Universidad de Ingeniería y Tecnología - UTEC
Departamento de Ingeniería Mecatrónica
Curso: Tópicos de Ingeniería Mecatrónica (2020-2)

Tema 5:
Support Vector Machines (SVM) - Parte 2

Prof. Oscar E. Ramos Ponce

1. SVM de Margen Blando

El SVM de Margen Blando, llamado *Soft* SVM en inglés, no hace la suposición que las clases son linealmente separables. Por ese motivo puede haber instancias de entrenamiento mal clasificadas a las cuales se asociará un error, denominado variable de holgura.

1.1. Variables de Holgura

La variable de holgura ξ_i (en inglés llamada *slack*) de una instancia i representa el grado de error en la clasificación y se mide como la distancia de la instancia al borde del margen que corresponde a la clase correcta. Si la instancia i es clasificada adecuadamente, entonces $\xi_i = 0$. De no ser clasificada adecuadamente, el plano que pasa por esta instancia i es $\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b = 1 - \xi_i$, si la instancia es $+1$, y $\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b = -(1 - \xi_i)$ si la instancia es -1. El plano en ambos casos se puede representar como $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) = 1 - \xi$. Utilizando esta expresión, se tiene que para toda instancia se debe cumplir

$$y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1 - \xi_i, \tag{1}$$

donde el valor de ξ puede ser:

- $\xi_i = 0$: cuando la instancia se encuentra en el lado correcto de la clasificación o justo sobre el margen de separación.
- $0 < \xi_i \leq 1$: cuando la instancia se encuentra entre el margen correcto y el plano $\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b = 0$.
- $\xi_i > 1$: cuando la instancia se encuentra más allá del plano $\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b = 0$, lo cual se interpreta como una mala clasificación.

Debido a que (1) siempre se cumple, esta restricción será utilizada en el problema de optimización.

1.2. Definición del Problema

Cuando no se hace la suposición que las clases son linealmente separables, se admite que puede ocurrir cierto error en la clasificación, medido por la variable de holgura ξ_i. Debido a que se desea el clasificador óptimo, se intenta minimizar todas estas variables de holgura, añadiendo un término al objetivo de optimización del SVM de margen duro. En este caso, las restricciones de cada instancia deben también contener el término de holgura como se muestra en (1). Utilizando estos criterios, el problema de optimización queda definido como

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.a.} \\ &\quad y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \cdots, n \\ &\quad \xi_i \geq 0, \qquad\qquad i = 1, \cdots, n, \end{aligned} \tag{2}$$

donde C es una constante llamada *penalidad de la holgura*, y $\boldsymbol{\xi}$ es un vector que contiene a todos los ξ_i. Si el valor de esta constante es muy grande, solamente se buscará separar las clases sin otorgar mucha importancia al margen γ. Si C es muy pequeño, ξ_i podrá tomar cualquier valor y básicamente se estaría ignorando los datos. Así, se debe escoger este parámetro de manera adecuada.

Como se puede ver en (2), se agrega una restricción adicional indicando que las variables de holgura no pueden ser negativas, ya que por definición el valor mínimo que pueden tener es cero. La sumatoria $\sum_{i=1}^{n}\xi_i$ es una medición de la cantidad de error en la que inevitablemente incurre el SVM de margen blando.

1.3. Función de Costo y Descenso del Gradiente

La restricción que impone la variable de holgura sobre cada una de las instancias se describe usando (1). Esta restricción se puede despejar para ξ_i quedando $\xi_i \geq 1 - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b)$, con $\xi_i \geq 0$, que son las restricciones de (2). Ambas restricciones implican que o ξ_i es mayor que $1 - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b)$, o es mayor que cero. Esto se expresa como

$$\xi_i = \text{máx}(0, 1 - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b)) = \ell(y^{(i)}, \boldsymbol{x}^{(i)}), \tag{3}$$

y se conoce también como la función de costo $l(y^{(i)}, \boldsymbol{x}^{(i)})$ asociada a la instancia i. A esta función de costo se le denomina de tipo *hinge-loss*. Denotando $z = \boldsymbol{w}^T\boldsymbol{x} + b$, sin incluir explícitamente la referencia a la muestra i por simplicidad de notación, es común graficar la función de costo $l(z) = \text{máx}(0, 1 - yz)$ usando z en el eje horizontal. Si $y = 1$, esta función es $1 - z$ desde $-\infty$ hasta $z = 1$, y para valores de $z > 1$ es cero. Si $y = -1$, esta función es cero hasta $z = -1$ y de ahí hacia ∞ es z. Graficando se puede observar que esta función tiene un margen de seguridad en $+1$ y -1, a partir del cual se incrementa el costo.

