{x}_i > - b \sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i$$

[2]El signo menos del segundo sumando es debido a que las restricciones de (12) están expresadas como restricciones del tipo $g(\boldsymbol{x}) \geq 0$ en lugar de $g(\boldsymbol{x}) \leq 0$.

Teniendo en cuenta que, según la condición (18), el tercer término de la expresión anterior es nulo, la substitución de (17) en dicha expresión resulta ser

$$L(\boldsymbol{\alpha}) = \frac{1}{2}\left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i\right)\left(\sum_{j=1}^{n}\alpha_j y_j \boldsymbol{x}_j\right) - \left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i\right)\left(\sum_{j=1}^{n}\alpha_j y_j \boldsymbol{x}_j\right) + \sum_{i=1}^{n}\alpha_i$$

$$L(\boldsymbol{\alpha}) = -\frac{1}{2}\left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i\right)\left(\sum_{j=1}^{n}\alpha_j y_j \boldsymbol{x}_j\right) + \sum_{i=1}^{n}\alpha_i$$

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j < \boldsymbol{x}_i, \boldsymbol{x}_j > \tag{19}$$

Es decir, hemos transformado el problema de minimización primal (12), en el problema dual, consistente en maximizar (19) sujeto a las restricciones (18), junto a las asociadas originalmente a los multiplicadores de Lagrange:

$$\begin{array}{ll} \text{máx} & L(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j < \boldsymbol{x}_i, \boldsymbol{x}_j > \\ \text{s.a.} & \sum_{i=1}^{n}\alpha_i y_i = 0 \\ & \alpha_i \geq 0,\ i = 1, \ldots, n \end{array} \tag{20}$$

Al igual que el problema primal, este problema es abordable mediante técnicas estándar de programación cuadrática. Sin embargo, como se puede comprobar, el tamaño del problema de optimización dual escala con el número de muestras, n, mientras que el problema primal lo hace con la dimensionalidad, d. Por tanto, aquí radica la ventaja del problema dual, es decir, el coste computacional asociado a su resolución es factible incluso para problemas con un número muy alto de dimensiones.

La solución del problema dual, $\boldsymbol{\alpha}^*$, nos permitirá obtener la solución del problema primal. Para ello, bastará sustituir dicha solución en la expresión (17) y, finalmente, sustituir el resultado así obtenido en (1), es decir:

$$D(\boldsymbol{x}) = \sum_{i=1}^{n}\alpha_i^* y_i < \boldsymbol{x}, \boldsymbol{x}_i > + b^* \tag{21}$$

Volviendo a las restricciones (16), resultantes de aplicar la segunda condición KKT, se puede afirmar que si $\alpha_i > 0$ entonces el segundo factor de la parte izquierda de dicha expresión tendrá que ser cero y, por tanto

$$y_i\left(< \boldsymbol{w}^*, \boldsymbol{x}_i > + b^*\right) = 1 \tag{22}$$

es decir, el correspondiente ejemplo, $(\boldsymbol{x}_i, y_i)$, satisface la correspondiente restricción del problema primal (12), pero considerando el caso "igual que". Por definición, los ejemplos que satisfacen las restricciones expresadas en (12), considerando el caso "igual que", son los vectores soporte y, por consiguiente, se puede afirmar que sólo los ejemplos que tengan asociado un $\alpha_i > 0$ serán vectores soporte. De este resultado, también puede afirmarse que el hiperplano de separación (21) se construirá como una combinación lineal de sólo los vectores soporte del conjunto de ejemplos, ya que el resto de ejemplos tendrán asociado un $\alpha_j = 0$.

Para que la definición del hiperplano (21) sea completa, es preciso determinar el valor del parámetro b^*. Su valor se calcula despejando b^* de (22):

$$b^* = y_{vs} - < \boldsymbol{w}^*, \boldsymbol{x}_{vs} > \tag{23}$$

donde $(\boldsymbol{x}_{vs}, y_{vs})$ representa la tupla de cualquier vector soporte, junto con su valor de clase, es decir, la tupla de cualquier ejemplo que tenga asociado un α_i distinto de cero. En la práctica, es más robusto obtener el valor de b^* promediando a partir de todos los vectores soporte, N_{vs}. Así, la expresión (23) se transforma en:

$$b^* = \frac{1}{N_{vs}} \sum_{i=1}^{N_{vs}} (y_{vs} - < \boldsymbol{w}^*, \boldsymbol{x}_{vs} >) \tag{24}$$

Finalmente, haciendo uso de (17) en (23), o en (24), permitirá también calcular el valor de b^*en función de la solución del problema dual.

Obsérvese que tanto la optimización de (20) como la evaluación de (21) dependen del producto escalar de los vectores ejemplos. Esta propiedad se utilizará más tarde (sección 4) para calcular hiperplanos de separación óptimos en espacios transformados de alta dimensionalidad.

3. SVM para clasificación binaria de ejemplos cuasi-separables linealmente

El problema planteado en la sección anterior tiene escaso interés práctico porque los problemas reales se caracterizan normalmente por poseer ejemplos ruidosos y no ser perfecta y linealmente separables. La estrategia para este tipo de problemas reales es relajar el grado de separabilidad del conjunto de ejemplos, permitiendo que haya errores de clasificación en algunos de los ejemplos del conjunto de entrenamiento.Sin embargo, sigue siendo un objetivo el encontrar un hiperplano óptimo para el resto de ejemplos que sí son separables.

Desde el punto de vista de la formulación vista en la sección anterior, un ejemplo es no-separable si no cumple la condición (11). Aquí se pueden dar dos casos. En el primero, el ejemplo cae dentro del margen asociado a la clase correcta, de acuerdo a la frontera de decisión que define el hiperplano de separación. En el otro caso, el ejemplo cae al otro lado de dicho hiperplano. En ambos casos se dice que el ejemplo es no-separable, pero en el primer caso es clasificado de forma correcta y, en el segundo, no lo es (ver fig. 4).

La idea para abordar este nuevo problema es introducir, en la condición (11), que define al hiperplano de separación, un conjunto de variables reales positivas, denominadas *variables de holgura*, ξ_i, $i = 1, \ldots n$, que permitirán cuantificar el número de ejemplos no-separables que se está dispuesto a admitir, es decir:

$$y_i \left(< \boldsymbol{w}, \boldsymbol{x}_i > + b\right) \geq 1 - \xi_i, \qquad \xi_i \geq 0,\ i = 1, \ldots, n \tag{25}$$

Por tanto, para un ejemplo $(\boldsymbol{x}_i, y_i)$, su variable de holgura, ξ_i, representa la desviación del caso separable, medida desde el borde del margen que corresponde a la clase y_i (ver fig. 4). De acuerdo a esta definición, variables de holgura de valor cero corresponden a ejemplos separables, mayores que cero corresponden a ejemplos no separables y mayores que uno corresponden a ejemplos no separables y, además, mal clasificados. Por tanto, la suma de todas las variables de holgura, $\sum_{i=1}^{n} \xi_i$, permite, de alguna manera, medir el coste asociado al número de ejemplos no-separables. Así, en una primera aproximación, cuanto mayor sea el valor de esta suma, mayor será el número de ejemplos no separables.

Relajadas las restricciones, según (25), ya no basta con plantear como único objetivo maximizar el margen, ya que podríamos lograrlo a costa de clasificar erróneamente muchos ejemplos. Por tanto, la función a optimizar debe incluir, de alguna forma, los errores de clasificación que

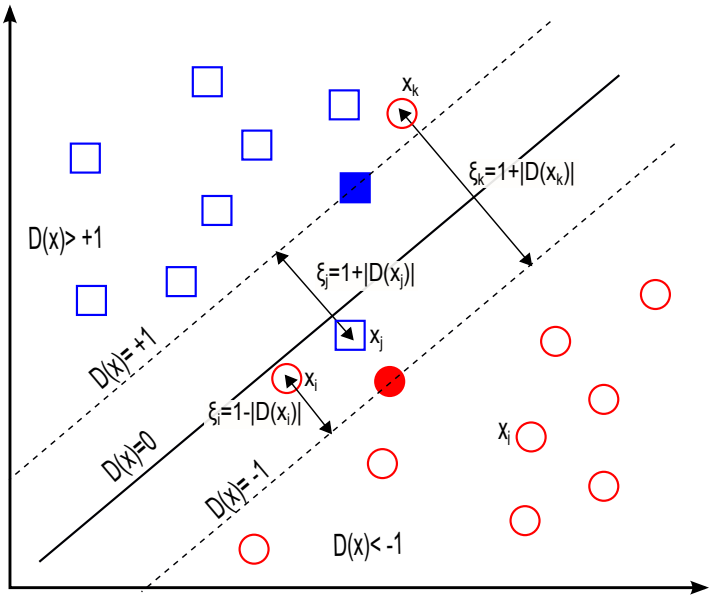


Figura 4: En el caso de ejemplos no-separables, las variables de holgura miden la desviación desde el borde del margen de la clase respectiva. Así, los ejemplos x_i, x_j y x_k son, cada uno de ellos, no-separables ($\xi_i, \xi_j, \xi_k > 0$. Sin embargo, x_i está correctamente clasificado, mientras que x_j y x_k están en el lado incorrecto de la frontera de decisión y, por tanto, mal clasificados.

está cometiendo el hiperplano de separación, es decir:

$$f(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i \tag{26}$$

donde C es una constante, suficientemente grande, elegida por el usuario, que permite controlar en qué grado influye el término del coste de ejemplos no-separables en la minimización de la norma, es decir, permitirá regular el compromiso entre el grado de sobreajuste del clasificador final y la proporción del número de ejemplos no separables. Así, un valor de C muy grande permitiría valores de ξ_i muy pequeños. En el límite ($C \to \infty$), estaríamos considerando el caso de ejemplos perfectamente separables ($\xi_i \to 0$). Por contra, un valor de C muy pequeño permitiría valores de ξ_i muy grandes, es decir, estaríamos admitiendo un número muy elevado de ejemplos mal clasificados. En el caso límite ($C \to 0$), se permitiría que todos los ejemplos estuvieran mal clasificados ($\xi_i \to \infty$).

