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



IIC2613 - Inteligencia Artificial

Support Vector Machines (SVM)

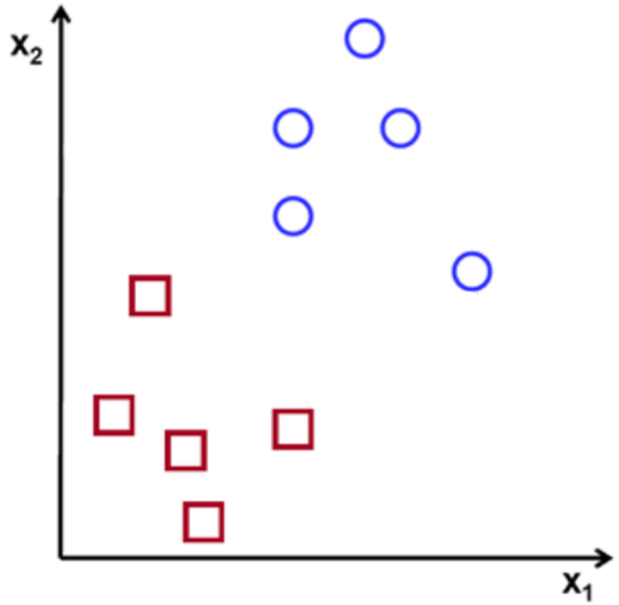
Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación

Recapitulemos nuevamente lo que hemos visto hasta ahora en el curso

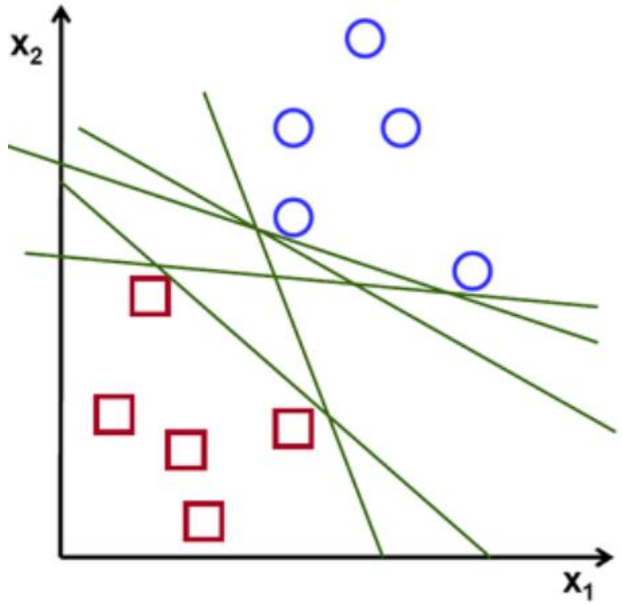
- Comenzamos con los conceptos fundamentales de ML.
- Luego exploramos árboles y ensambles, siempre teniendo en consideración el *trade-off* entre complejidad y sobreajuste.
- Hoy veremos una técnica que fue el estado del arte durante largo tiempo, y captura algunos conceptos centrales de lo que es ML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



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

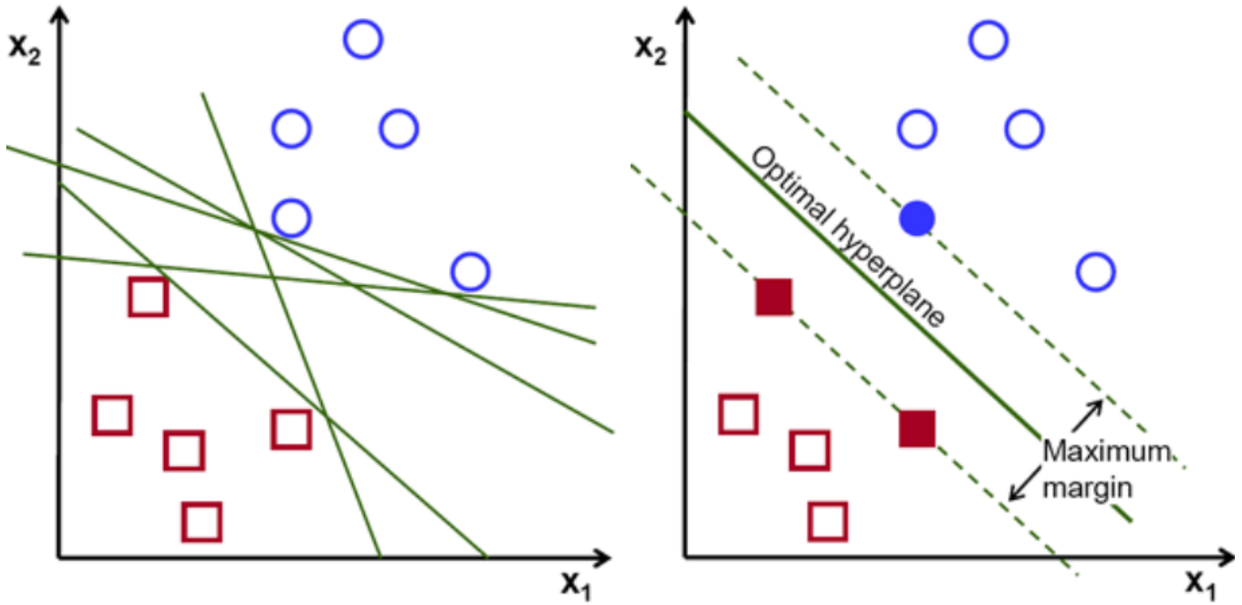


Existen múltiples soluciones, muchas incluso con rendimiento perfecto. Pero ¿existe una “mejor” que el resto?



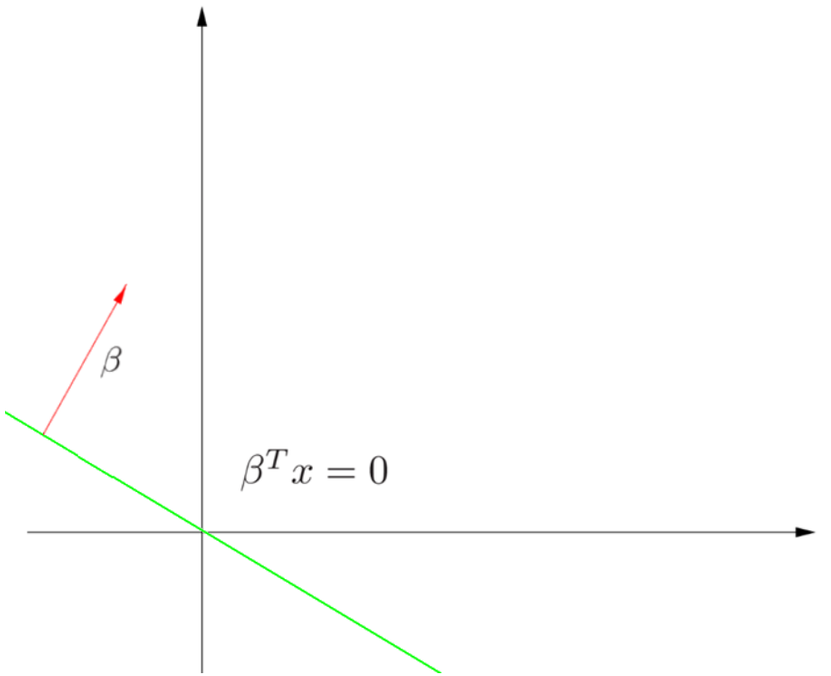
¿Cómo podríamos incorporar algo relacionado con la capacidad de generalización en la solución?

