

# Máquinas de Vector Soporte

Norberto Corral, Carlos Carleos

12 de diciembre de 2024

# Máquinas de Vector Soporte

Dados  $\vec{w} \in \mathbb{R}^p$  y  $b \in \mathbb{R}$ , fijos, se define una aplicación lineal  $f : \mathbb{R}^p \rightarrow \mathbb{R}$  como:

$$f(\vec{x}) = \langle \vec{w}, \vec{x} \rangle + b = \vec{w}' \vec{x} + b = \sum_{j=1}^p w_j x_j + b$$

Se llama hiperplano,  $\pi$ , al subconjunto de puntos  $\mathbb{R}^p$  que verifican:

$$\pi = \{ \vec{x} \mid f(\vec{x}) = 0 \}$$

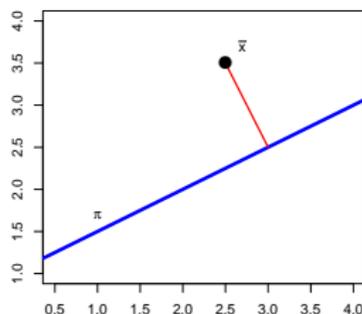
Cualquier hiperplano de  $\mathbb{R}^2$  define una recta

# Máquinas de Vector Soporte

El vector  $\vec{w}$  es ortogonal a todos los puntos del hiperplano

La distancia entre un hiperplano  $\pi$  y un punto  $\vec{x} \in \mathbb{R}^P$  viene dada por la expresión:

$$\text{dist}(\pi, \vec{x}) = \frac{f(\vec{x})}{\|\vec{w}\|} \quad \text{con} \quad \|\vec{w}\| = \sqrt{\sum_{j=1}^P w_j^2}$$



## Conjunto linealmente separable

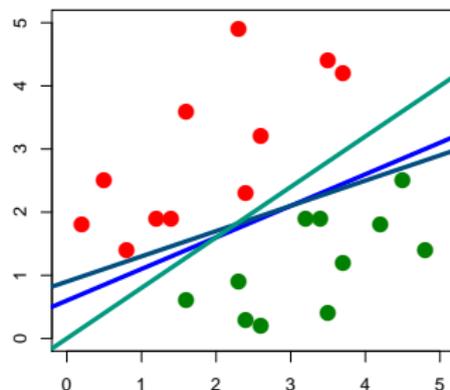
Un conjunto de puntos  $(\vec{x}_i, y_i)_{i=1}^n$  con  $\vec{x}_i \in \mathbb{R}^p$ ,  $y_i \in \{-1, +1\}$ , es linealmente separable si existe un hiperplano  $\pi$  tal que:

- ▶  $f(\vec{x}_i) = (\langle \vec{w}, \vec{x}_i \rangle + b) \geq 0$  si  $y_i = +1$
- ▶  $f(\vec{x}_i) = (\langle \vec{w}, \vec{x}_i \rangle + b) \leq 0$  si  $y_i = -1$

Esas dos condiciones son equivalentes a:

$$y_i f(\vec{x}_i) = y_i (\langle \vec{w}, \vec{x}_i \rangle + b) \geq 0 \quad i = 1, \dots, n$$

# Máquinas de Vector Soporte



Cuando existe un hiperplano de separación un criterio razonable para clasificar un punto  $\vec{x}$  es:

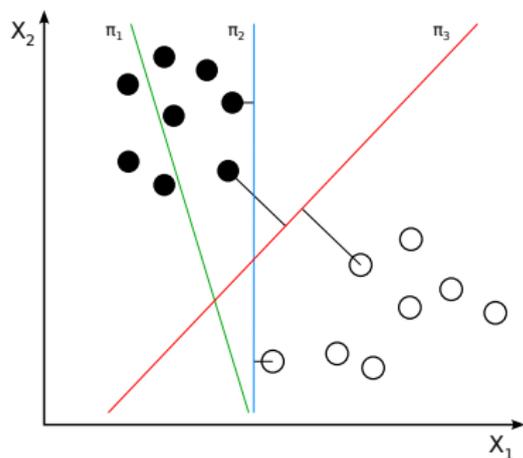
$$G(\vec{x}) = \text{signo}(f(\vec{x}))$$

