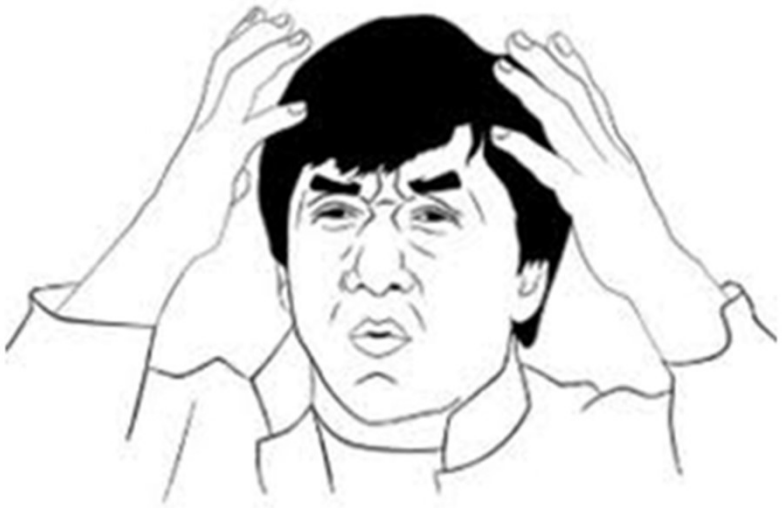
Máquinas de Vectores de Soporte (SVM)

Agenda

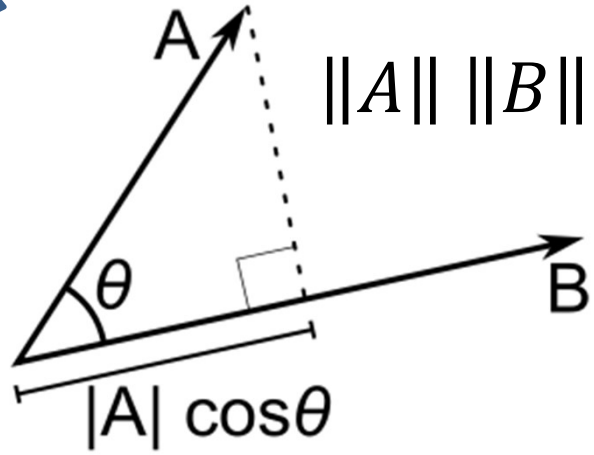
- Antecedentes
  - Conceptos geométricos
  - Algoritmo simple de clasificación
  - Aprendizaje estadístico
- Máquinas de Vectores de Soporte
  - Retrato Generalizado.
  - SVM.
    - Lineal
    - No-lineal
    - Con Ruido
- Discusión



B. Schölkopf, A. Smola (2002). *Learning with Kernels: SVM, Regularization, Optimization, and Beyond*. MIT Press.

Repaso de algunos conceptos geométricos

1 Producto Punto: **A . B**



$$\|A\| \, \|B\| cos\theta$$

2 Magnitud de un vector **x**

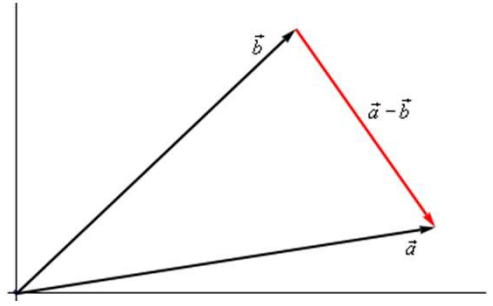
$$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \sqrt{\boldsymbol{x}_1^2 + \boldsymbol{x}_2^2 + \cdots + \boldsymbol{x}_N^2}$$

3 Vector unitario

$$\hat{\boldsymbol{x}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$$

4 Distancia entre vectores

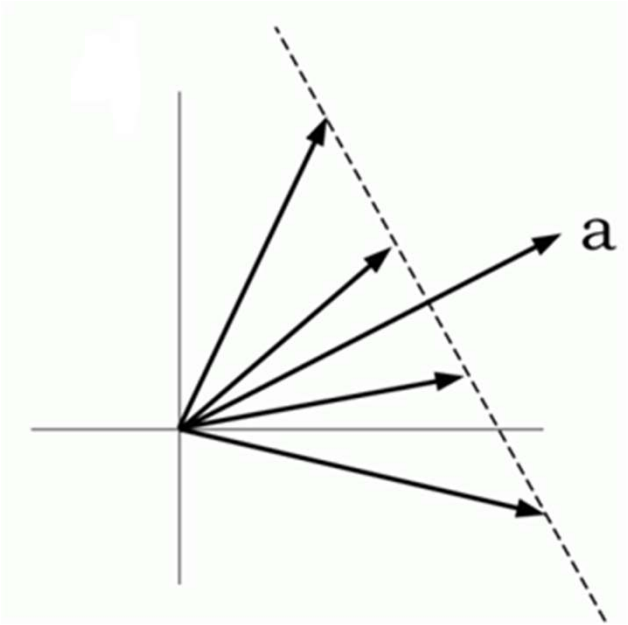
$$\boldsymbol{d} = \|\boldsymbol{a} - \boldsymbol{b}\|$$

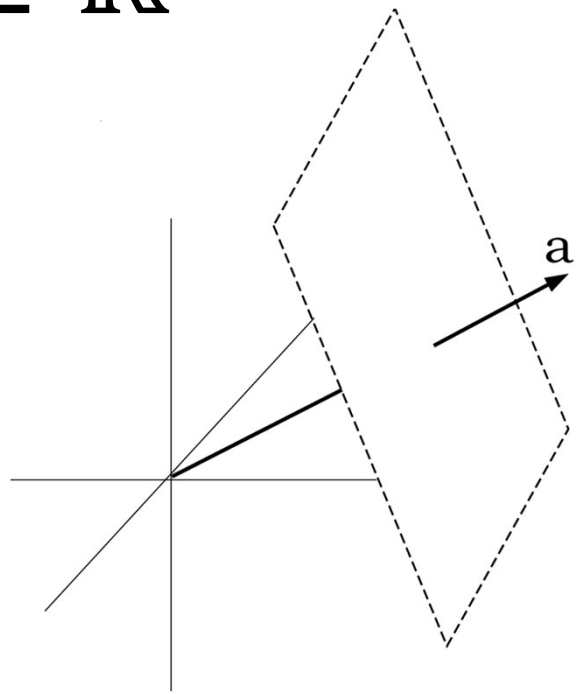


Definición geométrica de un Hiperplano

$$\{\boldsymbol{x} \in \mathrm{H} | \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0\}$$

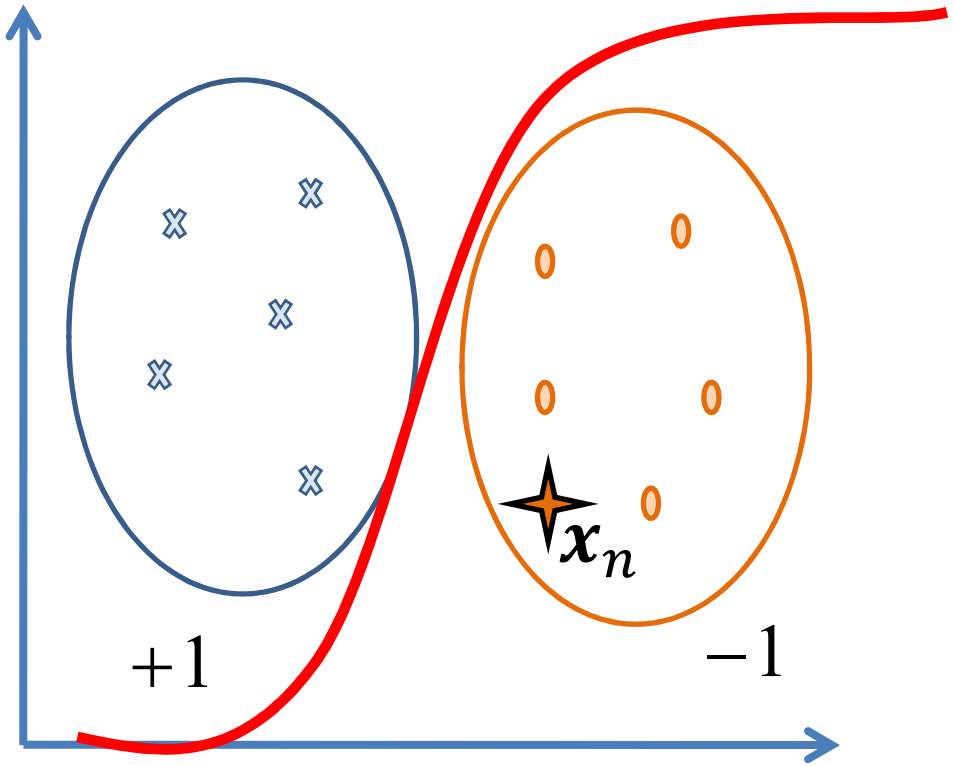
$$\boldsymbol{x} \in H,\ b \in \mathbb{R}$$



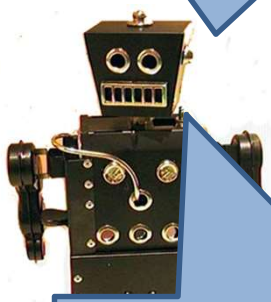


Dado un problema de clasificación binaria …

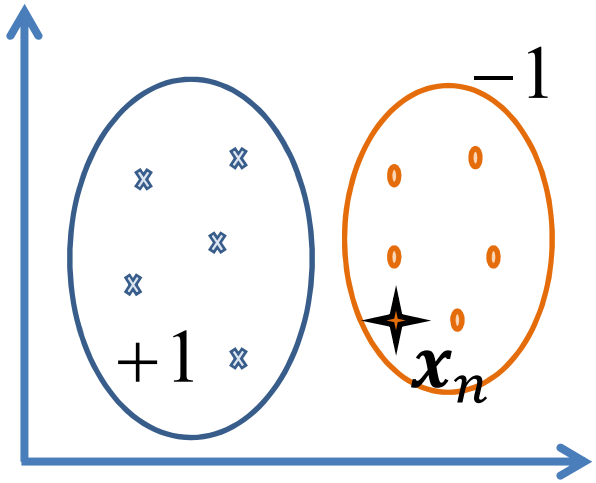
$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)$
$\boldsymbol{x} \in \mathbb{R}^N,\ y \in \{\pm 1\}$



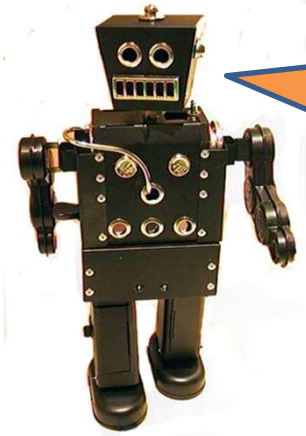
¿Cómo podemos asignar cada nuevo patrón a la clase a la que pertenece?



