Machine Learning I

Tema 2. Aprendizaje Supervisado

Máquinas de Vectores Soporte

Profesor: José Luis Sancho Gómez

Curso 2023-2024

Máquinas de Vectores Soporte (SVM)

Contenido

1.1. SVM lineales para datos separables linealmente ("hard margin optimization")

Discriminante lineal de máximo margen

"Support Vector Machines"

Vapnik, V. (Boser, Guyon, and Vapnik, 1992; Cortes and Vapnik, 1995, Vapnik, 1995, 1998)

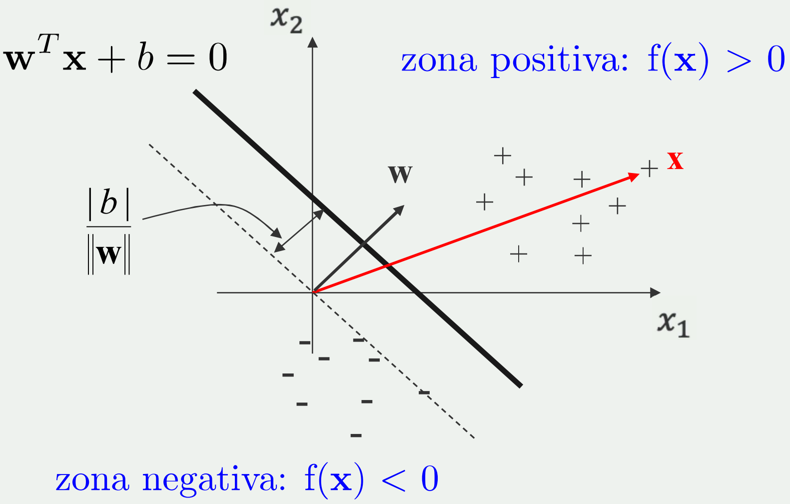
Básicamente son "máquinas lineales" con ciertas propiedades atractivas

Supóngase que se dispone de:

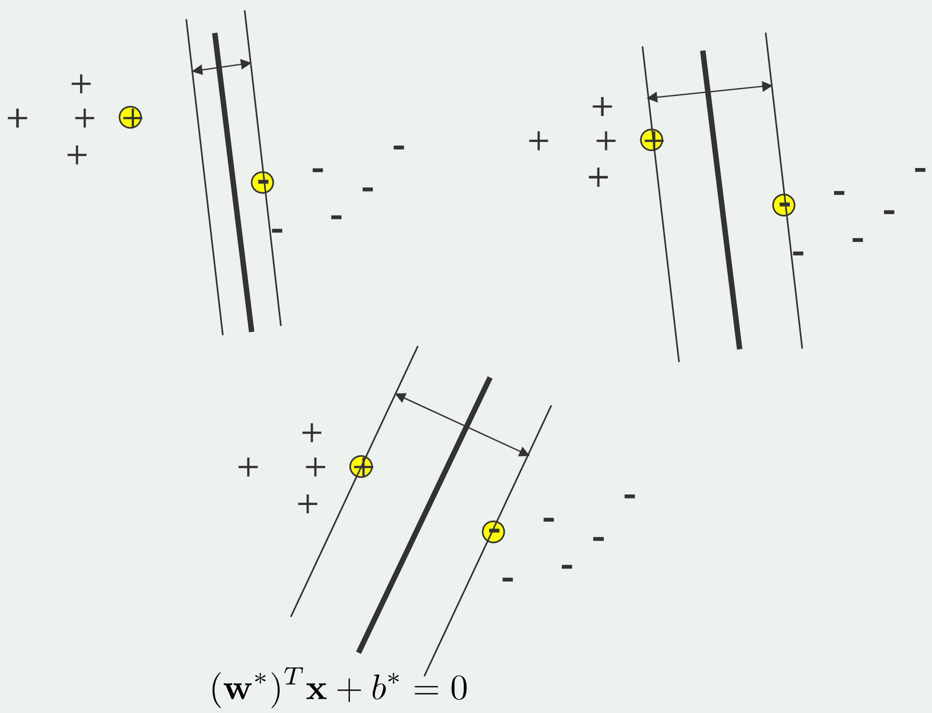
- Conjunto de datos: $X = \{\mathbf{x}_n, d_n\}_{n=1}^{N}, \text{ con } d = \pm 1$
- *Linealmente separables*

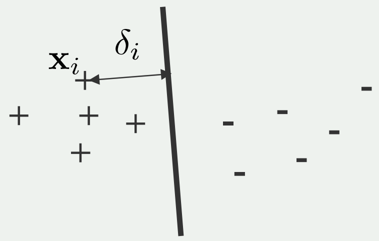
Función discriminante:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$



Concepto de margen (de error)





Margen:

$$\gamma = \min_i \delta_i$$

Discriminante de máximo margen:

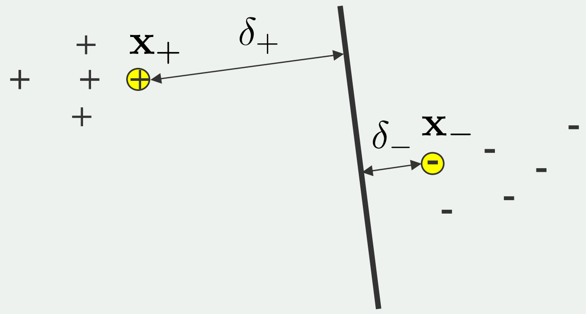
$$(\mathbf{w}^*, b^*) = \arg\max_{\mathbf{w},b} \gamma$$

$$\Downarrow$$

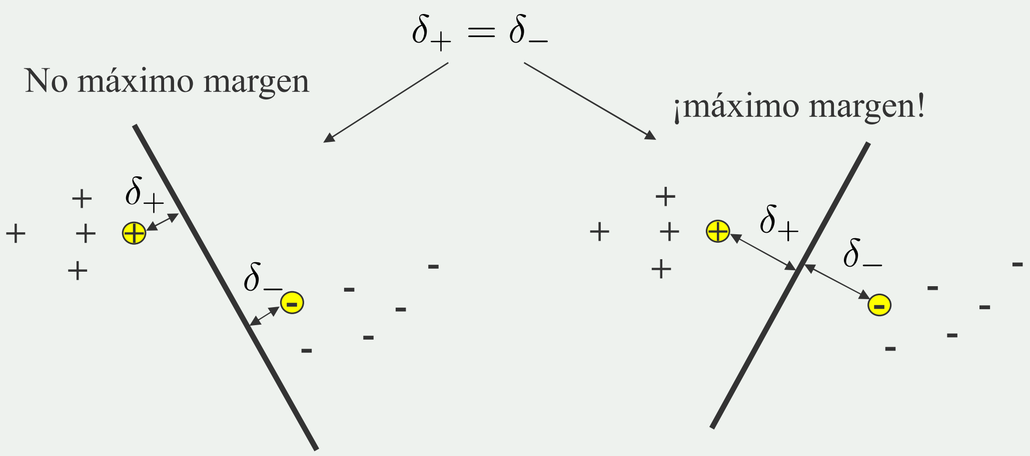
$$\delta_+ = \delta_-$$

$$\Downarrow$$

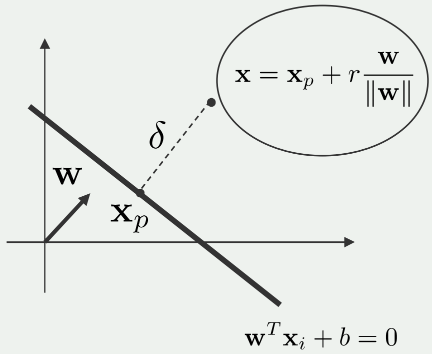
$$|f(\mathbf{x}_+)| = |f(\mathbf{x}_-)|$$



Condición necesaria (no suficiente)



