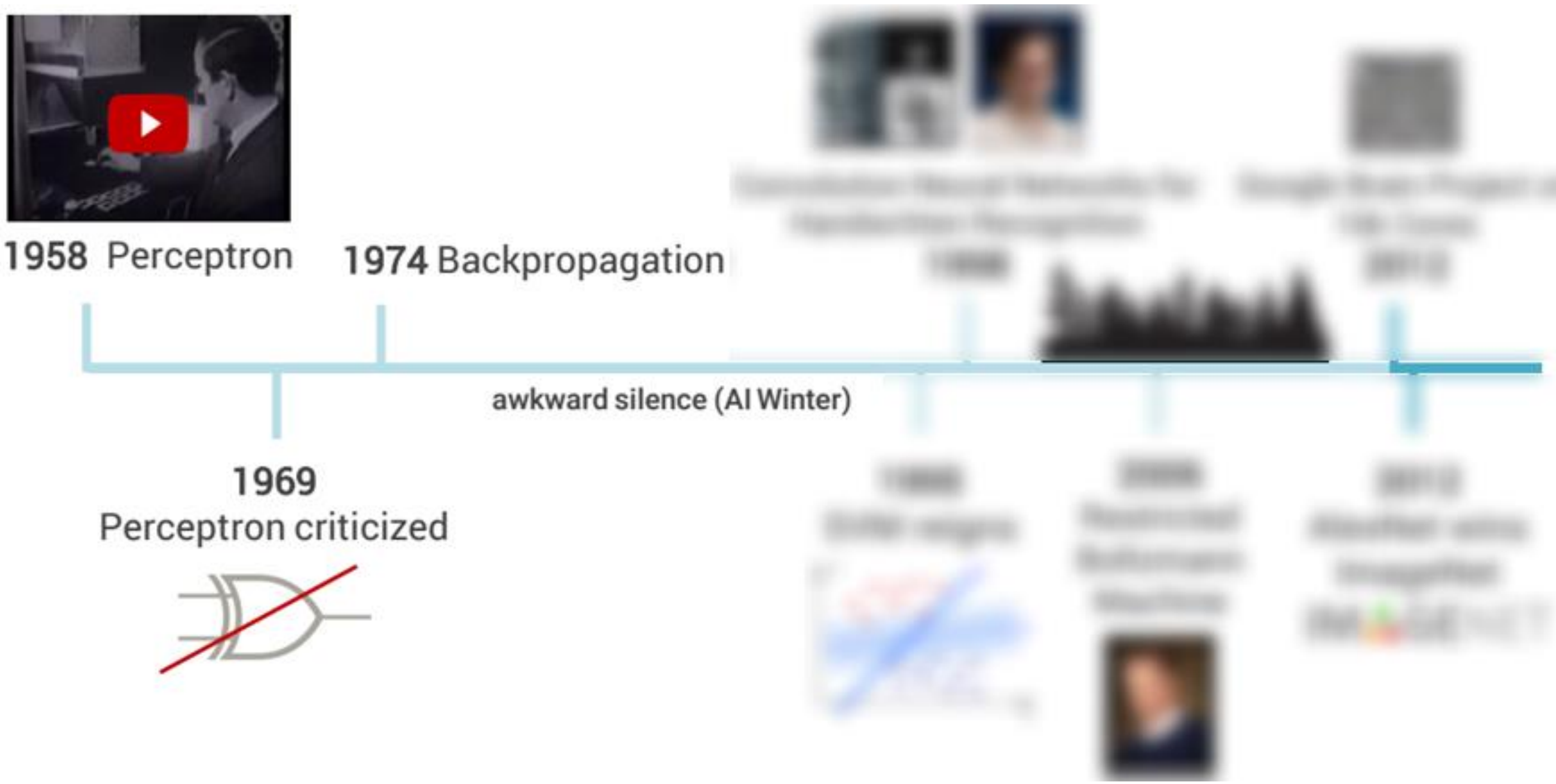
Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2613 – Inteligencia Artificial

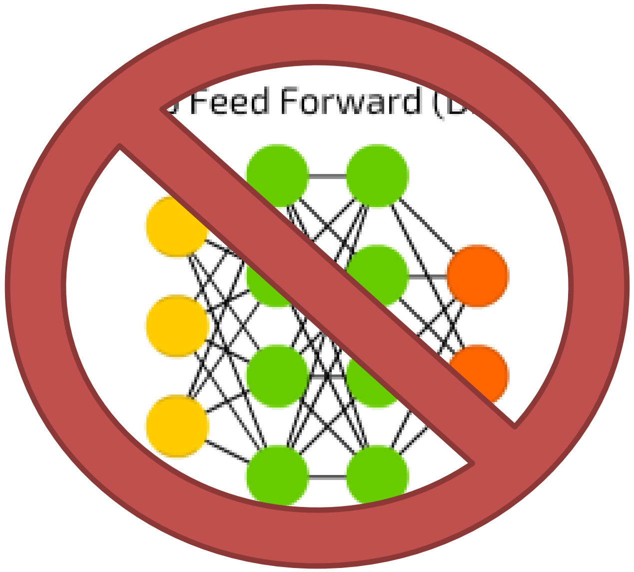
Support Vector Machines (SVM)