Encontremos una superficie de decisión!!



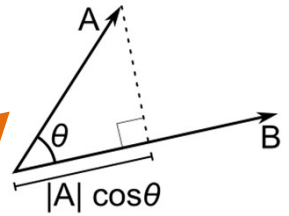
$$(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$$
$$x \in \mathbb{R}^N,\ y \in \{\pm 1\}$$



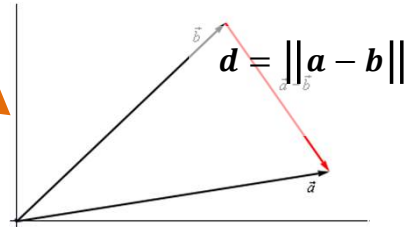
Para elegir la clase de (x_n, y_n), necesitamos una medida de similitud con $\mathcal{X}$, y designar la clase respecto a las muestras más similares.

$$k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$
$$(x_i, x_j) \rightarrow k(x_i, x_j)$$
$$k(x_i, x_j) = k(x_j, x_i)$$

$$\langle x_i, x_j \rangle = \sum_{s=1}^{N} x_s^i \cdot x_s^j$$



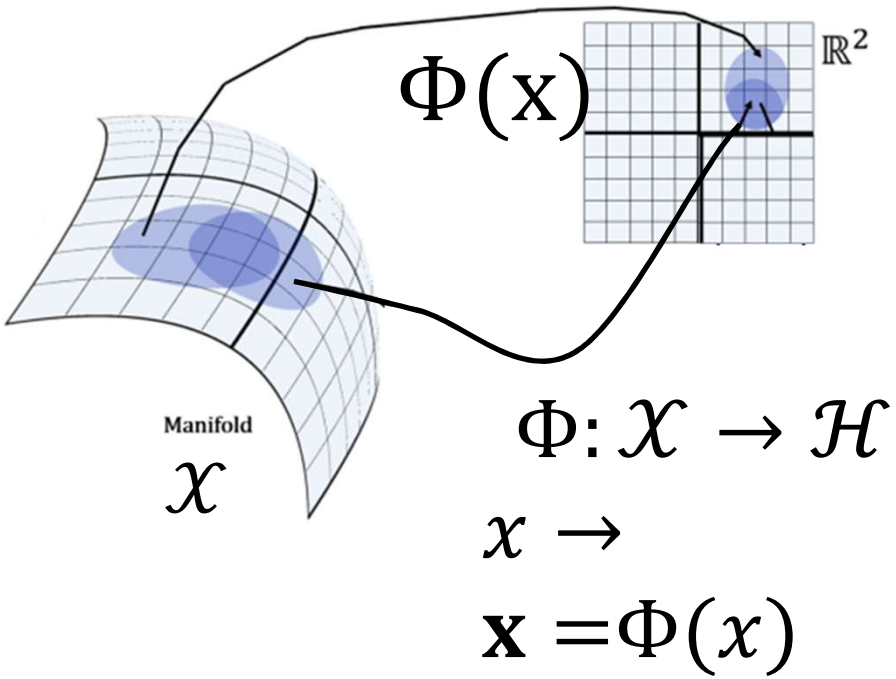
$$\|x\|$$



El **producto punto** es una función kernel

Función Kernel

La función de similitud solo funciona en un espacio Hilbert



1. Medida de Similitud en $\mathcal{H}$
   $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle = \langle \Phi(\boldsymbol{x_i}), \Phi(\boldsymbol{x_j}) \rangle$

2. Análisis Geométrico usando algebra lineal y geometría analítica.

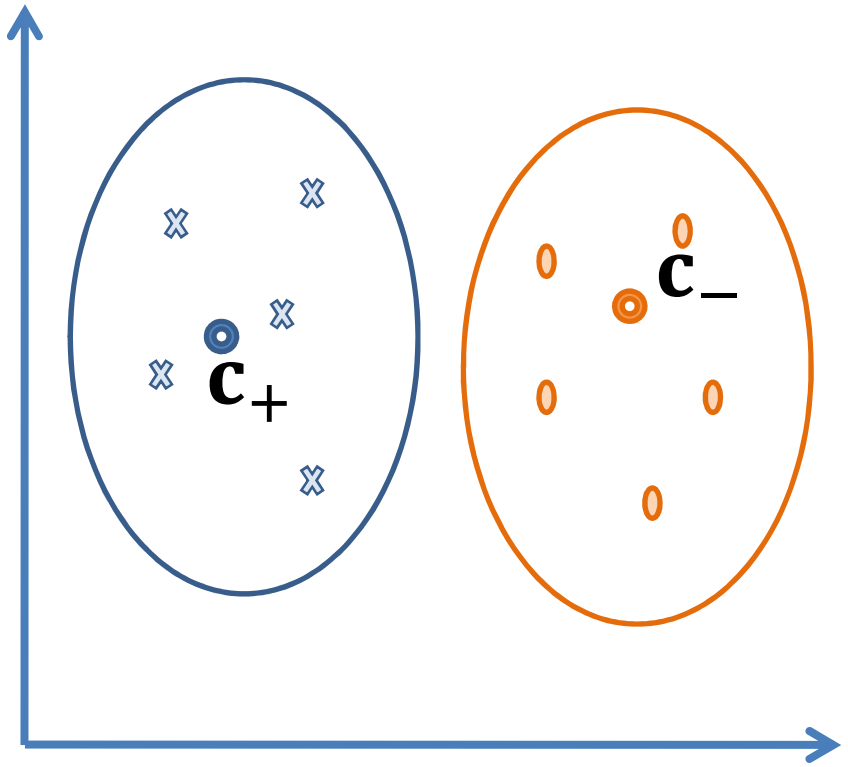
3. El producto punto, es un caso especial de Φ. Por lo que es posible realizar un análisis no-lineal.

Un algoritmo simple para clasificación binaria (2)

Paso 1: Calcular las medias de las dos clases

$$\boldsymbol{c}_+ = \frac{1}{m_+} \sum_{\{i|y_i=+1\}} \boldsymbol{x}_i$$

$$\boldsymbol{c}_- = \frac{1}{m_-} \sum_{\{i|y_i=-1\}} \boldsymbol{x}_i$$



m_+, m_- # de elementos de la clase +1 y -1 respectivamente

Un algoritmo simple para clasificación binaria (3)

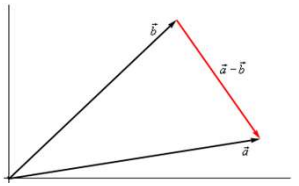
Paso 2: Construimos una superficie de decisión

$$\mathbf{c} = \left(\frac{\mathbf{c}_+ + \mathbf{c}_-}{2}\right)$$

$$\mathbf{w} = \mathbf{c}_+ - \mathbf{c}_-$$

$$\boldsymbol{x}_{\text{new}} = \boldsymbol{x}_n$$

Al proyectar ($\boldsymbol{x}_n - \boldsymbol{c}$) en **w** podemos clasificar el nuevo vector



Construcción formulada en términos del producto punto

Un algoritmo simple para clasificación binaria (4)

Paso 3: Para un $\mathbf{x}_n$ calculamos la clase al proyectar $\mathbf{x}_n$ - **c** en **w**

$$y = sgn(\langle \mathbf{x}_n - \boldsymbol{C}, \mathbf{w} \rangle)$$

¿Cómo?

Un algoritmo simple para clasificación binaria (4)

Paso 3: Para un $\mathbf{x}_n$ calculamos la clase al proyectar $\mathbf{x}_n$ - $\mathbf{c}$ en $\mathbf{w}$

$$
\begin{aligned}
y &= sgn(\langle \mathbf{x}_n - \boldsymbol{C}, \mathbf{w} \rangle) \\
&= sgn(\langle \mathbf{x}_n, \mathbf{w} \rangle - \langle \boldsymbol{C}, \mathbf{w} \rangle) \\
&= sgn(\langle \mathbf{x}_n, (\mathbf{c}_+ - \mathbf{c}_-) \rangle - \langle \boldsymbol{C}, (\mathbf{c}_+ - \mathbf{c}_-) \rangle) \\
&= sgn(\langle \mathbf{x}_n, \mathbf{c}_+ \rangle - \langle \mathbf{x}_n, \mathbf{c}_- \rangle - \langle \boldsymbol{C}, (\mathbf{c}_+ - \mathbf{c}_-) \rangle) \\
&= sgn\left(\langle \mathbf{x}_n, \boldsymbol{c}_+ \rangle - \langle \mathbf{x}_n, \boldsymbol{c}_- \rangle - \left\langle \left(\frac{\mathbf{c}_+ + \boldsymbol{c}_-}{2}\right), (\boldsymbol{c}_+ - \boldsymbol{c}_-) \right\rangle\right) \\
&= sgn\left(\langle \mathbf{x}_n, \boldsymbol{c}_+ \rangle - \langle \mathbf{x}_n, \boldsymbol{c}_- \rangle - \left\langle \left(\frac{\boldsymbol{c}_+}{2} + \frac{\boldsymbol{c}_-}{2}\right), (\boldsymbol{c}_+ - \boldsymbol{c}_-) \right\rangle\right)
\end{aligned}
$$

