Support Vector Machines (SVMs) - Parte I

Aprendizaje Automático INF-398 II-2021

Ricardo Ñanculef

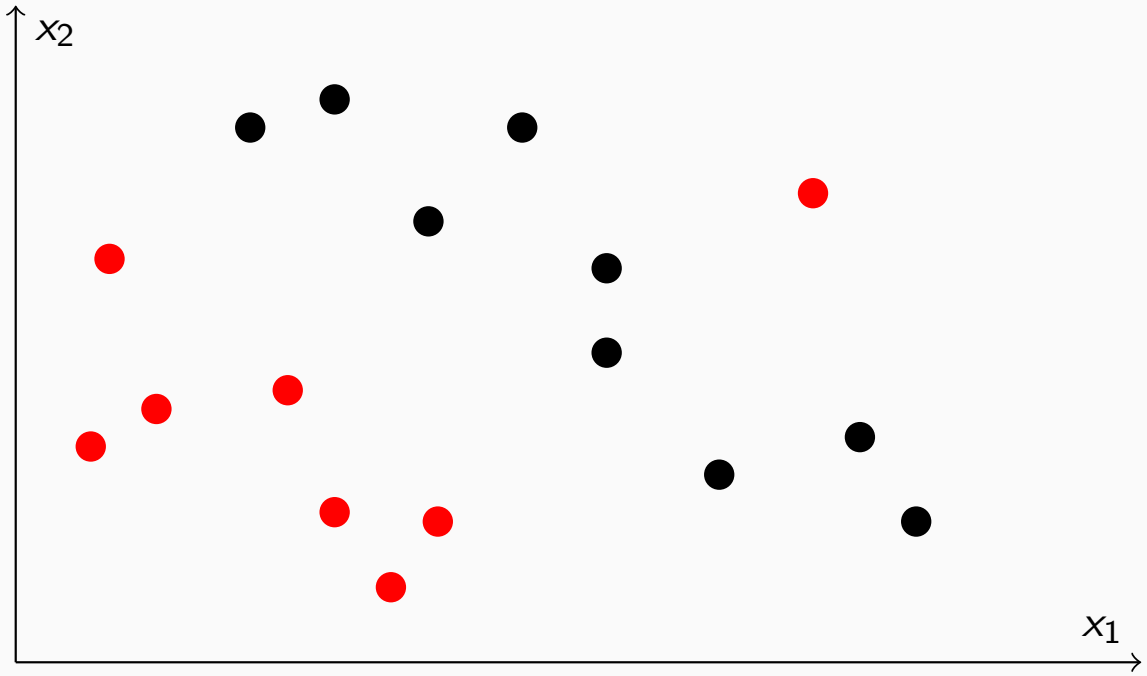
UTFSM Campus San Joaquín

Table of contents

Introducción

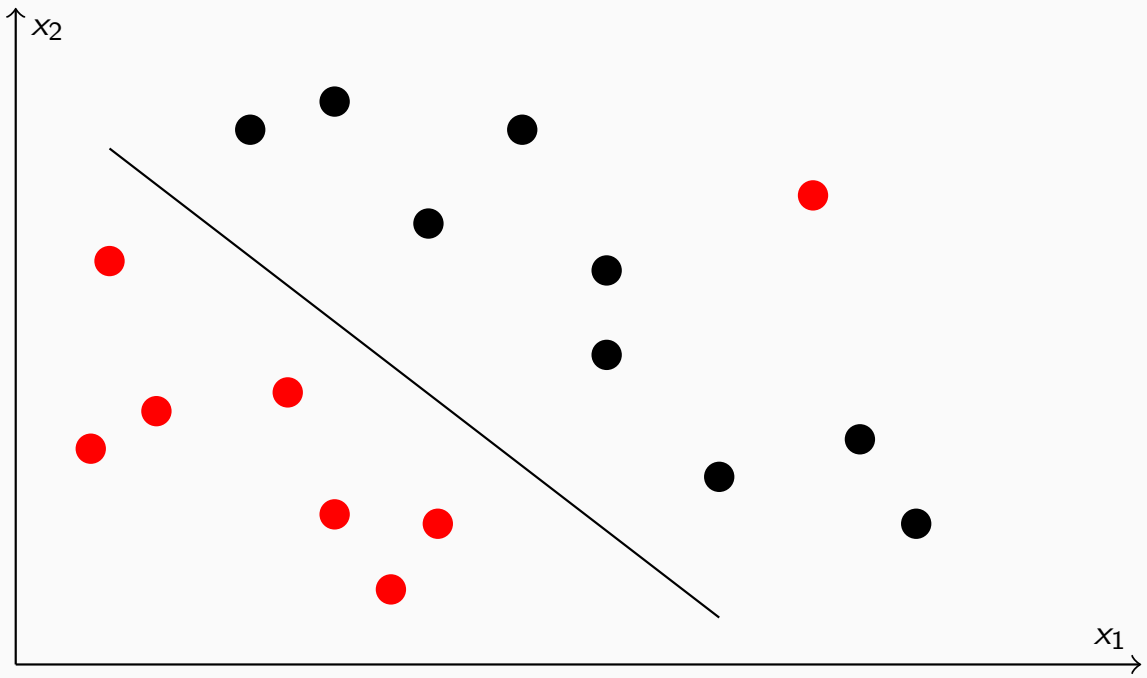
Clasificación (Caso Binario, Supervisado)

Consideremos un problema de clasificación con datos representados en $\mathbb{X} \subset \mathbb{R}^d$ y categorías codificadas como ± 1, de manera que el conjunto de ejemplos es de la forma $S = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^{n}$ con $x^{(\ell)} \in \mathbb{X}, y^{(\ell)} \in \{\pm 1\}$.



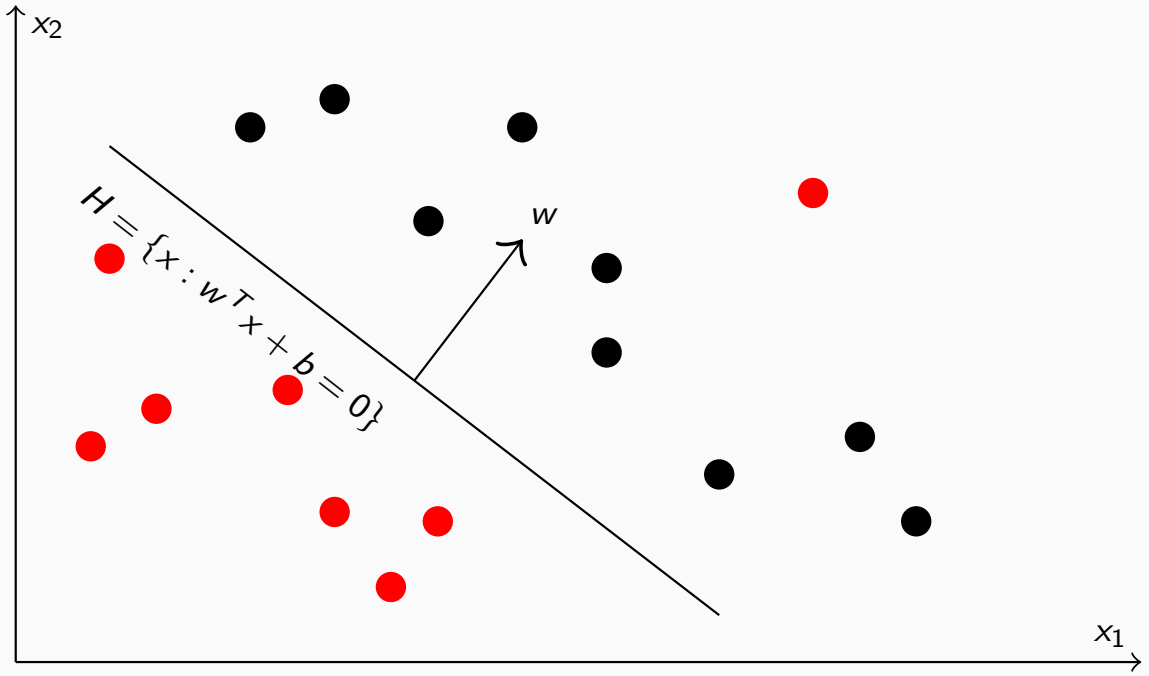
Hiperplanos Separadores

Hemos visto que una forma de implementar un clasificador $f : \mathbb{X} \to \{\pm 1\}$ para este problema consiste aproximar la frontera entre las dos clases mediante una función lineal de la forma $h(x) = w^T x + b = \sum_i w_i x_i + b$, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$.



Hiperplanos Separadores

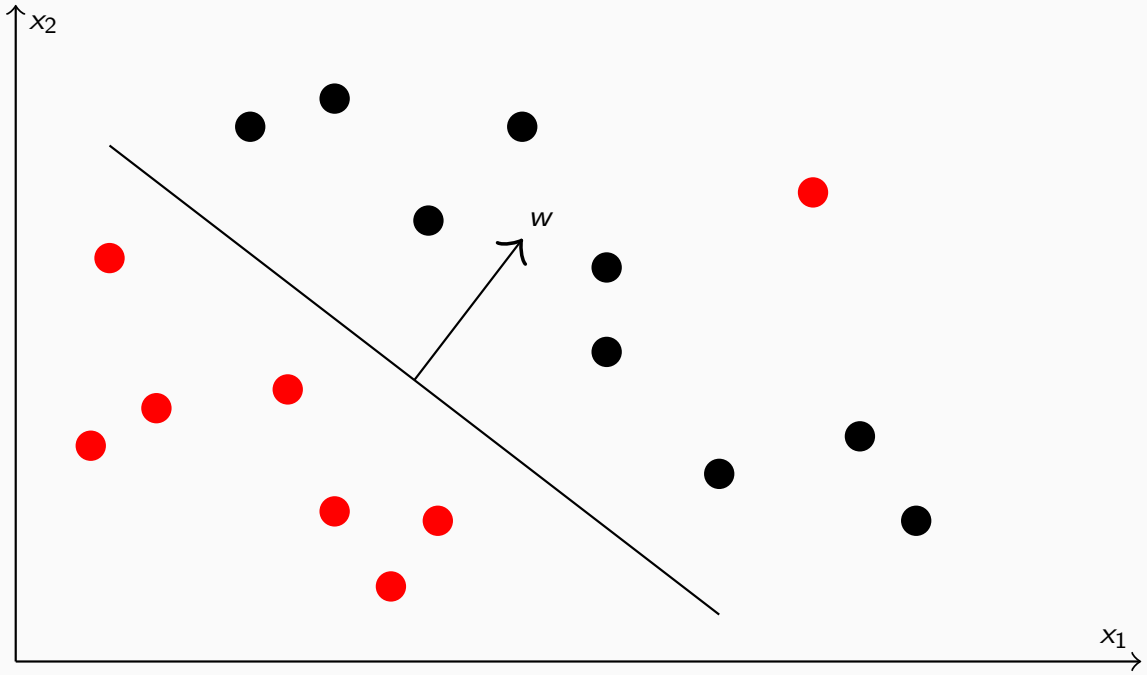
Dados ciertos valores para $w \in \mathbb{R}^d$ y $b \in \mathbb{R}$, el conjunto de puntos $\mathcal{H} = \{x \in \mathbb{X} : h(x) = 0\}$ corresponde a la frontera definida por $h(x)$.



Lados del Hiperplano

Si un punto $x \notin \mathcal{H}$, hay dos posibilidades:

$$w^T x + b > 0 \;\Rightarrow\; x \text{ está en el "lado positivo" de } \mathcal{H}. \tag{1}$$
$$w^T x + b < 0 \;\Rightarrow\; x \text{ está en el "lado negativo" de } \mathcal{H}$$



Clasificación vía Hiperplanos

Idea: Codificando las categorías como ± 1, podemos entonces usar los lados del hiperplano para identificar las clases.

Regla de Clasificación: Si no conocemos la clase de un punto $x \in \mathbb{X}$, basta evaluar $f(x) = w^T x + b$. Si $f(x) > 0$ la clase que corresponde a x es $+1$. En otro caso, la clase asignada a x es -1. El clasificador toma entonces la forma

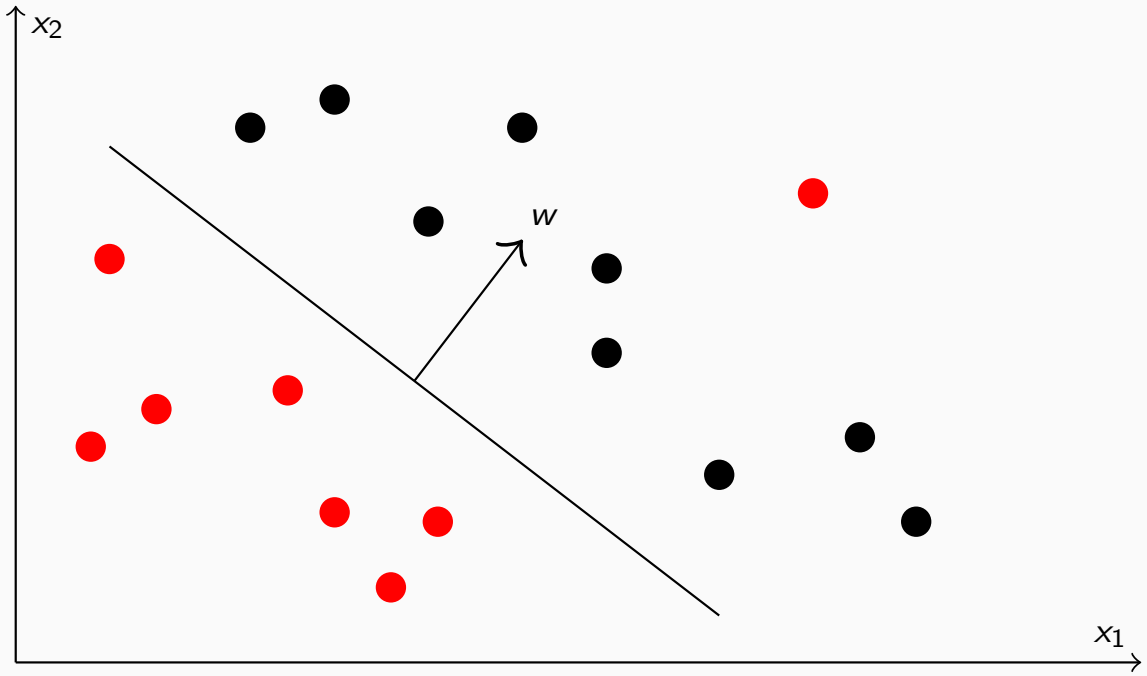
$$f(x) = \text{sign}\left(h(x)\right) = \text{sign}\left(w^T x + b\right) \tag{2}$$

Aprendizaje: Naturalmente, no conocemos los valores "correctos" de w y b. Debemos aprenderlos usando el conjunto de ejemplos $S = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^{n}$.

Parámetros Libres (Aprendibles)

La forma de la frontera estará controlada por los valores de w y b.

- El vector $w \in \mathbb{R}^d$, determina la orientación de la frontera (la superficie definida por $\mathcal{H}$ es ortogonal a w).
- El escalar $b \in \mathbb{R}$, determina la distancia del hiperplano al origen.



Nota: Con frecuencia diremos "El hiperplano (w, b)" para abreviar la expresión "El hiperplano $\mathcal{H}$ definido por los parámetros (w, b)".

Parámetros Libres (Aprendibles)

A partir de la regla de clasificación, podemos interpretar el significado de los parámetros libres

$$f(x) = \text{sign}\left(w^T x + b\right) = \text{sign}\left(\sum_i w_i x_i + b\right) .$$

- Si $w_i > 0$, un aumento en el valor del atributo x_i aumenta la probabilidad de que x sea asignado a la clase positiva.
- Si $w_i < 0$, un aumento en el valor del atributo x_i aumenta la probabilidad de que x sea asignado a la clase negativa.
- Si $w_i = 0$, cambios en el valor del atributo x_i no son relevantes para clasificar x.
- El valor (con signo) del parámetro b cambia la "probabilidad a priori" de clasificar un dato en la clase positiva o negativa.

¿Cómo aprender el hiperplano separador?

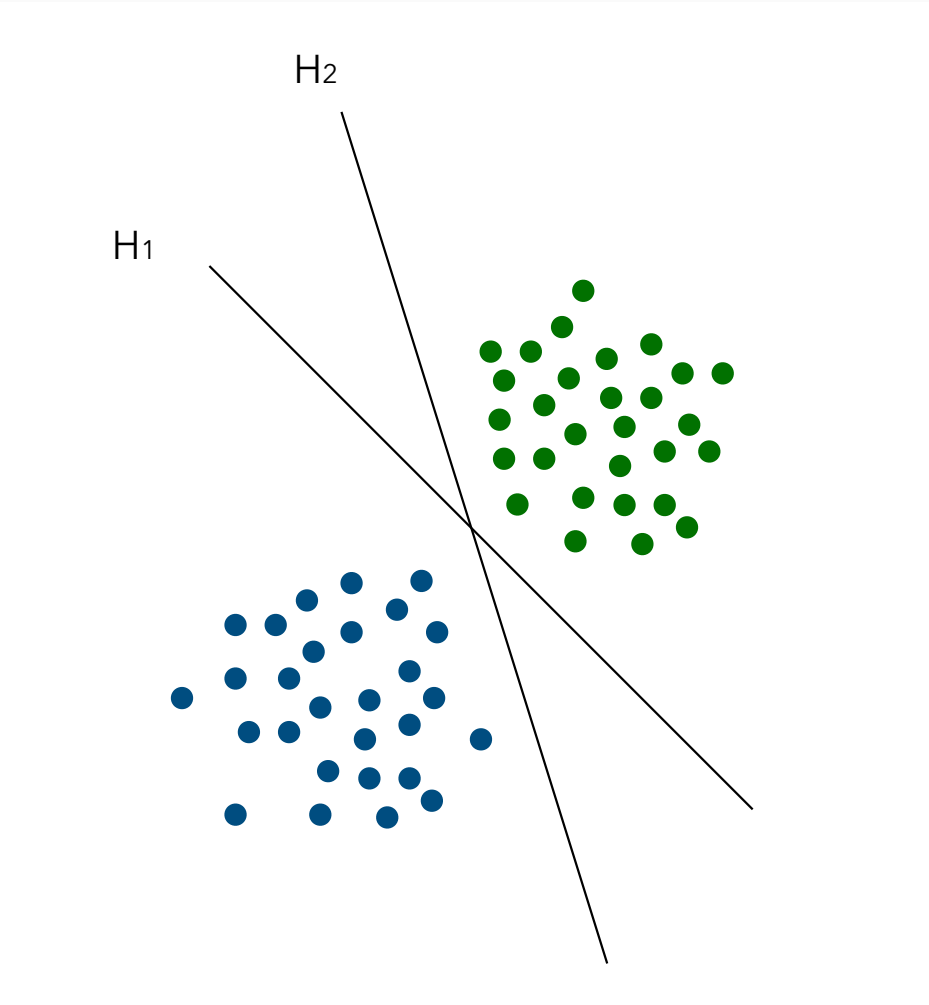
¿Cómo aprender automáticamente los parámetros w y b usando los ejemplos $S = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^{n}$?

Soluciones Posibles: Hemos visto varias formas de hacerlo, directa o indirectamente:

- El perceptrón: aprende directamente la frontera para separar consistentemente los ejemplos.
- El regresor logístico: aprende la frontera definiendo a partir de ésta un modelo $q(y|x)$ para $p(y|x)$. Luego minimiza la divergencia entre $q(y|x)$ y $p(y|x)$.
- LDA: aprende implícitamente fronteras lineales modelando $p(x|y)$ con gausianas geométricamente alineadas.

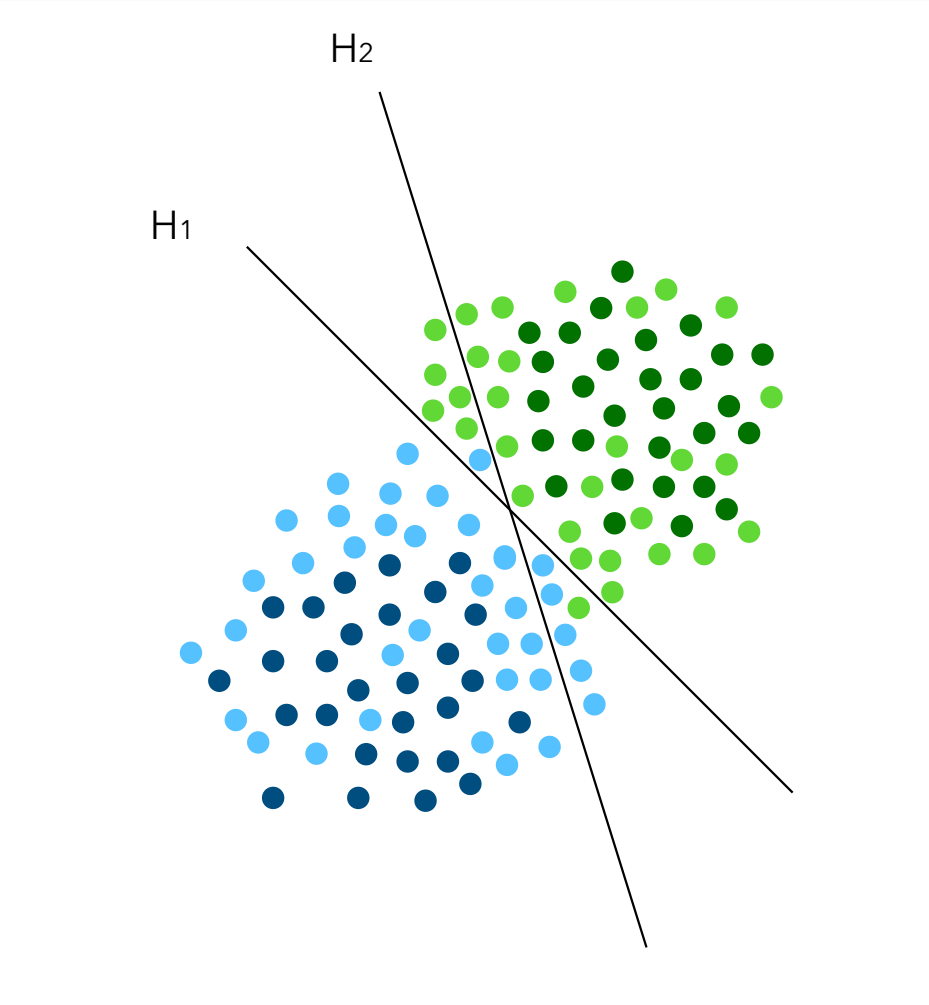
¿Qué hiperplano es mejor?

¿Es alguna de las soluciones obtenidas mejor que las otras? Depende de qué entendamos por "mejor". Los hiperplanos de más abajo son igualmente buenos en términos del error de entrenamiento ($= 0$).



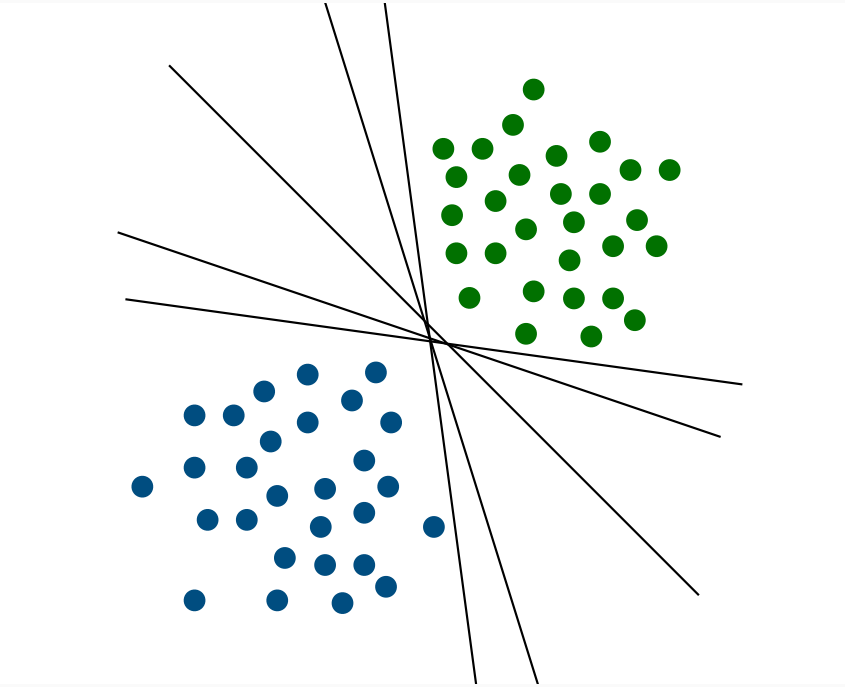
¿Qué hiperplano es mejor?

Sin embargo, no son iguales en términos del error de test (datos de test en colores claros).



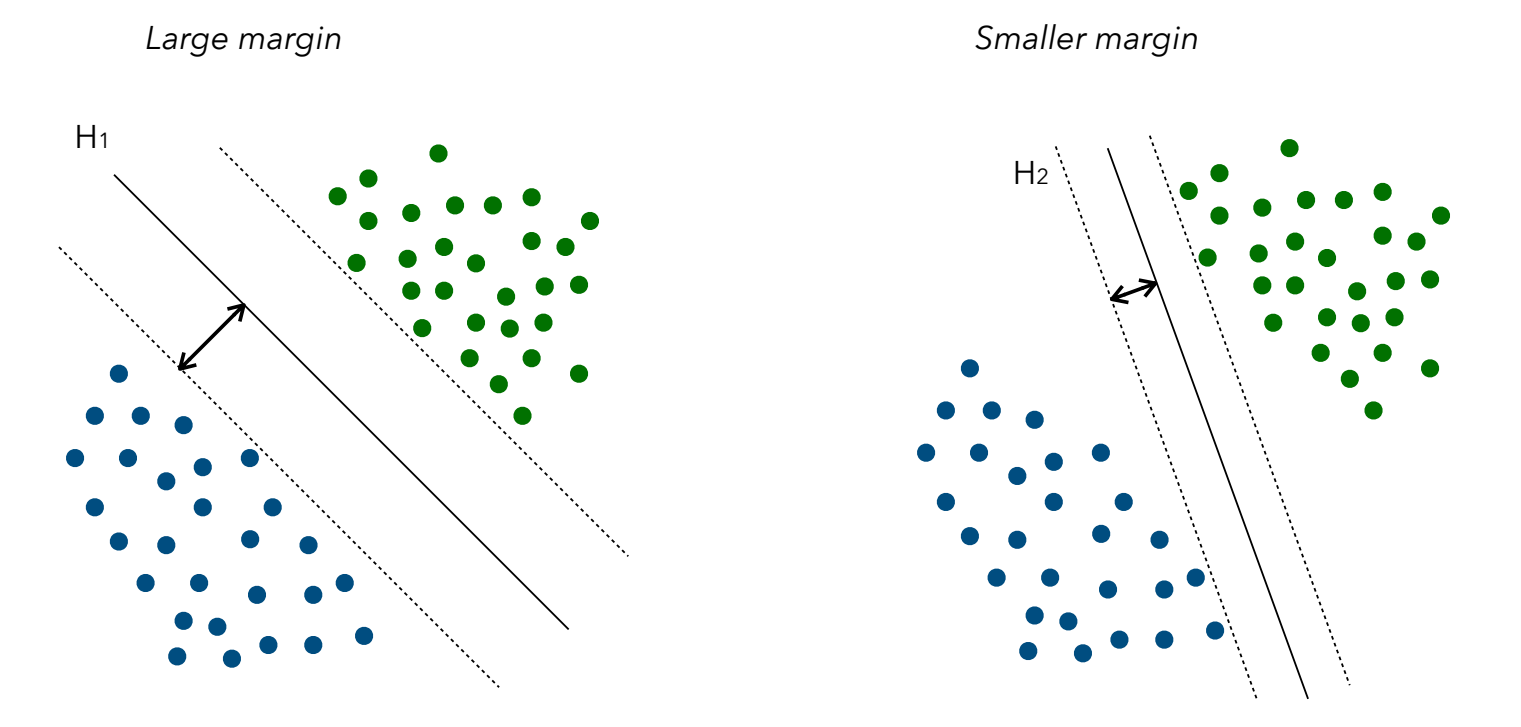
¿Qué hiperplano es mejor?

De hecho, asumiendo que el conjunto de ejemplos $S = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^{n}$ es **linealmente separable**, existen infinitos hiperplanos consistentes (error 0) con el conjunto de entrenamiento, pero sabemos que no todos son iguales en términos del error de test. ¿Cómo podemos anticipar cuál será mejor?



Margen

Entre dos hiperplanos consistentes, uno debiese preferir aquél que separa los ejemplos con mayor **margen**, es decir, aquél que **maximiza la distancia**[1] a los datos de entrenamiento. Intuitivamente, esto maximiza la probabilidad de clasificar correctamente nuevos datos.

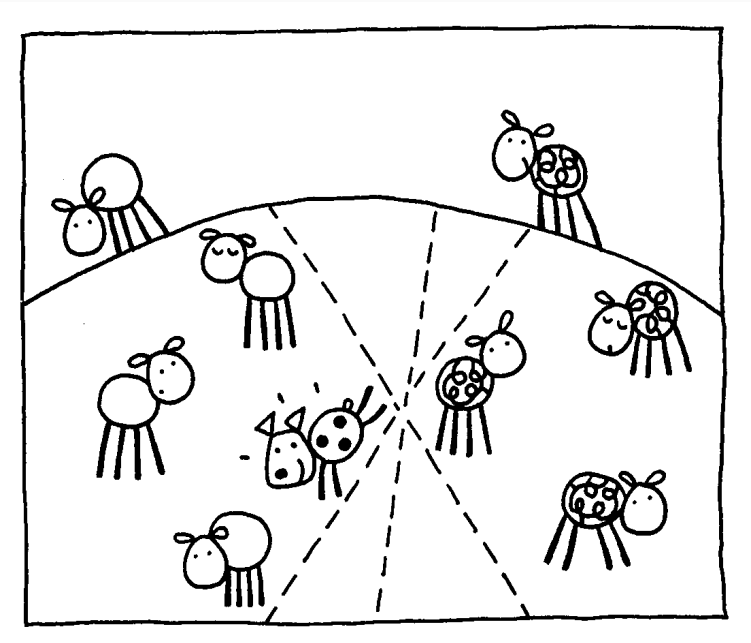


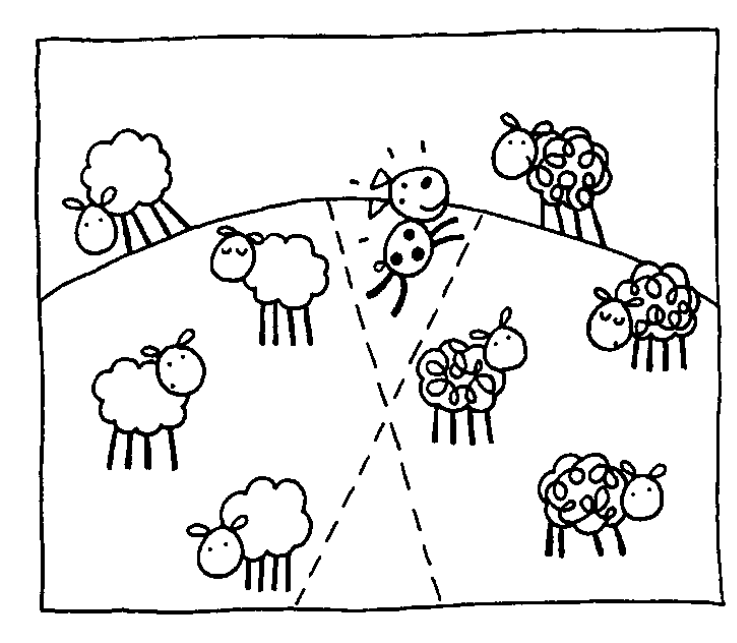
[1] Daremos definiciones precisas en la próxima sección

Una Analogía Clásica

Si los datos de entrenamiento son pocos (ovejas flacas), efectivamente muchos hiperplanos pueden clasificarlos consistentemente.

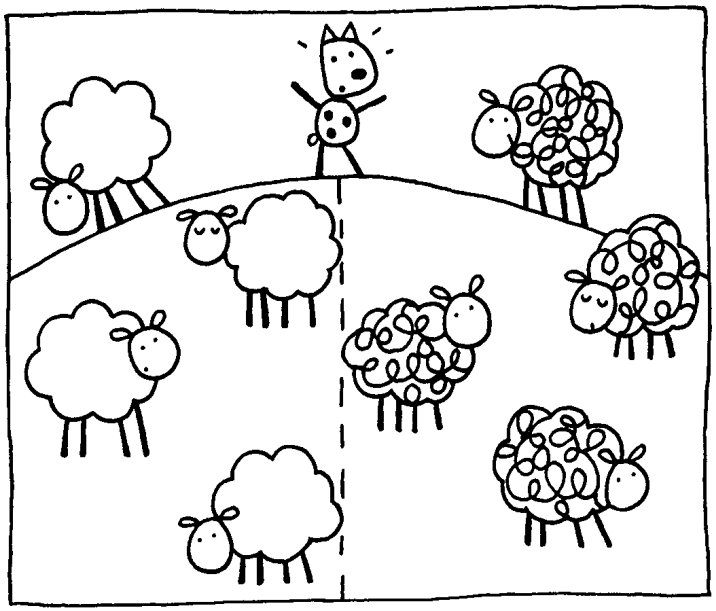
Sin embargo, si la data de entrenamiento cambia un poco o aumenta (ovejas más gordas), muchas de las soluciones que eran consistentes dejan de serlo.

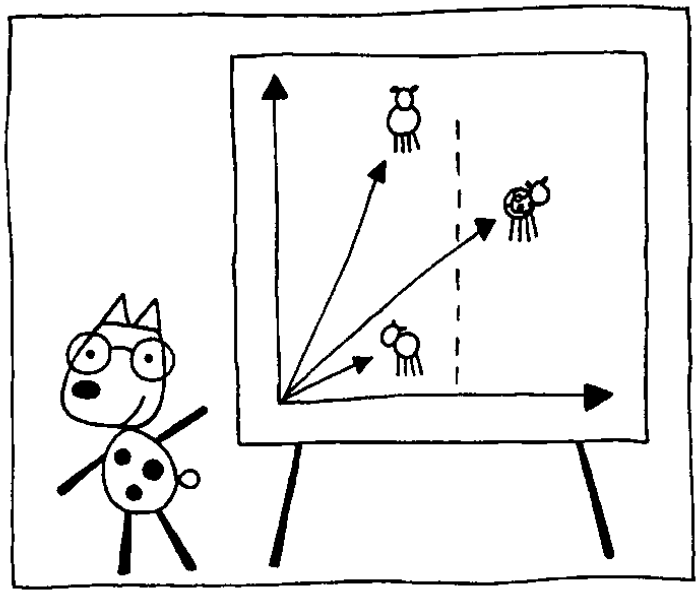




Una Analogía Clásica

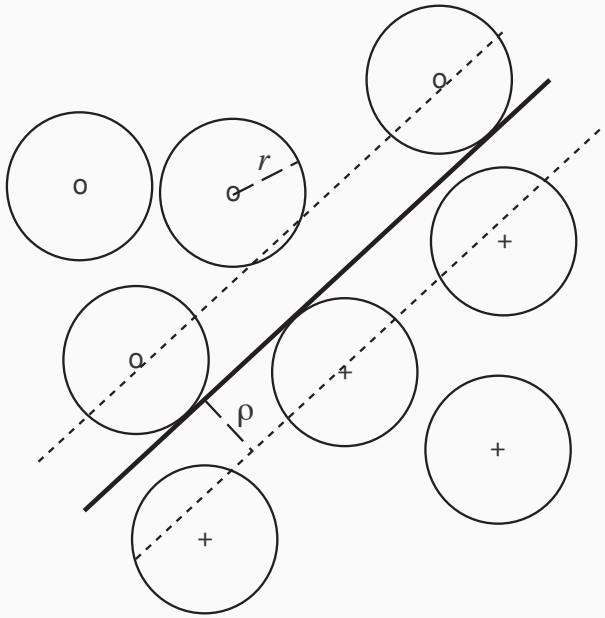
Intuitivamente, la solución más "robusta" a cambios en los datos de entrenamiento, es aquella que les deja "más espacio para crecer". Esta solución es precisamente el **hiperplano de margen máximo**.





Margen y Error de Predicción

Tratando de formalizar la intuición anterior, podemos decir que es razonable suponer que los datos futuros aparecerán en la vecindad de los datos de entrenamiento. Las ovejas gordas de la analogía, representan zonas de alta densidad de probabilidad con respecto a la aparición de de nuevos datos.



Margen y Error de Predicción

En efecto, todas estas ideas se demuestran correctas matemáticamente:

Teorema[2]

Consideremos una familia $\mathcal{F}$, formada por funciones $f : \mathbb{R}^n \to \{\pm 1\}$ de la forma $f(x) = \text{sign}(w^T x + b)$ con $\|w\|^2 \leq 1$ y sea S un conjunto de ejemplos $\{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^n$ muestreados independientemente de acuerdo a $P(x, y)$. Si $f(x) \in \mathcal{F}$ separa S con margen al menos ρ, entonces tenemos que con probabilidad $1 - \delta$ (sobre la elección de S)

$$\text{Err}_{ts} \leq \lambda \cdot \sqrt{\frac{\frac{R^2}{\rho^2} \ln n + \ln \frac{1}{\delta}}{n}},$$

donde Err_{ts} es el error de predicción (test) de f, es decir $P(f(x) \neq y)$, λ es una constante independiente de f y R es el radio de la menor bola que contiene S.

[2]Demostración disponible en Vapnik, W. (1998). *Statistical Learning Theory*. Wiley. ó (corrige un error técnico) Zhang, T. (2002). *Covering Number Bounds of Certain Regularized Linear Function Classes*. JMLR, 2, 527-550.

Margen y Error de Predicción

Maximizando el margen de clasificación ρ, minimizamos una cota superior para el error de predicción del clasificador.

Teorema[3]

$$\mathrm{Err}_{ts} \leq \lambda \cdot \sqrt{\frac{\frac{R^2}{\rho^2} \ln n + \ln \frac{1}{\delta}}{n}},$$

La idea básica de una SVM será entonces implementar un algoritmo para encontrar este hiperplano de máximo margen.

[3]Demostración disponible en Vapnik, W. (1998). *Statistical Learning Theory*. Wiley. ó (corrige un error técnico) Zhang, T. (2002). *Covering Number Bounds of Certain Regularized Linear Function Classes*. JMLR, 2, 527-550.

Derivación de una SVM Lineal Básica (Márgenes Rígidos)

En Búsqueda del Margen Máximo

- Dado un hiperplano $\mathcal{H} : \{x \in \mathbb{R}^d : w^T x + b = 0\}$ con parámetros w, b, la distancia $\rho(x; w, b)$ de un punto $x \in \mathbb{R}^d$ a $\mathcal{H}$ se define como

$$\rho(x; w, b) = \min_{z \in \mathcal{H}} \|x - z\|, \tag{3}$$

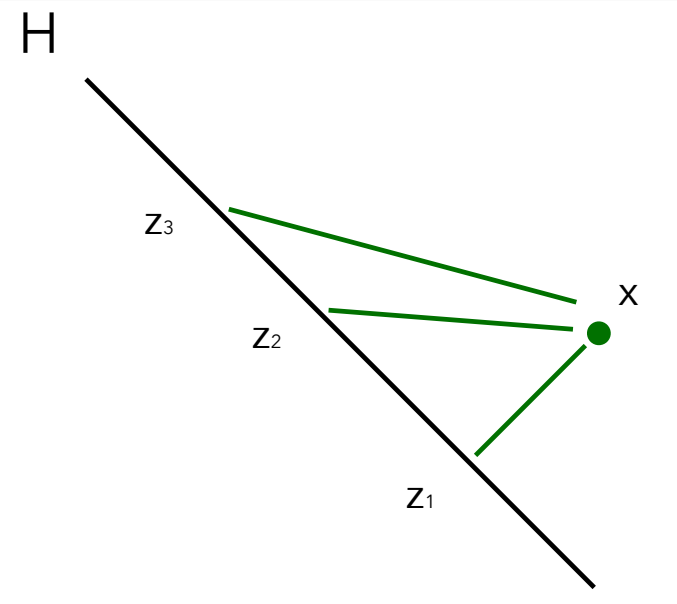
  es decir, la *distancia más pequeña* entre x y puntos z que pertenecen al hiperplano (están sobre la frontera implementada por $\mathcal{H}$)[4].

[4] Por ahora, podemos asumir que la norma utilizada es la norma euclidiana ordinaria.

Margen Geométrico respecto a 1 Punto

- La distancia de un punto x a $\mathcal{H}$ se denomina e a veces el margen (geométrico) de $\mathcal{H}$ con respecto a ese punto.

$$\rho(x; w, b) = \min_{z \in \mathcal{H}} \|x - z\|,$$



Margen Geométrico respecto a un Conjunto de Puntos

- Dado ahora un conjunto finito de puntos $S_x = \{x^{(\ell)}\}_{\ell=1}^{n}$, **el margen** de $\mathcal{H}$ con respecto a S_x se define como

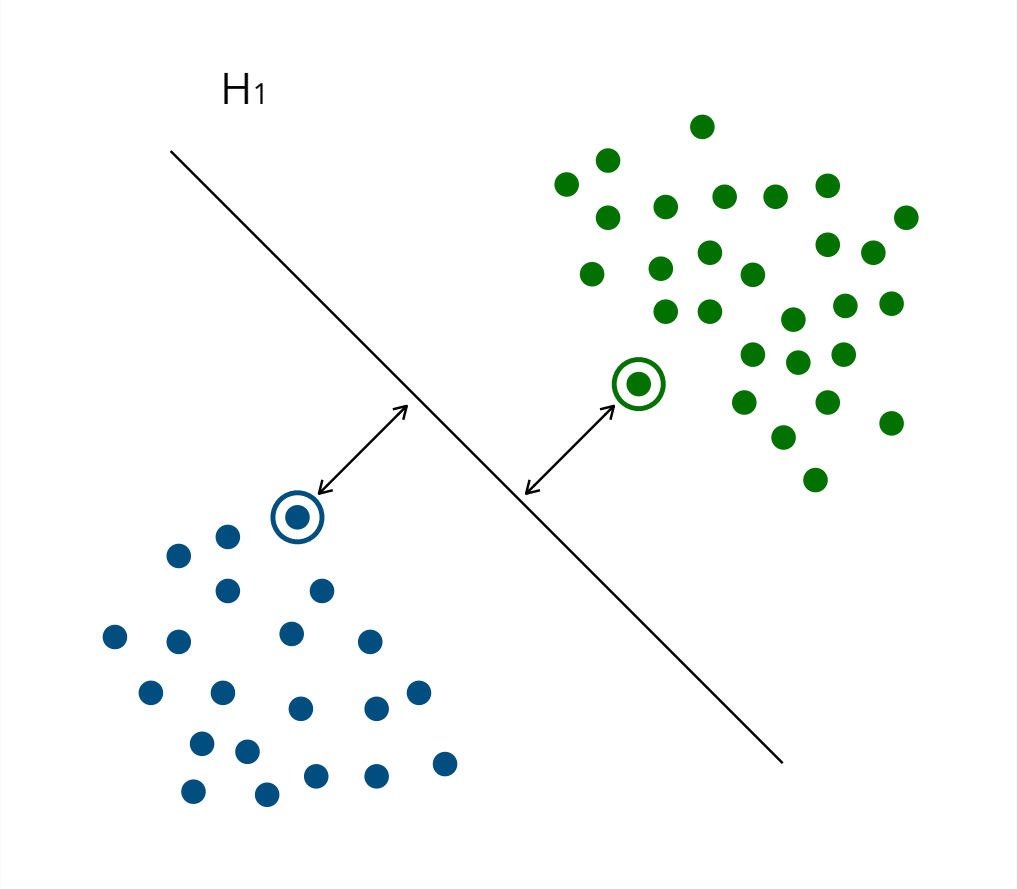
$$\rho(w, b) = \min_{\ell} \rho(x^{(\ell)}; w, b) = \min_{\ell} \min_{z \in \mathcal{H}} \|x^{(\ell)} - z\|, \tag{4}$$

  es decir, la *menor distancia* entre $\mathcal{H}$ y puntos $x^{(\ell)}$ en el conjunto.

- La necesidad de considerar la distancia más pequeña es que un margen amplio asegure una buena separación con respecto a todos los puntos de un dataset.

Margen Geométrico respecto a un Conjunto de Puntos

Gráficamente,



Primera Formulación (Ingenua)

- Podríamos entonces formular el problema de encontrar el hiperplano de margen máximo como la solución del siguiente problema de optimización:

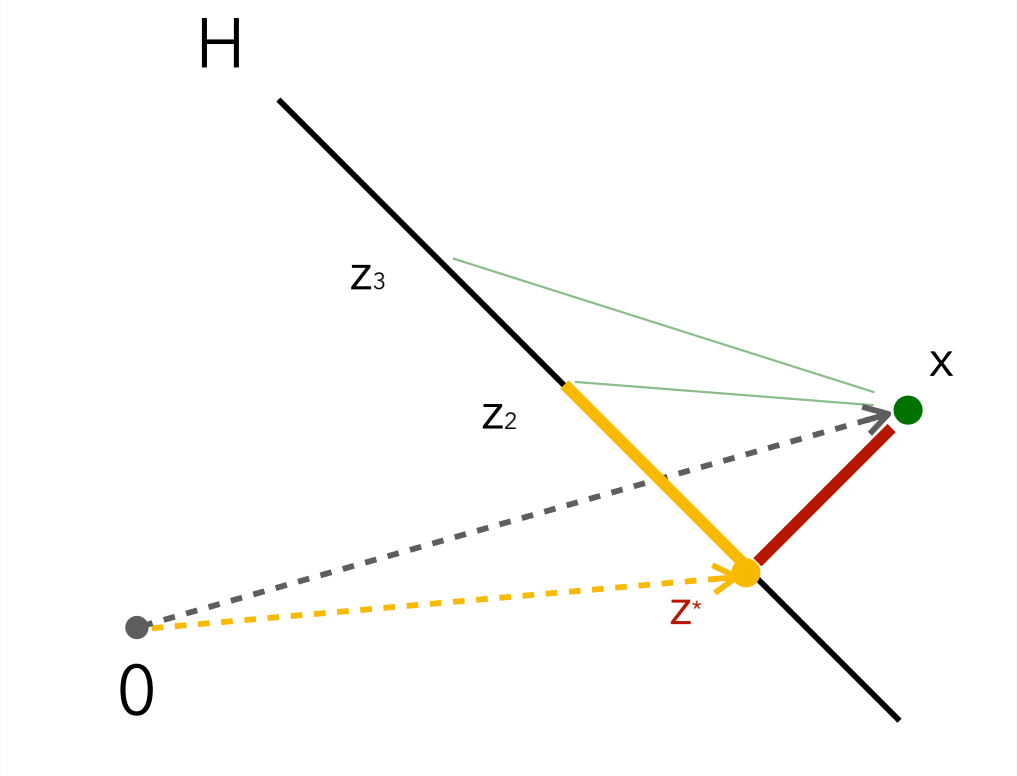
$$\max_{w,b} \min_{\ell} \min_{z \in \mathcal{H}} \|x^{(\ell)} - z\| \quad \text{s.t. sign}\left(w^T x^{(\ell)} + b\right) = y^{(\ell)} \ \forall \ell \,. \tag{5}$$

- La restricción $\text{sign}\left(w^T x^{(\ell)} + b\right) = y^{(\ell)}$ simplemente asegura que el hiperplano sea un hiperplano separador.
- Indudablemente, el problema anterior no tiene una forma simple.
- Un ingrediente crucial en el éxito de las SVMs formular el problema en un modo que sea fácil de resolver algorítmicamente.

Si se quieren saltar las derivaciones geométricas, ir directamente a la formulación (6).

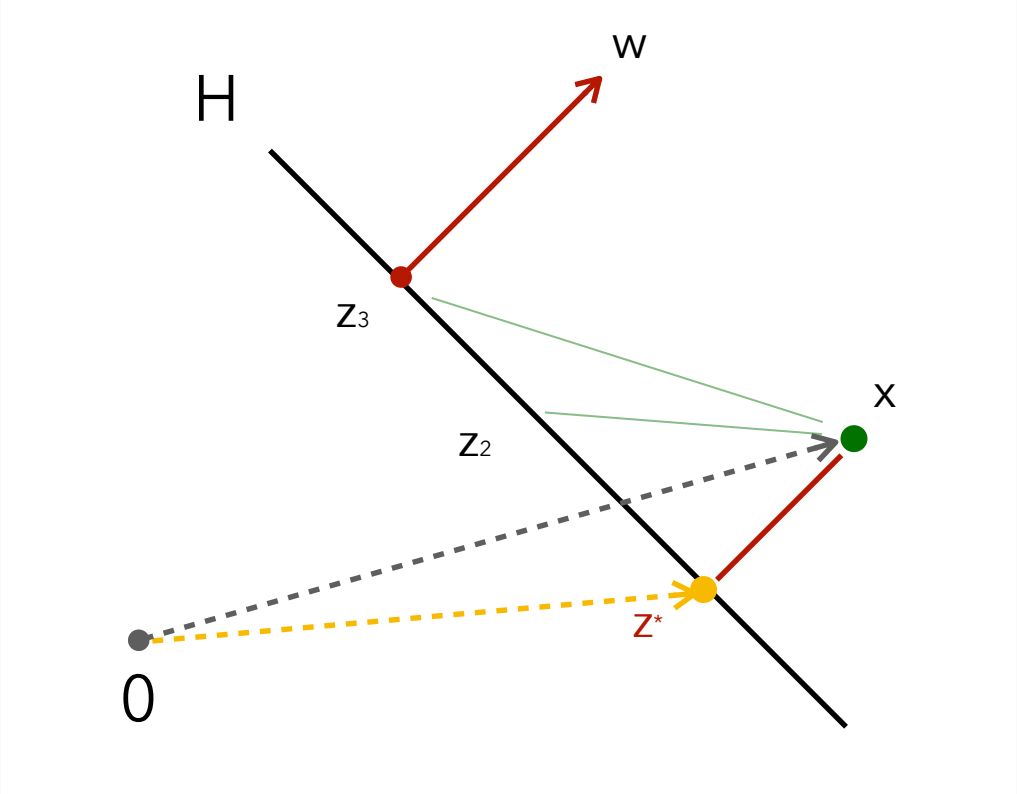
Cómputo del Margen Máximo

No es difícil demostrar que un punto z^* que pertenece al hiperplano es el más cercano a x si y sólo si se satisface la condición $(x - z^*) \perp (z - z^*)$ $\forall z \in \mathcal{H}$.



Cómputo del Margen Máximo

Equivalentemente, podemos decir que $(x - z^*) \parallel w$, donde w es el vector que define la orientación del hiperplano.



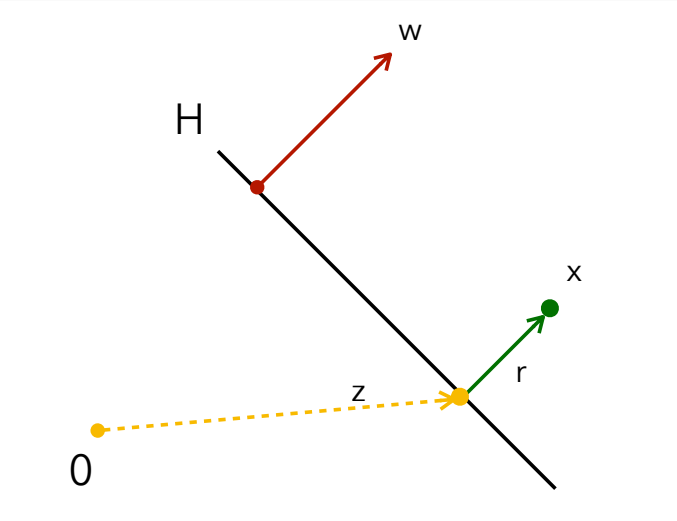
Esta observación nos permitirá simplificar el problema.

Cómputo del Margen Máximo

En efecto, siempre podemos escribir x como suma de dos vectores $x = z^* + r$, tal que (i) $\boldsymbol{z}^*$ está sobre el hiperplano, (ii) $r \parallel w$ y (iii) $\|r\|$ es exactamente la distancia de x al hiperplano. Tenemos entonces que

$$w^T r = \|w\| \, \|r\| \cos(\angle_w^r) = \pm \|w\| \, \|r\|$$

$$\Rightarrow \rho(x; w, b) = \|r\| = \left| \frac{w^T r}{\|w\|} \right| = \left| \frac{w^T x - w^T z^*}{\|w\|} \right| = \left| \frac{w^T x + b}{\|w\|} \right| .$$



Segunda Formulación

La optimización de más a la derecha en el problema original,

$$\max_{w,b} \min_{\ell} \min_{z \in \mathcal{H}} \|x^{(\ell)} - z\|$$
$$\text{s.t. sign}\left(w^T x^{(\ell)} + b\right) = y^{(\ell)} \ \forall \ell\,,$$

se puede entonces eliminar,

$$\max_{w,b} \min_{\ell} \left| \frac{w^T x^{(\ell)} + b}{\|w\|} \right|$$
$$\text{s.t. sign}\left(w^T x^{(\ell)} + b\right) = y^{(\ell)} \ \forall \ell\,.$$

Además, si $\text{sign}\left(w^T x^{(\ell)} + b\right) = y^{(\ell)}$,

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

Segunda Formulación

Por lo tanto nuestro problema se puede escribir como

$$\begin{aligned} \max_{w,b} \min_{\ell} \; & y^{(\ell)} \left(\frac{w^T x^{(\ell)} + b}{\|w\|} \right) \\ \text{s.t. } & \text{sign} \left(w^T x^{(\ell)} + b \right) = y^{(\ell)} \; \forall \ell \,, \end{aligned} \tag{6}$$

Este problema es mucho más simple que el de partida.

Sin embargo, podemos todavía hacerlo mucho mejor. Si se quieren saltar también estas derivaciones, ir directamente a la formulación (7).

Margen Máximo como Conjunto de Restricciones

Notemos ahora que el problema

$$\max_{w,b} \min_{\ell} \; y^{(\ell)} \left(\frac{w^T x^{(\ell)} + b}{\|w\|} \right)$$
$$\text{s.t. } \text{sign}\left(w^T x^{(\ell)} + b \right) = y^{(\ell)} \; \forall \ell \,,$$

es equivalente a

$$\max_{w,b} \rho \quad \text{s.t.} \quad \rho = \min_{\ell} y^{(\ell)} \left(\frac{w^T x^{(\ell)} + b}{\|w\|} \right)$$
$$\text{s.t.} \quad \text{sign}\left(w^T x^{(\ell)} + b \right) = y^{(\ell)} \; \forall \ell \,.$$

Margen Máximo como Conjunto de Restricciones

Pero la restricción adicional

$$\rho = \min_{\ell} y^{(\ell)} \left(\frac{w^T x^{(\ell)} + b}{\|w\|} \right) ,$$

es equivalente a escribir

$$y^{(\ell)} \left(\frac{w^T x^{(\ell)} + b}{\|w\|} \right) > \rho \; \forall \ell \, .$$

Esto produce un problema mucho más canónico:

$$\begin{aligned} \max_{w,b} \rho \quad & \text{s.t.} \quad y^{(\ell)} \left(\frac{w^T x^{(\ell)} + b}{\|w\|} \right) > \rho \; \forall \ell \\ & \text{s.t.} \quad \text{sign} \left(w^T x^{(\ell)} + b \right) = y^{(\ell)} \; \forall \ell \, . \end{aligned}$$

Tercera Formulación (casi una SVM)

Finalmente, si asumimos que el valor objetivo óptimo ρ^* del problema anterior es positivo (problema linealmente separable), la restricción

$$y^{(\ell)} \left(\frac{w^T x^{(\ell)} + b}{\|w\|} \right) > \rho \; \forall \ell \,,$$

ya asegura que

$$\text{sign} \left(w^T x^{(\ell)} + b \right) = y^{(\ell)} \; \forall \ell \,,$$

y por lo tanto podemos reducir el problema a:

$$\max_{w,b} \rho \quad \text{s.t.} \quad y^{(\ell)} \left(\frac{w^T x^{(\ell)} + b}{\|w\|} \right) > \rho \; \forall \ell \,, \tag{7}$$

Arreglos Finales

El problema anterior tiene dos inconvenientes:

- Dada una solución w, b candidata, cualquier escalamiento por una constante, produce el mismo resultado (espacio de búsqueda redundante).
- Las restricciones no son convexas.

Una solución que permite eliminar ambos problemas consiste en agregar una restricción de la forma $\|w\| = \kappa\,(= \text{cte})$.

$$\max_{w,b} \rho \quad \text{s.t.} \quad y^{(\ell)}\left(w^T x^{(\ell)}\right) + b > \rho \cdot \kappa \;\forall \ell \quad \|w\| = \kappa\,,$$

Como la elección de la constante es arbitraria, podemos elegir una conveniente $\kappa = 1/\rho$ y obtener

$$\max_{w,b} \rho \quad \text{s.t.} \quad y^{(\ell)}\left(w^T x^{(\ell)} + b\right) > 1 \;\forall \ell \quad \|w\| = 1/\rho\,,$$

Este problema está en una forma mucho más amena para utilizar un solver estándar de optimización con restricciones.

SVM de Margen Duro

Una última modificación se obtiene al observar que la solución w^* de este problema debe ser la misma si reemplazamos la f.o. por $\frac{1}{2}\|w\|^2$.

$\star$ Obtenemos una formulación conocida como Hard-Margin SVM:

$$\begin{aligned} &\min_{w,b} \ \frac{1}{2}\|w\|^2 \\ &\text{s.t. } y^{(\ell)}\left(w^T x^{(\ell)} + b\right) \geq 1 \ \forall \ell \end{aligned} \tag{8}$$

A diferencia de los anteriores, éste es un problema de optimización convexo (objetivo convexo y restricciones lineales) que:

- Puede ser resuelto eficientemente.
- Nos permite usar toda la teoría de optimización convexa para estudiar las propiedades de la solución.

Márgenes Blandos: Construcción de la C-SVM.

Márgenes Rígidos

En la sección anterior obtuvimos la formulación básica de una SVM

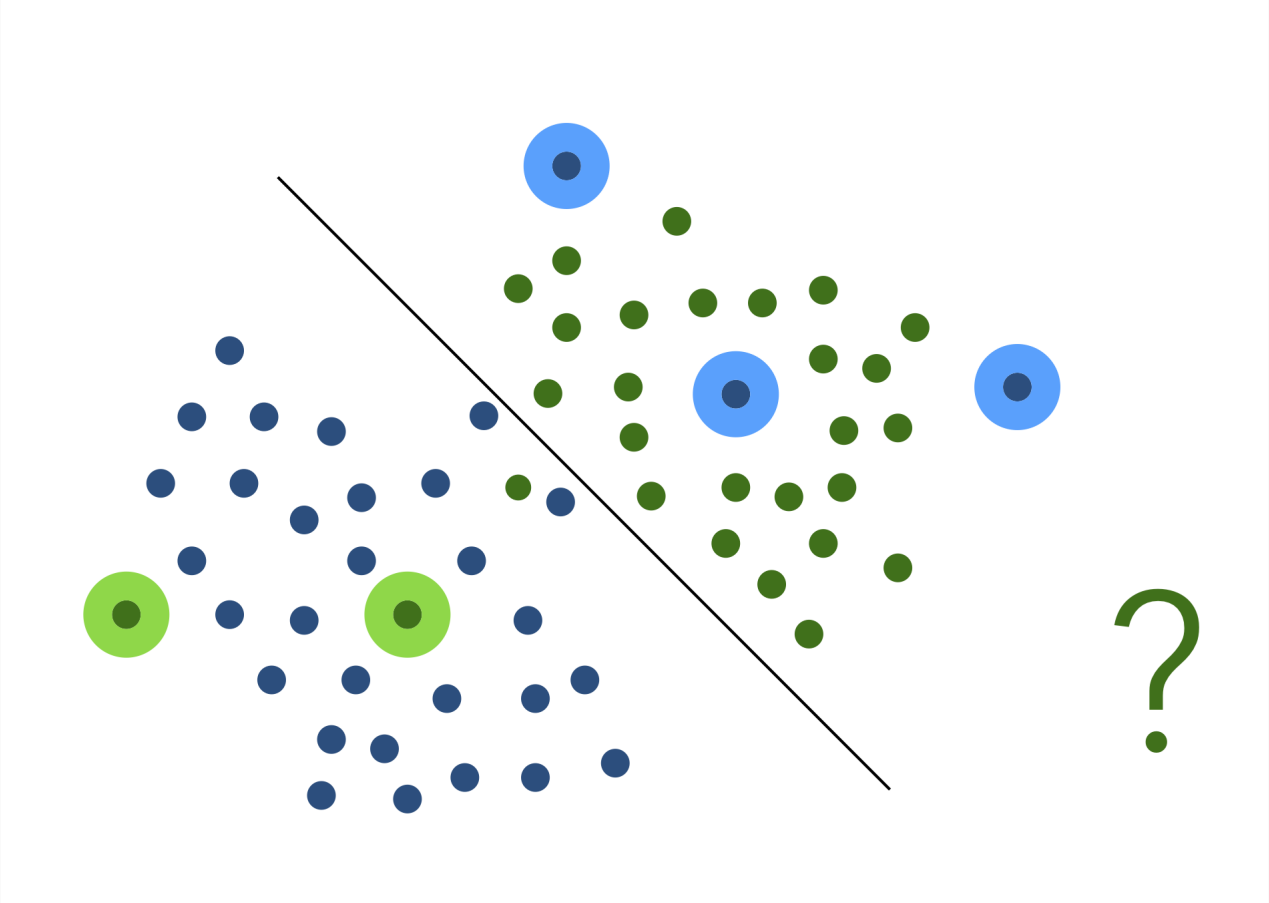
$$\begin{aligned} &\min_{w,b} \ \frac{1}{2}\|w\|^2 \\ &\text{s.t. } y^{(\ell)}\left(w^T x^{(\ell)} + b\right) \geq 1 \ \forall \ell \end{aligned}$$

que se basa en una definición "rígida" del concepto de margen, rigidez que trae consigo algunos problemas:

- Es poco robusta en el sentido de que basta 1 punto mal posicionado para que margen cambie completamente.
- Funciona sólo para problemas linealmente separables.

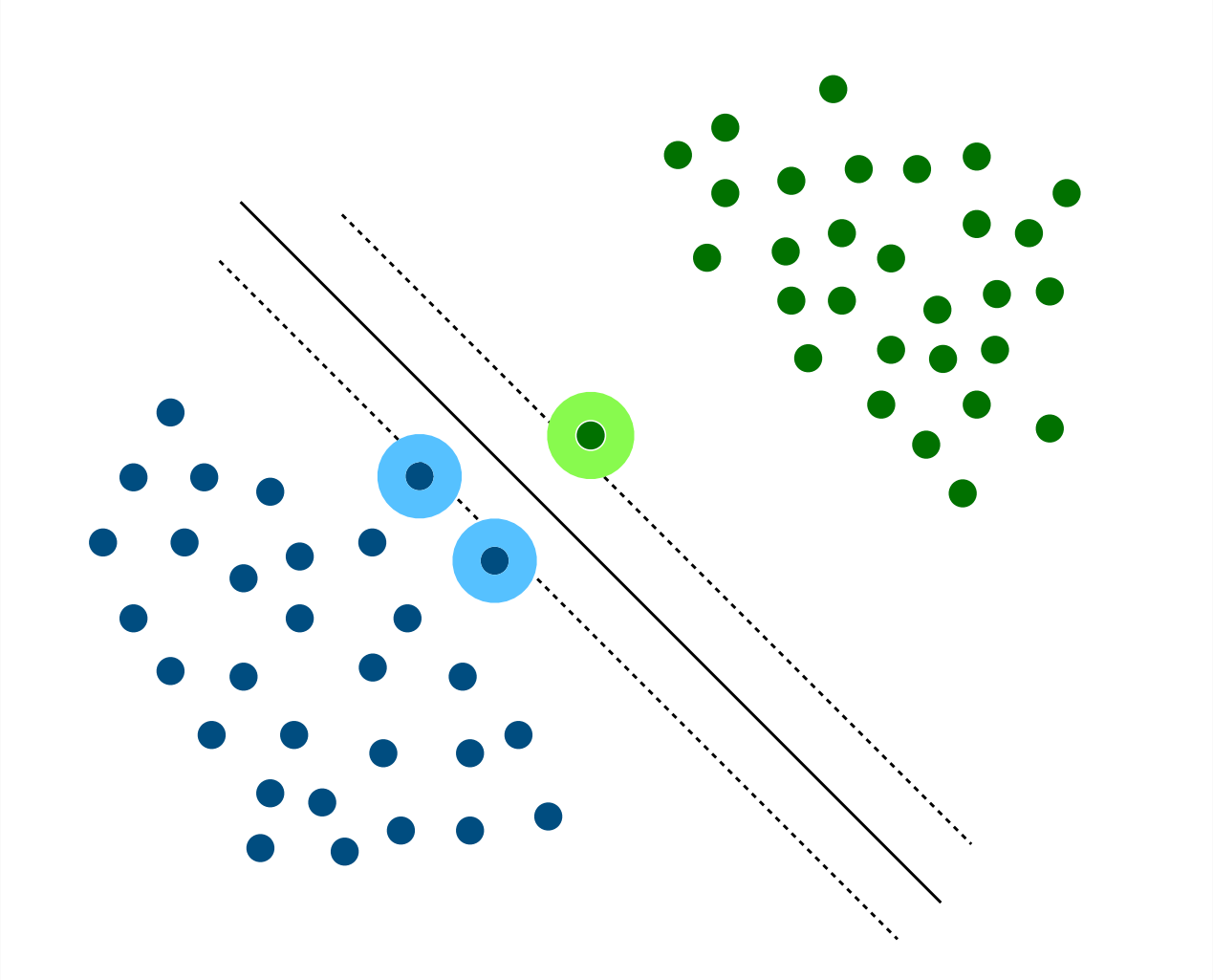
Márgenes Rígidos

Los datos podrían ser linealmente inseparables.



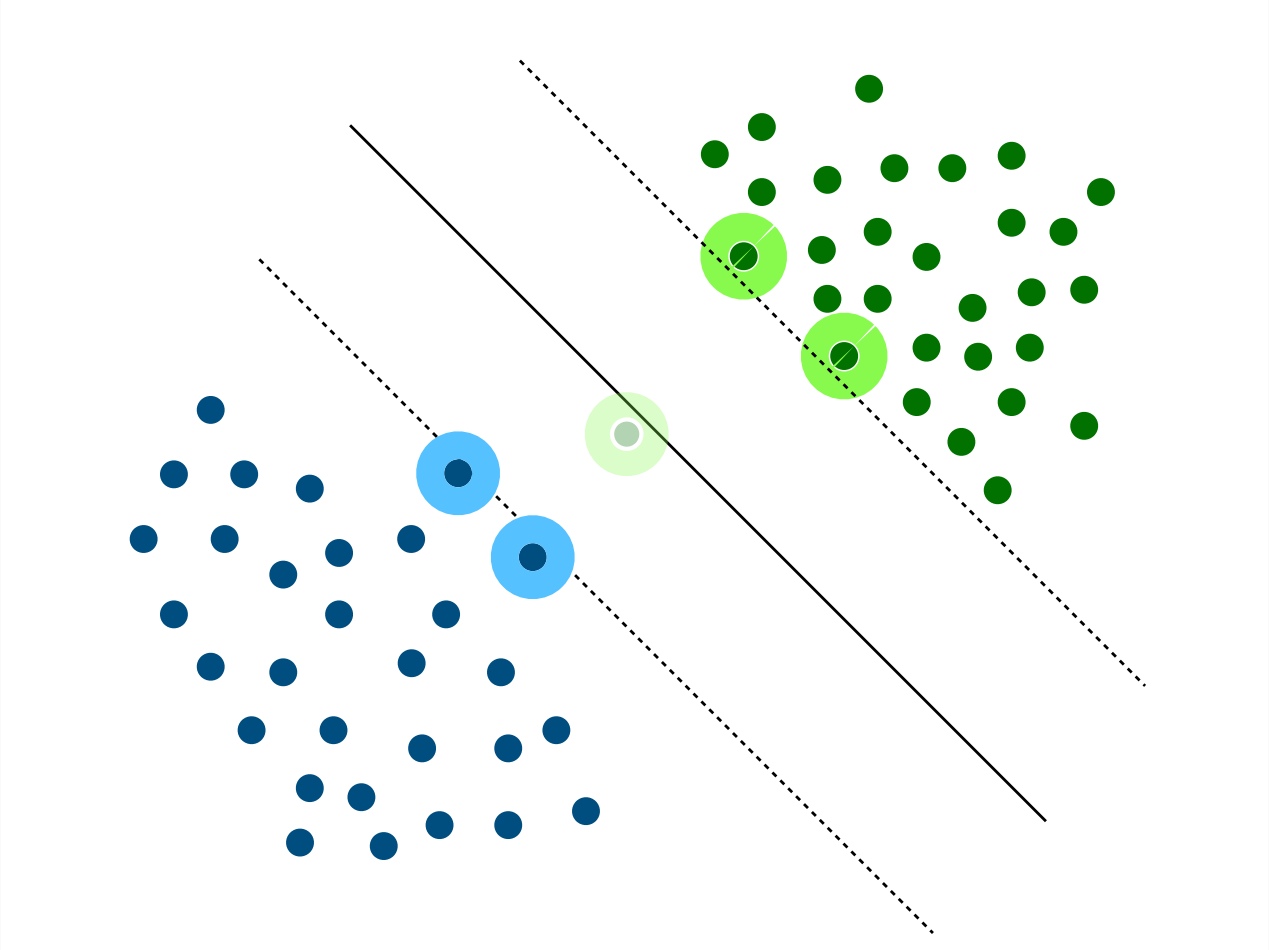
Márgenes Rígidos

Los puntos que determinan el margen podrían ser poco representativos de la clase.



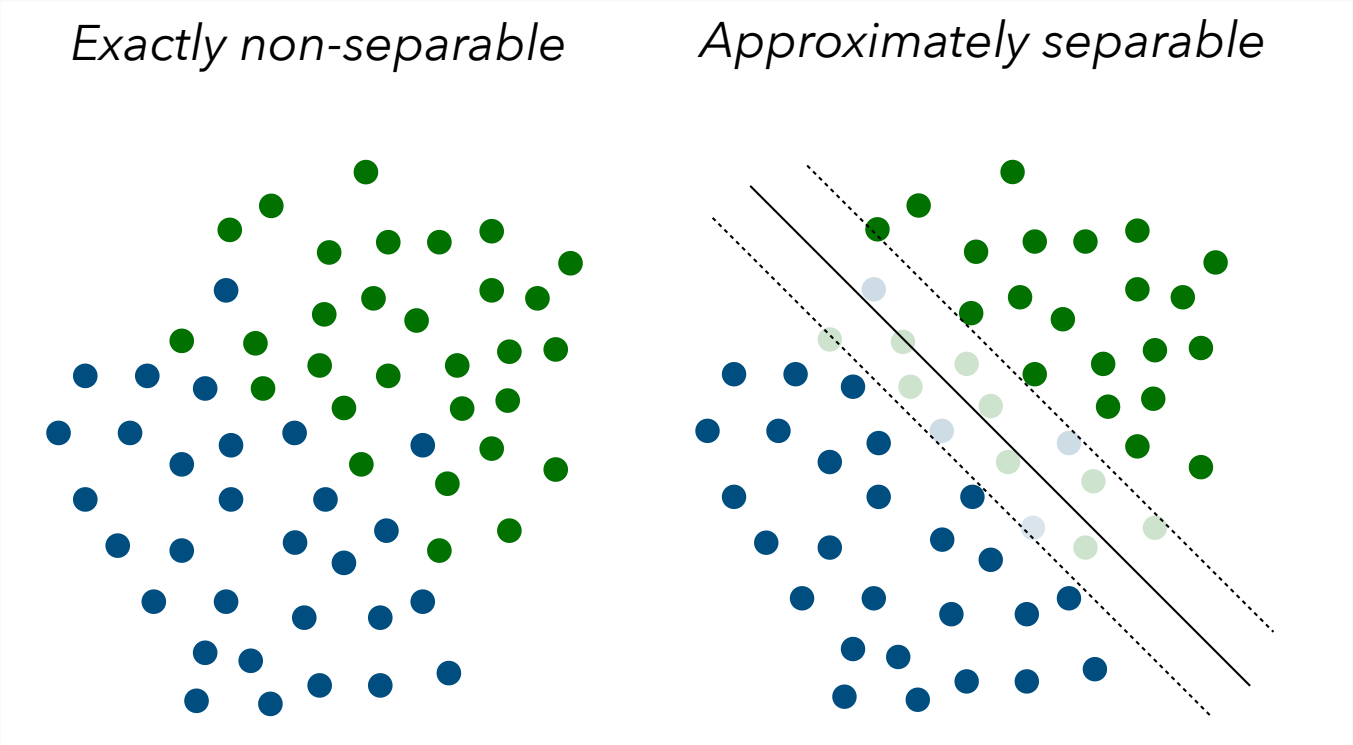
Márgenes Blandos

Idea: Si permitimos que la máquina ignore una pequeña fracción de los datos de entrenamiento, nuestro clasificador sería más robusto frente a pequeños cambios en el conjunto de entrenamiento.



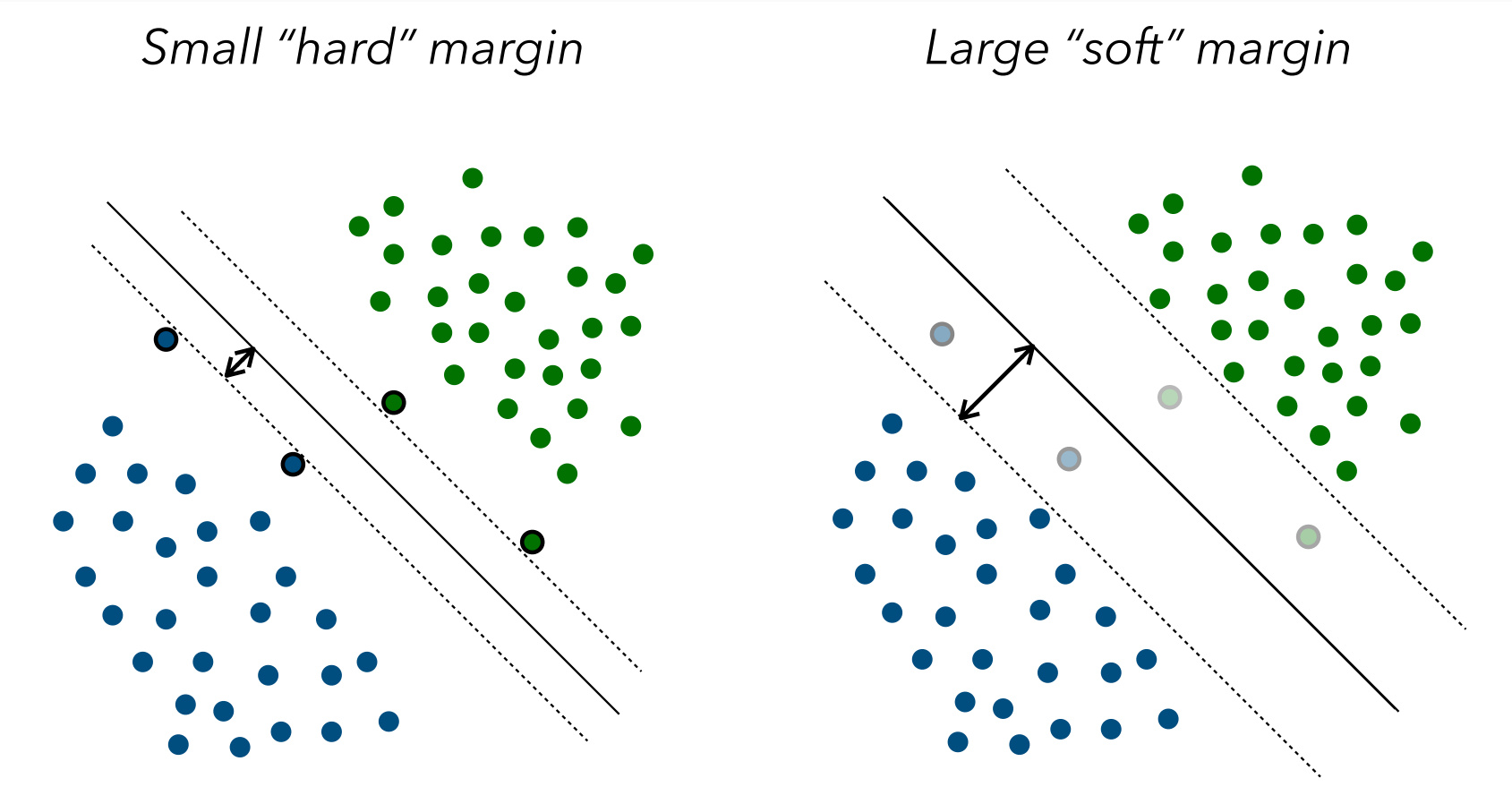
Márgenes Blandos

Idea: Ignorando una pequeña fracción de los datos de entrenamiento, un problema linealmente inseparable podría ser volverse linealmente separable.



Márgenes Blandos

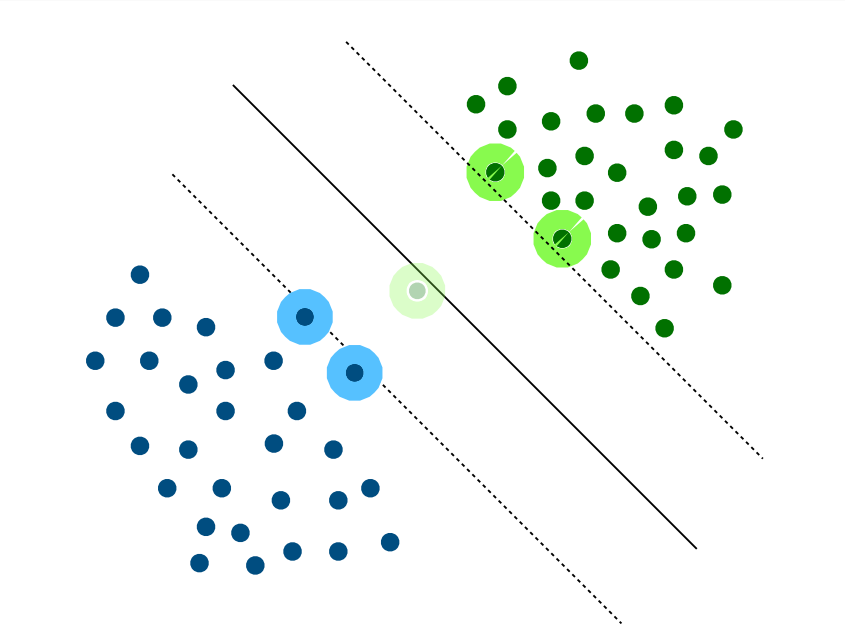
Idea: De hecho, podríamos encontrar un margen más grande, que vale de modo aproximado.



Márgenes Blandos

Podemos re-definir **el margen** de un hiperplano (w, b), para soportar cualquier restricción A del conjunto total de puntos en consideración $S = \{x^{(\ell)}\}_{\ell=1}^{n}$

$$\rho_A(w, b) = \min_{\ell:\, x^{(\ell)} \in A} \rho(x^{(\ell)}; w, b)\,. \tag{9}$$



Márgenes Blandos y Error de Predicción

La relación entre el margen y el error de predicción, que motiva las SVMs, se puede generalizar para considerar márgenes blandos.

Teorema

Consideremos una familia $\mathcal{F}$, formada por funciones $f : \mathbb{R}^n \to \{\pm 1\}$ de la forma $f(x) = \text{sign}(w^T x + b)$ con $\|w\|^2 \leq 1$ y sea S un conjunto de ejemplos $\{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^n$ muestreados independientemente de acuerdo a $P(x, y)$. Para cualquier función $f(x) \in \mathcal{F}$, tenemos que con probabilidad $1 - \delta$ (sobre la elección de S)

$$\text{Err}_{ts} \leq \text{Err}_\rho + \lambda \cdot \sqrt{\frac{\frac{R^2}{\rho^2} \ln n + \ln \frac{1}{\delta}}{n}} \ ,$$

donde Err_{ts} es el error de predicción (test) de f, λ, R son constantes y Err_ρ es la fracción de datos en S clasificados con margen menor que η por $f(x)$.

Márgenes Blandos y Error de Predicción

• Existe un tradeoff la fracción de puntos que se ignoran y el margen obtenido.

$$\mathrm{Err}_{ts} \leq \mathrm{Err}_{\rho} + \lambda \cdot \sqrt{\frac{\frac{R^2}{\rho^2} \ln n + \ln \frac{1}{\delta}}{n}},$$

• Ignorando muchos puntos, se aumenta valor del primer miembro (Err_{ρ}), pero se disminuye el valor del segundo (porque se puede elegir un margen ρ más grande).

• Ignorando muy pocos puntos, el valor del primer miembro puede acercarse a 0, pero se aumenta el valor del segundo (porque nos acercamos a la definición rígida de margen).

Reconstrucción de la SVM

• Las ideas anteriores, se pueden incorporar fácilmente en la SVM de la Eqn.(8) usando variables de holgura ξ_ℓ que permitan violaciones de las restricciones y cuantifiquen el grado de la violación.

Hard-margin constraints	Soft-margin constraints
$y^{(\ell)}(w^T x^{(\ell)} + b) \geq 1$	$y^{(\ell)}(w^T x^{(\ell)} + b) \geq 1 - \xi_\ell$

• Si $\xi_\ell = 0$, el punto $x^{(\ell)}$ satisface la restricción clásica (no está siendo ignorado)

• Si $\xi_\ell > 0$, el punto $x^{(\ell)}$ viola la restricción clásica (está siendo ignorado para el cálculo del margen)

• Si $\xi_\ell > 1$, el punto $x^{(\ell)}$ viola la restricción clásica y además está siendo mal clasificado.

• No tiene sentido que $\xi_\ell < 0$. En este caso, el punto $x^{(\ell)}$ podría respetar la restricción clásica.

SVM de Margen Blando

• El balance entre cantidad de puntos ignorados y el margen obtenido se puede controlar penalizando el valor de las variables de holgura en el objetivo.

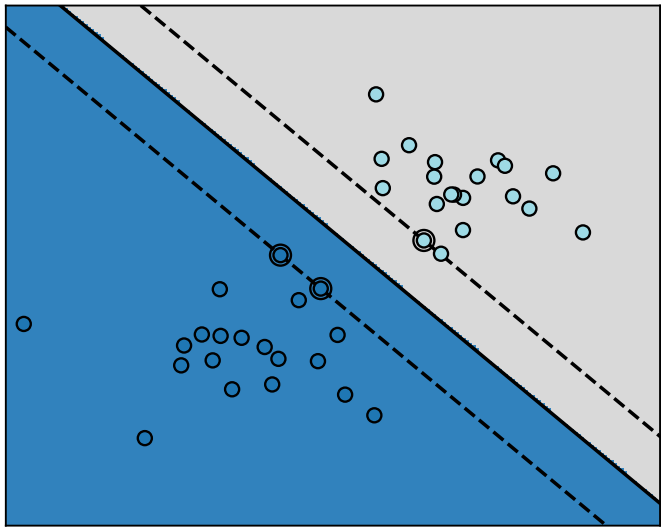
$$\begin{aligned} \min_{w,b,\xi} \ & \frac{1}{2}\|w\|^2 + \boldsymbol{C} \cdot \sum\nolimits_{\ell} \xi_\ell \\ \text{s.t. } & y^{(\ell)}(w^T x^{(\ell)} + b) \geq 1 - \xi_\ell \, \forall \ell \\ & \xi_\ell \geq 0 \, \forall \ell \end{aligned} \tag{10}$$

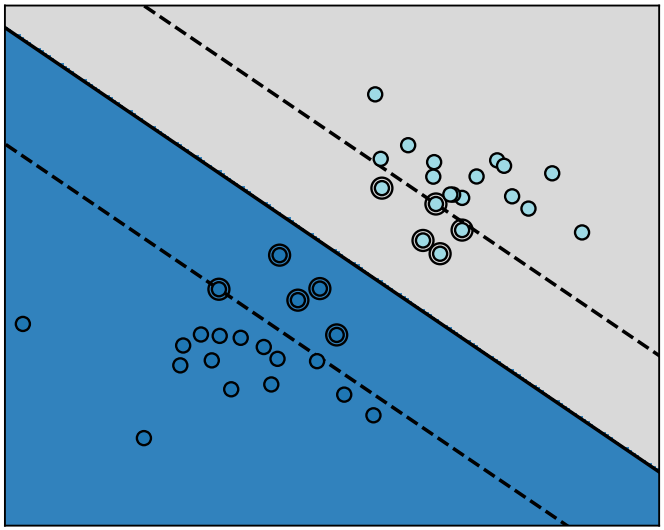
• Esta nueva formulación se conoce como Soft-Margin SVM (concretamente C-SVM).

• El parámetro C se denomina *parámetro de regularización*.

Rol del Parámetro C

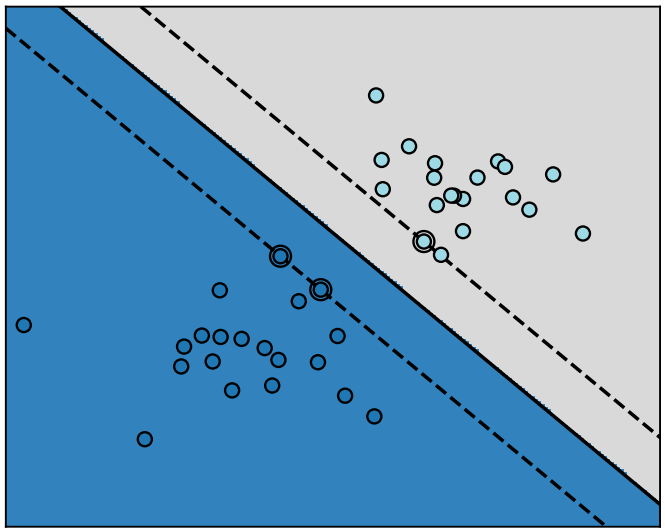
Un valor pequeño de C permite ignorar un gran conjunto de puntos de entrenamiento y por lo tanto enfoca el aprendizaje en *la maximización del margen*. Muchos ejemplos de entrenamiento podrían ser clasificados incorrectamente.

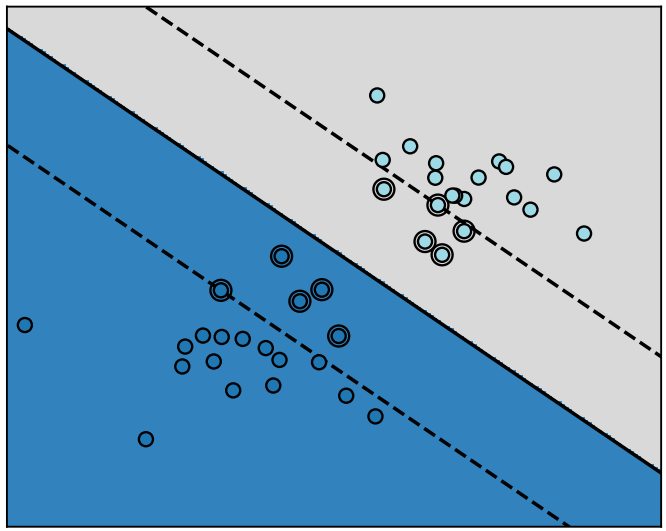




Rol del Parámetro C

Un valor grande de C no permite muchas violaciones de las restricciones originales. Enfoca el entrenamiento en la minimización del *error de entrenamiento*.





Rol del Parámetro C

El valor óptimo de C depende por lo tanto del problema y debe ser elegido usando un estimador del error de predicción (conjunto de validación, validación cruzada, etc).

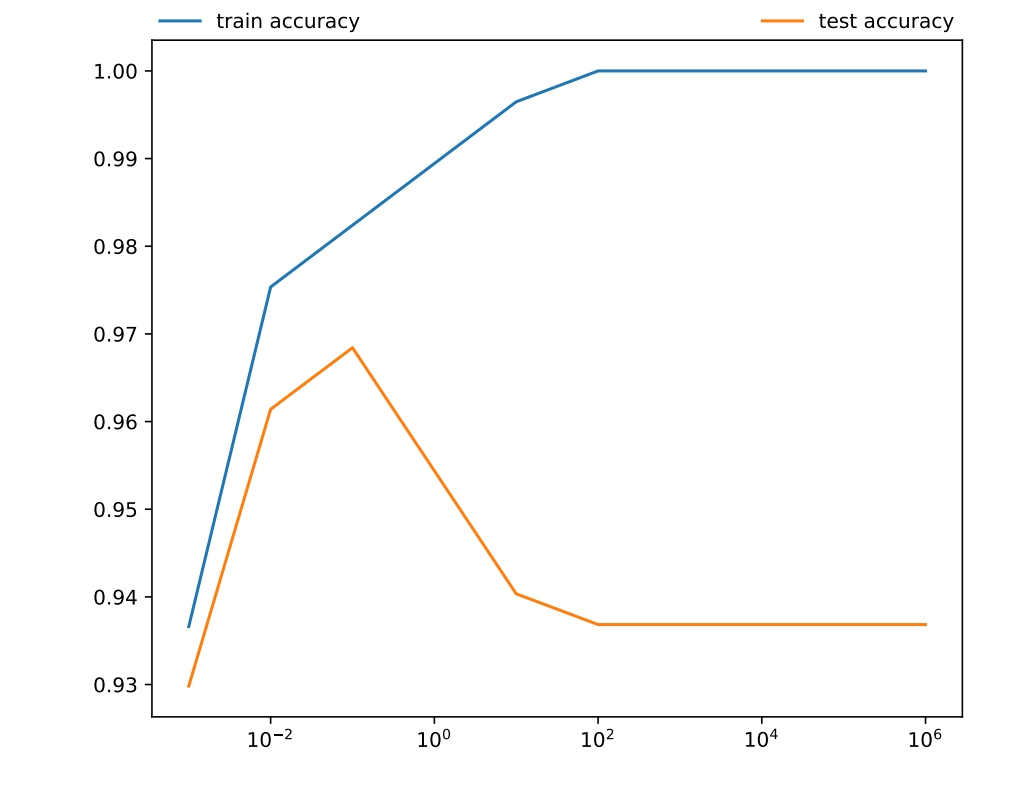


Figure 1: Eje X: valor de C. Eje Y: accuracy de test (naranja) y train (azul).

Algunas Propiedades del Clasificador de Margen Máximo

Función de Decisión & Support Vectors

- Porqué support vector machine?

- Es posible demostrar que si w (ver Apéndice) es la solución del problema de optimización que define nuestras SVMs, se satisface

$$w = \sum_{\ell} \alpha_\ell \, y^{(\ell)} x^{(\ell)} \quad \text{con} \quad \alpha_\ell \in \mathbb{R}_0^+ \, .$$

- Recordemos que la función de decisión implementada por el clasificador es $f(x) = \text{sign}(w^T x + b)$. Usando la expansión anterior, vemos que la función de decisión tiene la forma

$$f(x) = \text{sign} \left(\sum_{\ell} \alpha_\ell \, y^{(\ell)} \langle x^{(\ell)}, x \rangle \right) \, ,$$

donde hemos usado $\langle a, b \rangle$ para denotar $a^T b$.

Función de Decisión & Support Vectors

- De esta expresión para la función de decisión

$$f(x) = \text{sign}\left(\sum_{\ell} \alpha_\ell \, y^{(\ell)} \langle x^{(\ell)}, x \rangle\right) ,$$

concluimos que cualquier ejemplo $x^{(\ell)}$ tal que $\alpha_\ell = 0$ no participa de la decisión sobre la clase de nuevos datos (test).

- Los puntos $x^{(\ell)}$ tales que $\alpha_\ell \neq 0$ se denominan **support vectors**. Cualquier ejemplo que no sea un support vector se puede remover de la función del conjunto de entrenamiento sin alterar la solución aprendida por la SVM.

- Vemos también que la eficiencia del clasificador (luego de que está entrenado) es proporcional al número de support vectors.

Función de Decisión & Support Vectors

- De la expresión para la función de decisión

$$f(x) = \text{sign}\left(\sum_{\ell} \alpha_\ell \, y^{(\ell)} \langle x^{(\ell)}, x \rangle\right) ,$$

concluimos también que una SVM es en esencia un **modelo no-paramétrico**: el número de parámetros del modelo crece cuando crece el conjunto de datos de entrenamiento.

- En otras palabras, la complejidad de una SVM es adaptativa: si disponemos de una mayor cantidad de datos de entrenamiento, nos permitimos una mayor cantidad de parámetros libres.

Función de Decisión & Support Vectors

- Otra observación interesante se obtiene al notar que

$$
\begin{aligned}
f(x) &= \text{sign}\left(\sum_{\ell} \alpha_\ell \, y^{(\ell)} \langle x^{(\ell)}, x \rangle\right) \\
&= \text{sign}\left(\sum_{\ell : y^{(\ell)} = +1} \alpha_\ell \, \langle x^{(\ell)}, x \rangle - \sum_{\ell : y^{(\ell)} = -1} \alpha_\ell \, \langle x^{(\ell)}, x \rangle\right) \\
&= \text{sign}\left(\langle p_+, x \rangle - \langle p_-, x \rangle\right) ,
\end{aligned}
$$

- En otras palabras, para decidir sobre la clase de un dato nuevo, la SVM compara x con los support vectors positivos (vía el producto punto), luego con los support vectors negativos (vía el producto punto) y finalmente decide predecir la clase cuyos support vectors "se parecen más" a x.

Geometría de los Support Vectors

- En el caso de la SVM dura, el análisis de la solución del problema de optimización (condiciones KKT), revela que

$$\alpha_\ell \left(y^{(\ell)} \left(w^T x^{(\ell)} + b \right) - 1 \right) = 0 \ \forall \ell \tag{11}$$

- Eso significa que si $\alpha_\ell \neq 0$, debe ocurrir,

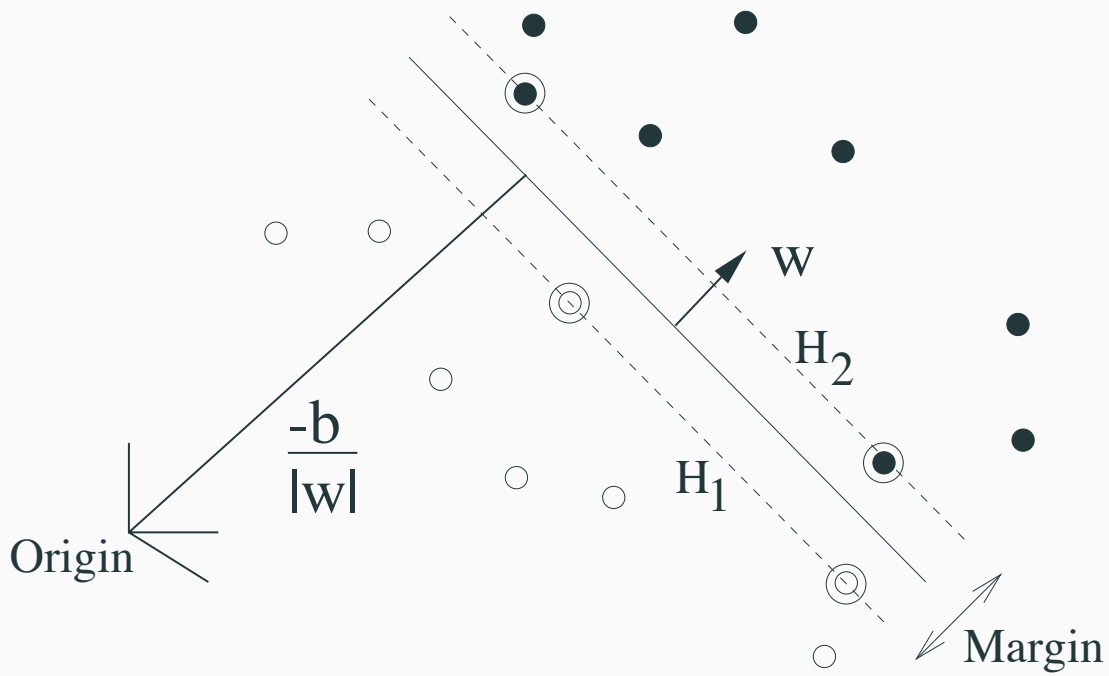
$$y^{(\ell)} \left(w^T x^{(\ell)} + b \right) = 1 \tag{12}$$

- Deshaciendo la normalización que nos había llevado a la formulación, obtenemos que

$$\frac{y^{(\ell)} \left(w^T x^{(\ell)} + b \right)}{\|w\|} = \rho \tag{13}$$

Geometría de los Support Vectors

• Esto significa que los support vectors son los puntos de cada clase que quedan más cerca de la frontera de decisión (conceptualmente, aquellos que se confunden más).



Geometría de los Support Vectors

- En el caso de la SVM blanda, el análisis de la solución del problema de optimización (condiciones KKT), revela que

$$\alpha_\ell \left(y^{(\ell)} \left(w^T x^{(\ell)} + b \right) - 1 + \xi_\ell \right) = 0 \ \forall \ell \tag{14}$$

$$C - \alpha_\ell = \beta_\ell \tag{15}$$

$$y^{(\ell)} \left(w^T x^{(\ell)} + b \right) \geq 1 - \xi_\ell \ \forall \ell \tag{16}$$

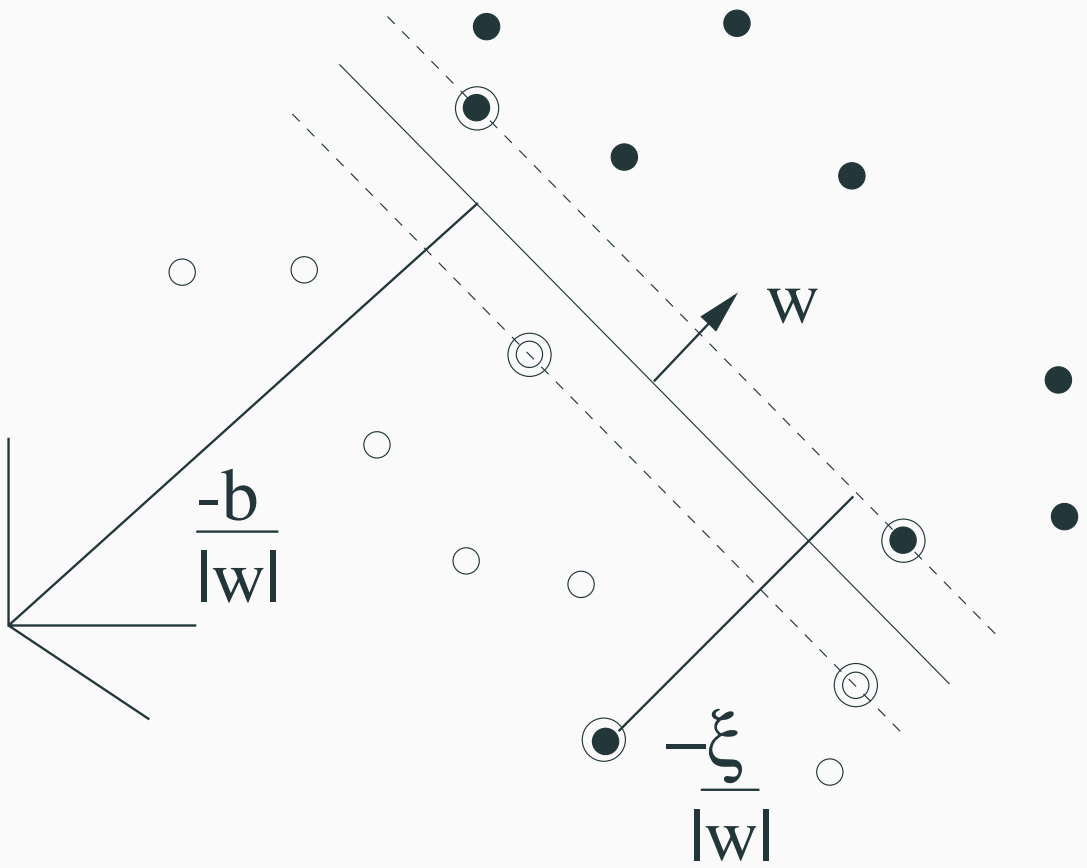
$$\beta_\ell \xi_\ell = 0 \ \forall \ell \tag{17}$$

$$\beta_\ell, \xi_\ell \geq 0 \ \forall \ell \tag{18}$$

- Por lo tanto, hay dos tipos de support vectors ...

Support Vectors

- Support vectors duros para los cuales $y^{(\ell)}\left(w^T x^{(\ell)} + b\right) = 1$.
- Support vectors blandos para los cuales $y^{(\ell)}\left(w^T x^{(\ell)} + b\right) = 1 - \xi_\ell$ con $\xi_\ell \neq 0$ (para estos, es seguro que $\alpha_\ell = C$ y sólo estos pueden corresponder a errores de clasificación).



Apéndice: Notas de Optimización Convexa

Optimización con Restricciones & Lagrangiana

En este y otros capítulos estamos aparecen con frecuencia problemas de optimización con restricciones de la forma

$$\min_z g(z) \;\; \text{s.t. } c_\ell(z) \geq 0 \; \forall \ell = 1, \ldots, n$$

En el estudio de estos problemas es frecuente considerar una función denominada *la Lagrangiana* asociada al problema y definida como:

$$\mathcal{L}(\alpha, z) = g(z) - \sum_\ell \alpha_\ell \, c_\ell(z) \tag{19}$$

donde las variables $\alpha_\ell \in \mathbb{R}_0^+$ se denominan *multiplicadores de Lagrange*.

Para entender la utilidad de esta función, consideremos la función de z

$$P(z) = \max_{\alpha \geq 0} \mathcal{L}(\alpha, z) \tag{20}$$

Notemos que como $\mathcal{L}(\alpha, z) = g(z) - \sum_\ell \alpha_\ell \, c_\ell(z)$, se satisface lo siguiente:

$$P(z) = \begin{cases} \infty & \text{si } \exists \ell : \; c_\ell(z) < 0 \\ g(z) & \text{si } \forall \ell : \; c_\ell(z) \geq 0 \end{cases}, \tag{21}$$

de modo que la solución de

$$\min_z P(z) = \min_z \max_{\alpha \geq 0} \mathcal{L}(\alpha, z) \tag{22}$$

es equivalente a resolver problema de optimización original, ya que para obtener el mini-max nos convendrá elegir siempre z tal que $c_\ell(z) \geq 0, \forall \ell$ (factibilidad) y α_ℓ tal que $\alpha_\ell \, c_\ell(z) = 0$ (holgura complementaria).

Optimización con Restricciones & Lagrangiana

El problema anterior tiene restricciones más simples que el original. Sin embargo, la ventaja más importante de lo anterior es que, con frecuencia, podremos hacer el siguiente truco: en vez de encontrar la solución (α^*, z^*) de

$$\min_{z} \max_{\alpha \geq 0} \mathcal{L}(\alpha, z) \tag{23}$$

uno busca la solución (α^*, z^*) de

$$\max_{\alpha \geq 0} \min_{z} \mathcal{L}(\alpha, z) \tag{24}$$

El primer problema de optimización se denomina problema primal, mientras que el segundo se denomina problema dual.

Resolver el problema dual puede resultar más fácil.

Como es fácil demostrar que

$$\max_{\alpha \geq 0} \min_{z} \mathcal{L}(\alpha, z) \quad \leq \quad \min_{z} \max_{\alpha \geq 0} \mathcal{L}(\alpha, z) \tag{25}$$

resolver el dual nos da siempre una cota inferior a la solución del primal.

Sin embargo, bajo algunas condiciones

$$\max_{\alpha \geq 0} \min_{z} \mathcal{L}(\alpha, z) \quad = \quad \min_{z} \max_{\alpha \geq 0} \mathcal{L}(\alpha, z) \tag{26}$$

Esas condiciones requieren la convexidad de $g(z)$, la linealidad de $c_\ell(z)$ y algunas condiciones técnicas adicionales (por ejemplo, que exista z, tal que $g(z) > 0$ es suficiente).

Optimización con Restricciones & Lagrangiana

Si además de las condiciones anteriores, se tiene que las funciones involucradas son diferenciables, se tiene que el par (α^*, z^*) que resuelve el dual y el primal, satisface

$$\begin{aligned} \frac{\partial}{\partial z}\mathcal{L}(\alpha^*, z^*) &= 0 \\ \alpha_\ell \, c_\ell(z) &= 0 \ \forall \ell \\ c_\ell(z) &\geq 0 \ \forall \ell \\ \alpha_\ell &\geq 0 \ \forall \ell \end{aligned} \tag{27}$$

Las condiciones anteriores se denominan condiciones de Karush-Kuhn-Tucker (KKT).

Usando las KKT para encontrar z como función de α, permite en ocasiones resolver el problema dual de modo más simple que el primal.

Solución de la SVM Dura

Nuestra SVM de margen rígido, corresponde al problema de optimización convexo

$$\begin{aligned} &\min_{w,b} \ \frac{1}{2}\|w\|^2 \\ &\text{s.t. } y^{(\ell)}\left(w^T x^{(\ell)} + b\right) \geq 1 \ \forall \ell \end{aligned}$$

cuya Lagrangiana toma la forma

$$\mathcal{L}(\alpha, w, b) = \frac{1}{2}\|w\|^2 - \sum_{\ell} \alpha_\ell \left(y^{(\ell)}\left(w^T x^{(\ell)} + b\right) - 1\right) \tag{28}$$

Solución de la SVM Dura

Las KKT corresponden a las siguientes condiciones

$$\frac{\partial}{\partial w}\mathcal{L}(\alpha, w, b) = w - \sum_{\ell} \alpha_\ell \, y^{(\ell)} x^{(\ell)} = 0 \tag{29}$$

$$\frac{\partial}{\partial b}\mathcal{L}(\alpha, w, b) = -\sum_{\ell} \alpha_\ell \, y^{(\ell)} = 0 \tag{30}$$

$$\alpha_\ell \left(y^{(\ell)} \left(w^T x^{(\ell)} + b \right) - 1 \right) = 0 \ \forall \ell \tag{31}$$

$$y^{(\ell)} \left(w^T x^{(\ell)} + b \right) \geq 1 \ \forall \ell \tag{32}$$

$$\alpha_\ell \geq 0 \ \forall \ell \tag{33}$$

La condición (29), implica que la solución del problema original satisface

$$w = \sum_{\ell} \alpha_\ell \, y^{(\ell)} x^{(\ell)} \, , \tag{34}$$

lo que nos permite obtener w como función de α.

Dual de la SVM Dura

Reemplazando (34) en la Lagrangiana, y usando (30), tenemos para la solución w, b del problema primal

$$\mathcal{L}(\alpha, w, b) = \sum_{\ell} \alpha_\ell - \frac{1}{2} \sum_{\ell,m} \alpha_\ell \, \alpha_m \, y^{(\ell)} y^{(m)} \langle x^{(\ell)}, x^{(m)} \rangle \tag{35}$$

Juntando esto con las otras KKT, tenemos que el problema dual de nuestra SVM tiene la forma

$$\begin{aligned} \max_{\alpha} \quad & \sum_{\ell} \alpha_\ell - \frac{1}{2} \sum_{\ell,m} \alpha_\ell \, \alpha_m \, y^{(\ell)} y^{(m)} \langle x^{(\ell)}, x^{(m)} \rangle \\ \text{s.t.} \quad & \sum_{\ell} \alpha_\ell \, y^{(\ell)} = 0, \quad \alpha_\ell \geq 0 \; \forall \ell \, . \end{aligned} \tag{36}$$

(Como ya habíamos observado, las KKT (31, holgura complementaria) y (32, factibilidad) las satisface siempre el óptimo del problema primal, que coincide con el dual).

Solución de la SVM Blanda

Nuestra SVM de margen blando, corresponde al problema de optimización convexo

$$\begin{aligned} \min_{w,b,\xi} \ & \frac{1}{2}\|w\|^2 + \boldsymbol{C} \cdot \sum\nolimits_{\ell} \xi_\ell \\ \text{s.t. } & y^{(\ell)}(w^T x^{(\ell)} + b) \geq 1 - \xi_\ell \, \forall \ell \\ & \xi_\ell \geq 0 \, \forall \ell \end{aligned}$$

cuya Lagrangiana toma la forma

$$\begin{aligned} \mathcal{L}(\alpha, \beta, w, b, \xi) = \frac{1}{2}\|w\|^2 + C \cdot \sum\nolimits_{\ell} \xi_\ell \\ - \sum_{\ell} \alpha_\ell \left(y^{(\ell)} \left(w^T x^{(\ell)} + b \right) - 1 + \xi_\ell \right) - \sum_{\ell} \beta_\ell \xi_\ell \end{aligned} \tag{37}$$

(38)

Solución de la SVM Blanda

Las KKT corresponden a las siguientes condiciones

$$\frac{\partial}{\partial w}\mathcal{L}(\alpha,\beta,w,b,\xi) = w - \sum_{\ell} \alpha_\ell\, y^{(\ell)} x^{(\ell)} = 0 \tag{39}$$

$$\frac{\partial}{\partial b}\mathcal{L}(\alpha,\beta,w,b,\xi) = -\sum_{\ell} \alpha_\ell\, y^{(\ell)} = 0 \tag{40}$$

$$\frac{\partial}{\partial \xi_\ell}\mathcal{L}(\alpha,\beta,w,b,\xi) = C - \alpha_\ell - \beta_\ell = 0 \tag{41}$$

$$\alpha_\ell \left(y^{(\ell)} \left(w^T x^{(\ell)} + b \right) - 1 + \xi_\ell \right) = 0 \ \forall \ell \tag{42}$$

$$y^{(\ell)} \left(w^T x^{(\ell)} + b \right) \geq 1 - \xi_\ell \ \forall \ell \tag{43}$$

$$\beta_\ell \xi_\ell = 0 \ \forall \ell \tag{44}$$

$$\xi_\ell \geq 0 \ \forall \ell \tag{45}$$

$$\alpha_\ell, \beta_\ell \geq 0 \ \forall \ell \tag{46}$$

Dual de la SVM Blanda

Reemplazando (39) en la Lagrangiana, y usando (40) y (41), tenemos para la solución w, b del problema primal

$$\mathcal{L}(\alpha, w, b) = \sum_{\ell} \alpha_\ell - \frac{1}{2} \sum_{\ell,m} \alpha_\ell \, \alpha_m \, y^{(\ell)} y^{(m)} \langle x^{(\ell)}, x^{(m)} \rangle \tag{47}$$

Juntando esto con las otras KKT, tenemos que el problema dual de la SVM blanda tiene la forma

$$\begin{aligned} \max_{\alpha} \quad & \sum_{\ell} \alpha_\ell - \frac{1}{2} \sum_{\ell,m} \alpha_\ell \, \alpha_m \, y^{(\ell)} y^{(m)} \langle x^{(\ell)}, x^{(m)} \rangle \\ \text{s.t.} \quad & \sum_{\ell} \alpha_\ell \, y^{(\ell)} = 0, \quad 0 \leq \alpha_\ell \leq C \ \forall \ell \, . \end{aligned} \tag{48}$$

La única diferencia con el dual de la SVM dura es que ahora aparece un límite superior para α_ℓ.

Referencias


Alex J Smola and Bernhard Schölkopf.
Learning with Kernels.
MIT press, 2001.


Sebastian Raschka.
Python Machine Learning.
Packt Publishing Ltd, 2015.


M Bishop Christopher.
Pattern Recognition and Machine Learning.
Springer-Verlag New York, 2016.