



# IIC 2433 Minería de Datos

<https://github.com/marcelomendoza/IIC2433>

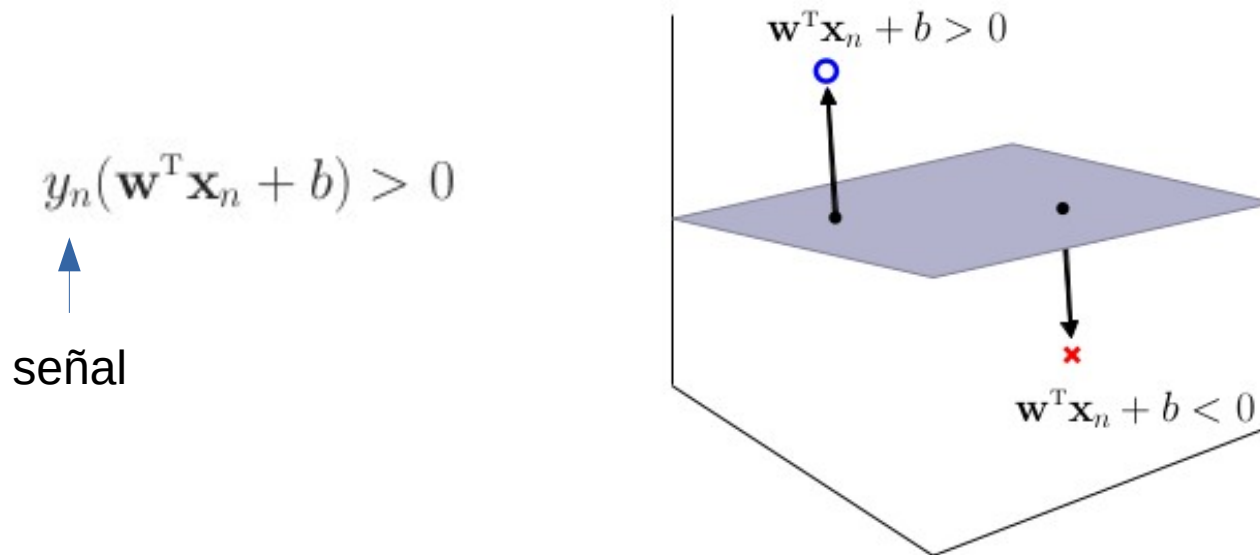
# - SUPPORT VECTOR MACHINES -

## Hiperplano separador

Queremos calcular el hiperplano más grueso que separa los datos. Es decir, queremos identificar el hiperplano separador de máximo margen.

El máximo margen es la distancia desde el hiperplano al dato más cercano.

Supongamos que tenemos un conjunto de datos que queremos separar del resto (son de la misma clase). El hiperplano separa estos ssi:



Notar que la magnitud de la señal no es relevante para la decisión, dado que podemos **reescalar** los pesos y el bias.

## Hiperplano separador

Encontramos **el dato más cercano**, y calculamos:

$$\rho = \min_{n=1,\dots,N} y_n(\mathbf{w}^T \mathbf{x}_n + b)$$

Y reescalamos con respecto a  $\rho$ :

$$\min_{n=1,\dots,N} y_n \left( \frac{\mathbf{w}^T}{\rho} \mathbf{x}_n + \frac{b}{\rho} \right) = \frac{1}{\rho} \min_{n=1,\dots,N} y_n(\mathbf{w}^T \mathbf{x}_n + b) = \frac{\rho}{\rho} = 1.$$

Es decir, es posible encontrar pesos tal que  $y_n(\mathbf{w}^T \mathbf{x}_n + b)$  es mayor o igual que 1 y al menos existe un dato que satisface la igualdad.

## Hiperplano separador

Encontramos **el dato más cercano**, y calculamos:

$$\rho = \min_{n=1,\dots,N} y_n(\mathbf{w}^T \mathbf{x}_n + b)$$

Y reescalamos con respecto a  $\rho$ :

$$\min_{n=1,\dots,N} y_n \left( \frac{\mathbf{w}^T}{\rho} \mathbf{x}_n + \frac{b}{\rho} \right) = \frac{1}{\rho} \min_{n=1,\dots,N} y_n(\mathbf{w}^T \mathbf{x}_n + b) = \frac{\rho}{\rho} = 1.$$

Es decir, es posible encontrar pesos tal que  $y_n(\mathbf{w}^T \mathbf{x}_n + b)$  es mayor o igual que 1 y al menos existe un dato que satisface la igualdad.

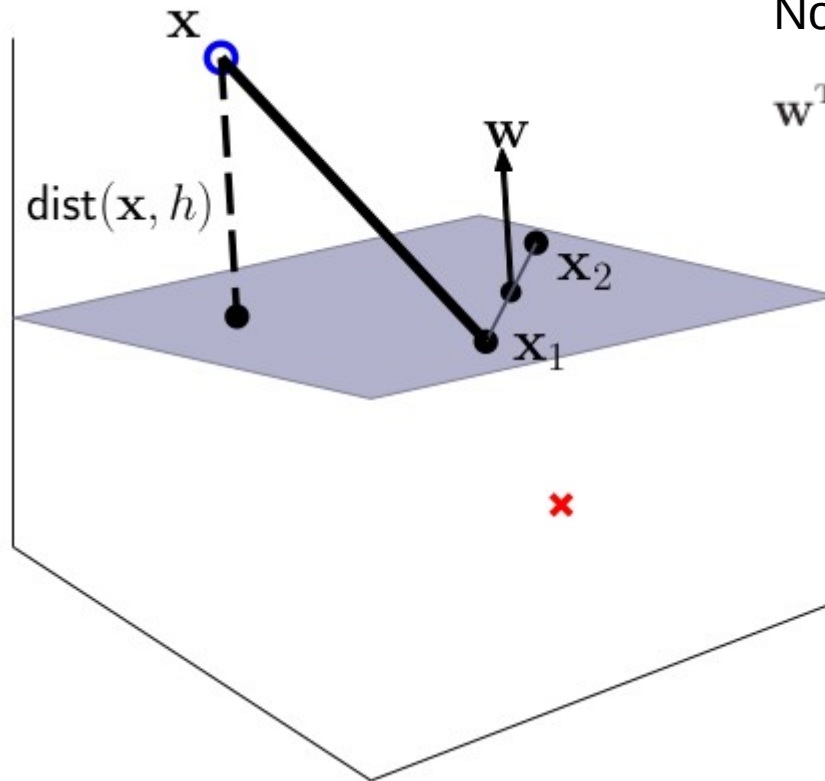
Definición. Un hiperplano separador (grueso) separa los datos si:

$$\min_{n=1,\dots,N} y_n(\mathbf{w}^T \mathbf{x}_n + b) = 1.$$

Básicamente, este es un esquema de normalización de pesos.