Un algoritmo simple para clasificación binaria (4)

Paso 3: Para un $\mathbf{x}_n$ calculamos la clase al proyectar $\mathbf{x}_n$ - $\mathbf{c}$ en $\mathbf{w}$

$$y_n = sgn(\langle \mathbf{x}_n, \mathbf{c}_+ \rangle - \langle \mathbf{x}_n, \mathbf{c}_- \rangle + b)$$

$$= sgn\left(\langle \mathbf{x}_n, \mathbf{c}_+ \rangle - \langle \mathbf{x}_n, \mathbf{c}_- \rangle - \left\langle \left(\frac{\mathbf{c}_+}{2} + \frac{\mathbf{c}_-}{2}\right), (\mathbf{c}_+ - \mathbf{c}_-) \right\rangle\right)$$

$$\mathrm{b} = -\left\langle \left(\frac{\mathbf{c}_+}{2} + \frac{\mathbf{c}_-}{2}\right), (\boldsymbol{c}_+ - \boldsymbol{c}_-) \right\rangle$$

$$\mathrm{b} = -\frac{1}{2}\langle \mathbf{c}_+, \mathbf{c}_+ \rangle - \frac{1}{2}\langle \mathbf{c}_-, \mathbf{c}_+ \rangle + \frac{1}{2}\langle \mathbf{c}_+, \mathbf{c}_- \rangle + \frac{1}{2}\langle \mathbf{c}_-, \mathbf{c}_- \rangle$$

$$\mathrm{b} = -\frac{1}{2}\langle \mathbf{c}_+, \mathbf{c}_+ \rangle + \frac{1}{2}\langle \mathbf{c}_-, \mathbf{c}_- \rangle$$

$$\mathrm{b} = \frac{1}{2}(\|\mathbf{c}_-\| - \|\mathbf{c}_+\|)$$

Si las medias de cada clase tienen la misma norma, $b \rightarrow 0$

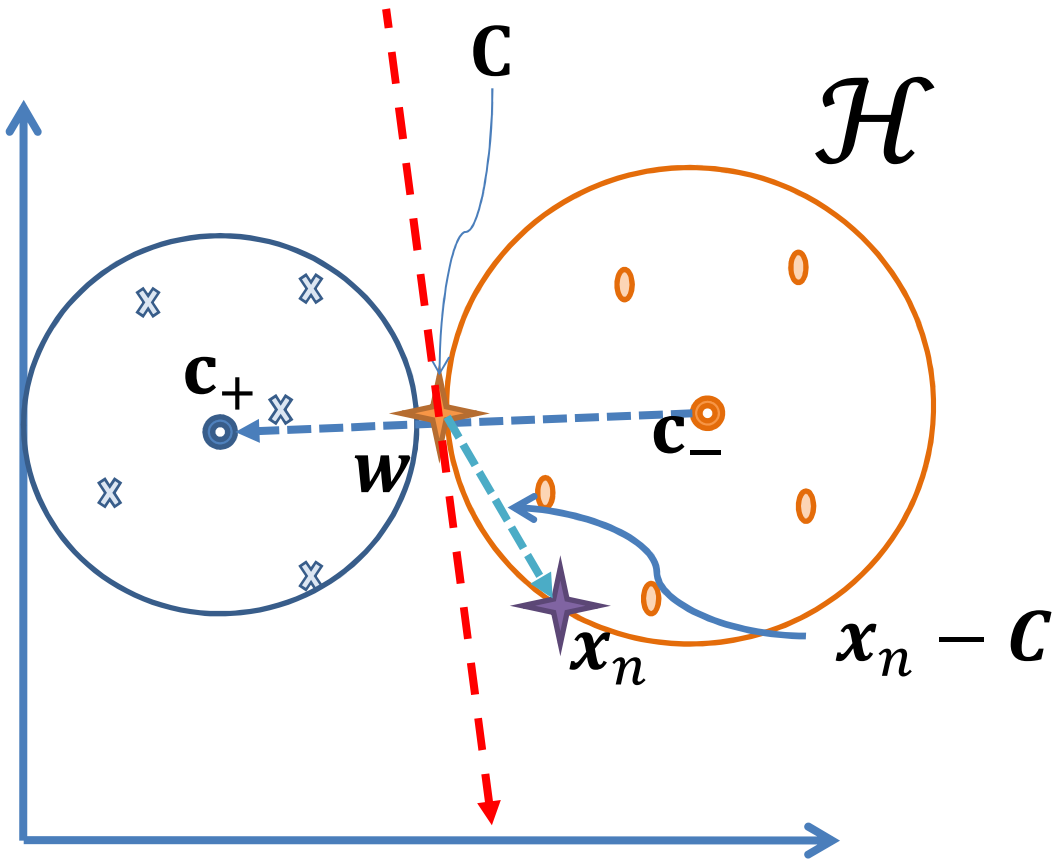
Límite de Decisión

$$y_n = sgn(\langle \mathbf{x}_n, \mathbf{c}_+ \rangle - \langle \mathbf{x}_n, \mathbf{c}_- \rangle + b)$$

Esta función también define un Hiperplano:

$$\{\boldsymbol{x} \in \mathrm{H} | \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0\}$$
$$\boldsymbol{x} \in H,\ b \in \mathbb{R}$$

Un nuevo vector $\mathbf{x}_n$ es clasificado como +1 o -1 si:
El vector $\mathbf{x}_n - C$ que conecta a C con $\mathbf{x}_n$, tiene un ángulo $\leq \frac{\pi}{2}$ con $\mathbf{w}_i$.
En la función del hiperplano, lo anterior corresponde a un cambio de signo.



Reescribamos el problema, es decir, **c y w** en términos de $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ y *k*.

$$y_n = sgn(\langle \mathbf{x}_n, \mathbf{c}_+ \rangle - \langle \mathbf{x}_n, \mathbf{c}_- \rangle + b)$$

$$y_n = sgn\left(\frac{1}{m_+} \sum_{\{i|y_i=+1\}} \langle \mathbf{x}_n, \mathbf{x}_i \rangle - \frac{1}{m_-} \sum_{\{i|y_i=-1\}} \langle \mathbf{x}_n, \mathbf{x}_i \rangle + b\right)$$

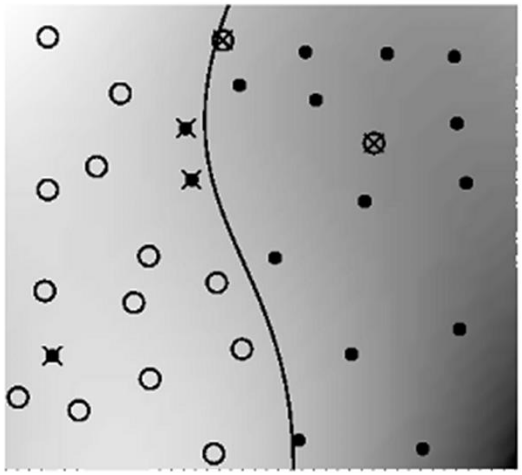
$$y_n = sgn\left(\frac{1}{m_+} \sum_{\{i|y_i=+1\}} k(\mathbf{x}_n, \mathbf{x}_i) - \frac{1}{m_-} \sum_{\{i|y_i=-1\}} k(\mathbf{x}_n, \mathbf{x}_i) + b\right)$$

$$b = \frac{1}{2}\left(\frac{1}{m_-^2} \sum_{\{(i,j)|y_i=y_j=-1\}} k(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{m_+^2} \sum_{\{(i,j)|y_i=y_j=+1\}} k(\mathbf{x}_i, \mathbf{x}_j)\right)$$

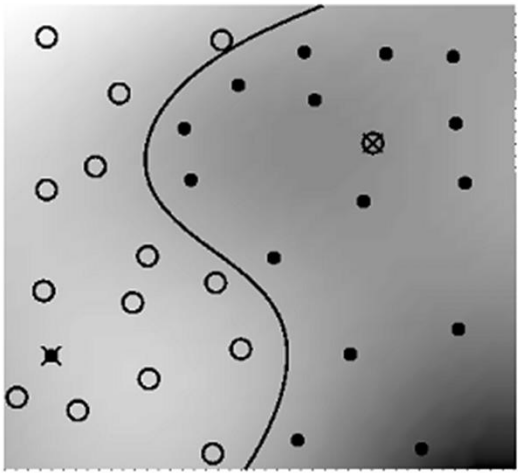
Observe que, si las medias de las clases tienen la misma distancia al origen $b \rightarrow 0$.

Tres clasificadores en 2-D

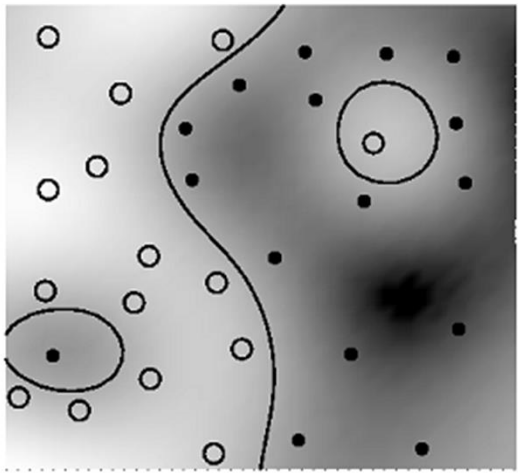
A



B



C



Buscamos $f: \mathcal{X} \to \{\pm 1\}$

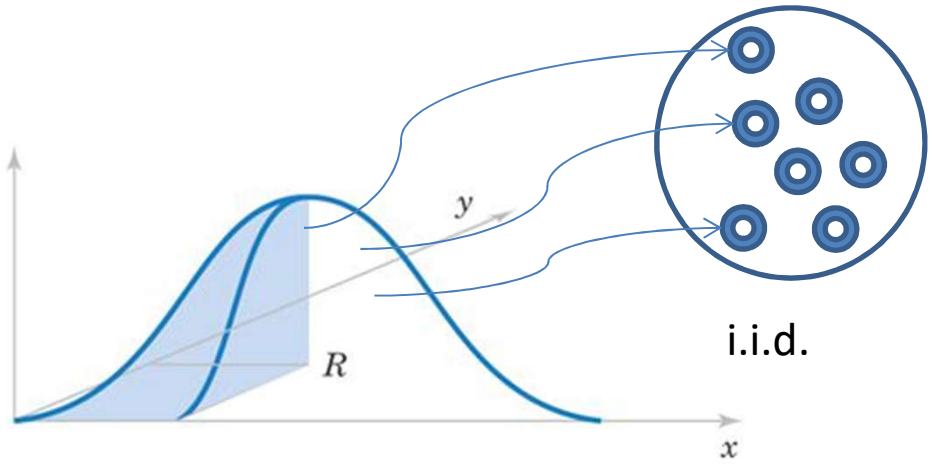
En lugar de graficar f_1, f_2 y f_3, solo se muestran las superficies de decisión.

