

# Máquinas de Vectores de Soporte (SVM)

Guillermo Santamaría-Bonfil

# Agenda

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

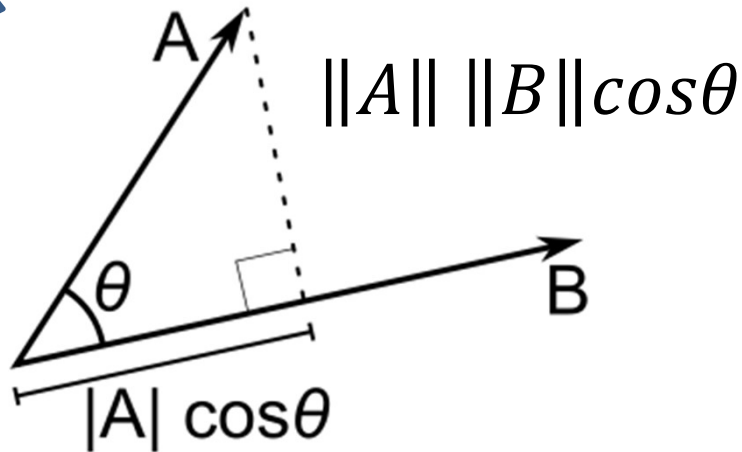


B. Schölkopf, A. Smola (2002). *Learning with Kernels: SVM, Regularization, Optimization, and Beyond*. MIT Press.

# Repaso de algunos conceptos geométricos

1

Producto Punto:  $\mathbf{A} \cdot \mathbf{B}$



3

Vector unitario

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

4

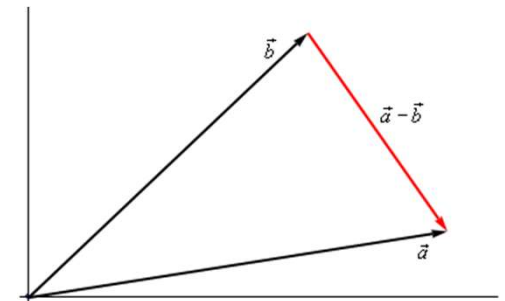
Distancia entre vectores

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

2

Magnitud de un vector  $\mathbf{x}$

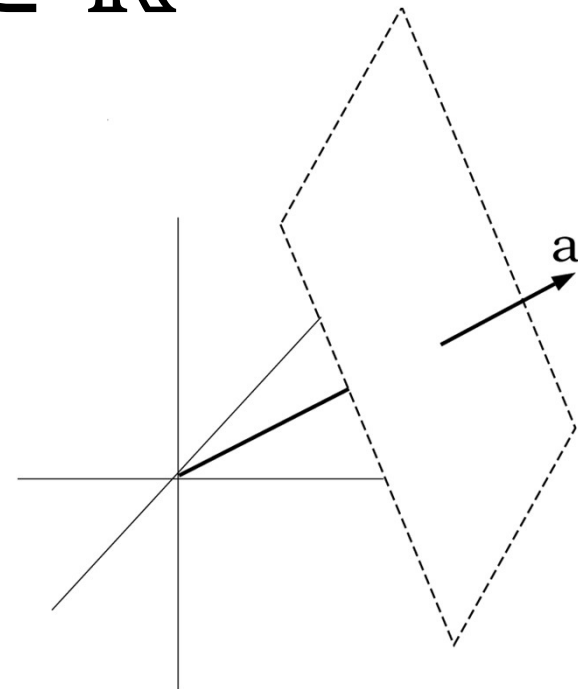
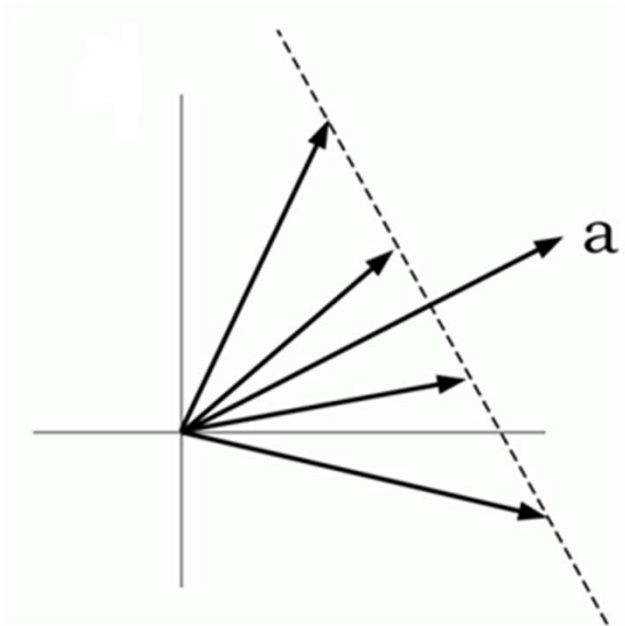
$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{x_1^2 + x_2^2 + \cdots + x_N^2}$$



# Definición geométrica de un Hiperplano

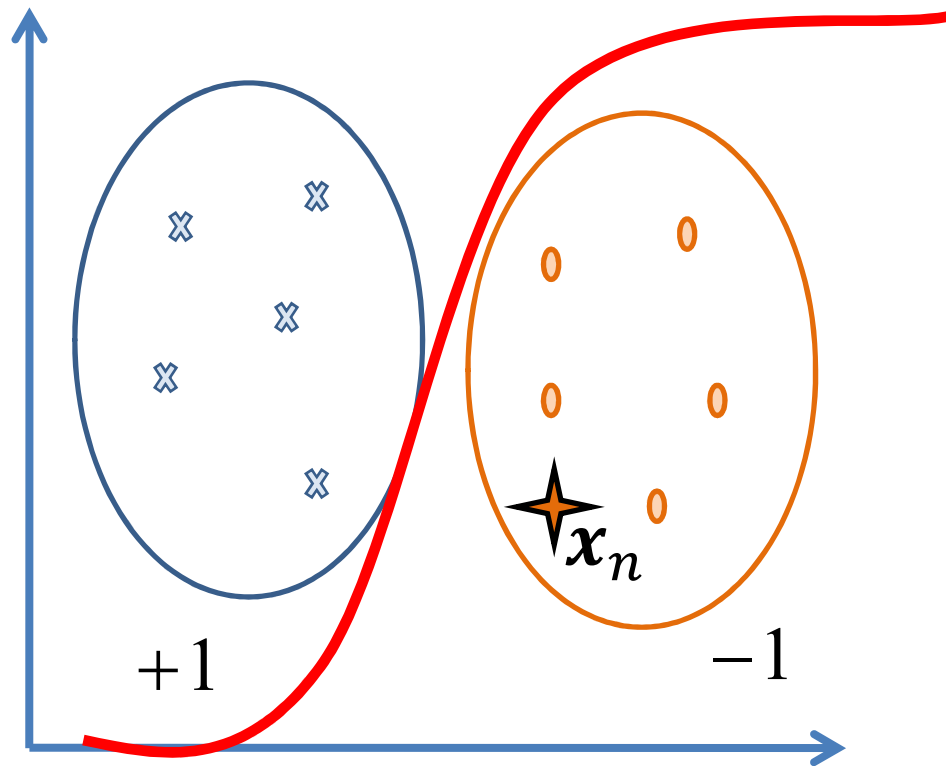
$$\{x \in H | \langle w, x \rangle + b = 0\}$$

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

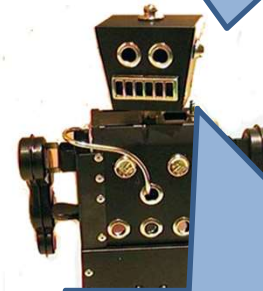


# Dado un problema de clasificación binaria ...

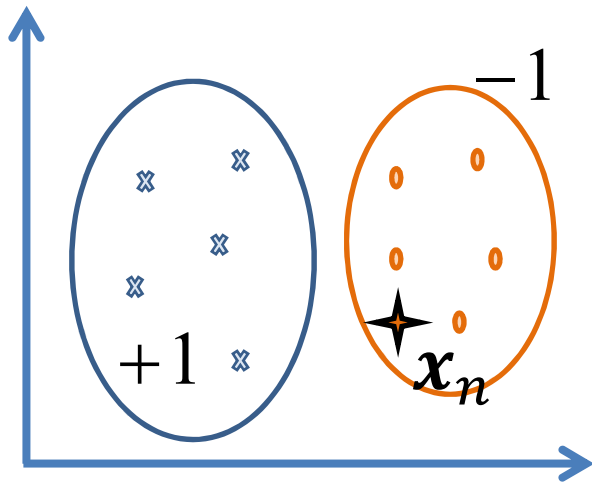
$$(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$$
$$x \in \mathbb{R}^N, y \in \{\pm 1\}$$



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

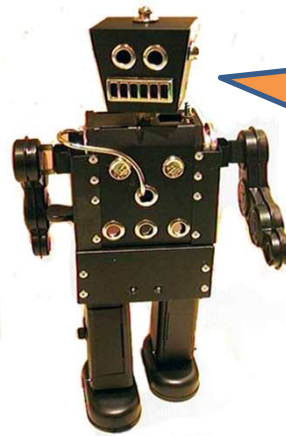


Encontremos una superficie de decisión!!



