

SUPPORT VECTOR MACHINE (SVM)

Vladimir Vapnik dijo: «No existe nada más práctico que una buena teoría».

DEFINICIÓN 1.

Sea un conjunto de entrenamiento $\{(x_i, y_i)\}_{i=1}^N$ donde $x_i \in \mathbb{R}^n$ son los vectores de características y $y_i \in \{-1, 1\}$ son las etiquetas de clase.

El modelo SVM busca determinar un hiperplano, definido por

$$w \cdot x + b = 0,$$

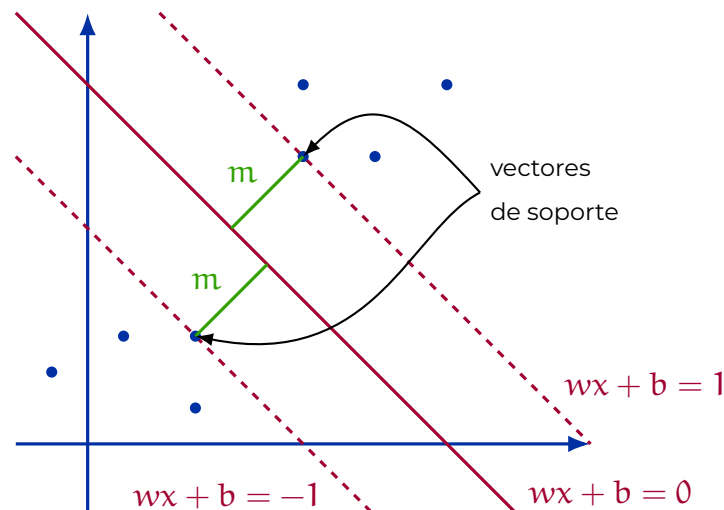
que separe los datos, es decir

$$w \cdot x_i + b \geq 1 \text{ si } y_i = 1 \quad \text{y} \quad w \cdot x_i + b \leq -1 \text{ si } y_i = -1.$$

La condición puede tomarse como:

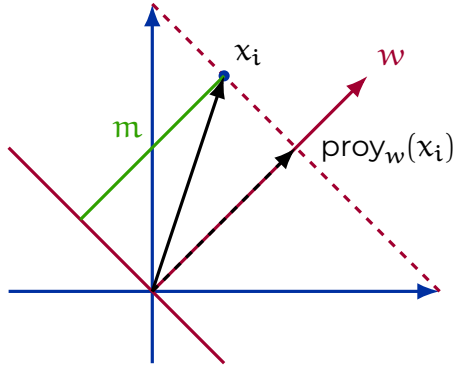
$$y_i(w \cdot x_i + b) \geq 1$$

A los vectores que alcanzan la igualdad de la condición, es decir, $y_i(w \cdot x_i + b) = 1$, se los llama **vectores de soporte**.



La distancia de los vectores de soporte al hiperplano separador es m (margen). El objetivo es maximizar el margen, que es la distancia entre los hiperplanos por los que pasan los vectores de soporte: $2m$.

Para determinar el valor de t , consideremos momentáneamente que el hiperplano pasa por el origen (es decir, $b = 0$).



La distancia m desde el punto x_i al hiperplano se puede calcular como la norma de la proyección del vector x_i sobre el vector normal al hiperplano w :

$$m = \|\text{proy}_w(x_i)\| = \left\| \frac{x_i \cdot w}{\|w\|^2} \cdot w \right\|,$$

además, como x_i está sobre el hiperplano de ecuación $w \cdot x = 1$, se cumple que $w \cdot x_i = 1$. Por lo tanto:

$$m = \left\| \frac{1}{\|w\|^2} \cdot w \right\| = \frac{1}{\|w\|}.$$

Con esto, lo que debemos maximizar es $2m = \frac{2}{\|w\|}$, lo cual es equivalente a minimizar $\frac{1}{2}\|w\|^2$; por conveniencia, se minimiza el cuadrado de la norma. Así, el problema de optimización queda:

$$\begin{cases} \text{mín} & f(w, b) = \frac{1}{2}\|w\|^2 \\ \text{sujeto a} & \\ & y_i(w \cdot x_i + b) \geq 1, \quad i = 1, \dots, N \end{cases}$$

Para resolver este problema de optimización con restricciones, se utiliza el método de los multiplicadores de Lagrange. Planteamos el Lagrangiano usando multiplicadores de Lagrange $\alpha_i \geq 0$:

$$L(w, b, \alpha_1, \dots, \alpha_N) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^N \alpha_i (y_i(w \cdot x_i + b) - 1),$$

desarrollando:

$$L(w, b, \alpha_1, \dots, \alpha_N) = \frac{1}{2}w \cdot w - w \cdot \sum_{i=1}^N \alpha_i y_i x_i - b \sum_{i=1}^N \alpha_i y_i + \sum_{i=1}^N \alpha_i.$$

Con esto, obtenemos las condiciones de optimalidad, derivando parcialmente el Lagrangiano con respecto a w y b e igualando a cero:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^N \alpha_i y_i x_i = 0 \quad \Rightarrow \quad w = \sum_{i=1}^N \alpha_i y_i x_i$$

y

$$\frac{\partial L}{\partial b} = - \sum_{i=1}^N \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^N \alpha_i y_i = 0.$$

Reemplazamos las condiciones de optimalidad en el Lagrangiano para obtener el problema dual:

$$L(w, b, \alpha_1, \dots, \alpha_N) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j).$$

El problema dual de maximización es:

$$\left\{ \begin{array}{l} \text{máx} \quad L_D(\alpha_1, \dots, \alpha_N) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ \text{sujeto a} \\ \alpha_i \geq 0, \quad i = 1, \dots, N, \\ \sum_{i=1}^N \alpha_i y_i = 0. \end{array} \right.$$

1. FUNCIONES KERNEL

DEFINICIÓN 2.

Una función kernel es una función

$$K: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$$

que cumple las propiedades distributiva, conmutativa y semidefinida positiva.

Dada una función de mapeo

$$\phi: \mathbb{R}^n \rightarrow \mathbb{R}^m,$$



se define el kernel asociado como

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j).$$

Al usar funciones kernel, el problema dual de maximización queda:

$$\left\{ \begin{array}{l} \text{máx} \quad L_D(\alpha_1, \dots, \alpha_N) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{sujeto a} \\ \alpha_i \geq 0, \quad i = 1, \dots, N, \\ \sum_{i=1}^N \alpha_i y_i = 0. \end{array} \right.$$

Algunas funciones kernel comunes son:

- Kernel lineal: $K(x_i, x_j) = x_i \cdot x_j$.

- Kernel polinomial: $K(x_i, x_j) = (\alpha x_i \cdot x_j + c)^p$, con $\alpha > 0$, $c \geq 0$ y $p \in \mathbb{N}$.
- Kernel radial: $K(x_i, x_j) = \exp(-\alpha \|x_i - x_j\|^2)$, con $\alpha > 0$.