Pregunta: ¿Cuál clasificador es mejor con los datos de entrenamiento?
¿Cuál será mejor con **datos nuevos**?

Objetivos del Aprendizaje Estadístico (SL)

Incorporar estos argumentos "intuitivos" en un marco formal, a través de las propiedades de f.

Dados $(\mathbf{x}_i, y_i)$ provenientes de $P(x, y)$



$P(x, y)$

i.i.d.

Buscamos $f(\mathbf{x}) = y$, tal que sirva para clasificar un $\mathbf{x}_{new}$ también provienete de $P(x, y)$

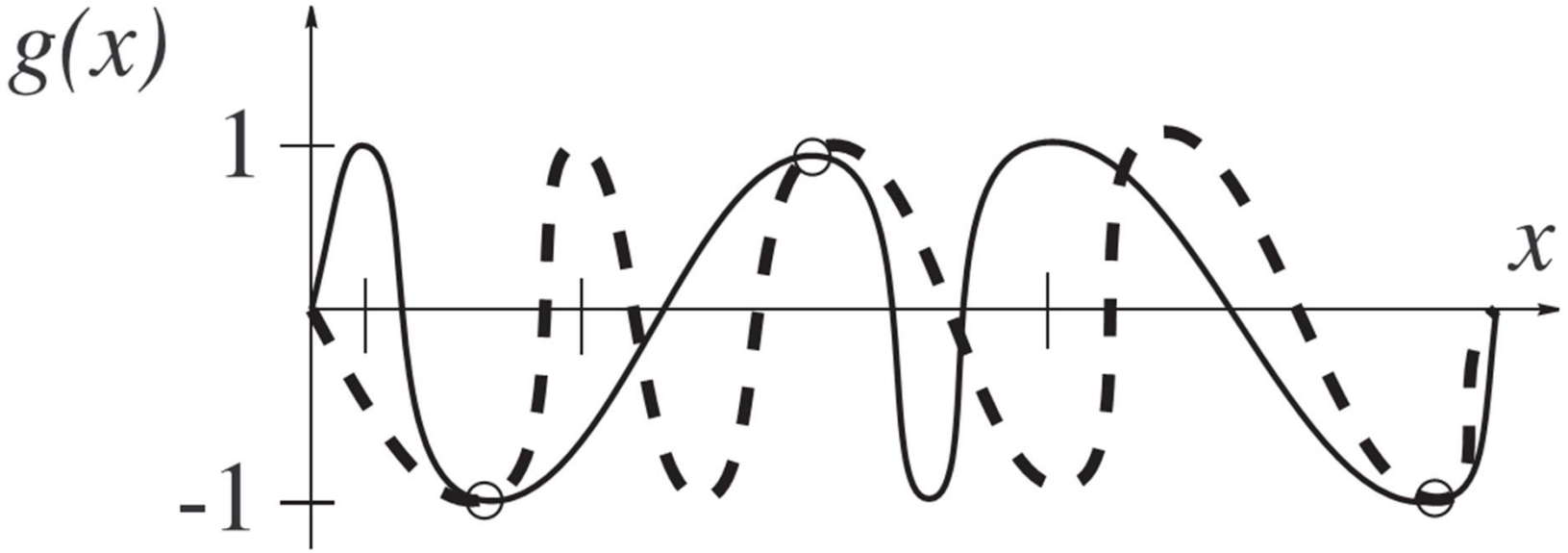
La precisión de clasificación de f es medida con la **función de pérdida** (i.e. riesgo, error, etc.) *cero-uno*:

$$c(\mathbf{x}, y, f(\mathbf{x})) = \frac{1}{2}|f(\mathbf{x}) - y|$$

¿Qué valores toma esta Función de pérdida?

http://michael.hahsler.net/SMU/EMIS7332/R/viz_classifier.html

Pero, ¿qué pasa si solo minimizamos la función de pérdida?



¿Qué función es preferible?, ¿Por qué?, ¿Es suficiente minimizar el error?

No es suficiente minimizar en promedio el Riesgo Empírico

Minimizar el **Riesgo empírico (ERM)**

$$R_{emp}[f] = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}|f(\mathbf{x}) - y| \qquad c(\mathbf{x}, y, f(\mathbf{x})) = \frac{1}{2}|f(\mathbf{x}) - y|$$

No implica minimizar el **riesgo verdadero** sobre una muestra de prueba también proveniente de $P(x, y)$

$$R[f] = \frac{1}{m}\int \frac{1}{2}|f(x) - y|\, dP(x, y)$$

P(x, y): Distribución desconocida

R[f] no se puede evaluar

La ERM obtiene *f* óptima cuando el problema esta **bien planteado**(well-posed)

- Un problema esta ***bien-planteado*** si la solución:
  - Existe
  - Es única
  - Es consistente. Es decir, conforme la cantidad de datos $\mathrm{n} \to \infty$ los parámetros θ de f convergen.

Riesgo Estructural

- SL o Teoría de Vapnik-Chervonenkis (VC) establecen que:
  - Se debe restringir la **CAPACIDAD** de f de acuerdo a los datos disponibles.

- SL provee **LIMITES** sobre los errores de prueba los cuales dependen:
  - $R_{emp}[f]$
  - Capacidad de f

Minimización del Riesgo Estructural

La dimensión VC como medida de Capacidad de f

- Para cada $f_i \in F$ los datos son separados de forma específica ∴ el etiquetando también es de forma específica. Además, si $y \in \{\pm 1\}$ existen 2^m diferentes formas de etiquetar los datos.

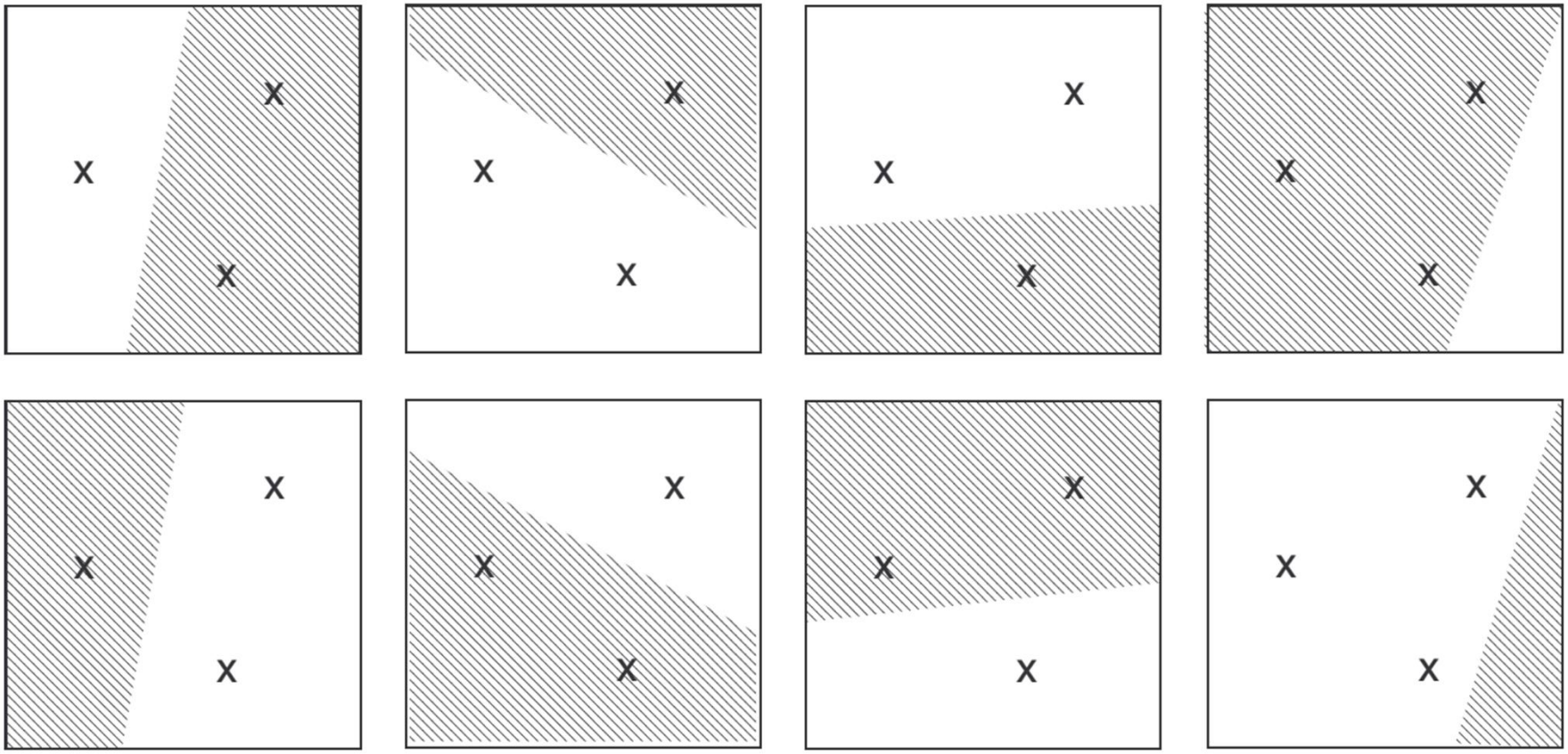
- La **capacidad** es una medida del número de puntos que ***F*** puede separar.

