Fundamentos de aprendizaje automático
Clasificadores basados en vectores soportes

Juan Miguel Santos
Centro de investigación y desarrollo en informática aplicada (CIDIA)
Universidad Nacional de Hurlingham
2023

Separabilidad lineal

Hiperplano

- En un espacio p-dimensional, un hiperplano es un subespacio $(p-1)$-dimensional.
- En R^2 es una recta.
- En R^3 es un plano.

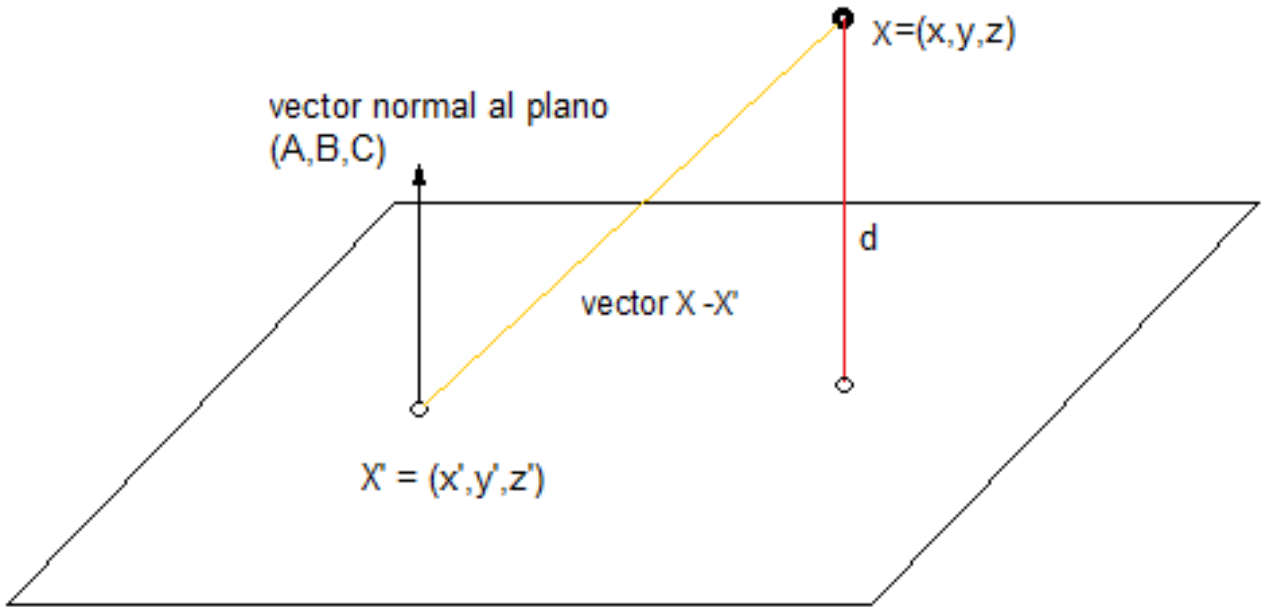
Separabilidad lineal

Ecuación de un hiperplano

- En un espacio p-dimensional un hiperplano está definido por $b_0 + b_1x_1 + b_2x_2 + .. + b_px_p = 0$
- En R^2, la recta queda definida por $b_0 + b_1x_1 + b_2x_2 = 0$ que quiere decir que todos los puntos $x = (x_1, x_2)$ que estén sobre la recta satisfacerán la ecuación.

Separabilidad lineal

Distancia de un punto X a un plano



Separabilidad lineal

Distancia de un punto X a un plano

Sea $Ax + By + Cz + D = 0$ la ecuación del plano.
En el punto $X' = (x', y', z')$ del plano se satisface que:

$$Ax' + By' + Cz' + D = 0$$

de donde se puede despejar D

$$D = -Ax' - By' - Cz'.$$

La distancia d será la proyección del vector $X - X'$ sobre el vector (A, B, C) ortogonal al plano cuya norma es 1:

$$d = \langle (X - X'), (A, B, C) \rangle$$

Separabilidad lineal

Distancia de un punto X a un plano

Desarrollando $d = \langle (X - X'), (A, B, C) \rangle$ tenemos:

$$d = (x - x')A + (y - y')B + (z - z')C =$$
$$d = xA - x'A + yB - y'B + zC - z'C$$

Agrupando los términos que tienen a x', y' y z', y usando que

$$D = -Ax' - By' - Cz':$$

podemos escribir a d como

$$d = D + xA + yB + zC =$$
$$d = \langle (x, y, z), (A, B, C) \rangle + D$$

Separabilidad lineal

Distancia de un punto X a un plano

Si consideramos el punto (x,y,z) en el origen (0,0,0) tenemos

$$d = \langle (x, y, z), (A, B, C) \rangle + D =$$
$$= \langle (0, 0, 0), (A, B, C) \rangle + D = D,$$

es decir, $d = D$.

Esto es, el término independiente **D** en

$$Ax + By + Cz + D = 0$$

es la **distancia del hiperplano al origen** (siempre, recordando que el vector (A,B,C) está normalizado).