Distancia de un punto a un plano



$r > 0 \to f(\mathbf{x}) > 0 \to$ zona positiva

$r < 0 \to f(\mathbf{x}) < 0 \to$ zona negativa

$r = 0 \to f(\mathbf{x}) = 0 \to$ plano

$$f(\mathbf{x}_p) = \mathbf{w}^T\mathbf{x}_p + b = 0$$

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = \mathbf{w}^T(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}) + b = r\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|} = r\|\mathbf{w}\| \longrightarrow r = \frac{f(\mathbf{x})}{\|\mathbf{w}\|}$$

$$\boxed{\delta = |r| = \frac{|f(\mathbf{x})|}{\|\mathbf{w}\|}}$$

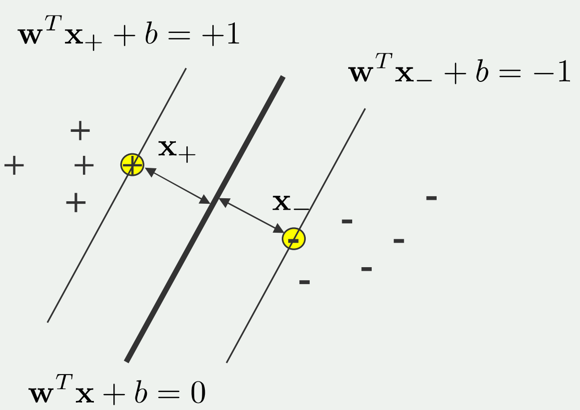
2.4. Máquinas de Vectores Soporte

$$\textbf{Escalado:} \quad \begin{aligned} \mathbf{w} &\to \alpha\mathbf{w} \\ b &\to \alpha b \end{aligned}$$

$$|f(\mathbf{x})| = d_i(\mathbf{w}^T\mathbf{x}_i + b) \quad \to \quad \delta_i(\alpha\mathbf{w},\ \alpha b) = \frac{\alpha d_i(\mathbf{w}^T\mathbf{x}_i + b)}{\alpha\|\mathbf{w}\|} = \frac{d_i(\mathbf{w}^T\mathbf{x}_i + b)}{\|\mathbf{w}\|} = \delta_i(\mathbf{w},\ b)$$

¡Las distancias no cambian por el escalado!
(tampoco el hiperplano: $\alpha\mathbf{w}^T\mathbf{x} + \alpha b = 0$)

Escalado especial:



Dos consecuencias:

$$\gamma = \frac{1}{\|\mathbf{w}\|}$$

$$\min_i\{d_i(\mathbf{w}^T\mathbf{x}_i + b)\} = 1$$

Problema de optimización

A)

$$\max_{\mathbf{w}} \frac{1}{\|\mathbf{w}\|}$$
$$\text{sujeta a: } \min_i\{d_i(\mathbf{w}^T\mathbf{x}_i + b)\} = 1$$

De muy difícil solución por el mínimo en las restricciones
SOLUCIÓN: escribirlo de otra forma

B)

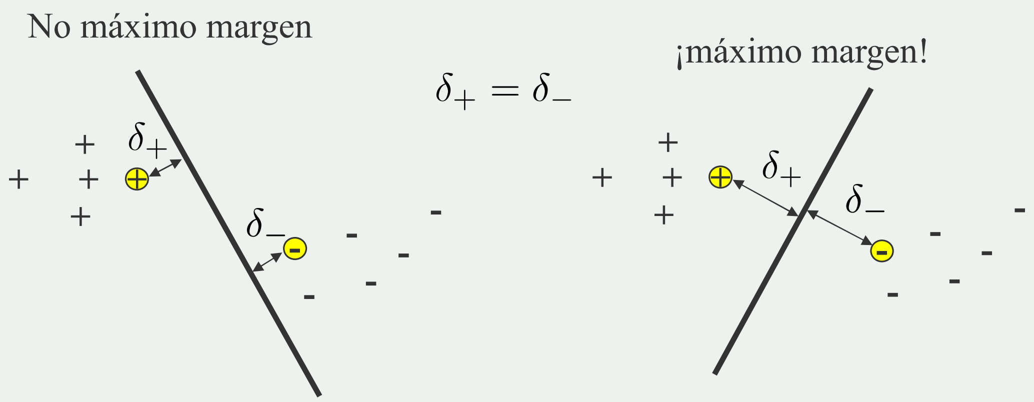
$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{sujeto a: } d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0$$

Problema "primal"

¿Son A) y B) realmente equivalentes?

CONCLUSIÓN

El discriminante de máximo margen es aquel que, de entre todos los discriminantes que cumplen $d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \ \forall i$, tiene un valor de $\|\mathbf{w}\|$ mínimo.



$$\|\mathbf{w_1}\| > \|\mathbf{w_2}\|$$

Problema original ('primal problem')

Dado el conjunto de entrenamiento separable linealmente $X = \{\mathbf{x}_n, d_n\}_{n=1}^N$, el hiperplano $(\mathbf{w}^*, b^*)$ que resuelve el problema de optimización cuadrática

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{sujeto a: } \; d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0$$

es el hiperplano de máximo margen.

2.4. Máquinas de Vectores Soporte

El problema original se puede resolver mediante **programación cuadrática** si la dimensión de los datos no es muy elevada. En general, es muy complejo.

Mejor **transformar el problema a su dual**. ¿Porqué?

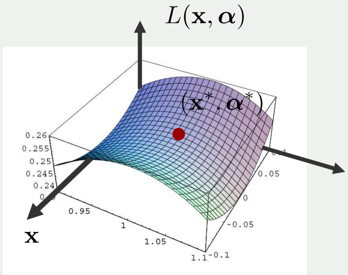
- Su complejidad depende del número de muestras N y no de su dimensión.
- Permite introducir la extensión a problemas no lineales mediante el denominado 'kernel trick'.

Principio de dualidad

Original	Dual
$\min_{\mathbf{x}} \phi(\mathbf{x})$ sujeto a: $g_i(\mathbf{x}) \geq 0,\ i = 1, \ldots, N$ con $\phi(\mathbf{x})$ función convexa en $\mathbf{x}$ $g_i(\mathbf{x})$ restriciones lineales en $\mathbf{x}$	$\max_{\boldsymbol{\alpha}} \theta(\boldsymbol{\alpha})$ sujeto a: $\boldsymbol{\alpha} \geq 0$ con $\theta(\boldsymbol{\alpha}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\alpha})$ $L(\mathbf{x}, \boldsymbol{\alpha}) = \phi(\mathbf{x}) - \sum_{i=1}^{N} \alpha_i g_i(\mathbf{x})$

L: Lagrangiano

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$: Multiplicadores de Lagrange



En nuestro caso:

$$\mathbf{x} \to (\mathbf{w}, b)$$
$$\phi(\mathbf{w}, b) \to \frac{1}{2}\|\mathbf{w}\|^2$$
$$g_i(\mathbf{w}, b) \to d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1, \quad i = 1, \ldots, N$$

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i \{d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1\}$$

$\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\alpha}) \longrightarrow$

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = 0 \ \to \ \mathbf{w} = \sum_{i=1}^{N} d_i \alpha_i \mathbf{x}_i$$
$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = 0 \ \to \ \sum_{i=1}^{N} d_i \alpha_i = 0$$

$$L(\boldsymbol{\alpha}) \ \to \ \theta(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{m=1}^{N} \alpha_i \alpha_m d_i d_m \mathbf{x}_i^T \mathbf{x}_m$$

Problema dual ('dual problem')

Dado el conjunto de entrenamiento separable linealmente $X = \{\mathbf{x_n}, d_n\}_{n=1}^{N}$, el conjunto de parámetros $\boldsymbol{\alpha}^*$ que resuelve el problema de optimización cuadrática

$$\max_{\boldsymbol{\alpha}} \left\{ \theta(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{m=1}^{N} \alpha_i \alpha_m d_i d_m \mathbf{x}_i^T \mathbf{x}_m \right\}$$

$$\text{Sujeto a:} \quad \sum_{i=1}^{N} d_i \alpha_i = 0$$

$$\alpha_i \geq 0$$

con $i = 1, \ldots, N$, determina el discriminante de máximo margen, es decir, $\gamma^* = \frac{1}{\|\mathbf{w}^*\|}$, con $\mathbf{w}^* = \sum_{i=1}^{N} d_i \alpha_i^* \mathbf{x}_i$.