Utilizando la definición de esta función de costo $l(y^{(i)}, \boldsymbol{x}^{(i)})$, que incluye todas las restricciones impuestas por cada ξ_i, el problema de optimización dado en (2) se puede escribir sin restricciones como

$$\min_{\boldsymbol{w},b} J(\boldsymbol{w}, b) \tag{4}$$

donde

$$J(\boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\text{máx}(0, 1 - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b)) \tag{5}$$

es la función de costo asociado a todas las instancias. Se puede demostrar que la función de costo (5) es convexa, y por tanto tiene un mínimo global.

Cuando se expresa el problema de optimización del SVM de margen blando sin restricciones como en (4), la solución se puede encontrar usando el método del descenso del gradiente visto anteriormente. Para ello se necesita las derivadas de la función de costo, las cuales están dadas por

$$\frac{\partial J}{\partial w_j} = w_j + C\sum_{i=1}^{n}\frac{\partial\,\ell(\boldsymbol{x}^{(i)}, y^{(i)})}{\partial w_j} \qquad \text{y} \qquad \frac{\partial J}{\partial b} = C\sum_{i=1}^{n}\frac{\partial\,\ell(\boldsymbol{x}^{(i)}, y^{(i)})}{\partial b}.$$

De manera estricta, las derivadas de ℓ no están definidas en todo el dominio de la función. Sin embargo, para fines prácticos se pueden definir como

$$\frac{\partial\,\ell(\boldsymbol{x}^{(i)}, y^{(i)})}{\partial w_j} = \begin{cases} 0, & \text{si } y^{(i)}(w^T x^{(i)} + b) \geq 1 \\ -y^{(i)}x_i^{(i)}, & \text{si } y^{(i)}(w^T x^{(i)} + b) < 1 \end{cases}$$

con respecto a cada w_j y como

$$\frac{\partial\,\ell(\boldsymbol{x}^{(i)}, y^{(i)})}{\partial b} = \begin{cases} 0, & \text{si } y^{(i)}(w^T x^{(i)} + b) \geq 1 \\ -y^{(i)}, & \text{si } -y^{(i)}(w^T x^{(i)} + b) < 1 \end{cases}$$

con respecto al offset b. Una vez que se tiene estas derivadas se puede aplicar directamente el descenso del gradiente en su forma batch, mini-batch o estocástico.

1.4. Solución al Problema de Optimización

La solución al problema de minimización del margen y de las variables de holgura, dado en (2), se puede también encontrar usando la resolución directa a través de las condiciones de Karush-Kuhn-Tucker. Dado que este problema general, dado en el apéndice A utiliza desigualdades de tipo $\leq$ en las restricciones, las restricciones de (2) se deben reescribir. Así, $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1 - \xi_i$ se escribirá como $1 - \xi_i - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \leq 0$, y $\xi_i \geq 0$ se escribirá como $-\xi_i \leq 0$. Usando estas modificaciones, el Lagrangiano será

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \alpha, \beta) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - \xi_i - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b)) - \sum_{i=1}^{n}\beta_i\xi_i, \tag{6}$$

donde α_i y β_i son los multiplicadores de Lagrange. Las condiciones KKT en este caso son las siguientes:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w}^* - \sum_{i=1}^{n}\alpha_i y^{(i)}\boldsymbol{x}^{(i)} = 0 \tag{7}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n}\alpha_i y^{(i)} = 0 \tag{8}$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \tag{9}$$

$$\alpha_i \geq 0 \tag{10}$$

$$\beta_i \geq 0 \tag{11}$$

$$\alpha_i(1 - \xi_i - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*)) = 0 \tag{12}$$

$$-\beta_i\xi_i = 0 \tag{13}$$

$$1 - \xi_i - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*) \leq 0 \tag{14}$$

$$-\xi_i \leq 0. \tag{15}$$

En las restricciones de KKT, los asteriscos hacen referencia a los valores óptimos.

