IIC 2433 Minería de Datos

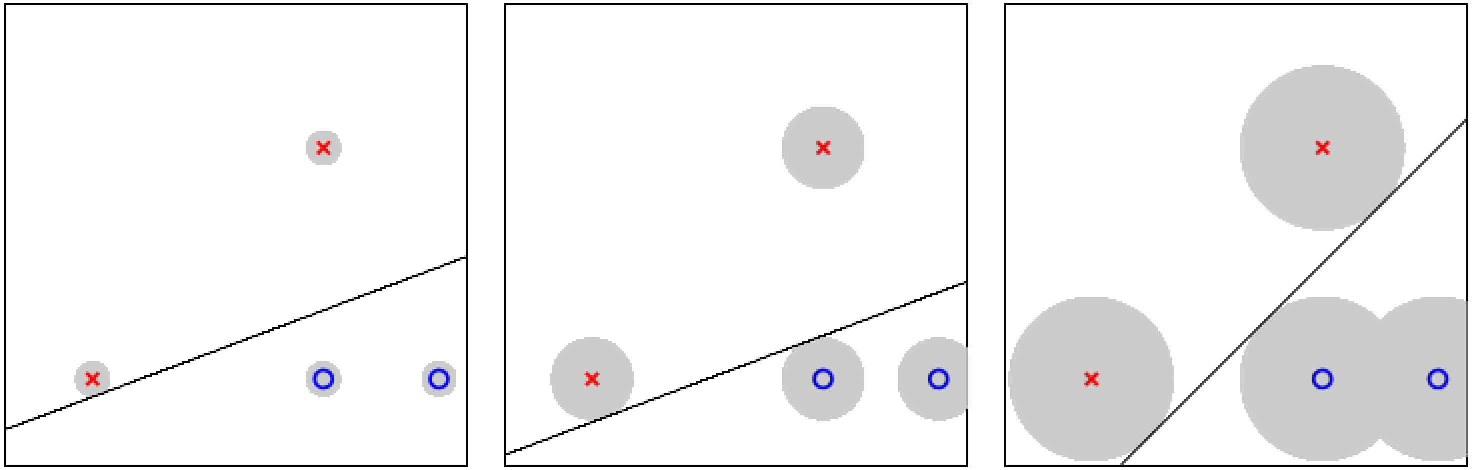
https://github.com/marcelomendoza/IIC2433

- SUPPORT VECTOR MACHINES -

Separadores y datos ruidosos

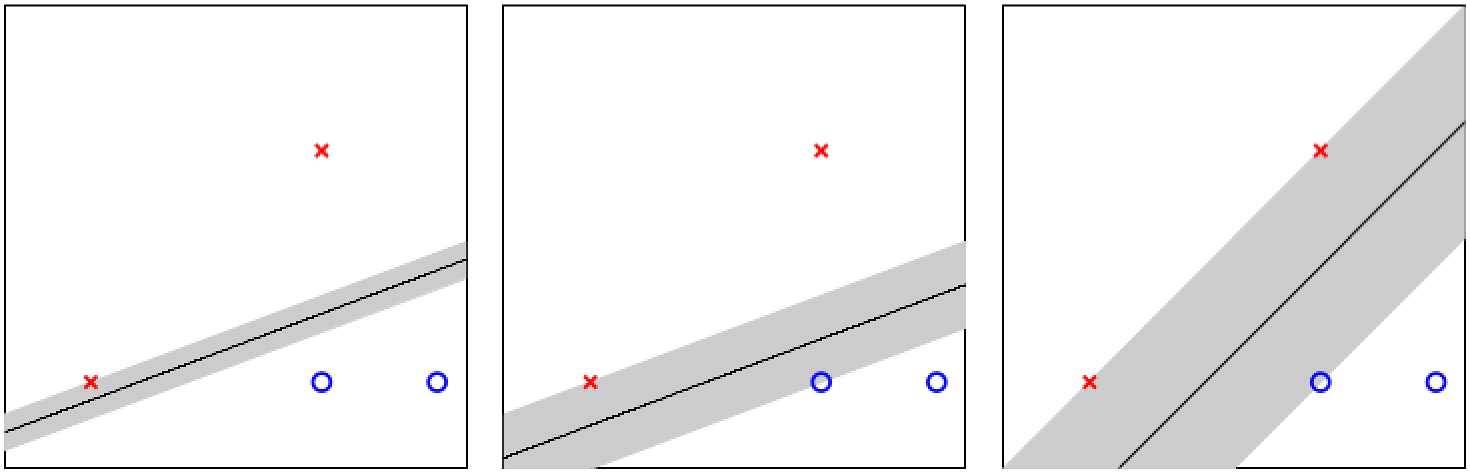
- Los datos pueden ser ruidosos (ruido de medición).
- Nuestros modelos debieran ser robustos a datos ruidosos.

