Modelos de Clasificación basados en Máquinas de Vectores Soporte

Article

1 author:

Luis Gonzalez-Abril
Universidad de Sevilla
238 PUBLICATIONS **735** CITATIONS

SEE PROFILE

Modelos de Clasificación basados en Máquinas de Vectores Soporte

L. González Abril
luisgon@us.es - Departamento de Economía Aplicada I
Universidad de Sevilla

Palabras clave: Problemas de Clasificación, SVM, Núcleos (Kernels).

Resumen

En este trabajo se presentan las características más importantes de los modelos de clasificación basados en máquinas de vectores soporte (SVM- Support Vector Machines).

Las máquinas de vectores soporte son sistemas de aprendizaje que utilizan como espacio de hipótesis, funciones lineales en espacios característicos de dimensión muy alta, ensayando algoritmos de aprendizaje de la teoría de la optimización que implementan un aprendizaje sesgado derivado a partir de la teoría del aprendizaje estadístico.

Algunas de las características de estos modelos tienen una especial significación en problemas de clasificación de corte económico y en este trabajo se presentan y comentan.

1. Introducción

La teoría de las SVMs fue desarrollada inicialmente por V. Vapnik [18] a principios de los años 80 y se centra en lo que se conoce como Teoría del Aprendizaje Estadístico[1]. El objeto de las SVMs es dar solución al problema fundamental que surge en distintos campos, donde se estudia, la relación entre sesgo y varianza [8], el control de la capacidad [13], sobreajuste en los datos [15], etc. Este problema consiste en buscar, para una tarea de aprendizaje dada, con una cantidad finita de datos, una adecuada función que permita llevar a cabo una buena generalización[2] que sea resultado de una adecuada relación entre la precisión alcanzada con un particular conjunto de entrenamiento y la capacidad del modelo[3].

[1]De forma simplificada indicar que esta teoría busca formas de estimar dependencias funcionales a partir de una colección de datos.

[2]Donde se entiende por generalización, la capacidad de una determinada función de explicar el comportamiento de los datos dentro de un dominio más amplio.

[3]Capacidad para aprender con cualquier conjunto de ensayo.

A continuación desarrollamos brevemente el planteamiento general de las SVMs para posteriormente centrarnos en los problemas de clasificación[4] desde la perspectiva de un aprendizaje supervisado, es decir, el conocimiento de las salidas de un conjunto de entradas nos permite cuantificar (supervisar) la bondad de los resultados del modelo.

El objetivo fundamental de este tipo de estudios es aprender a partir de los datos y para ello busca la existencia de alguna dependencia funcional entre un conjunto de vectores inputs (o de entrada)

$$\{x_i,\ i = 1, \cdots, n\} \subseteq \mathcal{X} \subseteq \mathbb{R}^d$$

y valores[5] outputs (o salidas)

$$\{y_i,\ i = 1, \cdots, n\} \subseteq \mathcal{Y} \subseteq \mathbb{R}$$

El modelo representado por la figura 1 (denominado modelo de aprendizaje a partir de ejemplos) recoge de manera clara el objetivo que se persigue. En este esquema, **G** representa un modelo

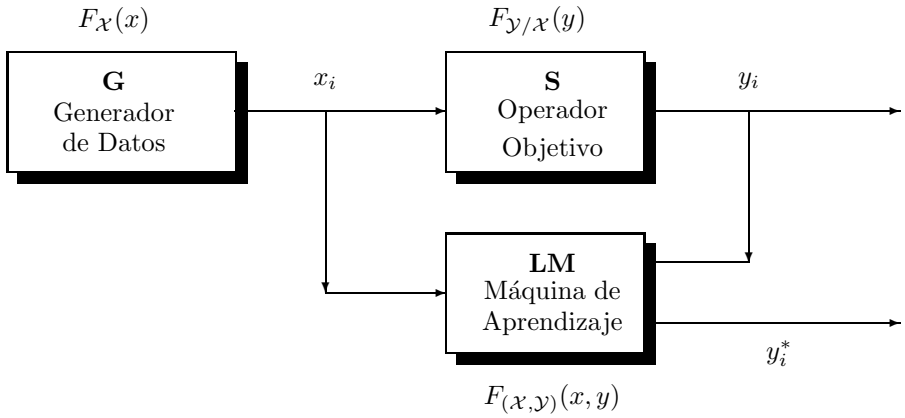

Figura 1: Esquema de configuración de una máquina de aprendizaje a partir de ejemplos.

generador de datos que nos proporciona los vectores $x_i \in X$, independientes e idénticamente distribuidos de acuerdo con una función de distribución $F_{\mathcal{X}}(x)$ desconocida pero que suponemos no varía a lo largo del proceso de aprendizaje. Cada vector x_i es la entrada del operador objetivo **S**, el cual lo transforma en un valor y_i según una función de distribución condicional $F_{\mathcal{Y}/\mathcal{X}=x_i}(y)$. Así la máquina[6] de aprendizaje, que denotamos **LM** (learning machine) recoge el siguiente conjunto de entrenamiento,

$$Z = \{(x_1, y_1), \cdots, (x_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$$

[4]Desde las SVM se puede abordar diferentes problemas como: regresión, estimación de densidades, aproximación de funciones, ...

[5]La generalización a $\mathbb{R}^d$ se sigue sin más que considerar el vector componente a compenente.

[6]Las SVMs tienen sus orígenes en problemas de corte industrial. La puesta en práctica de estos modelos en la industria finaliza cuando se implementa un programa informático dentro de un procesador que recoja datos y los procese. Este procesador es un objeto físico y tangible y por ello la denominación de máquina.

el cual es obtenido independiente e idénticamente distribuido siguiendo la función de distribución conjunta:

$$F_{(\mathcal{X},\mathcal{Y})}(x,y) = F_{\mathcal{X}}(x) \cdot F_{\mathcal{Y}/\mathcal{X}=x}(y)$$

A partir del conjunto de entrenamiento Z, la máquina de aprendizaje "construye" una aproximación al operador desconocido la cual proporcione para un generador dado $\mathbf{G}$, la mejor aproximación (en algún sentido) a las salidas proporcionadas por el supervisor. Formalmente construir un operador significa que la máquina de aprendizaje implementa un conjunto de funciones, de tal forma que durante el proceso de aprendizaje, elige de este conjunto una función apropiada siguiendo una determinada regla de decisión.