Análisis del valor de α. La restricción (9) establece que para los multiplicadores de lagrange α_i, β_i asociados con la instancia i se cumple que

$$C = \alpha_i + \beta_i. \tag{16}$$

Equivalentemente, se tiene que $\alpha_i = C - \beta_i$. En esta última expresión se puede ver que, dado que $\beta_i \geq 0$ según (15), el valor máximo de α_i es C. Más aún, de (14) se tiene que $\alpha_i \geq 0$, por lo que usando ambas desigualdades se llega a

$$0 \leq \alpha_i \leq C. \tag{17}$$

Existen tres casos distintos para α_i y se muestran a continuación.

- Si $\alpha_i = 0$, de (16) se tiene que $C = \beta_i > 0$, y para satisfacer la restricción (13) se necesita $\xi_i = 0$, lo cual implica que no se tiene error de clasificación. Es decir, las instancias $\boldsymbol{x}^{(i)}$ que tengan asociado un $\alpha_i = 0$ se encuentran correctamente clasificadas como +1 o -1.

- Si $\alpha_i = C$, de (16) se desprende que $\beta_i = 0$. Por otro lado, cuando $\epsilon_i > 0$, de (13) se necesita que $\beta_i = 0$, implicando, por lo anterior, que $\alpha_i = C$. Esto significa que las instancias $\boldsymbol{x}^{(i)}$ que tienen asociado un $\alpha_i = C$, o se encuentran bien clasificadas pero más allá del margen de tolerancia, o se encuentran mal clasificadas ($\epsilon_i > 0$). Más aún, cuando $\alpha_i \neq 0$, como en este caso, de (12) se desprende que $1 - \xi_i - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*) = 0$, o escrito de otra forma:

  $$y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*) = 1 - \xi_i,$$

  que es un caso especial de (1). Si $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*) \geq 0$, se tiene $\xi_i \in [0, 1]$ y la instancia $\boldsymbol{x}^{(i)}$ se encuentra adecuadamente clasificada. Si $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*) < 0$, se tiene $\xi_i > 1$, lo cual indica una clasificación incorrecta.

- Si $0 < \alpha_i < C$, de (16) se desprende que $\beta_i \neq 0$, lo cual implica $\xi_i = 0$ para satisfacer (13). Usando (12), en este caso se tendrá $\alpha_i(1 - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*)) = 0$, pero como $\alpha \neq 0$, se tiene $1 - y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*) = 0$. Despejando, se llega a $y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b^*) = 1$, que es la definición de un vector de soporte. Es decir, $\boldsymbol{x}^{(i)}$ es un vector de soporte, $\boldsymbol{x}_{sv}^{(i)}$, cuando $0 < \alpha < C$.

Adicionalmente, la condición (8) brinda una restricción sobre el valor de α_i según los valores deseados $\boldsymbol{y}^{(i)}$.

Parámetros Óptimos. A partir de (7), se obtiene una condición similar a la del SVM de margen duro. Concretamente, el valor óptimo del vector $\boldsymbol{w}$ debe ser:

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i y^{(i)} \boldsymbol{x}^{(i)}. \tag{18}$$

De manera similar, el valor de b óptimo se obtiene despejando la ecuación de la recta que pasa por un vector de soporte $y_{sv}^{(i)}(\boldsymbol{w}^T\boldsymbol{x}_{sv}^{(i)} + b^*) = 1$, y utilizando cualquier vector de soporte, de tal modo que

$$b^* = y_{sv}^{(i)} - \boldsymbol{w}^{*T}\boldsymbol{x}_{sv}^{(i)}.$$

Debe tenerse en cuenta que, en este caso, los vectores de soporte no se dan cuando $\alpha_i = 0$, sino cuando $0 < \alpha_i < C$. Igual que en el caso del SVM de margen duro, resulta conveniente promediar todos los valores de b obtenidos a partir de los diferentes vectores de soporte:

$$b^* = \frac{1}{N_{sv}} \sum_{i=1}^{N_{sv}} y_{sv}^{(i)} - \boldsymbol{w}^{*T} \boldsymbol{x}_{sv}^{(i)}, \tag{19}$$

donde N_{sv} es el número total de vectores de soporte.