$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)$$

$$\mathbf{x} \in \mathbb{R}^N, y \in \{\pm 1\}$$



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

$$k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

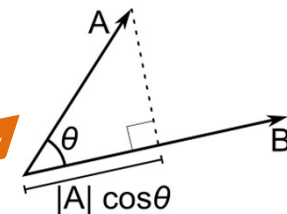
$$(\mathbf{x}_i, \mathbf{x}_j) \rightarrow k(\mathbf{x}_i, \mathbf{x}_j)$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_j, \mathbf{x}_i)$$

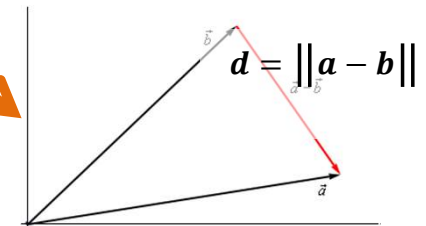
Función Kernel

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \sum_{s=1}^N x_s^i \cdot x_s^j$$

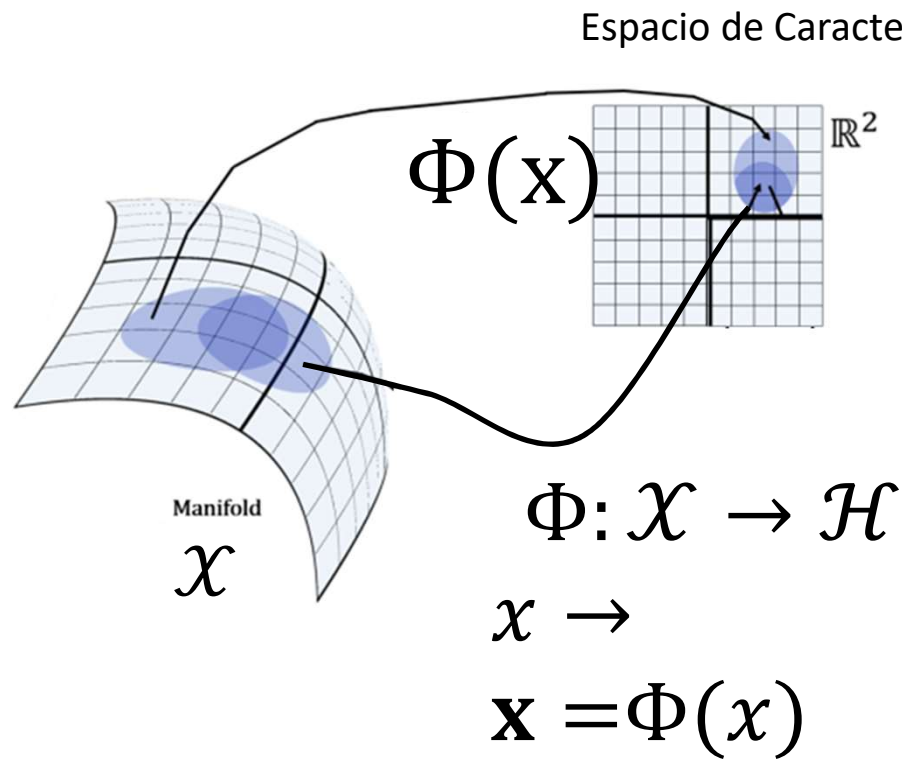
El producto punto es una función kernel



$$\|\mathbf{x}\|$$



# La función de similitud solo funciona en un espacio Hilbert



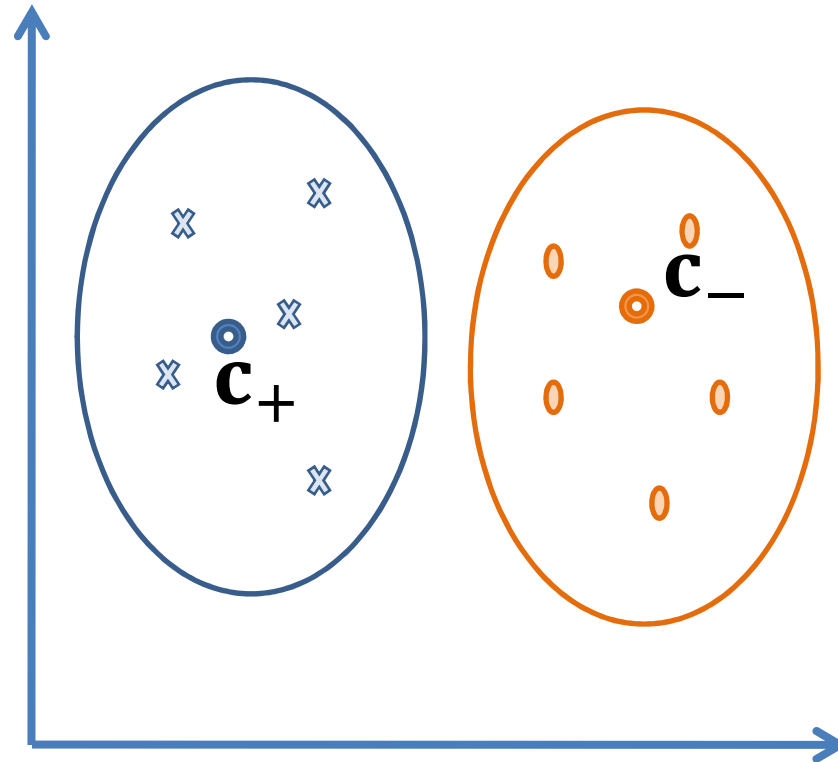
1. Medida de Similitud en  $\mathcal{H}$   
 $k(x_i, x_j) = \langle x_i, x_j \rangle = \langle \Phi(x_i), \Phi(x_j) \rangle$
2. Análisis Geométrico usando algebra lineal y geometría analítica.
3. El producto punto, es un caso especial de  $\Phi$ . Por lo que es posible realizar un análisis no-lineal.

# Un algoritmo simple para clasificación binaria (2)

**Paso 1:** Calcular las medias de las dos clases

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

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



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



# Un algoritmo simple para clasificación binaria (3)

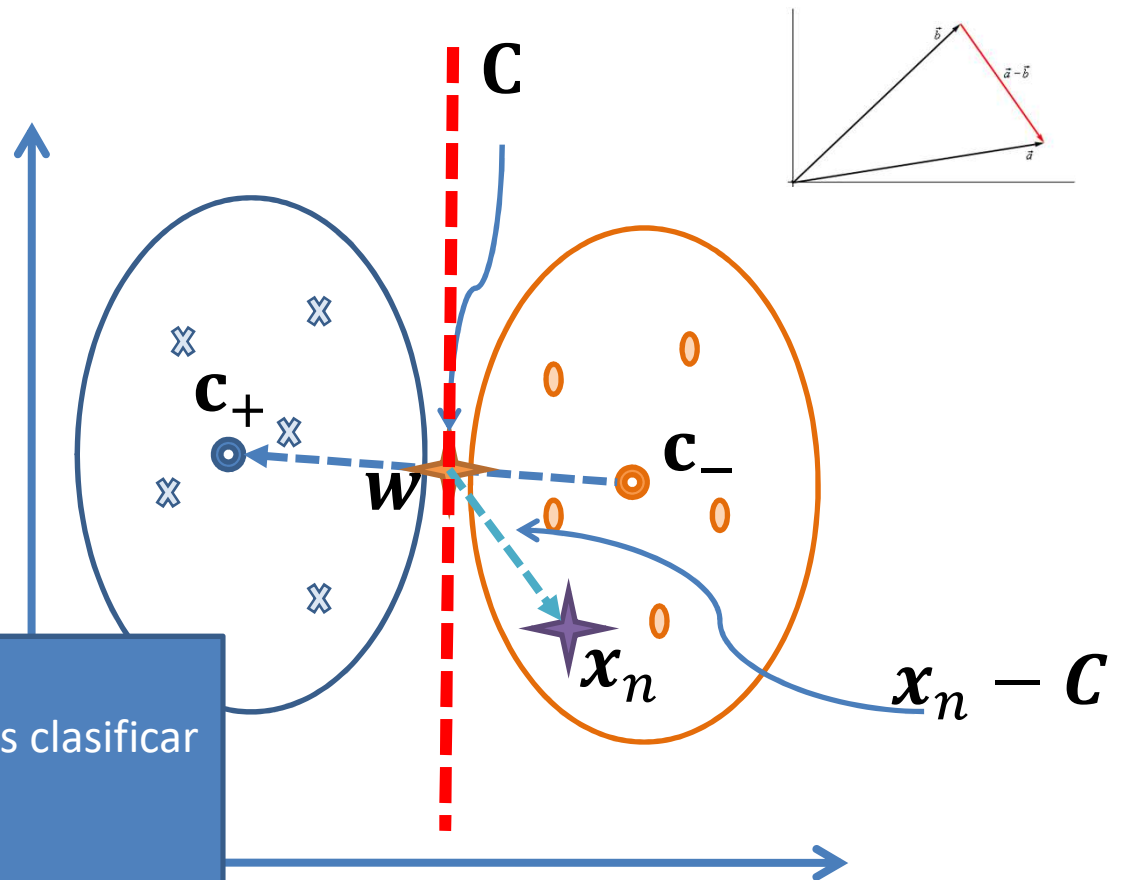
**Paso 2:** Construimos una superficie de decisión