Profesor: Hans Löbel

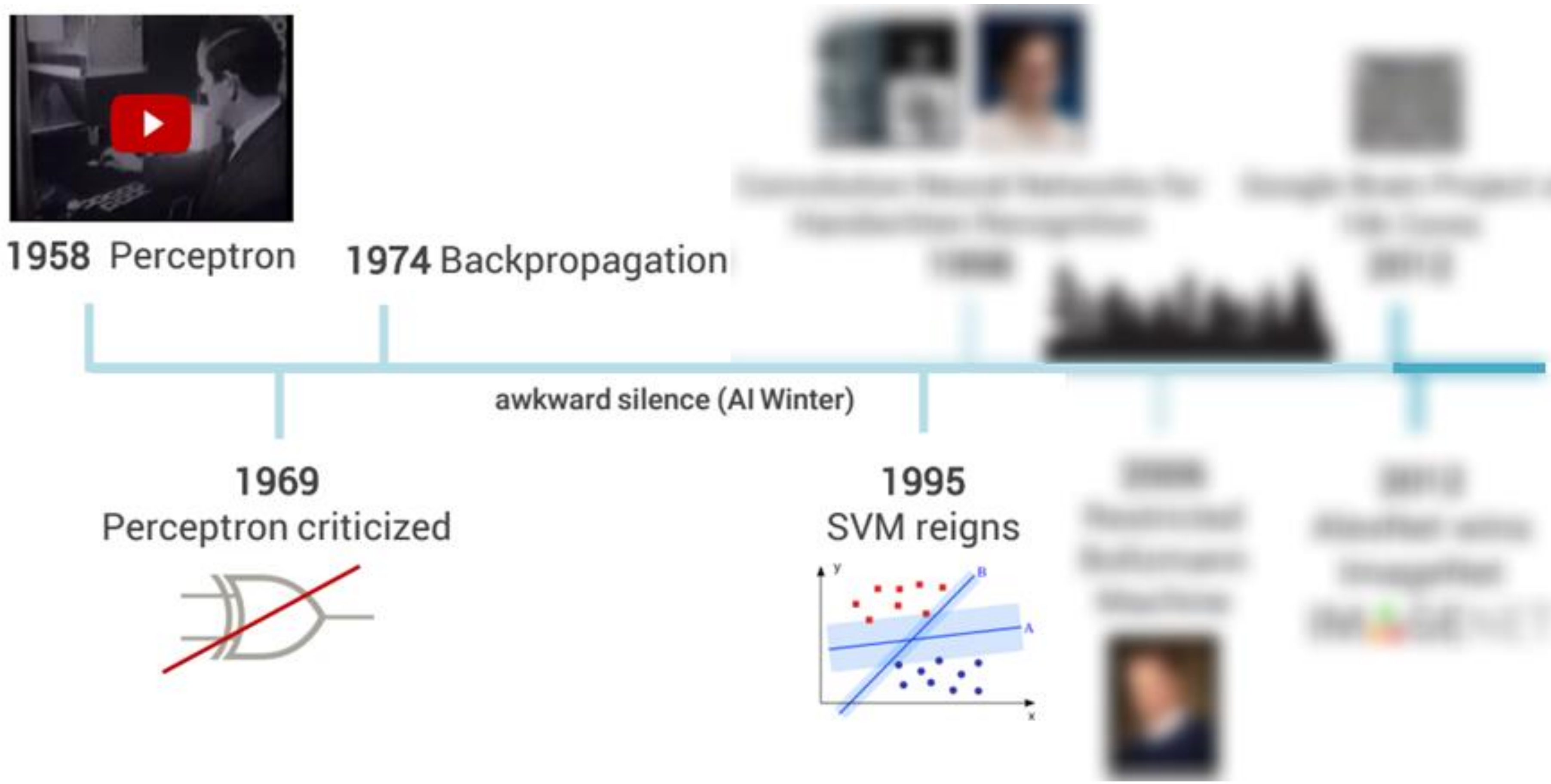

1958 Perceptron
1974 Backpropagation
awkward silence (AI Winter)
1969
Perceptron criticized

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

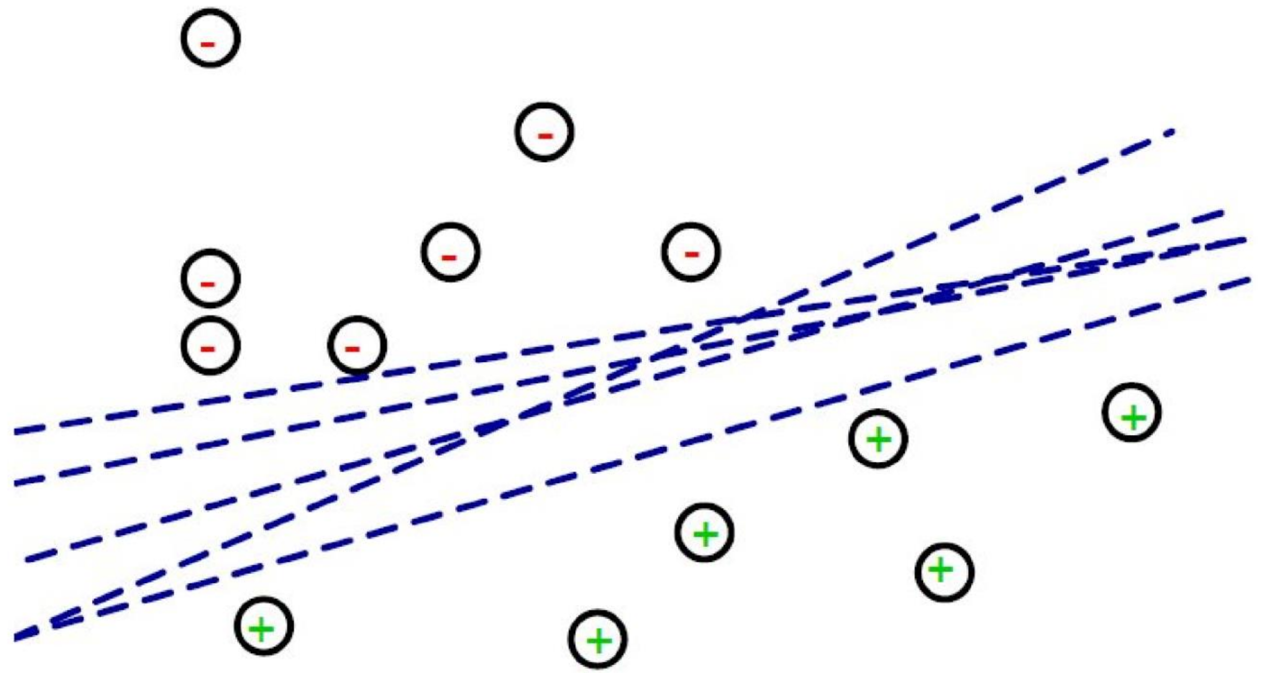
- Redes presentan problema no convexo y mínimos locales.
- Rendimiento no era sustancialmente superior al resto de las técnicas.
- Interpretación de los modelos es altamente compleja.



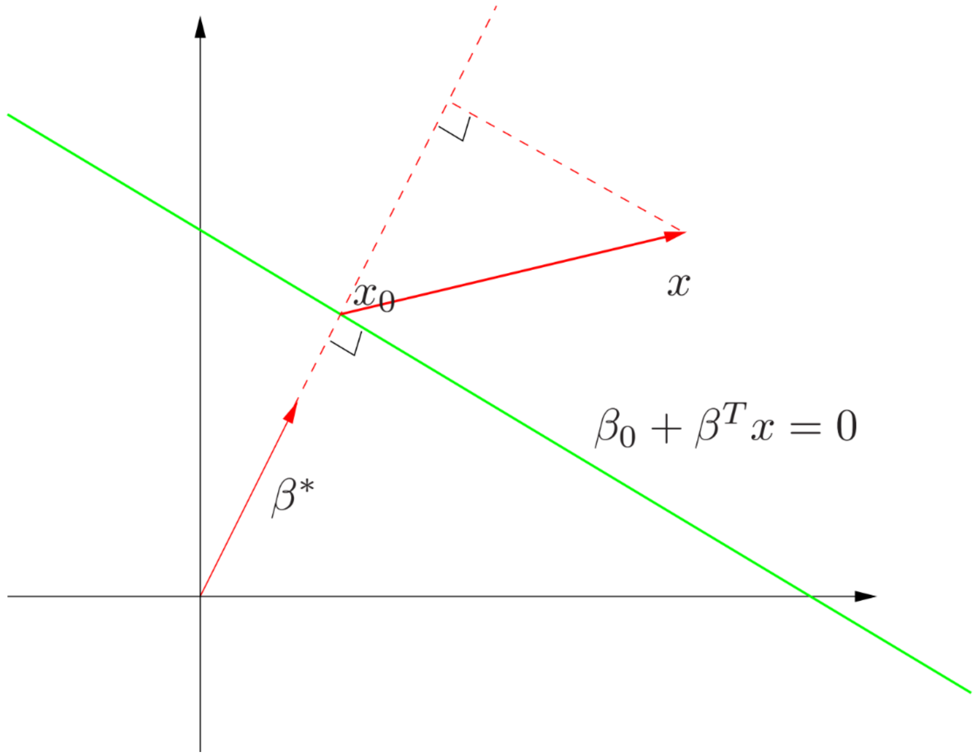
Dadas las restricciones de la época (≈1990), los modelos lineales seguían siendo atractivos, pero requerían mejor rendimiento.


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

¿Cuál es el (hiper)plano que mejor separa dos categorías?



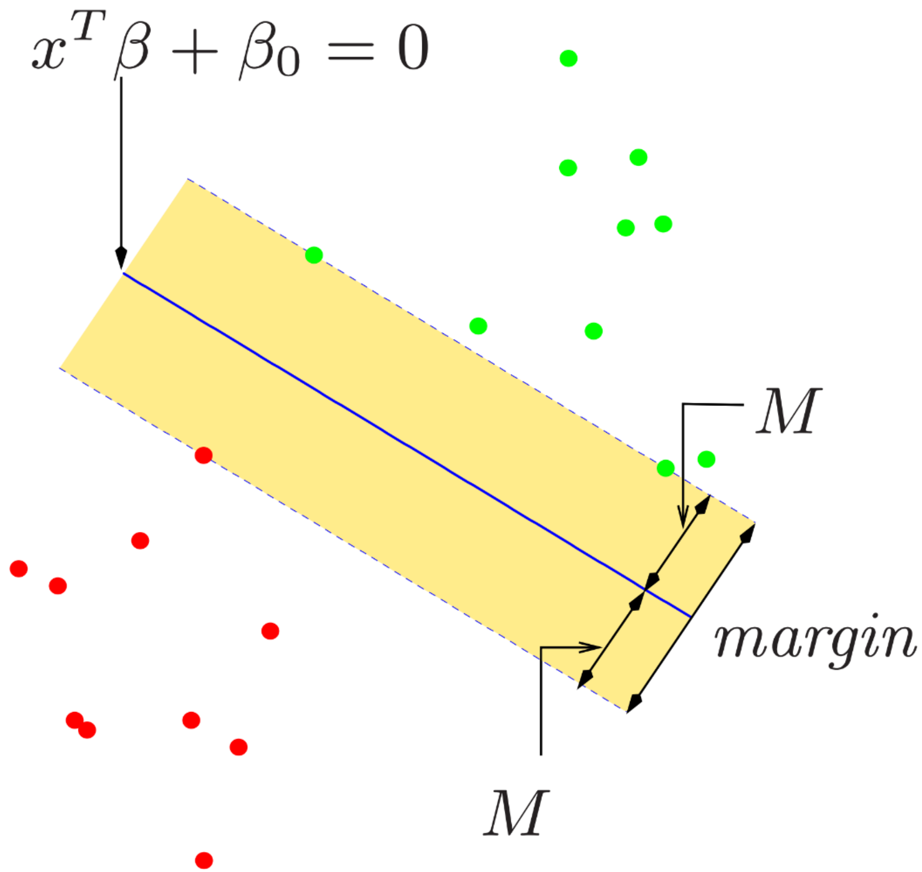
Hagamos un breve repaso de álgebra lineal