Desarrollo del Lagrangiano. El desarrollo del Lagrangiano (6) es similar al del SVM de margen duro. La principal diferencia es la expansión de los dos últimos terminos de (6), los cuales contienen los multiplicadores de Lagrange α_i y β_i:

$$\sum_{i=1}^{n} \alpha_i (1 - \xi_i - y^{(i)}(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b)) - \sum_{i=1}^{n} \beta_i \xi_i.$$

Estos términos se pueden escribir como

$$\sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i \xi_i - \sum_{i=1}^{n} \alpha_i y^{(i)}(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b) - \sum_{i=1}^{n} \beta_i \xi_i.$$

Usando (16), se tiene $-\sum_{i=1}^{n} \beta_i \xi_i - \sum_{i=1}^{n} \alpha_i \xi_i = -C \sum_{i=1}^{n} \xi_i$, que es un término que se anula con el término $C \sum_{i=1}^{n} \xi_i$ que aparece en (6). Reemplazando este desarrollo en (6) y realizando las mismas operaciones que para el SVM de margen duro, se llega a:

$$L(\boldsymbol{\alpha}) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \boldsymbol{x}^{(i)T} \boldsymbol{x}^{(j)} + \sum_{i=1}^{n} \alpha_i, \tag{20}$$

que es exactamente el mismo Lagrangiano que para el caso del margen duro.

Problema Dual. Siguiendo un procedimiento similar al utilizado para el SVM de margen duro, se puede demostrar que el problema dual para el caso del clasificador SVM de margen blando es

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \boldsymbol{x}^{(i)T} \boldsymbol{x}^{(j)} \right\} \tag{21}$$

s.a.

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0, \quad i = 1, \cdots, n$$
$$0 \leq \alpha_i \leq C, \quad i = 1, \cdots, n,$$

donde las restricciones sobre α provienen de (8) y (17). Este problema es similar al problema dual del SVM de margen duro, a excepción de la última restricción, que en este caso indica que cada valor de la variable α_i se encuentra, además, acotado superiormente por el valor de C.

La implementación de la solución a este problema cuadrático, usando un paquete de QPs estándar, se puede realizar de manera similar a como se realizó para el SVM de margen duro. Las diferencias significativas son que en este caso el cálculo de $\boldsymbol{w}^*$ implica el uso de

todos los valores de $\boldsymbol{x}^{(i)}$ y de α_i, y que el cálculo de b^* utiliza los vectores de soporte que están dados cuando $0 < \alpha_i < C$. Además, la restricción correspondiente a la desigualdad se puede separar en dos desigualdades como: $\alpha_i \leq C$ y $-\alpha_i \leq 0$. Siguiendo el mismo esquema usado para el *Hard SVM*, se puede definir un vector $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2 & \cdots \alpha_n\end{bmatrix}^T$, tal que las desigualdades separadas se representen como

$$\begin{bmatrix}\boldsymbol{\alpha} \\ -\boldsymbol{\alpha}\end{bmatrix} \preceq \begin{bmatrix}C_n \\ 0_n\end{bmatrix}, \qquad \text{o} \qquad \begin{bmatrix}I_n \\ -I_n\end{bmatrix}\boldsymbol{\alpha} \preceq \begin{bmatrix}C_n \\ 0_n\end{bmatrix},$$

donde I_n es una matriz identidad de tamaño $n \times n$, C_n es un vector de dimensión n con valores iguales a C, y 0_n es un vector de tamaño n relleno de ceros. Dado que el QP genérico utilizado tiene la restricción de desigualdad representada por $G\boldsymbol{x} \preceq \mathbf{h}$, se puede fácilmente concluir que $G = \begin{bmatrix}I_n \\ -I_n\end{bmatrix}$ y $\boldsymbol{h} = \begin{bmatrix}C_n \\ 0_n\end{bmatrix}$. Todo el resto del QP queda igual que en el caso de SVM de margen duro.

2. Kernels

El SVM de margen blando tiene la ventaja, sobre el SVM de margen duro, que puede clasificar clases que no son linealmente separables, haciendo un balance entre el ancho del margen y el error de clasificación. Este balance está dado por el hiperparámetro C. Sin embargo, este SVM, como tal, no puede clasificar adecuadamente clases cuya separación sea no lineal, dado que es, en su forma básica, un clasificador lineal.