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

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

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

Al proyectar  $(\mathbf{x}_n - \mathbf{c})$  en  $\mathbf{w}$  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  $\mathbf{x}_n - \mathbf{c}$  en  $\mathbf{w}$

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

¿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  $\mathbf{w}$

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

# 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  $\mathbf{w}$

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

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

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

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

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

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

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

# Límite de Decisión

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

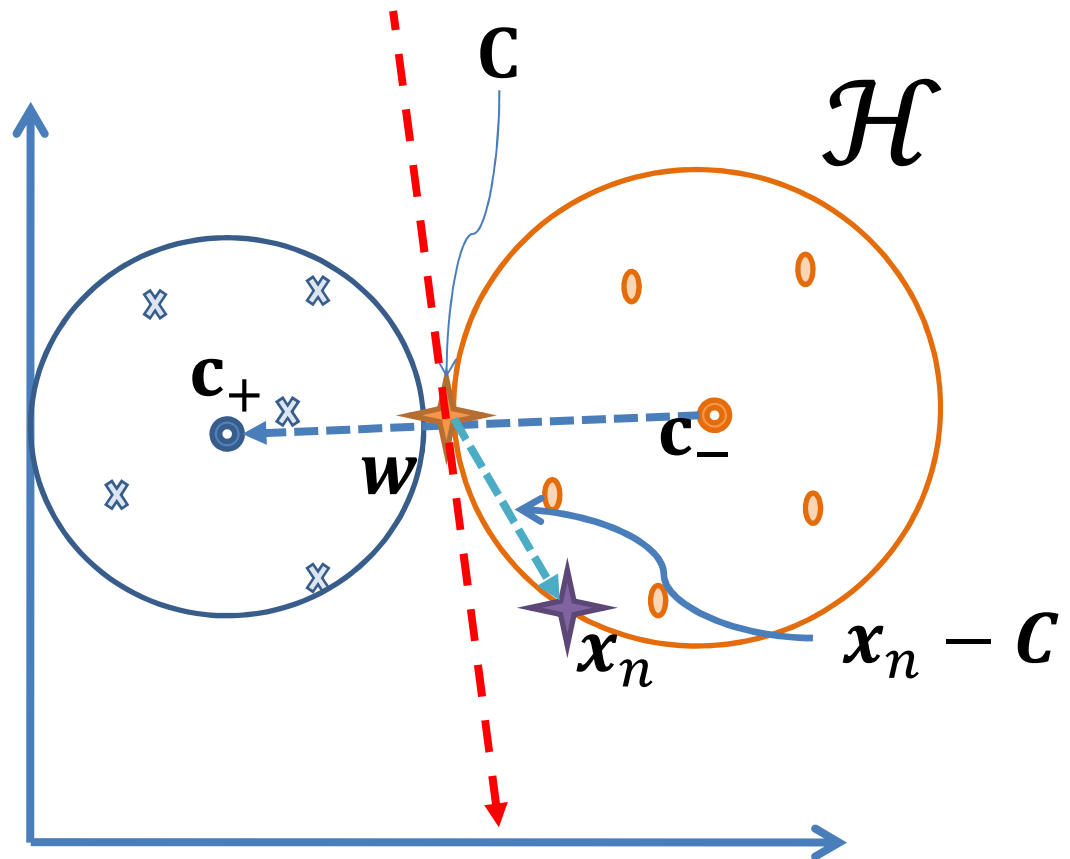
Esta función también define un Hiperplano:

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

Un nuevo vector  $\mathbf{x}_n$  es clasificado como +1 o -1 si:

El vector  $\mathbf{x}_n - \mathbf{C}$  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,  $\mathbf{c}$  y  $\mathbf{w}$  en términos de  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$  y  $k$ .

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

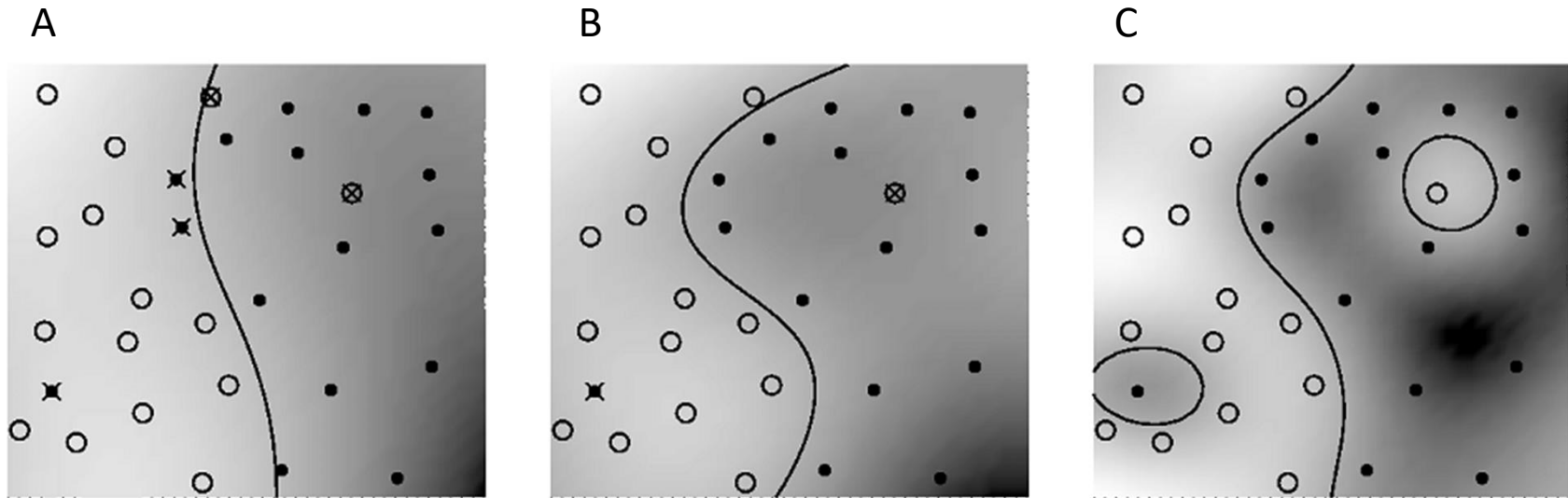
$$y_n = \text{sgn} \left( \frac{1}{m_+} \sum_{\{i|y_i=+1\}} \langle \mathbf{x}_n, \mathbf{x}_i \rangle - \frac{1}{m_-} \sum_{\{i|y_i=-1\}} \langle \mathbf{x}_n, \mathbf{x}_i \rangle + b \right)$$

