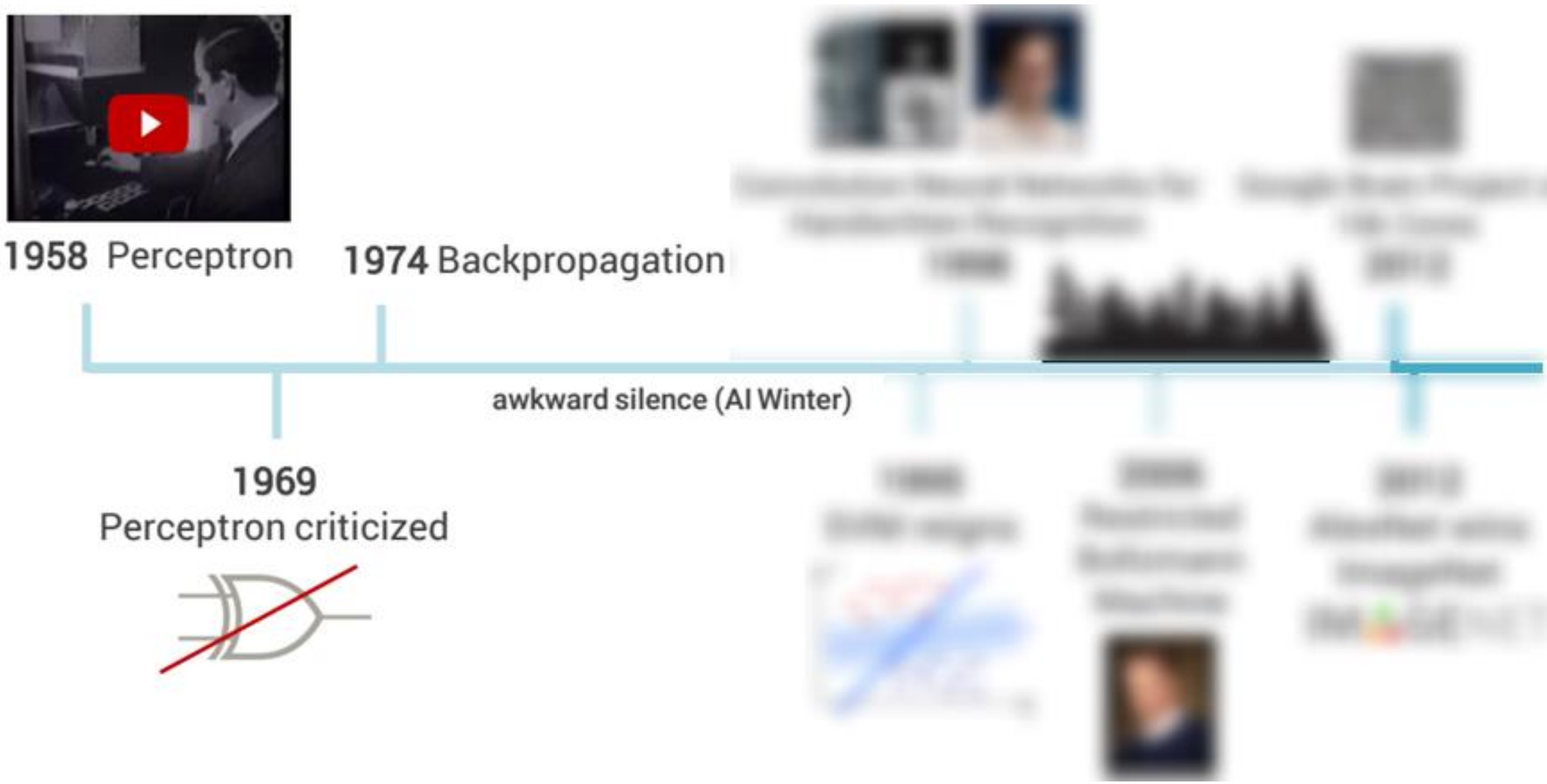
Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2613 – Inteligencia Artificial

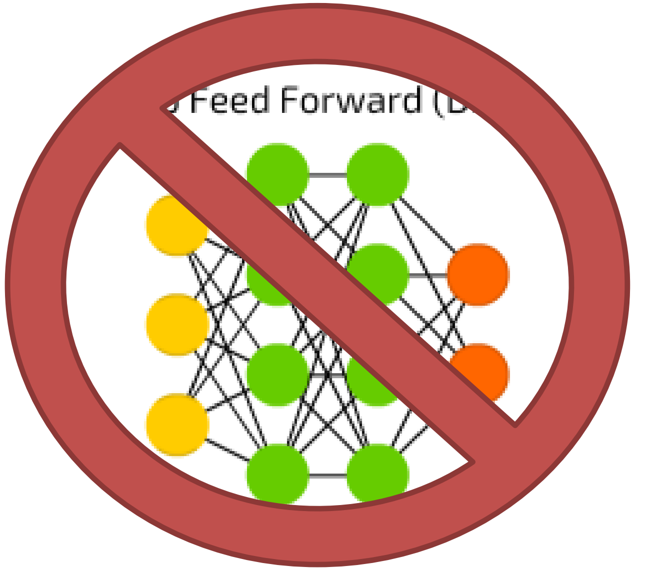
Support Vector Machines (SVM)