La estimación de esta dependencia estocástica basada en un conjunto de datos trata de aproximar la función de distribución condicional $F_{\mathcal{Y}/X}(y)$, lo cual en general lleva a un problema realmente complicado [18, 10]. Sin embargo, el conocimiento de la función $F_{\mathcal{Y}/X}(y)$ no siempre es necesario; a menudo se esta interesado solo en alguna de sus características. Por ejemplo se puede buscar estimar la función de esperanza matemática condicional:

$$E[Y/X = x] \stackrel{def}{=} \int y \, dF_{\mathcal{Y}/x}(y)$$

Por ello el objetivo del problema es la construcción de una función $f(x,y)$ dentro de una determinada clase de funciones[7] $\mathcal{F}$ elegida a priori, la cual debe cumplir un determinado criterio de la mejor manera posible. Formalmente el problema se plantea como sigue:

> *Dado un subespacio vectorial $\mathcal{Z}$ de $\mathbb{R}^{d+1}$ donde se tiene definida una medida de probabilidad $F_{\mathcal{Z}}(z)$, un conjunto $\mathcal{F} = \{f(z),\, z \in \mathcal{Z}\}$ de funciones reales y un funcional $R : \mathcal{F} \to \mathbb{R}$.*
>
> *Buscar una función $f^* \in \mathcal{F}$ tal que*[8]
>
> $$R[f^*] = \min_{f \in \mathcal{F}} R[f]$$

Con objeto de ser lo más general posible sería bueno elegir el funcional R de tal manera que se pudiese plantear con él, el mayor número de problemas posibles. Por ello se define $R[\cdot]$ como sigue:

Definición 1.1 *Dada una clase $\mathcal{F} = \{f(z),\, z \in \mathcal{Z}\}$ de funciones reales y una medida de probabilidad $F_{\mathcal{Z}}(z)$ se define el* **riesgo**, *$R : \mathcal{F} \to \mathbb{R}$, como:*

$$R[f] = \int_{\mathcal{Z}} c(z, f(z))\, dF_{\mathcal{Z}}(z) \tag{1}$$

donde $c(\cdot,\cdot)$ se denomina función de pérdida (o de coste) y tomará valores no negativos.

A la vista de la figura 1 se llega a la conclusión que los valores y_i e y_i^* no necesariamente coinciden. Cuando esto sea así, la máquina de aprendizaje habrá cometido un error que se debe cuantificar de alguna forma y este es precisamente el sentido que tiene la función de perdida.

[7] En este contexto, una clase de funciones (espacio de hipótesis) es sinónimo de una máquina (o modelo) de aprendizaje.

[8] La existencia del mínimo se garantiza cuando se concluye las hipótesis de esta teoría [18].

Así, en este planteamiento, dado un conjunto $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, el principal problema consiste es formular un criterio constructivo para elegir una función de $\mathcal{F}$ puesto que el funcional (1) por si mismo no sirve como criterio de selección, ya que la función $F_{\mathcal{Z}}(z)$ incluida en él es desconocida. Para elaborar dicho criterio se debe tener en cuenta que el riesgo se define como la esperanza matemática de una variable aleatoria respecto a una medida de probabilidad, por tanto es lógico elegir como estimación, la media muestral y de aquí la siguiente definición:

Definición 1.2 *Dado un riesgo definido por (1), un conjunto de funciones $\mathcal{F}$ y una muestra $\{z_1, \cdots, z_n\}$. Al funcional $R_{emp} : \mathcal{F} \to \mathbb{R}$ definido como*

$$R_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} c(z_i, f(z_i)), \qquad f \in \mathcal{F}$$

se le denomina **riesgo empírico**[9].

La forma clásica de abordar estos problemas es: si el valor mínimo del riesgo se alcanza con una función f_0 y el mínimo del riesgo empírico con f_n para una muestra dada de tamaño n, entonces se considera que f_n es una aproximación a f_0 en un determinado espacio métrico. El principio que resuelve este problema se denomina principio de minimización del riesgo empírico, Este es el principio utilizado en los desarrollos clásicos por ejemplo cuando se plantea a partir de un conjunto de datos la regresión lineal mínimo cuadrática.

La pregunta que surge es: ¿se puede asegurar que el riesgo $R[f_n]$ está cerca del $\min_{f \in \mathcal{F}} R[f]$? La respuesta es que en general esto no es cierto [10], lo cual lleva a la conclusión que la clase de funciones $\mathcal{F}$ no puede ser arbitraria, necesariamente se debe imponer algunas condiciones de regularidad a sus elementos, vía un funcional $Q[f]$. Así en la elaboración del problema se debe buscar una adecuada relación entre la precisión alcanzada con un particular conjunto de entrenamiento, medido a través de $R_{emp}[f]$, y la capacidad de la máquina medida por $Q[f]$. Ello lleva a considera el problema de minimizar un riesgo regularizado, donde este se define para algún $\lambda > 0$ en la forma:

$$R_{reg}[f] = R_{emp}[f] + \lambda\, Q[f]$$

Indicar que en las SVMs, el espacio de trabajo es

$$\mathcal{F} = \Big\{ f(x) = w \cdot x + b,\, w \in \mathbb{R}^d,\, b \in \mathbb{R} \Big\}$$

(funciones lineales) y la restricción se impone sobre el parámetro w.

1.1. Problemas de clasificación

En los problemas de clasificación dicotómicos (con etiquetas $\{0, 1\}$) es posible dar una interpretación probabilística de los funcionales riesgos.

Sea $\mathcal{F}$ la clase de todas las funciones reales definidas en $\mathcal{X} = \mathbb{R}^d$ que solo pueden tomar los valores $\{0, 1\}$. Entonces dada una función $f \in \mathcal{F}$ existe un subconjunto $A \subset \mathbb{R}^d$ tal que

[9] Es importante notar que la medida de probabilidad $F(z)$ aparece dada implícitamente a través de los datos $z_1, \cdots, z_n$.

$f(x) = I_A(x)$ (función indicadora del conjunto A). Por tanto se tiene que si $f(x) = y$ entonces la pérdida es $c(x, y, f(x)) = 0$ y si $f(x) \neq y$ se cuantifica el error como $c(x, y, f(x)) = 1$, y se sigue:

$$R[f] = P\left\{(x, y) \in \mathbb{R}^{d+1} \,/\, f(x)) \neq y\right\} = P(A_f)$$

y se tiene que el riesgo coincide con la probabilidad de un conjunto A_f respecto a una medida de probabilidad. Además elegida $f \in \mathcal{F}$ se tiene que $A_f \in \mathbb{R}^{d+1}$ es el conjunto de vectores donde se realiza una clasificación errónea, y de aquí que el riesgo proporciona la probabilidad de una clasificación errónea y cobra un mayor sentido la minimización de éste, es decir, determinar f^* tal que:

$$R[f^*] = P(A_{f^*}) = \min_{f \in \mathcal{F}} P(A_f)$$

Veamos como es el riesgo empírico:

Definición 1.3 *Sea un espacio probabilístico* $(\Omega, \mathcal{A}, P)$. *Se define la* **probabilidad empírica** *de* $A \in \mathcal{A}$ *a partir de la muestra* $\{z_1, \cdots, z_n\}$

$$v_n(A) \stackrel{def}{=} v(A; z_1, \cdots, z_n) = \frac{n(A)}{n}$$

donde $n(A)$ *es el número de elementos de la muestra que pertenecen a* A.

De la definición se tiene que:

$$R_{emp}[f] = \frac{1}{n} \sum_{i=1}^{n} c(x_i, y_i, f(x_i)) = v_n(A_f)$$

es decir, el riesgo empírico es la frecuencia relativa de ocurrencia del suceso A_f en una muestra. Luego según el principio de minimización del riesgo empírico el problema que se plantea es $\min_{f \in \mathcal{F}} v_n(A_f)$.