- Para cualquier par de puntos x_1 y x_2 en el hiperplano, se cumple: $\beta^T(x_1 - x_2) = 0$
- El vector unitario normal al hiperplano está dado por: $\beta^* = \beta/||\beta||$
- Para cualquier punto x, la distancia signada entre él y el hiperplano esta dada por:

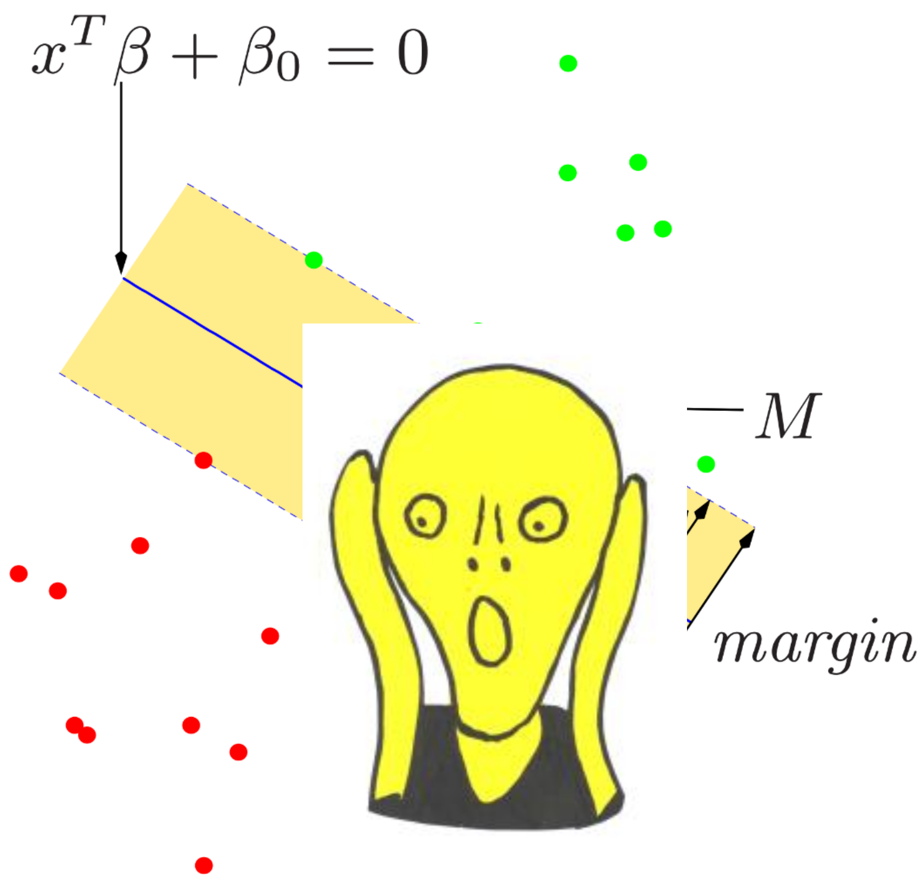
$$\beta^{*T}(x - x_0) \quad = \quad \frac{1}{||\beta||}(\beta^T x + \beta_0) \quad = \quad \frac{1}{||f'(x)||} f(x)$$

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



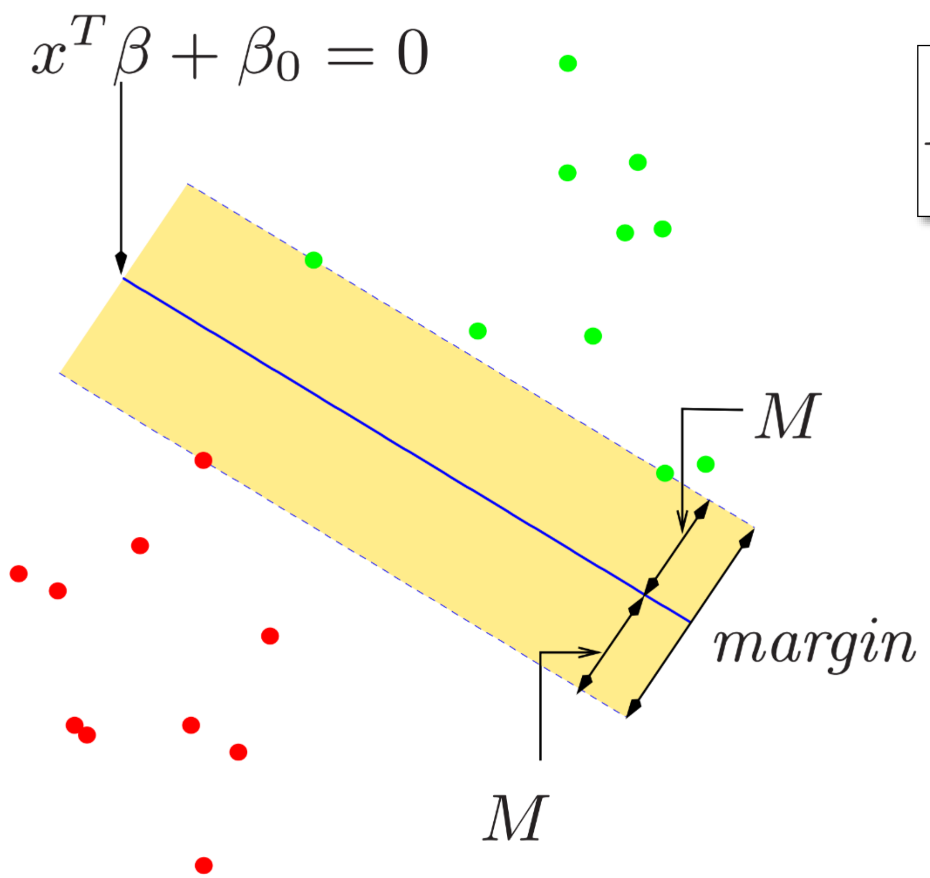
$$D = \frac{y}{||\beta||}(x^T\beta + \beta_0)$$

¿Cómo podemos plantear esto como un problema de optimización?



$$\max \quad M$$

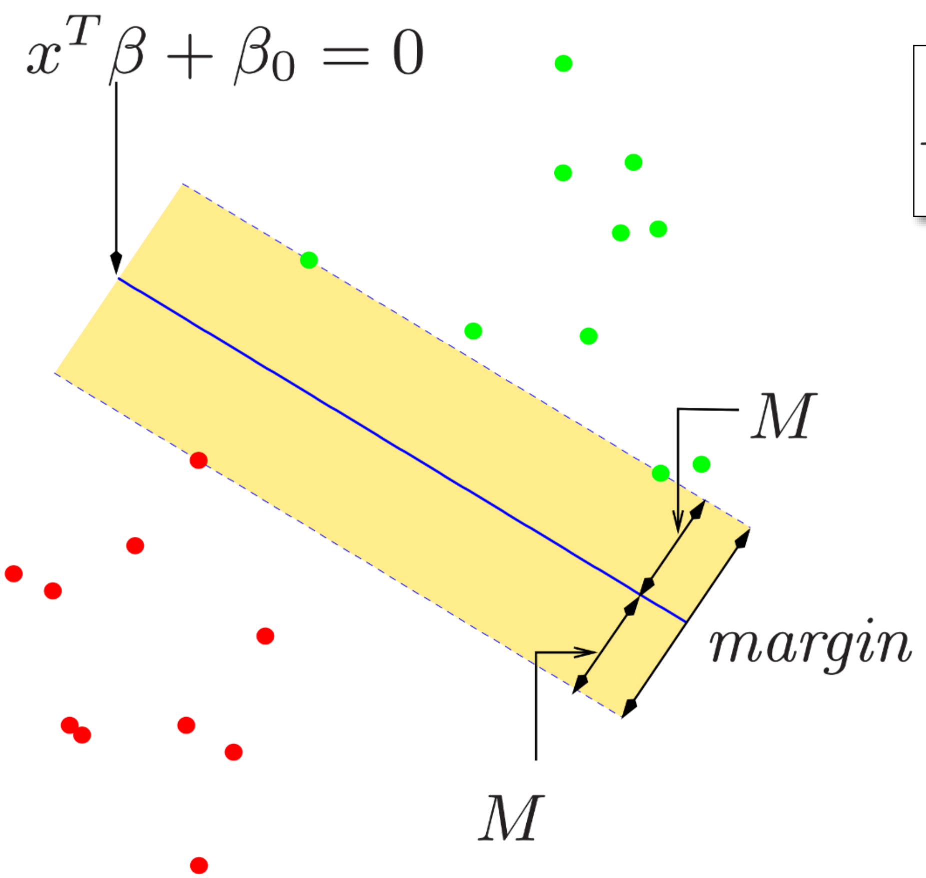
¿Cómo podemos plantear esto como un problema de optimización?