Separabilidad lineal

- Ahora bien, consideremos la recta en R^2
$$b_0 + b_1 x_1 + b_2 x_2 = 0,$$
y el punto $x' = (x'_1, x'_2)$ tal que
$$b_0 + b_1 x'_1 + b_2 x'_2 > 0.$$
Esto quiere decir que el punto x' no está sobre la recta sino que está **de un lado** de la recta.
- En el caso que
$$b_0 + b_1 x'_1 + b_2 x'_2 < 0,$$
esto quiere decir que el punto x' está **del otro lado** de la recta.

Separabilidad lineal

- Ahora supongamos que tenemos un conjunto de n ejemplos de p atributos x_i y clase y_i con $(1 \leq i \leq n)$:

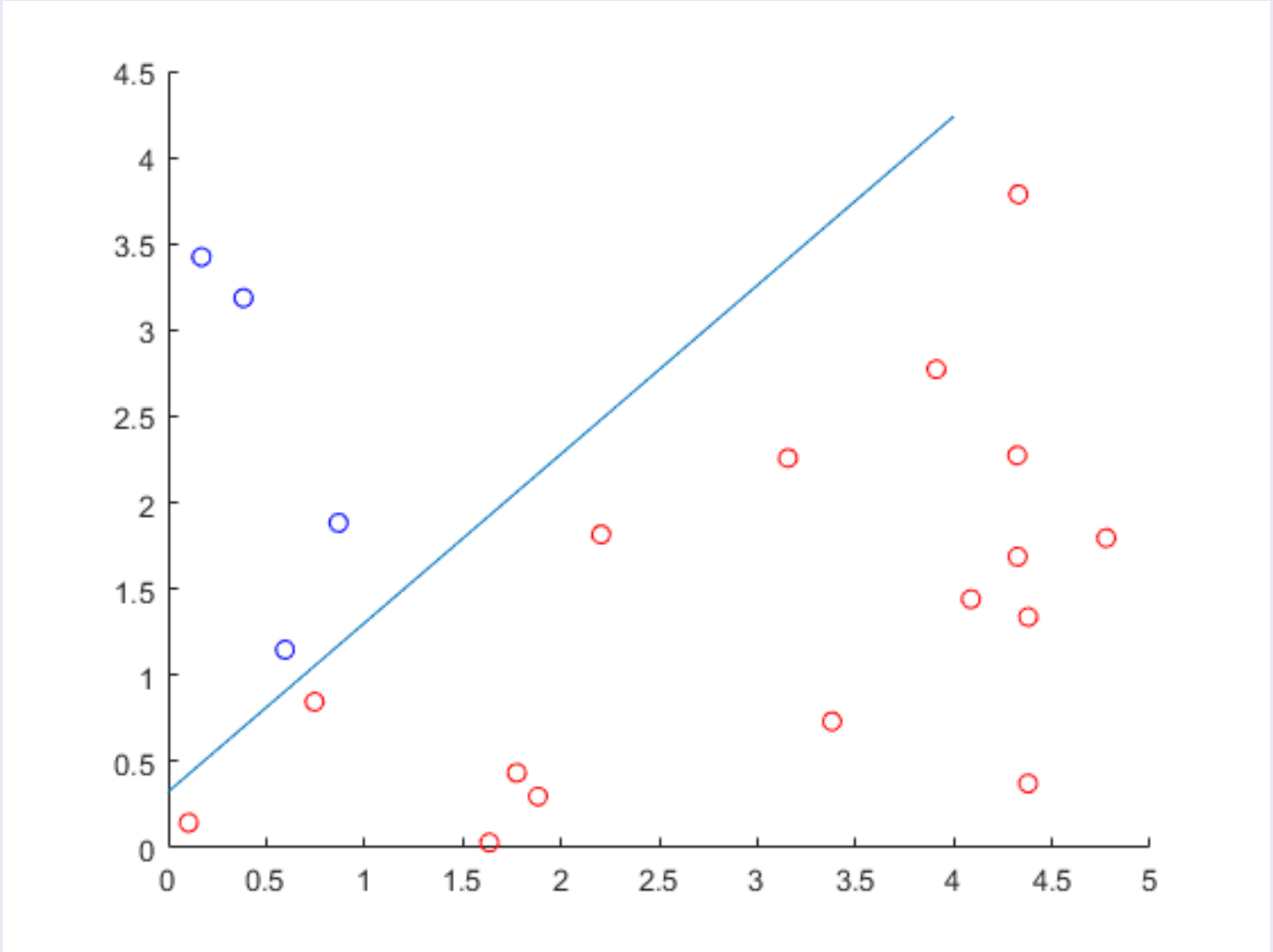
$$x_1 = (x_{1,1}, x_{1,2}, .., x_{1,p}), y_1$$
$$x_2 = (x_{2,1}, x_{2,2}, .., x_{2,p}), y_2$$
$$\ldots$$
$$x_n = (x_{n,1}, x_{n_2}, .., x_{n,p}), y_n$$

  donde cada y_i pertenece a $\{1, -1\}$.
- Queremos encontrar un clasificador que clasifique correctamente cada uno de estos ejemplos de acuerdo a su clase.

Separabilidad lineal

- Si obtenemos un
  hiperplano
  de tal forma que
  - todos los ejemplos cuya clase es -1 queden de un lado del hiperplano y
  - todos los ejemplos cuya clase es $+1$ queden del otro
  habremos **alcanzado dicho objetivo**.