Importante

Condiciones de Karush-Kuhn-Tucker (KKT)

Sea un problema de optimización en un conjunto convexo $\mathbf{x} \in S$

$$\min_{\mathbf{x}} \phi(\mathbf{x})$$

$$\text{sujeto a:} \quad g_i(\mathbf{x}) \geq 0, \; i = 1, \ldots, N$$

con ϕ convexa y g_i lineales (afines). Son condiciones necesarias y suficientes para que $\mathbf{x}^*$ sea solución las siguientes:

$$\frac{\partial L(\mathbf{x}^*, \boldsymbol{\alpha})}{\partial \mathbf{x}} = 0$$

$$g_i(\mathbf{x}^*) \geq 0, \; i = 1, \ldots, N$$

$$\alpha_i \geq 0, \; i = 1, \ldots, N$$

$$\alpha_i g_i(\mathbf{x}^*) = 0, \; i = 1, \ldots, N$$

Condiciones KKT

con $\quad L(\mathbf{x}, \alpha) = \phi(\mathbf{x}) - \sum_{i=1}^{N} \alpha_i g_i(\mathbf{x})$

Condiciones KKT en nuestro caso

El Lagrangiano es

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i\{d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1\}$$

y la solución debe cumplir KKT:

$$\frac{\partial L(\mathbf{x}^*, \boldsymbol{\alpha})}{\partial \mathbf{x}} = 0$$
$$d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0$$
$$\alpha_i \geq 0$$
$$\alpha_i(d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1) = 0 \longleftarrow \text{Importante}$$

Ahora nos interesa especialmente la última condición KKT:

$$\alpha_i^*(d_i(\mathbf{x}^{*T}\mathbf{x}_i + b^*) - 1) = 0, \ i = 1, \ldots, N$$

que dice que los parámetros α_i^* son distintos de cero sólo si la muestra correspondiente $(\mathbf{x}_i, d_i)$ satisface $d_i(\mathbf{w}^*\mathbf{x}_i + b^*) = 1$. A estas muestras se las llama **vectores soporte**.

Solución:

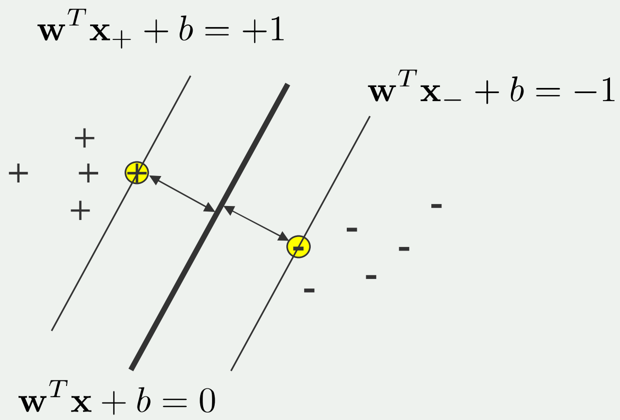
$$\mathbf{w}^* = \sum_{i \in SV} d_i \alpha_i^* \mathbf{x}_i$$
$$b^* = 1 - \mathbf{w}^*\mathbf{x}^{s+}$$

¡Sólo aparecen los vectores soporte!

$$f(\mathbf{x}, \mathbf{w}^*, b^*) = \mathbf{w}^{*T}\mathbf{x} + b^* = \sum_{i \in SV} d_i \alpha_i^* \mathbf{x}_i^T \mathbf{x} + b^*$$

CONCLUSIÓN

La solución del problema de máximo margen para el caso de separabilidad lineal viene expresada como una combinación lineal de unos cuantos datos que son los vectores soporte.



Resultado importante sobre la generalización

$$E_{\text{Gen}} \leq \frac{\#\text{ de SV's}}{N-1}$$

El número de SVs (algo medible) constituye una cota superior del error de generalización (algo difícil de medir).

El problema dual se resuelve mediante *programación cuadrática* (QP) que opera sobre minimizaciones, no maximizaciones. Por tanto, haciendo: $\max_{\boldsymbol{\alpha}} \theta(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}} -\theta(\boldsymbol{\alpha})$ se obtiene

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{N} \alpha_i \alpha_m d_i d_m \mathbf{x}_i^T \mathbf{x}_m - \sum_{i=1}^{N} \alpha_i$$

$$\text{Sujeto a:} \quad \sum_{i=1}^{N} d_i \alpha_i = 0$$

$$\alpha_i \geq 0$$

$\Longrightarrow$ Programación cuadrática $\Longrightarrow$ $\boldsymbol{\alpha}^*$

Matricialmente: $\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} + (-\mathbf{1})^T \boldsymbol{\alpha}$ sujeto a $\mathbf{d}^T \boldsymbol{\alpha} = 0; \quad \boldsymbol{\alpha} \geq 0$

$$Q = \begin{pmatrix} d_1 d_1 \mathbf{x}_1 \mathbf{x}_1 & d_1 d_2 \mathbf{x}_1 \mathbf{x}_2 & \dots & d_1 d_N \mathbf{x}_1 \mathbf{x}_N \\ d_2 d_1 \mathbf{x}_1 \mathbf{x}_1 & d_2 d_2 \mathbf{x}_1 \mathbf{x}_2 & \dots & d_2 d_N \mathbf{x}_1 \mathbf{x}_N \\ \dots & \dots & \dots & \dots \\ d_N d_1 \mathbf{x}_N \mathbf{x}_1 & d_N d_2 \mathbf{x}_N \mathbf{x}_2 & \dots & d_N d_N \mathbf{x}_N \mathbf{x}_N \end{pmatrix}$$

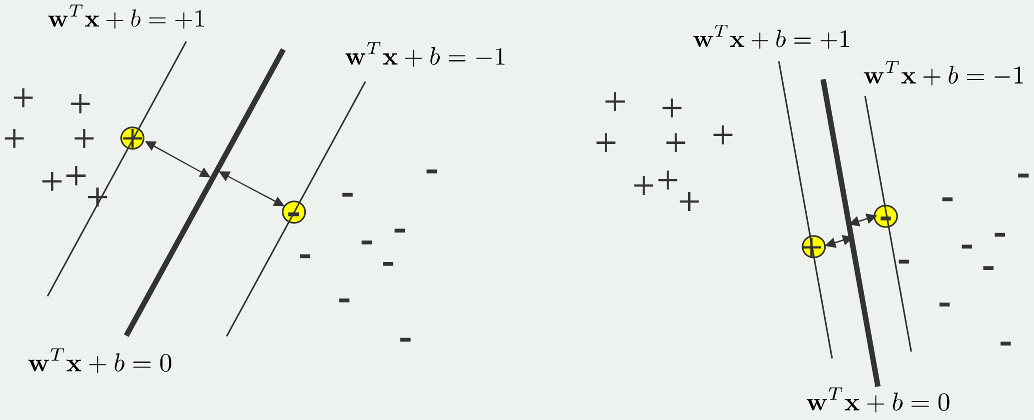
$\Longrightarrow$ La complejidad de la QP depende de N.

1.2. SVM lineales para datos no separables linealmente ("soft margin optimization")

"Soft Margin Optimization"

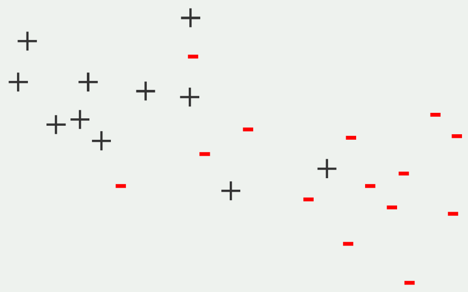
Dos problemas hasta el momento:

1. Sensibilidad a los datos (compromiso sesgo-varianza)



2. La solución no permite errores (‘**hard margin**’)

¿Qué ocurre si **no hay separabilidad lineal**?