¿Cómo se interpretaría el riesgo empírico en este caso? Sea un conjunto de vectores $\{x_1, \cdots, x_n\}$ y sus correspondientes etiquetas $\{y_1, \cdots, y_n\}$. Si nos olvidamos de ellas y sobre los vectores $\{x_1, \cdots, x_n\}$ se aplica una función $f(x)$ se obtendrá un conjunto de salidas $\{y'_1, \cdots, y'_n\}$. Entonces cuando se tenga para $i \in \{1, \cdots, n\}$ que $y_i \neq y'_i$ se dirá que se ha producido un error de entrenamiento. Así pues la minimización del riesgo empírico consiste en buscar la función $f^* \in \mathcal{F}$ que minimiza los errores de entrenamiento.

2. Modelos lineales de vectores soporte

Denotamos las dos posibles etiquetas por $Y = \{-1, 1\}$ y

Definición 2.1 *Un conjunto de vectores* $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ *donde* $x_i \in I\!R^d$ *e* $y_i \in \{-1, 1\}$ *para* $i = 1, \cdots, n$ *se dice separable si existe algún hiperplano en* $I\!R^d$ *que separa*[10] *los vectores* $X = \{x_1, \cdots, x_n\}$ *con etiqueta* $y_i = 1$ *de aquellos con etiqueta* $y_i = -1$.

[10] En el sentido de dejar en dos regiones del espacio diferentes.

Dado un conjunto separable existe (al menos) un hiperplano[11]

$$\pi : \; w \cdot x + b = 0$$

que separa los vectores x_i, $i = 1, \cdots, n$ (ver figura 2). Las SVMs buscan entre todos los hiper-

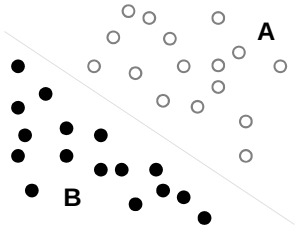

Figura 2: Conjunto e hiperplano separador en $\mathbb{R}^2$. Los puntos huecos representan los vectores con etiqueta $y = 1$ y los restantes $y = -1$.

planos separadores aquel que maximice la distancia de separación entre los conjuntos $\{(x_i, 1)\}$ y $\{(x_i, -1)\}$ (las dos clases posibles).

Veamos como se plantea el problema de optimización correspondiente: Fijado un hiperplano separador siempre es posible reescalar (ver [9]) los parámetros w y b de tal forma que:

$$\begin{aligned} x_i \cdot w + b \geq +1 \quad & \text{para } y_i = +1 \quad (\text{región } \mathbf{A}) \\ x_i \cdot w + b \leq -1 \quad & \text{para } y_i = -1 \quad (\text{región } \mathbf{B}) \end{aligned}$$

De esta forma la mínima separación entre los vectores y el hiperplano separador es la unidad[12]; y las dos desiguladades se pueden expresar en una sola de la forma:

$$y_i \left(x_i \cdot w + b \right) - 1 \geq 0, \qquad i = 1, \cdots, n \tag{2}$$

Sean los vectores de etiqueta 1 para los cuales se cumple la igualdad en (2). Estos puntos pertenecen al hiperplano $\pi_1 : x_i \cdot w + b = 1$ con vector normal w y distancia perpendicular hasta el origen igual a $|1 - b| / \|w\|$ donde $\|w\|$ es la norma euclidea de w. Análogamente, los vectores de etiqueta -1 que cumplen la igualdad en (2) pertenecen al hiperplano $\pi_2 : \; x_i \cdot w + b = -1$ con vector normal w y distancia perpendicular al origen de coorcdenadas igual a $|-1 - b| / \|w\|$. Así se tiene que los hiperplanos π_1 y π_2 son paralelos, la separación entre ellos es $2/\|w\|$ y ningún vector del conjunto de entrenamiento se encuentra entre ellos.

De entre las posibles elecciones del los hiperplanos π_1 y π_2, parece natural elegir aquella que proporcione una mayor separación entre ellos, ya que de esta forma permitiría distinguir de forma más clara las regiones donde caen los puntos con distintas etiquetas. Así se plantea[13]:

$$\begin{aligned} & \min_{w \in \mathbb{R}^d} \frac{1}{2} \|w\|^2 \\ & s.a. \quad y_i \left(x_i \cdot w + b \right) - 1 \geq 0, \; i = 1, \cdots, n \end{aligned} \tag{3}$$

[11] Se utilizará indistintamente la notación $w \cdot x$ y $\langle w, x \rangle$ para denotar el producto escalar de dos vectores.

[12] Sería una especie de normalización.

[13] Otra formulación, que proporciona la misma solución pero cuyo tratamiento es más complicado, consiste en maximizar la separación pero fijando la norma del vector w a la unidad.

La solución para el caso de dimensión dos se puede interpretar gráficamente a partir de la figura 3. A la vista de esta figura, es fácil darse cuenta de una característica muy importante de las SVMs y es que si se añade o elimina cualquier número de vectores que cumplan la desigualdad estricta (2), la solución del problema de optimización no se ve afectada. Sin embargo, basta con añadir un vector que se encuentre entre los dos hiperplanos, para que la solución cambie totalmente.

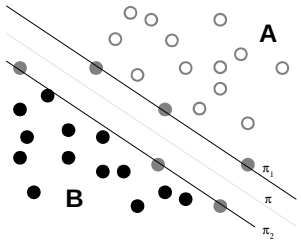

Figura 3: Hiperplanos paralelos y vectores soporte en $\mathbb{R}^2$.

Para resolver el problema de optimización con restricciones (3) se utiliza los multiplicadores de Lagrange[14]. Así la función objetivo es:

$$L_P(w, b, \alpha_i) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i \ (y_i \ (x_i \cdot w + b) - 1)$$

El problema queda como un problema de programación cuadrática donde la función objetivo es convexa, y los vectores que satisfacen las restricciones forman un conjunto convexo. Esto significa que se puede resolver el siguiente problema dual asociado al problema primal: maximizar la función $L_P(w, b, \alpha_i)$ respecto a las variables duales α_i sujeta a las restricciones impuestas para que los gradientes de L_P con respecto a w y b sean nulos, y sujeta también al conjunto de restricciones $C_2 = \{\alpha_i \geq 0, \ \ i = 1, \cdots, n\}$. La solución de este problema se expresa en la forma:

$$w = \sum_{i=1}^{n} \alpha_i \, y_i \, x_i, \qquad \sum_{i=1}^{n} \alpha_i \, y_i = 0$$

y la función objetivo dual:

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \, \alpha_j \, y_i \, y_j \, x_i \, x_j \tag{4}$$

Los vectores del conjunto de entrenamiento que proporcionan un multiplicador $\alpha_i > 0$ son denominados **vectores soporte** y claramente, estos vectores se encuentran en uno de los hiperplanos π_1 o π_2. Para este tipo de modelo de aprendizaje, los vectores soporte son los elementos críticos ya que ellos son los que proporcionan la aproximación del problema, puesto que si todos los restantes elementos del conjunto de entrenamiento son eliminados (o son cambiados por otros que no se encuentren entre los dos hiperplanos), y se repite el problema de optimización, se encuentran los mismos hiperplanos separadores.