## Margen de un hiperplano de separación

Dado un conjunto de puntos  $(\vec{x}_i)_{i=1}^n$  con  $\vec{x}_i \in \mathbb{R}^P$ , y un hiperplano de separación  $\pi = \{\vec{x} \mid \vec{w}' \cdot \vec{x} + b = 0\}$  se define su margen como el valor  $M$  tal que

$$M = \min_{i=1, \dots, n} \text{dist}(\vec{x}_i, \pi)$$

- ▶  $\pi_1$  no es hiperplano de separación
- ▶  $\pi_2$  es hiperplano de separación con margen pequeño
- ▶  $\pi_3$  es el hiperplano de separación con margen máximo



## Hiperplano óptimo

Es el hiperplano de separación con margen máximo.

El problema se puede plantear como la búsqueda del hiperplano óptimo de separación, es decir, :

Maximizar  $M$

$$y_i \cdot \frac{\langle \vec{w}, \vec{x} \rangle + b}{\|\vec{w}\|} \geq M$$

Imponiendo la condición

$$M \cdot \|\vec{w}\| = 1 \iff M = \frac{1}{\|\vec{w}\|}$$

$$\text{Maximizar } \frac{1}{\|\vec{w}\|} \iff \text{Minimizar } \|\vec{w}\|$$
$$y_i [\langle \vec{w}, \vec{x} \rangle + b] \geq 1 \iff y_i [\langle \vec{w}, \vec{x} \rangle + b] - 1 \geq 0$$

# Máquinas de Vector Soporte

- ▶ Problema de optimización cuadrática con restricciones lineales.
- ▶ La función lagrangiana es

$$L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2} \langle \vec{w}, \vec{w} \rangle - \sum_{i=1}^n \alpha_i [y_i (\langle \vec{w}, \vec{x}_i \rangle + b) - 1]$$

- ▶ Aplicando las condiciones de Karush-Kuhn-Tucker se obtiene

$$\frac{\partial L}{\partial \vec{w}} = 0 \iff \vec{w}^* = \sum_{i=1}^n \alpha_i \cdot y_i \cdot \vec{x}_i \quad i = 1, \dots, n$$

$$\frac{\partial L}{\partial b} = 0 \iff \sum_{i=1}^n \alpha_i \cdot y_i = 0 \quad i = 1, \dots, n$$

$$\alpha_i [1 - y_i (\langle \vec{w}^*, \vec{x}_i \rangle + b^*)] = 0$$

$$L(\vec{\alpha}) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \vec{x}_i, \vec{x}_j \rangle$$

Se puede plantear el problema inicial en su forma dual:

Maximizar  $L(\alpha)$

$$\sum_{i=1}^n \alpha_i \cdot y_i = 0$$

$$\alpha_i \geq 0 \quad i = 1, \dots, n$$

Una vez calculado  $\alpha^*$ , se obtiene

$$f(\vec{x}) = \sum_{i=1}^n \alpha_i^* \cdot y_i \cdot \langle \vec{x}, \vec{x}_i \rangle + b^*$$

Analizando la condición

$$\alpha_i - \alpha_j \cdot y_i \cdot \langle \vec{w}^*, \vec{x}_i \rangle + b^* = 0$$

- ▶  $\alpha_i > 0 \iff 1 = y_i \cdot (\langle \vec{w}^*, \vec{x}_i \rangle + b^*)$   
que corresponde a los vectores soporte

$$b^* = y_{vs} - \langle \vec{w}^*, \vec{x}_{vs} \rangle$$

- ▶  $\alpha_i = 0$  el punto ' $x_i$ ' no interviene en el hiperplano óptimo

La estimación de  $b^*$  se hace con todos los vectores soporte:

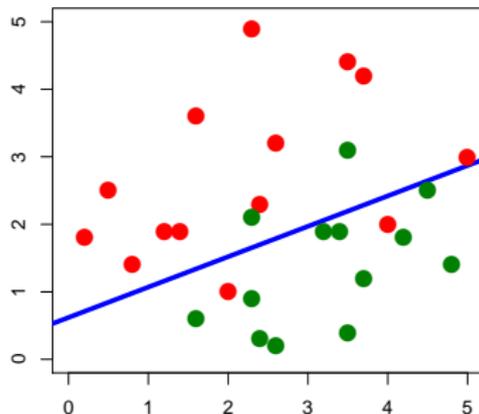
$$b^* = \frac{1}{N_{vs}} \sum_{k=1}^{N_{vs}} (y_{vs} - \langle \vec{w}^*, \vec{x}_{vs} \rangle)$$

# Máquinas de Vector Soporte

En la mayor parte de los problemas reales no existen hiperplanos de separación y se plantea la búsqueda de hiperplanos que verifiquen unas condiciones más suaves:

$$y_i \cdot (\langle \vec{w}, \vec{x}_i \rangle + b) \geq 1 - \xi_i \quad \text{con } \xi_i \geq 0 \quad i = 1, \dots, n$$

- ▶ Las variables  $\xi_i$  se denominan de holgura y permiten la existencia de puntos mal clasificados.
- ▶ Cuando  $\xi_i = 0$ , es el problema anterior.



El nuevo problema de optimización viene dado por:

$$L(\vec{w}, \vec{\xi}) = \frac{1}{2} \langle \vec{w}, \vec{w} \rangle + C \sum_{i=1}^n \xi_i$$

sujeto a las restricciones

$$y_i \cdot (\langle \vec{w}, \vec{x}_i \rangle + b) + \xi_i - 1 \geq 0; \quad \xi_i \geq 0; \quad i = 1, \dots, n$$

$C$  representa la penalización de los puntos mal clasificados.

Estos hiperplanos se denominan de margen blando.

# Máquinas de Vector Soporte

- ▶ La función lagrangiana del hiperplano de margen blando es:

$$\begin{aligned} L(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) &= \\ &= \frac{1}{2} \langle \vec{w}, \vec{w} \rangle + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [y_i (\langle \vec{w}, \vec{x}_i \rangle + b) + \xi_i - 1] - \sum_{i=1}^n \beta_i \xi_i \end{aligned}$$

- ▶ Aplicando las condiciones de Karush-Kuhn-Tucker se obtiene

$$\forall i = 1, \dots, n : \quad \frac{\partial L}{\partial \vec{w}} = 0 \iff \vec{w}^* = \sum_{i=1}^n \alpha_i \cdot y_i \cdot \vec{x}_i$$

$$\frac{\partial L}{\partial b} = 0 \iff \sum_{i=1}^n \alpha_i \cdot y_i = 0$$

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

$$\alpha_i [1 - y_i (\langle \vec{w}^*, \vec{x}_i \rangle + b^*) - \xi_i] = 0$$

$$\beta_i \xi_i = 0$$

$$L(\vec{\alpha}) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \vec{x}_i, \vec{x}_j \rangle$$

Se puede plantear el problema inicial en su forma dual:

Maximizar  $L(\alpha)$

$$\sum_{i=1}^n \alpha_i \cdot y_i = 0$$

$$0 \leq \alpha_i \leq C \quad i = 1, \dots, n$$

Una vez calculado  $\alpha^*$ , se obtiene

$$f(\vec{x}) = \sum_{i=1}^n \alpha_i^* \cdot y_i \cdot \langle \vec{x}, \vec{x}_i \rangle + b^*$$

# Máquinas de Vector Soporte

Teniendo en cuenta las condiciones que debe cumplir la solución

$$C = \alpha_i - \beta_i; \quad \beta_i \xi_i = 0, \quad i = 1 \dots, n$$

$$\alpha_i [1 - y_i (\langle \vec{w}^*, \vec{x}_i \rangle + b^*) - \xi_i] = 0, \quad i = 1 \dots, n$$

se pueden caracterizar los puntos  $\vec{x}_i$ :