'Soft margin'

Ambos problemas se solucionan permitiendo errores

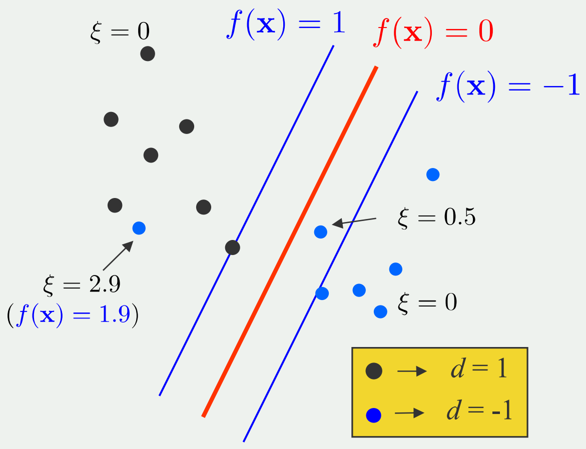
Ahora $d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$ no se puede satisfacer $\forall i$, apareciendo dos formas de violación de margen:

- Muestras bien clasificadas
- Muestras mal clasificadas

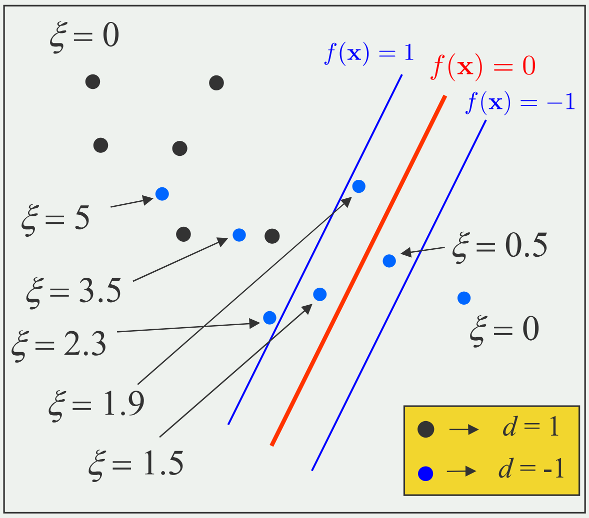
Se introducen las '**slack variables**'

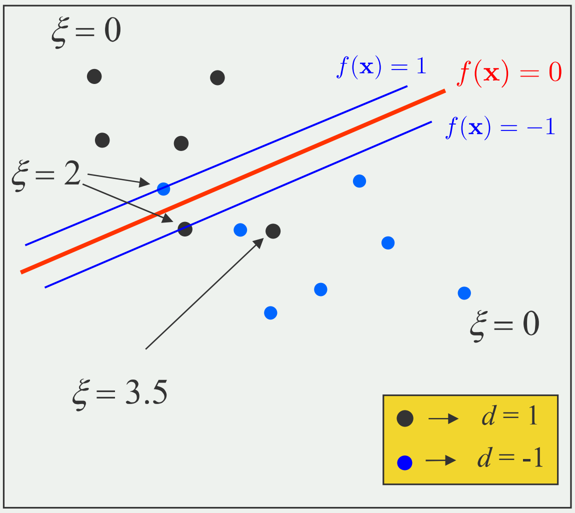
$$\xi_i \geq 0, \ i = 1, \ldots, N$$

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$$



$\xi = 0$	$\rightarrow$	$\mathbf{x}_i$ bien clasificada (fuera del margen)
$0 < \xi \leq 1$	$\rightarrow$	$\mathbf{x}_i$ bien clasificada (dentro del margen)
$1 < \xi \leq 2$	$\rightarrow$	$\mathbf{x}_i$ mal clasificada (dentro del margen)
$2 < \xi$	$\rightarrow$	$\mathbf{x}_i$ mal clasificada (fuera del margen)





Solución "mala": $\sum_{j=1}^{N=13} \xi_j = 14.7$

Solución "buena": $\sum_{j=1}^{N=13} \xi_j = 7.5$

Compromiso margen-error

- Seguimos queriendo **maximizar el margen** pues se maximiza la distancia de las muestras bien clasificadas a la frontera, es decir,

$$\min \|\mathbf{w}\|$$

$$\delta_i(\mathbf{w},\ b) = \frac{d_i(\mathbf{w}^T\mathbf{x}_i + b)}{\|\mathbf{w}\|} = d_i\left(\frac{\|\mathbf{w}\|\|\mathbf{x_i}\|\cos(\theta)}{\|\mathbf{w}\|} + \frac{b}{\|\mathbf{w}\|}\right) = d_i\left(\|\mathbf{x_i}\|\cos(\theta) + \frac{b}{\|\mathbf{w}\|}\right)$$

- Y, al mismo tiempo, buscamos **minimizar el error** de las muestras que violan el margen, es decir,

$$\min \sum_i \xi_i$$

Problema original (‘primal problem’)

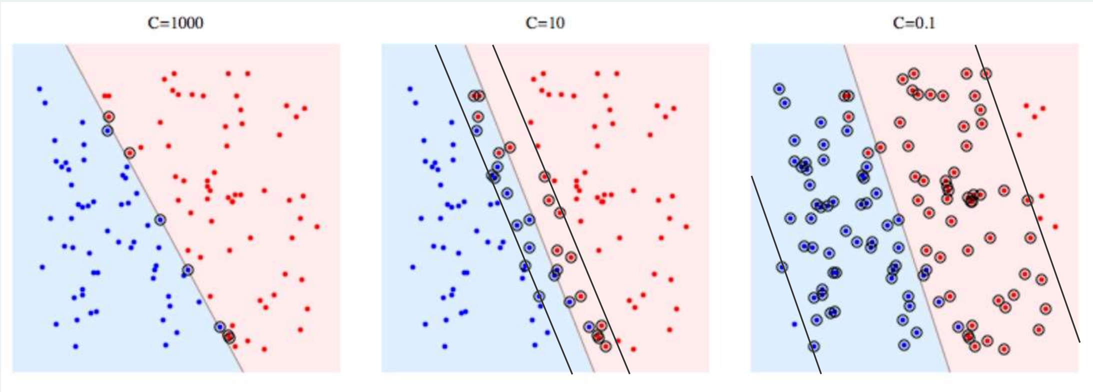
$$\min_{\mathbf{w},b,\xi} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i \right\}$$

$$\text{sujeto a:} \quad d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \quad i = 1, \ldots, N$$

$$\xi_i \geq 0$$

- C es un **parámetro de regularización** que controla el compromiso margen-error.
- C se especifica por el usuario normalmente mediante validación cruzada.
- Tamaño:
  - C pequeño: promueve un margen más amplio, permitiendo más clasificaciones erróneas. En SVM no lineales, origina modelos sencillos.
  - C grande: promueve la clasificación correcta de los datos, permitiendo márgenes pequeños. En SVM no lineales, origina modelos complejos que tiende al ‘**overfitting**’.

Ejemplo: problema separable resuelto con 'soft-margin' SVM

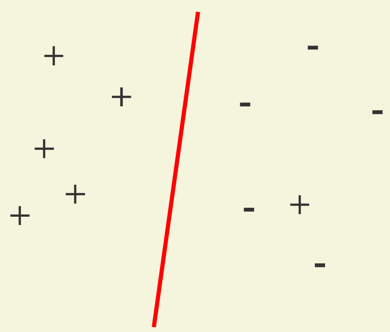


Margen muy pequeño
(casi nulo)
No permite errores

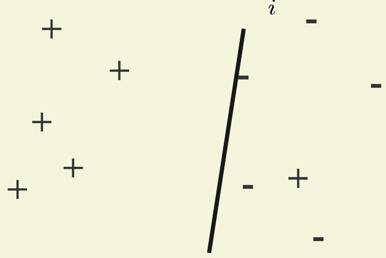
Margen muy grande.
Permite errores

SVM lineales

$$C \downdownarrows \;\Rightarrow\; \min \|\mathbf{w}\|$$

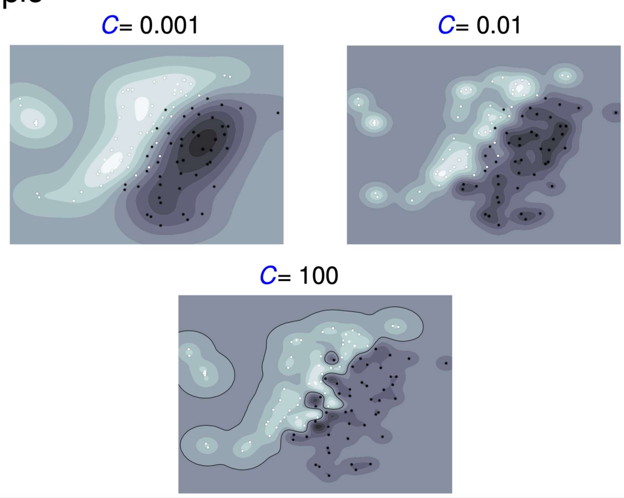


$$C \upuparrows \;\Rightarrow\; \min \sum_i \xi_i$$



SVM no lineales

Ejemplo



Principio de dualidad

Original

$\min_{\mathbf{x}} \phi(\mathbf{x})$

sujeto a: $g_i(\mathbf{x}) \geq 0, \; i = 1, \ldots, N$

con $\phi(\mathbf{x})$ función convexa en $\mathbf{x}$
$g_i(\mathbf{x})$ restriciones lineales en $\mathbf{x}$

