

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



IIC2613 - Inteligencia Artificial

Support Vector Machines (SVM)

Hans Löbel

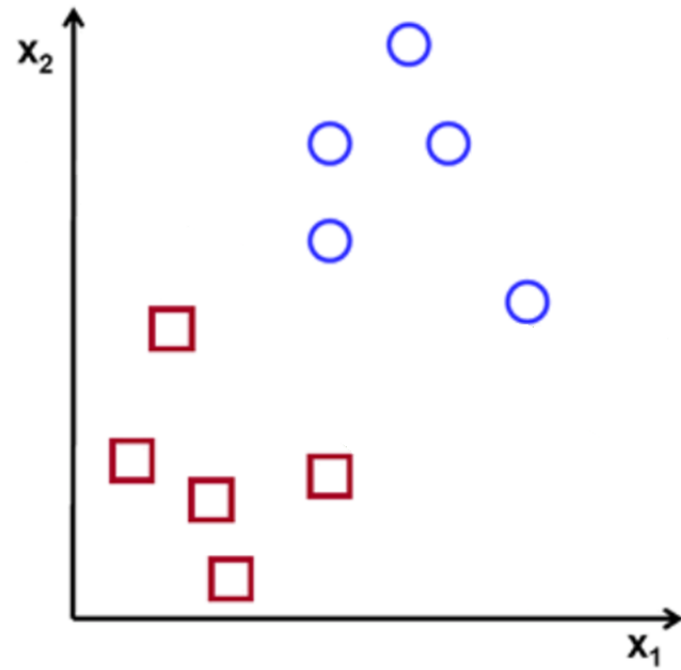
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación

Recapitulemos un poco lo que hemos visto hasta ahora en el curso

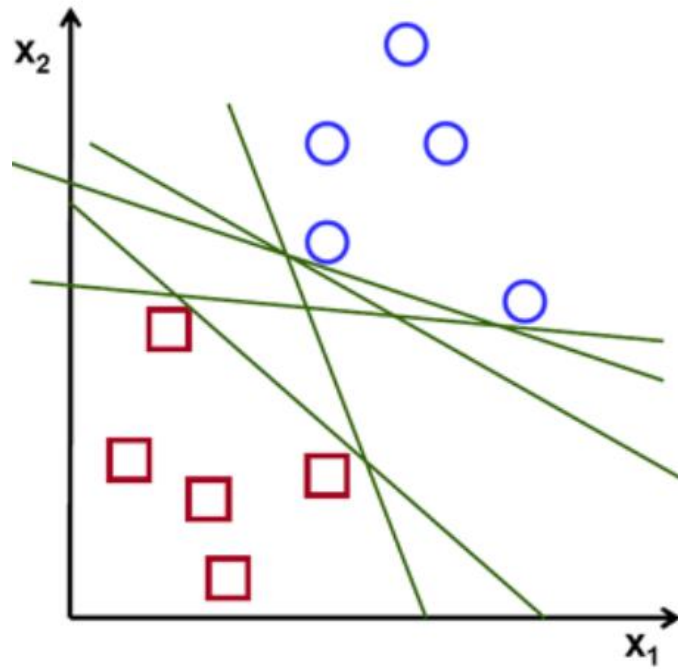
- Comenzamos con los conceptos fundamentales de ML.
- En las siguientes sesiones exploramos árboles y ensambles, siempre teniendo en consideración el *trade-off* entre complejidad y sobreajuste.
- Hoy veremos una técnica que fue el estado del arte durante largo tiempo, y captura algunos conceptos centrales de lo que es ML moderno.



Visualicemos un problema binario de clasificación en 2D,
¿cómo podríamos separar estas clases usando una línea?

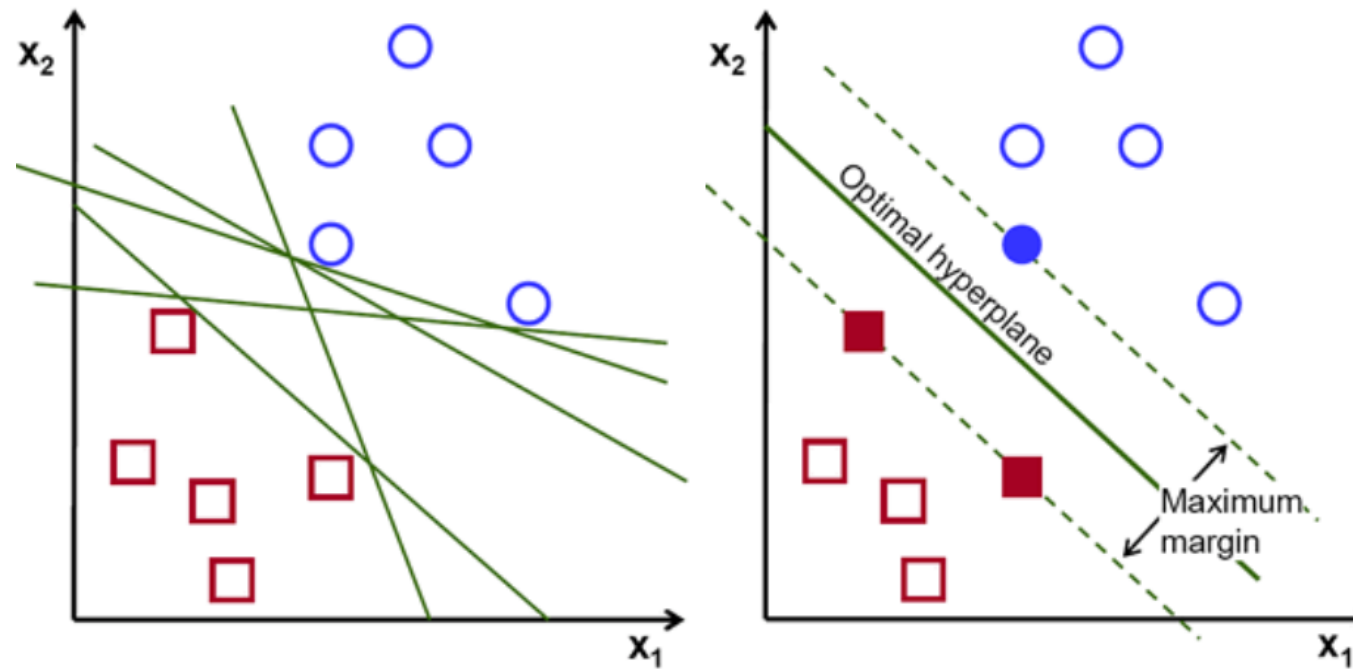


Existen múltiples soluciones, muchas incluso con rendimiento perfecto. Pero ¿existe una “mejor” que el resto?



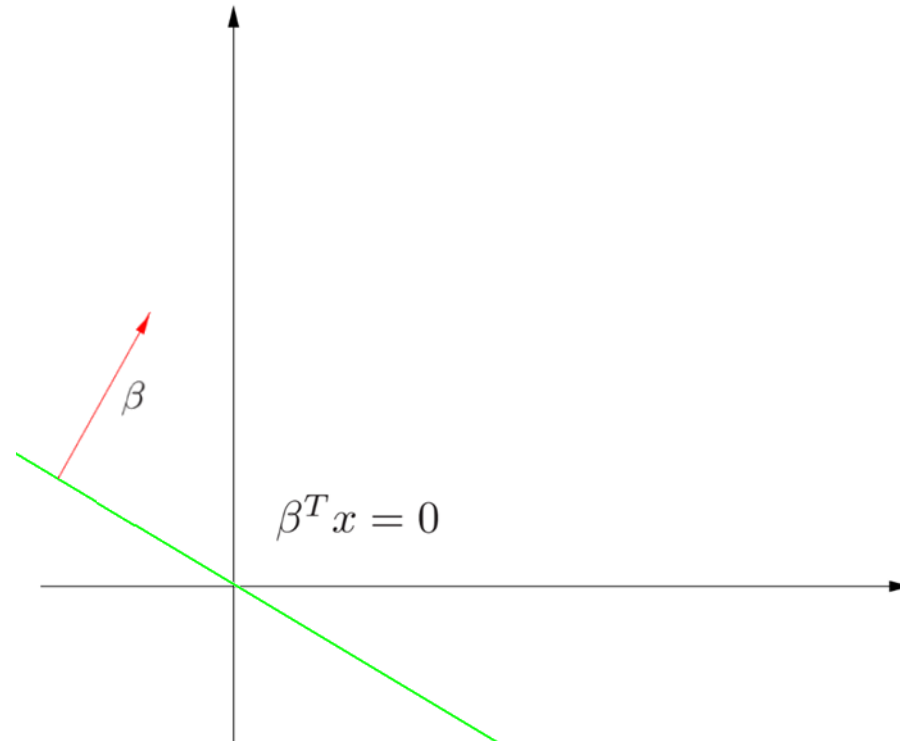
¿Cómo podríamos incorporar algo relacionado con la capacidad de generalización en la solución?