[14]En la literatura clásica se denominan de forma errónea de esta manera a pesar de plantearse un problema de optimización cons restricciones de desigualdades, La denominación correcta sería multiplicadores de Karush-Kuhn-Tucker, sin embargo utilizaremos las denominación clásica.

Esta característica de los modelos de vectores soporte puede ser utilizada en muchos problemas económicos donde se desea destacar la importancia de determinadas entradas. También si se trabaja con una gran cantidad de entradas, es útil trabajar con los vectores soporte ya que estos forman un esquema de comprensión[15] que permite reconstruir la solución del problema, es decir, si consideramos exclusivamente los vectores soporte y descartamos el resto de vectores de entrenamiento tendríamos un problema de optimización con menos restricciones que proporciona la misma información.

Las condiciones del problema de optimización nos lleva a que se cumple la condición (ver [9]):

$$\alpha_i \cdot (y_i \cdot (x_i \cdot w + b) - 1) = 0$$

denominada condición (complementaria) de Karush-Kuhn-Tucker (KKT). Estas restricciones indican que el producto de las restricciones del problema primal ($y_i \cdot (x_i \cdot w + b) - 1 \geq 0$) y las restricciones del problema dual ($\alpha_i \geq 0$) se anulan en todos los vectores de entrenamiento. De esta forma se sigue que las condiciones KKT, véase [7], para el problema primal definido a partir de la función objetivo son las siguientes:

$$\begin{aligned}
w_j - \sum_{i=1}^{n} \alpha_i \, y_i \, x_{ij} &= 0 \qquad j = 1, \cdots, d \\
\frac{\partial}{\partial b} L_P = -\sum_{i=1}^{n} \alpha_i \, y_i &= 0 \\
y_i(x_i \cdot w + b) - 1 &\geq 0 \qquad \forall i = 1, \cdots, n \\
\alpha_i &\geq 0 \qquad \forall i = 1, \cdots, n \\
\alpha_i(y_i(x_i \cdot w + b) - 1) &= 0 \qquad \forall i = 1, \cdots, n.
\end{aligned}$$

De los desarrollos iniciales no se sigue una forma explícita de determinar el valor b, sin embargo, la condición KKT complementaria nos permite determinarlo. Para ello, basta elegir un $\alpha_i > 0$ y despejar el valor de b obteniendo $b = y_i - x_i \cdot w$. Aunque se ha determinado b, es más adecuado realizar los cálculos con todos los $\alpha_i > 0$ y elegir como valor de b un valor promedio de los resultados obtenidos, con objeto de redondear los errores intrínsecos asociados a todo método de cálculo numérico[16]:

$$b = \frac{1}{\#\{\alpha_i > 0\}} \sum_{\alpha_i > 0} (y_i - x_i \cdot w)$$

Una vez obtenidos el vector w y la constante b la solución al problema de optimización se interpreta a partir de una función Θ de la siguiente forma:

$$\Theta(x) = \begin{cases} 1 & \text{si } \pi(x) > 0 \\ -1 & \text{si } \pi(x) < 0 \end{cases} \tag{5}$$

[15] Es una regla $\rho : S \to \rho(S)$ que permite construir una clasificación de un conjunto S a partir de un conjunto $\rho(S)$ más pequeño.

[16] Por $\#\{A\}$ se denota el número de elementos del conjunto A.

2.1. Ejemplo

Sean los vectores inputs X dados por[17] las dos primeras columnas de la tabla 1 y donde la tercera columna de esa misma tabla representa la etiqueta sexo.

La solución y otros resultados interesantes aparecen recogidos en la tabla 1. Se observa que la solución se determina a partir de sólo tres vectores, que el conjunto de entrenamiento es separable y por tanto la solución clasifica correctamente todos los vectores. De estos valores se sigue que el

x_{1i}	x_{2i}	y_i	α_i	$\pi(x_{i1}, x_{i2})$	$signo(\pi)$
180	80	1	0	21.2230	1
173	66	1	1.2346	1.0000	1
170	80	1	0	23.4450	1
176	70	1	0	6.5558	1
160	65	1	0	2.3330	1
160	61	-1	0	-3.8894	-1
162	62	-1	0	-2.7782	-1
168	64	-1	0.3210	-1.0000	-1
164	63	-1	0	-1.6670	-1
175	65	-1	0.9136	-1.0000	-1

Tabla 1: Resultados del ejemplo.

hiperplano separador es :

$$\pi(x_1, x_2) : -0{,}2222x_1 + 1{,}5556x_2 - 63{,}22893 = 0$$

(x_1 y x_2 denotan la primera y segunda componente de un vector $x \in \mathbb{R}^2$). Nótese como la imagen de cada vector soporte según el plano π es la unidad lo que significa que se encuentran en uno de los dos hiperplanos separadores.

Si se tiene un nuevo vector input (altura y peso de una persona), por ejemplo $\{195, 95\}$, la solución se interpreta a partir del signo de $\pi(195, 95) = 41{,}224 > 0$ lo que significa que la etiqueta que nosotros le asignamos es $y = 1$, es decir, se indicaría que esta nueva persona es un hombre.

2.2. SVMs para datos no separable

En la práctica no es habitual trabajar con conjuntos separables. En estos casos (ver en la figura 4) se encuentran vectores de una clase dentro de la región correspondiente a los vectores de otra clase y por tanto nunca podrán ser separados de esta clase por medio de hiperplanos. En estas situaciones se dirá que el conjunto es **no separable**. Ante estos casos, el problema de optimización (3), no encuentra una solución posible y ello es evidente sin más que observar como la función objetivo (4) crece de forma arbitraria ya que el multiplicador de Lagrange correspondiente a este vector se puede tomar arbitrariamente grande sin que viole las restricciones. Sin embargo,

[17]Este ejemplo esta tomado de [16]. Los inputs son de dimensión dos, la primera componente representa la altura en centímetros y la segunda el peso en kilogramos de una persona. Las salidas son $y = 1$ si es hombre e $y = -1$ si es mujer.

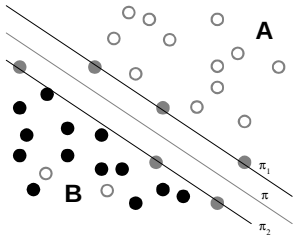

Figura 4: Ejemplo de hiperplanos separadores para el caso de datos no separables.

no es difícil ampliar las ideas generales del caso separable al caso no separable introduciendo una variable ξ de holgura en las restricciones y plantear un nuevo conjunto de restricciones:

$$\begin{aligned} x_i \cdot w + b &\geq +1 - \xi_i \qquad \text{para } y_i = +1 \\ x_i \cdot w + b &\leq -1 + \xi_i \qquad \text{para } y_i = -1 \\ \xi_i &\geq 0 \quad \forall i = 1, \cdots, n \end{aligned}$$

Se tiene ahora que para que se produzca un error en la clasificación de un vector de entrenamiento (una entrada no es ubicada en la clase correcta) es necesario que el valor correspondiente a ξ_i sea superior a la unidad. Así, si en el vector x_i se comete un error entonces $\xi_i \geq 1$ y por tanto $\sum_i \xi_i$ es una cota superior del número de errores que se cometen dentro del conjunto de entrenamiento.