## Margen del hiperplano separador

Margen: distancia desde el hiperplano al dato más cercano.



Notar que el vector de pesos cumple:

$$\mathbf{w}^T(\mathbf{x}_2 - \mathbf{x}_1) = \mathbf{w}^T\mathbf{x}_2 - \mathbf{w}^T\mathbf{x}_1 = -b + b = 0.$$

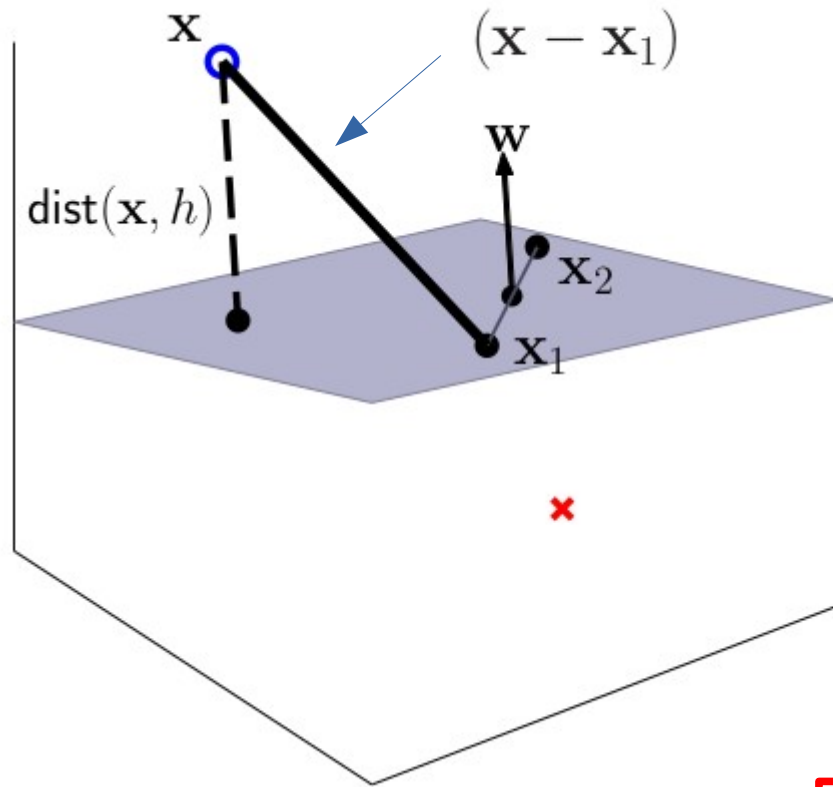
Esto porque para cualquier punto que este sobre el hiperplano se da que:

$$\mathbf{w}^T\mathbf{x}' + b = 0.$$

Es decir,  $\mathbf{w}$  es ortogonal al hiperplano.

## Margen del hiperplano separador

Usando un vector ortonormal  $\mathbf{u} = \mathbf{w} / \|\mathbf{w}\|$ , obtenemos:



Proyección de la hipotenusa a la dirección del vector ortonormal

$$\begin{aligned} \text{dist}(\mathbf{x}, h) &= |\mathbf{u}^T (\mathbf{x} - \mathbf{x}_1)| \\ &= \frac{1}{\|\mathbf{w}\|} \cdot |\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mathbf{x}_1| \\ &= \frac{1}{\|\mathbf{w}\|} \cdot |\mathbf{w}^T \mathbf{x} + b| \end{aligned}$$

$$\mathbf{w}^T \mathbf{x}' + b = 0.$$

Entonces:

$$\text{dist}(\mathbf{x}, h) = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|}$$

## Margen del hiperplano separador

$$\text{dist}(\mathbf{x}, h) = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|}$$

Como el hiperplano separa según el signo, es decir,  $y_n = \pm 1$  se cumple que:

$$|\mathbf{w}^T \mathbf{x}_n + b| = |y_n(\mathbf{w}^T \mathbf{x}_n + b)| = y_n(\mathbf{w}^T \mathbf{x}_n + b), \quad \begin{array}{l} \text{Datos} \\ \text{separados} \end{array}$$

$\nearrow \mathbf{x}_1, \dots, \mathbf{x}_N$



## Margen del hiperplano separador

$$\text{dist}(\mathbf{x}, h) = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|}$$

Como el hiperplano separa según el signo, es decir,  $y_n = \pm 1$  se cumple que:

$$|\mathbf{w}^T \mathbf{x}_n + b| = |y_n(\mathbf{w}^T \mathbf{x}_n + b)| = y_n(\mathbf{w}^T \mathbf{x}_n + b), \quad \begin{array}{l} \text{Datos} \\ \text{separados} \end{array}$$

$\nearrow \mathbf{x}_1, \dots, \mathbf{x}_N$

La distancia la podemos reescribir según:

$$\text{dist}(\mathbf{x}_n, h) = \frac{y_n(\mathbf{w}^T \mathbf{x}_n + b)}{\|\mathbf{w}\|}.$$

## Margen del hiperplano separador

$$\text{dist}(\mathbf{x}, h) = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|}$$

Como el hiperplano separa según el signo, es decir,  $y_n = \pm 1$  se cumple que:

$$|\mathbf{w}^T \mathbf{x}_n + b| = |y_n(\mathbf{w}^T \mathbf{x}_n + b)| = y_n(\mathbf{w}^T \mathbf{x}_n + b), \quad \begin{array}{l} \text{Datos} \\ \text{separados} \end{array}$$

$\mathbf{x}_1, \dots, \mathbf{x}_N$

La distancia la podemos reescribir según:

$$\text{dist}(\mathbf{x}_n, h) = \frac{y_n(\mathbf{w}^T \mathbf{x}_n + b)}{\|\mathbf{w}\|}.$$