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

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

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

# Tres clasificadores en 2-D



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

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

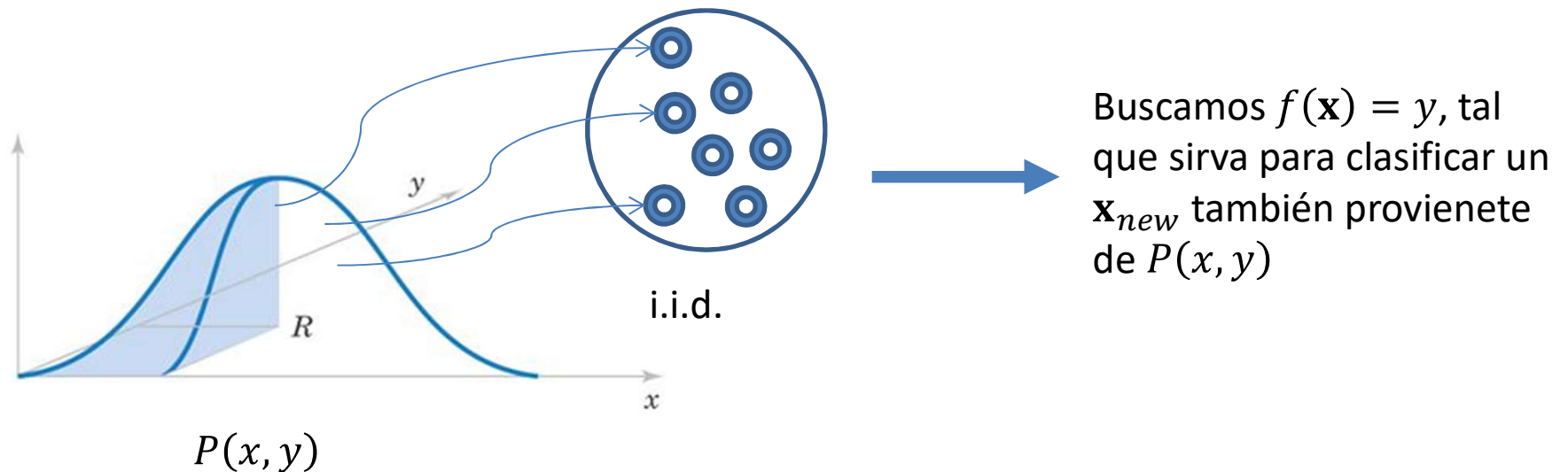
Pregunta: ¿Cuál clasificador es mejor con los datos de entrenamiento?

¿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  $P(x, y)$



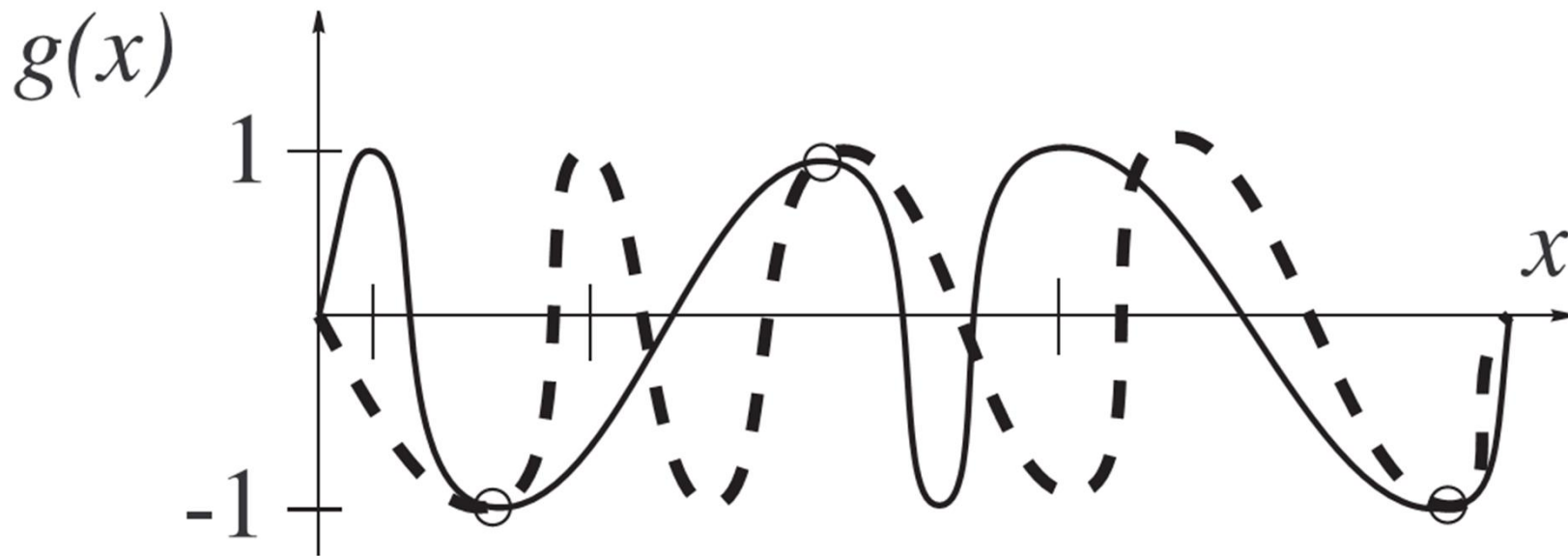
La precisión de clasificación de  $f$  es medida con la **función de pérdida** (i.e. riesgo, error, etc.) *cero-uno*:

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

**¿Qué valores toma esta Función de pérdida?**



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  $n \rightarrow \infty$  los parámetros  $\theta$  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  $f$

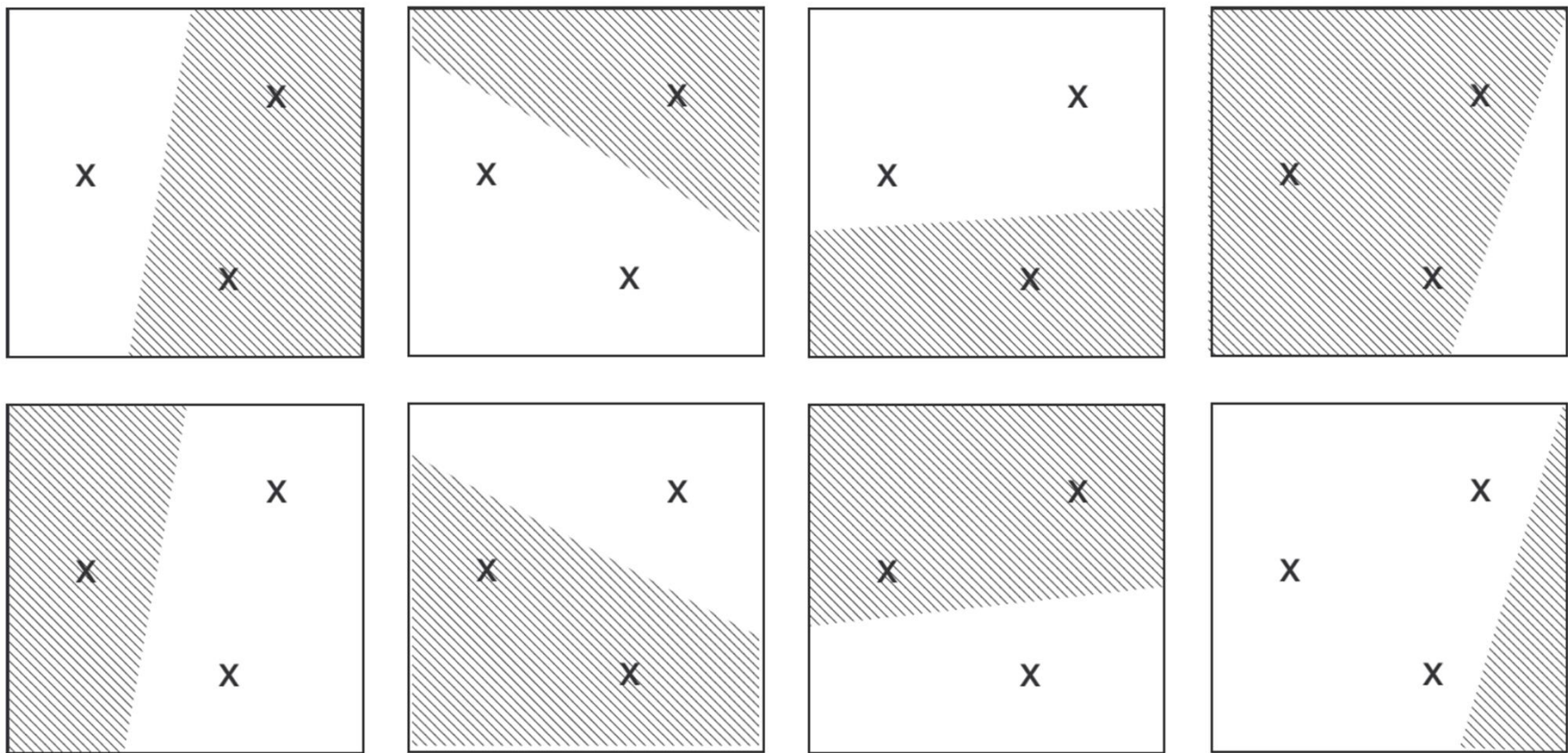


Minimización del Riesgo Estructural
--

# La dimensión VC como medida de Capacidad de $f$

- Para cada  $f_i \in F$  los datos son separados de forma específica  
 $\therefore$  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?