Ya que en el caso no separable, necesariamente se han de cometer errores, parece natural asignar a la función objetivo un coste extra que en "cierto modo" penalice los errores (función de pérdida). Por todo ello, una opción lógica sería plantear el problema de minimizar

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i^k$$

con $k \geq 1$. ¿Cómo se interpreta esta función objetivo? Si se considera un valor C grande, significa que el investigador está asignando un peso a los errores muy alto frente a $\|w\|^2$, y por el contrario si C es pequeño asigna un mayor peso a $\|w\|^2$. Esta interpretación resulta más intuitiva si se interpreta que $\|w\|^2$ es un factor de suavizamiento de la solución buscada. Por otro lado si k es grande lo que hacemos es dar mucho más peso a los errores cuantos mayores sean éstos. Se llega por tanto a plantear un problema de programación convexa para cualquier valor de k. Si $k = 2$ ó $k = 1$ se tiene un problema de programación convexa cuadrático. En el presente trabajo se considera $k = 1$ ya que en este caso se tiene la ventaja de que ningún valor ξ_i, ni ninguno de sus correspondientes multiplicadores de Lagrange, aparecen en el problema dual. Por tanto el problema de optimización que se plantea es:

$$\begin{aligned} &\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \\ &s.a. \begin{cases} y_i\,(x_i \cdot w + b) - 1 + \xi_i \geq 0, \ \forall\, i \\ \xi_i \geq 0, \ \forall\, i \end{cases} \end{aligned} \tag{6}$$

Utilizando la técnica de los multiplicadores de Lagrange se llega a la función objetivo dual

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \, \alpha_j \, y_i \, y_j \, x_i \, x_j$$

la cual hay que maximizar respecto α_i sujeta a

$$0 \leq \alpha_i \leq C, \;\; i = 1, \cdots, n \qquad \sum_{i=1}^{n} \alpha_i \, y_i = 0$$

y cuya solución final viene dada por

$$w = \sum_{i=1}^{N_{SV}} \alpha_i \, y_i \, s_i$$

donde N_{SV} denota el número de vectores soporte y por s_i los vectores soporte del conjunto $\{x_1, \cdots, x_n\}$. Claramente $N_{SV} \leq n$ y una de las características más interesante de estos modelos es que eligiendo adecuadamente los parámetros es posible conseguir que

$$N_{SV} \text{ sea muy inferior a } n$$

con lo que se consigue una representación "corta" de la solución en función de los vectores de entrada sin perder capacidad de generalización.

En problemas económicos, donde se trabaje con una cantidad ingente de datos, y se necesita dar una rápida respuesta frente a una nueva entrada, este esquema resulta muy atractivo puesto que la cantidad de recursos necesarios para proporcionar una salida es "pequeña" en comparación con la cantidad de datos de entrenamiento[18].

Nótese, que la única diferencia en la solución con respecto a la dada en el caso separable es que los multiplicadores de Lagrange α_i, están acotados superiormente por la constante C.

Las condiciones de KKT asociada a este problema son las siguientes:

$$\begin{aligned}
\frac{\partial}{\partial w_j} L_P = w_j - \sum_{i=1}^{n} \alpha_i \, y_i \, x_{ij} &= 0 \\
\frac{\partial}{\partial b} L_P = -\sum_{i=1}^{n} \alpha_i \, y_i &= 0 \\
\frac{\partial}{\partial \xi_i} L_P = C - \alpha_i - \mu_i &= 0 \qquad (7) \\
y_i \, (x_i \cdot w + b) - 1 + \xi_i &\geq 0 \\
\xi_i, \, \alpha_i, \, \mu_i &\geq 0 \\
\alpha_i (y_i \, (x_i \cdot w + b) - 1 + \xi_i) &= 0 \qquad (8) \\
\mu_i \, \xi_i &= 0 \qquad (9)
\end{aligned}$$

Como se comentó en el caso separable, se pueden usar las condiciones complementarias de KKT (las igualdades (8) y (9)), para determinar el valor de b. Nótese que la ecuación (7) combinada

[18] En este punto es importante reseñar que del análisis por ordenador de problemas multidimensionales altos resultó el descubrimiento que R. Bellman denominó, "la maldición de la dimensionalidad", ya que observo que incrementando el número de factores que debían ser tomados en consideración, estos requerían un incremento exponencial de la cantidad de recursos computacional.

con la ecuación (9) muestra que, si $\xi_i = 0$ entonces $\alpha_i < C$. Así se puede simplificar el cálculo de b tomando los vectores de entrenamiento tales que $0 < \alpha_i < C$ y usar la ecuación (8) con $\xi_i = 0$ (como ya se indicó anteriormente es más adecuado promediar este valor entre todos los vectores de entrenamiento con $\xi_i = 0$).

2.3. Modelos teóricos SVMs

Como resumen de lo señalado hasta ahora en esta sección se sigue que en el problema de minimización (6) se busca determinar un hiperplano separador óptimo:

$$\pi \equiv f(x; w, b) = \langle w, x \rangle + b = 0$$

donde $w \in \mathbb{R}^d$ y $b \in \mathbb{R}$ son parámetros que deben ser estimados. Por ello se puede considerar que el conjunto de funciones $\mathcal{F}$ sobre la que se busca la solución al problema (6) es

$$\mathcal{F} = \left\{ f(x; w, b) = \langle w, x \rangle + b, \;\; w \in \mathbb{R}^d, \, y \in \mathbb{R} \right\}$$

es decir, que

$$\min_{w \in \mathbb{R}^d} L_P(x; w, b) = \min_{f \in \mathcal{F}} L_P(x, f)$$

y $L_P(x; f)$ puede ser considerado como un riesgo regularizado (ver en [9]) $R_{reg}[f]$ por lo que

$$\min_{w \in \mathbb{R}^d} L_P(x; w, b) = \min_{f \in \mathcal{F}} R_{reg}[f]$$

Por otro lado, ya que una vez determinado el valor óptimo de w, a partir de las condiciones de KKT se obtiene el parámetro b, el objeto principal de estudio se reduce en esencia al vector de parámetros $w \in \mathbb{R}^d$. Este vector de parámetros se obtiene según una combinación lineal de los vectores del conjunto de entrenamiento por la expresión $w = \sum_{i=1}^{N} \alpha_i \, y_i \, x_i$ de donde se sigue que la función solución al problema de clasificación se expresa:

$$f(x) = \sum_{i=1}^{n} \alpha_i \, y_i \, \langle x_i, x \rangle + b. \tag{10}$$