El (hiper)plano óptimo es aquel que separa las dos categorías con el máximo margen

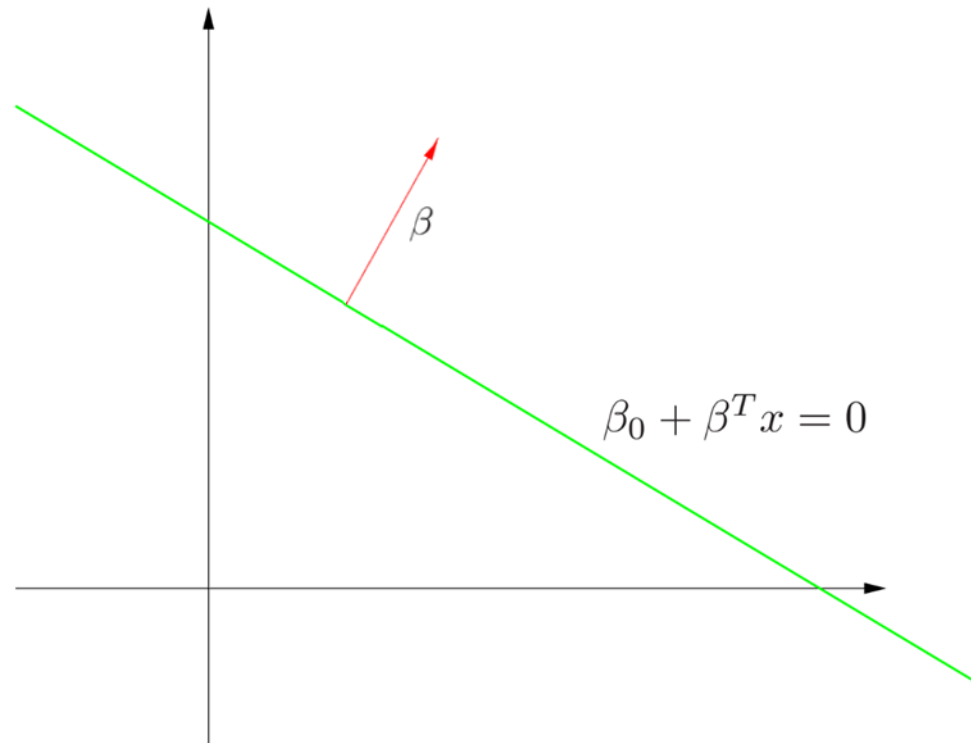


Mientras mayor sea la distancia entre el hiperplano y los puntos, mayor espacio existe para los datos difíciles

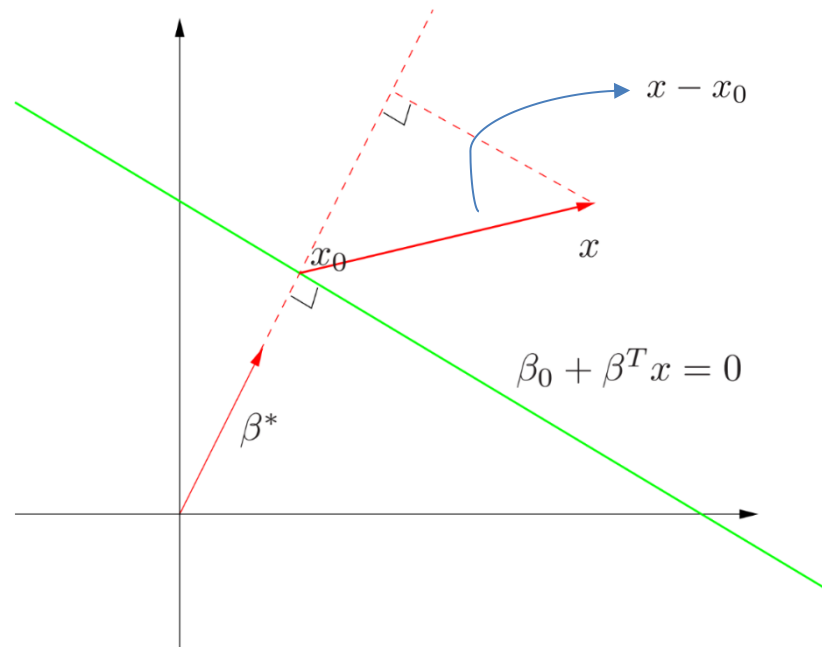
Hagamos un breve repaso de álgebra lineal



Hagamos un breve repaso de álgebra lineal



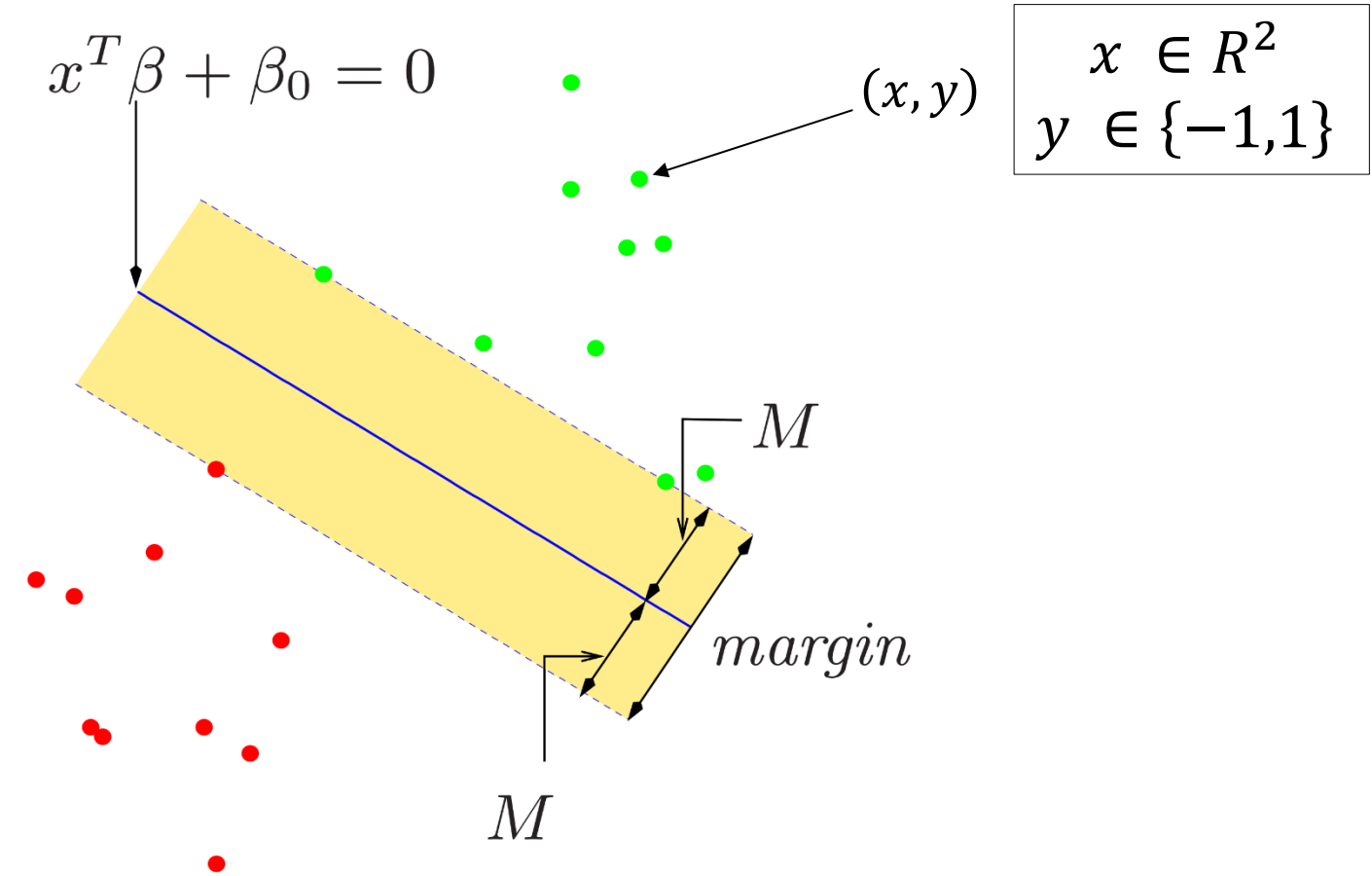
Hagamos un breve repaso de álgebra lineal