Profesor: Hans Löbel

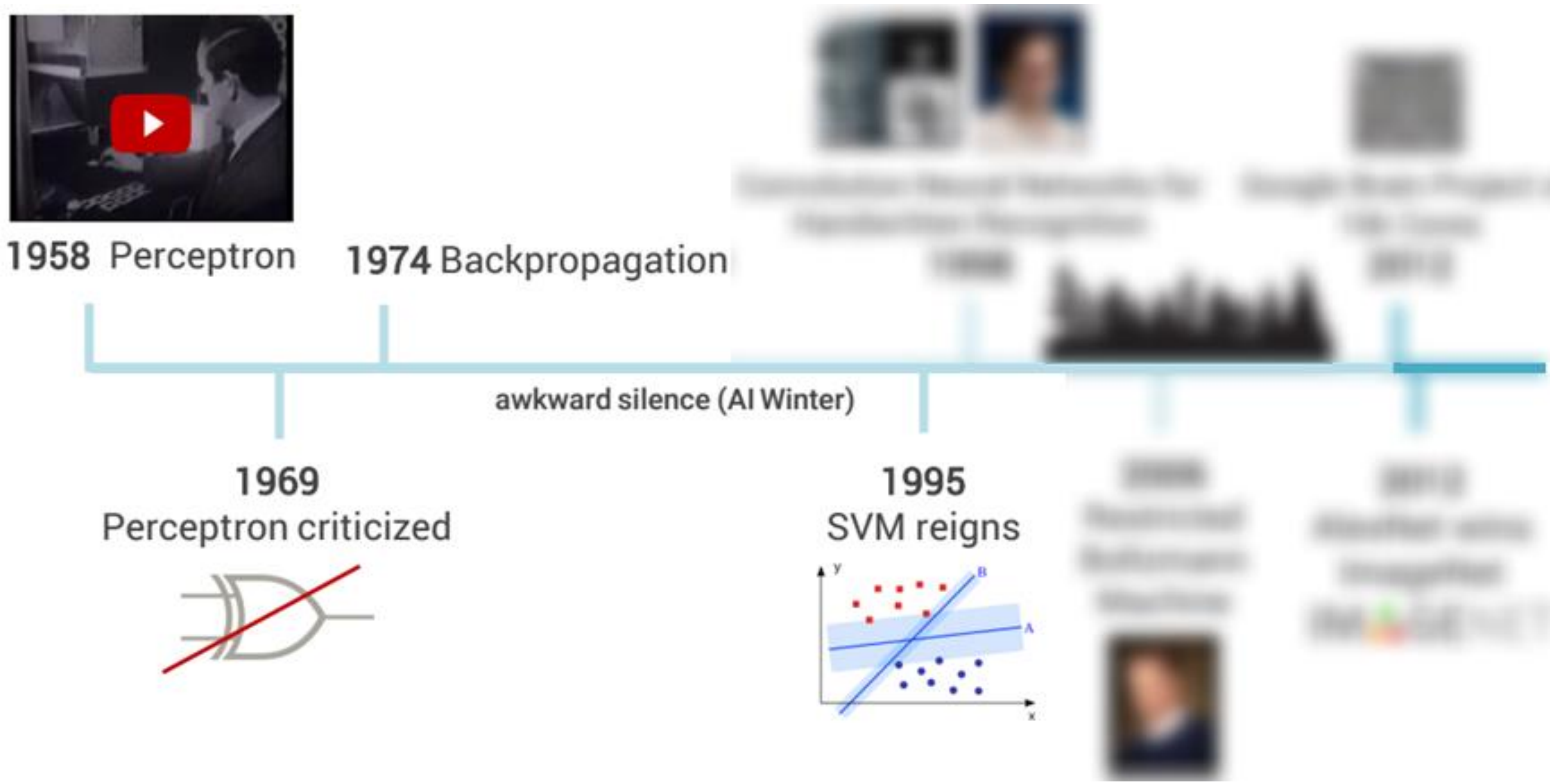

1958 Perceptron
1974 Backpropagation
awkward silence (AI Winter)
1969
Perceptron criticized

Dificultades de redes neuronales hicieron que el foco se centrara en otras técnicas

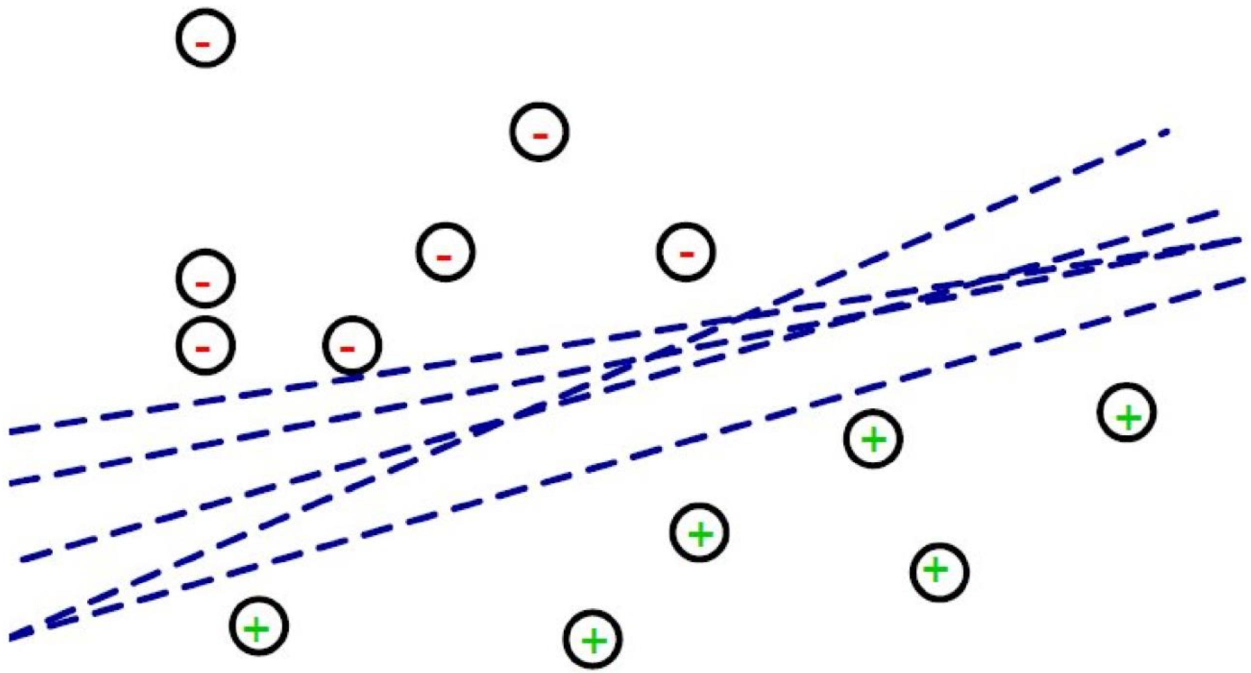
- Redes presentan problema no convexo y mínimos locales.
- Rendimiento no era sustancialmente superior al resto de las técnicas.
- Interpretación de los modelos es altamente compleja.



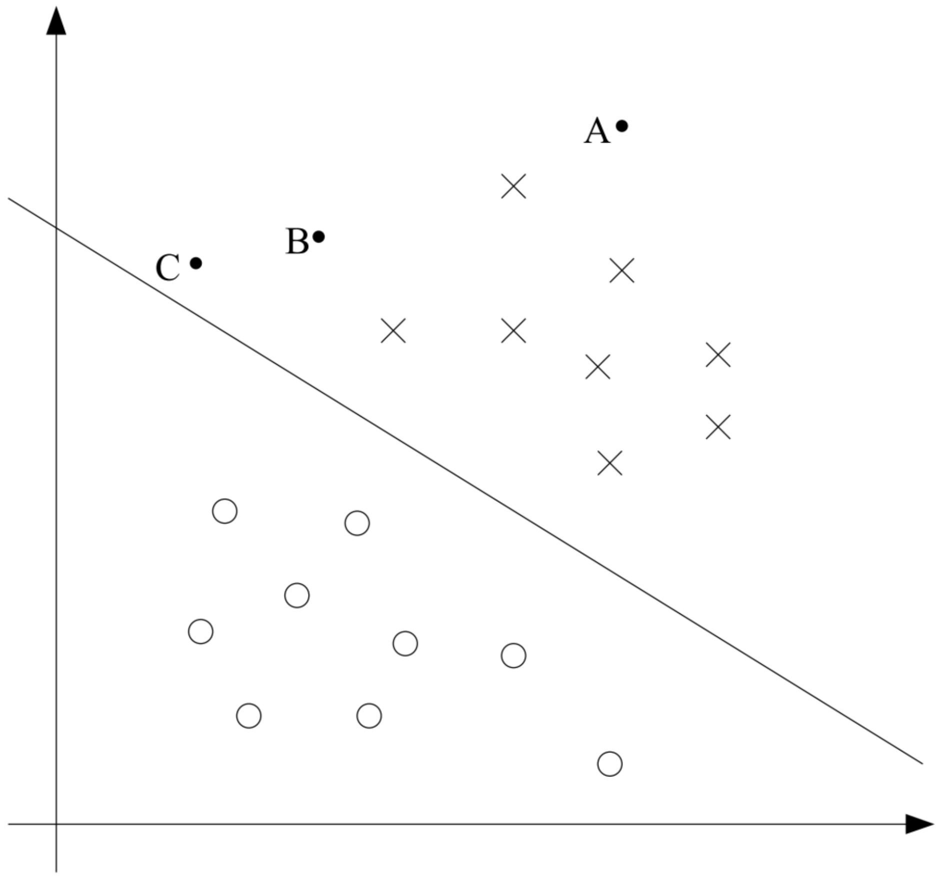
Dadas las restricciones de la época (≈1990), los modelos lineales seguían siendo atractivos, pero requerían mejor rendimiento.


1958 Perceptron
1974 Backpropagation
awkward silence (AI Winter)
1969
Perceptron criticized
1995
SVM reigns

¿Cuál es el hiperplano que mejor separa dos categorías?

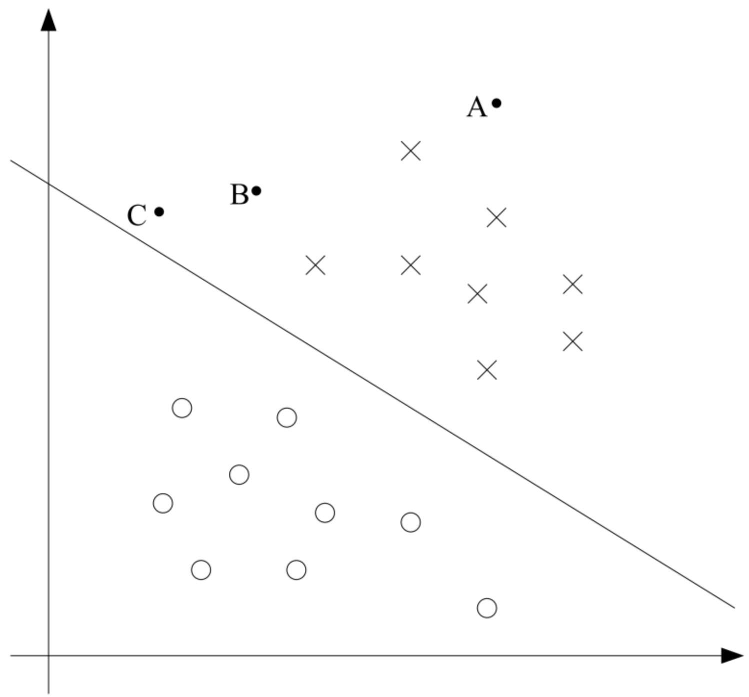


¿Cuál es el hiperplano que mejor separa dos categorías?



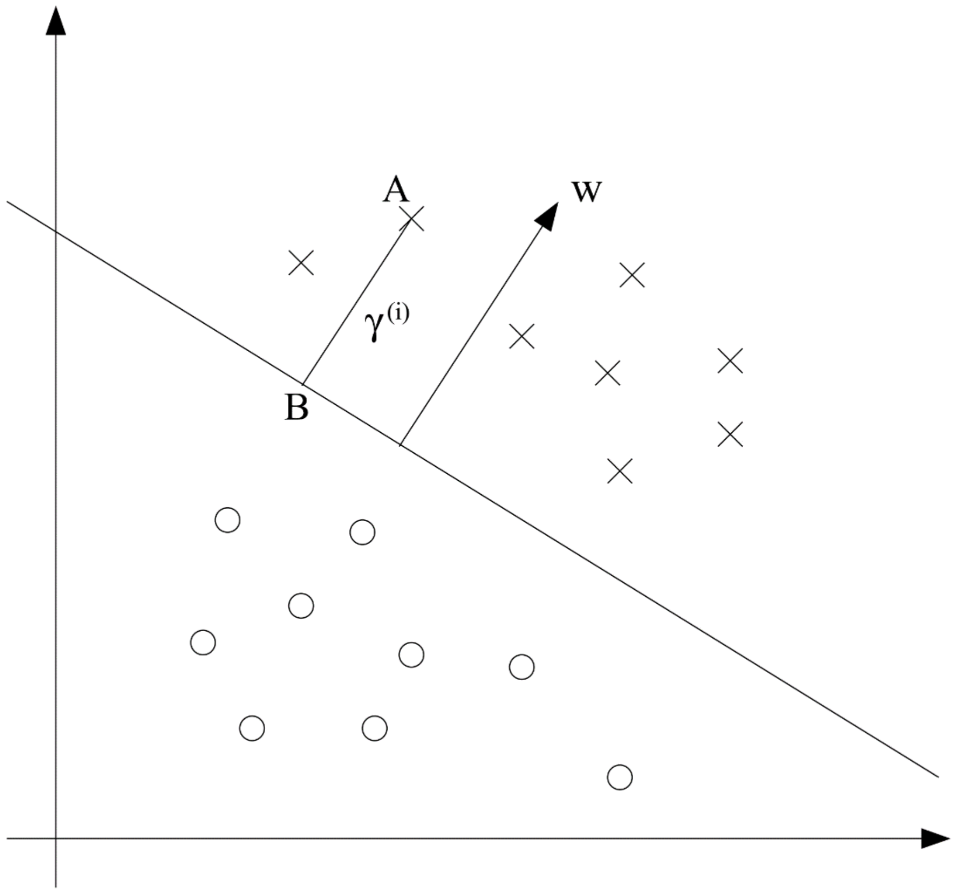
$$h_{w,b}(x) = g(w^T x + b)$$

Margen funcional da respuesta parcial, ya que tiene problemas con escalamiento



$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$$

Distancia de un punto al hiperplano (margen geométrico) es la clave de los SVM



Distancia de un punto al hiperplano (margen geométrico) es la clave de los SVM

$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{w}{||w||} \right) + b = 0$$

$$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{||w||} = \left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||}$$

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$$

Problema de aprendizaje relaciona ambos márgenes

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \ldots, m \\ & ||w|| = 1. \end{aligned}$$

$$\begin{aligned} \max_{\hat{\gamma},w,b} \quad & \frac{\hat{\gamma}}{||w||} \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, m \end{aligned}$$

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m \end{aligned}$$

Veamos como podemos resolver este problema

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m \end{aligned}$$

$$g_i(w) = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0$$

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

El problema dual, permite una interpretación más clara de como funciona un SVM

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial}{\partial b} \mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

El problema dual, permite una interpretación más clara de como funciona un SVM

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

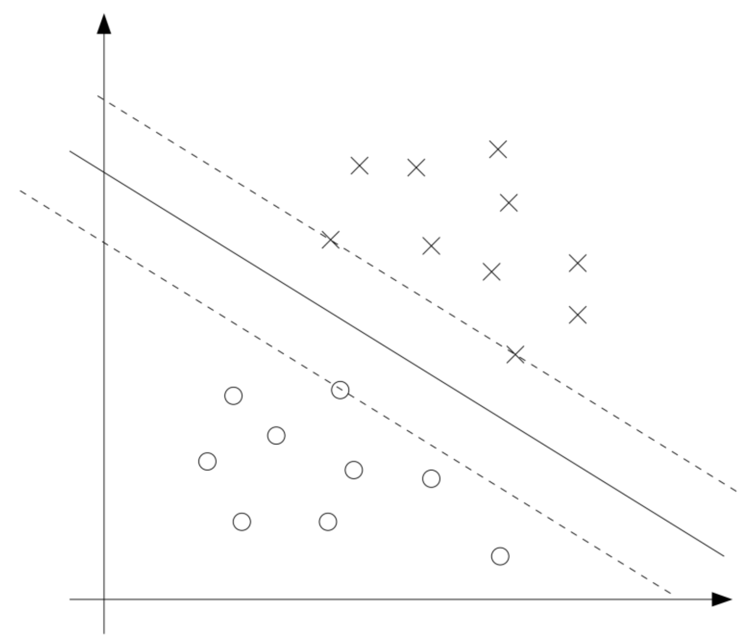
$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$

El problema dual, permite una interpretación más clara de como funciona un SVM

$$\begin{aligned} \max_\alpha \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0, \end{aligned}$$

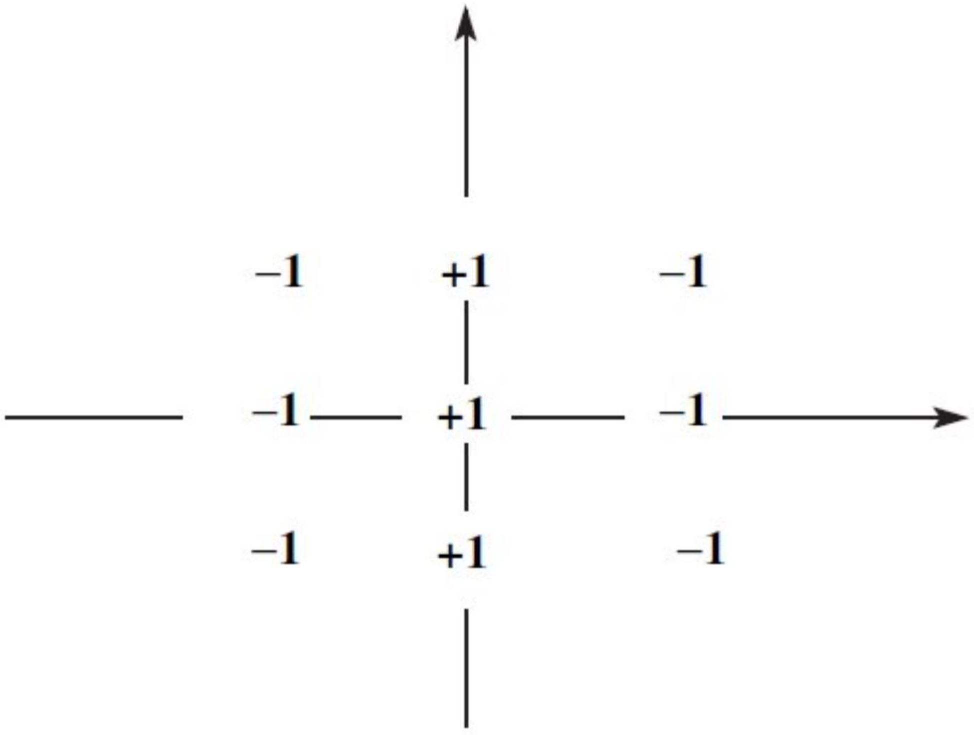
$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$



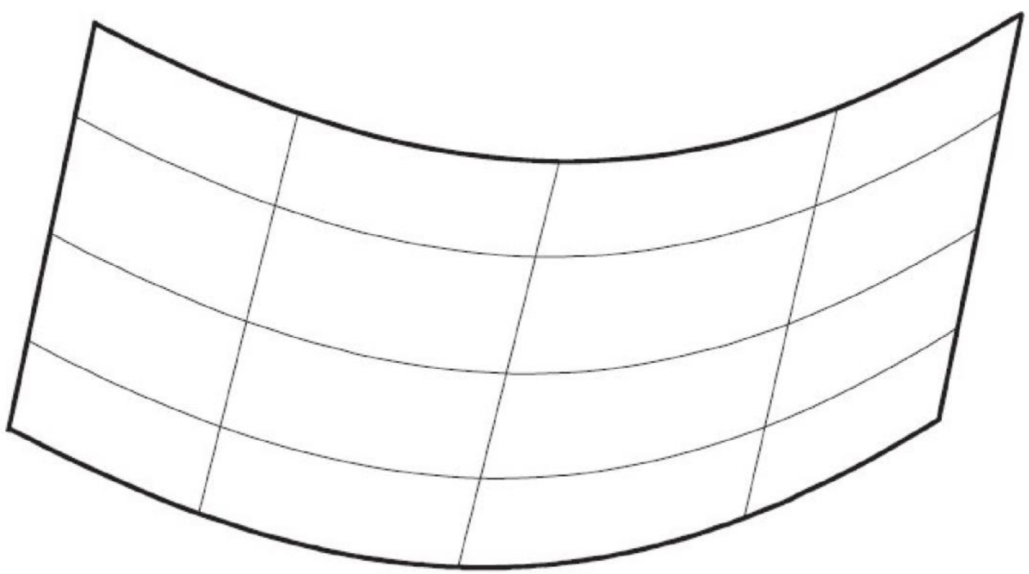
El problema dual, permite una interpretación más clara de como funciona un SVM

$$
\begin{aligned}
w^T x + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T x + b \\
&= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.
\end{aligned}
$$

Súper lindo, pero sigue siendo un clasificador lineal

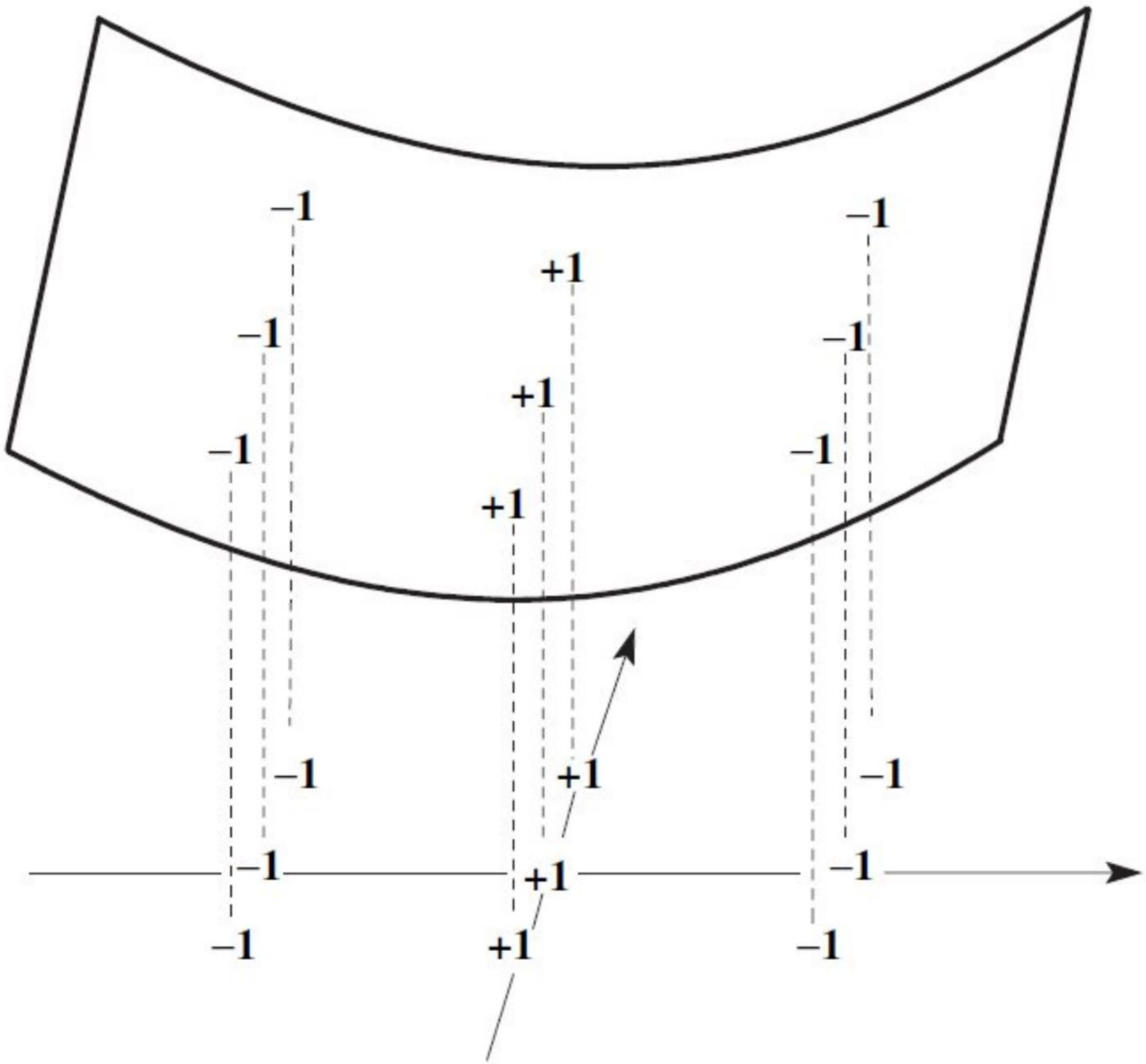


Una solución es generar un cambio de variables (transformación de espacio de características)

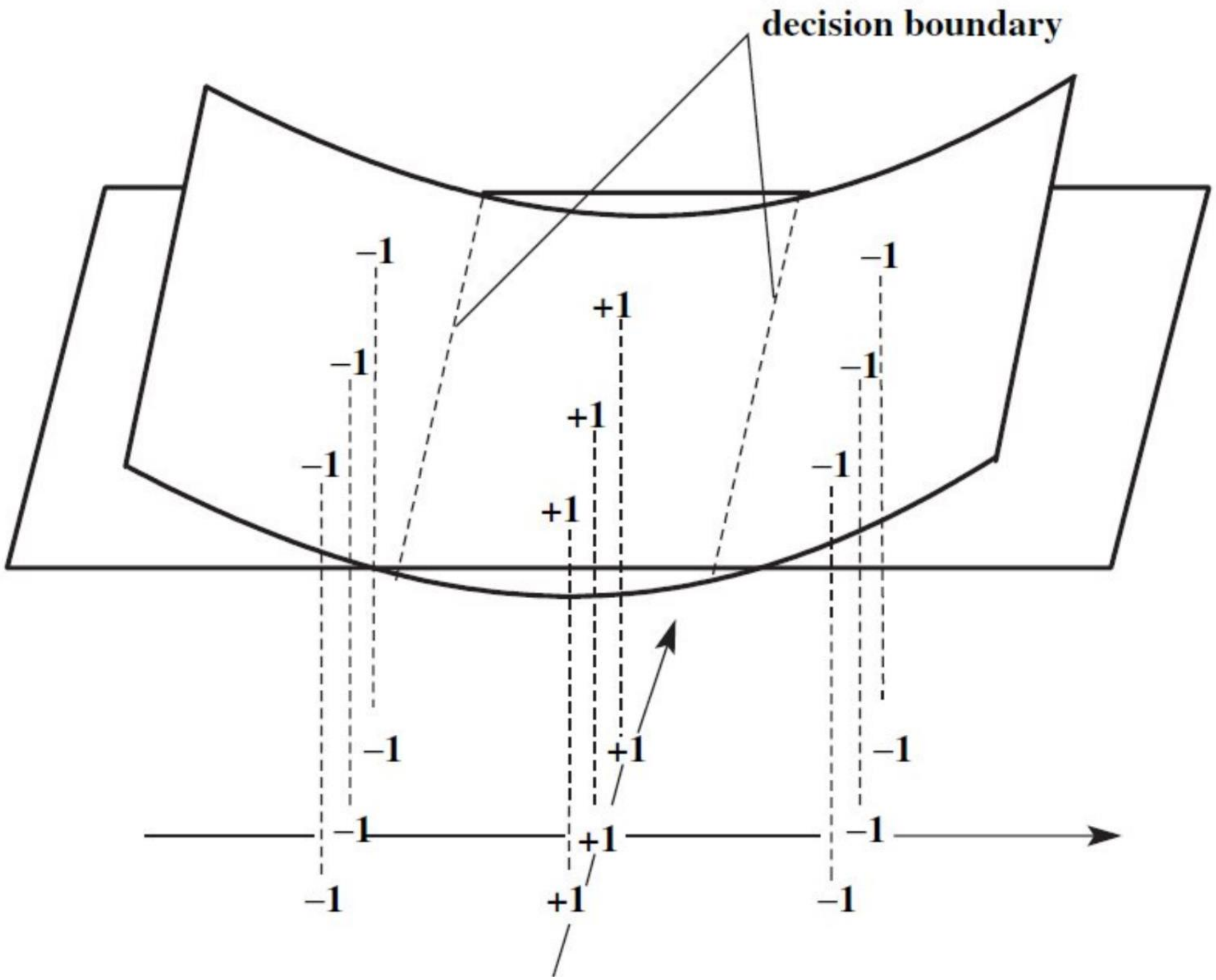


Espacio de características $f(x_1, x_2) = x_1^2$

Una solución es generar un cambio de variables (transformación de espacio de características)



No es muy distinto a regresión lineal con polinomios de mayor grado



Afortunadamente, existe un teorema que relaciona estas *features* con *kernels*

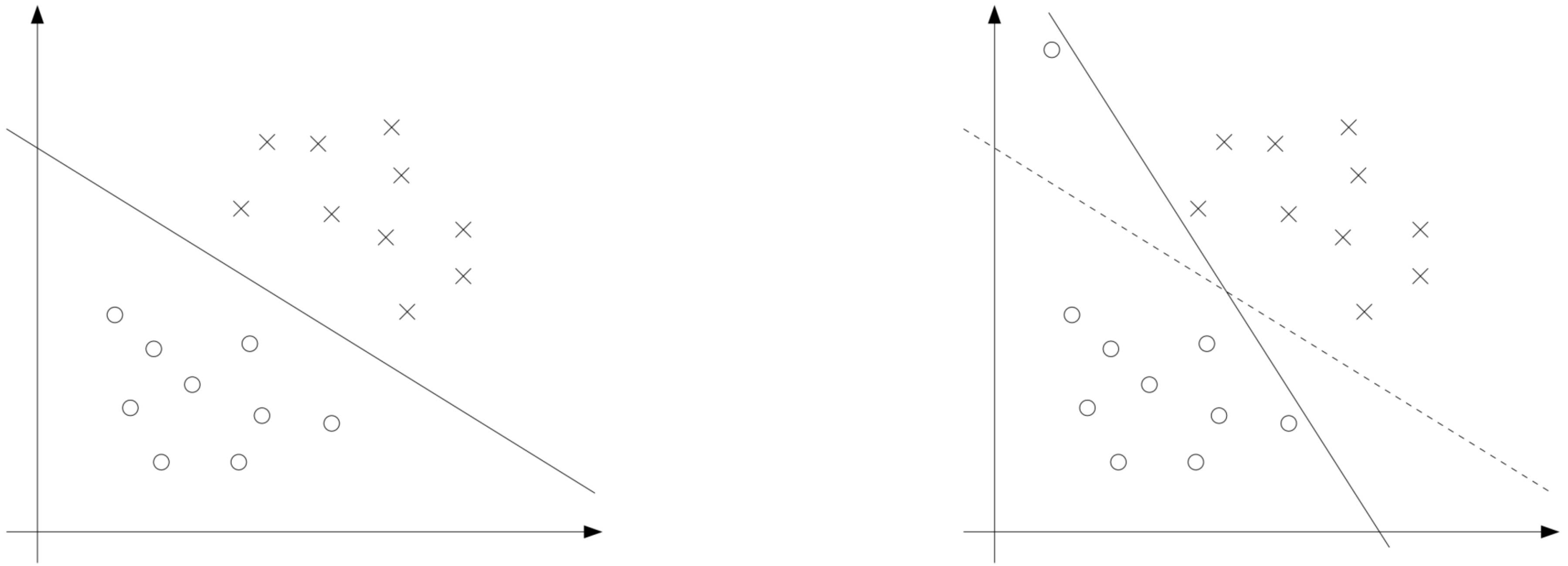
$$
\begin{aligned}
w^T x + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}\right)^T x + b \\
&= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.
\end{aligned}
$$

$$\langle x, z \rangle \Rightarrow \langle \phi(x), \phi(z) \rangle$$

$$K(x, z) = \phi(x)^T \phi(z)$$

Generalmente, calcular K es mucho más fácil/eficiente que calcular las features

¿Y qué hacemos si igualmente el problema no es linealmente separable? ¿O si hay *outliers*?



Podemos relajar la noción de margen (*soft-margin*) mediante nuevas variables

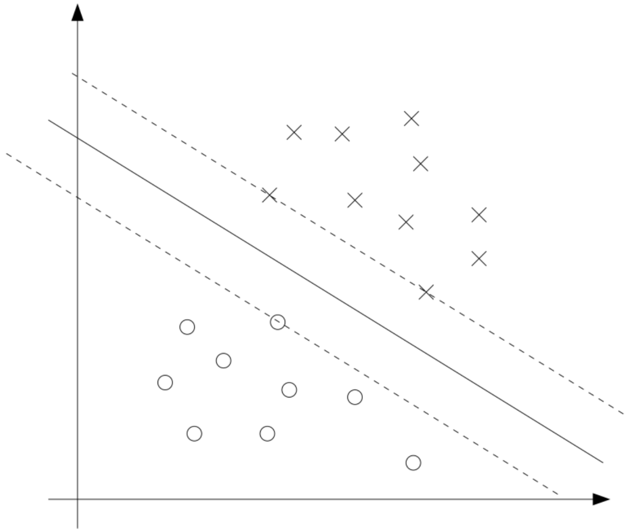
$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ & \xi_i \geq 0, \quad i = 1, \ldots, m. \end{aligned}$$

SVMs continúan siendo relevantes en *machine learning*

- SVMs son de los algoritmos *off-the-shelf* con mejor rendimiento.
- Simpleza y concepto de margen son sus grandes fortalezas.
- Han perdido fuerza últimamente debido a técnicas de Deep Learning.



Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2613 – Inteligencia Artificial

Support Vector Machines (SVM)

Profesor: Hans Löbel