Separabilidad lineal



Separabilidad lineal

- Como mencionamos antes, si para todo ejemplo x_i con $(1 \leq i \leq n)$ ocurre que

  $b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_p x_{i,p} > 0$ cuando $y_i = 1$, y
  $b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_p x_{i_p} < 0$ cuando $y_i = -1$,

  entonces podemos garantizar que

  $$y_i(b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_p x_{i,p}) > 0.$$

Separabilidad lineal

- Luego, dada una observación de test x' con p atributos podemos clasificarla de acuerdo a:
- x' será de la clase $+1$ si $b_0 + b_1 x'_1 + b_2 x'_2 + \cdots + b_p x'_p > 0$, y
- x' será de la clase -1 si $b_0 + b_1 x'_1 + b_2 x'_2 + \cdots + b_p x'_p < 0$

Separabilidad lineal

- Sea $f(x') = b_0 + b_1 x'_1 + b_2 x'_2 + .. + b_p x'_p$,
  además, podemos decir que
  si $f(x')$ es un valor cercano a 0,
  x' estará *cerca* del hiperplano
  y contrariamente,
  si $f(x')$ es un valor lejano del 0,
  x' estará *lejos* del hiperplano.

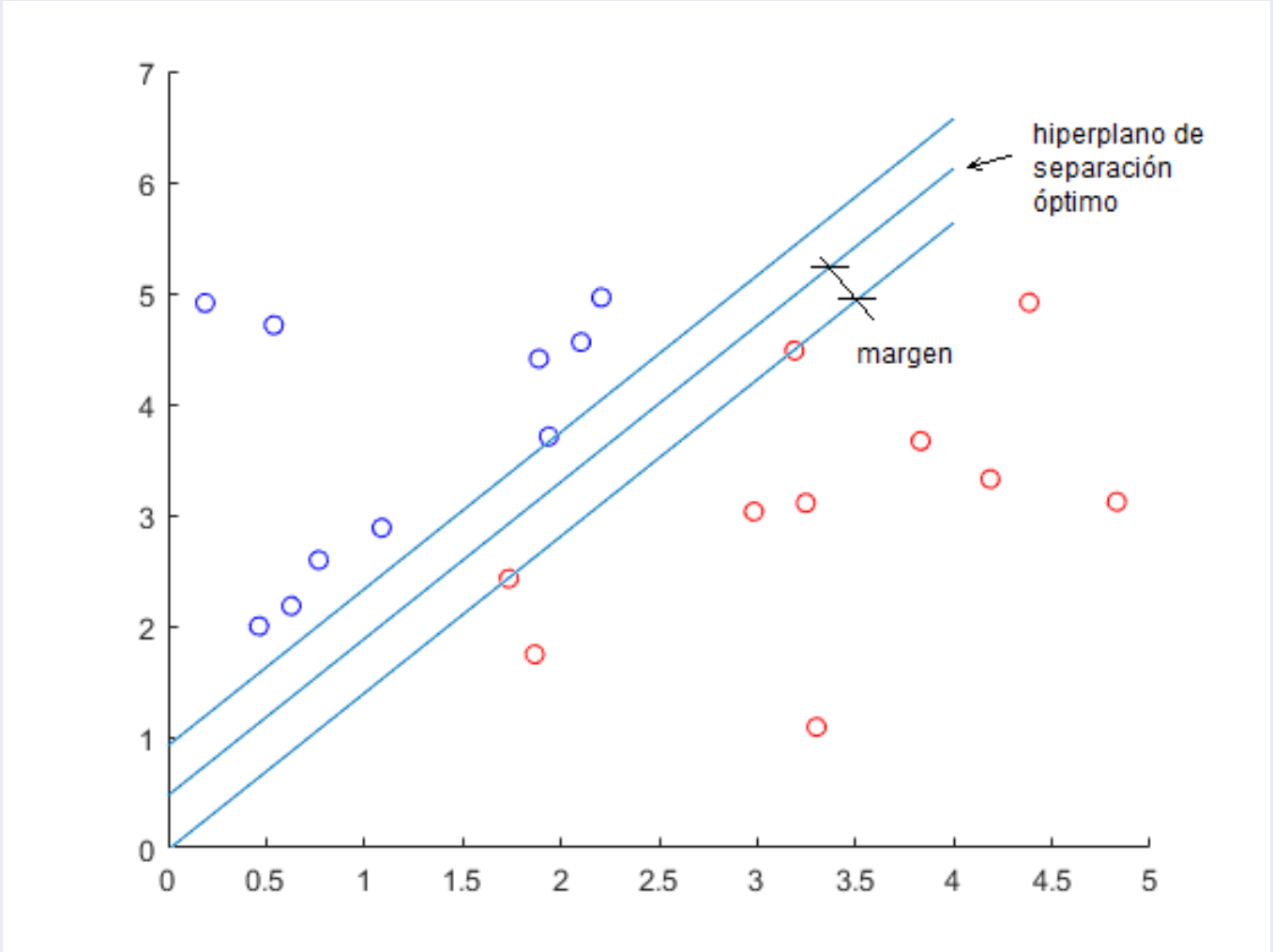
El clasificador de margen maximal

- Cuando un conjunto de ejemplos puede ser clasificado por un hiperplano entonces,
  puede haber más de un hiperplano que los separe,
  en general, infinitos hiperplanos que separan ambas clases.
- Sin embargo, hay un hiperplano que posee una propiedad en particular.