2.1. Bases no lineales y Kernel

Para los casos donde la separación de los datos es no lineal, se puede utilizar una base no lineal $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^m$, donde $m \geq d$, que transforme las entradas $\boldsymbol{x}$ del sistema a vectores no lineales $\boldsymbol{m} = \boldsymbol{\phi}(\boldsymbol{x})$. Si se aplica esta base no lineal a cada instancia $\boldsymbol{x}^{(i)}$, se tendrá entradas $\boldsymbol{m}^{(i)}$. Dado que todo lo que se necesita para resolver el problema de optimización en un SVM es calcular los valores de α_i que maximizan la función objetivo dada en (21), es importante ver el efecto que tiene el usar una base no lineal en esta función objetivo. Por simple inspección, puede verse que el efecto es directo, ya que en lugar de tener $\boldsymbol{x}$ se tendrá su transformación no lineal. De este modo, la función objetivo de (21) queda como

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{m}^{(i)T}\boldsymbol{m}^{(j)} \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\boldsymbol{\phi}(\boldsymbol{x}^{(i)})^T\boldsymbol{\phi}(\boldsymbol{x}^{(j)}), \end{aligned}$$

donde $\boldsymbol{m}^{(i)} = \boldsymbol{\phi}(\boldsymbol{x}^{(i)})$ es la base no lineal. De manera alternativa, esta función objetivo puede escribirse como:

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}),$$

donde

$$K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \boldsymbol{\phi}(\boldsymbol{x}^{(i)})^T\boldsymbol{\phi}(\boldsymbol{x}^{(j)}) \tag{22}$$

es llamado *kernel*. Este kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, representa el producto punto entre sus dos argumentos de entrada luego de que ambos argumentos han sido transformados a través

de la base no lineal $\boldsymbol{\phi}$. La ventaja del uso del kernel, con respecto al uso directo de las bases no lineales radica en que es un número real que puede ser, usualmente, calculado de manera fácil. Además, la función objetivo $L(\alpha)$ solo necesita este producto punto, y no necesita saber de manera explícita el valor de cada uno de los vectores transformados mediante $\boldsymbol{\phi}$. A este hecho de usar el kernel directamente, en lugar de la base no lineal, se le suele denominar el *truco del kernel*. Nótese que al usar el kernel, las restricciones del problema de optimización no se modifican ya que solamente contienen valores de α_i.

Ejemplo 1. Considérese la base no lineal $\boldsymbol{\phi} : \mathbb{R}^2 \to \mathbb{R}^6$ que mapea los vectores bidimensionales $\boldsymbol{x}$ a un espacio de dimensión 6 según: $\boldsymbol{\phi}(\boldsymbol{x}) = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_2^2 & x_1x_2 \end{bmatrix}^T$. El kernel asociado con esta base no lineal es

$$\begin{aligned} K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{z}) &= \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_2^2 & x_1x_2 \end{bmatrix} \begin{bmatrix} 1 & z_1 & z_2 & z_1^2 & z_2^2 & z_1z_2 \end{bmatrix}^T \\ &= 1 + x_1z_1 + x_2z_2 + x_1^2z_1^2 + x_2^2z_2^2 + x_1x_2z_1z_2, \end{aligned}$$

donde $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ y $\boldsymbol{z} = \begin{bmatrix} z_1 & z_2 \end{bmatrix}^T$.

Ejemplo 2. Considérese el kernel $K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ definido como

$$\begin{aligned} K(\boldsymbol{x}, \boldsymbol{z}) &= (1 + \boldsymbol{x}^T \boldsymbol{z})^2 \\ &= (1 + x_1z_1 + x_2z_2)^2 \\ &= 1 + x_1^2z_1^2 + x_2^2x_2^2 + 2x_1z_1 + 2x_2z_2 + 2x_1z_1x_2z_2, \end{aligned}$$

donde $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ y $\boldsymbol{z} = \begin{bmatrix} z_1 & z_2 \end{bmatrix}^T$. Se puede demostrar, de manera operativa, que este kernel surge como el producto punto de dos elementos transformados a través de la base no lineal definida como $\boldsymbol{\phi}(\boldsymbol{x}) = \begin{bmatrix} 1 & x_1^2 & x_2^2 & \sqrt{2}x_1 & \sqrt{2}x_2 & \sqrt{2}x_1x_2 \end{bmatrix}$.