- 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 está dada por:

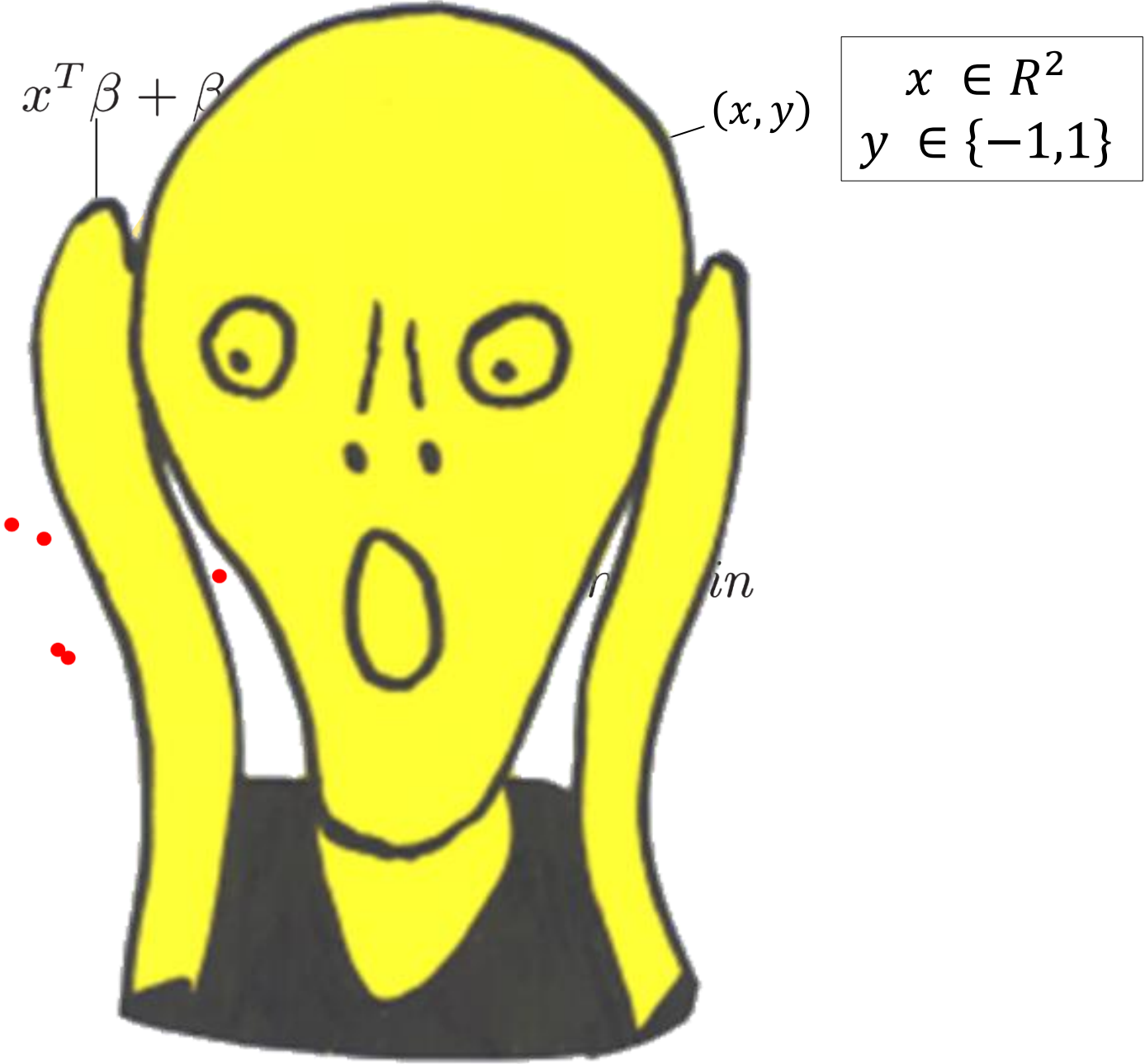
$$\beta^{*T}(x - x_0) = \frac{1}{\|\beta\|}(\beta^T x + \beta_0)$$

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



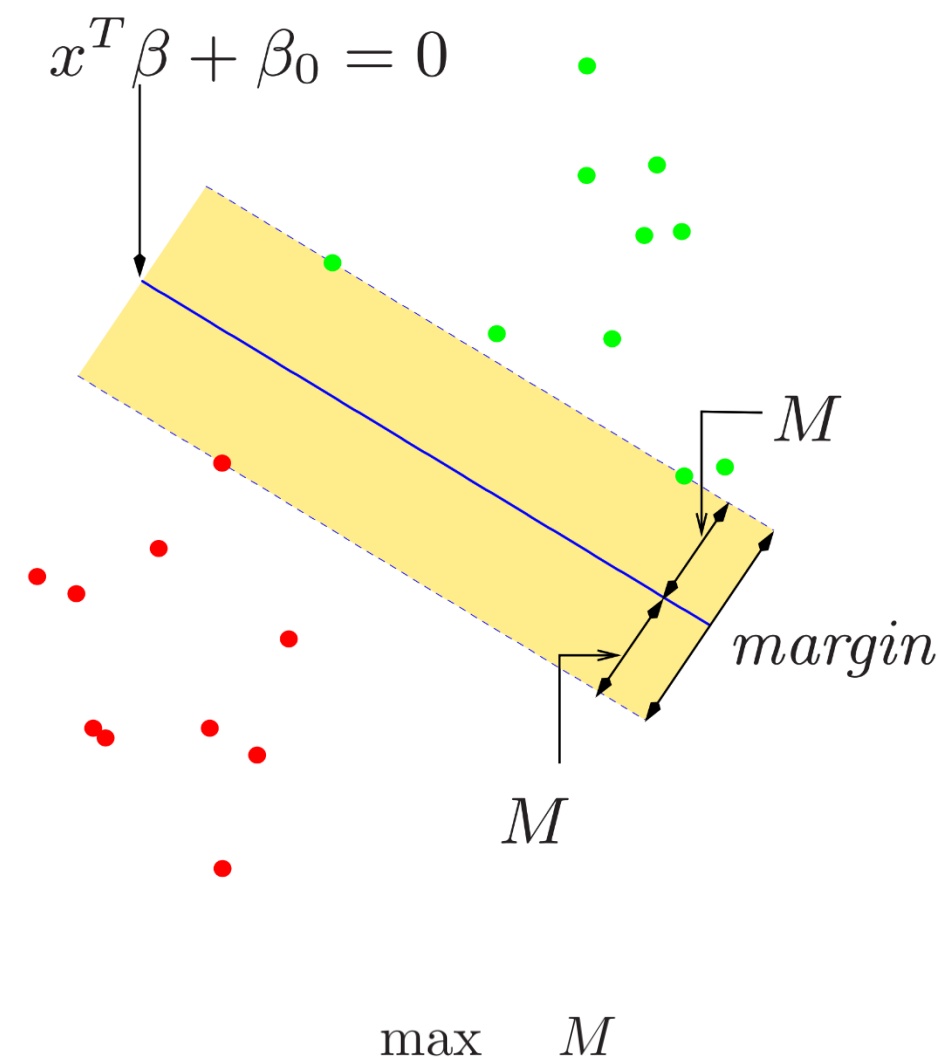
$$D = \frac{y}{\|\beta\|} (x^T \beta + \beta_0)$$