El clasificador de margen maximal

- Sea H un hiperplano que separa ambas clases. Consideremos las distancias de cada uno de los ejemplos a H.
- Definiremos **margen** como la distancia del ejemplo más cercano a H.
- Lo que nos interesa es escoger el hiperplano H tal que el margen sea máximo.
- A dicho hiperplano lo llamaremos **hiperplano de separación óptimo** o también **hiperplano de margen maximal**.

Separabilidad lineal



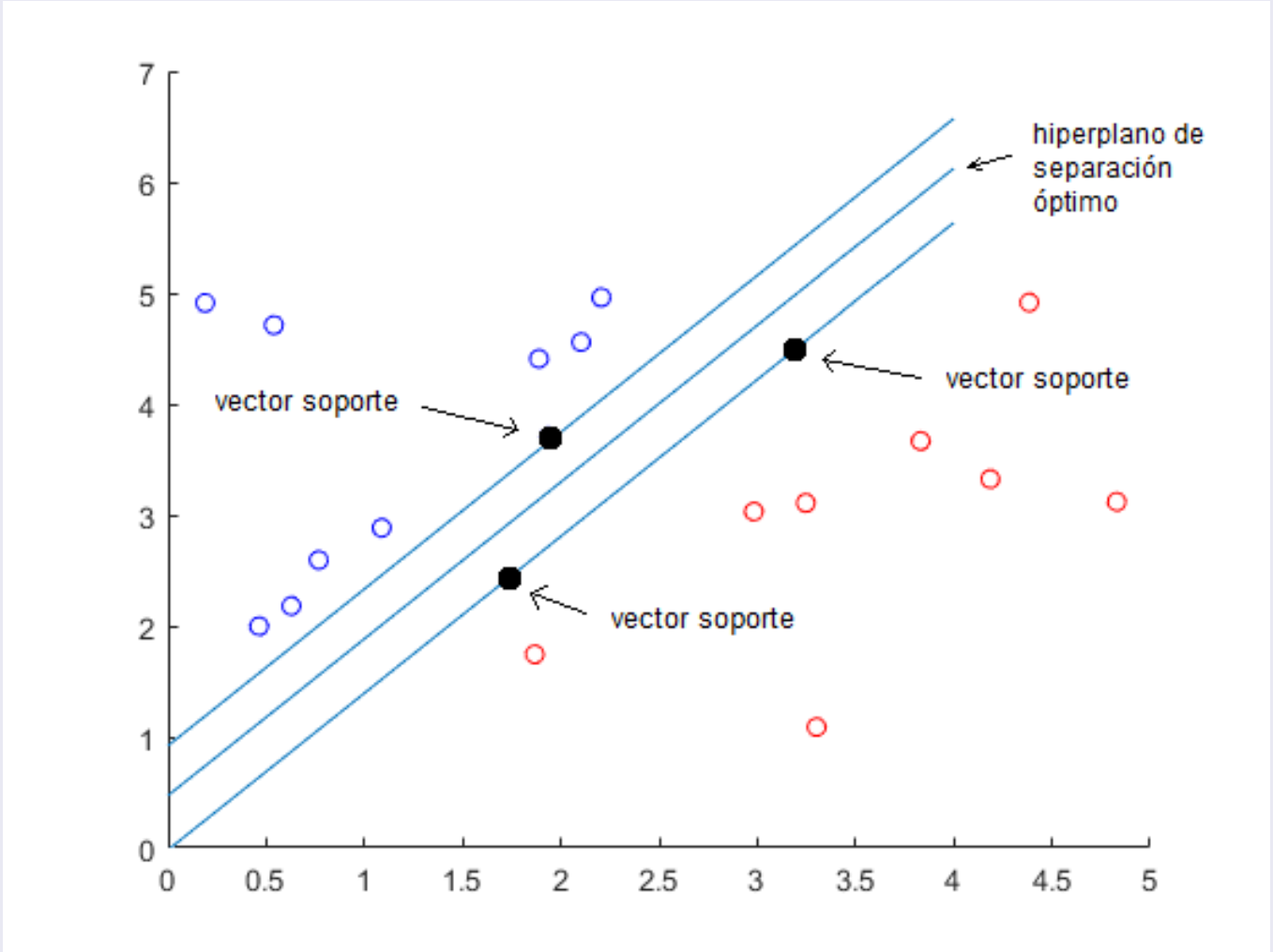
El clasificador de margen maximal

- Si nosotros usamos el **hiperplano de margen maximal** para separar ambas clases, el clasificador se llama **Clasificador de margen maximal**.
- Podemos ver que además de elegir el hiperplano de separación óptimo H también quedan definidos dos hiperplanos equidistantes de H, H+ y H-.
  La distancia entre H+ y H, y entre H- y H, es la misma: el margen.