Dual

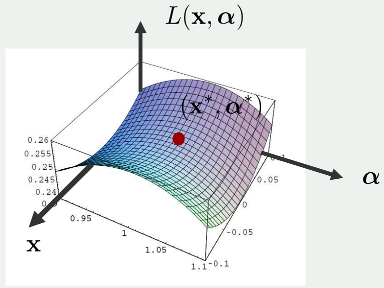
$\max_{\boldsymbol{\alpha}} \theta(\boldsymbol{\alpha})$

sujeto a: $\boldsymbol{\alpha} \geq 0$

con

$\theta(\boldsymbol{\alpha}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\alpha})$

$L(\mathbf{x}, \boldsymbol{\alpha}) = \phi(\mathbf{x}) - \sum_{i=1}^{N} \alpha_i g_i(\mathbf{x})$



Original

$$\min_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{N} \xi_i \right\}$$

sujeto a: $d_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad i = 1, \ldots, N$
$\xi_i \geq 0$

$\mathbf{x} \to (\mathbf{w}, b, \xi_i)$
$\alpha \to (\alpha_i, \beta_i)$

Dual

$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$

sujeto a: $\alpha_i \geq 0, \; \beta_i \geq 0$

$\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\boldsymbol{w}, b, \boldsymbol{\xi}} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (d_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{N} \beta_i \xi_i$$

Minimización de L :

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \;\to\; \mathbf{w} = \sum_{i=1}^{N} d_i \alpha_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b} = 0 \;\to\; \sum_{i=1}^{N} d_i \alpha_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = 0 \to \; C - \alpha_i - \beta_i = 0 \quad \longrightarrow$$

$$\left\{ \begin{array}{l} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i (d_i(\mathbf{w}^T \mathbf{x}_i + b) - 1) \quad \textbf{el mismo} \text{ que en el caso separable} \\ 0 \leq \alpha_i \leq C \quad (\text{ya que } \beta_i \geq 0 \text{ y } \alpha_i = C - \beta_i \Rightarrow \alpha_i \leq C) \end{array} \right.$$

Problema dual

$$\max_{\boldsymbol{\alpha}} \left\{ \theta(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{m=1}^{N} \alpha_i \alpha_m d_i d_m \mathbf{x}_i^T \mathbf{x}_m \right\}$$

Sujeto a: $\sum_{i=1}^{N} d_i \alpha_i = 0$

$0 \leq \alpha_i \leq C$ ← ¡Única diferencia!

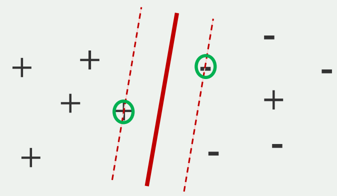
Programación cuadrática $\longrightarrow \alpha_i^* \longrightarrow$ **Solución:** $\mathbf{w}^* = \sum_{i \in SV} d_i \alpha_i^* \mathbf{x}_i$

$b^* = 1 - \mathbf{w}^* \mathbf{x}^{s+}$

Tipos de vectores soporte

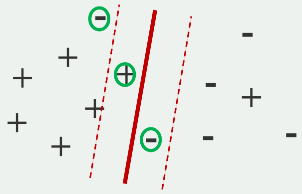
Margin support vectors $(0 < \alpha_i < C)$

$$\left.\begin{array}{l}\left.\begin{array}{l} 0 < \alpha_i < C \implies \beta_i > 0 \\ \beta_i \xi_i = 0 \quad \text{(condición KKT)} \end{array}\right\} \implies \xi_i = 0 \\ \alpha_i (d_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i) = 0 \quad \text{(condición KKT)} \end{array}\right\} \Rightarrow d_i(\mathbf{w}^T \mathbf{x}_i + b) = 1$$



Non-margin support vectors $(\alpha_i = C)$

$$\left.\begin{array}{l}\left.\begin{array}{l} \alpha_i = C \implies \beta_i = 0 \\ \beta_i \xi_i = 0 \quad \text{(condición KKT)} \end{array}\right\} \implies \xi_i \geq 0 \\ \alpha_i (d_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i) = 0 \quad \text{(condición KKT)} \end{array}\right\} \Rightarrow d_i(\mathbf{w}^T \mathbf{x}_i + b) = 1 - \xi_i$$

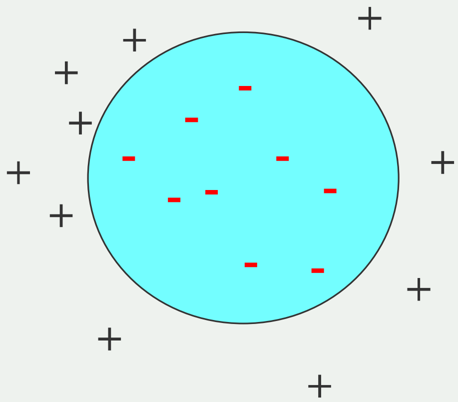


1.3. SVM no lineales (métodos ”kernel”)

‘Kernel methods’

Hemos analizado los discriminantes lineales de máximo margen para los casos separables y no separables.

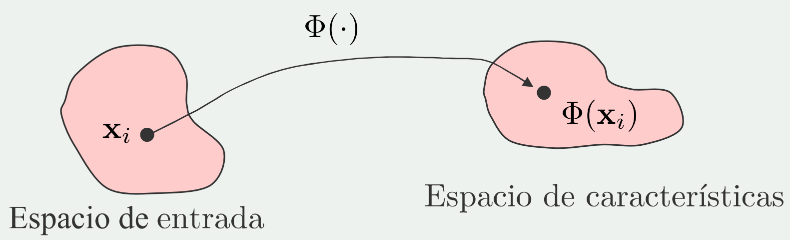
¿Cómo construir discriminantes no lineales?



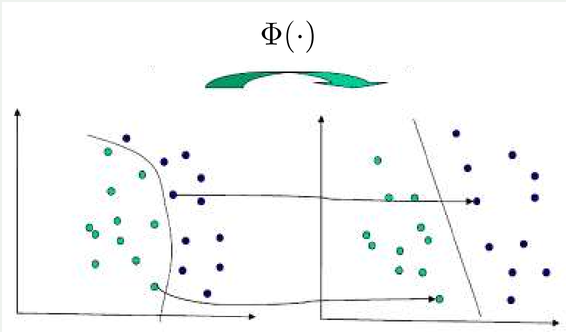
Teorema de Cover: *un espacio multidimensional puede transformarse en un espacio de características donde los patrones son linealmente separables si se verifican dos propiedades: a) la transformación es no lineal, y b) la dimensión del nuevo espacio es lo suficientemente alta.*

La **idea básica** consta de dos operaciones matemáticas:

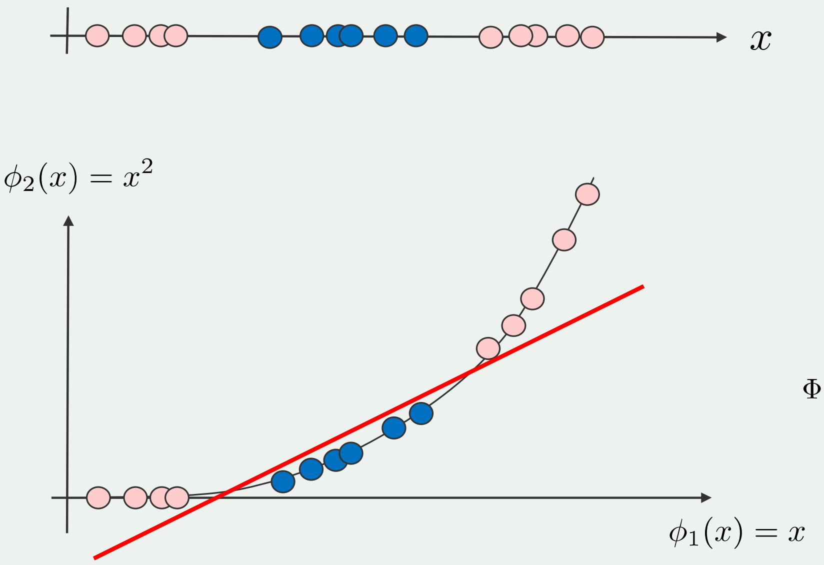
1. Un “mapping” no lineal del espacio de los datos L (m_0-dimensional) a un espacio de alta dimensión H (m_1-dimensional), llamado *espacio de características*
2. Construcción del hiperplano de máximo margen en el espacio de características