- Si F es muy rica, podrá obtener las 2^m separaciones. En otras palabras, F puede "romper" los *m* puntos. Sin embargo, usualmente puede representar menos separaciones.

- La dimensión VC es:

$$VC(F) = \begin{cases} m & \text{máximo \# } m \text{ que } F \text{ puede romper} \\ \infty & \text{Si no existe } m \end{cases}$$

$f_i \in F$



¿Cuál es la dimensión VC de *F*?, ¿Por qué?

Apreciad los resultados de la Teoría VC

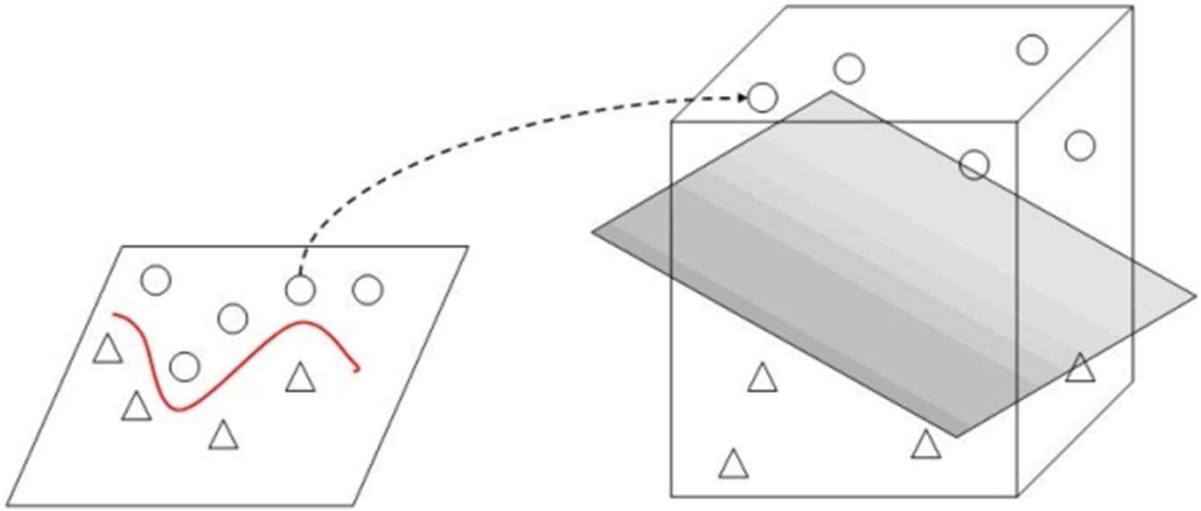
- Si $h < m$ es la dimensión VC de F
  - Entonces, para $f_i \in F | \forall i$
  - Independientemente de $P(x, y)$
  - Con una probabilidad de al menos $1 - \delta$ sobre la muestra de entrenamiento

Entonces:

$$R[f] \leq R_{emp}[f] + \phi(h, m, \delta)$$

Tal que ϕ (Término de Confianza o Capacidad) se define como:

$$\phi(h, m, \delta) = \sqrt{\frac{1}{m}\left(h\left(\ln\frac{2m}{h} + 1\right) + \ln\left(\frac{4}{\delta}\right)\right)}$$



SUPPORT VECTOR MACHINES

Retrato Generalizado (1)

- Propuesto por Vapnik, 1963, asume separación Lineal. Esta basado en 2 supuestos:

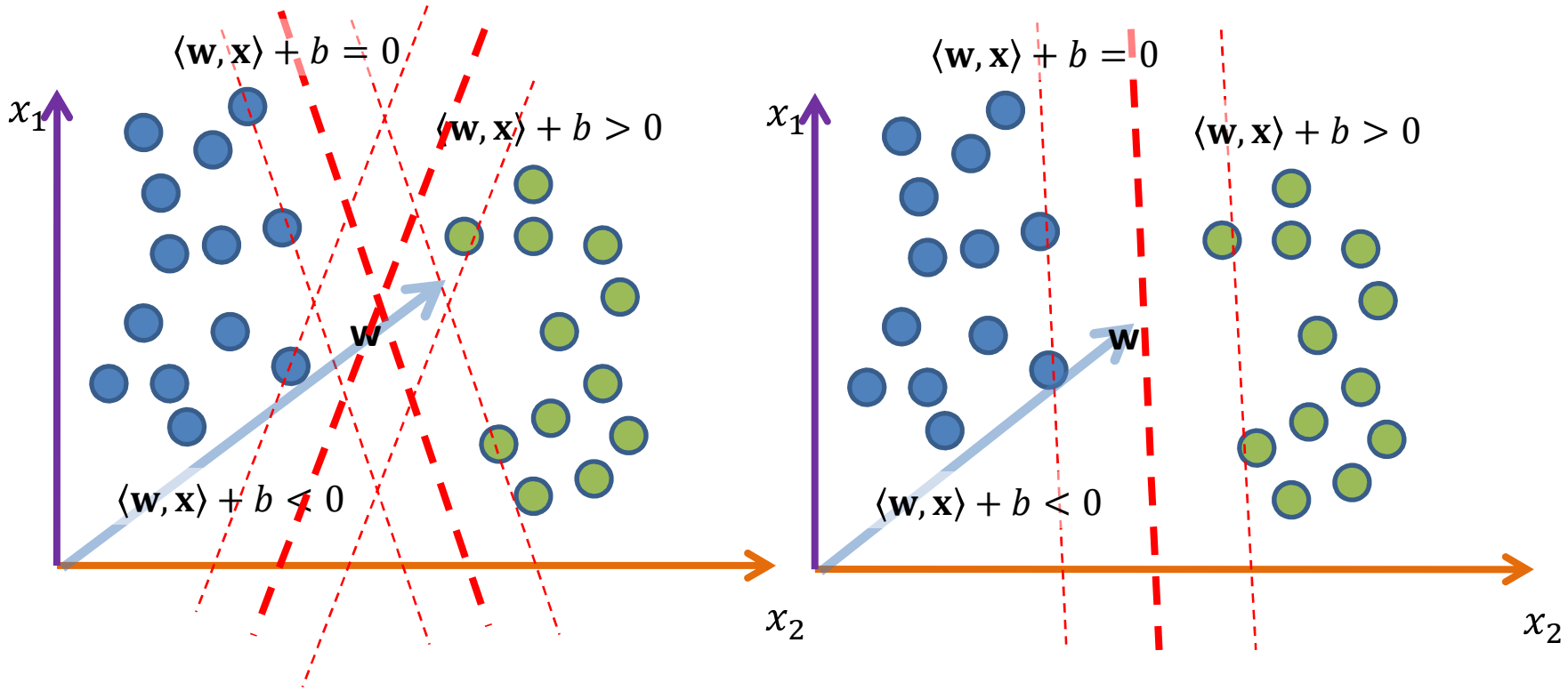
1. Entre todos los hiperplanos que separan los datos, **existe un único hiperplano óptimo** que se distingue por tener el máximo margen de separación entre las clases

$$\underset{\mathrm{x}\in\mathcal{H},\mathrm{b}\in\mathbb{R}}{\text{Maximizar}} \ \min. \|\mathbf{x}-\mathbf{x}_i\|$$

$$\text{Sujeto a}$$
$$\mathbf{x}\in\mathcal{H}$$
$$\langle \mathbf{w},\mathbf{x}\rangle + b = 0, \forall i = i,\dots,m.$$

Retrato Generalizado (2)

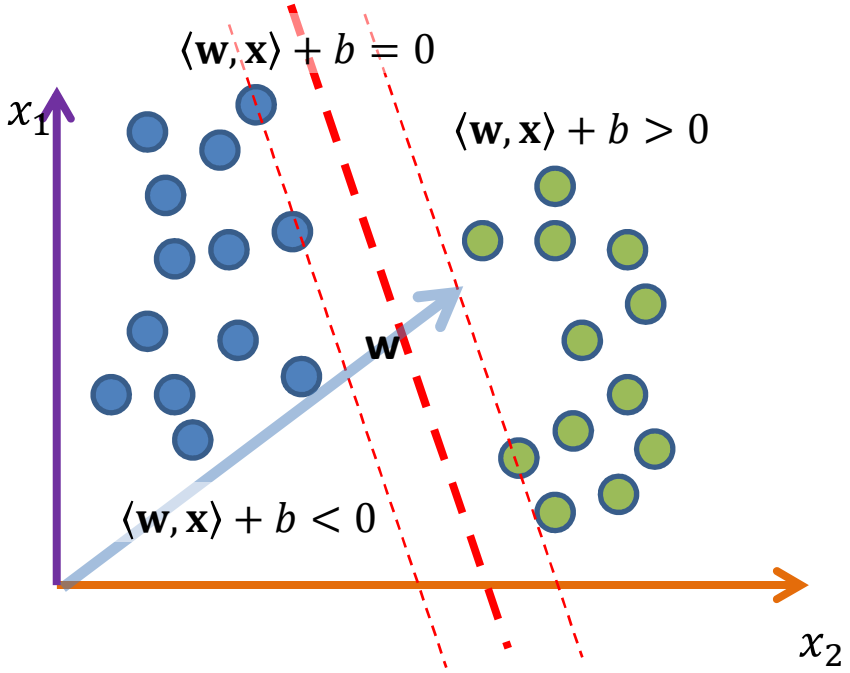
2. La capacidad de la clase F de hiperplanos, decrece conforme el margen crece.