El clasificador de margen maximal

- Sobre H+ y H- se pueden observar ejemplos de ambas clases que están sobre ellos. Dichos ejemplos se llaman **vectores de soporte** (support vectors).
- Fijarse también que:
  el hiperplano de separación óptimo **depende solamente** de los vectores de soporte y no del resto de los ejemplos de las clases.

Separabilidad lineal



El clasificador de margen maximal

Construcción del Clasificador de margen maximal

- Recordemos que $b = (b_1, b_2, ..., b_p)$ es un vector ortogonal al hiperplano dado por $b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p = 0$
- Por lo tanto, si se obtiene el producto interno entre x y b, y b está normalizado, este dará la distancia entre x y el hiperplano.

El clasificador de margen maximal

Construcción del Clasificador de margen maximal

Como lo que queremos es maximizar dicha distancia, y b define al hiperplano de separación óptimo entonces lo que queremos hacer es encontrar los valores de b tales que maximicen M sujeto a

- $y_i * (b_0 + b_1 x_{i,1} + b_2 x_{i,2} + ... + b_p x_{i,p}) \geq M$, $\forall i, 1 \leq i \leq n$,
- $\sum_{j=1}^{p} b_j^2 = 1$.

Clasificador con vectores de soporte

Clasificador con vectores de soporte

- La distancia de una observación de test al hiperplano de separación óptimo nos da una idea de la confianza que podemos tener en la clasificación.
- Si la distancia es *grande* tendremos más confianza (la observación está *bastante adentro* de la clase).
- Si la distancia es *pequeña*, cerca a 0, tendremos menos confianza (la observación está *cerca* del límite de la clase y por lo tanto cerca de la otra clase).

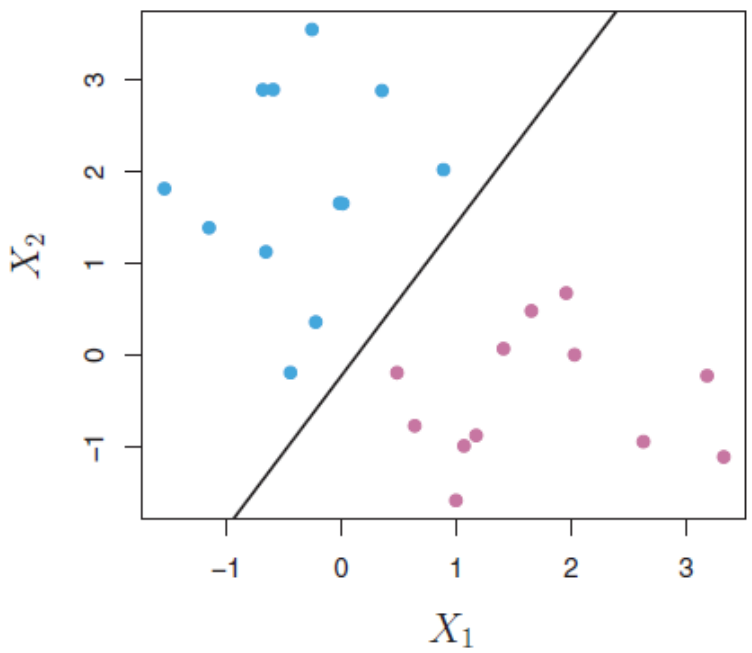
Clasificador con vectores de soporte

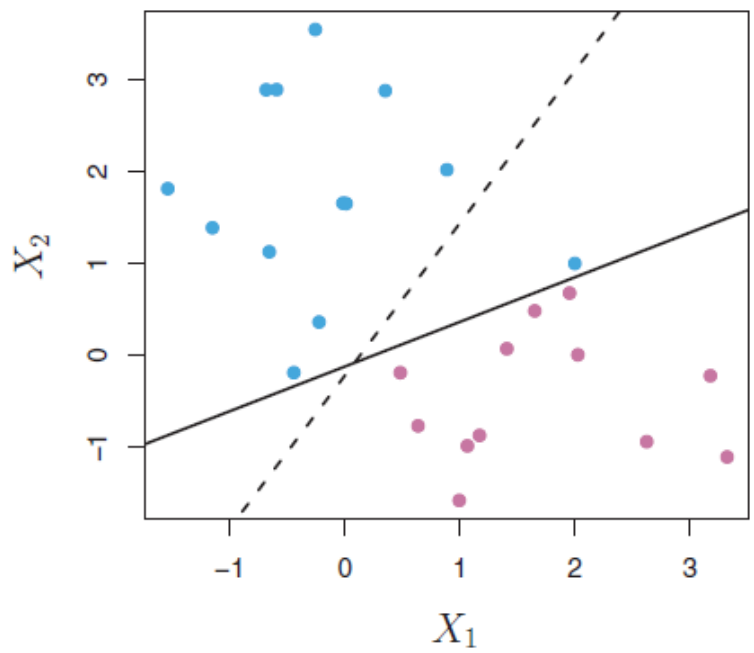
¿Es óptimo siempre el hiperplano de separación óptimo?

- Supongamos dos clases que están bien separadas, es decir, tienen un margen considerable, y agregamos un ejemplo que hace que el margen se reduzca considerablemente.
- Ejemplos que con el hiperplano inicial hubieran sido clasificados con mayor confianza ahora estarán clasificados con menos confianza.