$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im_0} \end{pmatrix} \xrightarrow{\quad\Phi(\cdot)\quad} \Phi(\mathbf{x}_i) = \begin{pmatrix} \phi_1(\mathbf{x}_i) \\ \phi_2(\mathbf{x}_i) \\ \vdots \\ \phi_{m_1}(\mathbf{x}_i) \end{pmatrix}, \text{ con } 0 \leq m_1 \leq \infty$$



Ejemplo 1:

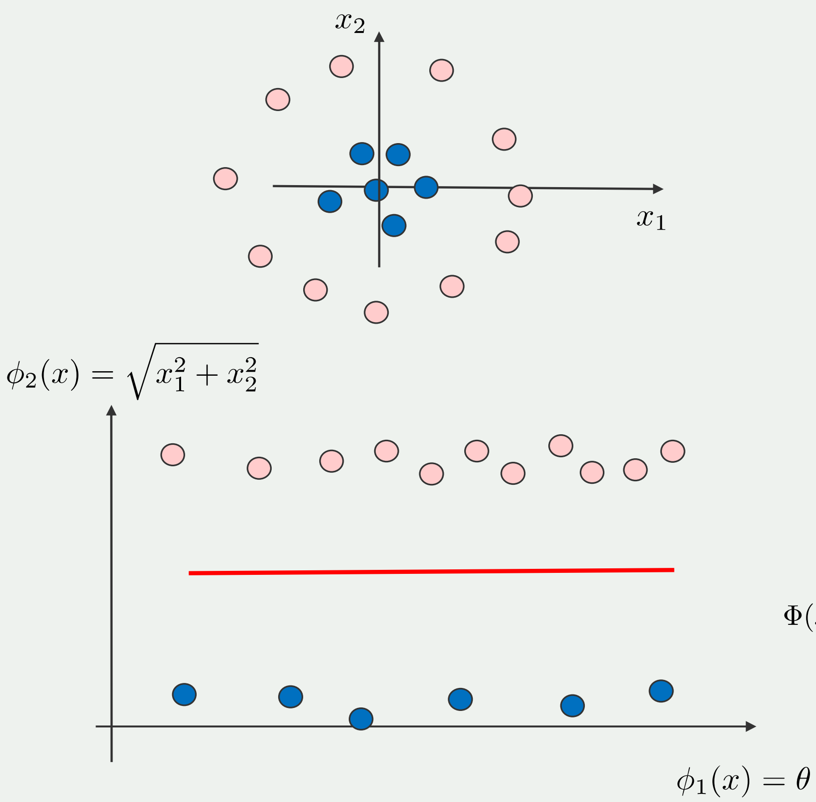


$$\mathbf{x} = x \in R$$

$$\Phi : \mathbf{x} \rightarrow \Phi(\mathbf{x})$$

$$\Phi(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix} = \begin{pmatrix} x \\ x^2 \end{pmatrix} \in R^2$$

Ejemplo 2: usando "Higher-order features"



$\mathbf{x} \in R^2$

$\Phi : \mathbf{x} \rightarrow \Phi(\mathbf{x})$

$$\Phi(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix} = \begin{pmatrix} \theta \\ \sqrt{x_1^2 + x_2^2} \end{pmatrix} \in R^2$$

Problema dual

$$\max_{\boldsymbol{\alpha}} \left\{ \theta(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{m=1}^{N} \alpha_i \alpha_m d_i d_m \Phi^T(\mathbf{x}_i)\Phi(\mathbf{x}_m) \right\}$$

Sujeto a: $\sum_{i=1}^{N} d_i \alpha_i = 0$

$0 \leq \alpha_i \leq C$

Programación cuadrática $\longrightarrow \alpha_i^* \longrightarrow$ $\mathbf{w}^* = \sum_{i \in SV} d_i \alpha_i^* \Phi(\mathbf{x}_i)$

$\mathbf{w}^*$ y $\Phi(\mathbf{x}_i)$ en el espacio de características

¿$\Phi(\mathbf{x})$?

Kernel Trick

a) Se introduce el kernel producto interno (*inner-product kernel*) definido como

$$\begin{aligned} K(\mathbf{x},\mathbf{x}_i) &= \Phi^T(\mathbf{x})\Phi(\mathbf{x}_i) \\ &= \sum_{j=0}^{m_1} \phi_j(\mathbf{x})\phi_j(\mathbf{x}_i) \\ &= K(\mathbf{x}_i,\mathbf{x}) \end{aligned}$$

b) Se establece la frontera de decisión en el espacio de características en términos del kernel:

$$f(\mathbf{w},b) = \mathbf{w}^T\Phi(\mathbf{x}) + b \quad \rightarrow \quad \left.\begin{aligned} &\mathbf{w}^T\Phi(\mathbf{x}) + b = 0 \\ &\mathbf{w}^* = \sum_{i\in SV} d_i\alpha_i^*\Phi(\mathbf{x}_i) \end{aligned}\right\} \longrightarrow \sum_{i\in SV} d_i\alpha_i^*\mathbf{\Phi}(\mathbf{x}_i)^T\mathbf{\Phi}(\mathbf{x}) + b^* = 0$$

$$\sum_{i\in SV} d_i\alpha_i^* K(\mathbf{x}_i,\mathbf{x}) + b^* = 0$$

Problema dual

$$\max_{\boldsymbol{\alpha}} \left\{ \theta(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{N}\alpha_i\alpha_m d_i d_m K(\mathbf{x}_i\mathbf{x}_m) \right\}$$

Sujeto a: $\sum_{i=1}^{N} d_i\alpha_i = 0$

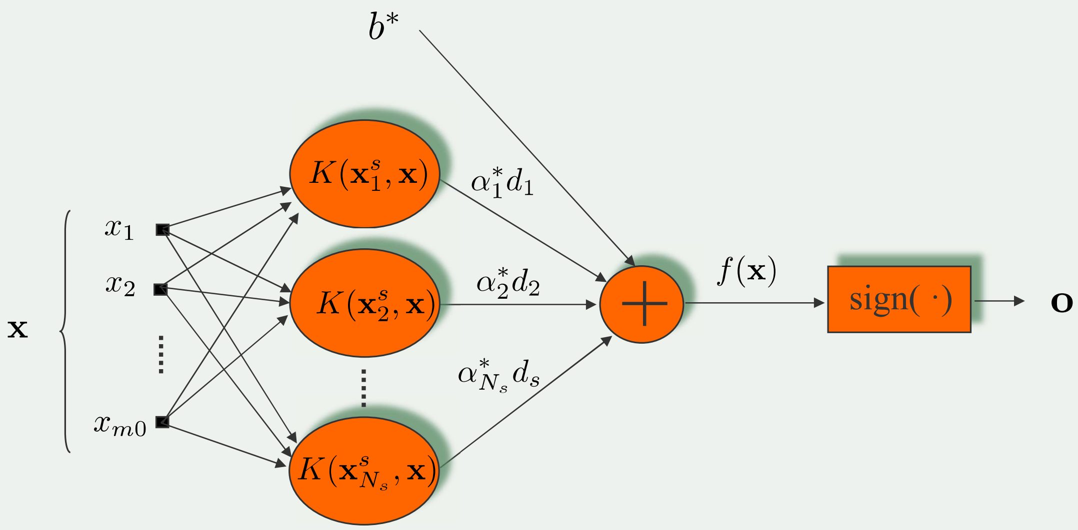
$0 \leq \alpha_i \leq C$

$$\downarrow$$

Programación cuadrática $\longrightarrow \alpha_i^* \longrightarrow$ $\sum_{i\in SV} d_i\alpha_i^* K(\mathbf{x}_i,\mathbf{x}) + b^* = 0$

¡Se construye el hiperplano óptimo en el espacio de caractarísticas sin emplear $\Phi(\mathbf{x})$!