3. Máquinas no lineales de vectores soporte

En esta sección se aborda el problema de generalizar los desarrollos anteriores para el caso de clases de funciones no necesariamente lineales en los parámetros. Para ello, obsérvese que los vectores de entrada forman parte de la solución (10) del problema de clasificación, a través de los productos escalares, $\langle x_i, x \rangle$. Utilizando la idea dada en [1] se sigue: Sea una aplicación Φ del conjunto de entradas $\mathcal{X}$, en un espacio $\mathcal{H}$ (denominado espacio característico) dotado de un producto escalar.

$$\Phi : \mathcal{X} \subset \mathbb{R}^d \longrightarrow \mathcal{H}.$$

Ahora, en lugar de considerar el conjunto de vectores $\{x_1, \cdots, x_n\}$ se considera los vectores transformados $\{\Phi(x_1), \cdots, \Phi(x_n)\}$ y si se plantea el problema de optimización original a estos

vectores se tiene que los nuevos vectores forman parte de la solución del problema solo a través del producto escalar definido en $\mathcal{H}$: $\langle \Phi(x_i), \Phi(x) \rangle$. Así, si se considera una función

$$k : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$$

(denominada función núcleo), tal que

$$k(x, x') = \Phi(x) \cdot \Phi(x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$$

solo es necesario conocer la función núcleo para resolver el algoritmo y no se necesita tener la forma explícita de la aplicación Φ.

Por tanto si se reemplaza $\langle x_i, x \rangle$ por $k(x_i, x)$ en la solución de los problemas de optimización, se habrá conseguido una máquina de vectores soporte planteada en un nuevo espacio, y además, lo que resulta muy importante en la práctica, la ejecución de un programa que lleve a cabo esta técnica no lineal, consume la misma cantidad de recursos computacionales que si la técnica fuese lineal.

Al resolver este problema en $\mathcal{H}$, donde se trabaja con funciones lineales, la solución que resulta es lineal en él, pero no es necesariamente lineal en el espacio de entradas $\mathcal{X}$, con lo cual se esta generalizando el problema a conjuntos de funciones no lineales.

Si se lleva a cabo la transformación de los datos, el vector solución es

$$w = \sum_{i=1}^{N_{SV}} \alpha_i \, y_i \, \Phi(s_i) \tag{11}$$

donde $\Phi(s_i)$ denota los vectores soporte del conjunto $\{\Phi(x_1), \cdots, \Phi(x_n)\}$. Es importante notar que los vectores soporte se encuentran dentro del conjunto $\{\Phi(x_1), \cdots, \Phi(x_n)\}$, se denotan por $\Phi(s_i)$ pero no son necesariamente los transformados de los vectores soporte s_i que se encuentran dentro del conjunto $\{x_1, \cdots, x_n\}$, entre otras cosas porque con los vectores sin transformar no se realiza ningún algoritmo, a pesar de esta indicación se sigue con esta notación, puesto que es la utilizada tradicionalmente.

Aunque la aplicación Φ aparece en la solución (11) se tiene que cuando se realiza la fase de prueba, no se necesita tener de forma explícita ya que la solución queda:

$$f(x) = \sum_{i=1}^{N_{SV}} \alpha_i \, y_i \, \langle \Phi(s_i), \Phi(x) \rangle + b$$

y escrita en términos de la función núcleo:

$$f(x) = \sum_{i=1}^{N_{SV}} \alpha_i \, y_i \, k(s_i, x) + b$$

Una importante consecuencia de la representación dual (en términos del núcleo) es que la dimensión del espacio característico no afecta a los cálculos ya que la única información necesaria se encuentra en la matriz de orden $n \times n$: $\{k(x_i, x_j)\}_{i,j=1}^{n}$ (denominada matriz de Gram). Un esquema de esta construcción se encuentra en la figura 5 y un ejemplo gráfico de la solución aportada por la SVM al ejemplo aparece en la figura 6.

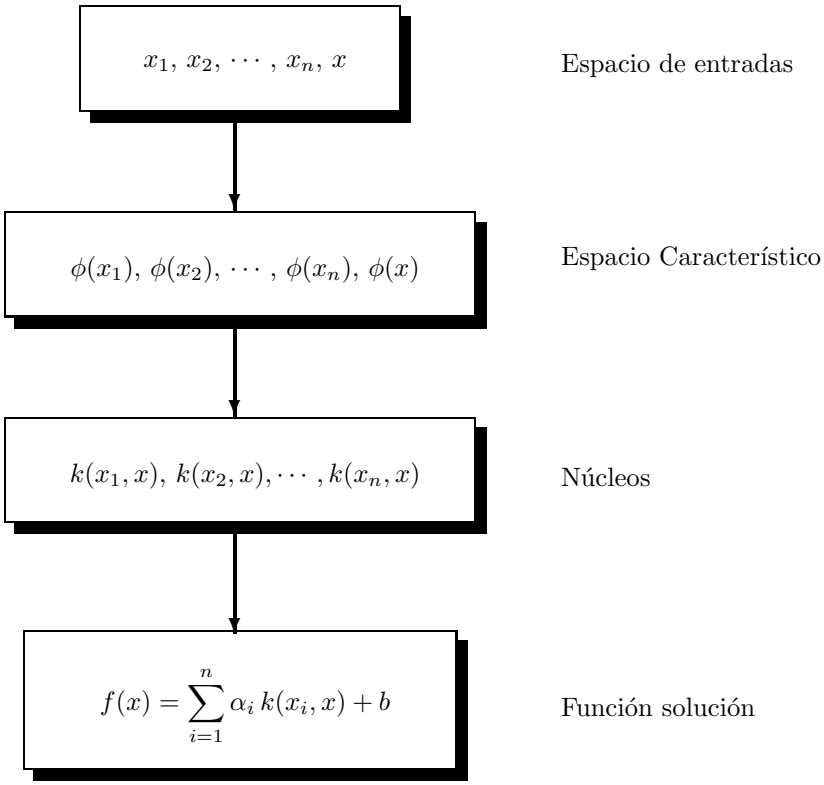

Figura 5: Las máquinas de vectores soporte transforman, inicialmente, el espacio de entradas en un espacio característico de dimensión superior y entonces construye la función de clasificación lineal óptima dentro de este nuevo espacio.

Al usar $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ se trabaja en un nuevo espacio $\mathcal{H}$ por lo cual el vector solución w se encuentra en este espacio. Por tanto, puede ocurrir que sobre el conjunto X inicial no se tenga definida ningún tipo de estructura, y la función Φ sirve para dar una estructura[19] a los datos y poder aplicar una adecuada clasificación. Esta idea aparece recogida en el trabajo [11] donde se presenta una función núcleo concreta y sobre todo una forma de hacer frente a la construcción de funciones núcleos a partir de una interpretación de similitud entre vectores de entrada.

Por otro lado, dado un espacio $\mathcal{H}$ dotado de un producto escalar habría que estudiar que propiedades deben cumplir las funciones

$$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{H}$$

que permiten construir un par $\{\Phi, \mathcal{H}\}$, con las propiedades descritas anteriormente, es decir, ser funciones núcleos. Un análisis detallado de este punto puede encontrarse entre otros en [9], [5] y [18].