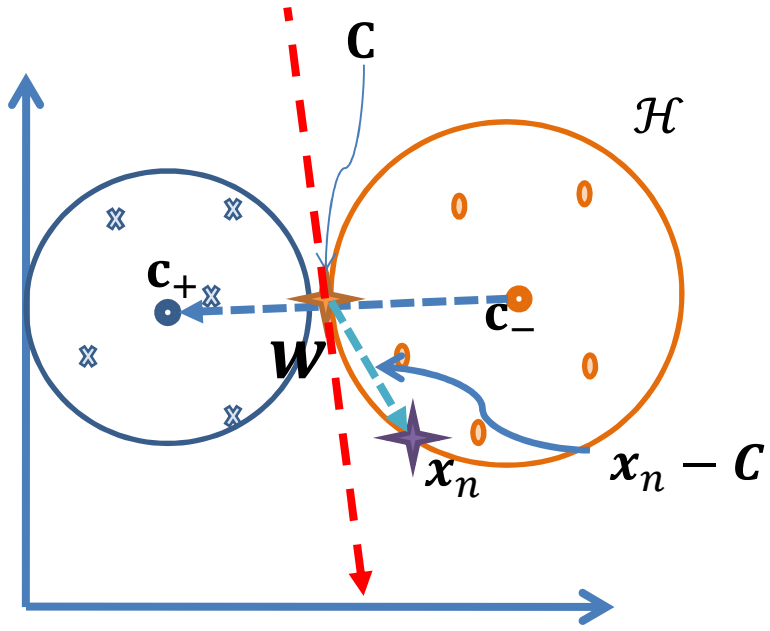


La superficie de decisión es prácticamente la misma que la del algoritmo simple

$$f(\mathbf{x}) = sgn(\langle \mathbf{w}, \mathbf{x} \rangle + b)$$

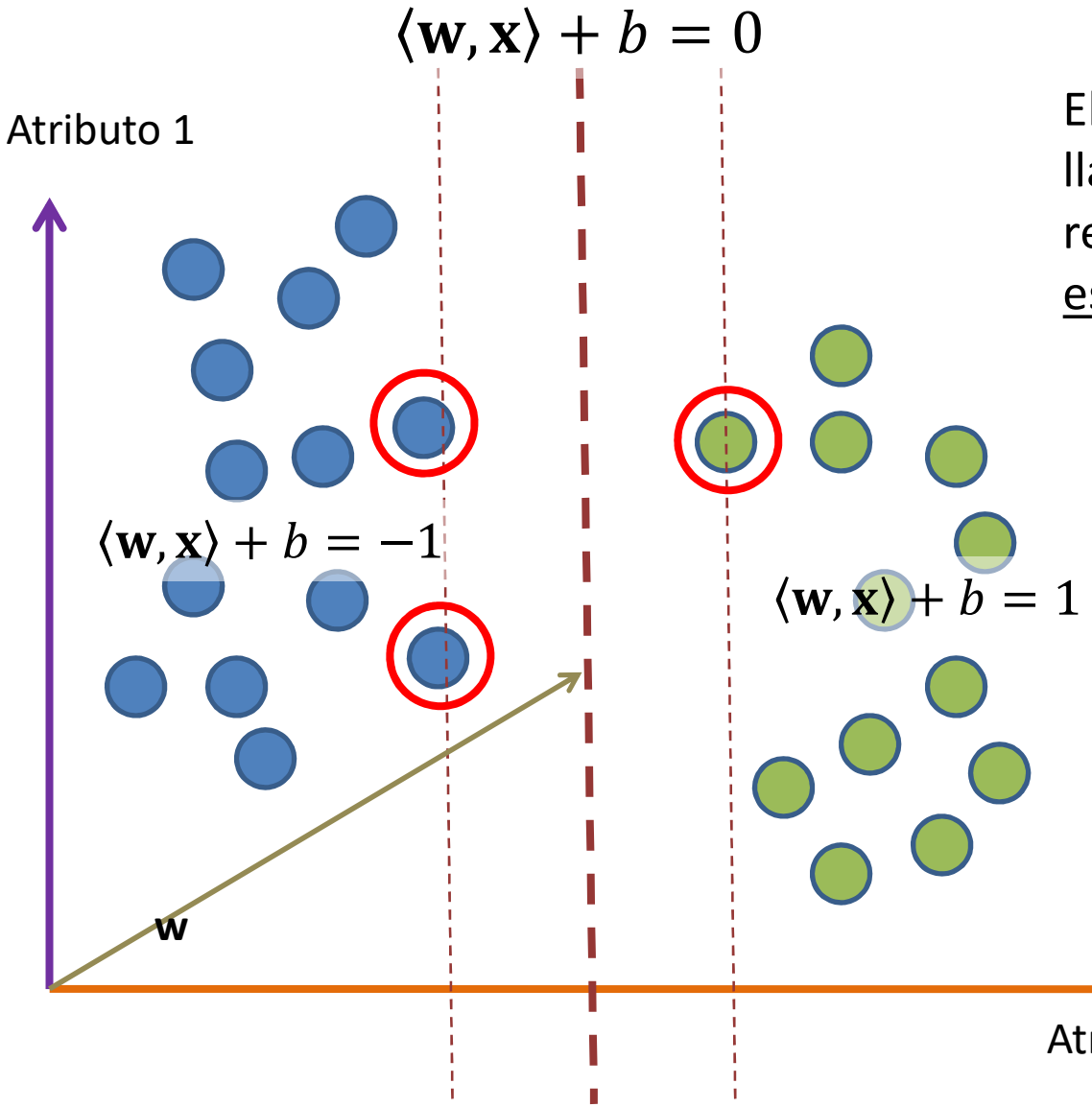
$$y_n = sgn(\langle \mathbf{x}_n, \mathbf{c}_+ \rangle - \langle \mathbf{x}_n, \mathbf{c}_- \rangle + b)$$





La diferencia es como estimamos **w.**

Definición: Hiperplano canónico



El par $\langle \mathbf{w}, b \rangle \in \mathcal{H} \times \mathbb{R}$ es llamado el *Hiperplano Canónico* respecto $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{H}$, si esta escalado tal que:

$$\min_{i=1,\ldots,m} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = 1$$

¿Qué distancia hay entre el hiperplano y los puntos más cercanos a este?

$\langle \mathbf{w}, b \rangle$ y $\langle -\mathbf{w}, -b \rangle$ satisfacen la definición de hiperplano canónico.

$\langle \mathbf{w}, b \rangle$ y $\langle -\mathbf{w}, -b \rangle$

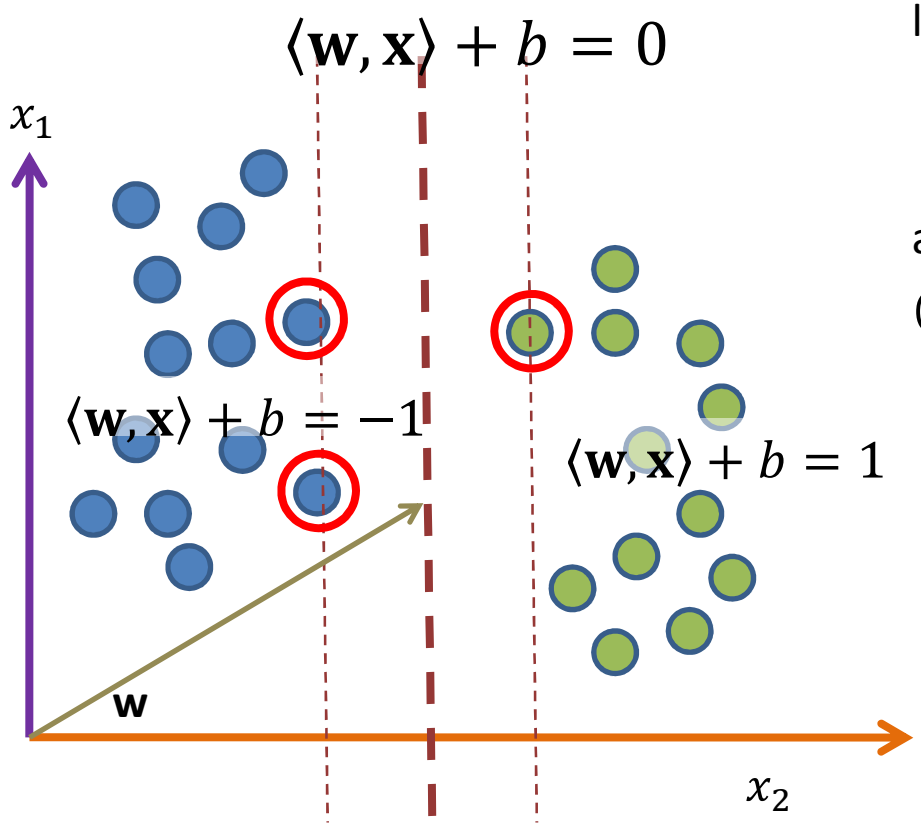
- Estas 2 tuplas forman dos superficies de decisión:

$$f_{\mathbf{w},b}: \mathcal{H} \rightarrow \{\pm 1\}$$

$$\mathbf{x} \mapsto f(\mathbf{x}) = sgn(\langle \mathbf{w}, \mathbf{x} \rangle + b)$$

- Si para cada $\mathbf{x}_i \exists y_i$, los planos definidos por $f_{\mathbf{w},b}$ realizan asignaciones de clase opuestas.

Definición: Margen geométrico



Dado el hiperplano $\{\mathbf{x} \in \mathcal{H} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$ llamamos

$$\rho_{(\mathbf{w},b)} = \frac{y(\langle \mathbf{w}, \mathbf{x} \rangle + b)}{\|\mathbf{w}\|}$$

al margen geométrico del punto $(\mathbf{x}, y) \in \mathcal{H} \times \{\pm 1\}$.

Y el valor mínimo

$$\mathrm{P}_{(\mathbf{w},b)} = \min_{i=1,\ldots,m} \rho_{(\mathbf{w},b)}(\mathbf{x}_\mathrm{i}, y_i)$$

es el margen geométrico del conjunto de entrenamiento $(\mathbf{x}_\mathrm{i}, y_i), \forall i = 1, \ldots, m$.