Idea: La robustez al ruido tiene relación con considerar un margen de error para las mediciones.



Separadores y datos ruidosos

Una idea análoga a márgenes para datos consiste en trabajar con **hiperplanos gruesos**, agregando un margen al separador.



Separadores y datos ruidosos

Para trabajar con un hiperlano grueso, podemos usar el sesgo de una manera ingeniosa.

Hiperplano estrecho

$$\mathbf{x} \in \{1\} \times \mathbb{R}^d; \;\; \mathbf{w} \in \mathbb{R}^{d+1}$$

$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}; \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}.$$

$$\text{signal} = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

El sesgo se codifica como una dimensión más

Hiperplano grueso

$$\mathbf{x} \in \mathbb{R}^d; \;\; b \in \mathbb{R}, \;\; \mathbf{w} \in \mathbb{R}^d$$

$$\text{bias } b$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}; \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}.$$

$$\text{signal} = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$$

El sesgo aditivo interviene en el espacio de representación

Separadores y datos ruidosos

Hiperplano grueso

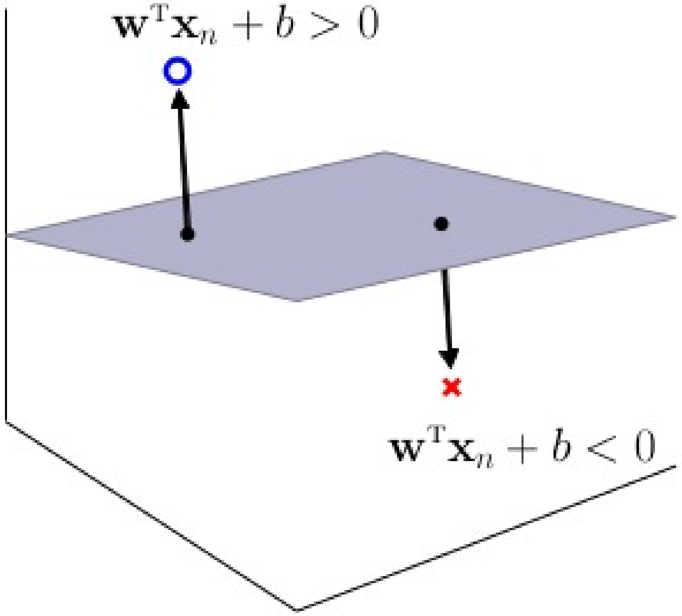
$\mathbf{x} \in \mathbb{R}^d;\ b \in \mathbb{R},\ \mathbf{w} \in \mathbb{R}^d$

bias b

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}; \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}.$$

$$\text{signal} = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$$

sesgo



Hiperplano separador

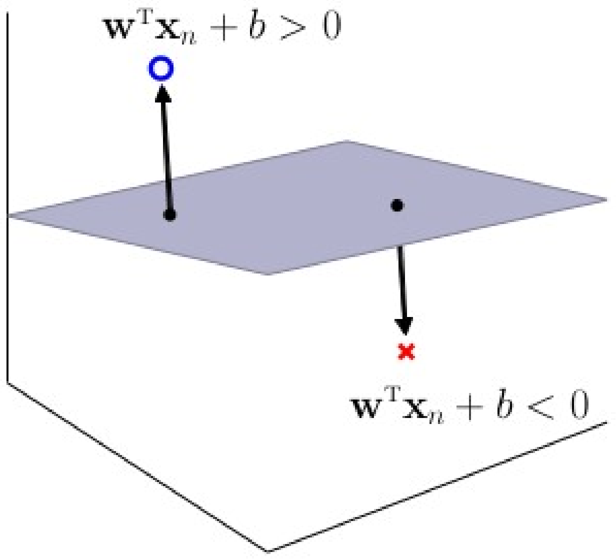
Queremos calcular el hiperplano más grueso que separa los datos. Es decir, queremos identificar el <u>hiperplano separador de máximo margen</u>.