$$D = \frac{y}{\|\beta\|}(x^T\beta + \beta_0)$$

$$\max_{\beta,\beta_0} \quad M$$

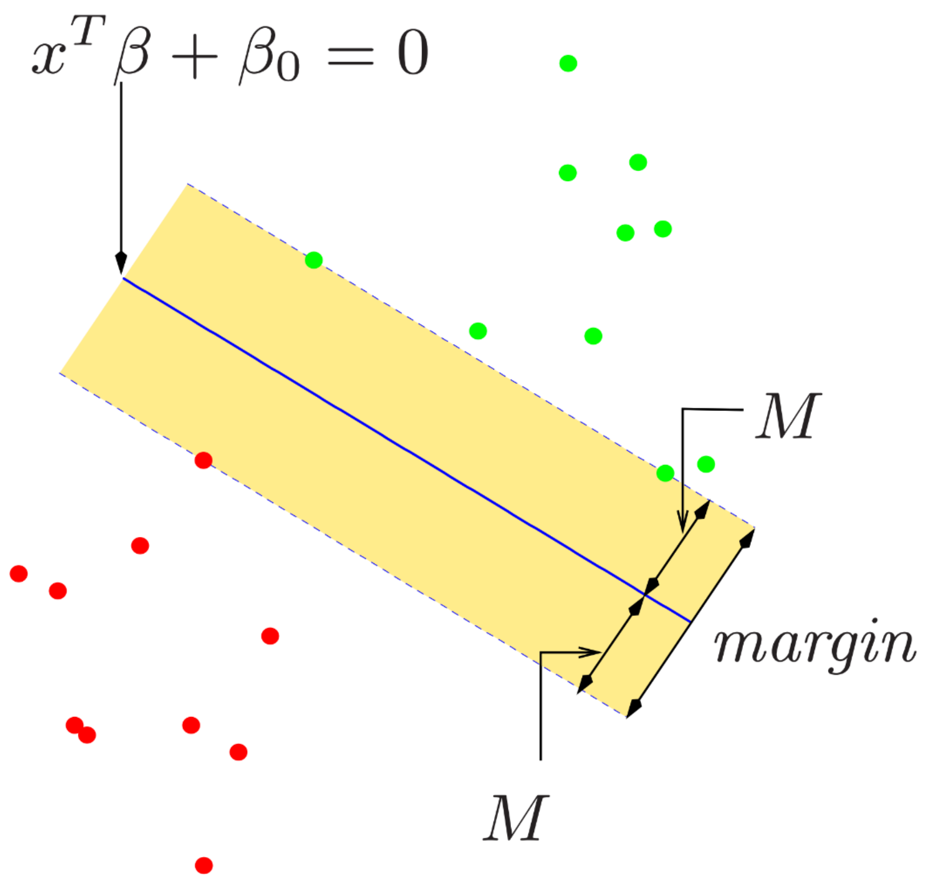
¿Cómo podemos plantear esto como un problema de optimización?



$$D = \frac{y}{||\beta||}(x^T\beta + \beta_0)$$

$$\max_{\beta,\beta_0} \quad M$$
$$\text{subject to } \frac{1}{||\beta||} y_i(x_i^T\beta + \beta_0) \geq M, \ i = 1, \ldots, N$$

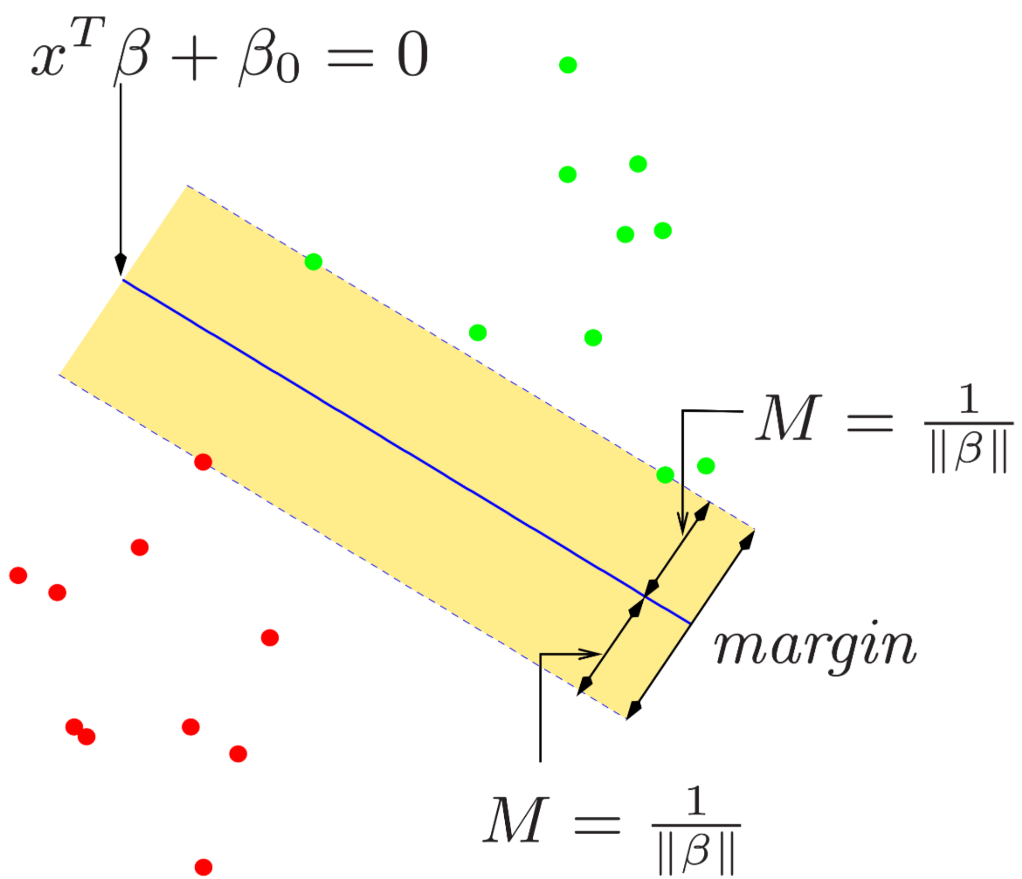
¿Cómo podemos plantear esto como un problema de optimización?



$$\max_{\beta,\beta_0} \quad M$$

$$\text{subject to} \quad y_i(x_i^T\beta + \beta_0) \geq M||\beta||, \; i = 1, \ldots, N$$

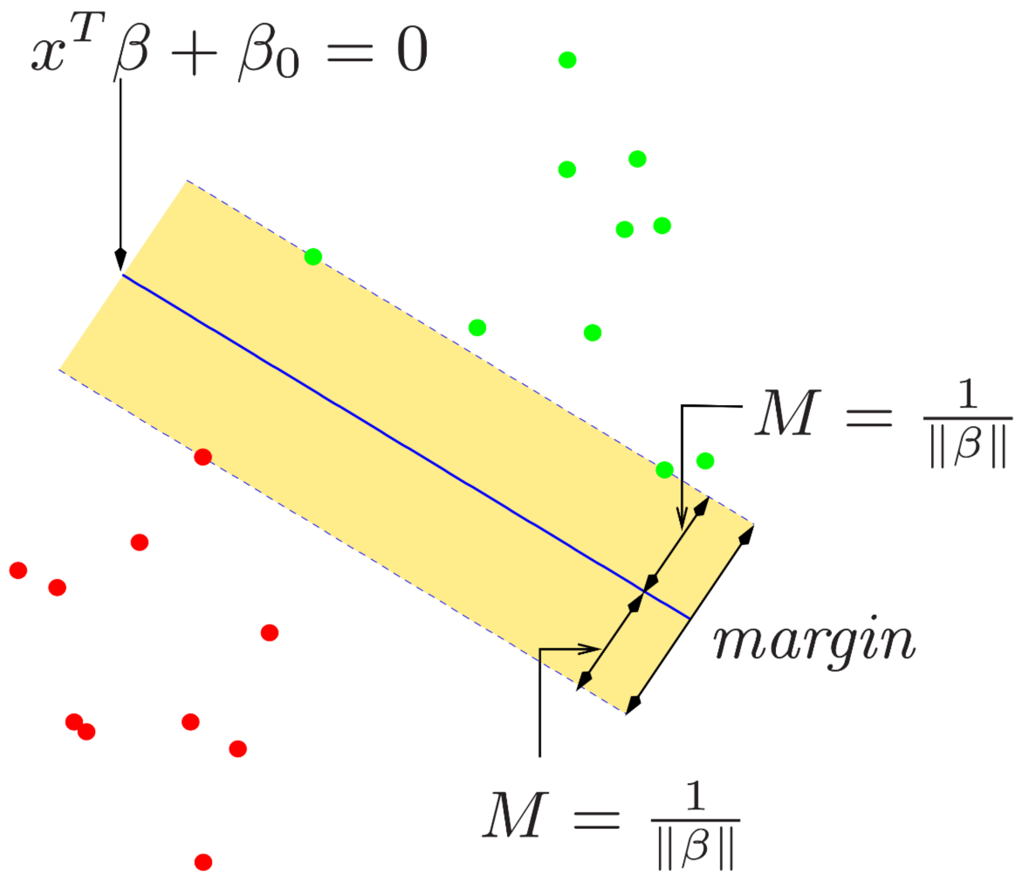
¿Cómo podemos plantear esto como un problema de optimización?



$$\max_{\beta,\beta_0} \quad M$$

$$\text{subject to} \;\; y_i(x_i^T\beta + \beta_0) \geq M\|\beta\|, \; i = 1, \ldots, N$$

¿Cómo podemos plantear esto como un problema de optimización?



$$\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2$$

$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq 1,\ i = 1, \ldots, N$$

Veamos como podemos resolver este problema (versión light)

$$\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2$$
$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq 1,\ i = 1, \ldots, N$$

$$\Downarrow$$

$$L_P = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{N} \alpha_i[y_i(x_i^T\beta + \beta_0) - 1]$$

Veamos como podemos resolver este problema (versión light)

$$L_P = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{N} \alpha_i [y_i(x_i^T \beta + \beta_0) - 1]$$

Derivando e igualando a cero, obtenemos:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i \qquad\qquad 0 = \sum_{i=1}^{N} \alpha_i y_i$$

Sustituyendo todo esto en el lagrangiano, obtenemos el dual:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k$$

$$\text{subject to } \alpha_i \geq 0 \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0$$

Veamos como podemos resolver este problema (versión light)

$$\max_{\alpha} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k$$

$$\text{subject to } \alpha_i \geq 0 \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\text{and } \alpha_i [y_i (x_i^T \beta + \beta_0) - 1] = 0 \;\forall i$$

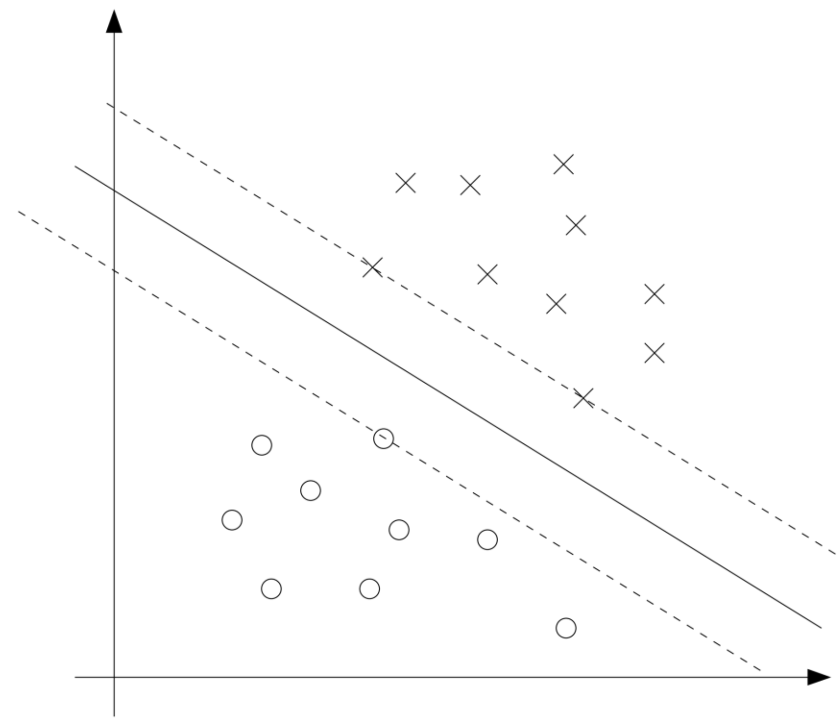
Esta última restricción (KKT) es fundamental para entender los SVM:

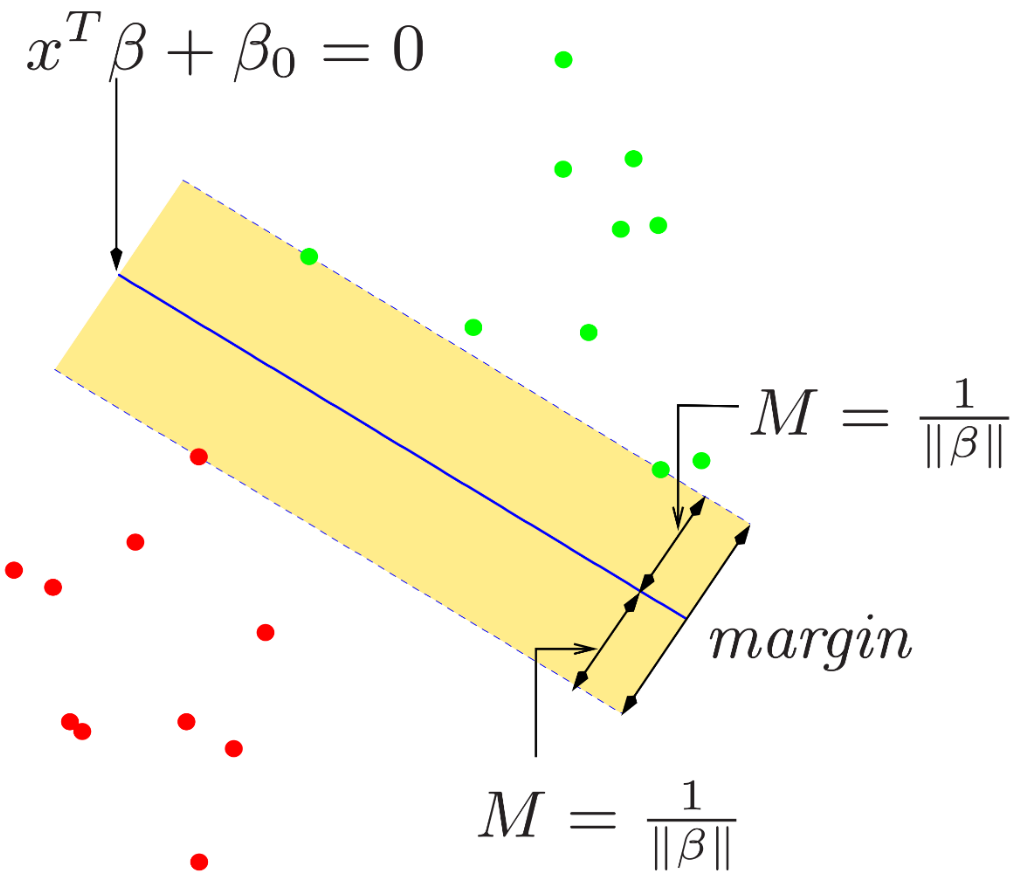
- Si $\alpha_i > 0,\; y_i(x_i^T \beta + \beta_0) = 1$ (el punto queda sobre el límite del margen)
- Si $y_i(x_i^T \beta + \beta_0) > 1$ (punto queda fuera del margen), $\alpha_i = 0$.