Clasificador con vectores de soporte

¿Es óptimo siempre el hiperplano de separación óptimo?





Clasificador con vectores de soporte

¿Es óptimo siempre el hiperplano de separación óptimo?

- ¿No valdría la pena usar el hiperplano inicial (el que no tenía en cuenta el ejemplo que hace el margen muy reducido) en vez de usar el nuevo hiperplano que divide pero acarrea un test menos confiable? Esto se puede resumir en estos dos objetivos:
- Que las observaciones individuales sean clasificadas con robustez (que la distancia al hiperplano no sea crítica)
- Una mejor clasificación de la mayoría de los ejemplos de entrenamiento (asumiendo clases no linealmente separables).

Clasificador con vectores de soporte

Clasificador con margen tolerante.

Estos dos objetivos están resumidos en el concepto de **Clasificador con margen tolerante** y se plantean siguiendo: encontrar los valores de b, y $\epsilon_1, .., \epsilon_n$ tales que maximicen M sujeto a

- $y_i * (b_0 + b_1 x_{i,1} + b_2 x_{i,2} + ... + b_p x_{i,p}) \geq M * (1 - \epsilon_i)$, $\forall i, 1 \leq i \leq n$,
- $\sum_{j=1}^{p} b_j^2 = 1$.
- $\forall_i, \epsilon_i \geq 0 \wedge \sum_{i=1}^{n} \epsilon_i \leq C$.
  donde C es un parámetro de ajuste del método.

Clasificador con vectores de soporte

Clasificador con margen tolerante.

- Cada ϵ_i permite clasificar el ejemplo x_i en un *lugar erróneo* si fuera necesario.

Este lugar erróneo bien podría ocurrir por:

- estar dentro del margen de la clase, o incluso
- estar del lado incorrecto del hiperplano.

Clasificador con vectores de soporte

Clasificador con margen tolerante.

- Dada una observación x_i
  $y_i * (b_0 + b_1 x_{i,1} + b_2 x_{i,2} + ... + b_p x_{i,p}) \geq M * (1 - \epsilon_i)$,
  $\forall i, 1 \leq i \leq n$, se satisface para
  - $\epsilon_i = 0 \Rightarrow$ la observación está del lado correcto del margen,
  - $\epsilon_i > 0$ pero $\epsilon_i < 1 \Rightarrow$ la observación está del lado incorrecto del margen,
  - $\epsilon_i > 1 \Rightarrow$ la observación está del lado incorrecto del hiperplano.

Clasificador con vectores de soporte

Clasificador con margen tolerante.

- C es un parámetro que dice cuánto se va a permitir que las observaciones (en su conjunto) violen el margen o el hiperplano.
- Si $C = 0$, el Clasificador con margen tolerante se convierte en un Clasificador de margen maximal.
- Si $C \neq 0$, y por ejemplo $C = 5$, quiere decir que sólo permitiríamos 5 ejemplos mal clasificados o 4 mal clasificados y dos (o más) ejemplos que estén en el lado incorrecto del margen, o etc.
- Una forma de encontrar C es con validación cruzada.

Máquina basada en vectores de soporte

Clasificación con límites de decisión no lineales

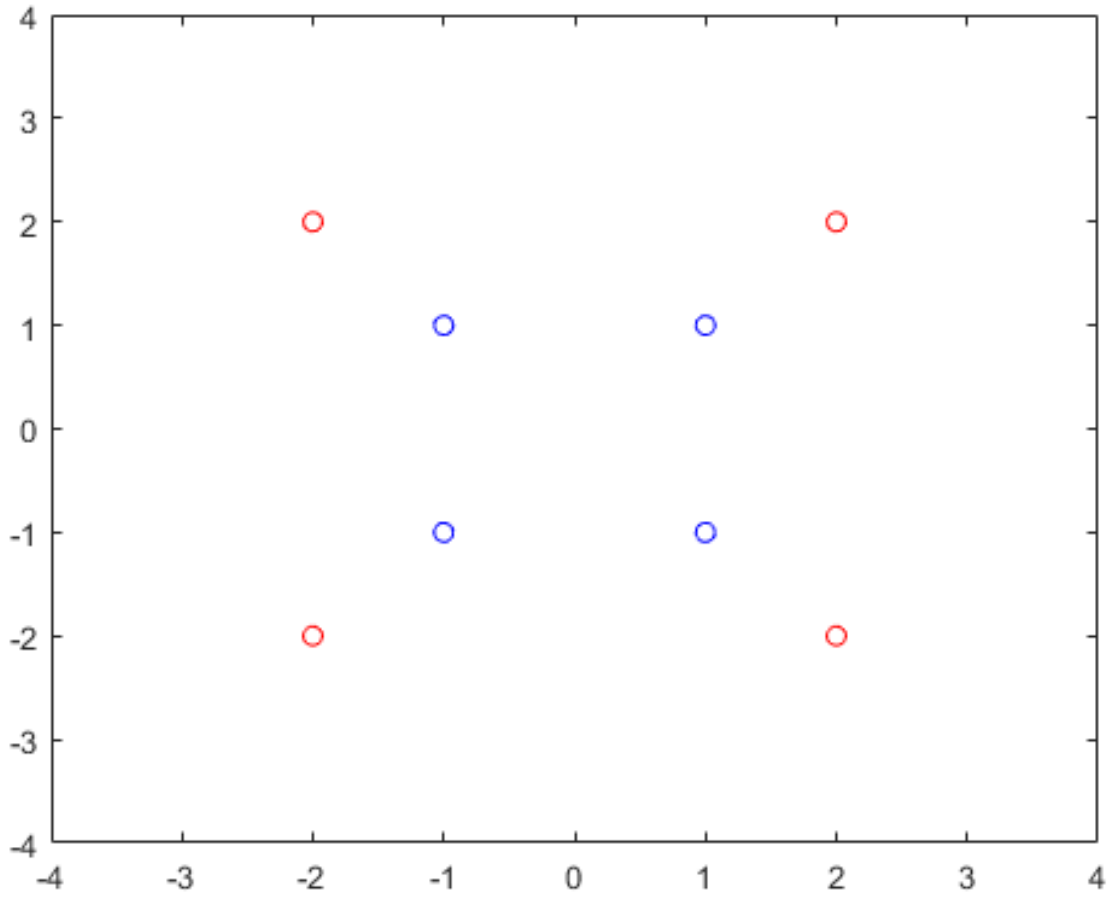
Consideremos un conjunto de entrenamiento donde sus ejemplos x_i son de dimensión p $= 2$ y su clase es $y_i \in \{-1, 1\}$ (-1 en rojo y 1 en azul):