Arquitectura de las SVM



SVM: Experimentación

- **Procesamiento de datos:** Los vectores describiendo los datos deben ser reales.
- Se suelen escalar los datos antes de aplicarlos.
  - Evitar que los atributos con rangos grandes dominen a los de rango más pequeño.
  - Se suele escalar al rango [-1,1] o [0,1].
- Seleccionar la función kernel.
- Determinar el parámetro C mediante validación.

Elección de los *kernels*

Los kernels deben cumplir el Teorema de Mercer

Teorema de Mercer

Sea $K(\mathbf{x}, \mathbf{x}')$ kernel simétrico definido en el intervalo cerrado $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$ y $\mathbf{a} \leq \mathbf{x}' \leq \mathbf{b}$. El kernel $K(\mathbf{x}, \mathbf{x}')$ se puede expandir en serie

$$K(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}')$$

con $\lambda_i > 0$. Para que esta expansión sea válida (con convergencia absoluta y uniforme), es condición necesaria y suficiente que

$$\int_b^a \int_b^a K(\mathbf{x}, \mathbf{x}') \Phi(\mathbf{x}) \Phi(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0$$

para toda función $\Phi(\cdot)$ que cumpla

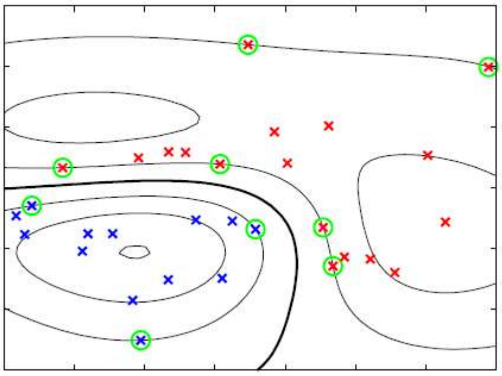
$$\int_b^a \Phi^2(\mathbf{x}) d\mathbf{x} > \infty$$

Ejemplos de kernels

Tipo de SVM	$K(\mathbf{x}, \mathbf{x}_i)$	Comentarios
Polinomial	$(\mathbf{x}\mathbf{x}_i + 1)^p$	p la especifica el usuario
RBF	$\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}_i\|^2\right)$	σ^2 la especifica el usuario
Two-Layer MLP	$\tanh\left(\beta_0 \mathbf{x}^T \mathbf{x}_i + \beta_1\right)$	Th. de Mercer se verifica sólo para algunos valores de β_0 y β_1

Ejemplo

Example of synthetic data from two classes in two dimensions showing contours of constant $y(\mathbf{x})$ obtained from a support vector machine having a Gaussian kernel function. Also shown are the decision boundary, the margin boundaries, and the support vectors.



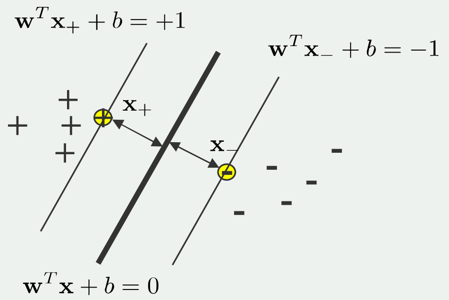
¡Frontera no lineal!

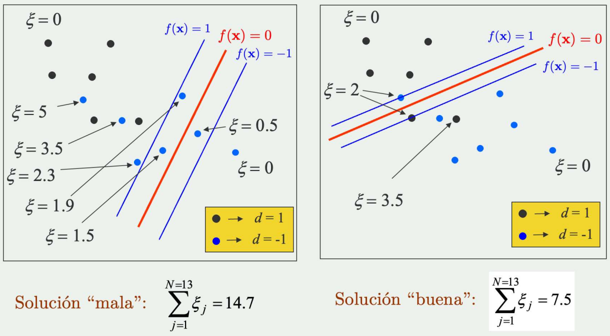
2.4. Máquinas de Vectores Soporte

2. SVM para regresión

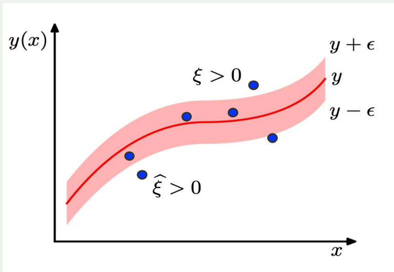
Support Vector Regression (SVR)

En clasificación, se busca que las muestras de ambas clases estén **FUERA** del margen



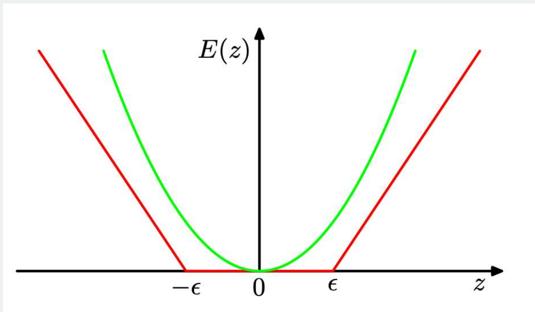


En regresión, se busca las muestras estén **DENTRO** del margen de una función de regresión.



Para obtener una solución dispersa, se define la función "ϵ-insensitive":

$$E(z) = \begin{cases} 0, & \text{si } |z| < \epsilon \\ |z| - \epsilon & \text{otro caso} \end{cases}$$



En nuestro caso: $$E_\epsilon(y(\mathbf{x}) - t) = \begin{cases} 0, & \text{si } |y(\mathbf{x}) - t| < \epsilon \\ |y(\mathbf{x}) - t| - \epsilon & \text{otro caso} \end{cases}$$

Minimizándose la función de error regularizada siguiente:

$$C \sum_{n=1}^{N} E_\epsilon(y(\mathbf{x}_n) - t_n) + \frac{1}{2}\|\mathbf{w}\|^2$$

donde C es el parámetro de regularización que aparece con el error por convenio, $y(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$, siendo $\phi(\mathbf{x})$ una transformación no lineal a un determinado espacio de características.

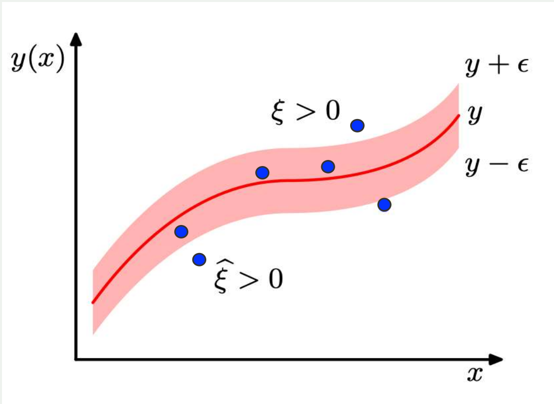
Al igual que antes, podemos re-expresar el problema de optimización, introduciendo las '**slack variables**'

$$\xi_n > 0, \text{ para un punto que verifica } t_n > y(\mathbf{x}_n) + \epsilon \text{ (por encima del tubo)}$$
$$\hat{\xi}_n > 0, \text{ para un punto que verifica } t_n < y(\mathbf{x}_n) - \epsilon \text{ (por debajo del tubo)}$$

Condiciones para los puntos fuera del tubo:

$$t_n \leq y(\mathbf{x}_n) + \epsilon + \xi_n$$
$$t_n \geq y(\mathbf{x}_n) - \epsilon - \hat{\xi}_n$$

con $\xi_n > 0$ y $\hat{\xi}_n > 0$.