El dato más cercano tiene distancia (recordar que  $\min_{n=1, \dots, N} y_n(\mathbf{w}^T \mathbf{x}_n + b) = 1$ ):

$$\min_n \text{dist}(\mathbf{x}_n, h) = \frac{1}{\|\mathbf{w}\|} \cdot \min_n y_n(\mathbf{w}^T \mathbf{x}_n + b)$$

## Margen del hiperplano separador

$$\text{dist}(\mathbf{x}, h) = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|}$$

Como el hiperplano separa según el signo, es decir,  $y_n = \pm 1$  se cumple que:

$$|\mathbf{w}^T \mathbf{x}_n + b| = |y_n(\mathbf{w}^T \mathbf{x}_n + b)| = y_n(\mathbf{w}^T \mathbf{x}_n + b), \quad \begin{array}{l} \text{Datos} \\ \text{separados} \end{array}$$

$\mathbf{x}_1, \dots, \mathbf{x}_N$

La distancia la podemos reescribir según:

$$\text{dist}(\mathbf{x}_n, h) = \frac{y_n(\mathbf{w}^T \mathbf{x}_n + b)}{\|\mathbf{w}\|}.$$

El dato más cercano tiene distancia (recordar que  $\min_{n=1, \dots, N} y_n(\mathbf{w}^T \mathbf{x}_n + b) = 1$ ):

$$\min_n \text{dist}(\mathbf{x}_n, h) = \frac{1}{\|\mathbf{w}\|} \cdot \min_n y_n(\mathbf{w}^T \mathbf{x}_n + b)$$
$$\min_n \text{dist}(\mathbf{x}_n, h) = \frac{1}{\|\mathbf{w}\|} \quad \leftarrow \text{Máximo margen}$$

## Margen del hiperplano separador

Objetivo:  $\min_n \text{dist}(\mathbf{x}_n, h) = \frac{1}{\|\mathbf{w}\|}$

Encontrar los pesos que maximizan la distancia al dato más cercano

¡Minimizar denominador!

## Margen del hiperplano separador

Objetivo:  $\min_n \text{dist}(\mathbf{x}_n, h) = \frac{1}{\|\mathbf{w}\|}$

Encontrar los pesos que maximizan la distancia al dato más cercano

¡Minimizar denominador!

$$\underset{b, \mathbf{w}}{\text{minimize}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

$$\text{subject to: } \min_{n=1, \dots, N} y_n (\mathbf{w}^T \mathbf{x}_n + b) = 1.$$

## Hiperplano separador de máximo margen

$$\underset{b, \mathbf{w}}{\text{minimize}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

$$\text{subject to: } y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \text{ for } n = 1, \dots, N.$$

## Hiperplano separador de máximo margen

$$\underset{b, \mathbf{w}}{\text{minimize}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

$$\text{subject to: } y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \text{ for } n = 1, \dots, N.$$

Ejemplo:

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$$

$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix} \quad \begin{array}{ll} -b \geq 1 & (i) \\ -(2w_1 + 2w_2 + b) \geq 1 & (ii) \\ 2w_1 + b \geq 1 & (iii) \\ 3w_1 + b \geq 1 & (iv) \end{array}$$

## Hiperplano separador de máximo margen

$$\underset{b, \mathbf{w}}{\text{minimize}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

$$\text{subject to: } y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \text{ for } n = 1, \dots, N.$$

Ejemplo:

$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix} \quad \begin{array}{ll} -b \geq 1 & (i) \\ -(2w_1 + 2w_2 + b) \geq 1 & (ii) \\ 2w_1 + b \geq 1 & (iii) \\ 3w_1 + b \geq 1 & (iv) \end{array}$$

Se resuelve usando (i), (ii) y (iii):

$$\left. \begin{array}{l} \text{- (i) y (iii) dan: } w_1 \geq 1 \\ \text{- (ii) y (iii) dan: } w_2 \leq -1 \end{array} \right\} \text{Tomamos el borde: } (b = -1, w_1 = 1, w_2 = -1)$$



## Hiperplano separador de máximo margen

$$\underset{b, \mathbf{w}}{\text{minimize}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

$$\text{subject to: } y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \text{ for } n = 1, \dots, N.$$

Ejemplo:

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$$

$$X = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix} \quad \begin{array}{ll} -b \geq 1 & (i) \\ -(2w_1 + 2w_2 + b) \geq 1 & (ii) \\ 2w_1 + b \geq 1 & (iii) \\ 3w_1 + b \geq 1 & (iv) \end{array}$$

Se resuelve usando (i), (ii) y (iii):

$$\left. \begin{array}{l} \text{- (i) y (ii) dan: } w_1 \geq 1 \\ \text{- (ii) y (iii) dan: } w_2 \leq -1 \end{array} \right\} \text{Tomamos el borde: } (b = -1, w_1 = 1, w_2 = -1)$$