# Apresiasi los resultados de la Teoría VC



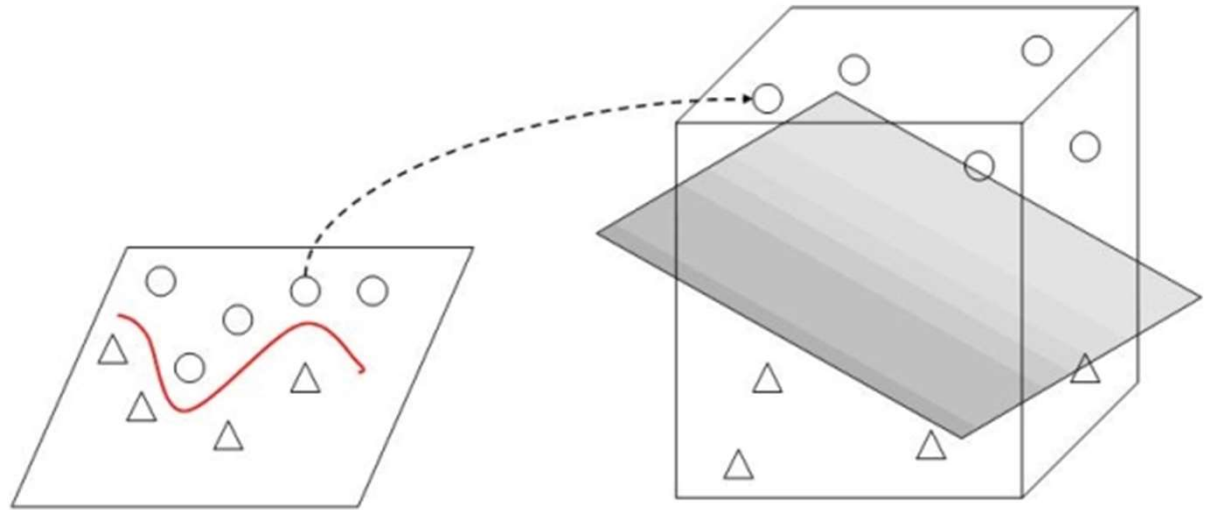
- Si  $h < m$  es la dimensión VC de  $F$ 
  - Entonces, para  $f_i \in F | \forall i$
  - Independientemente de  $P(x, y)$
  - Con una probabilidad de al menos  $1 - \delta$  sobre la muestra de entrenamiento

Entonces:

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

Tal que  $\phi$  (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

$$\text{Maximizar } \min_{\mathbf{x} \in \mathcal{H}, b \in \mathbb{R}} \|\mathbf{x} - \mathbf{x}_i\|$$

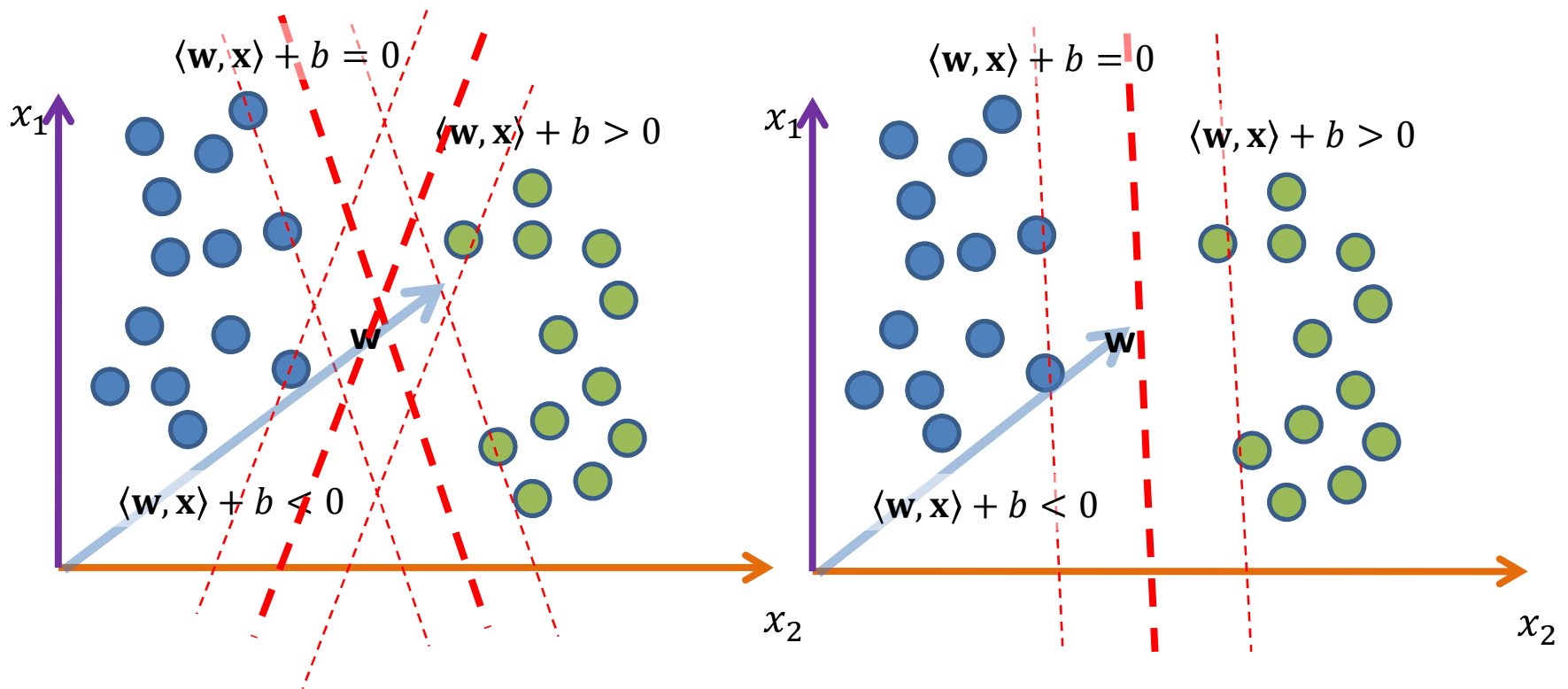
Sujeto a

$$\mathbf{x} \in \mathcal{H}$$

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0, \forall i = 1, \dots, m.$$

# Retrato Generalizado (2)

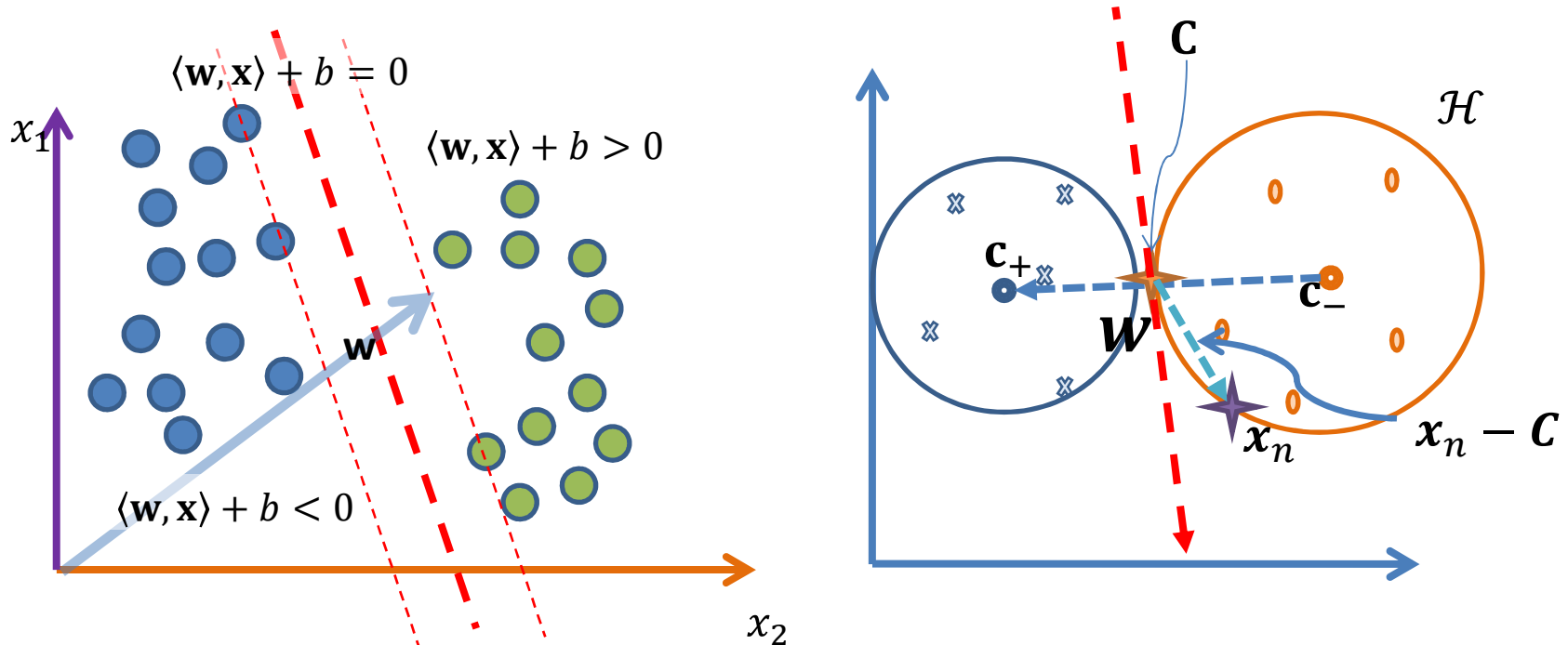
2. La capacidad de la clase F de hiperplanos, decrece conforme el margen crece.