2.2. Función de Hipótesis

Para el caso del SVM usando kernels, la predicción se realiza considerando el hiperplano de separación; es decir, lo que quede a un lado del hiperplano constituye la clase positiva y lo que quede al otro lado constituye la clase negativa. Puesto que se está utilizando la base no lineal dada por $\boldsymbol{\phi}(\boldsymbol{x})$, el hiperplano se define como $\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) + b = 0$, y los elementos de una clase o de la otra tendrán valores diferentes de cero; es decir, o valores positivos, o valores negativos. Así, el signo se asocia directamente con la clase, y la función de hipótesis se expresa como

$$h_w(\boldsymbol{x}) = \operatorname{sgn}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}) + b), \tag{23}$$

donde $\boldsymbol{\phi}(\boldsymbol{x})$ es la base no lineal aplicada a la instancia $\boldsymbol{x}$. El vector de pesos óptimo en (23) es similar al obtenido en (18), pero en este caso se debe reemplazar el valor de $\boldsymbol{x}$ por el valor de la base no lineal $\boldsymbol{\phi}(x)$, quedando

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i y^{(i)} \boldsymbol{\phi}(\boldsymbol{x}^{(i)}). \tag{24}$$

Al reemplazar esta expresión en (23) se tiene

$$h_w(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y^{(i)} \boldsymbol{\phi}(\boldsymbol{x}^{(i)})^T \boldsymbol{\phi}(\boldsymbol{x}) + b \right),$$

donde se observa que solo interesa el producto punto entre $\boldsymbol{\phi}(\boldsymbol{x}^{(i)})$ y $\boldsymbol{\phi}(\boldsymbol{x})$. Habiendo definido este producto punto como el kernel, en (22), la función de hipótesis se puede expresar como

$$h_w(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y^{(i)} K(\boldsymbol{x}^{(i)}, \boldsymbol{x}) + b\right). \tag{25}$$

Nuevamente, se puede ver que no se requiere conocer la expresión de la función $\boldsymbol{\phi}$, sino solamente su producto punto, representado por el kernel.

Para el cálculo de b se sigue un procedimiento similar al usado anteriormente: se busca el hiperplano que contiene los vectores de soporte y a partir de allí se despeja el valor de b. En este caso, dicho hiperplano está dado por $y_{sv}^{(i)}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_{sv}^{(i)}) + b) = 1$, donde el subíndice sv indica que son los vectores de soporte. El hiperplano se puede también expresar como $(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_{sv}^{(i)}) + b) = y_{sv}^{(i)}$, ya que $y_{sv}^{(i)} = \{1, -1\}$. Despenjando el valor de b se tiene $b = y_{sv}^{(i)} - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_{sv}^{(i)})$, y reemplazando el valor del vector de pesos (24), se llega a

$$b = y_{sv}^{(i)} - \sum_{j=1}^{n} \alpha_j y^{(j)} \boldsymbol{\phi}(\boldsymbol{x}^{(j)})^T \boldsymbol{\phi}(\boldsymbol{x}_{sv}^{(i)}).$$

En esta expresión se puede reemplazar el valor del producto punto de las bases no lineales haciendo uso del kernel, como:

$$b = y_{sv}^{(i)} - \sum_{j=1}^{n} \alpha_j y^{(j)} K(\boldsymbol{x}^{(j)}, \boldsymbol{x}_{sv}^{(i)}). \tag{26}$$

Como se observa, ahora para calcular el valor óptimo de b no se requiere la expresión de la base no lineal $\boldsymbol{\phi}(\boldsymbol{x})$ de manera explícita, sino solamente se requiere el kernel. Como se hizo anteriormente, es usual obtener el valor de b promediando todos los valores de b obtenidos usando los diversos vectores de soporte (aquellos que tienen $0 < \alpha_i < C$) que pudiesen existir.

2.3. Condiciones para la existencia de un Kernel

Como se muestra en las secciones anteriores, para realizar una clasificación no lineal con SVMs de margen blando, no es necesario la definición explícita de una base no lineal $\boldsymbol{\phi}(\boldsymbol{x})$ que actúe sobre el vector de entrada $\boldsymbol{x}$, sino que basta con escoger un kernel $K(\boldsymbol{x}, \boldsymbol{z})$ adecuado, donde $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$. Esto trae como consecuencia que se puede definir un kernel incluso sin saber cuál es la base no lineal que lo genera. Sin embargo, no cualquier función que realiza el mapeo de $\mathbb{R}^d \times \mathbb{R}^d$ a $\mathbb{R}$ puede ser un kernel, sino que tiene que satisfacer ciertas restricciones, las cuales se encuentran asociadas con la existencia de una solución para el problema cuadrático de optimización a resolver.