En consecuencia, el nuevo problema de optimización consistirá en encontrar el hiperplano, definido por $\boldsymbol{w}$ y b, que minimiza el funcional (26) y sujeto a las restricciones dadas por (25), es decir

$$\begin{aligned} \text{mín} \quad & \tfrac{1}{2} < \boldsymbol{w}, \boldsymbol{w} > + C \textstyle\sum_{i=1}^{n} \xi_i \\ \text{s.a.} \quad & y_i \left(< \boldsymbol{w}, \boldsymbol{x}_i > + b \right) + \xi_i - 1 \geq 0 \\ & \xi_i \geq 0, \ i = 1, \ldots, n \end{aligned} \tag{27}$$

El hiperplano así definido recibe el nombre de *hiperplano de separación de margen blando* (del inglés *soft margen*), en oposición al obtenido en el caso perfectamente separable, también conocido como *hiperplano de separación de margen duro* (del inglés *hard margen*). Como en el caso de la sección anterior, si el problema de optimización a ser resuelto corresponde a un espacio de características de muy alta dimensionalidad, entonces, para facilitar su resolución,

puede ser transformado a su forma dual. El procedimiento para obtener el hiperplano de separación es similar al allí utilizado. Por tanto, aquí sólo se reproducirán de forma esquemática y secuencial los pasos necesarios para realizar dicha transformación.

Paso 1: Obtención de la función Lagrangiana[3]

$$L\left(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{1}{2} < \boldsymbol{w}, \boldsymbol{w} > + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left[y_i \left(< \boldsymbol{w}, \boldsymbol{x}_i > + b\right) + \xi_i - 1\right] - \sum_{i=1}^{n} \beta_i \xi_i \quad (28)$$

Paso 2: Aplicación de las condiciones de KKT:

$$\frac{\partial L}{\partial \boldsymbol{w}} \equiv \boldsymbol{w}^* - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = 0 \quad (29)$$

$$\frac{\partial L}{\partial b} \equiv \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (30)$$

$$\frac{\partial L}{\partial \xi_i} \equiv C - \alpha_i - \beta_i = 0 \quad (31)$$

$$\alpha_i \left[1 - y_i \left(< \boldsymbol{w}^*, \boldsymbol{x}_i > + b^*\right) - \xi_i\right] = 0,\ i = 1, \ldots, n \quad (32)$$

$$\beta_i \cdot \xi_i = 0,\ i = 1, \ldots, n \quad (33)$$

Paso 3: Establecer las relaciones entre las variables del problema primal $(\boldsymbol{w}, b, \boldsymbol{\xi})$ con las del problema dual $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Para ello, hacemos uso de la relacion (29):

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \quad (34)$$

Paso 4: Establecer restricciones adicionales de las variables duales. Para ello se hace uso de las relaciones (30-31):

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad (35)$$

$$C = \alpha_i + \beta_i \quad (36)$$

Paso 5: Del resultado obtenido en el paso 3, eliminar las variables primales de la función Lagrangiana para obtener así el problema dual que queremos maximizar:

$$L\left(\boldsymbol{\alpha}\right) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j < \boldsymbol{x}_i, \boldsymbol{x}_j >$$

Finalmente, se obtiene la formalización buscada del problema dual[4]:

$$\begin{array}{ll} \text{máx} & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j < \boldsymbol{x}_i, \boldsymbol{x}_j > \\ \text{s.a.} & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C,\ i = 1, \ldots, n \end{array} \quad (37)$$

[3] Obsérvese que, en este caso, aparecen dos familias de multiplicadores de Lagrange, $\alpha_i \geq 0$ y $\beta_i \geq 0$, con $i = 1, \ldots, n$, como consecuencia de las dos familias de restricciones que aparecen en (27). Nuevamente, el signo menos del tercer y cuarto sumando obedece a que las dos familias de restricciones en (27) están expresadas como restricciones del tipo $g(\boldsymbol{x}) \geq 0$ en lugar de $g(\boldsymbol{x}) \leq 0$.

[4] La restricción de que $\alpha_i \leq C$ se obtiene de (36) y de las condiciones $\alpha_i, \beta_i \geq 0$

Como en el caso anterior, la solución del problema dual nos permitirá expresar el hiperplano de separación óptima en términos de $\boldsymbol{\alpha}^*$. Para ello, bastará tener en cuenta dicha solución y sustituir la expresión (34) en (1), es decir:

$$D(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i < \boldsymbol{x}, \boldsymbol{x}_i > + b^* \tag{38}$$

Antes de obtener una expresión para el cálculo del valor de b^*, se considerarán algunos resultados interesantes. Así, de la restricción (36) es fácil deducir que si $\alpha_i = 0$, entonces $C = \beta_i$. De este último resultado y de la restricción (33) se deduce que $\xi_i = 0$. Por tanto, se puede afirmar que todos los ejemplos $\boldsymbol{x}_i$ cuyo α_i asociado sea igual a cero corresponden a ejemplos separables ($\xi_i = 0$).

Por otro lado, todo ejemplo no separable, x_i, se caracteriza por tener asociado un $\xi_i > 0$ (ver fig. 4). En este caso, y de la restricción (33), se deduce que $\beta_i = 0$. A su vez, de este último resultado y la restricción (36), se deduce que $\alpha_i = C$. Por tanto, se puede afirmar que todos los ejemplos $\boldsymbol{x}_i$ cuyo $\alpha_i = C$ corresponderán a ejemplos no-separables ($\xi_i > 0$). Además, dado que, en este caso, $\alpha_i \neq 0$, de la restricción (32) se deduce que

$$1 - y_i \left(< \boldsymbol{w}^*, \boldsymbol{x}_i > + b^*\right) - \xi_i = 0$$

es decir

$$1 - y_i D(\boldsymbol{x}_i) = \xi_i$$

Aquí se pueden considerar dos casos (ver fig. 4). En el primero, el ejemplo, $\boldsymbol{x}_i$, aunque no separable, está bien clasificado, es decir, $y_i D(\boldsymbol{x}_i) \geq 0$, entonces $\xi_i = 1 - |D(\boldsymbol{x}_i)|$. En el segundo caso, el ejemplo, $\boldsymbol{x}_i$, es no separable y está mal clasificado, es decir, $y_i D(\boldsymbol{x}_i) < 0$, entonces $\xi_i = 1 + |D(\boldsymbol{x}_i)|$.

Finalmente, consideremos el caso $0 < \alpha_i < C$. Así, en esta situación, la restricción (36) permite afirmar que $\beta_i \neq 0$ y, de este resultado y la restricción (33), se deduce que $\xi_i = 0$. Igualmente, si $0 < \alpha_i < C$, de la restricción (32) y del resultado obtenido anteriormente ($\xi_i = 0$), se deduce que

$$1 - y_i \left(< \boldsymbol{w}^*, \boldsymbol{x}_i > + b^*\right) = 0$$

Por tanto, se puede afirmar que un ejemplo, $\boldsymbol{x}_i$, es vector soporte si y solo si $0 < \alpha_i < C$.

De la expresión anterior, se esta en disposición de calcular el valor b^*, es decir

$$b^* = y_i - < \boldsymbol{w}^*, \boldsymbol{x}_i > \quad \forall i \text{ t.q. } 0 < \alpha_i < C \tag{39}$$

Obsérvese que, a diferencia del caso perfectamente separable, ahora, para el cálculo de b^*, no es suficiente con elegir cualquier ejemplo $\boldsymbol{x}_i$ que tenga asociado un $\alpha_i > 0$. Ahora, se habrá de elegir cualquier ejemplo $\boldsymbol{x}_i$ que tenga asociado un α_i que cumpla la restricción $0 < \alpha_i < C$.