[19] Una forma de verlos.

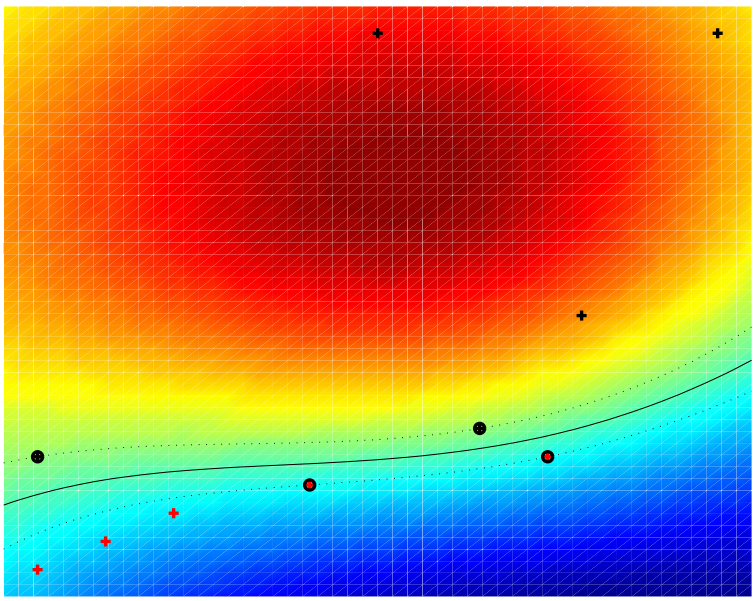

Figura 6: Solución gráfica al problema planteado en la sección 2.1 a partir de una función núcleo (gaussiano). Los puntos "+"en negro representan los vectores de entrada con etiqueta 1 (hombres) y los puntos "+"en rojo los de etiqueta -1 (mujeres). Los puntos en círculos representan los vectores soporte.

3.1. Clasificación múltiple

Generalizamos los problemas de clasificación a etiquetas múltiples. Así sea el conjunto de etiquetas posibles $\{\theta_1, \cdots, \theta_\ell\}$, siendo $\ell > 2$ y sin una relación de orden definida entre ellas. Sea Z el conjunto de entrenamiento y se construyen los subconjuntos

$$Z_k = \{(x_i, y_i),\ \text{tales que}\ y_i = \theta_k\}$$

que determinan una partición de Z. Denotamos por n_k el número de vectores de entrenamiento del conjunto Z_k ($n = n_1 + n_2 + \cdots + n_\ell$); y por I_k el conjunto de índices i tales que $(x_i, y_i) \in Z_k$ de donde se sigue que $\bigcup_{i \in I_k} \{(x_i, y_i)\} = Z_k$.

La forma, más habitual, de utilización de las máquinas de vectores soporte para resolver problemas de multiclasificación admite dos tipos de arquitectura:

- SVMs biclasificadoras generalizadas: Construyen una función clasificadora global a partir de un conjunto de funciones clasificadoras dicotómicas (biclasificadoras).
- SVMs multiclasificadoras: Construyen una función clasificadora global directamente considerando todas las clases a la vez.

Las SVMs multiclasificadoras plantea un problema de optimización similar a las SVMs biclasificadoras que posteriormente resuelve. Tienen el inconveniente que proporciona una salida final y el proceso de obtención de esta salida es como una caja negra que no admite ninguna reconstrucción posterior.

Las SVMs biclasificadoras generalizadas dan solución al problema de la multiclasificación transformando las ℓ particiones del conjunto de entrenamiento en un conjunto de L biparticiones, en las cuales construye la correspondiente función clasificadora (es lo que se denomina esquema de descomposición) obteniendo $f_1, \cdots f_L$ clasificadores dicotómicos o biclasificadores. A continuación, mediante un esquema de reconstrucción, realiza la fusión de los biclasificadores f_i, $i = 1, \cdots, L$ con objeto de proporcionar como salida final, una de las ℓ clases posibles, $\{\theta_1, \cdots, \theta_\ell\}$.

Dentro del esquema de descomposición, las máquinas más utilizadas son:

- Máquinas 1-v-r SV (iniciales de *one- versus- rest*). Máquinas de vectores soporte, donde cada función clasificadora parcial f_i, enfrenta los vectores de la clase θ_i contra los vectores de las restantes clases.
- Máquinas 1-v-1 SV (iniciales de *one- versus- one*). Máquinas de vectores soporte, donde cada función clasificadora parcial f_{ij}, enfrenta los vectores de la clase θ_i contra los de la clase θ_j, sin considerar las restantes clases.

En estos casos para la reconstrucción del problema se utiliza algún esquema de votación que tenga en cuenta la distribución de las etiquetas asignadas por las máquinas parciales:

Etiquetas	Votos
θ_1	m_1
$\vdots$	$\vdots$
θ_k	m_k
$\vdots$	$\vdots$
θ_ℓ	m_ℓ
	$\frac{\ell \cdot (\ell - 1)}{2}$

donde m_k es el número de votos que las máquinas f_i, $i = 1, \cdots, \frac{\ell \cdot (\ell - 1)}{2}$ dan a la etiqueta θ_k.

4. Comentarios

Los núcleos proporcionan el lenguaje descriptivo usado por la máquina para ver los datos. Frecuentemente, el diseñador puede trabajar con una noción más intuitiva de similitud en el espacio de los inputs y es ahí donde los expertos pueden dar una guía inigualable de una apropiada medida de similitud. Dentro de los problemas económicos, la elección de los núcleos nos permitirá tener a nuestra disposición un gran número de modelos no paramétricos con la ventaja de poder interpretar las características del modelo, gracias a la linealidad del espacio característico.

La introducción de los núcleos aumenta significativamente la potencia de las máquinas de aprendizaje a la vez que retiene la linealidad que asegura que los aprendizajes resulten comprensibles. El incremento de la flexibilidad, sin embargo, incrementa el riesgo de sobreajuste con la elección de hiperplanos separables que aumentan los problemas debido al número de grados de libertad. El control adecuado de la flexibilidad del espacio característico inducido por los núcleos requiere una teoría de generalización, la cual sea capaz de describir con precisión que factores han

de ser controlados en las máquinas de aprendizaje con objeto de garantizar unas buenas generalizaciones. Este control de la capacidad de generalización queda recogido en [5] donde se dan diferentes teoremas que permiten afirmar que para cualquier distribución seguida por los datos y considerando el dominio de las funciones $f \in \mathcal{F}$, una bola de radio R centrada en el origen que contengan todos los vectores x_i, se sigue que con probabilidad al menos $1 - \eta$ sobre una muestra de tamaño n, se tiene para todas $f \in \mathcal{F}$, el error de ensayo $\text{mín}_{f \in \mathcal{F}} P(A_f)$ puede ser controlado a partir de $\|w\|^2 + \lambda \sum_{i=1}^{n} \xi_i$ lo cual da pleno sentido a las máquinas de vectores soporte.