$X = \{(-2,-2),(-2,2),(2,-2),(2,2),(-1,-1),(-1,1),(1,-1),(1,1)\}$

$Y = \{-1,-1,-1,-1,1,1,1,1\}$

donde no se puede establecer un hiperplano de separación entre una clase y la otra.

Máquina basada en vectores de soporte



Máquina basada en vectores de soporte

Clasificación con límites de decisión no lineales

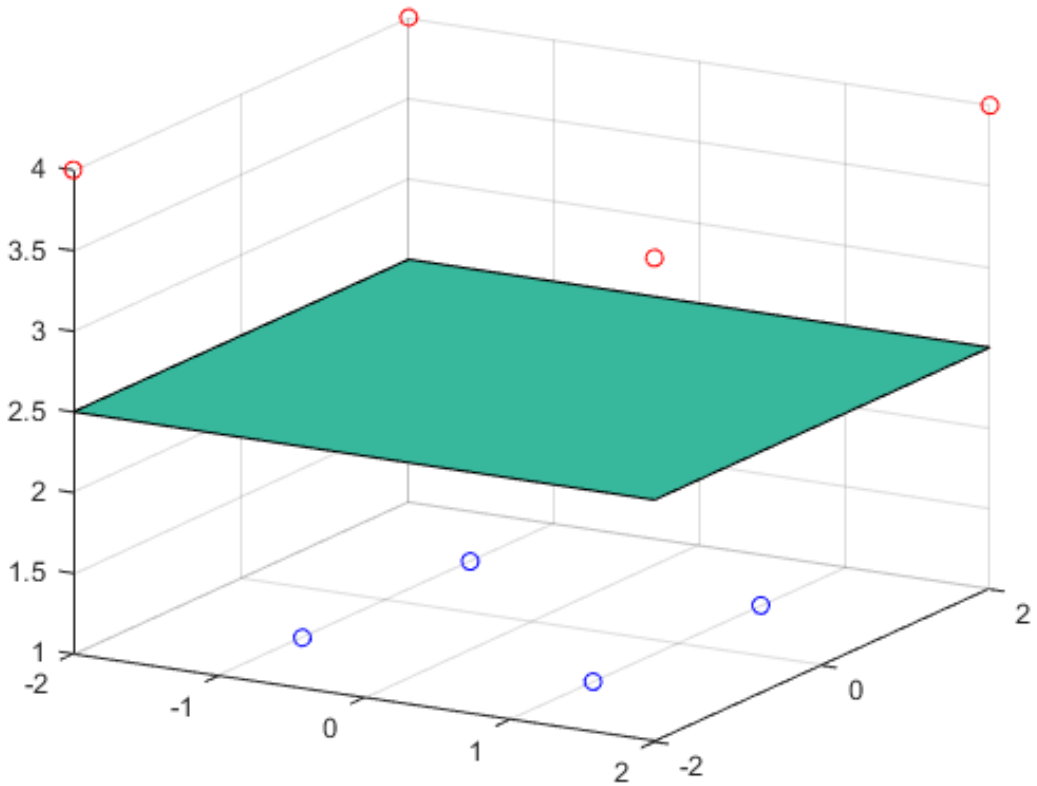
Pero, fijarse que si los ejemplos $x_i = (x_{i1}, x_{i2})$ de X en vez de representarlos en R^2 lo hacemos en R^3 como $x_i = (x_{i1}, x_{i2}, x_{i1}^2)$ con el mismo Y:

$X' = \{(-2,-2,4),(-2,2,4),(2,-2,4),(2,2,4),$
$(-1,-1,1),(-1,1,1),(1,-1,1),(1,1,1)\}$

$Y' = \{-1,-1,-1,-1,1,1,1,1\}$

obtenemos: (ver próxima transparencia)

Máquina basada en vectores de soporte



donde sí existe un hiperplano de separación entre ambas clases.