Finalmente, haciendo uso de (34), es posible expresar b^* en términos de las variables duales:

$$b^* = y_i - \sum_{j=1}^{n} \alpha_i^* y_i < \boldsymbol{x}_j, \boldsymbol{x}_i > \quad \forall \alpha_i \text{ t.q. } 0 < \alpha_i < C \tag{40}$$

donde los coeficientes α_i^*, $i = 1, \ldots, n$, corresponden a la solución del problema dual.

A modo de resumen, en el caso de ejemplos cuasi-separables, hay dos tipos de ejemplos para los que los $\alpha_i^* \neq 0$. Aquellos, para los que $0 < \alpha_i^* < C$, que corresponderían a vectores soporte "normales" y, aquellos para los que $\alpha_i^* = C$, asociados a ejemplos no separables.

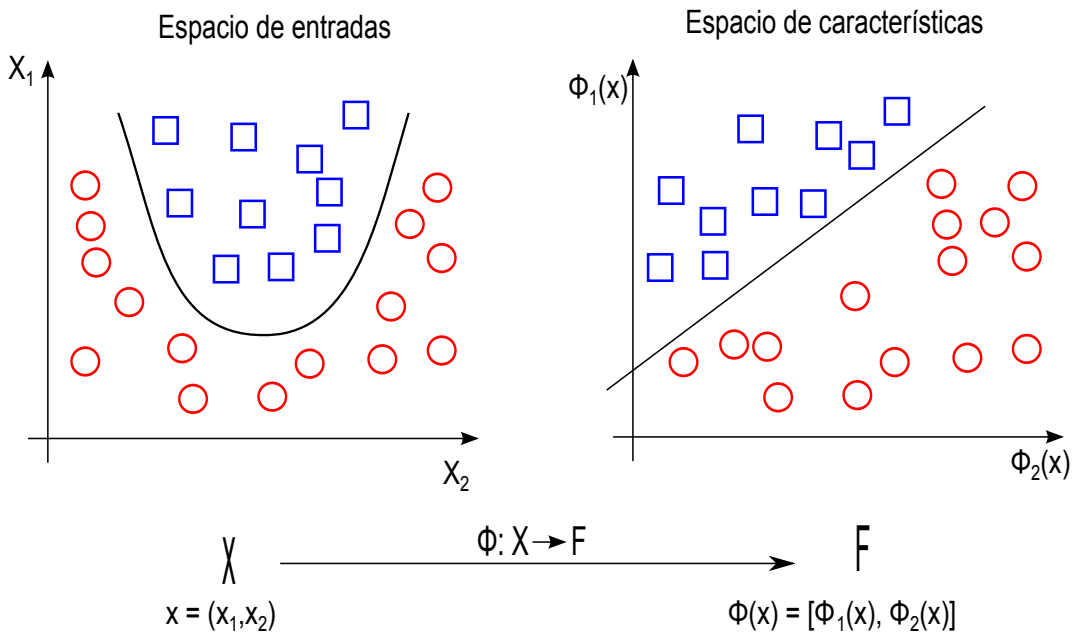


Figura 5: El problema de la búsqueda de una función de decisión no lineal en el espacio del conjunto de ejemplos original (espacio de entradas), se puede transformar en un nuevo problema consistente en la búsqueda de una función de decisión lineal (hiperplano) en un nuevo espacio transformado (espacio de características)

Esto últimos reciben el nombre de vectores soporte "acotados" (del inglés *bounded support vectors*). Ambos tipos de vectores (ejemplos) intervienen en la construcción del hiperplano de separación. El problema dual del caso cuasi-separable (37) y el correspondiente al caso perfectamente separable, (20), son prácticamente iguales. La única diferencia radica en la inclusión de la constante C en las restricciones del primero.

4. SVM para clasificación binaria de ejemplos no separables linealmente

En las dos secciones anteriores se ha mostrado que los hiperplanos de separación son buenos clasificadores cuando los ejemplos son perfectamente separables o cuasi-perfectamente separables. También se vio que el proceso de búsqueda de los parámetros que definen dichos hiperplanos se puede hacer independientemente de la dimensionalidad del problema a resolver. Así, si ésta es baja, basta con resolver directamente el problema de optimización primal asociado. En cambio, si la dimensionalidad es muy alta, basta con transformar el problema primal en su problema dual equivalente y resolver este último. Sin embargo, hasta ahora, se ha asumido la idea de que los ejemplos eran separables o cuasi-separables y, por tanto, los hiperplanos se definían como funciones lineales en el espacio-$\boldsymbol{x}$ de los ejemplos. En esta sección se describirá cómo usar de forma eficiente conjuntos de funciones base, no lineales, para definir espacios transformados de alta dimensionalidad y cómo buscar hiperplanos de separación óptimos en dichos espacios transformados. A cada uno de estos espacios se le denomina *espacio de características*, para diferenciarlo del espacio de ejemplos de entrada (espacio-$\boldsymbol{x}$).

Sea $\Phi : \mathbb{X} \rightarrow \mathcal{F}$ la función de transformación que hace corresponder cada vector de entrada $\boldsymbol{x}$ con un punto en el espacio de características $\mathcal{F}$, donde $\Phi(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x})]$ y

$\exists \phi_i(\boldsymbol{x})$, $i = 1, ..., m$, tal que $\phi_i(\boldsymbol{x})$ es una función no lineal. La idea entonces es construir un hiperplano de separación lineal en este nuevo espacio. La frontera de decisión lineal obtenida en el espacio de características se transformará en una frontera de decisión no lineal en el espacio original de entradas (ver fig. 5).

En este contexto, la función de decisión (1) en el espacio de características vendrá dada por[5]

$$D(\boldsymbol{x}) = (w_1\phi_1(\boldsymbol{x}) + ... + w_m\phi_m(\boldsymbol{x})) =< \boldsymbol{w}, \Phi(\boldsymbol{x}) > \tag{41}$$

y, en su forma dual, la función de decisión se obtiene transformando convenientemente la expresión de la frontera de decisión (38) en:

$$D(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i K(\boldsymbol{x}, \boldsymbol{x}_i) \tag{42}$$

donde $K(\boldsymbol{x}, \boldsymbol{x}')$ se denomina *función kernel*.

Por definición, una función kernel es una función $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ que asigna a cada par de elementos del espacio de entrada, $\mathbb{X}$, un valor real correspondiente al producto escalar de las imágenes de dichos elementos en un nuevo espacio $\mathcal{F}$ (espacio de características), es decir,

$$K(\boldsymbol{x}, \boldsymbol{x}') =< \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') >= \left(\phi_1(\boldsymbol{x})\phi_1(\boldsymbol{x}') + ... + \phi_m(\boldsymbol{x})\phi_m(\boldsymbol{x}')\right) \tag{43}$$

donde $\Phi : \mathbb{X} \to \mathcal{F}$.

Por tanto, una función kernel puede sustituir convenientemente el producto escalar en (38). Así, dado el conjunto de funciones base, $\Phi = \{\phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x})\}$, el problema a resolver en (42) sigue siendo encontrar el valor de los parámetros α_i^*, $i = 1, \ldots n$, que optimiza el problema dual (37), pero expresado ahora como:

$$\begin{array}{ll} \text{máx} & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{s.a.} & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \ i = 1, \ldots, n \end{array} \tag{44}$$

Se está ya en disposición de describir la metodología necesaria para resolver un problema de clasificación de ejemplos no separables linealmente. Concretamente, la función de decisión vendrá dada por la expresión (42), donde el valor de los parámetros $\alpha_i, i = 1, \ldots n$, se obtendrán como solución al problema de optimización cuadrática dado por (44), conocidos el conjunto de ejemplos de entrenamiento (x_i, y) , $i = 1, \ldots, n$, el kernel K, y el parámetro de regularización C. Actualmente, no existe una forma teórica de encontrar el valor de C. Sólo existe la heurística de usar un valor grande (recuérdese que $C = \infty$ para el caso linealmente separable).