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

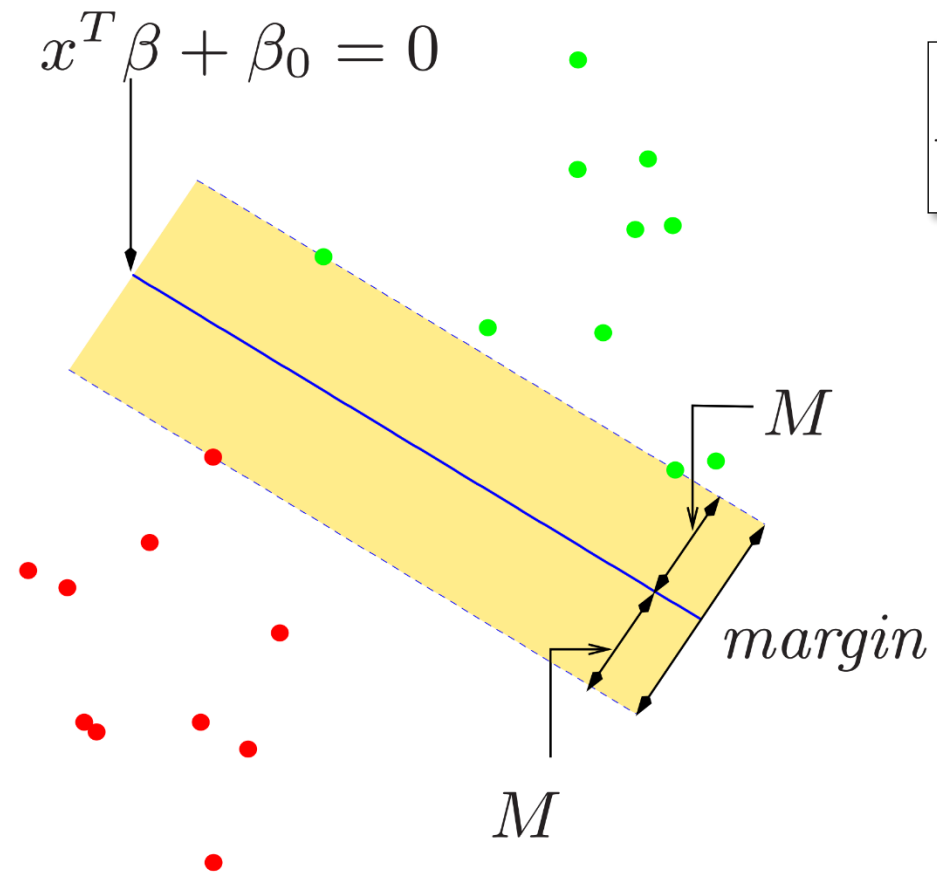


¿Cómo podemos plantear todo esto como un problema de optimización?

¿Cómo podemos plantear todo esto como un problema de optimización?



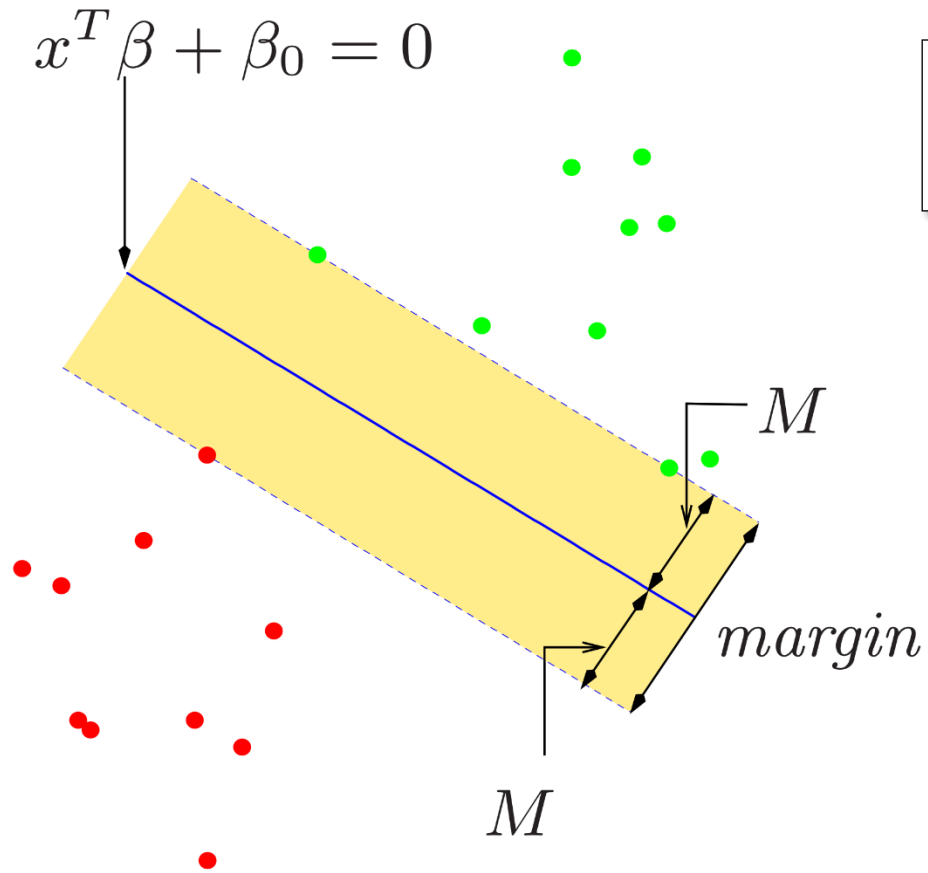
¿Cómo podemos plantear esto como un problema de optimización?



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

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

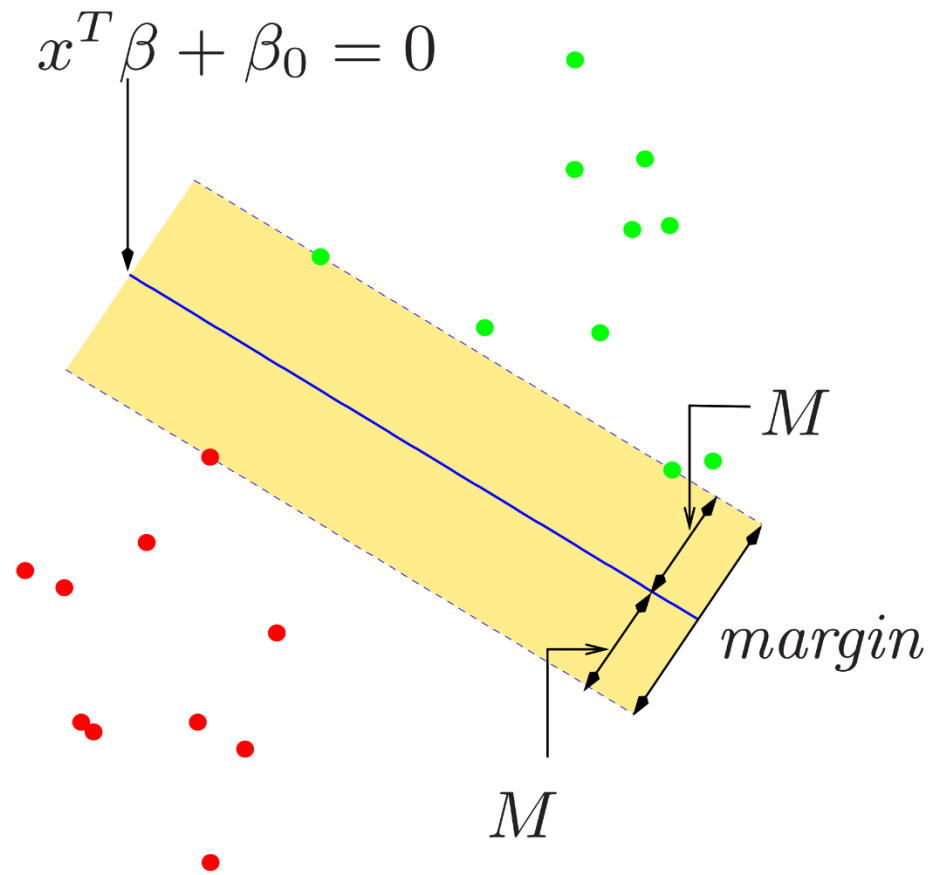
¿Cómo podemos plantear esto como un problema de optimización?