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:

$$y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b) > 0$$

señal



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:

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

Y reescalamos con respecto a ρ:

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

Es decir, es posible encontrar pesos tal que $y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)$ 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:

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

Y reescalamos con respecto a ρ:

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

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

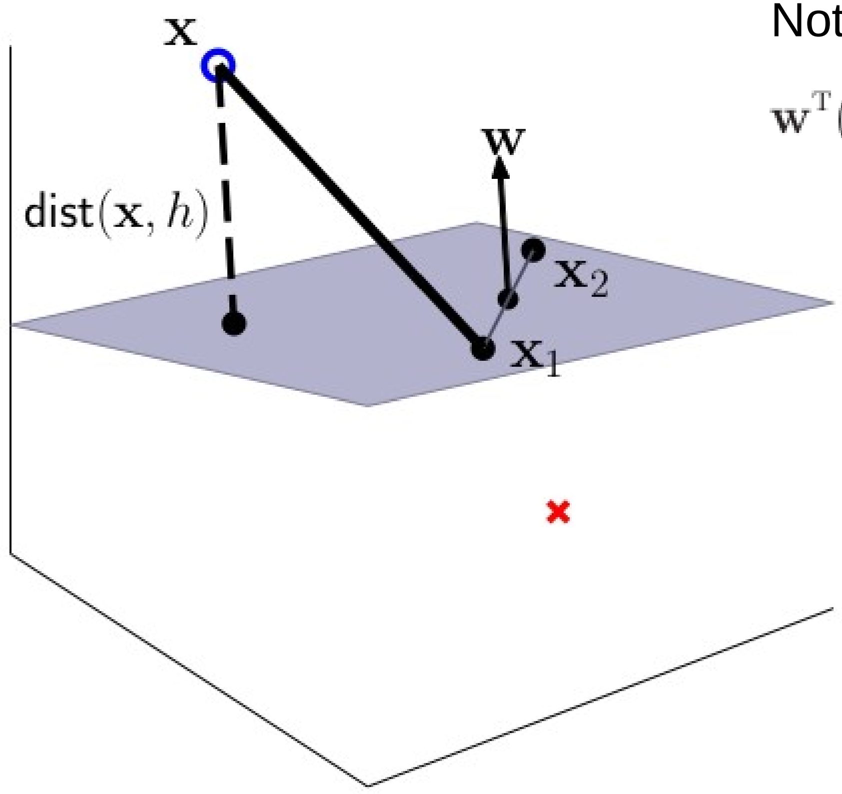
<u>Definición</u>. Un hiperplano separador (grueso) separa los datos si:

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

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:

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

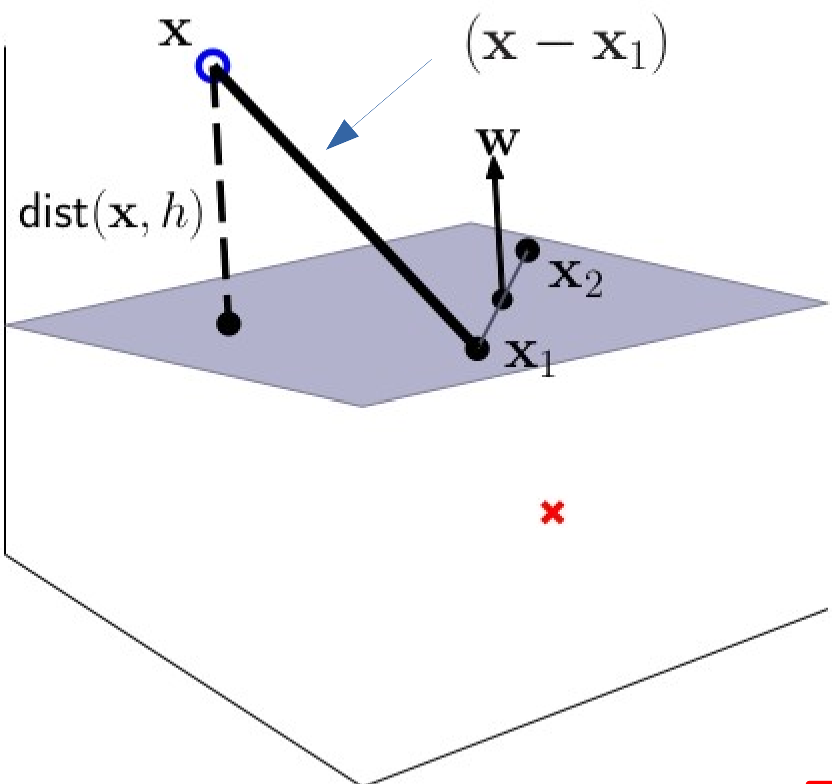
Esto porque para cualquier punto que este sobre el hiperplano se da que:

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

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

Margen del hiperplano separador

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



Proyección de la hipotenusa a la dirección del vector ortonormal

- b

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

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

Entonces: $\boxed{\text{dist}(\mathbf{x}, h) = \dfrac{|\mathbf{w}^{\mathrm{T}}\mathbf{x} + b|}{\|\mathbf{w}\|}}$

Margen del hiperplano separador

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

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

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

$$|\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b| = |y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)| = y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b),$$

Datos separados

Margen del hiperplano separador

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

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

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

$$|\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b| = |y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)| = y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b),$$

Datos separados

La distancia la podemos reescribir según:

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

Margen del hiperplano separador

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

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

$$|\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b| = |y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)| = y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b),$$

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

Datos separados

La distancia la podemos reescribir según:

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

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

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

Margen del hiperplano separador

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

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

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

$$|\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b| = |y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)| = y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b),$$

Datos separados

La distancia la podemos reescribir según:

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

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

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

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

Máximo margen

Margen del hiperplano separador

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

Encontrar los pesos que maximizan la distancia al dato más cercano

¡Minimizar denominador!