¿Qué distancia hay entre el hiperplano y los puntos más cercanos a este?

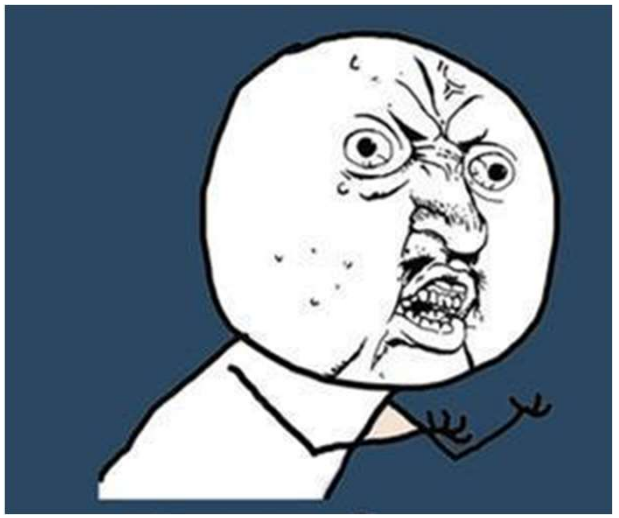
¿Qué relación existe entre minimizar **w** y una buena clasificación?

1. Para un punto $(\mathbf{x}_i, y_i)$ bien clasificado,
   ρ es la distancia al hiperplano.
   $\rho = 0$ si **x** esta sobre el hiperplano.

2. Trabajamos con el hiperplano normalizado
   $\left(\hat{\mathbf{w}}, \hat{b}\right) = \left(\frac{\mathbf{w}}{\|\mathbf{w}\|}, \frac{b}{\|\mathbf{w}\|}\right)$ con $\frac{\mathbf{w}}{\|\mathbf{w}\|} = 1$ y
   solo debemos calcular $y\left(\langle \hat{\mathbf{w}}, \mathbf{x}_\mathrm{i} \rangle + \hat{b}\right)$.

   $\langle \hat{\mathbf{w}}, \mathbf{x}_\mathrm{i} \rangle$ calcula $\mathbf{x}_\mathrm{i} \perp \mathrm{hiperplano}$, y al sumar $\hat{b}$ tenemos la distancia al hiperplano.

3. Al multiplicar $y(hiperplano)$ obtenemos valores del margen
   + si esta bien clasificado el punto
   - de otra forma.



Al minimizar **w** maximizamos la distancia entre las clases, *ergo*, la mejor clasificación posible dado el conjunto de entrenamiento

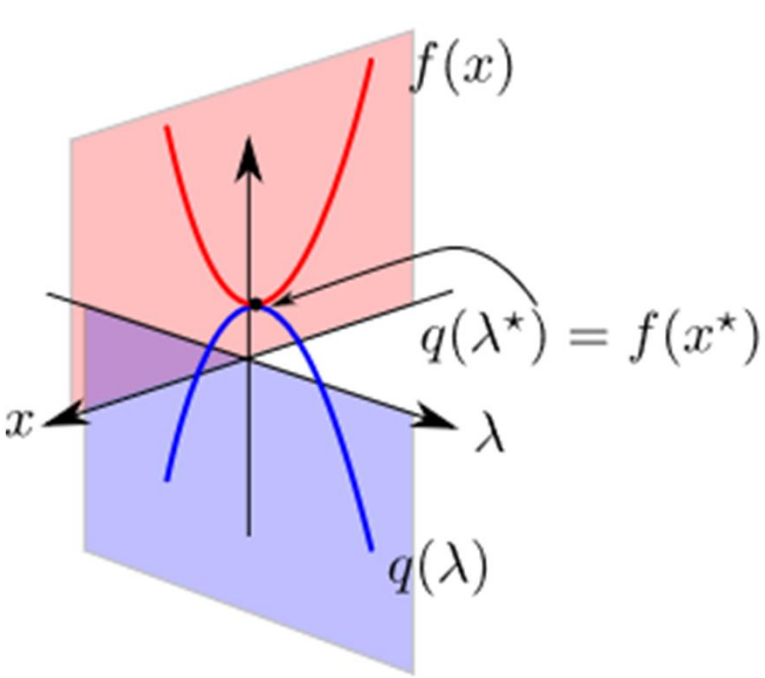
Formulación primal de SVM

$$\text{Min } \tau(w) = \frac{1}{2}\|w\|^2$$

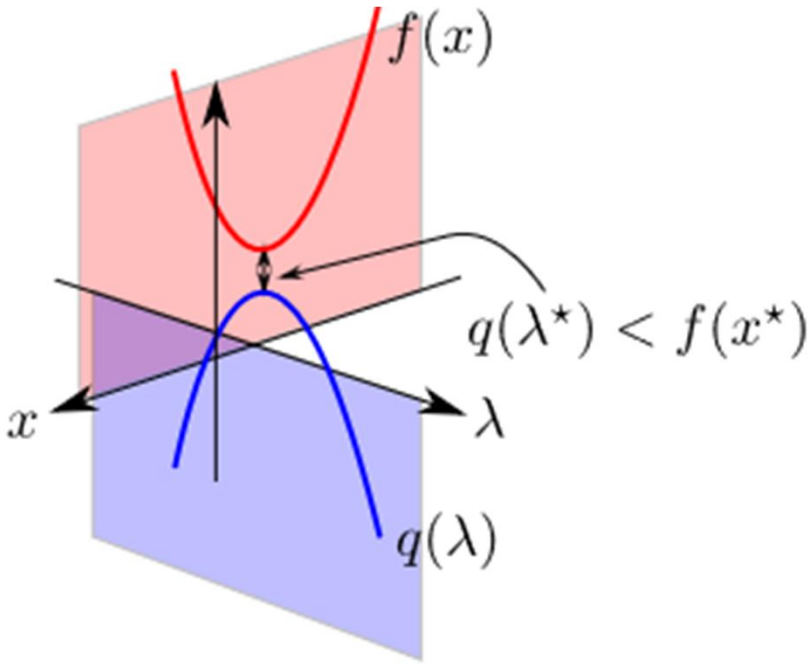
Sujeto a

$$y_i\left(\langle \vec{x}_i, \vec{w} \rangle + b\right) \geq 1, \forall\, i = 1, \ldots, m$$

Optimización, Primal y Dualidad.



strong duality



weak duality

Hacia la formulación dual de SVM

- $\tau(w)$ es la función objetivo
- $y_i(\langle \mathbf{w}, \mathbf{x_i} \rangle + b) \geq 1$ son las restricción de desigualdad.

- Utilizando los multiplicadores de Lagrange $\alpha_i \geq 0$ podemos construir el lagrangeano:

$$L(\vec{w}, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i \left(\langle \vec{x}_i, \vec{w} \rangle + b \right) - 1 \right)$$

Hacia la formulación dual de SVM

1. Minimizar con respecto a las variables *primales* w y b
2. Maximizar con respecto a las variables *duales* α_i

- Para obtener la solución debemos derivar L con respecto a las variables primales:

$$\frac{\partial L}{\partial b}(w,b,\alpha) = 0 \qquad \frac{\partial L}{\partial w}(w,b,\alpha) = 0$$

Condiciones de Optimalidad KKT

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

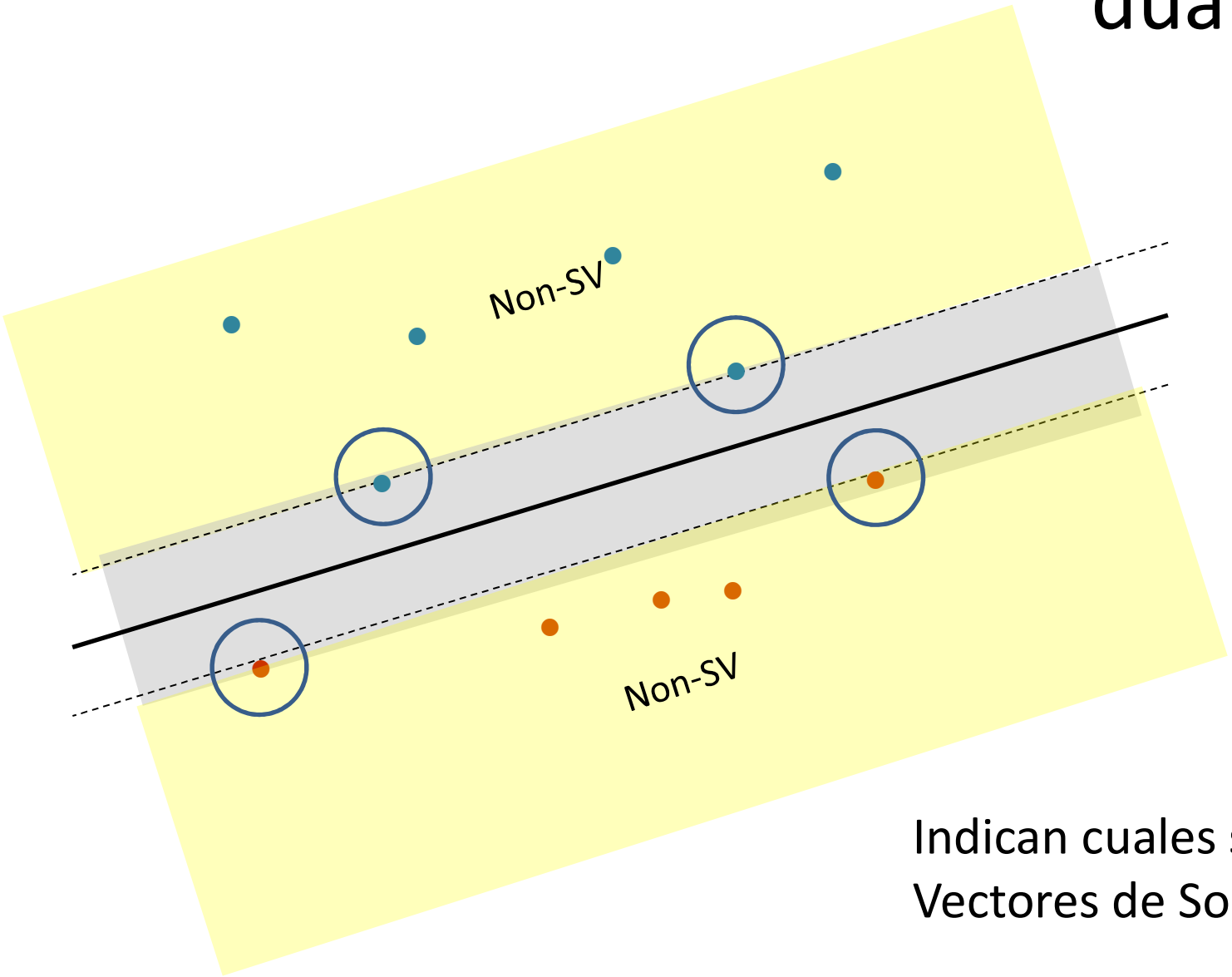
→Expresados como una combinación lineal del conjunto de entrenamiento

$$\frac{\partial L(\mathbf{w},b,\alpha)}{\partial b} = \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$KKT\ cond : \alpha_i [y_i (\mathbf{w}\mathbf{x}_i + b) - 1] = 0$$

→Únicamente los SVs tendrán un $\propto_i$ mayor a cero

¿Qué significado tienen las variables duales?