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

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

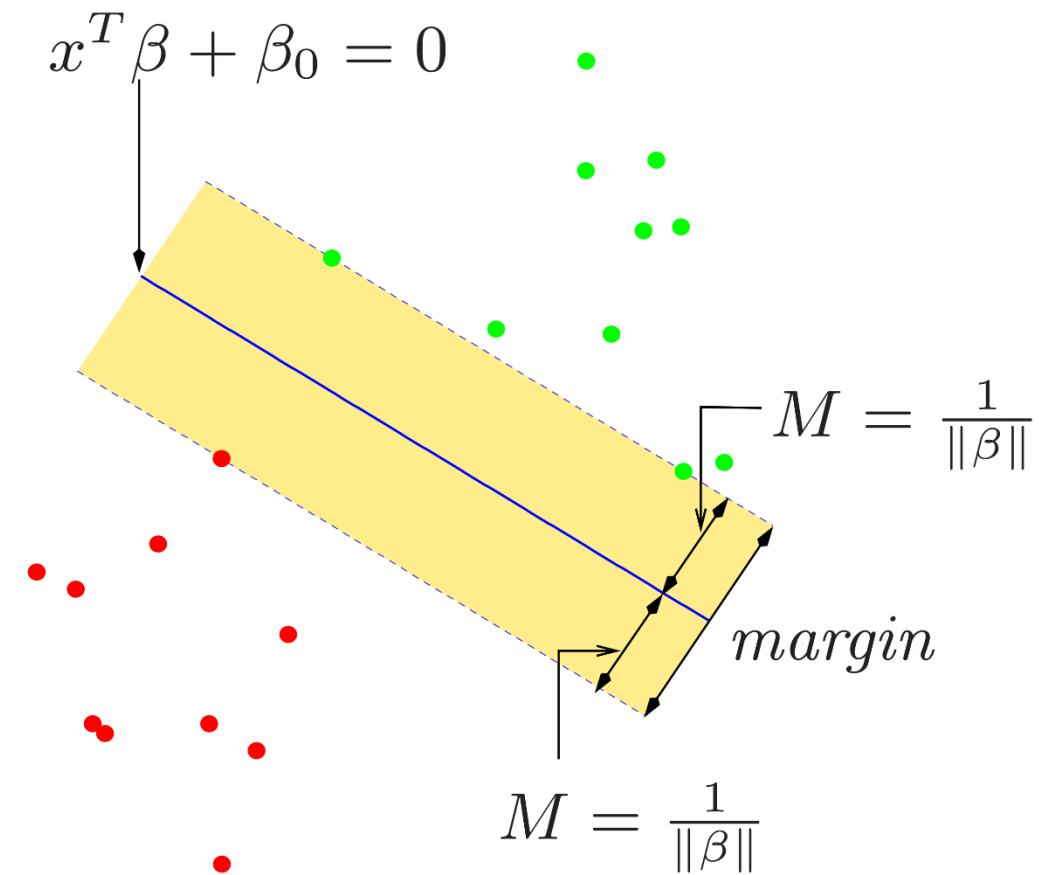
¿Cómo podemos plantear esto como un problema de optimización?



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

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

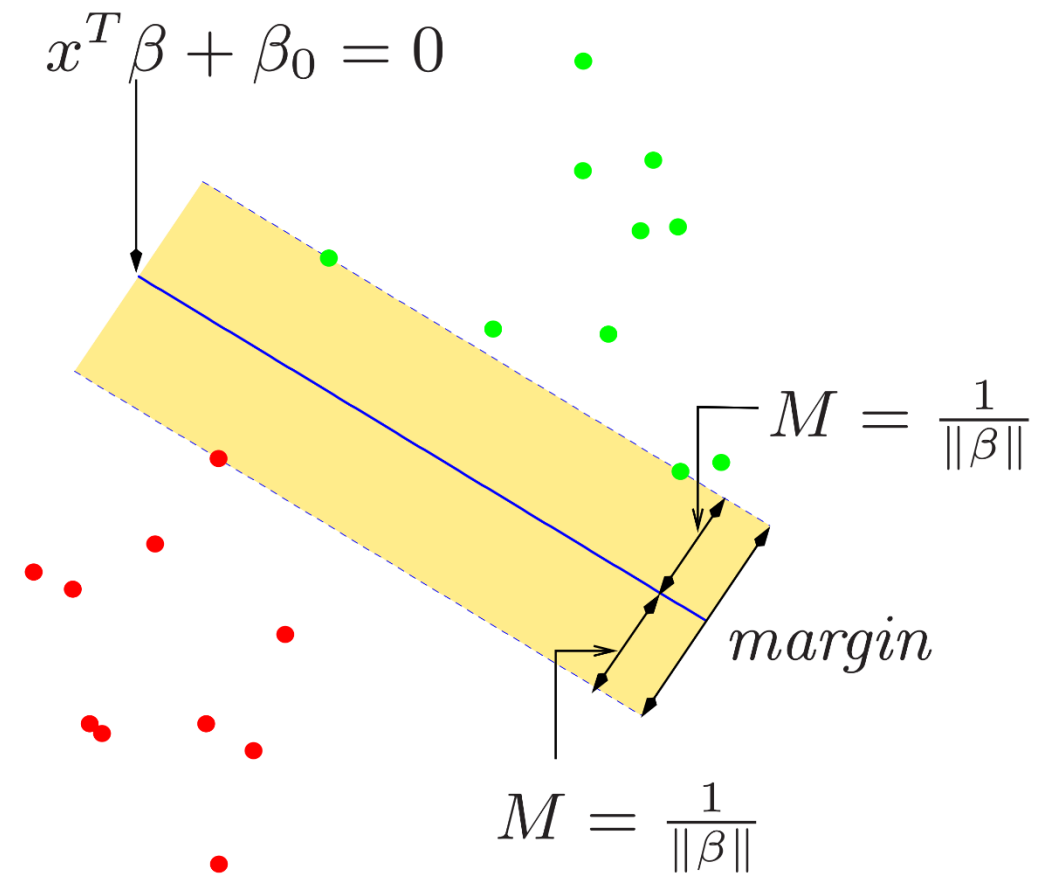
¿Cómo podemos plantear esto como un problema de optimización?



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

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

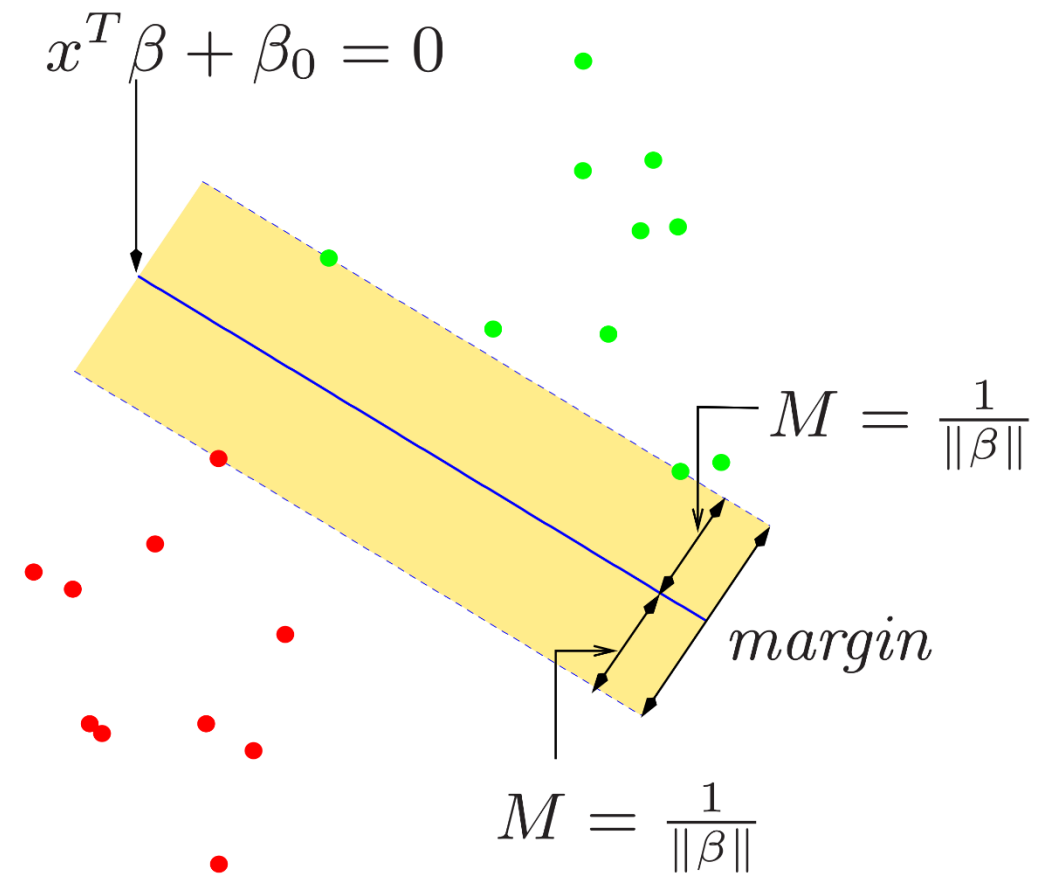
¿Cómo podemos plantear esto como un problema de optimización?



$$\min_{\beta, \beta_0} \|\beta\|$$

subject to $y_i(x_i^T \beta + \beta_0) \geq 1, i = 1, \dots, N$

¿Cómo podemos plantear esto como un problema de optimización?