- ▶ Si  $\vec{x}_i$  no es separable  $\iff \xi_i > 0$ . Por tanto  $\beta_i = 0$ ,  $\alpha_i = C$  y

$$y_i (\langle \vec{w}^*, \vec{x}_i \rangle + b^*) = 1 + \xi_i$$

- ▶ Si  $\alpha_i = 0$  entonces  $\beta_i = C$ ,  $\xi_i = 0$  y el punto  $\vec{x}_i$  es separable.
- ▶ Cuando  $0 < \alpha_i < C \iff \beta_i \neq 0 \iff \xi_i = 0$ . Por lo tanto

$$y_i (\langle \vec{w}^*, \vec{x}_i \rangle + b^*) = 1$$

es decir,  $\vec{x}_i$  es un vector soporte y permiten estimar  $b^*$

$$b^* = \frac{1}{N_{vs}} \sum_{k=1}^{N_{vs}} (y_k - \langle \vec{w}^*, \vec{x}_k \rangle)$$

Cuando las clases no son separables se plantea “pasar” los datos a un espacio de dimensión mayor en el que se cumpla la condición de separabilidad.

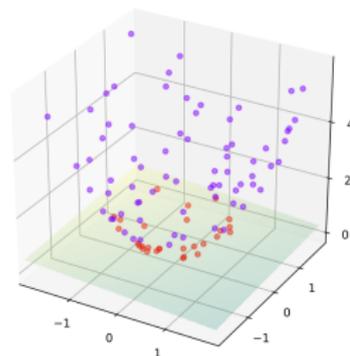
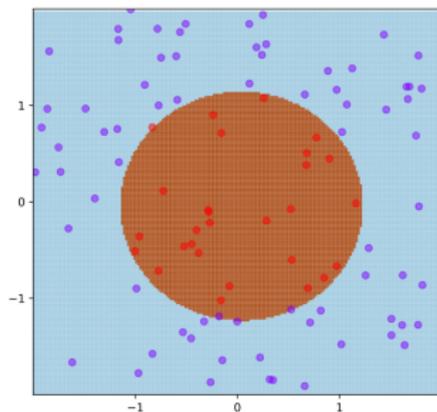
El espacio transformado se denomina espacio de características.

$$\begin{aligned}\Phi : \mathbb{X} &\rightarrow F \\ \vec{x} &\mapsto \Phi(\vec{x}) = (\phi_1(\vec{x}), \dots, \phi_m(\vec{x}))\end{aligned}$$

donde alguna de las funciones  $\phi_j$  no es lineal.

El objetivo es buscar el hiperplano óptimo, en el espacio de características, que produce una frontera no lineal en el espacio inicial.

# Máquinas de Vector Soporte



Transformación (núcleo):  $(a, b) \rightarrow (a, b, a^2 + b^2)$

## Funciones Núcleo

Sea  $K$  una función  $K : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$   
que verifica las condiciones

1.  $K(\vec{x}_1, \vec{x}_2) = K(\vec{x}_2, \vec{x}_1)$  (simétrica)
2.  $K(\vec{x}, \vec{x}) \geq 0$  (semidefinida positiva)

Algunos ejemplos de funciones núcleo ( $\lambda > 0$ ,  $\gamma > 0$ ):

- ▶  $K(\vec{x}_1, \vec{x}_2) = \langle \vec{x}_1, \vec{x}_2 \rangle$
- ▶  $K(\vec{x}_1, \vec{x}_2) = (\lambda \langle \vec{x}_1, \vec{x}_2 \rangle + \gamma)^p$
- ▶  $K(\vec{x}_1, \vec{x}_2) = \exp(-\lambda \langle \vec{x}_1 - \vec{x}_2, \vec{x}_1 - \vec{x}_2 \rangle) = \exp(-\lambda \|\vec{x}_1 - \vec{x}_2\|^2)$

## Teorema de Moore-Aronszajn

Para cualquier función  $K : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$  simétrica y semidefinida positiva, existe un espacio de Hilbert (normado y completo) y una función  $\Phi : \mathbb{X} \rightarrow F$  tal que

$$K(\vec{x}, \vec{x}') = \langle \Phi(\vec{x}), \Phi(\vec{x}') \rangle$$

Este resultado permite calcular el producto escalar  $\langle \Phi(\vec{x}), \Phi(\vec{x}') \rangle$  sin necesidad de conocer la función  $\Phi$ .