El problema dual, permite una interpretación más clara de los vectores de soporte.

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\alpha_i [y_i(x_i^T \beta + \beta_0) - 1] = 0 \; \forall i$$





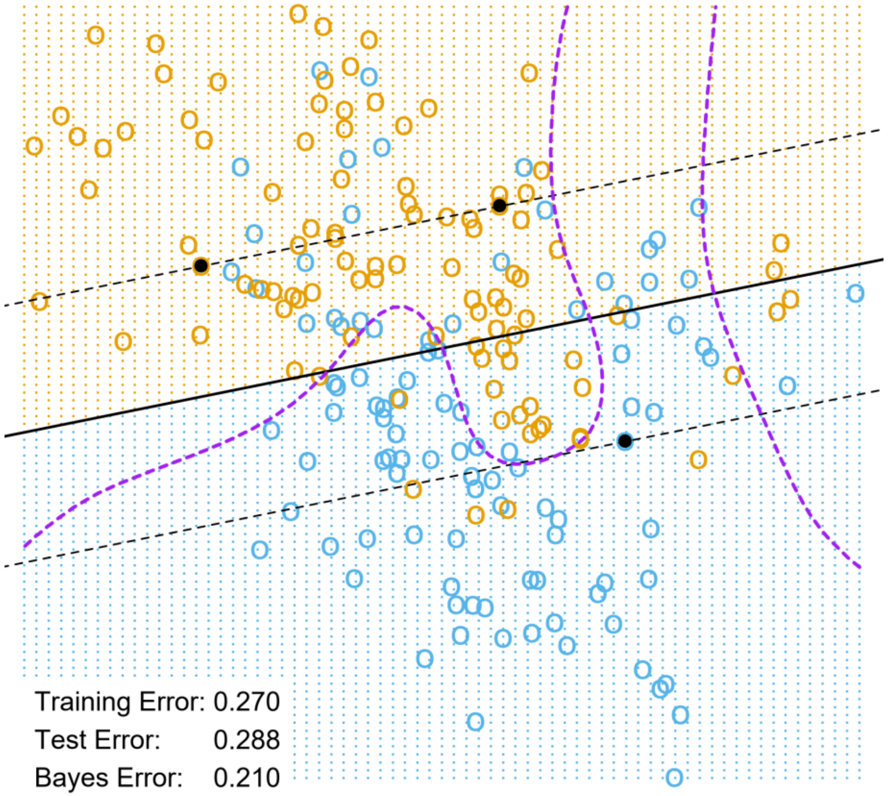
$$\frac{1}{||\beta||} y_i(x_i^T\beta + \beta_0) \quad \geq \quad M(1 - \xi_i)$$

$$\min \|\beta\| \quad \text{subject to} \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \ \forall i, \\ \xi_i \geq 0, \ \sum \xi_i \leq \text{constant}. \end{cases}$$

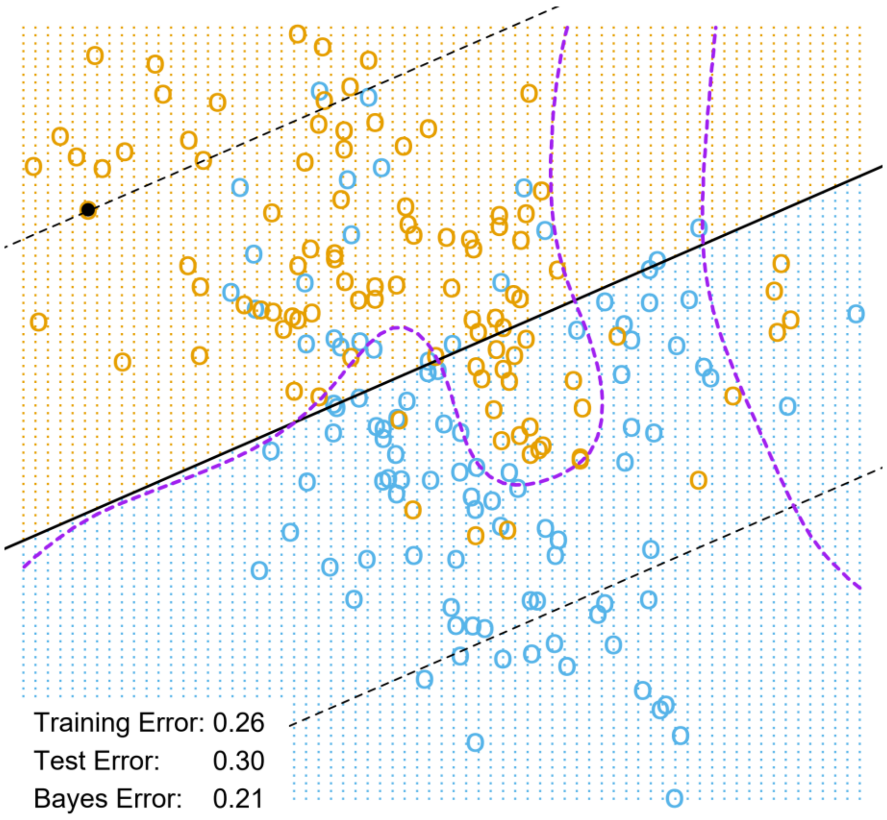
$$\Downarrow$$

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{subject to} \quad \xi_i \geq 0, \ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \ \forall i$$