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

subject to $y_i(x_i^T \beta + \beta_0) \geq 1, i = 1, \dots, N$

Veamos cómo podemos resolver este problema (hard-margin SVM)

$$\begin{aligned} & \min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 \\ & \text{subject to } y_i(x_i^T \beta + \beta_0) \geq 1, \quad i = 1, \dots, N \end{aligned}$$



Lagrangiano

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

Veamos cómo podemos resolver este problema (**hard-margin SVM**)

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

Derivando e igualando a cero, obtenemos:

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

Sustituyendo todo esto en el **lagrangiano**, obtenemos el **dual**:

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

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

Veamos cómo podemos resolver este problema (**hard-margin SVM**)

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{k=1}^N \alpha_i \alpha_k y_i y_k x_i^T x_k \\ \text{subject to } & \alpha_i \geq 0 \text{ and } \sum_{i=1}^N \alpha_i y_i = 0 \end{aligned}$$

Condición de optimalidad (KKT) \rightarrow 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:

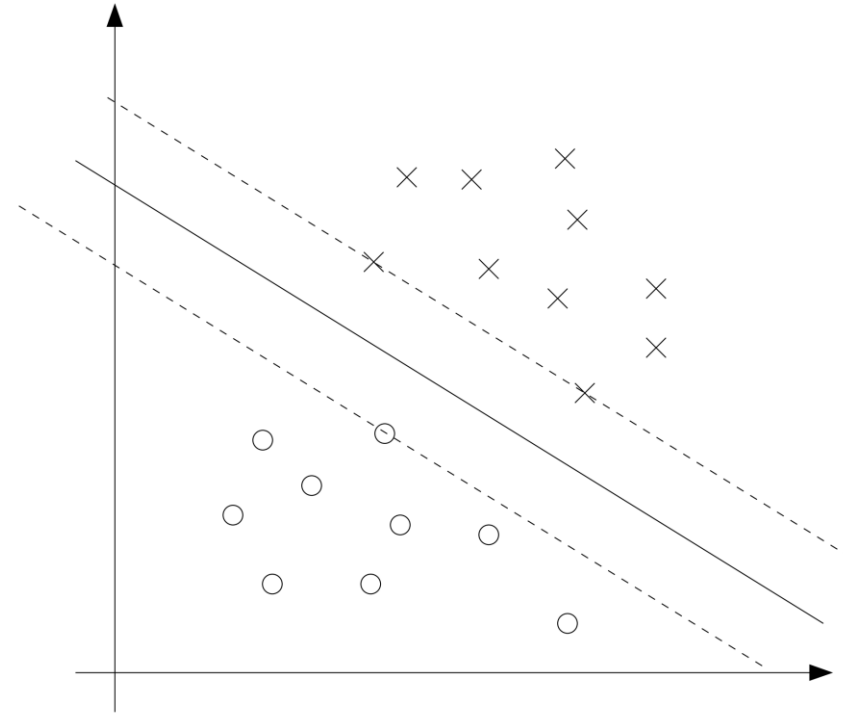
- Si $\alpha_i > 0$, $y_i (x_i^T \beta + \beta_0) = 1$ (el punto queda sobre el límite del margen)
- Si $y_i (x_i^T \beta + \beta_0) > 1$ (punto queda fuera del margen), $\alpha_i = 0$.

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

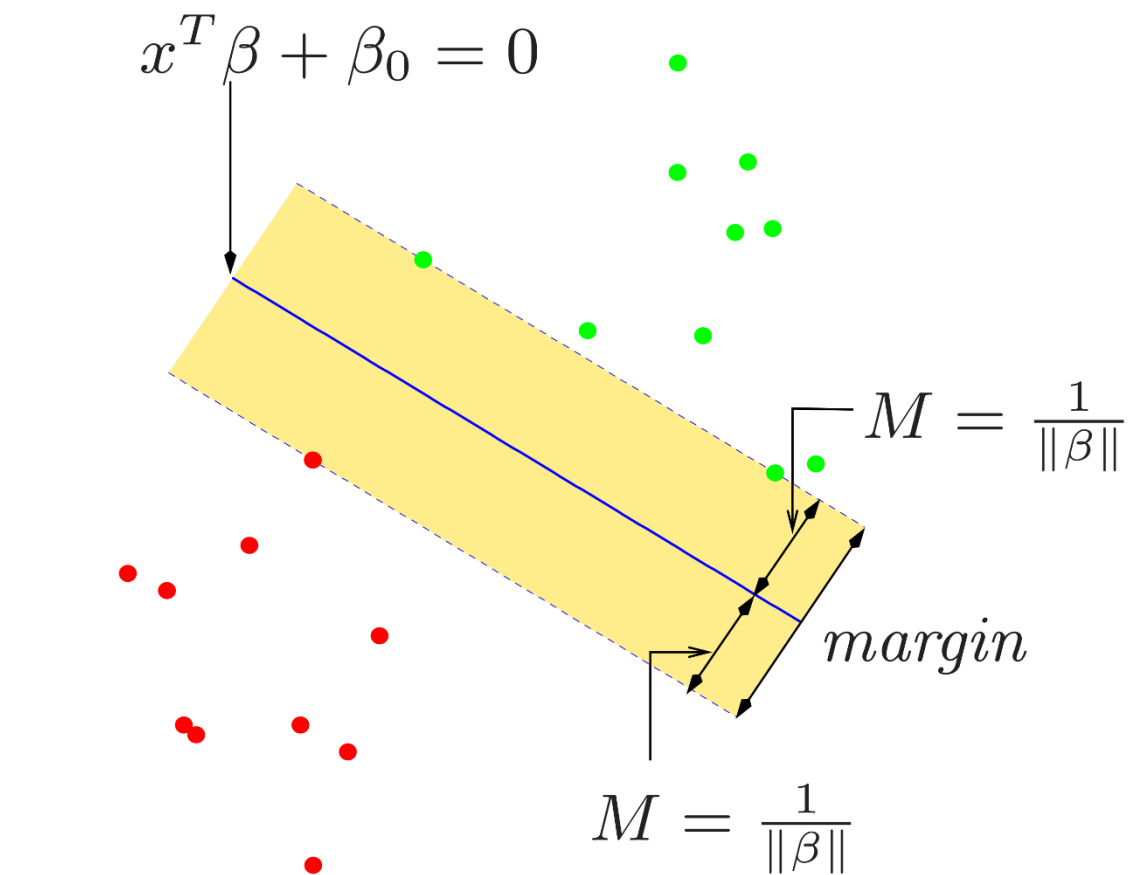
$$\hat{y}_i = \text{sgn}(x_i^T \beta + \beta_0)$$