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

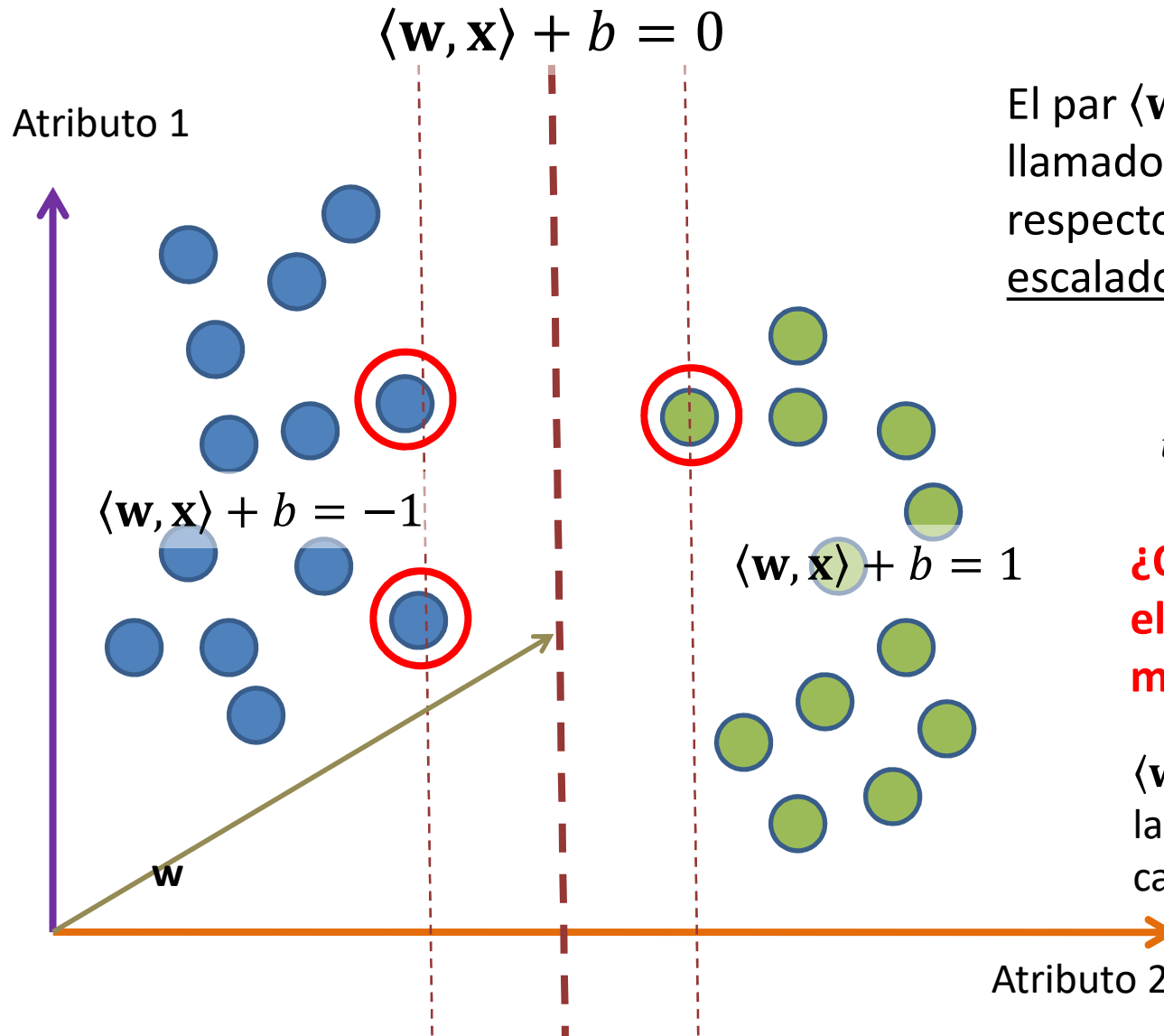
$$f(\mathbf{x}) = \text{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$$

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



La diferencia es como estimamos  $\mathbf{w}$ .

# Definición: Hiperplano canónico



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

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

**¿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.

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

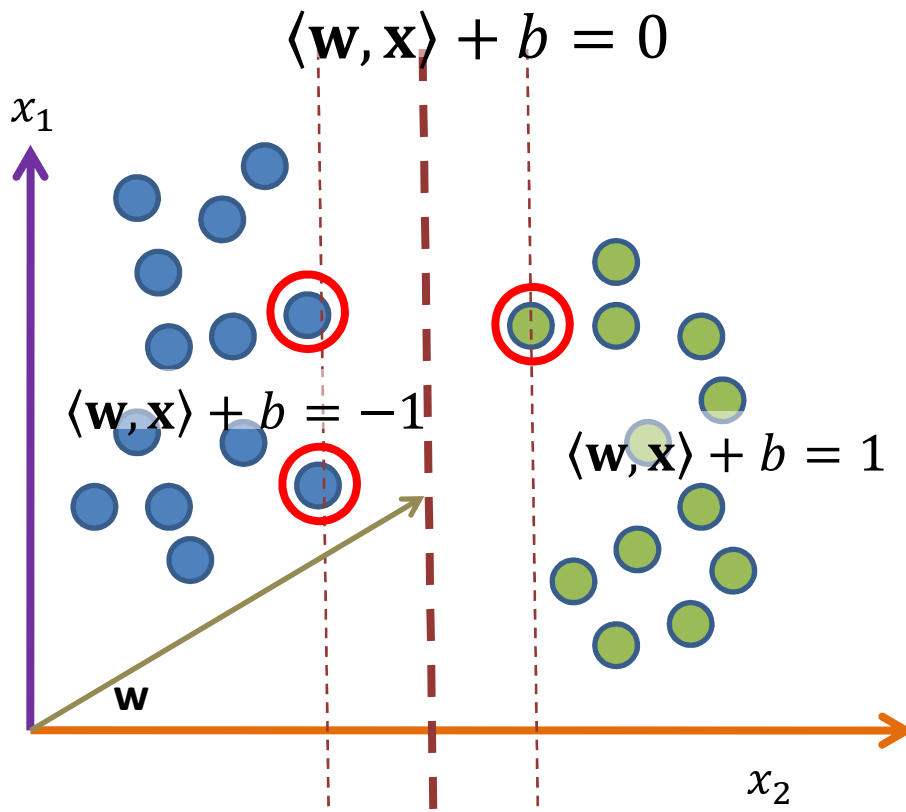
- Estas 2 tuplas forman dos superficies de decisión:

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

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

- 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  $\{\mathbf{x} \in \mathcal{H} \mid \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$   
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

$$P_{(\mathbf{w}, b)} = \min_{i=1, \dots, m} \rho_{(\mathbf{w}, b)}(\mathbf{x}_i, y_i)$$

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

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

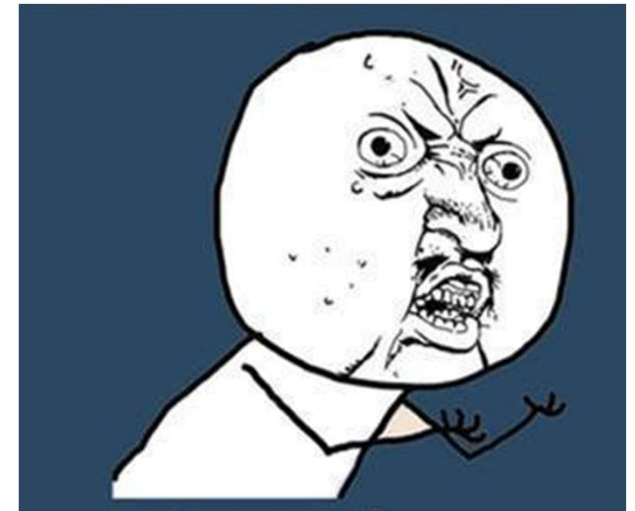
# ¿Qué relación existe entre minimizar $\mathbf{w}$ y una buena clasificación?