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

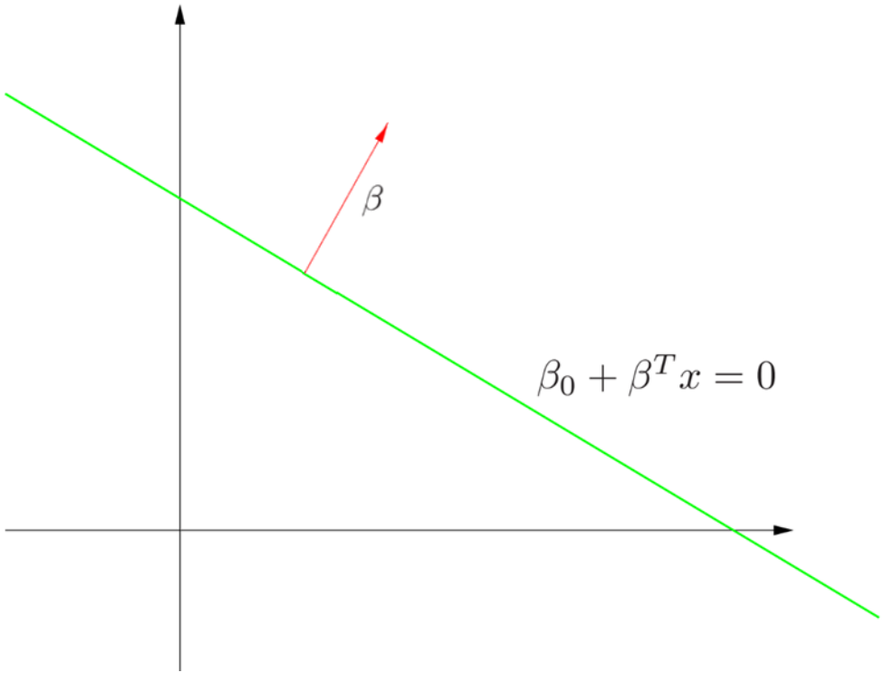


Mientras mayor sea la distancia entre el hiperplano y los puntos, mayor espacio existe para los datos difíciles

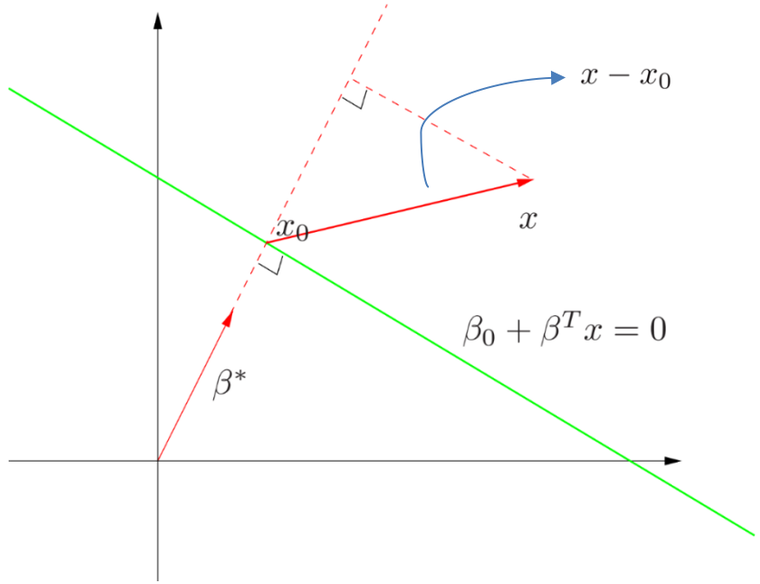
Hagamos un breve repaso del álgebra lineal necesaria para este capítulo



Hagamos un breve repaso del álgebra lineal necesaria para este capítulo



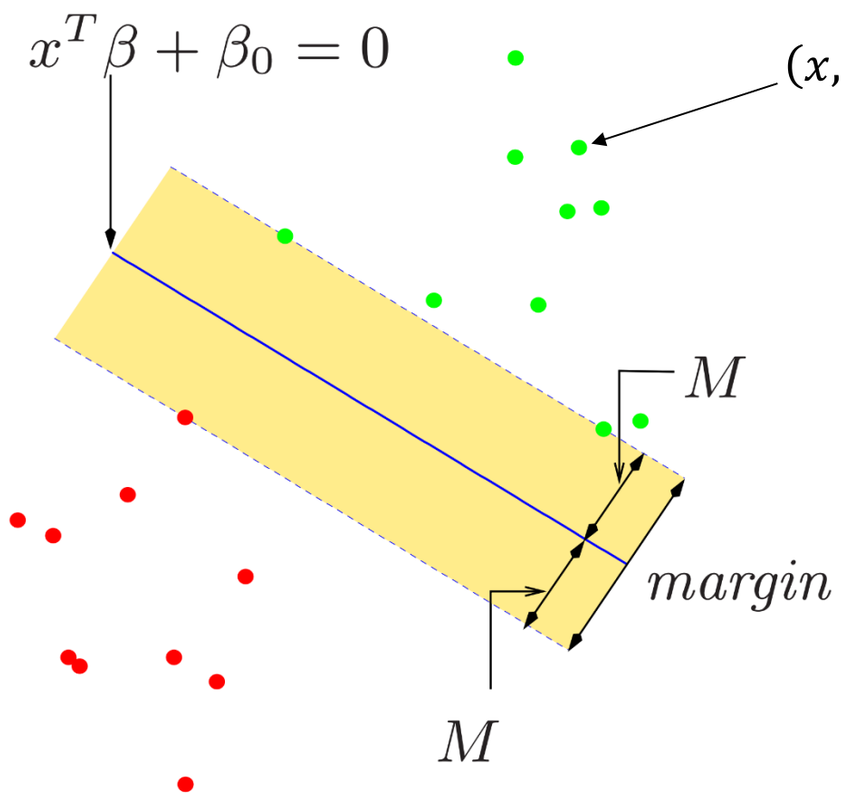
Hagamos un breve repaso del álgebra lineal necesaria para este capítulo



- 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:

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

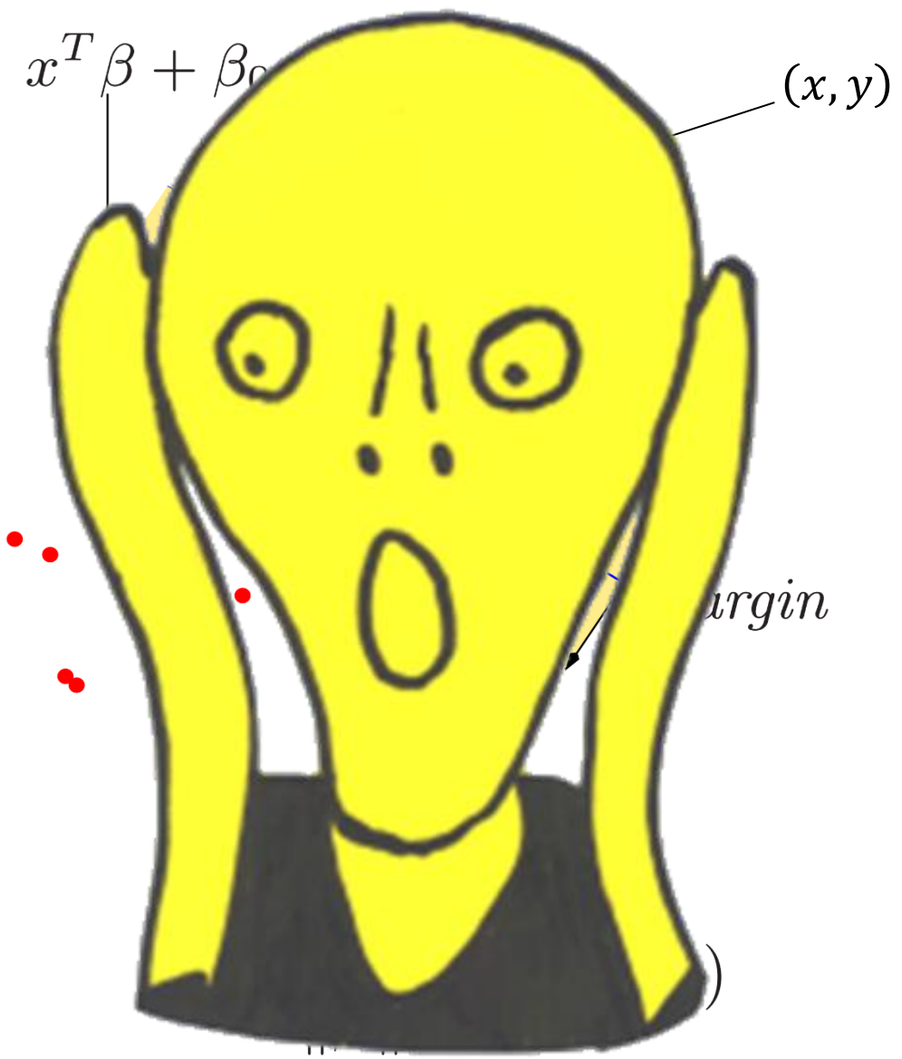
Distancia de un punto al hiperplano (margen) es la clave de los SVM



$$x \in R^2$$
$$y \in \{-1,1\}$$

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

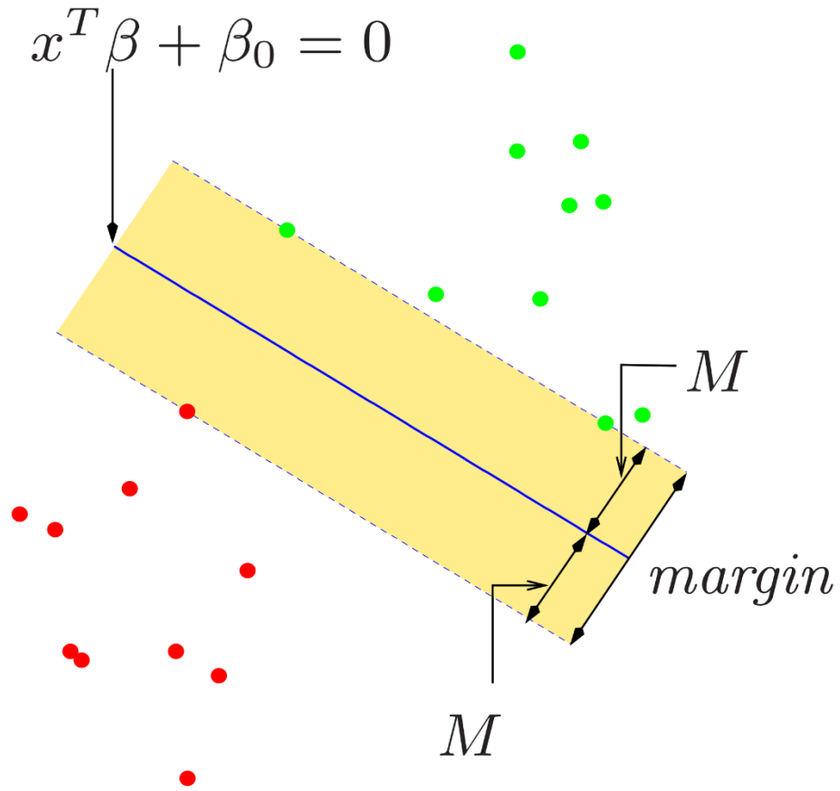
Distancia de un punto al hiperplano (margen) es la clave de los SVM



$$x \in R^2$$
$$y \in \{-1,1\}$$

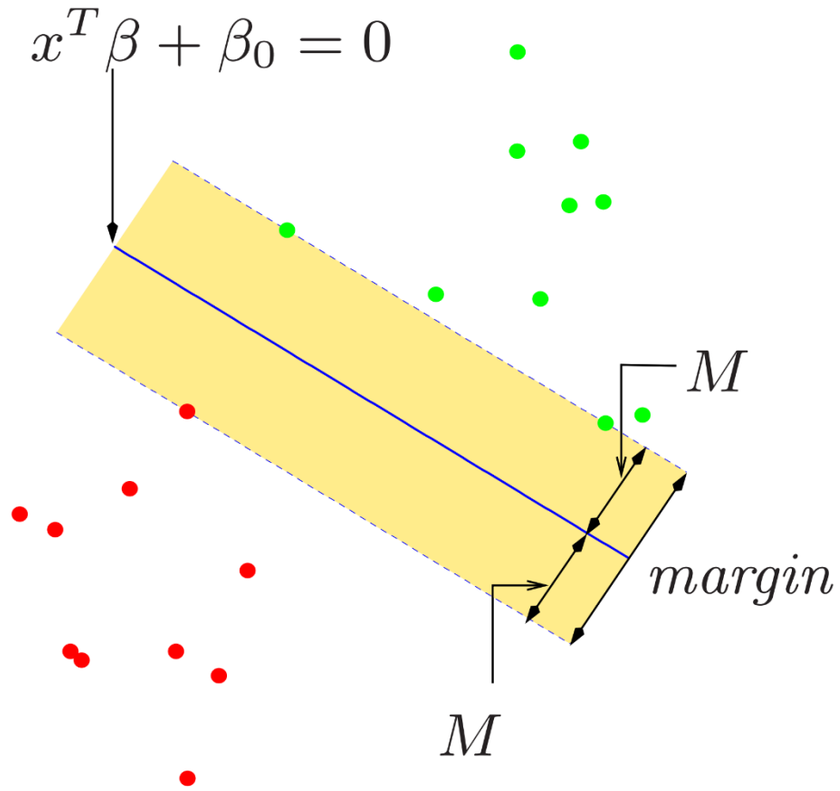
¿Cómo puedo tomar esto y plantear un problema de optimización para encontrar el hiperplano?

Partamos por lo básico, ¿qué es lo que queremos maximizar o minimizar?



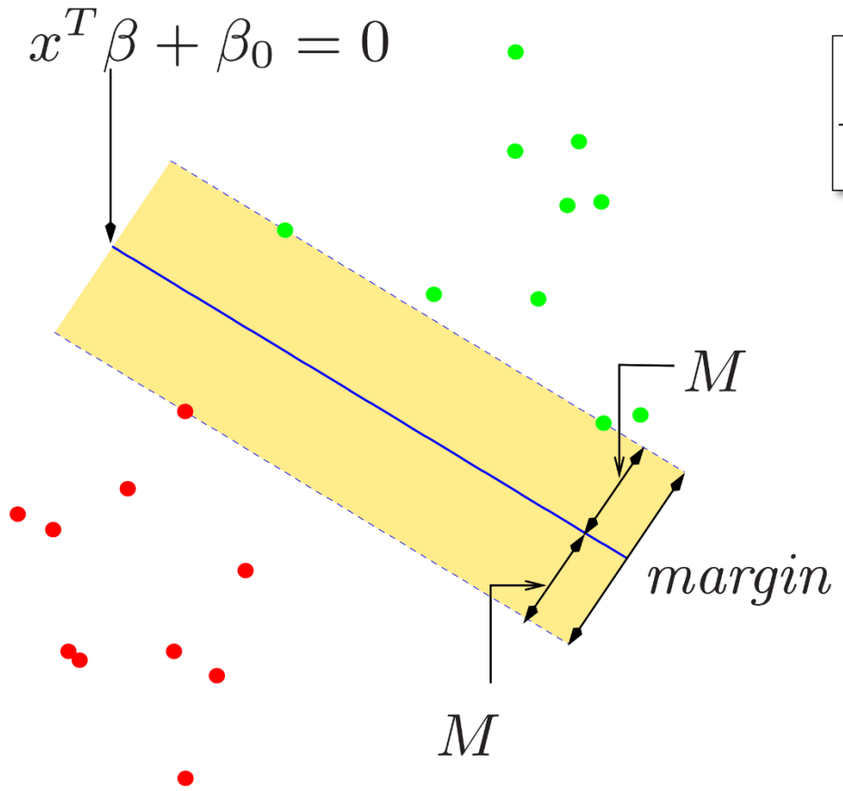
$$\max \quad M$$

A continuación, ¿qué es lo que queremos obtener como resultado?



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

¿Qué condiciones debe cumplir nuestro hiperplano óptimo?

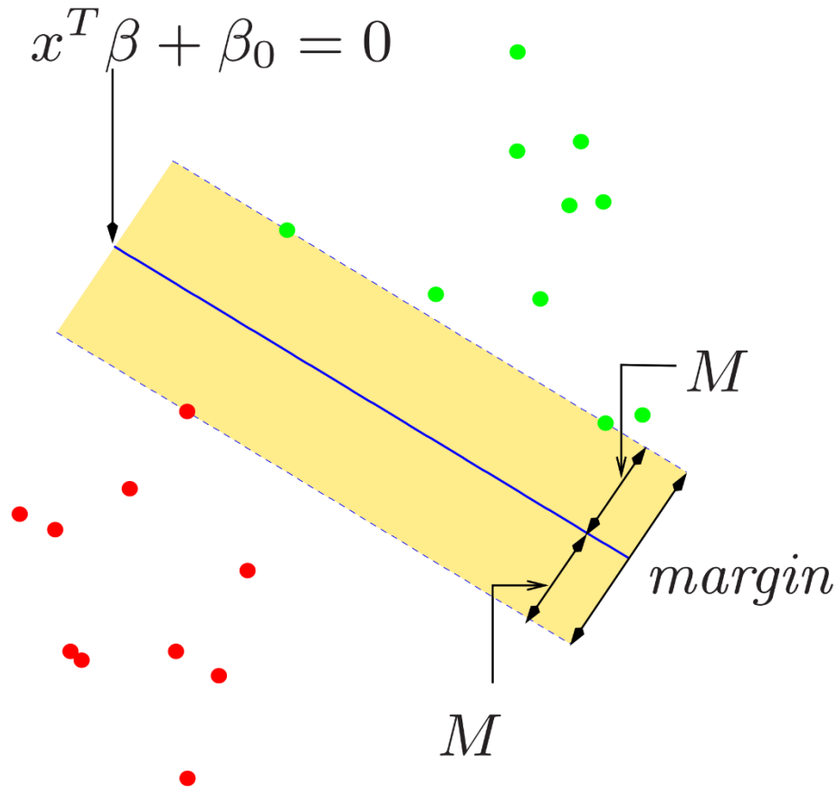


$$D = \frac{y}{\|\beta\|}(x^T\beta + \beta_0)$$

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

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

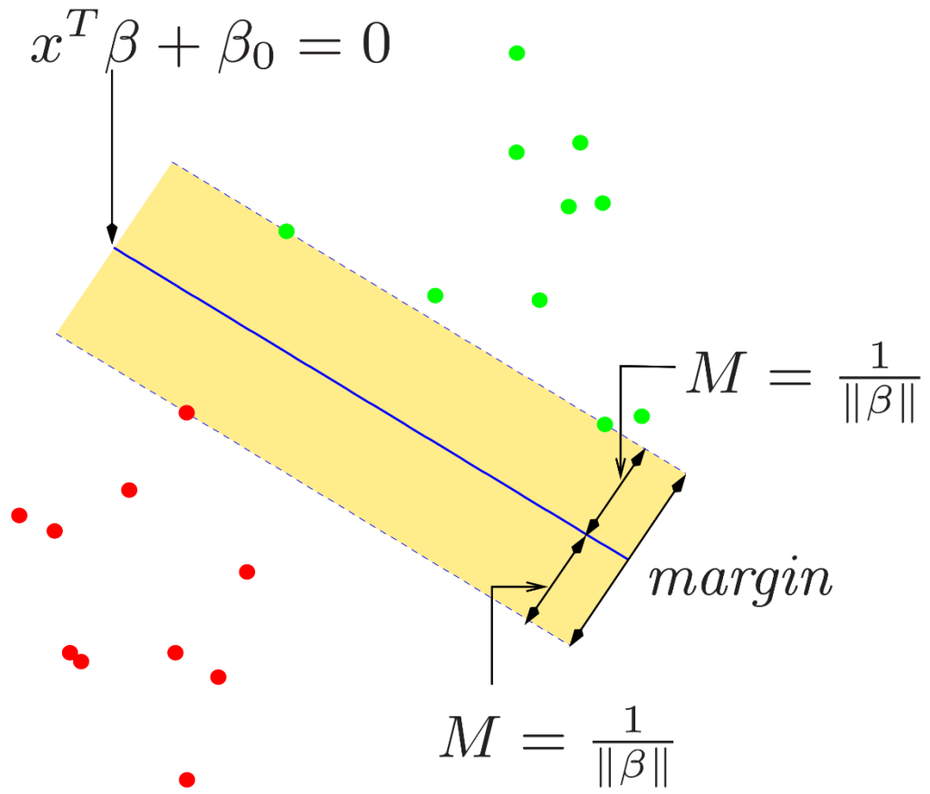
Ahora, acomodaremos los términos para facilitar la resolución



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

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

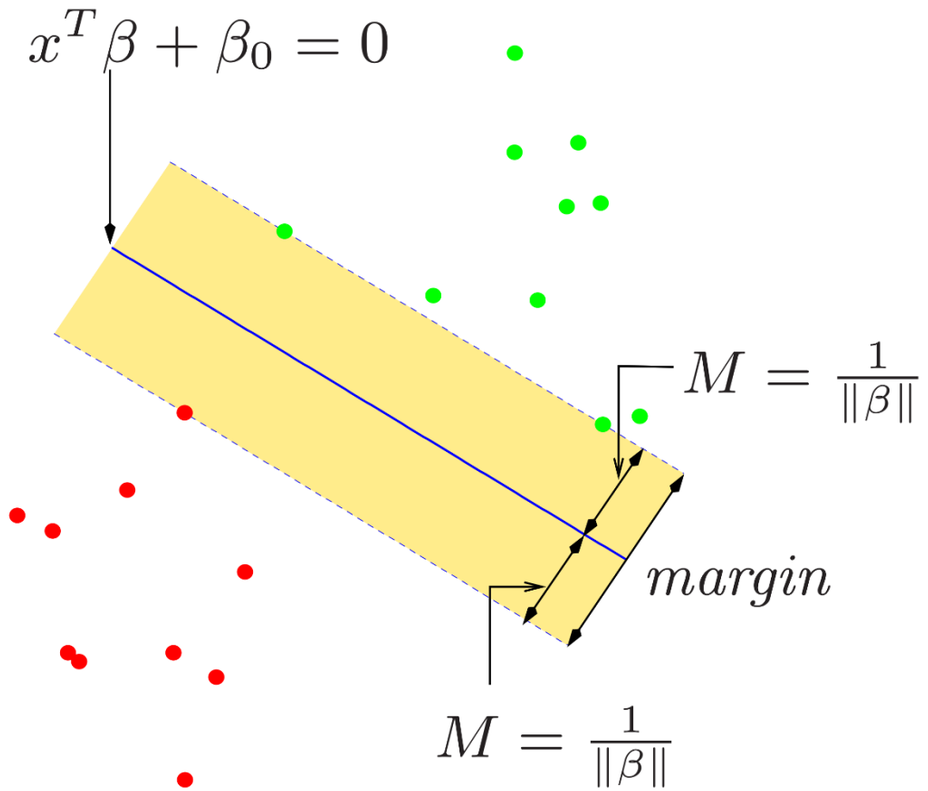
Ahora, acomodaremos los términos para facilitar la resolución



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

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

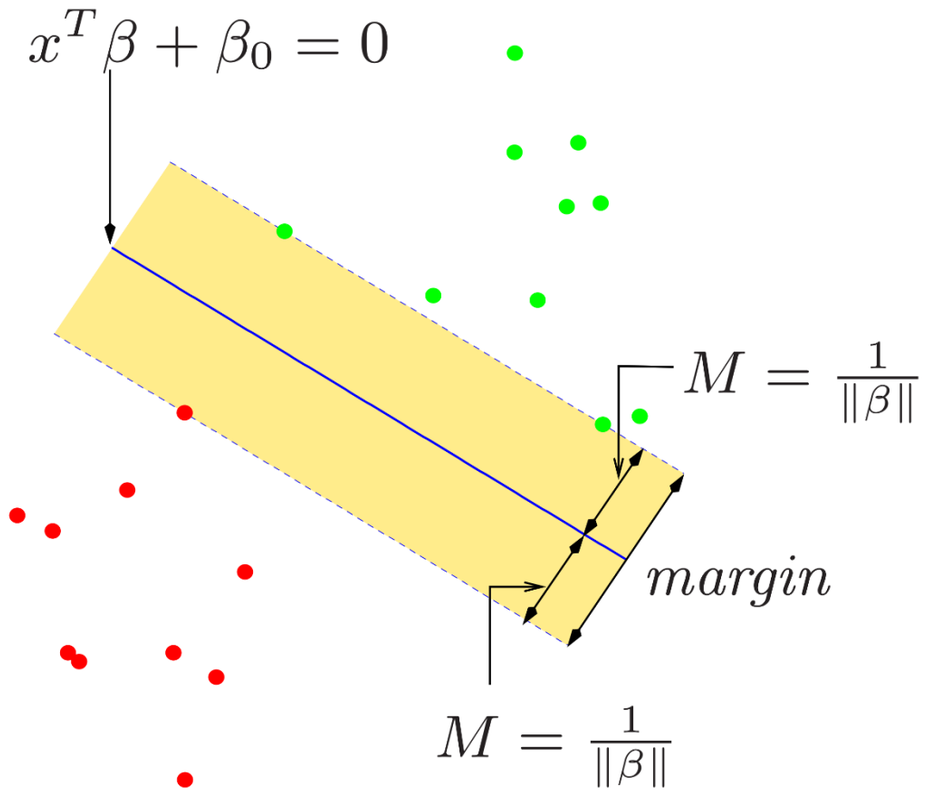
Ahora, acomodaremos los términos para facilitar la resolución



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

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

Ahora, acomodaremos los términos para facilitar la resolución



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

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

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

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

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



Lagrangiano

$$\min_{\beta,\beta_0} \max_{\alpha_i \geq 0} L_P = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{N} \alpha_i[y_i(x_i^T\beta + \beta_0) - 1]$$

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 con respecto a β y β_0, respectivamente, 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 problema dual:

$$\max_{\alpha_i \geq 0} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k$$

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

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

$$\max_{\alpha_i \geq 0} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k$$

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

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

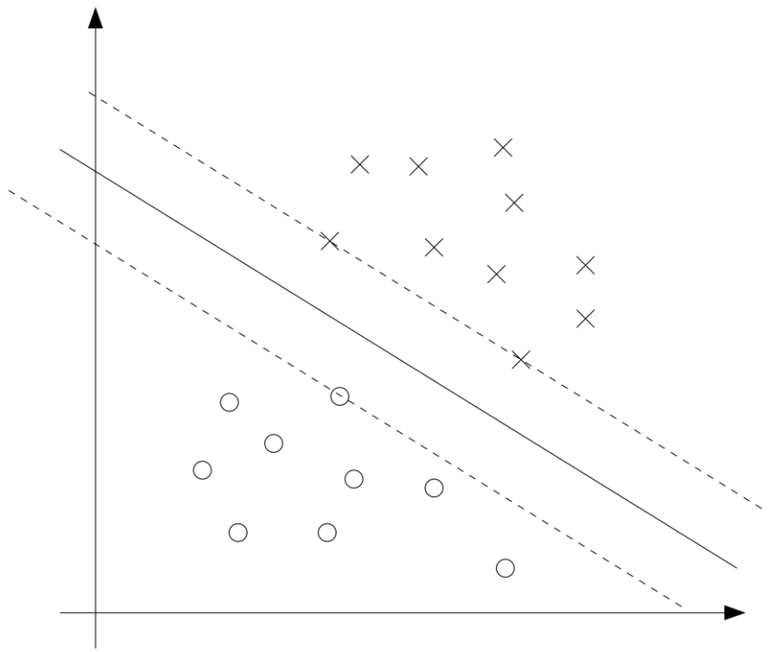
Esta última restricción (KKT) es fundamental para entender los SVM:

- 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

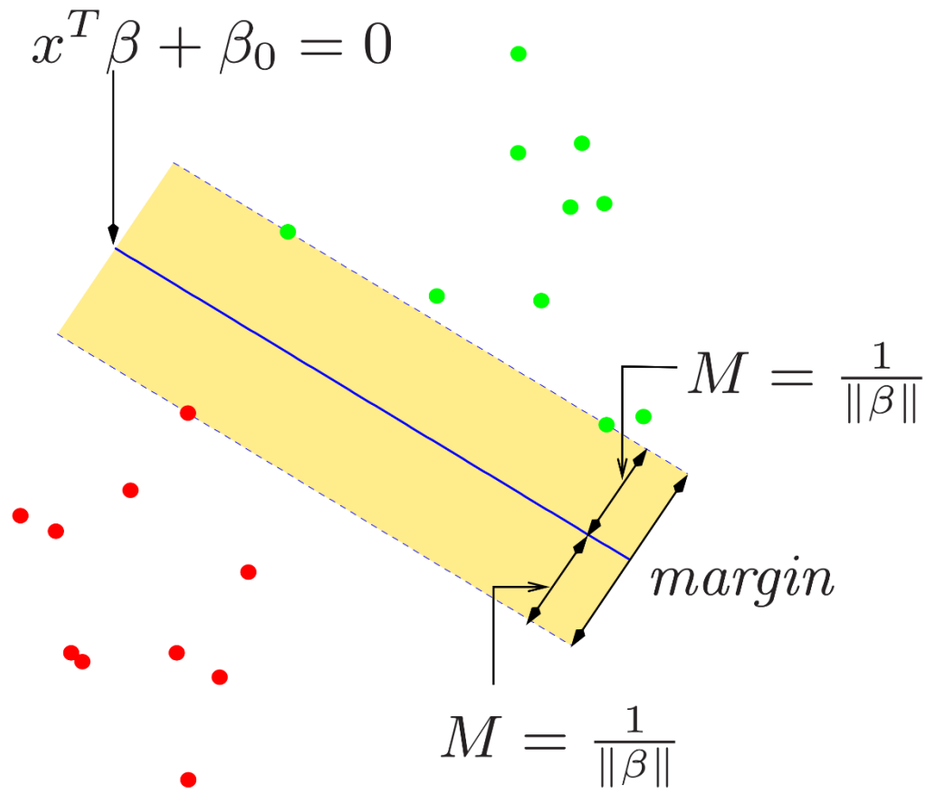
$$\hat{y}_i = \text{sgn}(x_i^T \beta + \beta_0)$$

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



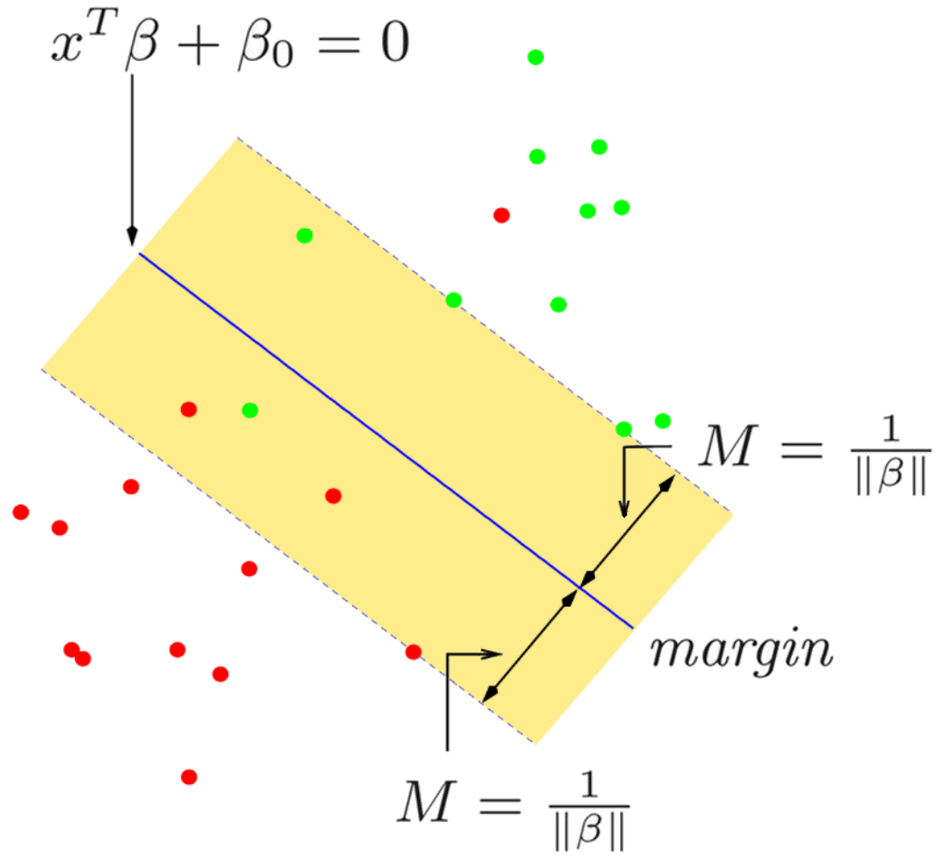
$$\alpha_i [y_i (x_i^T \beta + \beta_0) - 1] = 0 \; \forall i$$

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



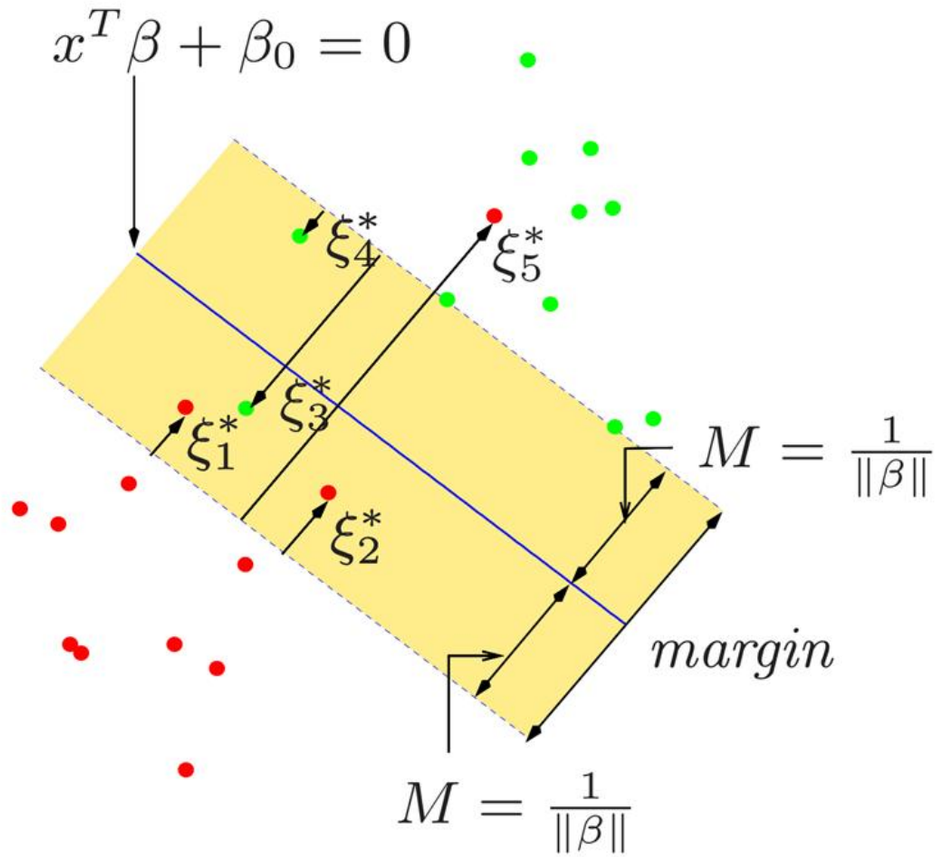
$$\frac{1}{||\beta||} y_i(x_i^T\beta + \beta_0) \quad \geq \quad M$$

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



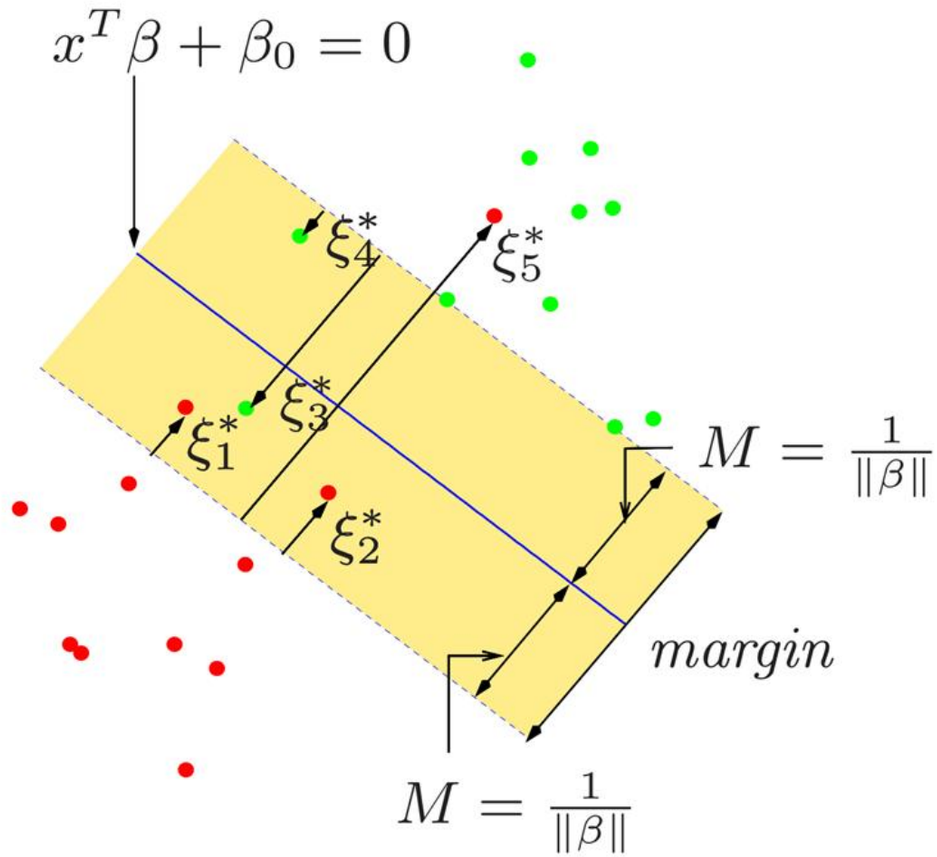
$$\frac{1}{||\beta||} y_i(x_i^T\beta + \beta_0) \quad \geq \quad M$$

Soft-margin SVM



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

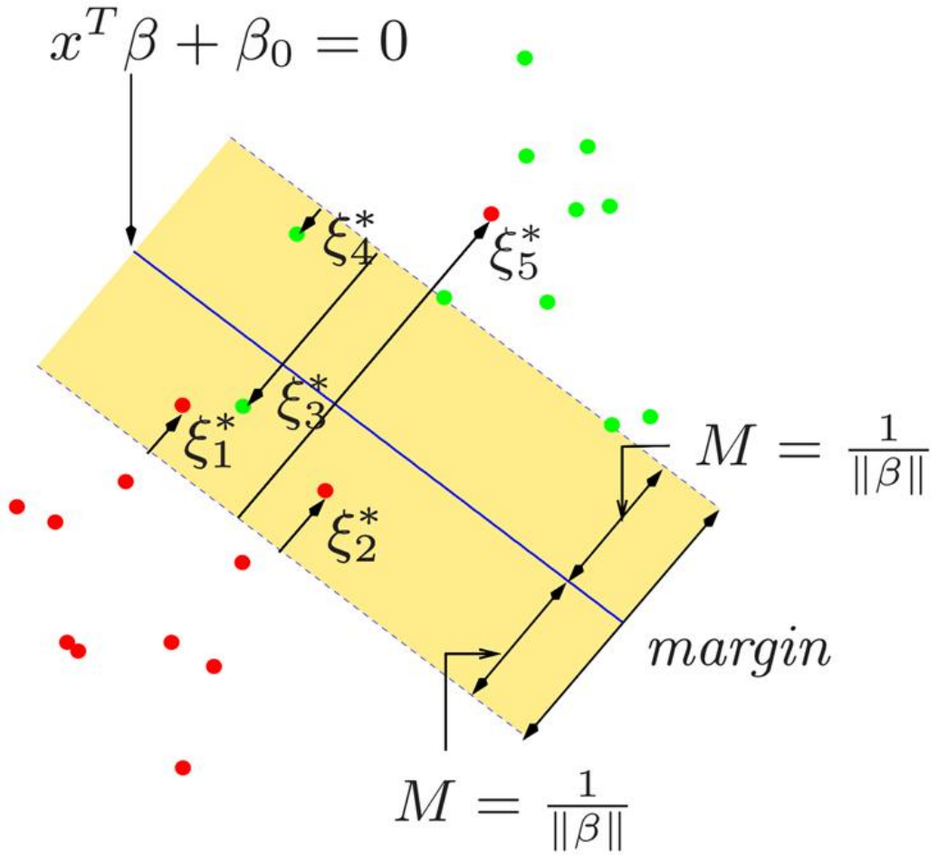
Soft-margin SVM



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

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

Soft-margin SVM



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

Soft-margin SVM

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



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

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

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

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

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

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

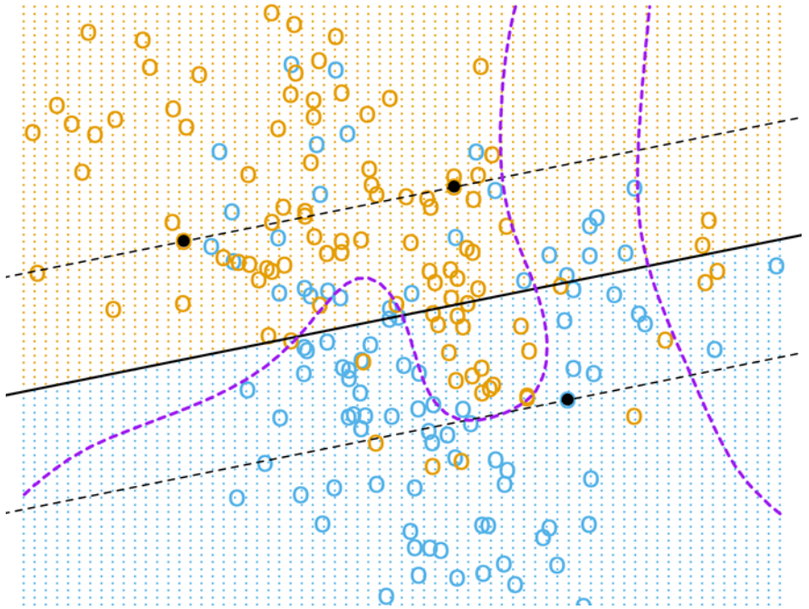
Regularizador: función que induce sesgo inductivo en el modelo predictivo a través de sus parámetros

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

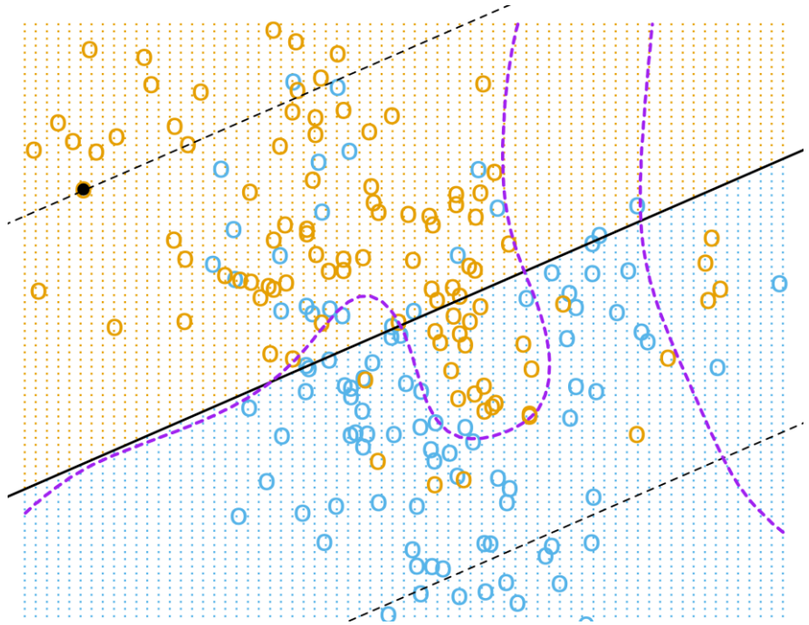
$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$

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

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?

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