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

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

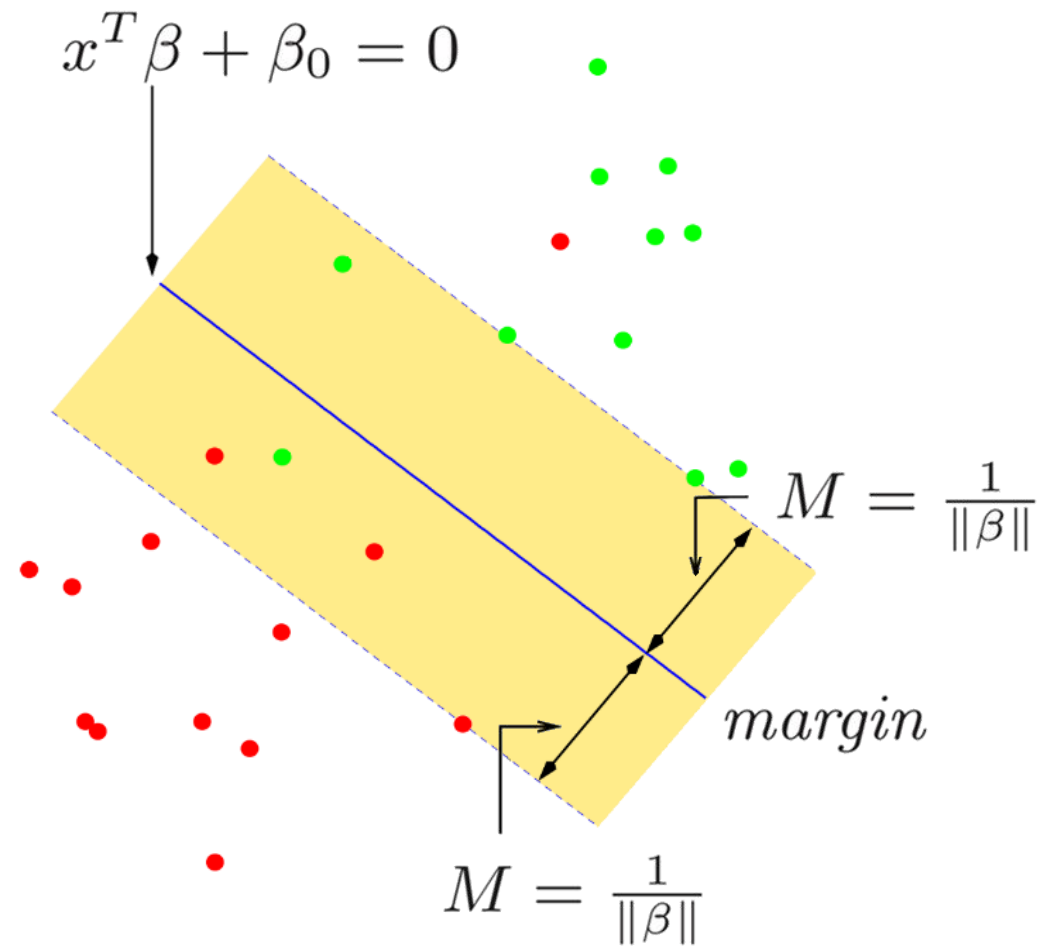


¿Qué pasa con el SVM si tenemos datos no son linealmente separables?



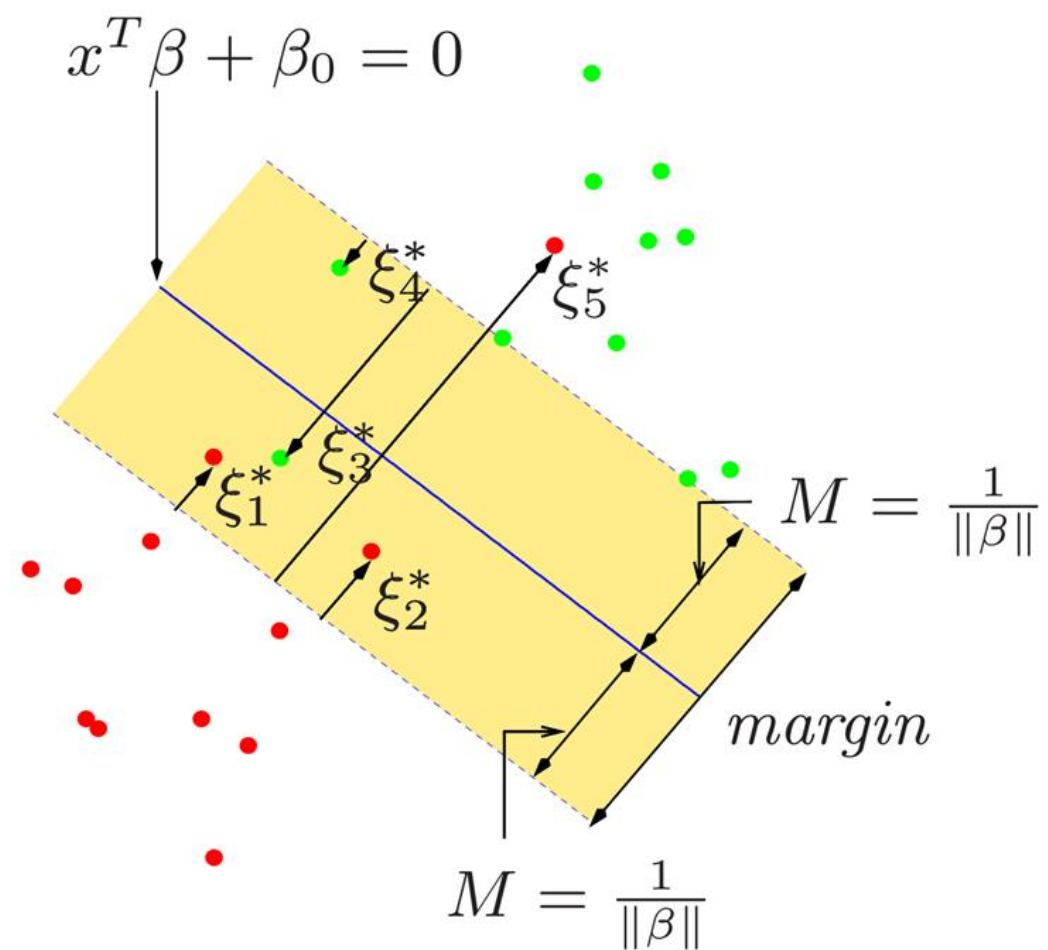
$$\frac{1}{\|\beta\|} y_i (x_i^T \beta + \beta_0) \geq M$$

¿Qué pasa con el SVM si tenemos datos no son linealmente separables?



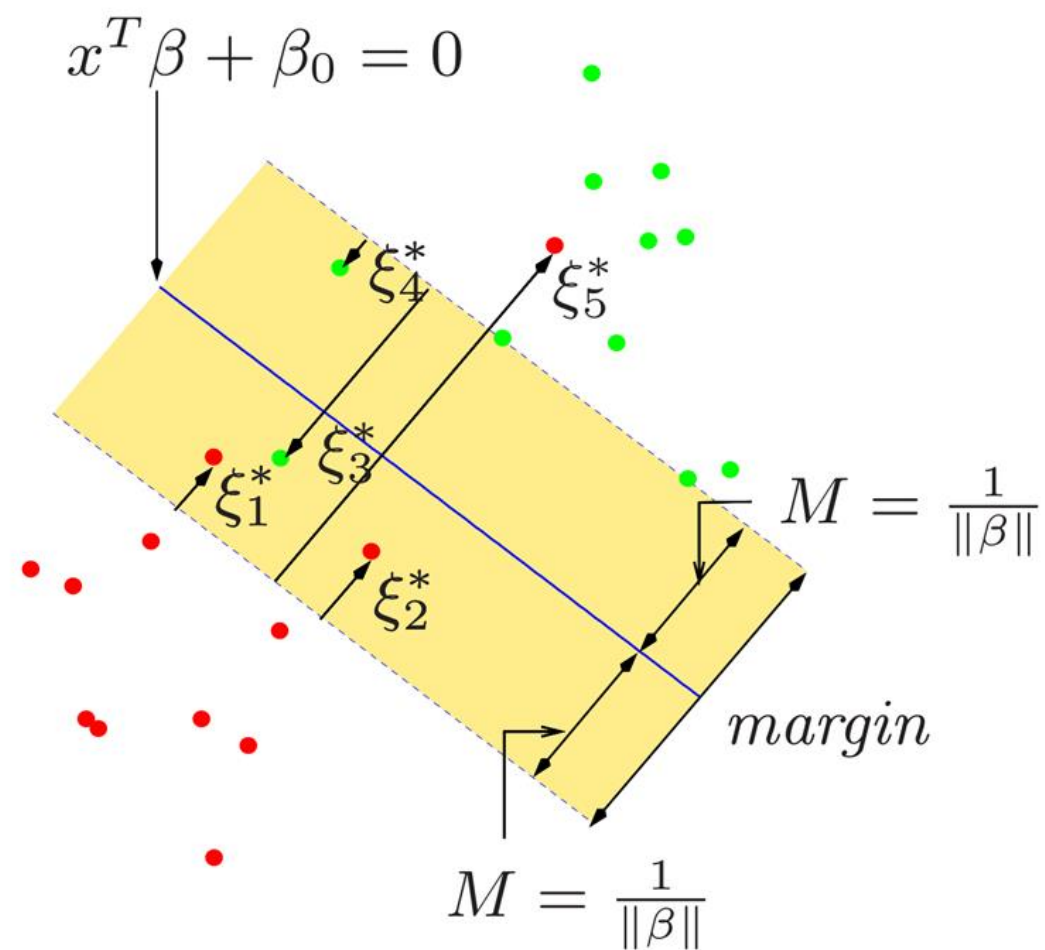
$$\frac{1}{\|\beta\|} y_i (x_i^T \beta + \beta_0) \geq M$$