Indicar que sólo después de llevar a cabo un proceso de aprendizaje se puede conocer cual es la complejidad de la hipótesis resultante. Este tipo de análisis, el cual proporciona una forma de explotar las buenas condiciones entre la función objetivo y la distribución de las entradas, es llamada minimización del riesgo estructural dependiente de los datos.

Los multiplicadores de Lagrange asociados con cada input, en la solución aportada por las máquinas de vectores soporte, llegan a ser las variables duales, dando con ello una interpretación intuitiva que cuantifica como de importante es un vector de entrenamiento dado en la formación de la solución final.

Para muchas clases de núcleos, siempre es posible encontrar un parámetro del núcleo para el cual los datos llegan a ser separables (en general forzar esta separación puede conducir fácilmente al sobreajuste). En estos casos, los outliers podrían ser caracterizados por tener multiplicadores de Lagrange muy grandes, y este procedimiento podría ser usado para depurar los datos ya que ello puede clasificar los datos de entrenamiento de acuerdo a la dificultad que ellos presentan para una clasificación correcta.

La idea de representar la solución por medio de un pequeño subconjunto del conjunto de entrenamiento tiene una enorme ventaja de cálculo. La implementación de estas técnicas pasa por plantear una función objetivo con tantas restricciones como número de entradas, lo cual conduce a un problema de una elevada complejidad de cálculo. Existen distintas direcciones de páginas web, por ejemplo:

http://www.support-vector.net

http://www.kernel-machines.org

http://svm.first.gmd.de

http://www.syseng.anu.edu.au/lsg/

donde se pueden encontrar muchos artículos donde se estudian diferentes formas de optimizar estas técnicas, así como comparativas con otras técnicas que resuelven problemas similares. Los modelos basados en SVMs pueden ser también aplicado a los problemas de estimación de la regresión, manteniendo las principales características de los problemas de clasificación: una función no lineal es buscada, a través de una máquina lineal, en un espacio característico inducido por un núcleo mientras la capacidad del sistema es controlada por un parámetro que no depende de la dimensionalidad del espacio.

En la página web:

http://www.clopinet.com/isabelle/Projects /SVM/applist.html

se puede encontrar diferentes aplicaciones de estos modelos a problemas reales, donde se obtie-

nen resultados muy satisfactorios. Algunos de los problemas que resuelven se encuentran en los siguientes campos: Clasificación de imágenes, Aproximación de funciones y regresión, Reconocimiento de objetos en 3D, Caracterización de textos, Reconocimiento de caracteres de escritura, Predicción de series temporales y Reconstrucción de sistemas caóticos, Árboles de decisión, ...

La aplicación de estas técnicas a problemas económicos no han sido aún muy utilizadas, sin embargo algunas referencias son:

- [2] estudia el problema de clasificación de empresas (rating)
- [14] donde se plantea un modelo de regresión que explica el precio de la casas en Boston, a partir de 15 variables explicativas continuas, y compara la solución aportada por una máquina de vectores soporte con una regresión en cordillera (ridge) y otra regresión en árbol;
- [4] plantea árboles de decisión en bases de datos comerciales y lo compara con otros métodos estándar de clasificación.
- [6] Plantea a través de un problema de clasificación una modelo de predecir bancarrotas aplicadas a un conjunto de bancos australianos.
- [17] se presentan algunos resultados sobre precios de stocks

En estos trabajos se compara los modelos SVMs con otros (redes neuronales, árboles de decisión,...) y se obtiene unos resultados muy satisfactorios.

5. Conclusiones y trabajos futuros

Poco a poco la comunidad científica esta cada vez más convencida de la utilidad de las SVMs y así se pone de manifiesto cuando, por ejemplo, buscamos una palabra clave relacionada con estás técnicas en los buscadores de internet y observamos el alto número de referencias que aparecen. Sin embargo, aún son pocas las aplicaciones de esta metodología en problemas económicos.

En la actualidad estoy trabajando con compañeros de otras universidades con el objetivo de generalizar las SVMs a problemas de multiclasificación en dos sentido: en primer lugar incorporando una interpretación probabilística de las salidas, parte de estas investigaciones se encuentran en [12]; y en segunda lugar, mejorando algunas deficiencias que presentan las SVMs 1-v-1 cuyos trabajos van por buen camino y nos han aceptado una comunicación en ESANN2003 [3].

Referencias

[1] M.A. Aizerman, E.M. Braverman, and L.I. Rozonoér. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, (25):821–837, 1964.

[2] C. Angulo. *Aprendizaje con máquinas núcleos en entornos de multiclasificación.* Tesis doctoral, Universidad Politécnica de Cataluña, Abril 2001.

[3] C. Angulo and L. González. 1-v-1 tri-class sv machine. *ESANN'2003*, 2003.

[4] K. Bennett and L. Auslender. On support vector decision trees for database marketing. *Bajado de http://www.clopinet.com/ isabelle/ Projects/ SVM/ applist.html*, Marzo 1998.

[5] N. Cristianini and J.Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University press 2000, 2000.

[6] A. Fan and M. Palaniswami. Selecting bankruptcy predictors using svm approach. *IEEE*, 2000.

[7] R. Fletcher. *Practical methods of optimization*. John Wiley and Sons, Inc, 2 edition, 1987.

[8] S. German and E. Bienenstock. Neural networks and the bias / variance dilemma. *Neural Computation, 4:1-58*, 1992.

[9] L. González. *Teoría del aprendizaje estadístico de la regresión. Máquinas de regresión de vector base*. Trabajo interno del departamento de economía aplicada i, Facultad de Ciencias Económicas y Empresariales, Universidad de Sevilla, Diciembre 2000.

[10] L. González. Análisis discriminante utilizando máquinas núcleos de vectores soporte. Función núcleo similitud. Tesis doctoral, Dpto. Economía Aplicada I. Universidad de Sevilla, Junio 2002.

[11] L. González and J.M. Alba. Similitud entre sucesos. *Terceras Jornadas de Trabajo sobre Metodologías Cualitativas Aplicadas a los Sistemas Socioeconómicos*, Julio 2001.

[12] L. González, C. Angulo, F. Velasco, and M. Vilchez. Máquina ℓ-SVCR con salidas probabilísticas. *Inteligencia Artificial*, (17):72–82, september 2002.

[13] I. Guyon, V. Vapnik, B. Boser, L. Bottou, and S. Solla. Structural risk minimization for character recognition. *Advances in Neural Information Processing Systems, 4:471-479*, 1992.

[14] K.R. Müller, A.J. Smola, G. Rätsch, B. Schölkopf, J. Kohlnorgen, and V. Vapnik. Predicting times series with support vector machine. Notas de trabajo, 1997.

[15] D. Montgomery and Peck E. *Introduction to Linear Regression Analysis*. John Wiley and Sons, Inc. 2nd edición, 1992.

[16] M. Stitson, J. Weston, A. Gammerman, V. Vovk, and V. Vapnik. Theory of support vector machines. Informe Técnico. Bajado de http://svm.first.gmd.de/, 1996.

[17] T. Trafalis and H. Ince. Svm for regression and applications to financial forecasting. *IEEE*, 2000.

[18] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc, 1998.