Indican cuales son los Vectores de Soporte

Formulación Dual de SVM

- Substituyendo las derivadas en el lagrangeano obtenemos la formulación Dual del P.O.:

$$\text{maximizar } f_0(x) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_j \alpha_i y_j y_i x_i, x_j$$

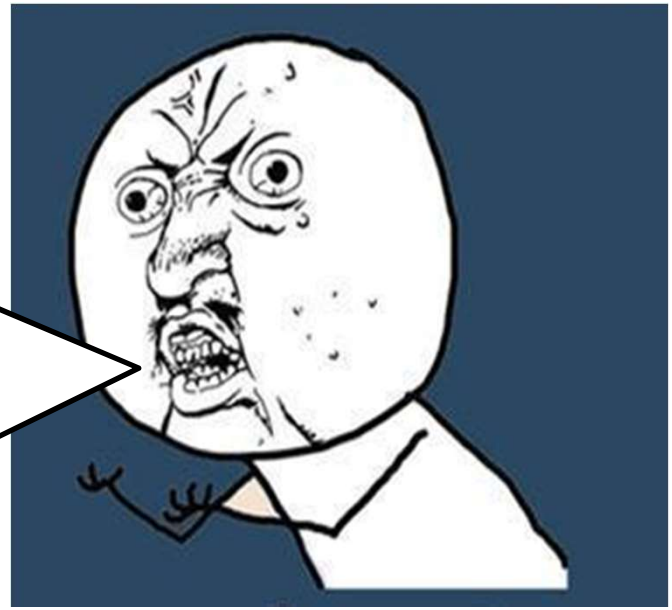
Sujeto a

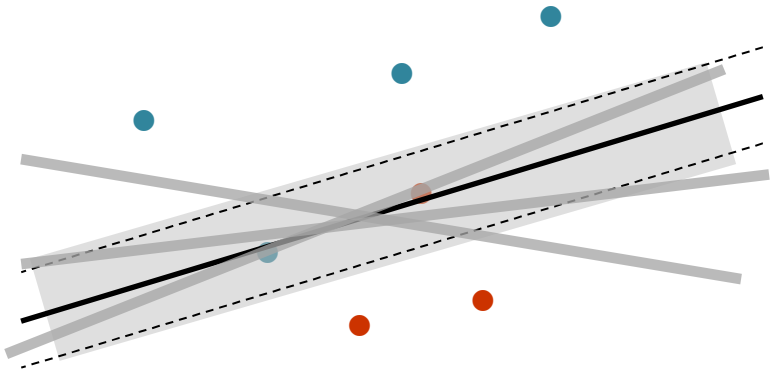
$$\alpha_i \geq 0, i = 1, \ldots, m$$

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

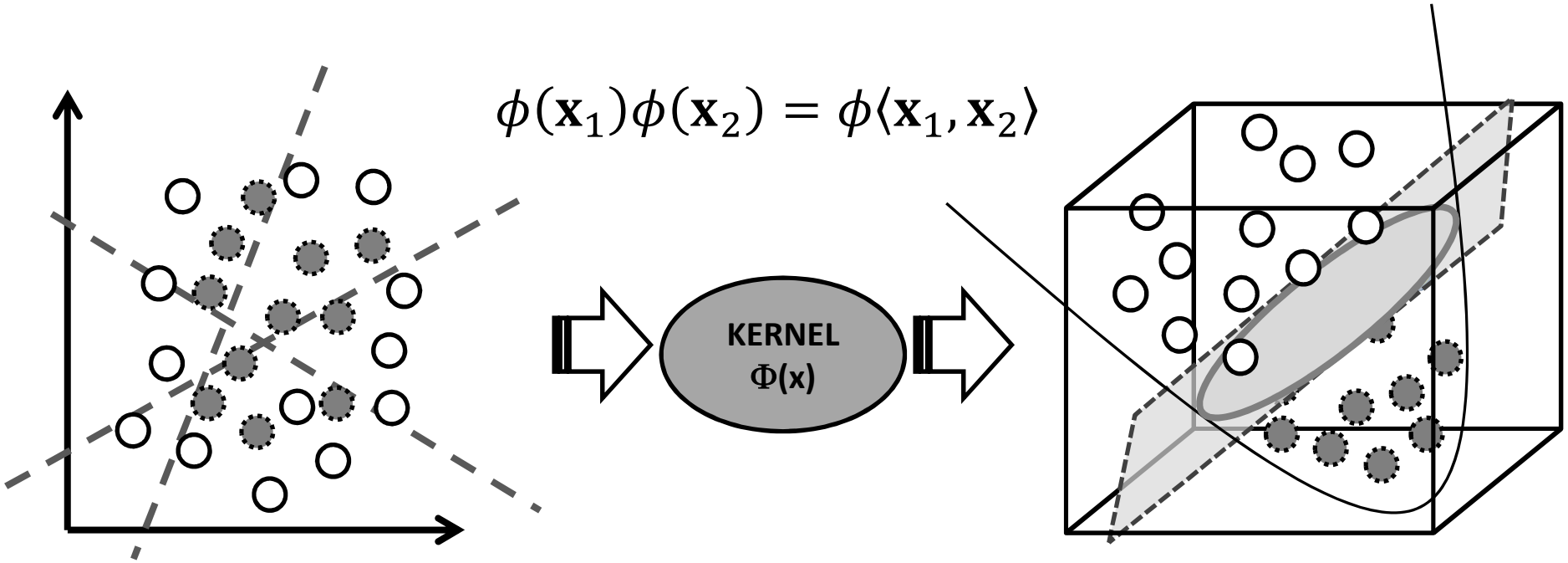
¿Qué sucede si no existe un hiperplano separador?

¿Cómo podemos manejar el ruido y datos etiquetados erróneamente?





Separación No-Lineal: Truco del Kernel



$$Lineal\ (K_{lineal})\quad K_{lineal}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i * \mathbf{x}_j$$

$$Polinomial\ (K_{pol})\quad K_{pol}(\mathbf{x}_i, \mathbf{x}_j) = ((\mathbf{x}_i * \mathbf{x}_j) + r)^d$$

$$Gausiano\ (K_{Gaus})\quad K_{Gaus}(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\gamma ||\mathbf{x}_i * \mathbf{x}_j||^2\right)$$

La formulación dual de SVM es preferida debido al número reducido de variables [Schölkopf, 2002]

- Caso Lineal:

$$\text{maximizar } f(x) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_j \alpha_i y_j y_i x_i, x_j$$

sujeto a

$$\alpha_i \geq 0, i = 1, \ldots, m$$

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

- Caso no-lineal:

$$\textbf{maximizar } f(x) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_j \alpha_i y_j y_i k(x_i, x_j)$$

sujeto a

$$\alpha_i \geq 0, i = 1, \ldots, m$$

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

α: *multiplicadores de Lagrange*
$K(x_i, x_j)$: *truco del kernel*

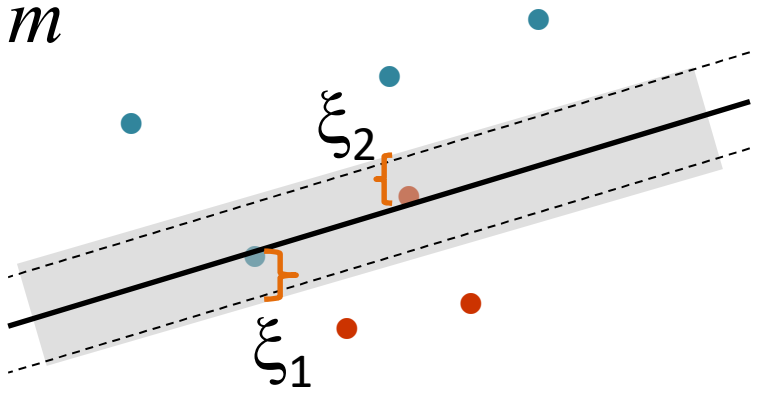
Forma primal de C-SVM

minimizar $f_0(x) = \frac{1}{2}\|\vec{w}\|^2 + \frac{C}{m}\sum_{i=1}^{m}\xi_i$

sujeto a

$$y_i\left(\langle \vec{w}, \vec{x} \rangle + b\right) \geq 1 - \xi_i, i = 1,\ldots,m$$

$$\xi_i \geq 0, \forall i = 1,\ldots,m$$



SVM Margen suave forma dual

maximizar $f_0(x) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_j \alpha_i y_j y_i k(x_i, x_j)$

sujeto a

$$\mathbf{0} \le \alpha_i \le \frac{C}{m}, i = 1, \ldots, m$$

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

C-Support Vector Classification
C-SVC/ C-SVM