Función de error para SVR:

$$\boxed{C \sum_{n=1}^{N} (\xi_n + \hat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2}$$

Problema original ('primal problem')

$$\min_{\mathbf{w},b,\xi,\hat{\xi}} \left\{ C \sum_{n=1}^{N} (\xi_n + \hat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2 \right\}$$

sujeto a: $\xi_n \geq 0, \quad n = 1, \ldots, N$

$\hat{\xi}_n \geq 0$

$t_n \leq y(\mathbf{x}_n) + \epsilon + \xi_n$

$t_n \geq y(\mathbf{x}_n) - \epsilon - \hat{\xi}_n$

Como ya se ha visto, aplicando el principio de dualidad y la teoría de Lagrange para optimización con restricciones, obtenemos el **problema dual.**

Principio de dualidad

Original

$$\min_{\mathbf{x}} \phi(\mathbf{x})$$

sujeto a: $g_i(\mathbf{x}) \geq 0, \ i = 1, \ldots, N$

con $\phi(\mathbf{x})$ función convexa en $\mathbf{x}$
$g_i(\mathbf{x})$ restricciones lineales en $\mathbf{x}$

Dual

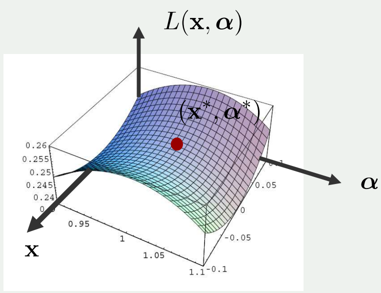
$$\max_{\boldsymbol{\alpha}} \theta(\boldsymbol{\alpha})$$

sujeto a: $\boldsymbol{\alpha} \geq 0$

con

$$\theta(\boldsymbol{\alpha}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\alpha})$$

$$L(\mathbf{x}, \boldsymbol{\alpha}) = \phi(\mathbf{x}) - \sum_{i=1}^{N} \alpha_i g_i(\mathbf{x})$$



Original

$$\min_{\mathbf{w},b} \ C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2$$

sujeto a: $\xi_n \geq 0, \quad n = 1, \ldots, N$

$$\hat{\xi}_n \geq 0$$
$$y(\mathbf{x}_n) + \epsilon + \xi_n - t_n \geq 0$$
$$-y(\mathbf{x}_n) + \epsilon + \hat{\xi}_n + t_n \geq 0$$

$\mathbf{x} \to (\mathbf{w}, b, \xi_n, \hat{\xi}_n)$
$\alpha \to (a_n, \hat{a}_n, \mu_n, \hat{\mu}_n))$

Dual

$$\max_{\boldsymbol{a},\hat{\boldsymbol{a}},\boldsymbol{\mu},\hat{\boldsymbol{\mu}}} \theta(\boldsymbol{a}, \hat{\boldsymbol{a}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$$

sujeto a: $a_n \geq 0, \ \hat{a}_n \geq 0 \ \mu_n \geq 0, \ \hat{\mu}_n \geq 0$

con

$$\theta(\boldsymbol{a}, \hat{\boldsymbol{a}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \min_{\boldsymbol{w},b,\boldsymbol{\xi},\hat{\boldsymbol{\xi}}} L(\boldsymbol{w}, b, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{a}, \hat{\boldsymbol{a}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$$

$$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{a}, \hat{\boldsymbol{a}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N}\mu_n\xi_n - \sum_{n=1}^{N}\hat{\mu}_n\hat{\xi}_n$$
$$- \sum_{n=1}^{N} a_n(y(\mathbf{x}_n) + \epsilon + \xi_n - t_n) - \sum_{n=1}^{N}\hat{a}_n(-y(\mathbf{x}_n) + \epsilon + \hat{\xi}_n + t_n)$$

Minimización de L :

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \ \rightarrow \ \mathbf{w} = \sum_{n=1}^{N}(a_n - \hat{a}_n)\phi(\mathbf{x}_n)$$

$$\frac{\partial L}{\partial b} = 0 \ \rightarrow \ \sum_{n=1}^{N}(a_n - \hat{a}_n) = 0$$

$$\frac{\partial L}{\partial \xi_n} = 0 \rightarrow \ a_n + \mu_n = C$$

$$\frac{\partial L}{\partial \hat{\xi}_n} = 0 \rightarrow \ \hat{a}_n + \hat{\mu}_n = C$$

$$L(\mathbf{a}, \ \hat{\mathbf{a}}) \ \rightarrow \ \theta(\mathbf{a}, \ \hat{\mathbf{a}}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m) - \epsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n$$

donde $k(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m)$

Problema dual

$$\max_{\mathbf{a}, \ \hat{\mathbf{a}}} \left\{ \theta(\mathbf{a}, \ \hat{\mathbf{a}}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m) - \epsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n \right\}$$

sujeto a: $0 \leq a_n \leq C$

$$0 \leq \hat{a}_n \leq C$$

$$\sum_{n=1}^{N}(a_n - \hat{a}_n) = 0$$

Programación cuadrática $\longrightarrow$ $a_n^*, \ \hat{a}_n^*$ $\longrightarrow$

$$y(\mathbf{x}) = \sum_{n=1}^{N}(a_n - \hat{a}_n)k(\mathbf{x}, \mathbf{x}_n) + b$$

Solución dispersa: solo los vectores soporte contribuyen a $y(\mathbf{x})$. Falta b.

Demo de las restricciones del problema dual:

Las restricciones son:

$a_n \geq 0, \quad n = 1, \ldots, N$
$\hat{a}_n \geq 0$
$\mu_n \geq 0,$
$\hat{\mu}_n \geq 0$
$\sum_{n=1}^{N}(a_n - \hat{a}_n) = 0$
$a_n + \mu_n = C$
$\hat{a}_n + \hat{\mu}_n = C$

$\longrightarrow$

$0 \leq a_n \leq C$
$0 \leq \hat{a}_n \leq C$
$\sum_{n=1}^{N}(a_n - \hat{a}_n) = 0$

Sabemos que la solución cumple las **Condiciones KKT**:

$$\frac{\partial L(\mathbf{x}^*, \boldsymbol{\alpha})}{\partial \mathbf{x}} = 0$$
$$g_i(\mathbf{x}^*) \geq 0, \; i = 1, \ldots, N$$
$$\alpha_i \geq 0, \; i = 1, \ldots, N$$
$$\alpha_i g_i(\mathbf{x}^*) = 0, \; i = 1, \ldots, N$$

La última condición significa que el producto de las variables duales y las restricciones se hace cero, es decir,

$$\begin{aligned} & && a_n(y(\mathbf{x}_n) + \epsilon + \xi_n - t_n) = 0 \\ & && \hat{a}_n(-y(\mathbf{x}_n) + \epsilon + \hat{\xi}_n + t_n) = 0 \\ \mu_n \xi_n = 0 & \quad \rightarrow && (C - a_n)\xi_n = 0 \\ \hat{\mu}_n \hat{\xi}_n = 0 & \quad \rightarrow && (C - \hat{a}_n)\hat{\xi}_n = 0 \end{aligned}$$

$$\begin{aligned} a_n(y(\mathbf{x}_n) + \epsilon + \xi_n - t_n) = 0 \\ \hat{a}_n(-y(\mathbf{x}_n) + \epsilon + \hat{\xi}_n + t_n) = 0 \end{aligned} \quad \Longrightarrow$$

- a_n puede ser no nula si $y(\mathbf{x}_n) + \epsilon + \xi_n - t_n = 0$, es decir, $\mathbf{x}_n$ es un punto que está en el borde superior del tubo ($\xi_n = 0$) o por encima ($\xi_n > 0$).
- $\hat{a}_n$ puede ser no nula si $-y(\mathbf{x}_n) + \epsilon + \xi_n + t_n = 0$, es decir, $\mathbf{x}_n$ es un punto en el borde inferior del tubo ($\hat{\xi}_n = 0$) o por encima ($\hat{\xi}_n > 0$).

Además, las restricciones $y(\mathbf{x}_n) + \epsilon + \xi_n - t_n = 0$ y $-y(\mathbf{x}_n) + \epsilon + \xi_n + t_n = 0$ son incompatibles, no se pueden satisfacer a la vez. Por tanto, para todo $\mathbf{x}_n$, o bien $a_n = 0$ o bien $\hat{a}_n = 0$ (o ambas).

Los vectores soporte son los puntos que contribuyen a la solución $y(\mathbf{x}) = \sum_{n=1}^{N} (a_n - \hat{a}_n) k(\mathbf{x}, \mathbf{x}_n) + b$. Por tanto, para ellos se cumple que $a_n \neq 0$ ó $\hat{a}_n \neq 0$, es decir son puntos en el borde o fuera del tubo. Los puntos dentro del tubo satisfacen $a_n = \hat{a}_n = 0$. De nuevo, solución dispersa.