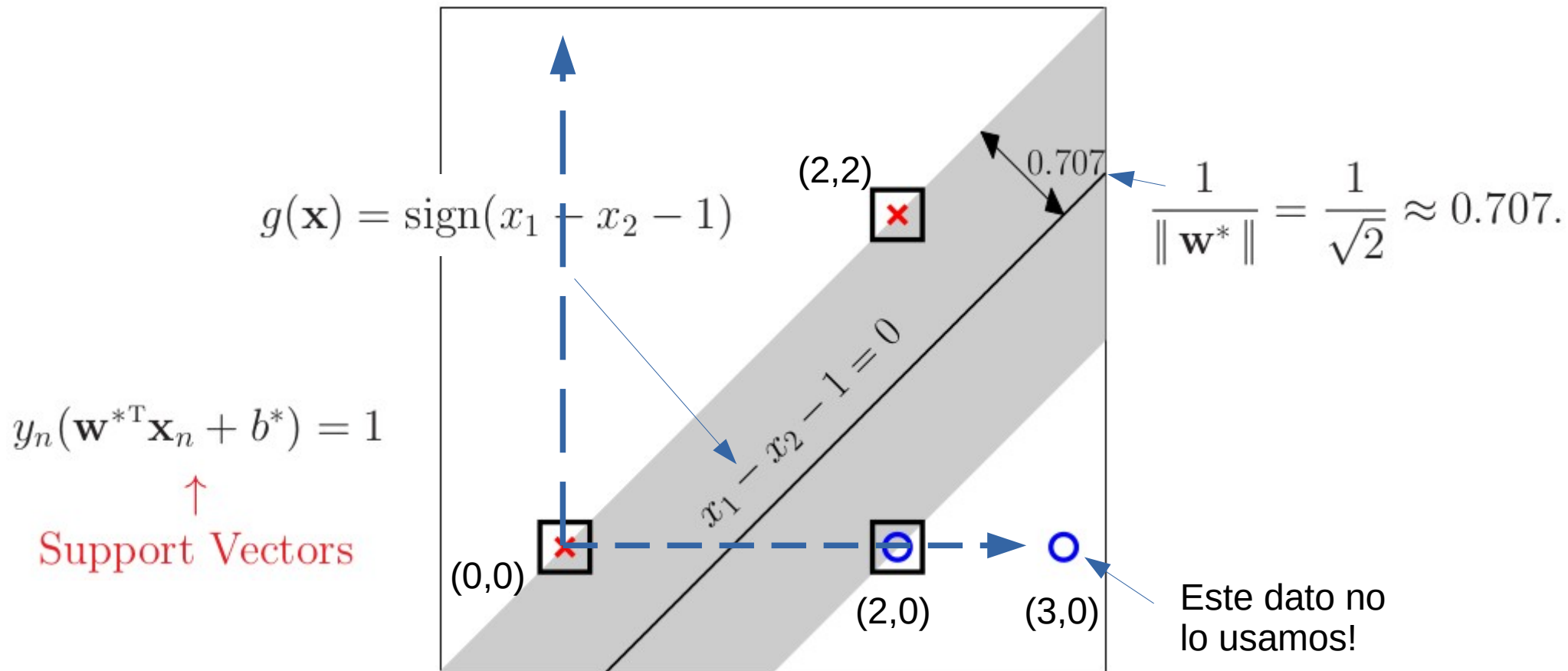
$$\rightarrow g(\mathbf{x}) = \text{sign}(x_1 - x_2 - 1)$$

$$\text{y} \quad \frac{1}{\|\mathbf{w}^*\|} = \frac{1}{\sqrt{2}} \approx 0.707.$$

# Hiperplano separador de máximo margen $y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$

$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

$$\begin{aligned} -b &\geq 1 & (i) \\ -(2w_1 + 2w_2 + b) &\geq 1 & (ii) \\ 2w_1 + b &\geq 1 & (iii) \\ 3w_1 + b &\geq 1 & (iv) \end{aligned}$$



## Support Vector Machines (forma matricial)

$$\underset{b, \mathbf{w}}{\text{minimize}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

$$\text{subject to: } y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \text{ for } n = 1, \dots, N.$$

Expresión  
matricial


$$\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \in \mathbb{R}^{d+1}$$

$$\frac{1}{2} \mathbf{w}^T \mathbf{w} = \begin{bmatrix} b & \mathbf{w}^T \end{bmatrix} \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix} \begin{bmatrix} b \\ \mathbf{w}^T \end{bmatrix} = \mathbf{u}^T \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix} \mathbf{u} \implies \mathbf{Q} = \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix}, \mathbf{p} = \mathbf{0}_{d+1}$$

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \equiv \begin{bmatrix} y_n & y_n \mathbf{x}_n^T \end{bmatrix} \mathbf{u} \geq 1 \implies \begin{bmatrix} y_1 & y_1 \mathbf{x}_1^T \\ \vdots & \vdots \\ y_N & y_N \mathbf{x}_N^T \end{bmatrix} \mathbf{u} \geq \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \implies \mathbf{A} = \begin{bmatrix} y_1 & y_1 \mathbf{x}_1^T \\ \vdots & \vdots \\ y_N & y_N \mathbf{x}_N^T \end{bmatrix}, \mathbf{c} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Se reduce a un problema de optimización cuadrática para encontrar  $\mathbf{Q}$  y  $\mathbf{P}$ .

# Support Vector Machines

Forma estándar de problema QP:

$$\begin{array}{ll}\text{minimize:} & \frac{1}{2} \mathbf{u}^T \mathbf{Q} \mathbf{u} + \mathbf{p}^T \mathbf{u} \\ \text{subject to:} & \mathbf{A} \mathbf{u} \geq \mathbf{c}.\end{array}$$

donde:

$$\mathbf{Q} = \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix}, \quad \mathbf{p} = \mathbf{0}_{d+1} \qquad \mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \in \mathbb{R}^{d+1}$$

$$\mathbf{A} = \begin{bmatrix} y_1 & y_1 \mathbf{x}_1^T \\ \vdots & \vdots \\ y_N & y_N \mathbf{x}_N^T \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

# - SUPPORT VECTOR MACHINES - (KERNEL TRICK)

## SVM en datos no separables (linealmente)

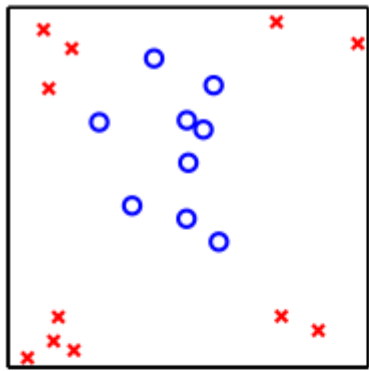
Consideremos una transformación  $\Phi: \mathbb{R}^d \rightarrow \mathbb{R}^d$  tal que  $\mathbf{z}_n = \Phi(\mathbf{x}_n)$ . Después de transformar los datos, el problema SVM (hard margin) es:

$$\begin{array}{ll} \underset{\tilde{b}, \tilde{\mathbf{w}}}{\text{minimize:}} & \frac{1}{2} \tilde{\mathbf{w}}^T \tilde{\mathbf{w}} \\ \text{subject to:} & y_n \left( \tilde{\mathbf{w}}^T \mathbf{z}_n + \tilde{b} \right) \geq 1 \quad (n = 1, \dots, N), \end{array}$$

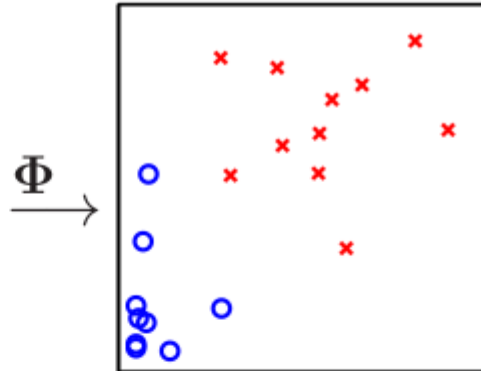
donde  $\tilde{\mathbf{w}}$  reside en el espacio de representación  $Z$ . Notemos que la dimensionalidad de  $Z$  puede ser distinta a la de la entrada.