# Máquinas de Vector Soporte

En el espacio de características se busca la función:

$$f(\vec{x}) = \langle \vec{w}, \Phi(\vec{x}) \rangle + b$$

que en su forma dual se expresa como:

$$f(\vec{x}) = \sum_{i=1}^n \alpha_i^* y_i K(\vec{x}, \vec{x}_i)$$

Se resuelve mediante el siguiente problema de optimización

$$\text{Max} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \vec{x}_i, \vec{x}_j \rangle$$

$$\sum_{i=1}^n \alpha_i \cdot y_i = 0$$

$$0 \leq \alpha_i \leq C \quad i = 1, \dots, n$$

El problema XOR

Caso	$(X_1, X_2)$	Y
1	(+1, +1)	-1
2	(-1, -1)	-1
3	(+1, -1)	+1
4	(-1, +1)	+1

Para representar el producto escalar en el espacio de las características se usa el siguiente kernel polinómico:

$$K(\vec{x}, \vec{x}') = (\langle \vec{x}, \vec{x}' \rangle + 1)^2$$

# Máquinas de Vector Soporte

El problema dual a resolver es:

$$\text{Max} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(\vec{x}_i, \vec{x}_j)$$

$$\sum_{i=1}^n \alpha_i \cdot y_i = 0$$

$$0 \leq \alpha_i \leq C \quad i = 1, \dots, n$$

La solución de  $\alpha$  es:

$$\alpha^* = 0,125 \quad i = 1, \dots, 4$$

La función de clasificación que se obtiene es:

$$f(\vec{x}) = 0,125 * \sum_{i=1}^4 y_i k(\vec{x}, \vec{x}_i)$$

En este caso se pueden obtener las funciones de transformación.

$$\langle \Phi(\vec{x}), \Phi(\vec{x}') \rangle = k(\vec{x}, \vec{x}') = (\langle \vec{x}, \vec{x}' \rangle + 1)^2 =$$

$$1 + x_1^2(x_1')^2 + x_2^2(x_2')^2 + 2x_1x_2x_1'x_2' + 2x_1x_1' + 2x_2x_2'$$

con

$$\Phi = (\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6)$$

$$f(\vec{x}) = \frac{1}{\sqrt{2}}\phi_4(\vec{x}) = x_1x_2$$

## Puntos débiles de las Máquinas de Vector Soporte

- ▶ Las MVS son muy sensibles a los parámetros que intervienen en su cálculo. Es aconsejable probar con diferentes valores y analizar los resultados y su estabilidad.
- ▶ En los problemas de clasificación es preferible usar un núcleo gaussiano y la función objetivo basada en el coste  $C$ . En este caso sólo son necesarios los parámetros  $\lambda$  y  $C$ . Estrategias posibles de búsqueda:
  - ▶ Buscar un  $C$  adecuado probando con valores entre 1 y 1000, usando validación cruzada. Con el  $C$  seleccionado buscar el  $\lambda$  adecuado.
  - ▶ Buscar simultáneamente sobre  $\lambda$  y  $C$ , usando una malla.
- ▶ Las MVS son sensibles a las unidades de las variables y es conveniente realizar una tipificación previa.

La librería `e1071` de R puede usarse para aplicar las máquinas de vector soporte.

```
library (e1071)
mvs_salida <- svm (y ~ x, datos.entrena)
mvs_pred   <- predict (mvs_salida, datos.prueba)
## para optimizar parámetros en una malla:
tune (svm, y ~ x, data = datos.entrena,
      ranges = list (cost=10^(-1:3), gamma=10^(-3:0)))
```

Por omisión, usa núcleo gaussiano (radial) donde  $\lambda = \text{gamma}$ .