A modo de ejemplo, supongamos el caso de vectores de entrada de dos dimensiones, $x = (x_1, x_2)$, y el conjunto de funciones base formado por todos los polinomios de grado tres, es decir,

$$\begin{array}{lll} \phi_1(x_1, x_2) = 1 & \phi_2(x_1, x_2) = x_1 & \phi_3(x_1, x_2) = x_2 \\ \phi_4(x_1, x_2) = x_1 x_2 & \phi_5(x_1, x_2) = x_1^2 & \phi_6(x_1, x_2) = x_2^2 \\ \phi_7(x_1, x_2) = x_1^2 x_2 & \phi_8(x_1, x_2) = x_1 x_2^2 & \phi_9(x_1, x_2) = x_1^3 \\ \phi_{10}(x_1, x_2) = x_2^3 & & \end{array}$$

En este caso, cada entrada de dos dimensiones es transformada en un espacio de características de diez dimensiones. La idea es entonces buscar un hiperplano en el espacio de características

[5]Obsérvese que se ha prescindido del termino b puesto que éste puede ser representado incluyendo en la base de funciones de transformación la función constante $\phi_1(\boldsymbol{x}) = 1$

que sea capaz de separar los ejemplos. La frontera de decisión lineal asociada a dicho hiperplano se transformará en un límite de decisión polinomial de grado tres en el espacio de entradas. Obsérvese también que si, en este ejemplo, un problema de tan solo dos dimensiones se transforma en uno de diez dimensiones, un pequeño aumento en la dimensionalidad del espacio de entrada puede provocar un gran aumento en la dimensionalidad del espacio de características. En el caso límite, existen incluso espacios de características de dimensión infinita. Es por esta razón por la que, ahora, el problema de optimización se expresa sólo en su forma dual, ya que, como se ha visto en las dos secciones anteriores, la solución de este problema no depende de la dimensionalidad del espacio sino de la cardinalidad del conjunto de vectores soporte.

Si la transformación del espacio de entradas al espacio de características puede definirse a partir de un conjunto infinito de funciones base, surge la pregunta de cómo transformar los ejemplos de entrada, de dimension finita, en otro espacio de dimensión infinita. El siguiente teorema responde a esta pregunta.

Teorema de Aronszajn. *Para cualquier función* $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ *que sea simétrica*[6] *y semidefinida positiva*[7]*, existe un espacio de Hilbert y una función* $\Phi : \mathbb{X} \to \mathcal{F}$ *tal que*

$$K(\boldsymbol{x}, \boldsymbol{x}') =< \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') > \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X} \tag{45}$$

Una consecuencia importante de este teorema es que para construir una función kernel no es necesario hacerlo a partir de un conjunto de funciones base, $\Phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x}))$, simplemente basta definir una función que cumpla las dos condiciones del teorema. Por tanto, para evaluar una función kernel no se necesitará conocer dicho conjunto de funciones base y, aún conocido éste, tampoco sería necesario realizar explícitamente el cálculo del producto escalar correspondiente, es decir, será suficiente con evaluar dicha función. En definitiva, para resolver el problema dual (44), no sólo no se necesita conocer el conjunto de funciones base de transformación, sino que tampoco es necesario conocer las coordenadas de los ejemplos transformados en el espacio de características. Sólo se necesitará conocer la forma funcional del kernel correspondiente, $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, aún cuando este pudiera estar asociado a un conjunto infinito de funciones base.

Ejemplos de funciones kernel

Se presentan aquí algunos ejemplos de funciones kernel:

- Kernel lineal:

$$K(\boldsymbol{x}, \boldsymbol{x}') =< \boldsymbol{x}, \boldsymbol{x}' > \tag{46}$$

- kernel polinómico de grado-p:

$$K_p(\boldsymbol{x}, \boldsymbol{x}') = \left[\gamma < \boldsymbol{x}, \boldsymbol{x}' > +\tau\right]^p \tag{47}$$

- kernel gaussiano:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2\right), \ \gamma > 0 \tag{48}$$

[6] *Una función* $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ *es simétrica si* $K(\boldsymbol{x}, \boldsymbol{x}') = K(\boldsymbol{x}', \boldsymbol{x}) \, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}$

[7] *Una función* $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ *es semidefinida positiva si* $\sum_{i=1}^{n} \sum_{i=1}^{n} c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$*, para cualesquiera conjuntos* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{X}$ *y* $c_1, \ldots c_n \in \mathbb{R}$*, siendo* $n > 0$.

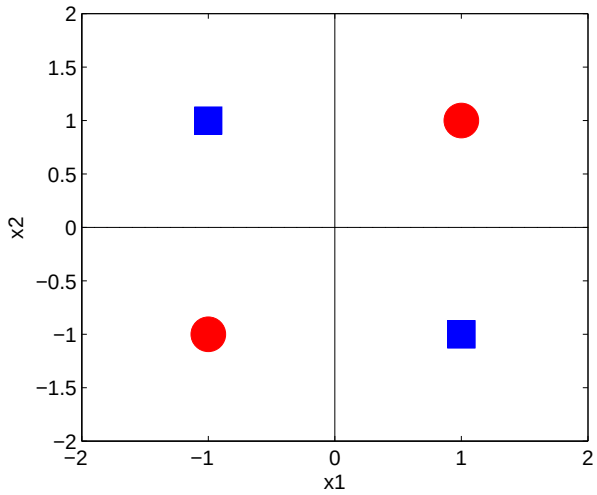


Figura 6: Representación del conjunto de datos perteneciente al problema XOR.

- kernel sigmoidal:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \tanh(\gamma < \boldsymbol{x}, \boldsymbol{x}' > +\tau) \tag{49}$$

A los parámetros γ, τ y p se les denomina parámetros del kernel.

Ejemplo: Solución del problema OR-exclusivo mediante SVMs

El problema or-exclusivo pertenece al caso de problemas separables no-linealmente ($C = \infty$) y se define como el problema de encontrar un hiperplano de separación que clasifique sin error los ejemplos de la tabla siguiente:

Ejemplo	(x_1, x_2)	y
1	$(+1, +1)$	$+1$
2	$(-1, +1)$	-1
3	$(-1, -1)$	$+1$
4	$(+1, -1)$	-1

De la fig. 6, resulta obvio que es imposible resolver este problema con un límite de decisión lineal en el espacio original de entradas. La solución que se propone es crear un clasificador SVM, usando un kernel polinómico 47, con $p = 2$, $\gamma = 1$ y $\tau = 1$:

$$K_2(\boldsymbol{x}, \boldsymbol{x}') = \left[< \boldsymbol{x}, \boldsymbol{x}' > +1\right]^2 \tag{50}$$

Los valores de α_i^*, $i = 1, \ldots n$, corresponderán a la solución del problema dual (44), particularizado para el problema que queremos resolver, es decir

$$\text{máx} \qquad \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{4} \alpha_i \alpha_j y_i y_j K_2(\boldsymbol{x}, \boldsymbol{x}_i)$$

$$s.a. \qquad \sum_{i=1}^{4} \alpha_i y_i = 0, \qquad 0 \leq \alpha_i \leq C,\ i = 1, \ldots, 4$$

La solución a este problema de optimización es $\alpha_i^* = 0{,}125$, $i = 1, \ldots 4$. Dado que no existe ningún i para el que $\alpha_i^* = 0$, se puede afirmar que todos los ejemplos del conjunto de entrenamiento corresponden a vectores soporte. Por tanto, la función de decisión vendrá dada por (42), particularizada para la solución obtenida y el kernel elegido, es decir:

$$D(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i K(\boldsymbol{x}, \boldsymbol{x}_i) = 0{,}125 \sum_{i=1}^{4} y_i K_2(\boldsymbol{x}, \boldsymbol{x}_i) \tag{51}$$

Obsérvese que con esta expresión habríamos resuelto el problema de clasificación planteado inicialmente, es decir, bastaría evaluarla con cualquier ejemplo (de clasificación conocida o desconocida) y asignarle la clase correspondiente, de acuerdo a lo indicado en (2). Sin embargo, aprovecharemos el problema XOR para obtener otros resultados relacionados con diferentes conceptos descritos anteriormente. Así, por ejemplo, de la definición de función kernel (43) y del kernel aquí empleado (50), es posible obtener el conjunto base de funciones de transformación:

$$K(\boldsymbol{x}, \boldsymbol{x}') = < \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') >= \left[< \boldsymbol{x}, \boldsymbol{x}' > +1 \right]^2 =$$

$$\left[< (x_1, x_2), (x_1', x_2') > +1 \right] =$$

$$x_1^2 (x_1')^2 + x_2^2 (x_2')^2 + 2x_1 x_2 x_1' x_2' + 2x_1 x_1' + 2x_2 x_2' + 1 =$$

$$< \left(1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2\right), \left(1, \sqrt{2}x_1', \sqrt{2}x_2', , \sqrt{2}x_1' x_2', (x_1')^2, (x_2')^2\right) >$$