Si la transformación es no lineal, puede ayudar a la SVM a separar datos no separables en el espacio original.

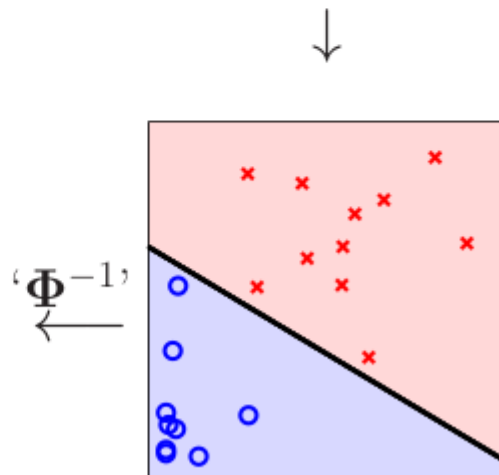
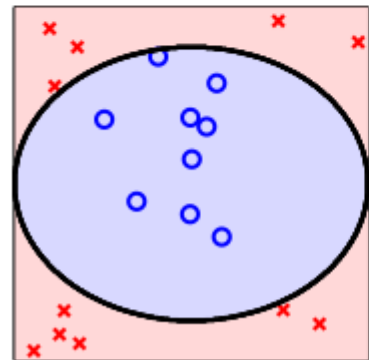
# SVM en datos no separables (linealmente)



1.  $\mathbf{x}_n \in \mathcal{X}$



2.  $\mathbf{z}_n = \Phi(\mathbf{x}_n) \in \mathcal{Z}$



$\Phi^{-1}$

4.  $g(\mathbf{x}) = \tilde{g}(\Phi(\mathbf{x})) = \text{sign}(\tilde{\mathbf{w}}^T \Phi(\mathbf{x}))$     3.  $\tilde{g}(\mathbf{z}) = \text{sign}(\tilde{\mathbf{w}}^T \mathbf{z})$

$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$$

$$y_1, y_2, \dots, y_N$$

$$\mathbf{z} = \Phi(\mathbf{x}) = \begin{bmatrix} 1 \\ \Phi_1(\mathbf{x}) \\ \vdots \\ \Phi_{\tilde{d}}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} 1 \\ z_1 \\ \vdots \\ z_{\tilde{d}} \end{bmatrix}$$

$$\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N$$

$$y_1, y_2, \dots, y_N$$

$$g(\mathbf{x}) = \text{sign}(\tilde{\mathbf{w}}^T \Phi(\mathbf{x}))$$

$$\tilde{\mathbf{w}} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{\tilde{d}} \end{bmatrix}$$

Este problema es difícil de resolver cuando  $\tilde{d}$  es muy grande.

## SVM en datos no separables (linealmente)

En la práctica, en lugar de trabajar sobre el problema original en el espacio  $Z$ , se aborda la formulación dual ya que es más fácil desde el punto de vista de optimización.

Teorema (KKT). El problema QP convexo en su forma primal:

$$\begin{array}{ll} \underset{\mathbf{u} \in \mathbb{R}^L}{\text{minimize:}} & \frac{1}{2} \mathbf{u}^T \mathbf{Q} \mathbf{u} + \mathbf{p}^T \mathbf{u} \\ \text{subject to:} & \mathbf{a}_m^T \mathbf{u} \geq c_m \end{array}$$

define la función dual (Lagrange):

$$\mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}) = \frac{1}{2} \mathbf{u}^T \mathbf{Q} \mathbf{u} + \mathbf{p}^T \mathbf{u} + \sum_{m=1}^M \alpha_m (c_m - \mathbf{a}_m^T \mathbf{u}).$$

La solución del primal es óptima ssi es óptima en el dual. Luego:


$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}).$$



## SVM en datos no separables (linealmente)

Apliquemos la formulación dual a la SVM:

$$\max_{\alpha \geq 0} \min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \alpha).$$


  $\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \in \mathbb{R}^{d+1}$

$$\begin{aligned} \mathcal{L}(b, \mathbf{w}, \alpha) &= \frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{n=1}^N \alpha_n (1 - y_n (\mathbf{w}^T \mathbf{x}_n + b)) \\ &= \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{n=1}^N \alpha_n y_n \mathbf{w}^T \mathbf{x}_n - b \sum_{n=1}^N \alpha_n y_n + \sum_{n=1}^N \alpha_n \end{aligned}$$

## SVM en datos no separables (linealmente)

Apliquemos la formulación dual a la SVM:

$$\max_{\alpha \geq 0} \min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \alpha).$$

  $\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \in \mathbb{R}^{d+1}$

$$\begin{aligned} \mathcal{L}(b, \mathbf{w}, \alpha) &= \frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{n=1}^N \alpha_n (1 - y_n (\mathbf{w}^T \mathbf{x}_n + b)) \\ &= \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{n=1}^N \alpha_n y_n \mathbf{w}^T \mathbf{x}_n - b \sum_{n=1}^N \alpha_n y_n + \sum_{n=1}^N \alpha_n \end{aligned}$$

Minimizamos w.r.t.  $(b, \mathbf{w})$

$$\frac{\partial \mathcal{L}}{\partial b} = 0:$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{n=1}^N \alpha_n y_n \quad \Rightarrow \quad \sum_{n=1}^N \alpha_n y_n = 0$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0:$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{n=1}^N \alpha_n y_n \mathbf{x}_n \quad \Rightarrow \quad \mathbf{w} = \sum_{n=1}^N \alpha_n y_n \mathbf{x}_n$$