1. Para un punto  $(\mathbf{x}_i, y_i)$  bien clasificado,  
 $\rho$  es la distancia al hiperplano.  
 $\rho = 0$  si  $\mathbf{x}$  esta sobre el hiperplano.
2. Trabajamos con el hiperplano normalizado  
 $(\hat{\mathbf{w}}, \hat{b}) = \left( \frac{\mathbf{w}}{\|\mathbf{w}\|}, \frac{b}{\|\mathbf{w}\|} \right)$  con  $\frac{\mathbf{w}}{\|\mathbf{w}\|} = 1$  y  
solo debemos calcular  $y(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b})$ .

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

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

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



# Formulación primal de SVM

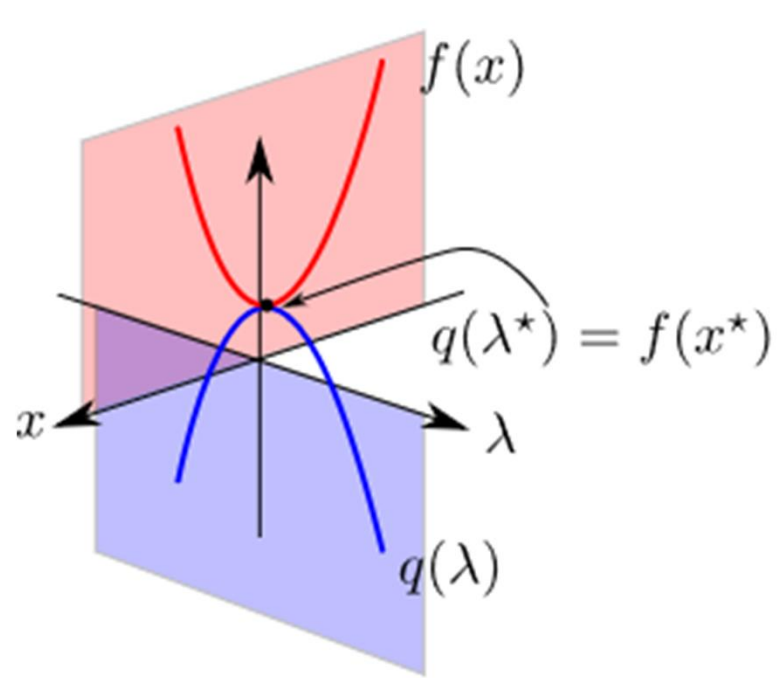
$$\text{Min } \tau(w) = \frac{1}{2} \|w\|^2$$

Sujeto a

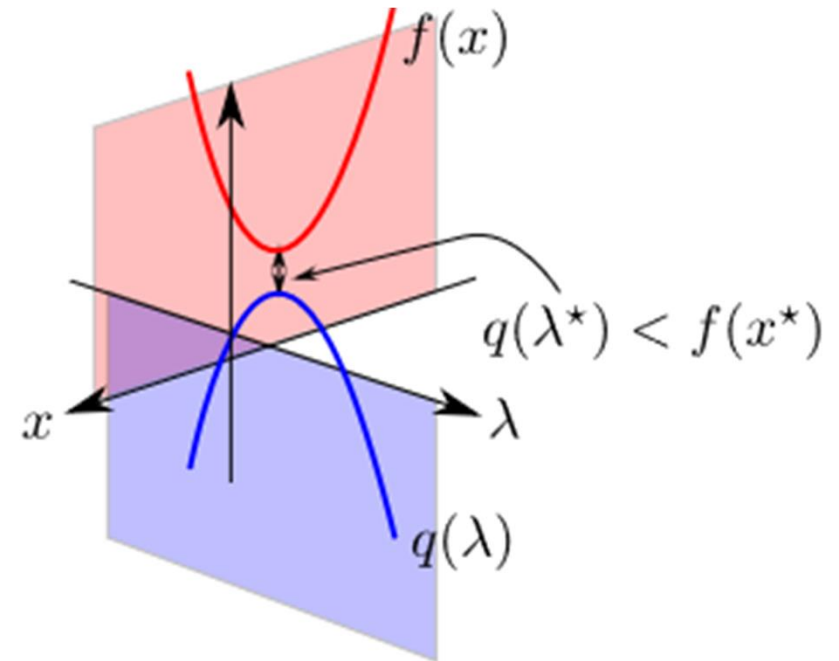
$$y_i (\langle \bar{x}_i, \bar{w} \rangle + b) \geq 1, \forall i = 1, \dots, m$$



# Optimización, Primal y Dualidad.



strong duality



weak duality

# Hacia la formulación dual de SVM

- $\tau(w)$  es la función objetivo
- $y_i(\langle w, x_i \rangle + b) \geq 1$  son las restricción de desigualdad.
- Utilizando los multiplicadores de Lagrange  $\alpha_i \geq 0$  podemos construir el lagrangeano:

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

# 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*  $\alpha_i$ 
  - Para obtener la solución debemos derivar  $L$  con respecto a las variables primales:

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

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

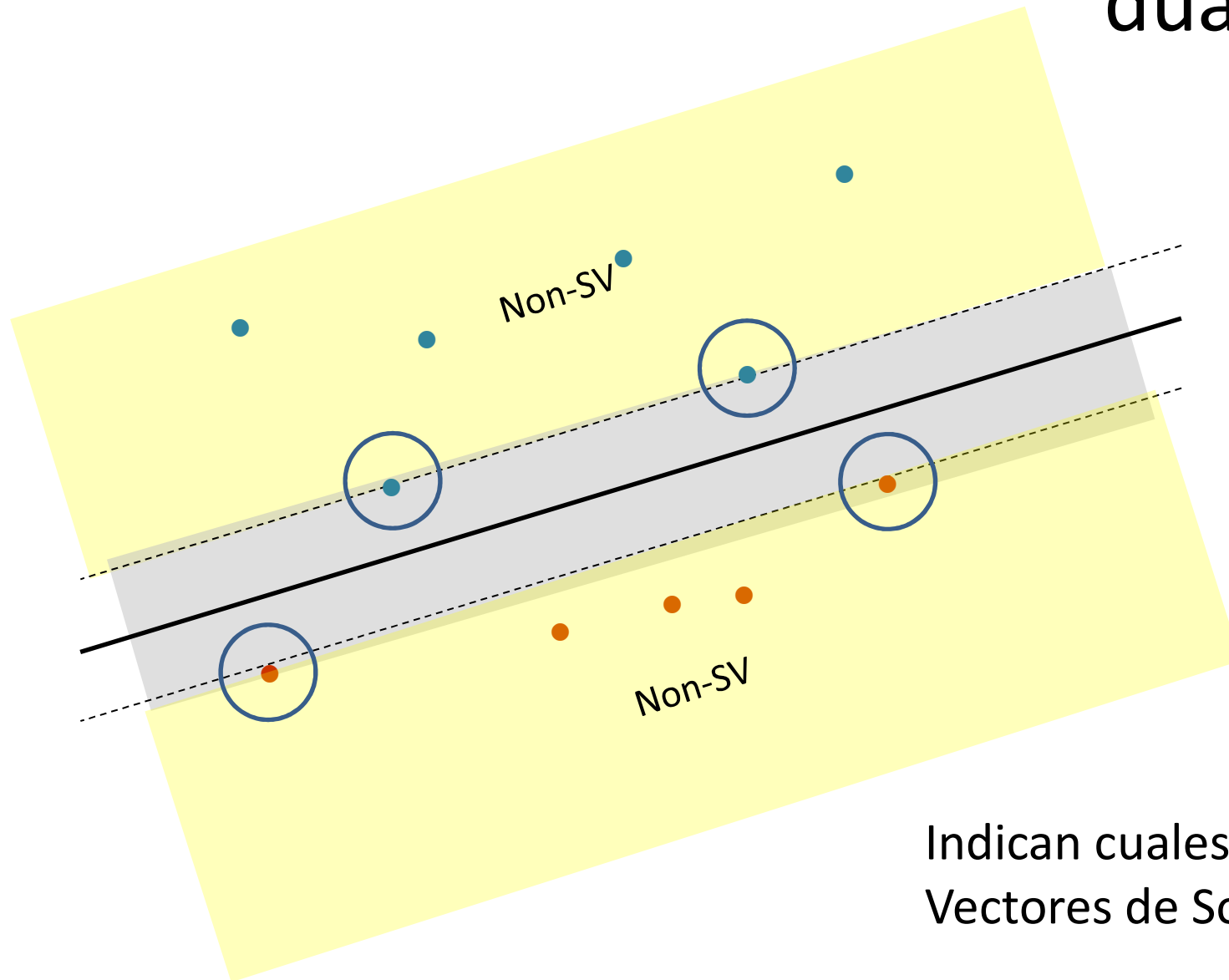
# Condiciones de Optimalidad KKT

$$\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = 0 \quad \rightarrow \text{Expresados como una combinación lineal del conjunto de entrenamiento}$$