Como se mencionó anteriormente, la expresión cuadrática de la función objetivo dual $L(\alpha)$ se puede expresar sin sumatorias de forma vectorial como

$$\boldsymbol{v}_1^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T M \boldsymbol{\alpha},$$

donde $M = X_y X_y^T$, como se definió anteriormente. Cuando se utiliza un kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, la expresión de $M \in \mathbb{R}^{n\times n}$ se encuentra dada por

$$M = \begin{bmatrix} y^{(1)}y^{(1)}K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}) & y^{(1)}y^{(2)}K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) & \cdots & y^{(1)}y^{(n)}K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(n)}) \\ y^{(2)}y^{(1)}K(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}) & y^{(2)}y^{(2)}K(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(2)}) & \cdots & y^{(2)}y^{(n)}K(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(n)}) \\ \vdots & \vdots & \ddots & \vdots \\ y^{(n)}y^{(1)}K(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(1)}) & y^{(n)}y^{(2)}K(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(2)}) & \cdots & y^{(n)}y^{(n)}K(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(n)}) \end{bmatrix},$$

y el resto del problema no cambia. Esta matriz se puede expresar como

$$M = I_y \overline{K} I_y,$$

donde la matriz $\overline{K}$ contiene todos los elementos asociados al kernel:

$$\overline{K} = \begin{bmatrix} K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}) & K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) & \cdots & K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(n)}) \\ K(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}) & K(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(2)}) & \cdots & K(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(n)}) \\ \vdots & \vdots & \ddots & \vdots \\ K(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(1)}) & K(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(2)}) & \cdots & K(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(n)}) \end{bmatrix},$$

y la matriz I_y es una matriz diagonal cuyos elementos son los valores de salida $y^{(i)}$:

$$I_y = \begin{bmatrix} y^{(1)} & 0 & \cdots & 0 \\ 0 & y^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y^{(n)} \end{bmatrix}.$$

A partir de este desarrollo, es posible demostrar que un kernel K será válido si cumple las siguientes dos condiciones:

1. El kernel es simétrico: $K(\boldsymbol{x}, \boldsymbol{z}) = K(\boldsymbol{z}, \boldsymbol{x})$
2. Se cumple la llamada *condición de Mercer*: la matriz $\overline{K}$ es semidefinida positiva para cualquier valor de $\boldsymbol{x}^{(i)}$, $\boldsymbol{x}^{(j)}$.

Sin embargo, en el uso práctico de SVMs, existen ya algunos kernels bien definidos, cuyas condiciones de validez han sido ya probadas, y que son usados con frecuencia.

2.4. Ejemplos de Kernels

Algunos de los kernels más utilizados para SVM son los siguientes

- *Kernel lineal.* Es un Kernel que no modifica el producto punto, y para fines prácticos es como si no existiese. Está dado por

$$K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}^T \boldsymbol{z}.$$

- *Kernel polinomial de grado Q.* Es un kernel que se comporta como un polinomio y está definido por

$$\begin{aligned} K(\boldsymbol{x}, \boldsymbol{z}) &= (1 + \boldsymbol{x}^T \boldsymbol{z})^Q \\ &= (1 + x_1 z_1 + x_2 z_2 + \cdots + x_d z_d)^Q, \end{aligned}$$

  donde se asume que $\boldsymbol{x}$ y $\boldsymbol{z}$ son vectores con d dimensiones.En algunas ocasiones, se puede ajustar la escala de este Kernel, dándole dos grados adicionales de libertad, definiéndolo como

$$K(\boldsymbol{x}, \boldsymbol{z}) = (r + \gamma \boldsymbol{x}^T \boldsymbol{z})^Q,$$

  donde r y γ son constantes consideradas hiperparámetros.

- *Kernel Gaussiano o RBF.* Es un kernel que se basa en la definición de una función gaussiana y se define matemáticamente como

$$K(\boldsymbol{x}, \boldsymbol{z}) = e^{-\gamma\|\boldsymbol{x}-\boldsymbol{z}\|^2}$$

  donde $\gamma > 0$ es un hiperparámetro. Este kernel también se denomina kernel *RBF* (del inglés *Radial Basis Function*) y es uno de los más utilizados para SVM.

- *Kernel sigmoidal.* Se define como:

$$K(\boldsymbol{x}, \boldsymbol{z}) = \tanh(\gamma \boldsymbol{x}^T \boldsymbol{z} + \tau),$$

  donde γ y τ son hiperparámetros.