Margen del hiperplano separador

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

Encontrar los pesos que maximizan la distancia al dato más cercano

¡Minimizar denominador!

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

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

Hiperplano separador de máximo margen

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

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

Hiperplano separador de máximo margen

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

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

Ejemplo:

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

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

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

Hiperplano separador de máximo margen

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

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

Ejemplo:

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

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

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

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

- (i) y (iii) dan: $w_1 \geq 1$
- (ii) y (iii) dan: $w_2 \leq -1$

Tomamos el borde: $(b = -1, w_1 = 1, w_2 = -1)$

Hiperplano separador de máximo margen

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

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

Ejemplo:

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

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

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

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

- (i) y (ii) dan: $w_1 \geq 1$
- (ii) y (iii) dan: $w_2 \leq -1$

Tomamos el borde: $(b = -1, w_1 = 1, w_2 = -1)$

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

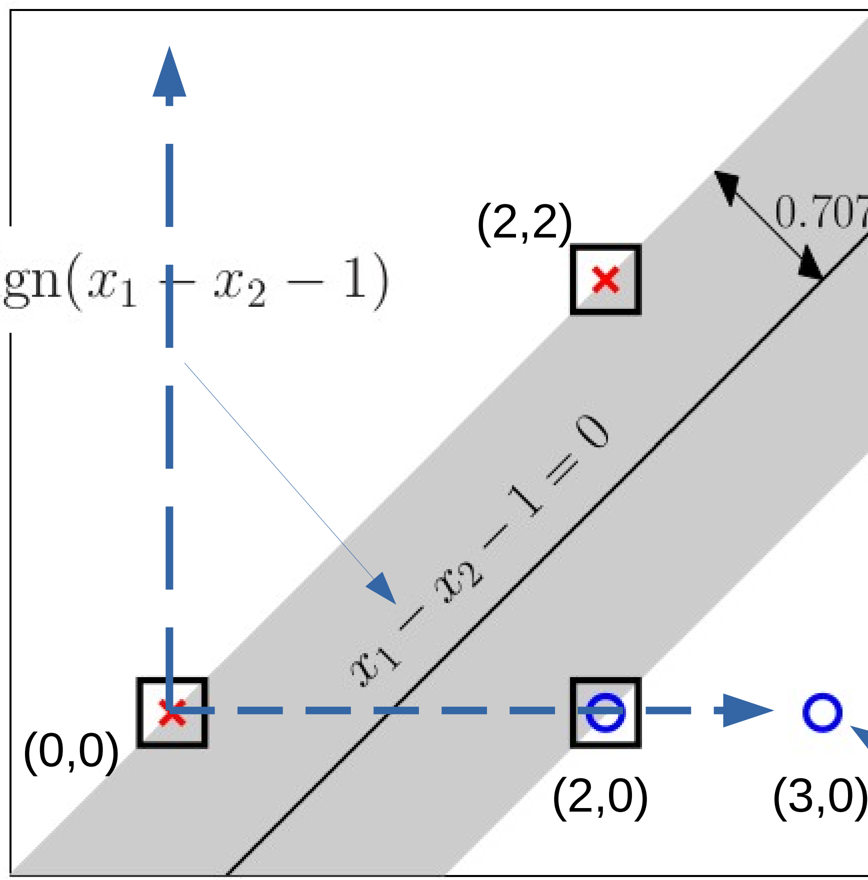
y $$\frac{1}{\|\mathbf{w}^*\|} = \frac{1}{\sqrt{2}} \approx 0.707.$$

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

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

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

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



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

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

$\uparrow$

Support Vectors

Este dato no lo usamos!

Support Vector Machines

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

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

Expresión matricial

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

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

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

Se reduce a un problema de optimización cuadrática para encontrar Q y P.

Support Vector Machines

Forma estándar de problema QP:

$$\begin{aligned} \underset{\mathbf{u}\in\mathbb{R}^L}{\text{minimize:}} \quad & \tfrac{1}{2}\mathbf{u}^{\mathrm{T}}\mathrm{Q}\mathbf{u} + \mathbf{p}^{\mathrm{T}}\mathbf{u} \\ \text{subject to:} \quad & \mathrm{A}\mathbf{u} \geq \mathbf{c}. \end{aligned}$$

donde:

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

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