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

$$KKT \text{ cond} : \alpha_i [y_i (\mathbf{w} \mathbf{x}_i + b) - 1] = 0 \quad \rightarrow \text{Únicamente los SVs tendrán un } \alpha_i \text{ mayor a cero}$$

# ¿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.:

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

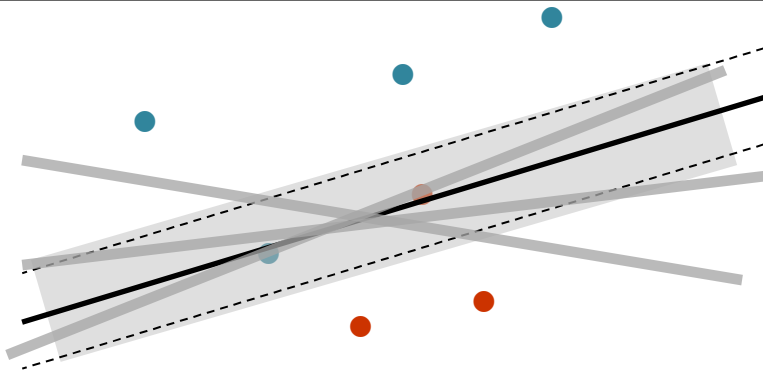
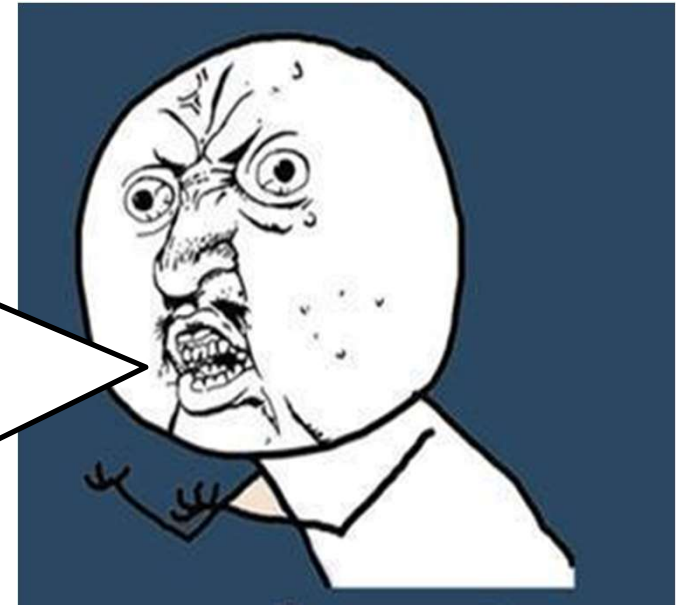
Sujeto a

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

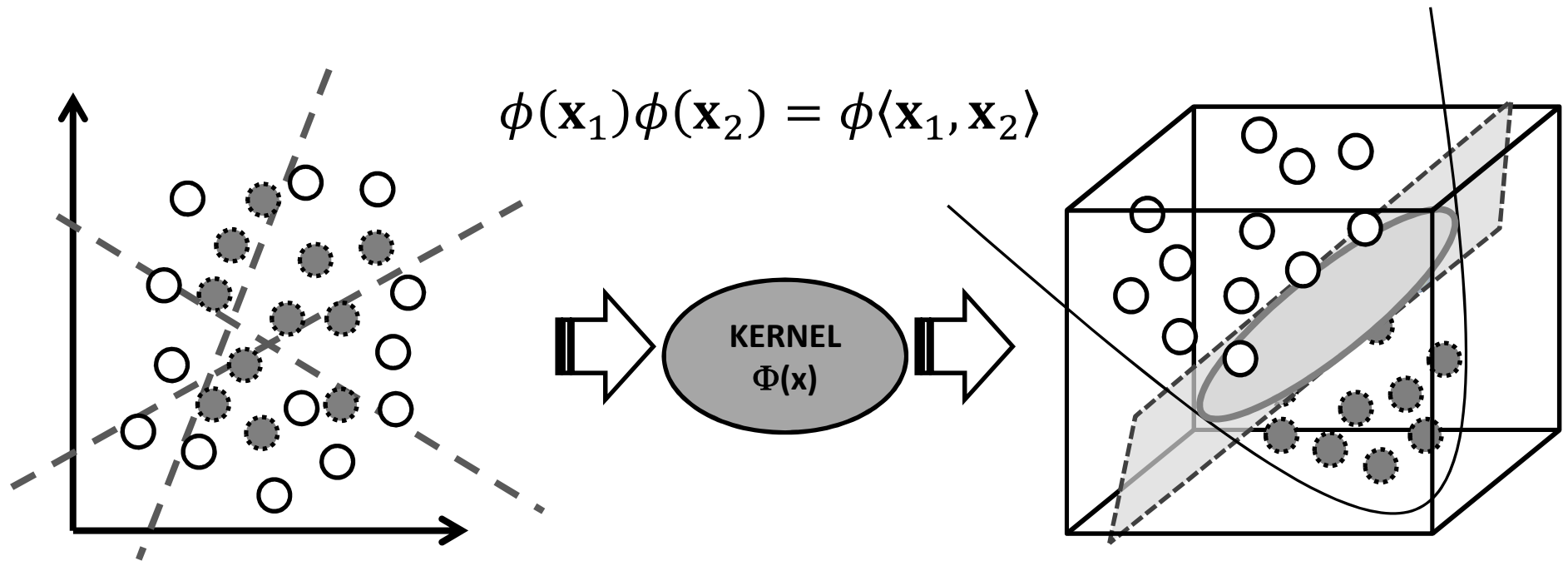
$$\sum_{i=1}^m \alpha_i y_i = 0$$

¿Qué sucede si no existe un  
hiperplano separador?

¿Cómo podemos manejar el  
ruido y datos etiquetados  
erróneamente?



# Separación No-Lineal: Truco del Kernel



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

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

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



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

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

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

- Caso no-lineal:

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

sujeto a

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

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

$\alpha$ : multiplicadores de Lagrange

$K(x_i, x_j)$ : truco del kernel

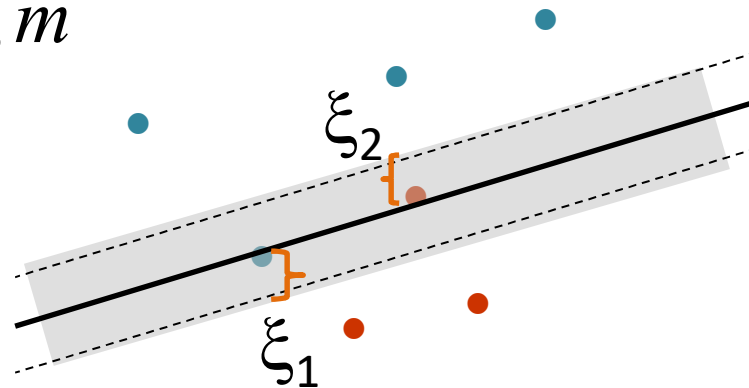
# Forma primal de C-SVM

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

sujeito a

$$y_i (\langle \bar{w}, \bar{x} \rangle + b) \geq 1 - \xi_i, i = 1, \dots, m$$

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



# SVM Margen suave forma dual

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

sueto a

$$0 \leq \alpha_i \leq \frac{C}{m} \quad i = 1, \dots, m$$

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

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