## SVM en datos no separables (linealmente)

Usaremos  $\mathbf{w} = \sum_{n=1}^N \alpha_n y_n \mathbf{x}_n$  en  $\frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{n=1}^N \alpha_n y_n \mathbf{w}^T \mathbf{x}_n$

$$\begin{aligned}
 \mathcal{L} &= \frac{1}{2} \mathbf{w}^T \mathbf{w} - \mathbf{w}^T \sum_{n=1}^N \alpha_n y_n \mathbf{x}_n - b \sum_{n=1}^N \alpha_n y_n + \sum_{n=1}^N \alpha_n \\
 &= -\frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{n=1}^N \alpha_n \\
 &= -\frac{1}{2} \sum_{m,n=1}^N \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m + \sum_{n=1}^N \alpha_n
 \end{aligned}$$

## SVM en datos no separables (linealmente)

Usaremos  $\mathbf{w} = \sum_{n=1}^N \alpha_n y_n \mathbf{x}_n$  en  $\frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{n=1}^N \alpha_n y_n \mathbf{w}^T \mathbf{x}_n + \sum_{n=1}^N \alpha_n$ .

$$\begin{aligned} \mathcal{L} &= \frac{1}{2} \mathbf{w}^T \mathbf{w} - \mathbf{w}^T \sum_{n=1}^N \alpha_n y_n \mathbf{x}_n - b \sum_{n=1}^N \alpha_n y_n + \sum_{n=1}^N \alpha_n \\ &= -\frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{n=1}^N \alpha_n \\ &= -\frac{1}{2} \sum_{m,n=1}^N \alpha_n \alpha_m y_n y_m \mathbf{x}_n^T \mathbf{x}_m + \sum_{n=1}^N \alpha_n \end{aligned}$$

Ojo que:

Es decir, debemos hacer lo siguiente:

$$\begin{aligned} &\text{minimize}_{\alpha \in \mathbb{R}^N} : \quad \frac{1}{2} \sum_{m=1}^N \sum_{n=1}^N y_n y_m \alpha_n \alpha_m \mathbf{x}_n^T \mathbf{x}_m - \sum_{n=1}^N \alpha_n \\ &\text{subject to:} \quad \sum_{n=1}^N y_n \alpha_n = 0 \\ &\quad \quad \quad \alpha_n \geq 0 \quad (n = 1, \dots, N). \end{aligned}$$


$$\max_{\alpha \geq 0} \min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \alpha).$$

← es lo mismo que

## SVM en datos no separables (linealmente)

El dual se puede reescribir en una forma más amable:

$$\begin{aligned} & \underset{\alpha \in \mathbb{R}^N}{\text{minimize}} : && \frac{1}{2} \sum_{m=1}^N \sum_{n=1}^N y_n y_m \alpha_n \alpha_m \mathbf{x}_n^T \mathbf{x}_m - \sum_{n=1}^N \alpha_n \\ & \text{subject to:} && \sum_{n=1}^N y_n \alpha_n = 0 \\ & && \alpha_n \geq 0 \quad (n = 1, \dots, N). \end{aligned}$$


$$\begin{aligned} & \underset{\alpha}{\text{minimize}} && \frac{1}{2} \alpha^T G \alpha - \mathbf{1}^T \alpha \\ & \text{subject to:} && \mathbf{y}^T \alpha = 0 \\ & && \alpha \geq 0 \end{aligned}$$

donde  $(G_{nm} = y_n y_m \mathbf{x}_n^T \mathbf{x}_m)$

## SVM en datos no separables (linealmente)

Podemos recuperar la SVM del primal desde el dual:

$$\mathbf{w} = \sum_{n=1}^N \alpha_n^* y_n \mathbf{x}_n$$

$$\begin{aligned} \alpha_s > 0 &\implies y_s(\mathbf{w}^T \mathbf{x}_s + b) - 1 = 0 \\ &\implies b = y_s - \mathbf{w}^T \mathbf{x}_s \end{aligned}$$

## SVM en datos no separables (linealmente)

Podemos recuperar la SVM del primal desde el dual:

$$\mathbf{w} = \sum_{n=1}^N \alpha_n^* y_n \mathbf{x}_n$$

$$\begin{aligned} \alpha_s > 0 &\implies y_s(\mathbf{w}^T \mathbf{x}_s + b) - 1 = 0 \\ &\implies b = y_s - \mathbf{w}^T \mathbf{x}_s \end{aligned}$$

$$(G_{nm} = y_n y_m \mathbf{x}_n^T \mathbf{x}_m)$$

Una forma simple de calcular G:

$$X = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix} \quad \longrightarrow \quad X_s = \begin{bmatrix} 0 & 0 \\ -2 & -2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \longrightarrow \quad G = X_s X_s^T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 8 & -4 & -6 \\ 0 & -4 & 4 & 6 \\ 0 & -6 & 6 & 9 \end{bmatrix}$$

signed data matrix



# SVM en datos no separables (linealmente)

$$\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \in \mathbb{R}^{d+1}$$

QP problem

primal desde el dual

Dual SVM

$$\begin{array}{ll} \underset{\mathbf{u}}{\text{minimize}} & \frac{1}{2} \mathbf{u}^T \mathbf{Q} \mathbf{u} + \mathbf{p}^T \mathbf{z} \\ \text{subject to:} & \mathbf{A} \mathbf{u} \geq \mathbf{c} \end{array}$$

$$\begin{array}{ll} \underset{\boldsymbol{\alpha}}{\text{minimize}} & \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha} \\ \text{subject to:} & \mathbf{y}^T \boldsymbol{\alpha} = 0 \\ & \boldsymbol{\alpha} \geq 0 \end{array}$$

Resuelvo para u

$$\left. \begin{array}{l} \mathbf{u} = \boldsymbol{\alpha} \\ \mathbf{Q} = \mathbf{G} \\ \mathbf{p} = -\mathbf{1}_N \\ \mathbf{A} = \begin{bmatrix} \mathbf{y}^T \\ -\mathbf{y}^T \\ \mathbf{I}_N \end{bmatrix} \\ \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{0}_N \end{bmatrix} \end{array} \right\} \xrightarrow{\text{QP}(\mathbf{Q}, \mathbf{p}, \mathbf{A}, \mathbf{c})} \begin{array}{l} \boldsymbol{\alpha}^* = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 1 \\ 0 \end{bmatrix} \\ \mathbf{w} = \sum_{n=1}^4 \alpha_n^* y_n \mathbf{x}_n = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ b = y_1 - \mathbf{w}^T \mathbf{x}_1 = -1 \\ \gamma = \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{2}} \end{array}$$



## SVM en datos no separables (linealmente)

$$\mathbf{w} = \sum_{n=1}^N \alpha_n^* y_n \mathbf{x}_n$$

$$\begin{aligned} \alpha_s > 0 &\implies y_s(\mathbf{w}^T \mathbf{x}_s + b) - 1 = 0 \\ &\implies b = y_s - \mathbf{w}^T \mathbf{x}_s \end{aligned}$$

QP

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbf{w} = \sum_{n=1}^4 \alpha_n^* y_n \mathbf{x}_n = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$b = y_1 - \mathbf{w}^T \mathbf{x}_1 = -1$$

$$\gamma = \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{2}}$$

## SVM en datos no separables (linealmente)

$$\mathbf{w} = \sum_{n=1}^N \alpha_n^* y_n \mathbf{x}_n$$

$$\begin{aligned} \alpha_s > 0 &\implies y_s(\mathbf{w}^T \mathbf{x}_s + b) - 1 = 0 \\ &\implies b = y_s - \mathbf{w}^T \mathbf{x}_s \end{aligned}$$

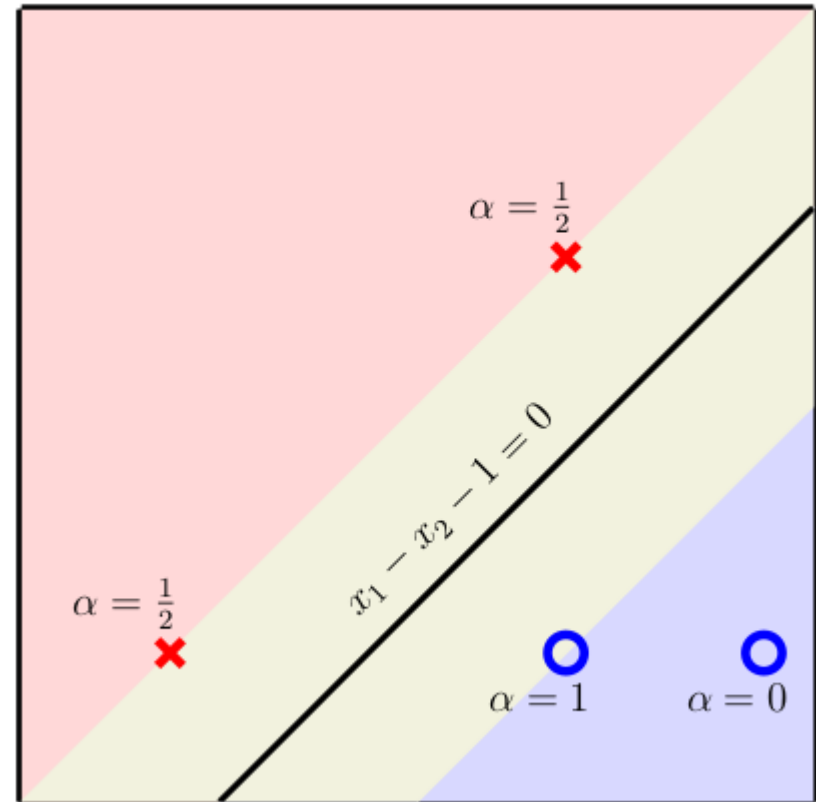
QP

$$\alpha^* = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbf{w} = \sum_{n=1}^4 \alpha_n^* y_n \mathbf{x}_n = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$b = y_1 - \mathbf{w}^T \mathbf{x}_1 = -1$$

$$\gamma = \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{2}}$$



Los que no son vectores de soporte tienen:  $\alpha_n = 0$

## SVM en datos no separables (linealmente)


Resolver el problema en el dual nos permite trabajar en el espacio  $Z$ .

$$\underset{\alpha}{\text{minimize}} \quad \frac{1}{2} \alpha^T G \alpha - \mathbf{1}^T \alpha$$

$$\text{subject to: } \mathbf{y}^T \alpha = 0$$

$$C \geq \alpha \geq 0$$

$$G_{nm} = y_n y_m (\mathbf{z}_n^T \mathbf{z}_m)$$

$$g(\mathbf{x}) = \text{sign} \left( \sum_{\alpha_n^* > 0} \alpha_n^* y_n (\mathbf{z}_n^T \mathbf{z}) + b^* \right)$$


producto interno

$$C > \alpha_s^* > 0$$

$$b^* = y_s - \sum_{\alpha_n^* > 0} \alpha_n^* y_n (\mathbf{z}_n^T \mathbf{z}_s)$$

## SVM en datos no separables (linealmente)

Un kernel es una función que combina tanto la transformación como el producto interno:

$$K_{\Phi}(\mathbf{x}, \mathbf{x}') \equiv \Phi(\mathbf{x})^T \Phi(\mathbf{x}').$$

El kernel toma dos vectores y entrega el producto interno en  $Z$ .

## SVM en datos no separables (linealmente)

Un kernel es una función que combina tanto la transformación como el producto interno:

$$K_{\Phi}(\mathbf{x}, \mathbf{x}') \equiv \Phi(\mathbf{x})^T \Phi(\mathbf{x}').$$

El kernel toma dos vectores y entrega el producto interno en  $Z$ .

Ejemplo: Kernel polinomial de segundo orden.

$$\Phi_2(\mathbf{x})^T \Phi_2(\mathbf{x}') = 1 + (\mathbf{x}^T \mathbf{x}') + (\mathbf{x}^T \mathbf{x}')^2.$$

## SVM en datos no separables (linealmente)

1: **Input:**  $X, y$ .

2: Compute  $G$ :  $G_{nm} = y_n y_m K(\mathbf{x}_n, \mathbf{x}_m)$ .

3: Solve (QP):

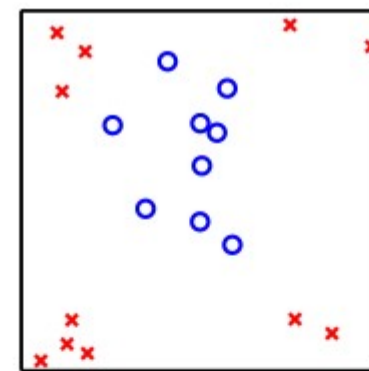
$$\left. \begin{array}{l} \underset{\boldsymbol{\alpha}}{\text{minimize:}} \quad \frac{1}{2} \boldsymbol{\alpha}^T G \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha} \\ \text{subject to:} \quad \mathbf{y}^T \boldsymbol{\alpha} = 0 \\ \quad \quad \quad \boldsymbol{\alpha} \geq \mathbf{0} \end{array} \right\} \rightarrow \boldsymbol{\alpha}^*$$

index  $s$  :  $\alpha_s^* > 0$

4:  $b^* = y_s - \sum_{\alpha_n^* > 0} \alpha_n^* y_n K(\mathbf{x}_n, \mathbf{x}_s)$

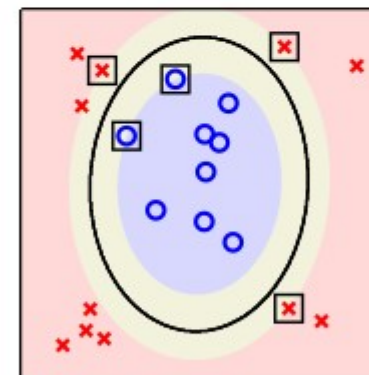
5: The final hypothesis is

$$g(\mathbf{x}) = \text{sign} \left( \sum_{\alpha_n^* > 0} \alpha_n^* y_n K(\mathbf{x}_n, \mathbf{x}) + b^* \right)$$



$\mathbf{x}_n \in \mathcal{X}$

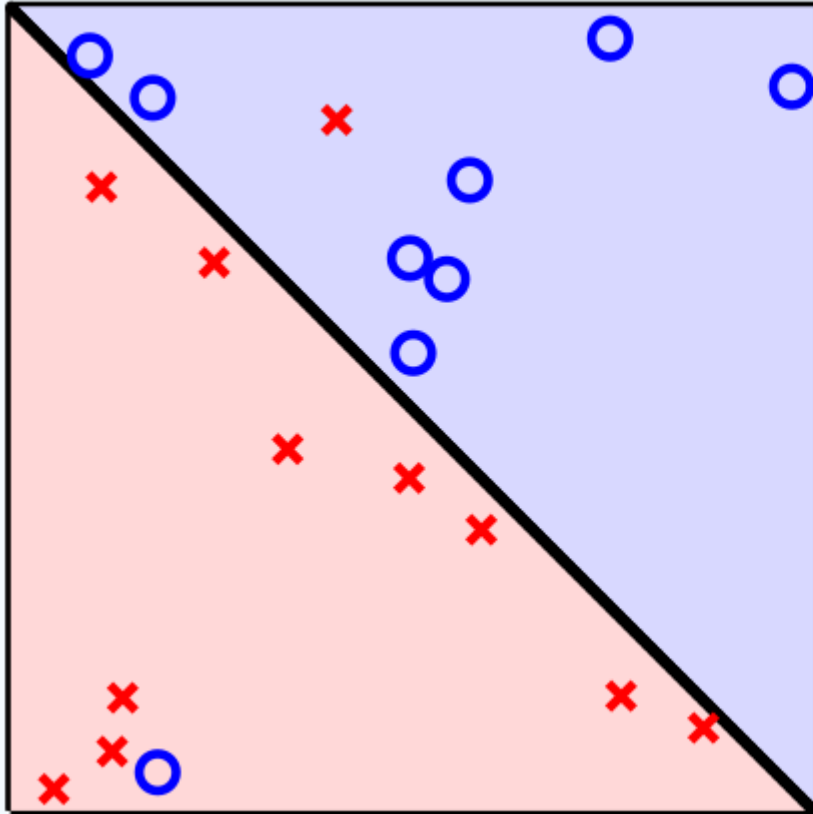
$\downarrow K(\cdot, \cdot)$



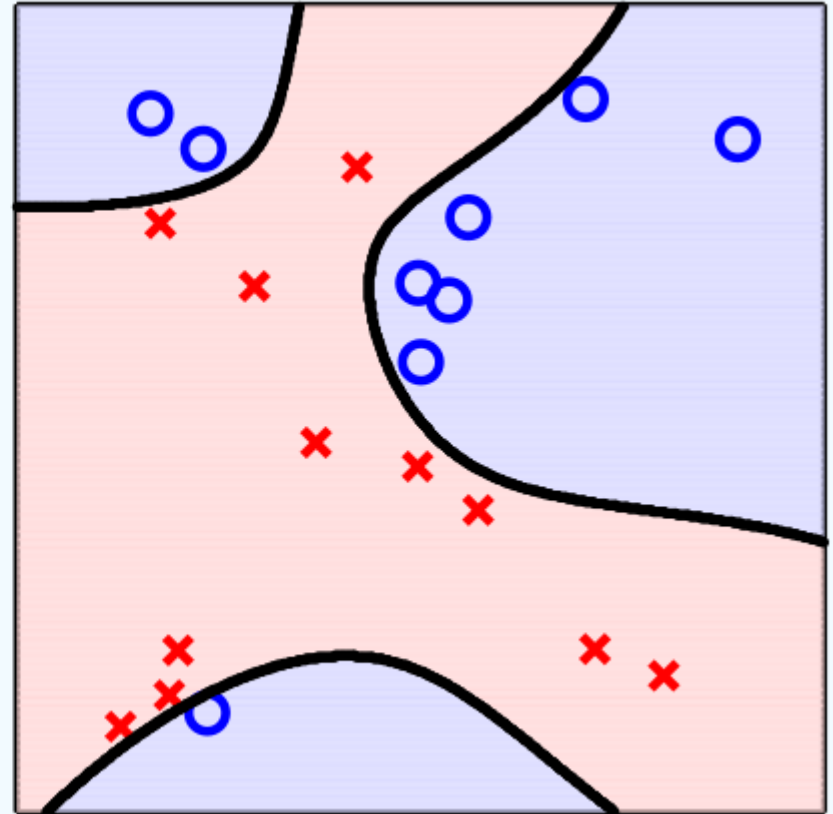
$$g(\mathbf{x}) = \text{sign} \left( \sum_{\alpha_n^* > 0} \alpha_n^* y_n K(\mathbf{x}_n, \mathbf{x}) + b^* \right)$$

$$b^* = y_s - \sum_{\alpha_n^* > 0} \alpha_n^* y_n K(\mathbf{x}_n, \mathbf{x}_s)$$

## SVM con kernel Gaussiano



a) SVM lineal



b) SVM con kernel Gaussiano