Soft-margin SVM



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

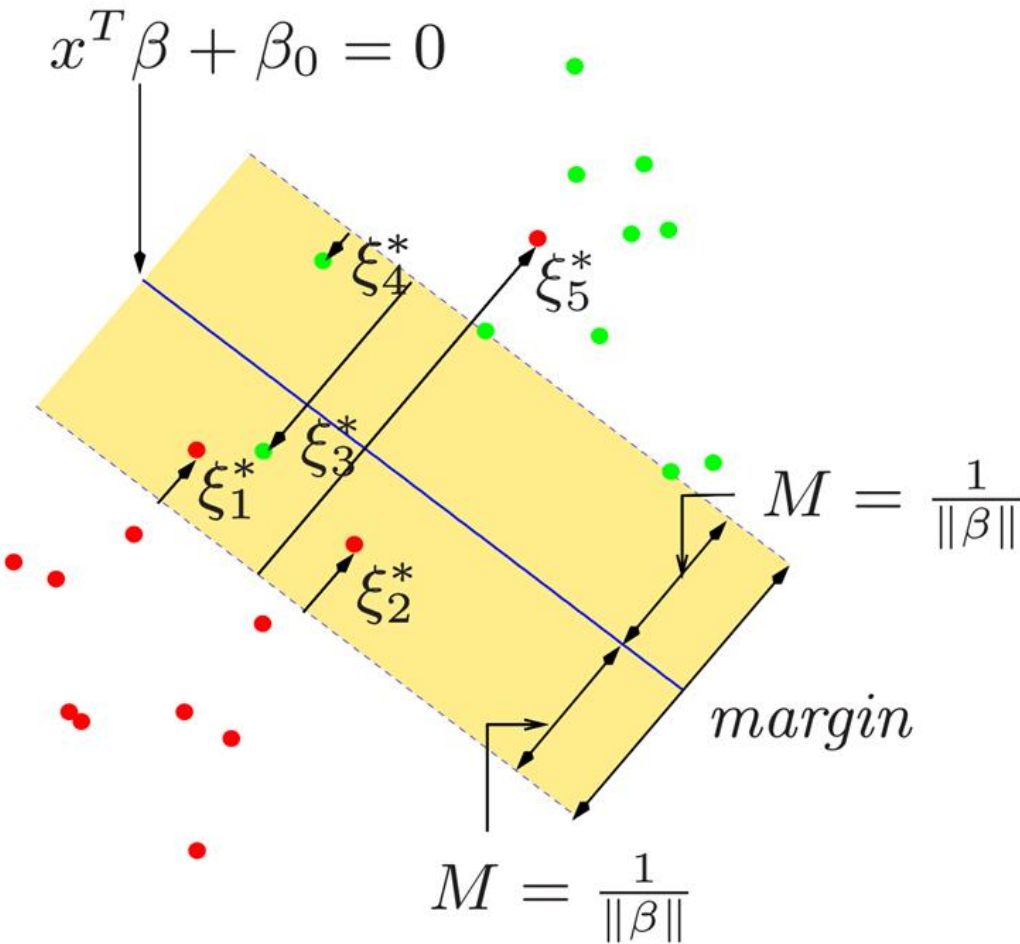
Soft-margin SVM



$$\xi_i^* = M \xi_i$$

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

Soft-margin SVM



$$y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i$$

Soft-margin SVM

$$\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 \quad \text{subject to} \quad \begin{cases} y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i \quad \forall i, \\ \xi_i \geq 0, \quad \sum \xi_i \leq \text{constant}. \end{cases}$$


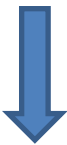


$$\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^N \xi_i$$

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

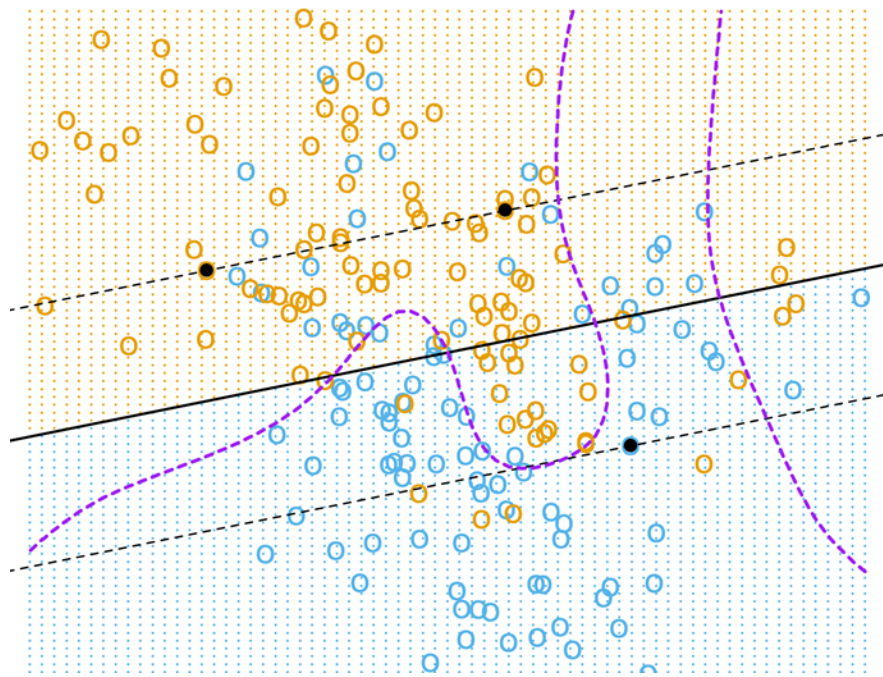
Función objetivo de los SVM calza con lo planteado al inicio del curso

Función de pérdida: captura el rendimiento del modelo al cuantificar su error en los datos de entrenamiento

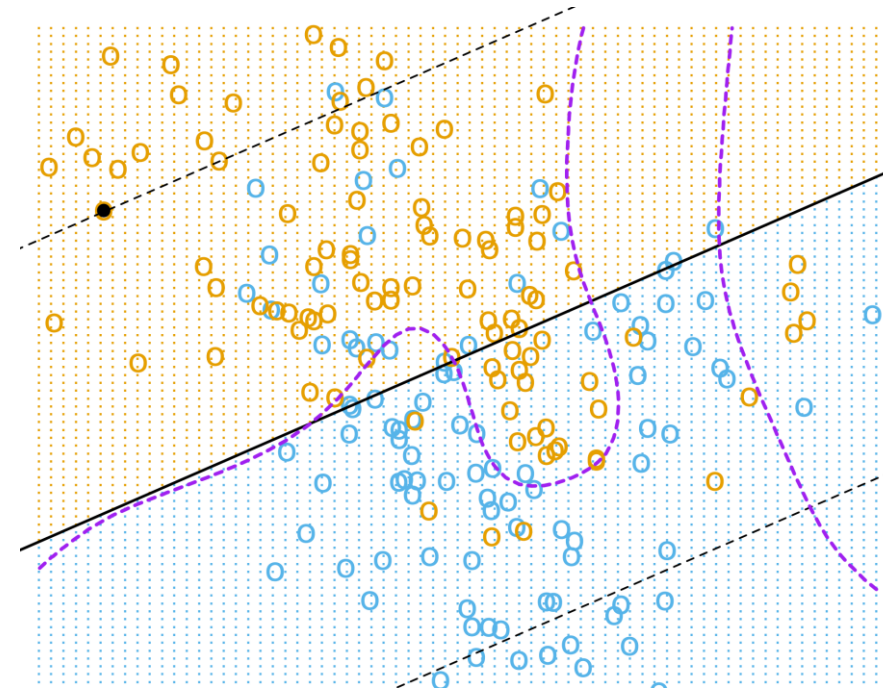
$$\operatorname{argmin}_W J(X, Y; W) = \lambda \mathcal{R}(W) + \sum_i^N \mathcal{L}(f(x_i; W), y_i)$$


Regularizador: función que induce sesgo inductivo en el modelo predictivo a través de sus parámetros

Soft-margin SVM



$C = 1000$



$C = 0.01$

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

Vamos a Colab...



Cuáles son los conceptos centrales de la clase

- Concepto de margen y su relación con la generalización.
- Bases geométricas de SVM
- Problema de aprendizaje como un problema de optimización, una característica del ML moderno.
- Función objetivo con 2 términos: regularización + pérdida.

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



IIC2613 - Inteligencia Artificial

Support Vector Machines (SVM)

Hans Löbel

Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación