Aprendizaje de máquinas
Support vector machines (parte 1)

Felipe Tobar

Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile

Otoño, 2021.

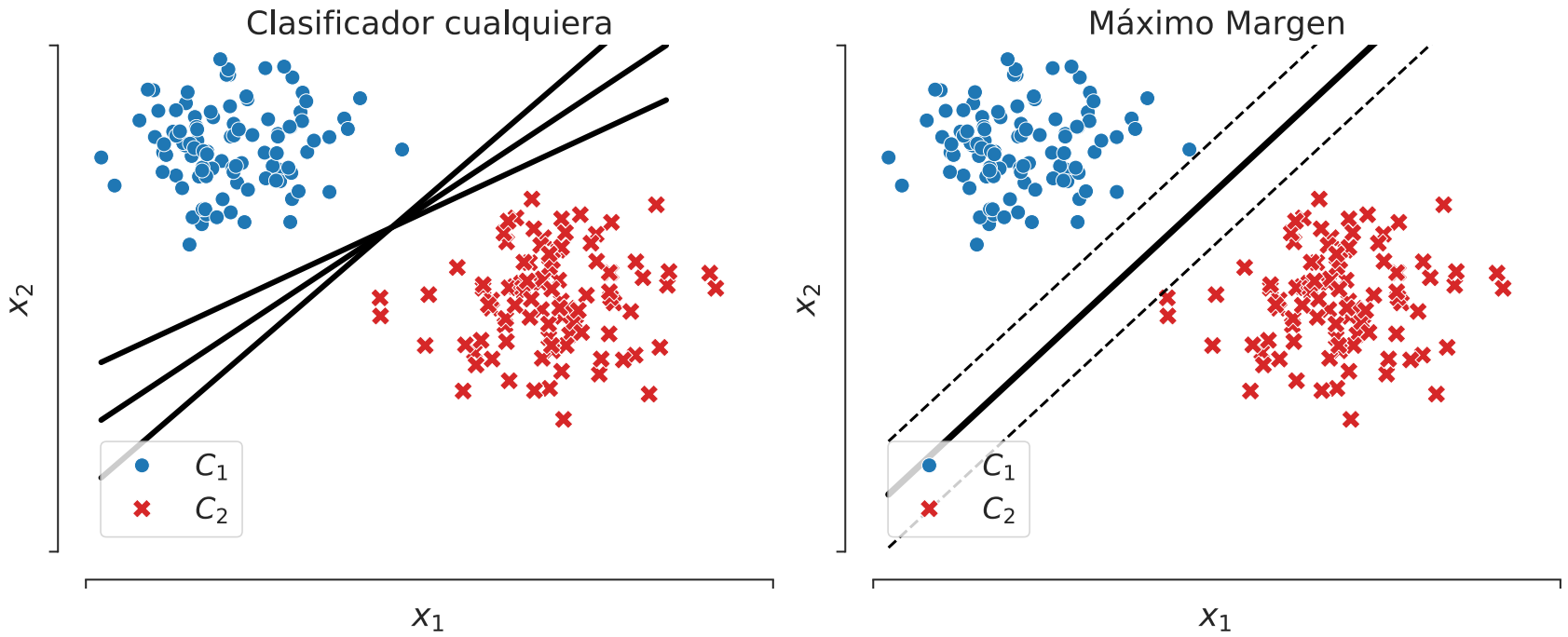
Idea general

Para un problema de clasificación donde las clases son linealmente separables, por lo general existen infinitos hiperplanos separadores.

Idea general

Para un problema de clasificación donde las clases son linealmente separables, por lo general existen infinitos hiperplanos separadores.

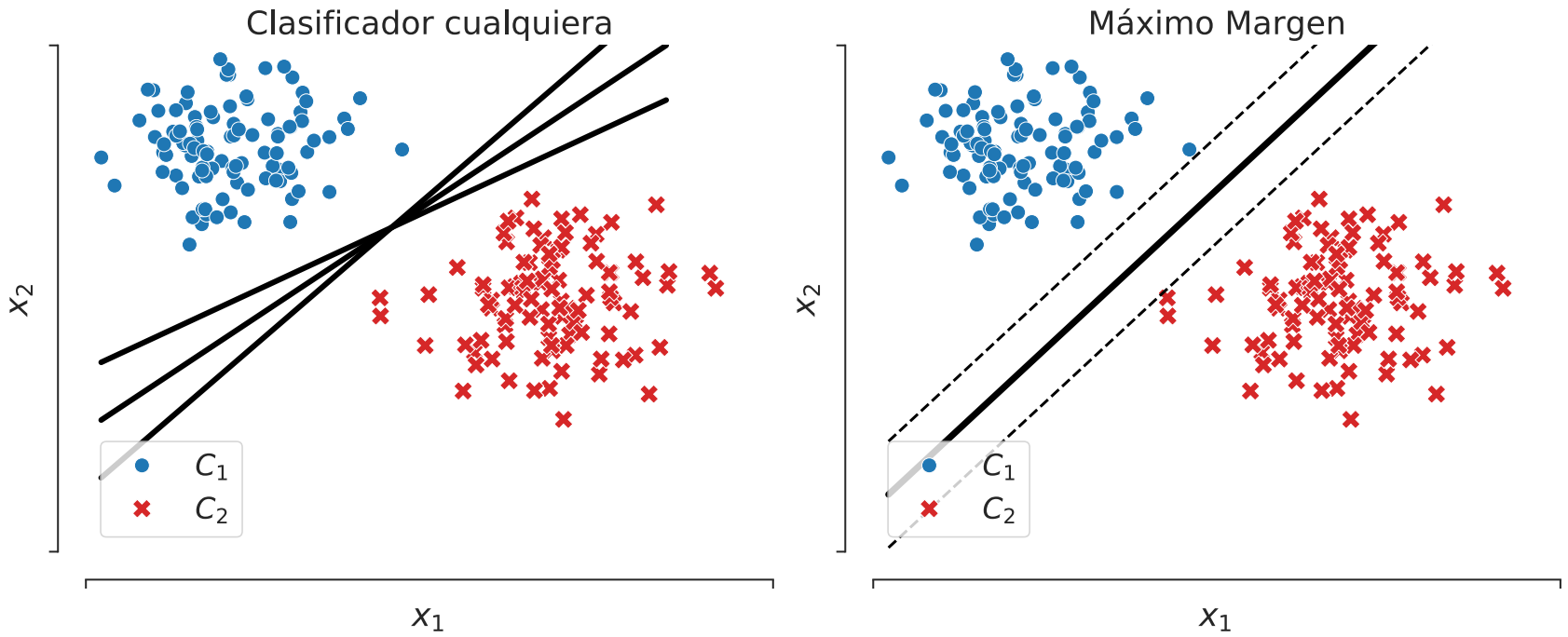
Como se puede ver en la siguiente figura, la asignación de clase es similar en la mayoría del espacio salvo la región cercana al límite de las clases.



Idea general

Para un problema de clasificación donde las clases son linealmente separables, por lo general existen infinitos hiperplanos separadores.

Como se puede ver en la siguiente figura, la asignación de clase es similar en la mayoría del espacio salvo la región cercana al límite de las clases.



SVM busca el hiperplano separador que maximice el margen entre las clases, lo cual es equivalente a buscar una cinta de ancho máximo que separe los datos.

Ventajas de SVM

Encontrar este clasificador tiene ventajas sobre los otros estudiados ya que:

- Si los datos generados en cada clase provienen de una distribución latente, es de esperar que si se obtienen nuevos datos desde la misma distribución, estos estén cerca de los datos observados inicialmente. De este forma, SVM tiene buenas propiedades de generalización.

Ventajas de SVM

Encontrar este clasificador tiene ventajas sobre los otros estudiados ya que:

- Si los datos generados en cada clase provienen de una distribución latente, es de esperar que si se obtienen nuevos datos desde la misma distribución, estos estén cerca de los datos observados inicialmente. De este forma, SVM tiene buenas propiedades de generalización.
- Como se verá a continuación, el clasificador de máximo margen queda definido únicamente por algunos datos. Esto inmediatamente resuelve el problema de desbalances de clase o de que las clases tengan formas distintas.

Ventajas de SVM

Encontrar este clasificador tiene ventajas sobre los otros estudiados ya que:

- Si los datos generados en cada clase provienen de una distribución latente, es de esperar que si se obtienen nuevos datos desde la misma distribución, estos estén cerca de los datos observados inicialmente. De este forma, SVM tiene buenas propiedades de generalización.
- Como se verá a continuación, el clasificador de máximo margen queda definido únicamente por algunos datos. Esto inmediatamente resuelve el problema de desbalances de clase o de que las clases tengan formas distintas.
- Por lo anterior, la solución de máximo margen se mantiene si agregamos datos que estén fuera del margen. Los datos que definen el margen los llamaremos vectores de soporte, cuya función será restringir la rotación y expansión del margen.

Formulación del problema: restricción sobre los vectores soporte

Dado un conjunto de entrenamiento $\{x_i\}_{i=1}^N$ linealmente separable, con clases $\{1, -1\}$, un hiperplano separador está definido mediante

$$\{x \in \mathbb{R}^n | w^\top x + b = 0\}$$

Donde si $w^\top x + b > 0$ se le asignará a x la clase 1, mientras que si $w^\top x + b < 0$ se le asignará a x la clase -1.

Formulación del problema: restricción sobre los vectores soporte

Dado un conjunto de entrenamiento $\{x_i\}_{i=1}^N$ linealmente separable, con clases $\{1, -1\}$, un hiperplano separador está definido mediante

$$\{x \in \mathbb{R}^n | w^\top x + b = 0\}$$

Donde si $w^\top x + b > 0$ se le asignará a x la clase 1, mientras que si $w^\top x + b < 0$ se le asignará a x la clase -1.

El par (w, b) no es único ya que si (w, b) forma un hiperplano separador, entonces $(\lambda w, \lambda b)$ también lo hará.

Formulación del problema: restricción sobre los vectores soporte

Dado un conjunto de entrenamiento $\{x_i\}_{i=1}^N$ linealmente separable, con clases $\{1, -1\}$, un hiperplano separador está definido mediante

$$\{x \in \mathbb{R}^n | w^\top x + b = 0\}$$

Donde si $w^\top x + b > 0$ se le asignará a x la clase 1, mientras que si $w^\top x + b < 0$ se le asignará a x la clase -1.

El par (w, b) no es único ya que si (w, b) forma un hiperplano separador, entonces $(\lambda w, \lambda b)$ también lo hará.

Para evitar este escalamiento se impondrá una restricción sobre los bordes del margen. Sean x_+ y x_- los vectores soporte de cada clase, se impone que estos pertenezcan a su respectiva clase, es decir:

$$w^\top x_+ + b = 1$$
$$w^\top x_- + b = -1,$$

donde estos vectores soportes puede no ser únicos. Observemos que si bien aún no sabemos cuales son los vectores de soporte, podemos imponer esta restricción de todas formas.

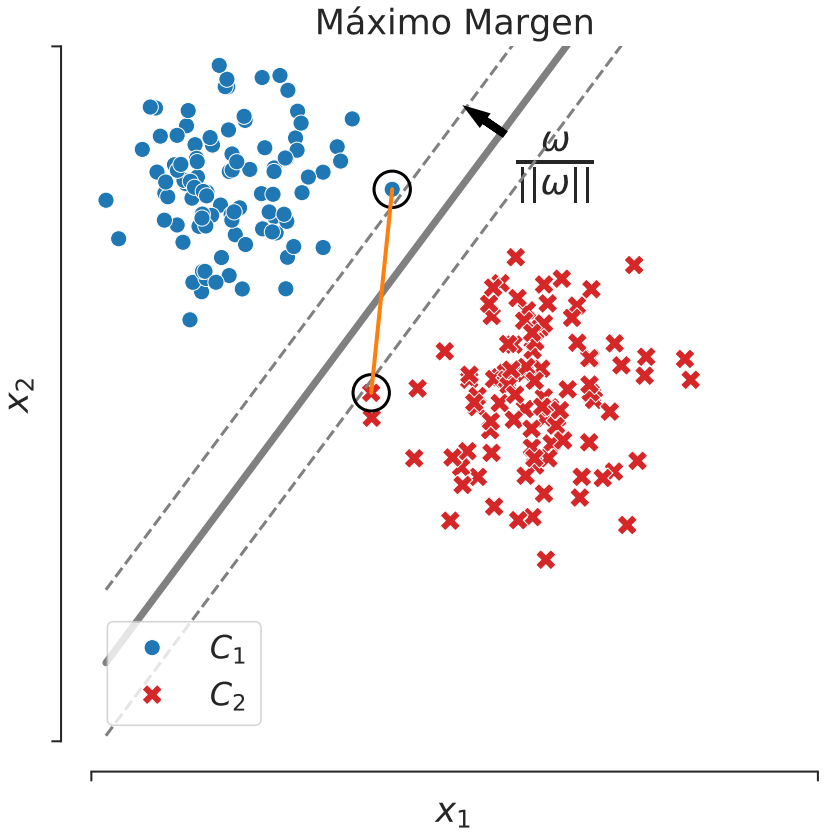
Formulación del problema: ancho del margen

Las restricciones anteriores definen hiperplanos paralelos a la región de decisión ya que tienen el mismo parámetro w.

Formulación del problema: ancho del margen

Las restricciones anteriores definen hiperplanos paralelos a la región de decisión ya que tienen el mismo parámetro w.

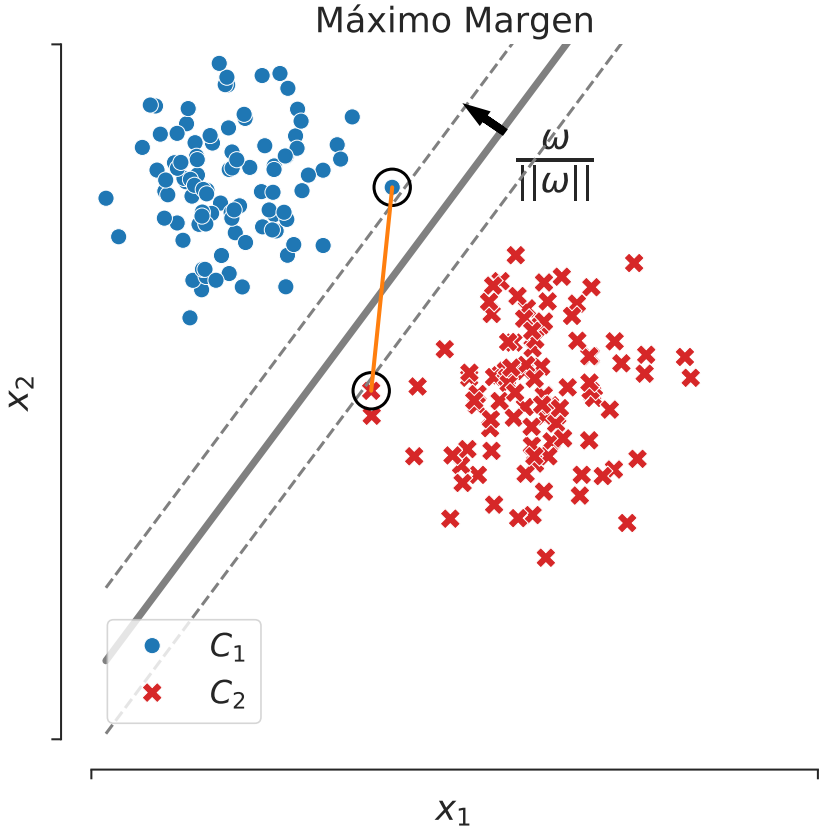
Por otro lado, el ancho del margen m, es la distancia entre la región de decisión y cualquiera de las clases.



Formulación del problema: ancho del margen

Las restricciones anteriores definen hiperplanos paralelos a la región de decisión ya que tienen el mismo parámetro w.

Por otro lado, el ancho del margen m, es la distancia entre la región de decisión y cualquiera de las clases.



En la figura se observa que el ancho del margen puede ser calculado como una proyección sobre la dirección normal del hiperplano, es decir,

$$
\begin{aligned}
m &= \frac{1}{2}\left\|\text{proy}_{\text{w}}(\text{x}_+ - \text{x}_-)\right\| \\
&= \frac{1}{2}\left\|x_+ - x_-\right\|\cos(\theta) \\
&= \frac{1}{2}\left\|x_+ - x_-\right\|\left(\frac{w^\top(x_+ - x_-)}{\|w\|\cdot\|x_+ - x_-\|}\right) \\
&= \frac{1}{2\|w\|}w^\top(x_+ - x_-)
\end{aligned}
$$

Donde se usó el hecho de que $\cos\left(\measuredangle(x,y)\right) = \frac{\langle x,y\rangle}{\|x\|\|y\|}$.

Formulación del problema: problema de optimización

Además, el ancho del margen m no depende explícitamente de los vectores soporte:

$$m = \frac{1}{2||w||}\left((w^\top x_+) - (w^\top x_-)\right) = \frac{1}{2||w||}\left((1-b) - (-1-b)\right) = \frac{1}{||w||}$$

Formulación del problema: problema de optimización

Además, el ancho del margen m no depende explícitamente de los vectores soporte:

$$m = \frac{1}{2||w||}\left((w^\top x_+) - (w^\top x_-)\right) = \frac{1}{2||w||}\left((1-b) - (-1-b)\right) = \frac{1}{||w||}$$

Por otra parte, se considerará la siguiente codificación para las clases:

$$y_i = +1 \Leftrightarrow w^\top x_i + b \geq +1$$
$$y_i = -1 \Leftrightarrow w^\top x_i + b \leq -1$$

Formulación del problema: problema de optimización

Además, el ancho del margen m no depende explícitamente de los vectores soporte:

$$m = \frac{1}{2||w||}\left((w^\top x_+) - (w^\top x_-)\right) = \frac{1}{2||w||}\left((1-b) - (-1-b)\right) = \frac{1}{||w||}$$

Por otra parte, se considerará la siguiente codificación para las clases:

$$y_i = +1 \Leftrightarrow w^\top x_i + b \geq +1$$
$$y_i = -1 \Leftrightarrow w^\top x_i + b \leq -1$$

Por lo tanto, se puede formular el problema de clasificación de máximo margen mediante el siguiente problema de optimización:

$$\begin{aligned} \max_{w,b} \quad & \frac{1}{||w||} \\ \text{s.a} \quad & y_i(w^\top x_i + b) \geq 1, \ i = 1, \ldots, N. \end{aligned}$$

Donde las restricciones exigen que todas las muestras sean bien clasificadas.

Formulación del problema: problema dual

Para evitar problemas de diferenciablidad, se considerará la siguiente formulación equivalente del problema anterior:

$$
\begin{aligned}
(P) \quad & \min_{w,b} \quad \frac{1}{2}||w||^2 \\
& \text{s.a} \quad y_i(w^\top x_i + b) \geq 1, \ i = 1, \ldots, N.
\end{aligned}
$$

Formulación del problema: problema dual

Para evitar problemas de diferenciablidad, se considerará la siguiente formulación equivalente del problema anterior:

$$\begin{aligned} (P) \quad & \min_{w,b} \quad \frac{1}{2}||w||^2 \\ & \text{s.a} \quad y_i(w^\top x_i + b) \geq 1, \ i = 1, \ldots, N. \end{aligned}$$

Este problema de optimización con restricciones puede ser resuelto mediante dualidad lagrangiana:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 + \sum_{i=1}^{N} \alpha_i \left(1 - y_i(w^\top x_i + b)\right)$$

Formulación del problema: problema dual

Para evitar problemas de diferenciablidad, se considerará la siguiente formulación equivalente del problema anterior:

$$\begin{aligned} (P) \quad & \min_{w,b} \quad \frac{1}{2}||w||^2 \\ & \text{s.a} \quad y_i(w^\top x_i + b) \geq 1, \ i = 1, \ldots, N. \end{aligned}$$

Este problema de optimización con restricciones puede ser resuelto mediante dualidad lagrangiana:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 + \sum_{i=1}^{N} \alpha_i \left(1 - y_i(w^\top x_i + b)\right)$$

El lagrangiano dual del problema viene dado por $\theta(\alpha) = \inf_{w,b} L(w, b, \alpha)$ y dado que L es convexo, basta aplicar la condición de primer orden:

$$\frac{\partial L}{\partial w} = w^\top - \sum_{i=1}^{N} \alpha_i y_i x_i^\top = 0 \implies \overline{w} = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \alpha_i y_i = 0 \implies \sum_{i=1}^{N} \alpha_i y_i = 0$$

Formulación del problema: problema final

Sustituyendo en L y simplificando, el lagrangiano dual tiene la siguiente forma:

$$\theta(\alpha) = \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i y_i \alpha_j y_j \langle x_i, x_j \rangle + \sum_{i=1}^{N} \alpha_i - \sum_{i,j=1}^{N} \alpha_i y_i \alpha_j y_j \langle x_j, x_i \rangle = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

Formulación del problema: problema final

Sustituyendo en L y simplificando, el lagrangiano dual tiene la siguiente forma:

$$\theta(\alpha) = \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i y_i \alpha_j y_j \langle x_i, x_j \rangle + \sum_{i=1}^{N} \alpha_i - \sum_{i,j=1}^{N} \alpha_i y_i \alpha_j y_j \langle x_j, x_i \rangle = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

Finalmente, el problema dual consiste en maximizar $\theta(\alpha)$ sujeto a que $\alpha \geq 0$, es decir:

$$\begin{aligned} (D) \quad \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{s.a} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \end{aligned}$$

Donde la primera restricción se heredó de la CPO impuesta sobre L al calcular $\theta(\alpha)$.

Observaciones de la formulación

La forma usual para resolver SVM es mediante la resolución de su problema dual. Para este problema se tienen las siguientes observaciones:

- La función objetivo es una forma cuadrática definida negativa.

Observaciones de la formulación

La forma usual para resolver SVM es mediante la resolución de su problema dual. Para este problema se tienen las siguientes observaciones:

- La función objetivo es una forma cuadrática definida negativa.
- Lo anterior implica que el problema tiene un único máximo.

Observaciones de la formulación

La forma usual para resolver SVM es mediante la resolución de su problema dual. Para este problema se tienen las siguientes observaciones:

- La función objetivo es una forma cuadrática definida negativa.
- Lo anterior implica que el problema tiene un único máximo.
- Una vez resuelta la formulación dual (i.e., se han encontrado los valores óptimos para α), la predicción de un nuevo punto $x_\star$ es de la forma

$$\hat{y}(x_\star) = \operatorname{sgn}(\overline{w}^\top x_\star + b) = \operatorname{sgn}\left(\left[\sum_{i=1}^{N} \alpha_i y_i \langle x_i, x_\star \rangle\right] + b\right),$$

Observaciones de la formulación

La forma usual para resolver SVM es mediante la resolución de su problema dual. Para este problema se tienen las siguientes observaciones:

- La función objetivo es una forma cuadrática definida negativa.
- Lo anterior implica que el problema tiene un único máximo.
- Una vez resuelta la formulación dual (i.e., se han encontrado los valores óptimos para α), la predicción de un nuevo punto $x_\star$ es de la forma

$$\hat{y}(x_\star) = \operatorname{sgn}(\overline{w}^\top x_\star + b) = \operatorname{sgn}\left(\left[\sum_{i=1}^{N} \alpha_i y_i \langle x_i, x_\star \rangle\right] + b\right),$$

- Por el teorema de holgura complementaria, para α óptimo se tiene que

$$\alpha_i \left(1 - y_i(\overline{w}^\top x_i + b)\right) = 0, \quad \forall i \in \{1, \ldots, N\}$$

  Por lo que $\alpha_i = 0$ para todo x_i fuera del margen y, consecuentemente, x_i no aporta en la predicción $\hat{y}$.

Problemas del planteamiento anterior

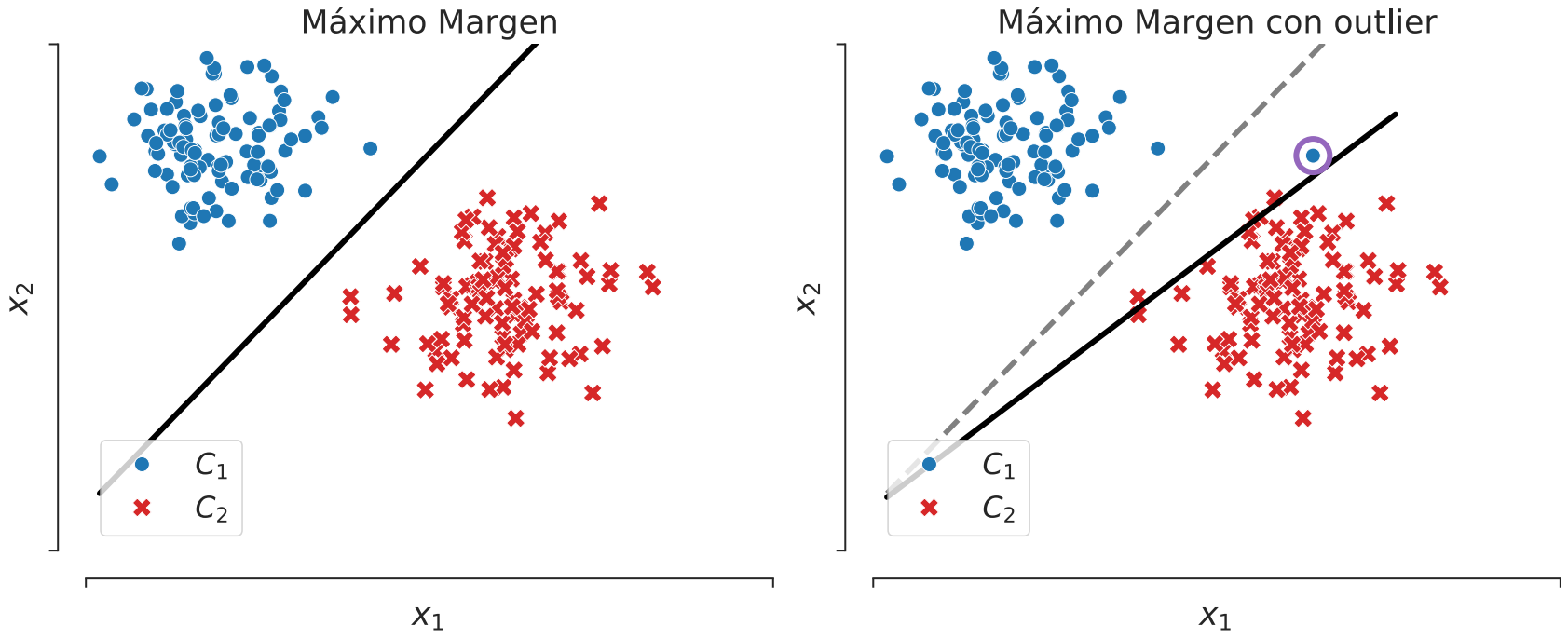
El planteamiento anterior tiene dos debilidades:

- Los datos no siempre serán separables por lo que el problema puede ser infactible.

Problemas del planteamiento anterior

El planteamiento anterior tiene dos debilidades:

- Los datos no siempre serán separables por lo que el problema puede ser infactible.
- Incluso si los datos son linealmente separables, el clasificador puede ser muy sensible a nuevos datos. Esto se observa en la siguiente figura:

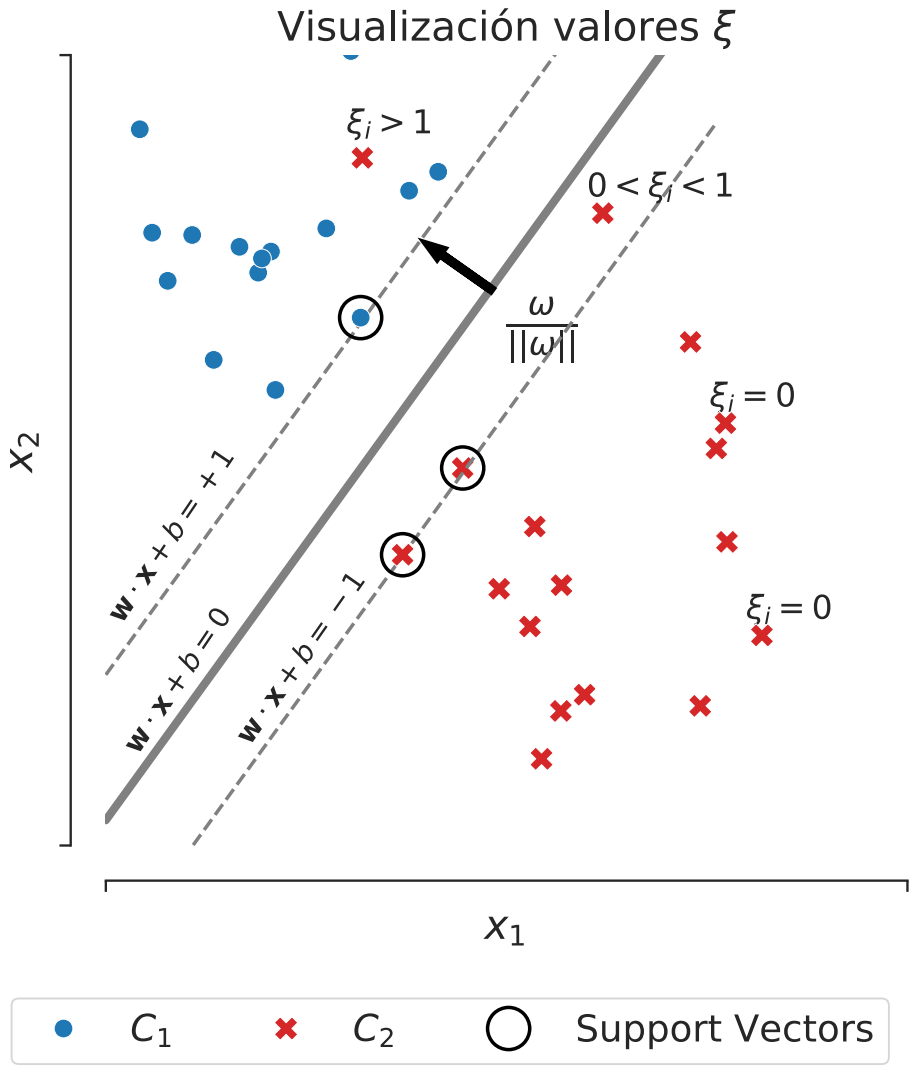


Margen suave: variables de holgura

Para corregir las debilidades anteriores, se pueden introducir "variables de holgura", las cuales tienen el objetivo de permitir al clasificador admitir algunos datos incorrectamente clasificados.

Margen suave: variables de holgura

Para corregir las debilidades anteriores, se pueden introducir "variables de holgura", las cuales tienen el objetivo de permitir al clasificador admitir algunos datos incorrectamente clasificados.



La formulación que *perdona* algunos datos mal clasificados mediante la utilización de variables de holgura $\{\xi_i\}_{i=1}^N$ es la siguiente:

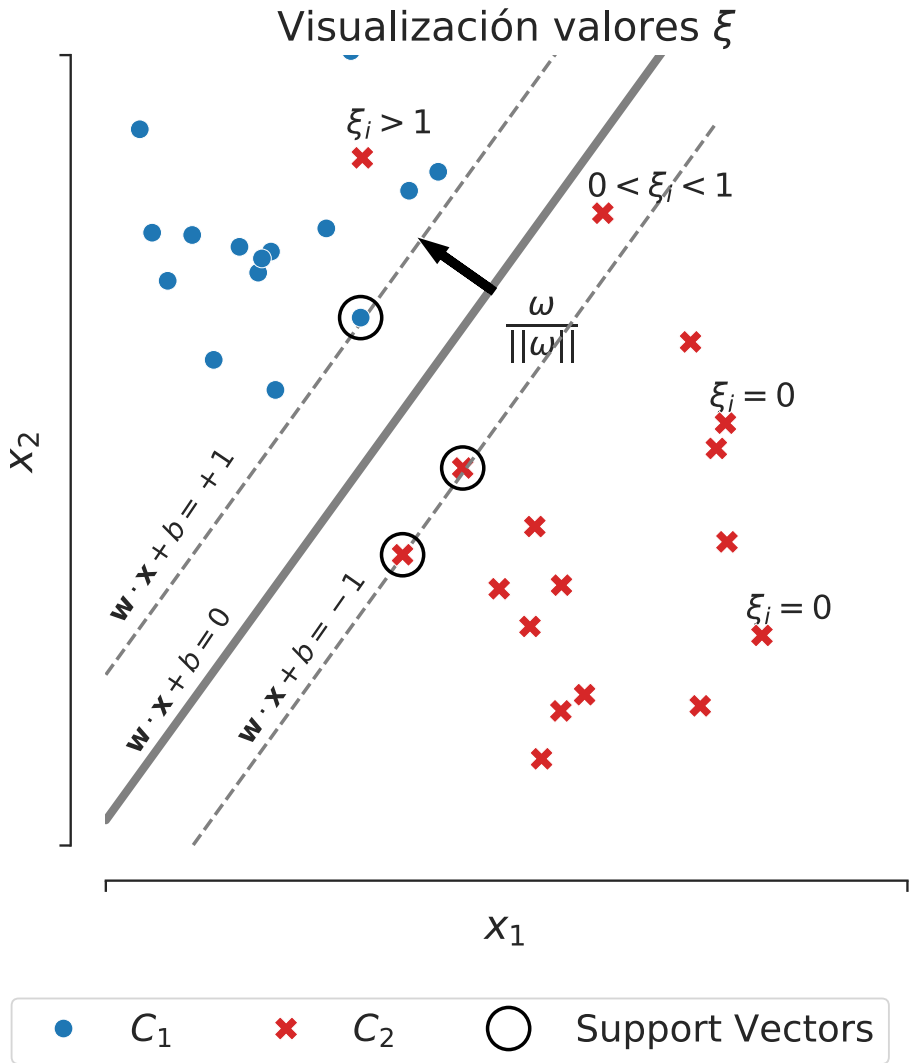
$$(P) \quad \min_{w,b,\xi} \quad \frac{1}{2}||w||^2 + c\sum_{i=1}^{N}\xi_i$$

$$\text{s.a} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0$$

donde $c > 0$ es un hiperparámetro.

Margen suave: variables de holgura

Para corregir las debilidades anteriores, se pueden introducir "variables de holgura", las cuales tienen el objetivo de permitir al clasificador admitir algunos datos incorrectamente clasificados.



La formulación que *perdona* algunos datos mal clasificados mediante la utilización de variables de holgura $\{\xi_i\}_{i=1}^N$ es la siguiente:

$$(P) \quad \min_{w,b,\xi} \quad \frac{1}{2}||w||^2 + c\sum_{i=1}^{N}\xi_i$$

$$\text{s.a} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0$$

donde $c > 0$ es un hiperparámetro.
Se observa que la introducción del término $c\sum_{i=1}^{N}\xi_i$ en la función de costo puede ser interpretada como una regularización al igual que en MCR.

Margen suave: formulación

Procediendo de la misma forma que en el caso anterior, el dual de este problema es:

$$
\begin{aligned}
(D) \quad \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
\text{s.a} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq c.
\end{aligned}
$$

Margen suave: formulación

Procediendo de la misma forma que en el caso anterior, el dual de este problema es:

$$
\begin{aligned}
(D) \quad \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
\text{s.a} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq c.
\end{aligned}
$$

Se observa que:

- La diferencia entre ambas formulaciones está dada por el hecho de que los multiplicadores de Lagrange ahora están acotados por el hiperparámetro c.

Margen suave: formulación

Procediendo de la misma forma que en el caso anterior, el dual de este problema es:

$$
\begin{aligned}
(D) \quad \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
\text{s.a} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq c.
\end{aligned}
$$

Se observa que:

- La diferencia entre ambas formulaciones está dada por el hecho de que los multiplicadores de Lagrange ahora están acotados por el hiperparámetro c.
- c representa la importancia que se da a la suma de las variables de holgura versus el ancho del margen.

Margen suave: formulación

Procediendo de la misma forma que en el caso anterior, el dual de este problema es:

$$
\begin{aligned}
(D) \quad \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
\text{s.a} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq c.
\end{aligned}
$$

Se observa que:

- La diferencia entre ambas formulaciones está dada por el hecho de que los multiplicadores de Lagrange ahora están acotados por el hiperparámetro c.
- c representa la importancia que se da a la suma de las variables de holgura versus el ancho del margen.
- Cuando $c \to \infty$ se recupera la formulación original (margen duro).

Margen suave: formulación

Procediendo de la misma forma que en el caso anterior, el dual de este problema es:

$$
\begin{aligned}
(D) \quad \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
\text{s.a} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq c.
\end{aligned}
$$

Se observa que:

- La diferencia entre ambas formulaciones está dada por el hecho de que los multiplicadores de Lagrange ahora están acotados por el hiperparámetro c.
- c representa la importancia que se da a la suma de las variables de holgura versus el ancho del margen.
- Cuando $c \to \infty$ se recupera la formulación original (margen duro).
- La elección de este hiperparámetro puede realizarse mediante validación cruzada.