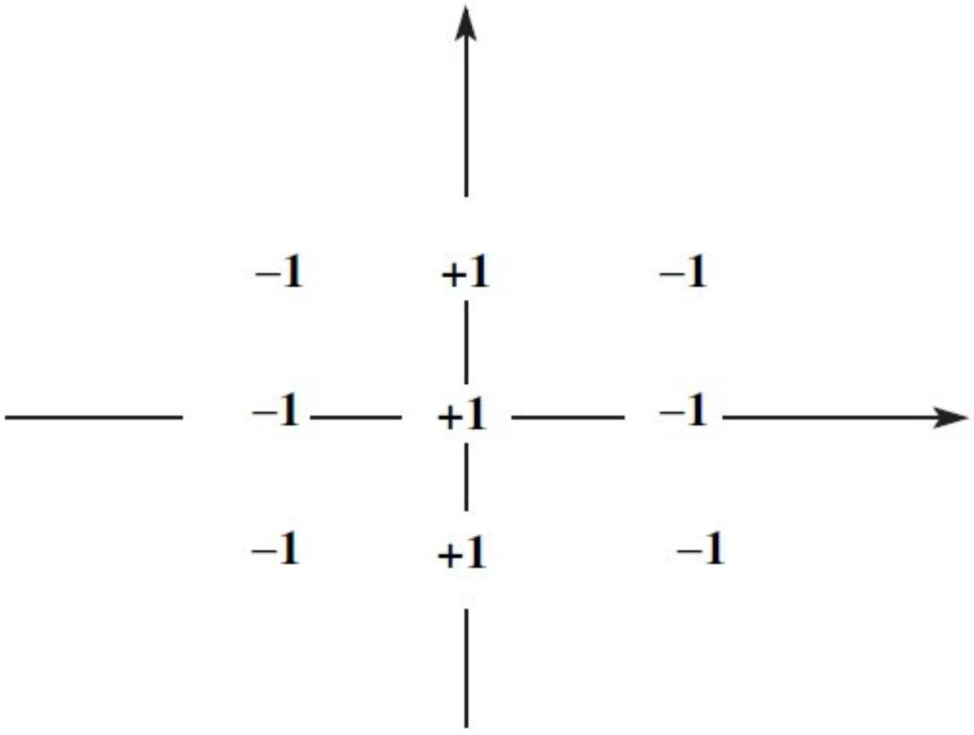
C = 1000



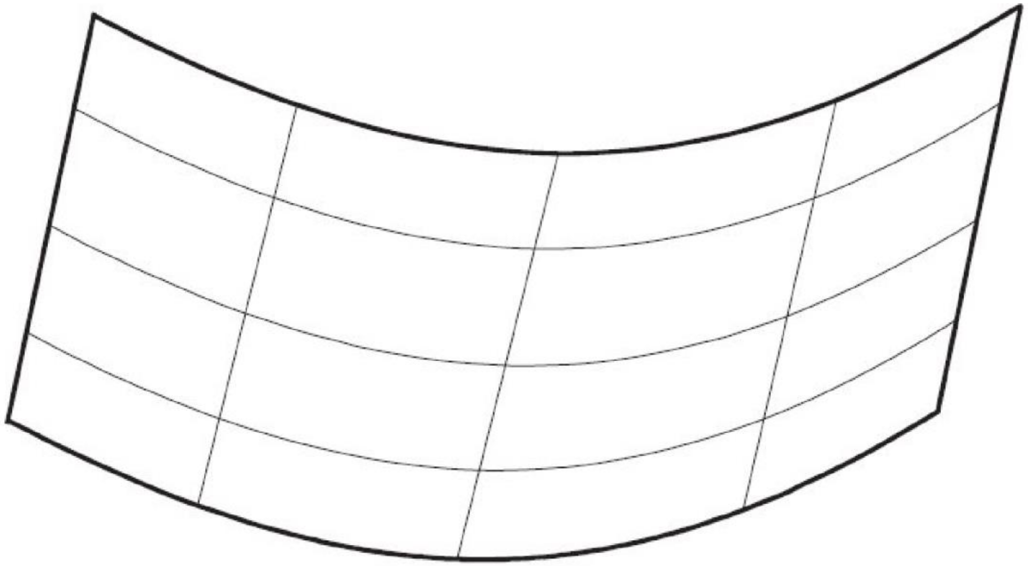
C = 0.01

- ¿Cuál de las dos soluciones tiene un mayor valor para la constante C?
- ¿Cómo puedo estimar el valor óptimo de C?

Súper lindo, pero sigue siendo un clasificador lineal

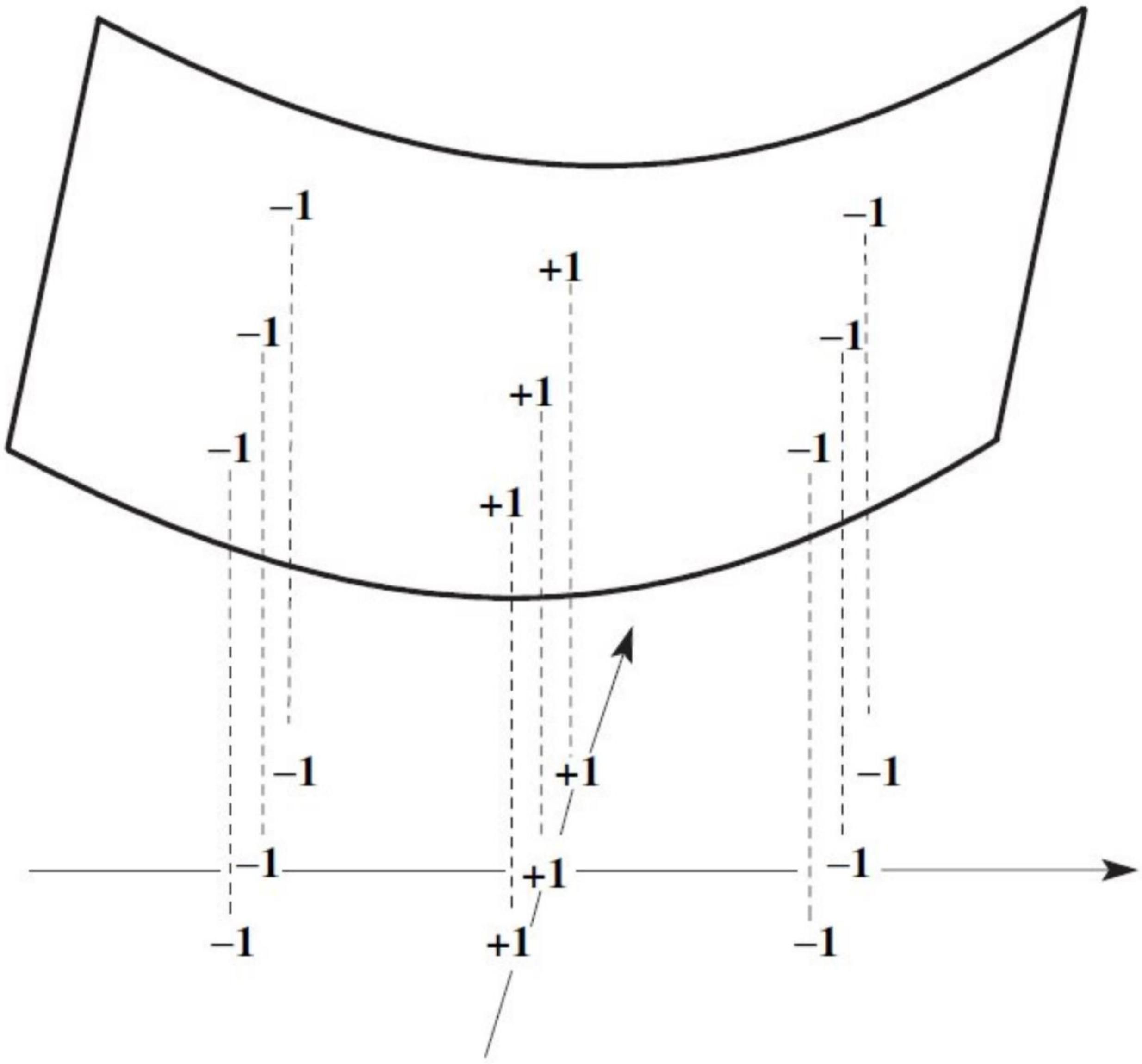


Una solución es generar un cambio de variables (transformación del espacio de características)

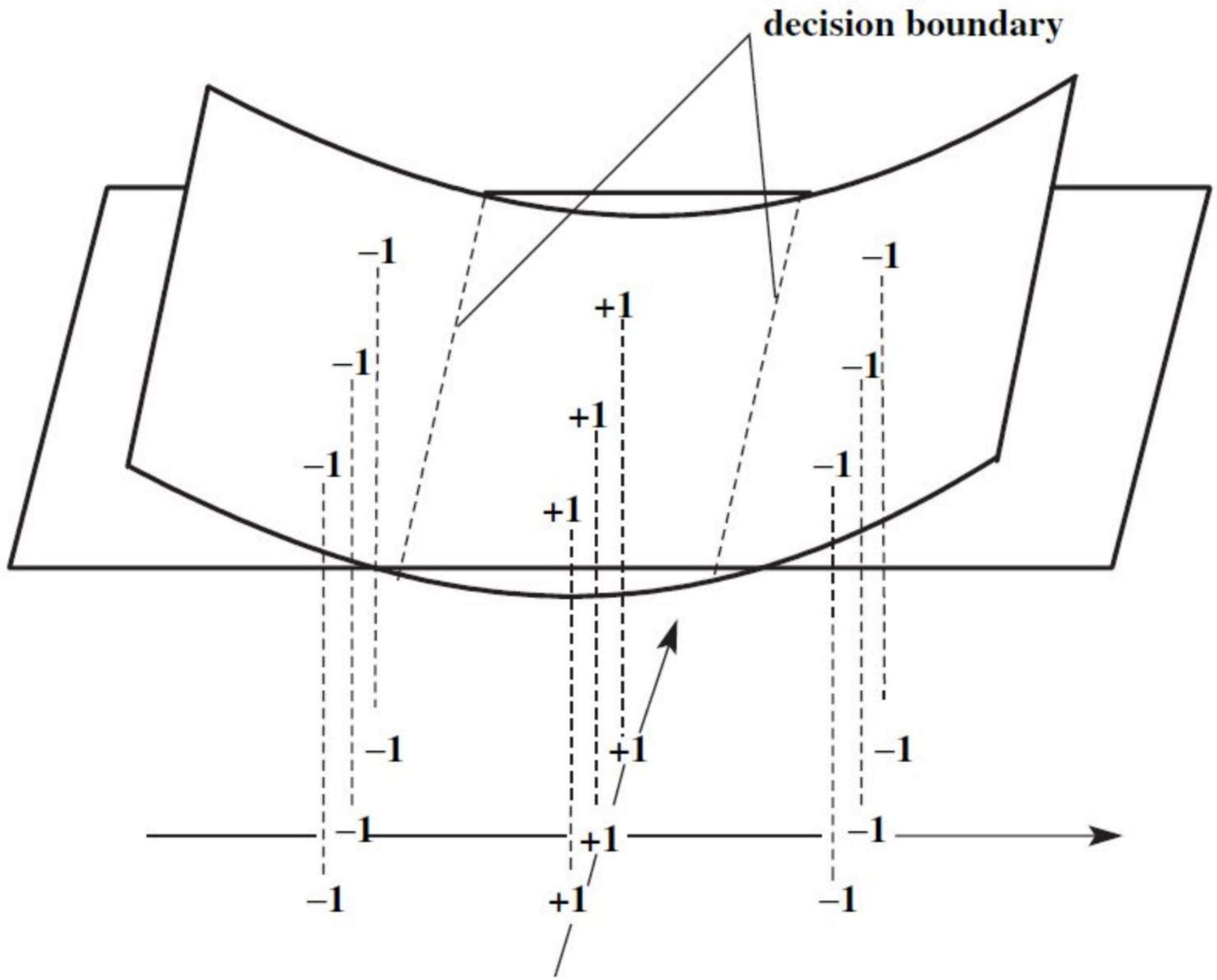


Espacio de características $f(x_1, x_2) = x_1^2$

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



Podemos incorporar esto en los SVMs mediante algo conocido como el *kernel trick*

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} x_i^T x_{i'}$$

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} \langle x_i, x_{i'} \rangle$$

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} \langle h(x_i), h(x_{i'}) \rangle$$

Podemos incorporar esto en los SVMs mediante algo conocido como el *kernel trick*

$$f(x) = h(x)^T\beta + \beta_0$$

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$$

Podemos incorporar esto en los SVMs mediante algo conocido como el *kernel trick*

$$\begin{aligned} f(x) &= h(x)^T\beta + \beta_0 \\ &= \sum_{i=1}^{N} \alpha_i y_i \langle h(x), h(x_i)\rangle + \beta_0 \end{aligned}$$

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$K(x, x') = \langle h(x), h(x')\rangle$$

Es posible construir una matriz K, conocida como la matriz de *kernel:*

- Si K es positiva semidefinida, entonces define un kernel de *Mercer* válido.
- Aplicar un kernel de Mercer es análogo a aplicar el producto punto entre features de mayor dimensionalidad que la original (potencialmente infinita).

Intuitivamente, kernels miden la similitud entre dos vectores

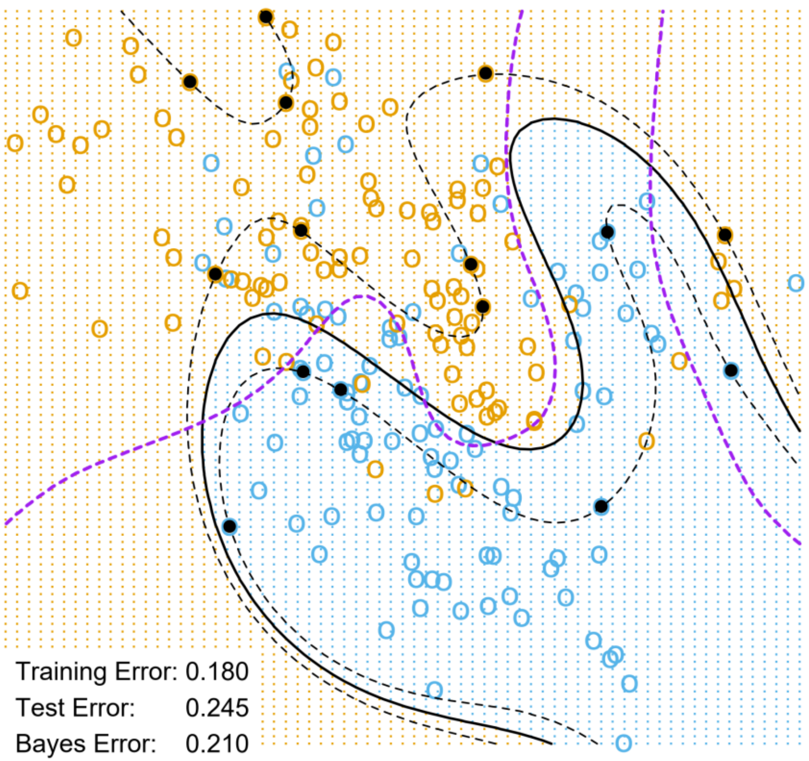
$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + \beta_0$$

$$\boxed{\begin{aligned} d\text{th-Degree polynomial: } & K(x,x') = (1 + \langle x, x' \rangle)^d, \\ \text{Radial basis: } & K(x,x') = \exp(-\gamma \|x - x'\|^2), \\ \text{Neural network: } & K(x,x') = \tanh(\kappa_1 \langle x, x' \rangle + \kappa_2). \end{aligned}}$$

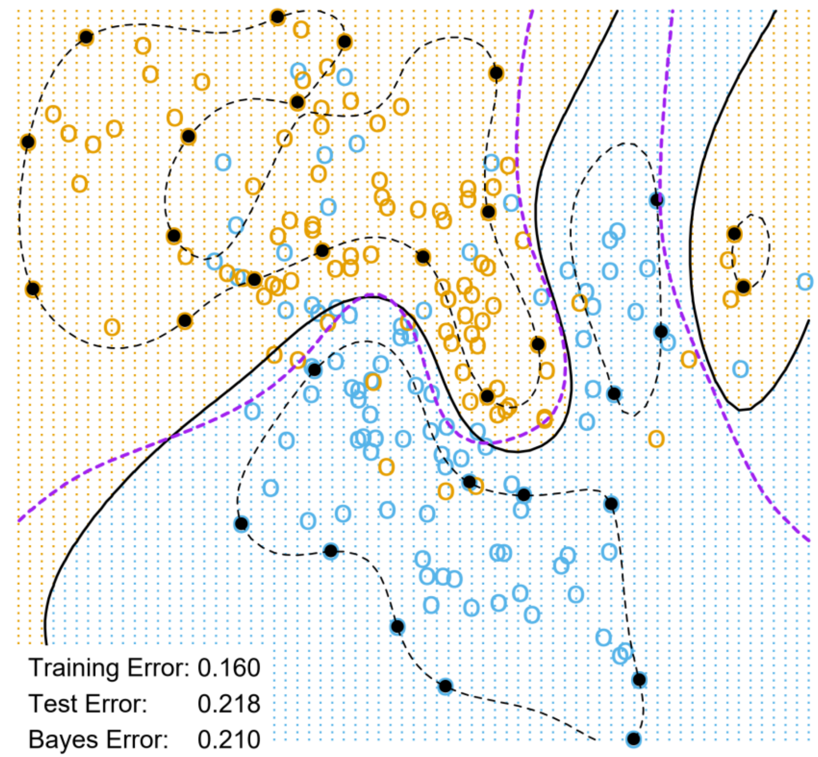
Un par de preguntas para terminar SVMs

- ¿Cuáles son los vectores de soporte en este caso?
- ¿Cómo se ve el margen en este caso?
- ¿Cuál es el kernel del SVM que vimos anteriormente?

SVM - Degree-4 Polynomial in Feature Space

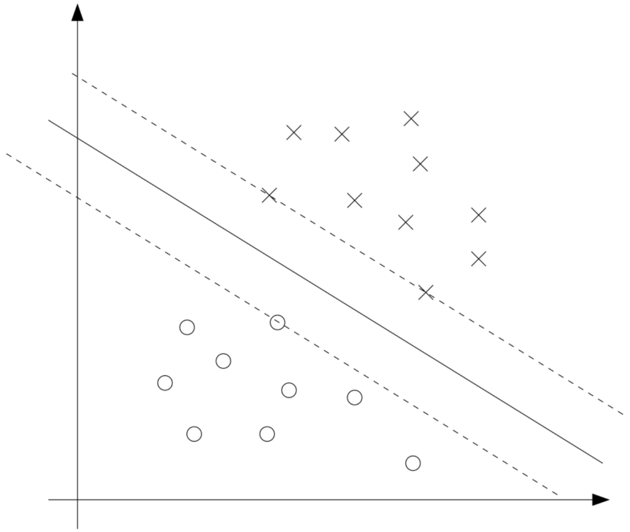


SVM - Radial Kernel in Feature Space



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