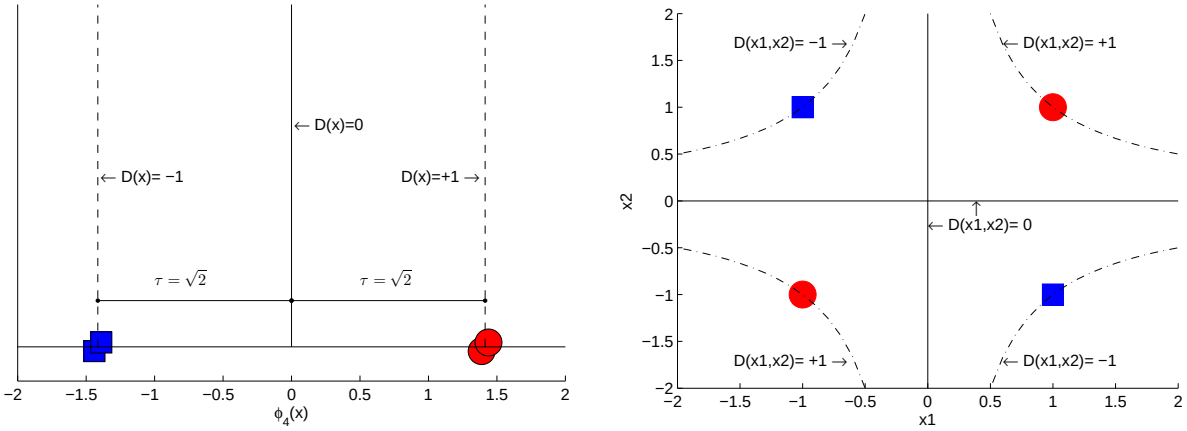
es decir, $\Phi_2 = \{\phi_1(\boldsymbol{x}), \ldots, \phi_6(\boldsymbol{x})\}$, donde

$$\begin{array}{lll} \phi_1(x_1, x_2) = 1 & \phi_2(x_1, x_2) = \sqrt{2}x_1 & \phi_3(x_1, x_2) = \sqrt{2}x_2, \\ \phi_4(x_1, x_2) = \sqrt{2}x_1 x_2 & \phi_5(x_1, x_2) = x_1^2 & \phi_6(x_1, x_2) = x_2^2 \end{array} \tag{52}$$

Utilizando este resultado y el obtenido en (53), la función de decisión lineal en el espacio transformado puede expresarse en función del conjunto de funciones base:

$$\begin{array}{c} D(\boldsymbol{x}) = 0{,}125 \cdot \sum_{i=1}^{4} y_i K_2(\boldsymbol{x}, \boldsymbol{x}_i) = 0{,}125 \cdot \sum_{i=1}^{4} y_i < \Phi_2(\boldsymbol{x}), \Phi_2(\boldsymbol{x}_i) >= \\ 0{,}125 \left[\phi_1(\boldsymbol{x}) + \sqrt{2}\phi_2(\boldsymbol{x}) + \sqrt{2}\phi_3(\boldsymbol{x}) + \sqrt{2}\phi_4(\boldsymbol{x}) + \phi_5(\boldsymbol{x}) + \phi_6(\boldsymbol{x}) + \right. \\ (-\phi_1(\boldsymbol{x})) + \sqrt{2}\phi_2(\boldsymbol{x}) - \sqrt{2}\phi_3(\boldsymbol{x}) + \sqrt{2}\phi_4(\boldsymbol{x}) - \phi_5(\boldsymbol{x}) - \phi_6(\boldsymbol{x}) + \\ \phi_1(\boldsymbol{x}) - \sqrt{2}\phi_2(\boldsymbol{x}) - \sqrt{2}\phi_3(\boldsymbol{x}) + \sqrt{2}\phi_4(\boldsymbol{x}) + \phi_5(\boldsymbol{x}) + \phi_6(\boldsymbol{x}) + \\ \left. (-\phi_1(\boldsymbol{x})) - \sqrt{2}\phi_2(\boldsymbol{x}) + \sqrt{2}\phi_3(\boldsymbol{x}) + \sqrt{2}\phi_4(\boldsymbol{x}) - \phi_5(\boldsymbol{x}) - \phi_6(\boldsymbol{x}) \right] = \\ 0{,}125 \left[4\sqrt{2}\phi_4(\boldsymbol{x}) \right] = \frac{1}{\sqrt{2}} \cdot \phi_4(\boldsymbol{x}) \end{array} \tag{53}$$

Del resultado obtenido, se puede afirmar que, de las seis dimensiones del espacio de características, la función lineal de decisión en dicho espacio se expresa en términos de sólo una de ellas, $\phi_4(\boldsymbol{x})$. Es decir, sólo se necesita una dimensión del espacio transformado para poder separar los ejemplos del conjunto de entrenamiento original (ver fig. 7a). Este hecho se confirma al calcular los ejemplos transformados de los ejemplos originales en el nuevo espacio de características, mostrados en la siguiente tabla:



(a) (b)

Figura 7: Solución al problema XOR: (a) hiperplano de separación en el espacio de características, junto con su margen asociado (los cuatro ejemplos son vectores soporte) (b) función de decisión no lineal en el espacio de ejemplos original resultante de transformar el hiperplano obtenido en (a) en coordenadas del espacio original.

Ejemplo #	Espacio de entradas (x_1, x_2)	Espacio de características $(\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots, \phi_6(\boldsymbol{x}))$	Clase y
1	$(+1, +1)$	$\left(1, \sqrt{2}, \sqrt{2}, \sqrt{2}, 1, 1\right)$	$+1$
2	$(-1, +1)$	$\left(1, -\sqrt{2}, \sqrt{2}, -\sqrt{2}, 1, 1\right)$	-1
3	$(-1, -1)$	$\left(1, -\sqrt{2}, -\sqrt{2}, \sqrt{2}, 1, 1\right)$	$+1$
4	$(+1, -1)$	$\left(1, \sqrt{2}, -\sqrt{2}, -\sqrt{2}, 1, 1\right)$	-1

Para obtener la ecuación del hiperplano de separación en el espacio de características, bastará hacer $D(\boldsymbol{x}) = 0$ en (53), es decir:

$$\frac{1}{\sqrt{2}}\phi_4(\boldsymbol{x}) = 0 \quad \Rightarrow \quad \phi_4(\boldsymbol{x}) = 0$$

y para obtener las ecuaciones de las fronteras que delimitan el margen, habrá que calcular $D(\boldsymbol{x}) = +1$ y $D(\boldsymbol{x}) = -1$, es decir:

$$\begin{aligned} \phi_4(\boldsymbol{x}) &= +\sqrt{2} \\ \phi_4(\boldsymbol{x}) &= -\sqrt{2} \end{aligned}$$

De la fig. 7a, resulta fácil deducir que el valor del margen máximo es $\tau = \sqrt{2}$. No obstante, el valor de dicho margen máximo se puede calcular matemáticamente. Para ello, bastará calcular el valor de $\|\boldsymbol{w}^*\|$ y aplicar (9). A su vez, el valor de $\boldsymbol{w}^*$ se puede obtener a partir de (17), conocidos los valores de α_i^*, $i = 1, \ldots 4$, es decir,

$$\boldsymbol{w}^* = \sum_{i=1}^{4} \alpha_i^* y_i \boldsymbol{x}_i = \left(0, 0, 0, \frac{1}{\sqrt{2}}, 0, 0\right)$$

donde los valores de $\boldsymbol{x}_i$, $i = 1, \ldots 4$ corresponden a los valores de los ejemplos de entrenamiento expresados respecto a las coordenadas del espacio de características. Del resultado anterior, resulta inmediato obtener el valor del margen máximo:

$$\tau_{máx} = \frac{1}{\|w^*\|} = \sqrt{2}$$

También es fácil obtener la función de decisión no lineal en el espacio original de entradas a partir transformar de la función lineal de decisión del espacio de características (ver fig. 7b). Para ello, basta sustituir el valor de $\phi_4(x)$, obtenido de (52), en (53), es decir

$$D(\boldsymbol{x}) = x_1 x_2$$

Así, las ecuaciones de las fronteras de separación vendrán dadas por $D(x) = 0$, es decir

$$x_1 x_2 = 0 \quad \Rightarrow \quad \begin{cases} x_1 = 0 \\ \\ x_2 = 0 \end{cases}$$

y la de las fronteras que delimitan los márgenes por $D(x) = +1$ y $D(x) = -1$, es decir

$$x_1 x_2 = +1 \quad \Rightarrow \quad x_2 = 1/x_1$$

$$x_1 x_2 = -1 \quad \Rightarrow \quad x_2 = -1/x_1$$

5. SVM para regresión

Las máquinas de vectores soporte pueden también adaptarse para resolver problemas de regresión. En estos casos, es muy común designarlas por el acrónimo SVR (del inglés *Support Vector Regression*). Así, dado un conjunto de ejemplos de entrenamiento $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, donde $\boldsymbol{x}_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$, en el que se asume que los valores y_i de todos los ejemplos de S se pueden ajustar (o cuasi-ajustar) mediante una función lineal, el objetivo de la tarea de regresión es encontrar los parámetros $w = (w_1, \ldots, w_d)$ que permitan definir dicha función lineal:

$$f(\boldsymbol{x}) = (w_1 x_1 + ... + w_d x_d) + b =< \boldsymbol{w}, \boldsymbol{x} > +b \tag{54}$$

Para permitir cierto ruido en los ejemplos de entrenamiento se puede relajar la condición de error entre el valor predicho por la función y el valor real. Para ello, se utiliza la denominada *función de pérdida ϵ-insensible*, L_ϵ, (ver fig. 8) caracterizada por ser una función lineal con una zona insensible, de anchura 2ϵ, en la que el error es nulo, y viene definida por:

$$L_\epsilon\left(y, f(\boldsymbol{x})\right) = \begin{cases} 0 & \text{si } |y - f(\boldsymbol{x})| \leq \epsilon \\ \\ |y - f(\boldsymbol{x})| - \epsilon & \text{en otro caso} \end{cases} \tag{55}$$

La principal razón para elegir esta función es la de permitir cierta dispersión en la función solución, de tal forma que todos los ejemplos que quedan confinados en la región *tubular* definida por $\pm\epsilon$ no serán considerados vectores soporte. De esta forma se reducirá significativamente el número de éstos.

Dado que en la práctica es muy difícil que los ejemplos de entrenamiento se ajusten al modelo lineal con un error de predicción igual a cero, se recurre al concepto de margen blando, ya utilizado anteriormente al resolver el problema de clasificación. Para ello, se definen dos variables de holgura, ξ_i^+ y ξ_i^-, que permitirán cuantificar la magnitud de dicho error (ver fig. 8). Así, la variable $\xi_i^+ > 0$ cuando la predicción del ejemplo, $f(\boldsymbol{x}_i)$ es mayor que su valor real, y_i, en una cantidad superior a ϵ, es decir, $f(\boldsymbol{x}_i) - y_i > \epsilon$. En otro caso, su valor será cero. De forma similar, la variable $\xi_i^- > 0$ cuando el valor real del ejemplo es mayor que su predicción en una cantidad superior a ϵ, es decir, $y_i - f(\boldsymbol{x}_i) > \epsilon$. En otro caso, su valor será cero. Dado

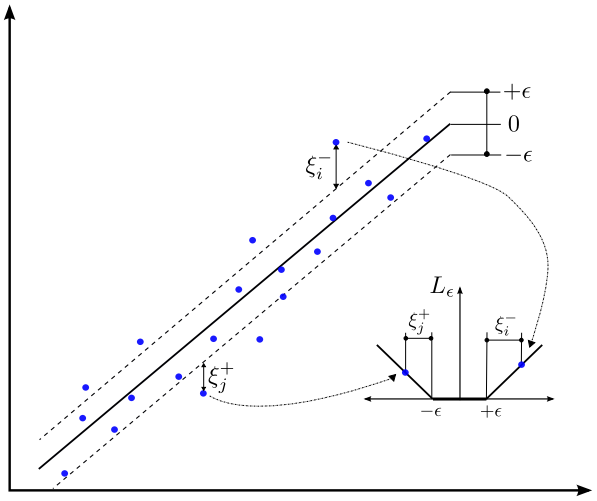


Figura 8: SVR con margen blando: se muestra la relación entre las variables de holgura, ξ_i^-, ξ_j^+, asociadas a ejemplos que quedan fuera de la zona tubular ϵ-insensible y la función de pérdida, L_ϵ.

que no puede ocurrir simultáneamente que la predicción de un ejemplo sea simultáneamente mayor ($\xi_i^+ > 0$) y menor ($\xi_i^- > 0$) que su valor real, se puede afirmar que $\xi_i^+ \cdot \xi_i^- = 0$.

Tal y como ocurría en el problema de clasificación con margen blando, aquí también la suma de todas las variables de holgura permitirá, de alguna manera, medir el coste asociado al número de ejemplos con un error de predicción no nulo. Por tanto, la función a optimizar será la misma que la del problema de clasificación con margen blando (26), con la salvedad de que aquí tenemos dos tipos de variables de holgura. En definitiva, el problema primal, en el caso de regresión, queda definido como:

$$\begin{array}{ll} \text{mín} & \frac{1}{2} < \boldsymbol{w}, \boldsymbol{w} > + C \sum_{i=1}^{n} \left(\xi_i^+ + \xi_i^-\right) \\ \\ \text{s.a.} & (< \boldsymbol{w}, \boldsymbol{x}_i > + b) - y_i - \epsilon - \xi_i^+ \leq 0 \\ & y_i - (< \boldsymbol{w}, \boldsymbol{x}_i > + b) - \epsilon - \xi_i^- \leq 0 \\ & \xi_i^+, \xi_i^- \geq 0, \ i = 1, \ldots, n \end{array} \tag{56}$$

La transformación al problema dual requiere los mismos pasos que se han seguido hasta ahora en secciones anteriores, es decir:

Paso 1: Obtención de la función Lagrangiana

$$\begin{array}{rl} L\left(\boldsymbol{w}, b, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-, \boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-, \boldsymbol{\beta}^+, \boldsymbol{\beta}^-\right) = & \frac{1}{2} < \boldsymbol{w}, \boldsymbol{w} > + C \sum_{i=1}^{n} \sum_{i=1}^{n} \left(\xi_i^+ + \xi_i^-\right) + \\ & \sum_{i=1}^{n} \alpha_i^+ \left[(< \boldsymbol{w}, \boldsymbol{x}_i > + b) - y_i - \epsilon - \xi_i^+\right] + \\ & \sum_{i=1}^{n} \alpha_i^- \left[y_i - (< \boldsymbol{w}, \boldsymbol{x}_i > + b) - \epsilon - \xi_i^-\right] - \\ & \sum_{i=1}^{n} \beta_i^+ \xi_i^+ - \sum_{i=1}^{n} \beta_i^- \xi_i^- \end{array} \tag{57}$$

Paso 2: Aplicación de las condiciones de KKT:

$$\frac{\partial L}{\partial \boldsymbol{w}} \equiv \boldsymbol{w} + \sum_{i=1}^{n} \alpha_i^+ \boldsymbol{x}_i - \sum_{i=1}^{n} \alpha_i^- \boldsymbol{x}_i = 0 \tag{58}$$

$$\frac{\partial L}{\partial b} \equiv \sum_{i=1}^{n} \alpha_i^+ - \sum_{i=1}^{n} \alpha_i^- = 0 \tag{59}$$

$$\frac{\partial L}{\partial \xi_i^+} \equiv C - \alpha_i^+ - \beta_i^+ = 0 \tag{60}$$

$$\frac{\partial L}{\partial \xi_i^-} \equiv C - \alpha_i^- - \beta_i^- = 0 \tag{61}$$

$$\alpha_i^+ \left[(< \boldsymbol{w}^*, \boldsymbol{x}_i > + b*) - y_i - \epsilon - \xi_i^+ \right] = 0 \tag{62}$$

$$\alpha_i^- \left[y_i - (< \boldsymbol{w}^*, \boldsymbol{x}_i > + b^*) - \epsilon - \xi_i^- \right] = 0 \tag{63}$$

$$\beta_i^+ \xi_i^+ = 0 \tag{64}$$

$$\beta_i^- \xi_i^- = 0 \tag{65}$$

Paso 3: Establecer las relaciones entre las variables del problema primal $\left(\boldsymbol{w}, b, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-\right)$ con las del problema dual $\left(\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-, \boldsymbol{\beta}^+, \boldsymbol{\beta}^-\right)$. Para ello, se hace uso de (58):

$$\boldsymbol{w} = \sum_{i=1}^{n} \left(\alpha_i^- - \alpha_i^+\right) \boldsymbol{x}_i \tag{66}$$

Paso 4: Establecer restricciones adicionales de las variables duales. Para ello se hace uso de (59)-(61), es decir:

$$\sum_{i=1}^{n} \left(\alpha_i^+ - \alpha_i^-\right) = 0 \tag{67}$$

$$\beta_i^+ = C - \alpha_i^+ \tag{68}$$

$$\beta_i^- = C - \alpha_i^- \tag{69}$$

Paso 5: Del resultado obtenido en el paso 3, eliminar las variables primales de la función Lagrangiana:

$$\begin{aligned} L\left(\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-\right) = \quad & \textstyle\sum_{i=1}^{n} \left(\alpha_i^- - \alpha_i^+\right) y_i - \epsilon \sum_{i=1}^{n} \left(\alpha_i^- + \alpha_i^+\right) - \\ & \tfrac{1}{2} \textstyle\sum_{i,j=1}^{n} \left(\alpha_i^- - \alpha_i^+\right) \left(\alpha_j^- - \alpha_j^+\right) < \boldsymbol{x}_i, \boldsymbol{x}_j > \end{aligned} \tag{70}$$

Finalmente, se obtiene la formalización buscada del problema dual[8]:

$$\begin{aligned} \text{máx} \quad & \textstyle\sum_{i=1}^{n} \left(\alpha_i^- - \alpha_i^+\right) y_i - \epsilon \sum_{i=1}^{n} \left(\alpha_i^- + \alpha_i^+\right) - \\ & \tfrac{1}{2} \textstyle\sum_{i,j=1}^{n} \left(\alpha_i^- - \alpha_i^+\right) \left(\alpha_j^- - \alpha_j^+\right) < \boldsymbol{x}_i, \boldsymbol{x}_j > \\ \text{s.a.} \quad & \textstyle\sum_{i=1}^{n} \left(\alpha_i^+ - \alpha_i^-\right) = 0 \\ & 0 \leq \alpha_i^+, \alpha_i^- \leq C, \; i = 1, \ldots, n \end{aligned} \tag{71}$$

El regresor asociado a la función lineal buscada resulta ser

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \left(\alpha_i^- - \alpha_i^+\right) < \boldsymbol{x}, \boldsymbol{x}_i > + b^* \tag{72}$$

[8]La restricción de que $\alpha_i^+ \leq C$ se obtiene de (68) y de $\alpha_i^+, \beta_i^+ \geq 0$. Igualmente, la restricción $\alpha_i^- \leq C$ se obtiene de (69) y de $\alpha_i^-, \beta_i^- \geq 0$

La obtención del valor de b^* ya no es tan inmediata como en el caso de clasificación. Para ello, se hace uso de las restricciones obtenidas como resultado de aplicar la segunda condición KKT (62)-(65). Utilizando las expresiones (68)-(69), las dos últimas se pueden reescribir como:

$$\left(C - \alpha_i^+\right) \xi_i^+ = 0 \tag{73}$$

$$\left(C - \alpha_i^-\right) \xi_i^- = 0 \tag{74}$$

De (62)-(63), se deduce que $\alpha_i^+ \alpha_i^- = 0$, es decir, ambas variables no pueden ser simultáneamente distintas de cero. De lo contrario, se obtendrían dos valores diferentes de b^* para cada una de las ecuaciones (62)-(63). Por otro lado, a partir de (73)-(74), se puede afirmar que si un ejemplo $(\boldsymbol{x}_i, y_i)$ esta fuera de la zona tubular ϵ-insensible, es decir, $\xi_i^- = 0$ y $\xi_i^+ > 0$ o $\xi_i^- > 0$ y $\xi_i^+ = 0$, entonces $\alpha_i^+ = C$ (primer caso) o $\alpha_i^- = C$ (segundo caso). Estas deducciones permiten, a su vez, concluir que:

$$\begin{aligned} &(< \boldsymbol{w}^*, \boldsymbol{x}_i > + b^*) - y_i - \epsilon \leq 0 \;\text{ y }\; \xi_i^+ = 0 \quad \text{si } \alpha_i^+ < C \\ &(< \boldsymbol{w}^*, \boldsymbol{x}_i > + b^*) - y_i - \epsilon \geq 0 \qquad\qquad\quad \text{si } \alpha_i^+ > 0 \end{aligned} \tag{75}$$

o lo que es lo mismo:

$$y_i - < \boldsymbol{w}^*, \boldsymbol{x}_i > + \epsilon \leq b^* \leq y_i - < \boldsymbol{w}^*, \boldsymbol{x}_i > + \epsilon \quad \text{si } 0 < \alpha_i^+ C \tag{76}$$

es decir:

$$b^* = y_i - < \boldsymbol{w}^*, \boldsymbol{x}_i > + \epsilon \quad \text{si } 0 < \alpha_i^+ < C \tag{77}$$

Trabajando de forma análoga para α_i^-, se puede obtener que

$$b^* = y_i - < \boldsymbol{w}^*, \boldsymbol{x}_i > - \epsilon \quad \text{si } 0 < \alpha_i^- < C \tag{78}$$

Obsérvese que el valor de b^* siempre será único porque las condiciones asociadas a las expresiones (77) y (78) no pueden ser ciertas simultáneamente, ya que $\alpha_i^+ \alpha_i^- = 0$. Así, si se cumple la condición de la primera expresión, es decir, $0 < \alpha_i^+ < C$, entonces $\alpha_i^- = 0$ y, por tanto, no se cumplirá la de la segunda. De la misma forma, si se cumpliera la condición de la segunda expresión ($0 < \alpha_i^- < C$), entonces $\alpha_i^+ = 0$ y no podría cumplirse la condición de la primera.

Kernelización de las SVR

En el caso de que los ejemplos no puedan ajustarse por una función lineal, se recurre a una metodología similar a la utilizada en el problema de clasificación para ejemplos no separables linealmente. Es decir, los ejemplos pertenecientes al espacio original de entradas se transforman en un nuevo espacio, denominado también espacio de características, en el que sí es posible ajustar los ejemplos transformados mediante un regresor lineal. El tipo de transformación dependerá del kernel utilizado. El regresor asociado a la función lineal en el nuevo espacio es:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \left(\alpha_i^- - \alpha_i^+\right) K(\boldsymbol{x}, \boldsymbol{x}_i) \tag{79}$$

Obsérvese que se prescinde del término b^* puesto que éste puede ser representado mediante la inclusión de una función constante en el conjunto de funciones base como, por ejemplo,

$\phi(\boldsymbol{x}) = 1$. Los coeficientes α_i^-, α_i^+ se obtienen como resultado de resolver el problema dual, expresado ahora como:

$$\begin{array}{ll} \text{máx} & \sum_{i=1}^n \left(\alpha_i^- - \alpha_i^+\right) y_i - \epsilon \sum_{i=1}^n \left(\alpha_i^- + \alpha_i^+\right) - \\ & \frac{1}{2} \sum_{i,j=1}^n \left(\alpha_i^- - \alpha_i^+\right) \left(\alpha_j^- - \alpha_j^+\right) K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{s.a.} & \sum_{i=1}^n \left(\alpha_i^+ - \alpha_i^-\right) = 0 \\ & 0 \leq \alpha_i^+, \alpha_i^- \leq C,\ i = 1, \ldots, n \end{array} \tag{80}$$

que no es más que el problema dual (71), en el que los productos escalares se sustituyen por funciones kernel.

A modo de resumen, puede decirse que para resolver problemas de regresión mediante SVRs hay que seleccionar, además del kernel más adecuado (en el caso de regresión no lineal), tanto ϵ como C. Ambos parámetros afectan a la complejidad del modelo. En el caso de problemas de regresión con ruido, el parámetro ϵ debería ser elegido de forma que refleje la varianza del ruido de los datos. En la mayoría de casos prácticos es posible obtener una medida aproximada de la varianza del ruido a partir de los datos de entrenamiento. Para problemas de regresión sin ruido (problemas de interpolación) el valor ϵ corresponde a la exactitud preestablecida de interpolación, de forma que, cuanto mayor sea el valor de ϵ, menor número de vectores soporte se necesitarán, y viceversa. Por otro lado, la metodología usada para seleccionar el valor de C más adecuado, se basa normalmente en técnicas de validación cruzada.

6. Software sobre SVMs

En la actualidad existe un número importante de repositorios web y de paquetes software de libre distribución dedicados a la implementación de SVMs y muchas de sus variantes. En esta sección vamos a describir algunos de estos paquetes software.

LIBSVM

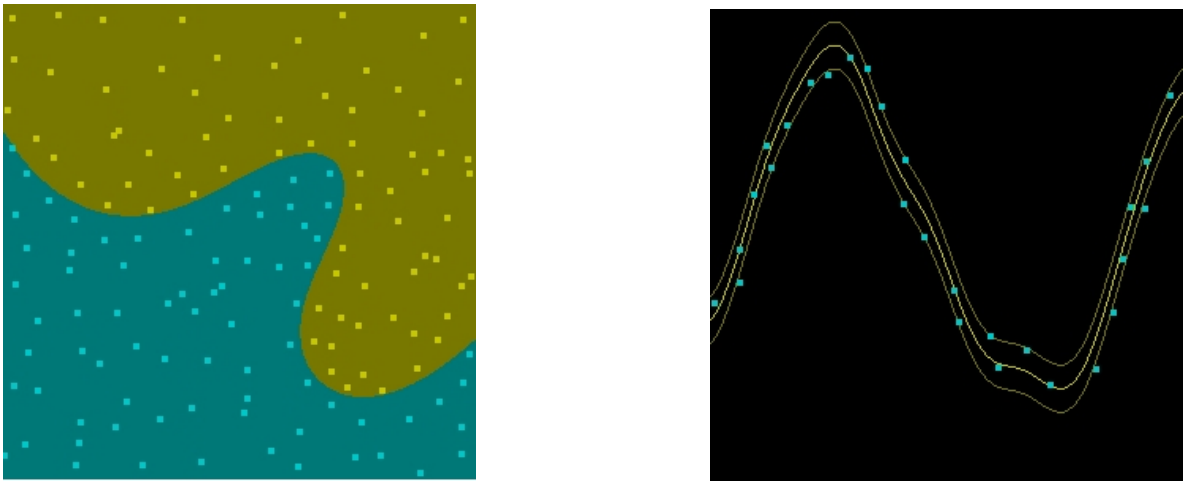
Enlace: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

La librería LIBSVM es un paquete software pensado para resolver problemas de clasificación y regresión mediante máquinas de vectores soporte. Entre sus principales características, cabe citar que es un software de código abierto (disponible en C++ y Java); implementa diferentes formulaciones SVM con la posibilidad de usar diferentes tipos de kernels; permite la clasificación multiclase y la posibilidad de usar técnicas de validación cruzada para la selección de modelos. También ofrece interfaces para una gran cantidad de lenguajes de programación (Python, R, MATLAB, Perl, Ruby, Weka, Common LISP, CLISP, Haskell, OCaml, LabVIEW, y PHP). Además, en su página web dispone de un *applet* para implementar sencillos problemas de clasificación y de regresión en dos dimensiones. La figura 9 muestra las soluciones de un ejemplo de problema de clasificación binaria (fig. 9a) y otro de regresión (fig. 9b), construidos ambos y resueltos mediante dicho applet.

SVM^{light}

Enlace: http://svmlight.joachims.org/

SVM^{light} es una implementación en C de máquinas de vectores soporte. Entre sus principales características destaca permitir resolver no sólo problemas de clasificación y de regresión,



(a) (b)

Figura 9: Ejemplos de salida del applet de la página web del software LIBSVM: (a) frontera de decisión no-lineal en un problema de clasificación binaria; (b) regresor no-lineal del conjunto de ejemplos mostrado en la figura.

sino también problemas de ranking; permite manejar varios cientos de miles de ejemplos de entrenamiento, junto con muchos miles de vectores soporte; soporta funciones kernel estándar y, además, permite al usuario definir sus propias funciones kernel. Como novedad presenta una implementación SVM, denominada SVM^{Struct}, para la predicción de salidas estructuradas o multivariable, tales como conjuntos, secuencias y árboles.

7. Anexo

El modelo matemático asociada a las SVMs da lugar a problemas de optimización con restricciones lineales. Este tipo de problemas se resuelven haciendo uso de la teoría de la optimización. En este anexo se hace un resumen de las principales ideas de esta teoría, orientado a la resolución de problemas asociados con el uso de SVMs.

Sea el siguiente problema de optimización denominado *problema primal*:

$$\begin{aligned} &\text{mín} \quad f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega \\ &\text{s.a.} \quad g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, n \end{aligned} \tag{81}$$

Si todas las funciones f y g_i fueran lineales estaríamos ante un problema de optimización de programación lineal. En cambio, si la función a optimizar es cuadrática y las restricciones siguen siendo lineales, estaríamos ante un problema de optimización cuadrática. La solución del problema primal, x^*, cumplirá que $g_i(\boldsymbol{x}^*) \leq 0$ y $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$, $\forall x$ t.q. $g_i(\boldsymbol{x}) \leq 0$, donde $i = 1, ..., n$.

Se define la función de Lagrange como:

$$L(\boldsymbol{x}, \boldsymbol{\alpha}) = f(\boldsymbol{x}) + \sum_{i=1}^{n} \alpha_i g_i(\boldsymbol{x}) \tag{82}$$

donde los coeficientes $\alpha_i \geq 0$ reciben el nombre de multiplicadores de Lagrange y, de forma intuitiva, indican la dificultad de cumplir cada restricción, es decir, a mayor valor de α_i, más

difícil sera de cumplir su restricción asociada g_i. La función lagrangiana tiene la particularidad de incorporar la función objetivo y las funciones restricción en una única función.

A partir de la función de Lagrange se puede definir el *problema dual* como:

$$\begin{aligned} \text{máx} \quad & \varphi(\boldsymbol{\alpha}) = \inf_{\boldsymbol{x}\in\Omega} L(\boldsymbol{x}, \boldsymbol{\alpha}) \\ \text{s.a.} \quad & \alpha_i(\boldsymbol{x}) \geq 0, \quad i = 1, \ldots, n \end{aligned} \tag{83}$$

El interés del problema dual es que, bajo determinadas condiciones, al resolverlo, obtenemos también la solución del problema primal asociado. La ventaja de esta transformación es que normalmente el problema dual es más fácil de resolver que el primal. Los dos siguiente teoremas ayudan a entender la relación existente entre las soluciones de los dos problemas.

Teorema 1. *Sean $\boldsymbol{x}$ y $\boldsymbol{\alpha}$ vectores tales que satisfacen las restricciones respectivas del problema primal y dual, es decir,$g_i(\boldsymbol{x}) \leq 0$ y $\alpha_i \geq 0$, con $i = 1, ..., n$, entonces $\varphi(\boldsymbol{\alpha}) \leq f(\boldsymbol{x})$.*

Del teorema anterior se pueden extraer dos corolarios. El primero establece que el problema dual está acotado superiormente por el problema primal. El segundo permite afirmar que si $\varphi(\boldsymbol{\alpha}) = f(\boldsymbol{x})$, entonces $\boldsymbol{\alpha}$ y $\boldsymbol{x}$ son soluciones, respectivamente, del problema dual y primal. El interés de este teorema es práctico, ya que permite establecer una heurística para resolver, simultáneamente, el problema primal y dual. Así, estaremos más cerca de la solución, a medida que la diferencia $|\varphi(\boldsymbol{\alpha}) - f(\boldsymbol{x})|$ sea más pequeña. La solución se alcanza cuando la diferencia es nula. Esta solución corresponde a un punto silla de la función lagrangiana, caracterizado por ser simultáneamente un mínimo de $L(\boldsymbol{x}, \boldsymbol{\alpha})$ respecto de $\boldsymbol{x}$ y un máximo de $L(\boldsymbol{x}, \boldsymbol{\alpha})$ respecto de $\boldsymbol{\alpha}$.

El segundo teorema, denominado teorema de Karush-Kuhn-Tucker establece las condiciones suficientes (también conocidas como condiciones KKT) para que un punto $\boldsymbol{x}^*$ sea solución del problema primal.

Teorema de Karush-Kuhn-Tucker. *Si en el problema primal* (81)*, las funciones $f : \mathbb{R}^d \to \mathbb{R}$ y $g_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \ldots, n$ son todas ellas funciones convexas, y existen constantes $\alpha_i \geq 0$, $i = 1, \ldots, n$ tales que:*

$$\frac{\partial f(\boldsymbol{x}^*)}{\partial x_j} + \sum_{i=1}^{n} \alpha_i \frac{\partial g_i(\boldsymbol{x}^*)}{\partial x_j} = 0 \quad j = 1, \ldots, d \tag{84}$$

$$\alpha_i g_i(\boldsymbol{x}^*) = 0 \quad i = 1, \ldots, n \tag{85}$$

entonces el punto $\boldsymbol{x}^*$ es un mínimo global del problema primal.

La primera condición surge como consecuencia de la definición de la función $\varphi(\boldsymbol{\alpha})$ como el ínfimo de la función Lagrangiana, punto en el que las derivadas parciales respecto de $\boldsymbol{x}$ deben ser cero. La segunda condición, denominada *condición complementaria*, es la que garantizará que los óptimos del problema primal y dual coincidan ($\varphi(\boldsymbol{\alpha}^*) = f(\boldsymbol{x}^*)$), ya que, de ser cierta la condición, todos los sumandos del sumatorio de la función Lagrangiana (82) serían nulos.

El interés de este teorema es que establece las condiciones que han de cumplirse para poder resolver el problema primal gracias al dual. Así, partiendo del problema primal, se construye la función lagrangiana. Seguidamente, se aplica la primera condición del teorema de KKT a dicha función y esto permite obtener un conjunto de relaciones que, sustituidas

en la función lagrangiana, harán desaparecer todas las variables primales de dicha función. Este paso es equivalente a calcular $\varphi(\boldsymbol{\alpha}) = \inf_{\boldsymbol{x}\in\Omega} L(\boldsymbol{x}, \boldsymbol{\alpha})$. La función dual así obtenida, sólo dependerá de los multiplicadores de Lagrange. También es posible que, del conjunto de relaciones obtenido al aplicar la primera condición KKT, surjan restricciones adicionales para las variables duales (multiplicadores de lagrange). En este punto queda definido el problema dual junto con sus restricciones. La solución del problema dual permitirá resolver también el problema primal. Para ello, bastará sustituir finalmente la solución dual en las relaciones que anteriormente se obtuvieron al aplicar la primera condición KKT a la función lagrangiana. Ésta es la estrategia que se ha utilizado en la secciones anteriores de este tutorial para abordar los diferentes problemas de optimización que surgen al abordar el problema de clasificación o de regresión mediante SVMs.

Referencias

[Boser et al., 1992] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92 (pp. 144–152). New York, NY, USA: ACM.

[Cortes & Vapnik, 1995] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.