Máquina basada en vectores de soporte

Clasificación con límites de decisión no lineales

Por ejemplo, si tenemos ejemplos en una dimensión p

$$x_{i1}, x_{i2}, .., x_{ip}$$

podríamos representarlos en una dimensión $2p$ de acuerdo a

$$x_{i1}, x_{i1}^2, x_{i2}, x_{i2}^2, .., x_{ip}, x_{ip}^2$$

donde podría haber un hiperplano de dimensión $2p - 1$ que los separe.

Máquina basada en vectores de soporte

Clasificación con límites de decisión no lineales

En este caso, el problema a resolver es encontrar los valores de $b = b_0, b_{11}, b_{12}, .., b_{p1}, b_{p2}$, y $\epsilon_1, .., \epsilon_n$ tales que maximicen M sujeto a

- $y_i * (b_0 + \sum_{j=1}^{p} b_{j1} * x_{ij} + \sum_{j=1}^{p} b_{j2} * x_{ij}^2 \geq (M - \epsilon_i)$, $\forall i, 1 \leq i \leq n$
- $\sum_{j=1}^{p} \sum_{k=1}^{2} b_{jk}^2 = 1$.
- $\epsilon_i \geq 0, \forall_i, 1 \leq i \leq n \wedge \sum_{i=1}^{n} \epsilon_i \leq C$.
  donde C es un parámetro de ajuste del método.

Máquina basada en vectores de soporte

Clasificación con límites de decisión no lineales

La idea es proponer un espacio donde exista separabilidad lineal aunque no haya separabilidad lineal en el espacio original de los ejemplo.

Máquina basada en vectores de soporte

SVM - Support vector machine

Resumen:
Cuando tenemos un conjunto de ejemplos que pertenecen a dos clases diferentes podemos:

- si ellos son linealmente separables, obtener un hiperplano de separación en el espacio que los ejemplos están representados,
- si ellos no son linealmente separables, proponer una nueva representación de los ejemplos de tal modo que sí sean linealmente separables es ese mnuevo espacio de representación.

Máquina basada en vectores de soporte

SVM - Support vector machine

Entonces,
ya sea que los ejemplos están representados en un espacio tal que ellos son linealmente separables o,
sea que logramos representarlos en un espacio donde sean linealmente separables,
en todos los casos, debemos resolver un problema de optimización no lineal con restricciones.
No hay una única forma de hacerlo y en general, ellas escapan al alcance de este curso.

Máquina basada en vectores de soporte

SVM - Support vector machine

Los métodos de resolución del problema de optimización no lineal con restricciones dan como salida la siguiente información:

- Una lista de vectores soporte x_i
- Una lista de constantes α_i
- y un término independiente b_0

Máquina basada en vectores de soporte

SVM - Support vector machine

La forma de utilizar dicha información para establecer a qué clase pertenece un nuevo ejemplo x' es calcular $f(x')$ según:

$$f(x') = b_0 + \sum_{i=1}^{n} \alpha_i \langle x', x_i \rangle$$

Si $f(x') > 0$ entonces x' pertenece a la clase 1
Si $f(x') \leq 0$ entonces x' pertenece a la clase -1

Fijarse que $\langle x', x_i \rangle$ expresa el producto interno entre x' y x_i

Máquina basada en vectores de soporte

SVM - Support vector machine

- La propuesta es dar una generalización de la clasificación con límites de decisión no lineales.
- La forma de generalizar esta idea es mediante la introducción del concepto de **Núcleo** (**Kernel**).
- Se propone un tipo de núcleo, se llama a una función que lleva a cabo el proceso de optimización no lineal con restricciones.

Máquina basada en vectores de soporte

SVM - Support vector machine

Luego, dada una observación x', si queremos saber a qué clase pertenece, calculamos $f(x')$ como:

$$f(x') = b_0 + \sum_{i=1}^{n} \alpha_i K(x', x_i)$$

donde K se lo denomina Núcleo (Kernel) propuesto y donde los α_i, los x_i y b_0 fueron obtenidos mediante el proceso de optimización no lineal con restricciones.

Máquina basada en vectores de soporte

SVM - Support vector machine

Existen varios tipos de núcleos:

- Núcleo lineal
$$Kx', x_i) = \langle x', x_i \rangle$$
- Núcleo polinómico
$$K(x', x_i) = (1 + \sum_{j=1}^{p} x_{ij} x'_j)^d$$
donde d es el grado del polinomio.
A medida que d se incrementa habrá mayor flexibilidad para encontrar una separación lineal en el nuevo espacio de los ejemplos (el espacio ampliado).

Máquina basada en vectores de soporte

SVM - Support vector machine

- Núcleo radial

$$K(x', x_i) = e^{-\gamma \sum_{j=1}^{p} (x_{ij} - x'_j)^2}$$

donde γ es una constante positiva.
Los ejemplos *lejos* de x' tendrán muy poco *peso* en el valor de $f(x')$.

- Entre otros.

Máquina basada en vectores de soporte

SVM - Support vector machine

- Fijarse que aunque cambiemos el núcleo, la formulación para el problema no cambia. El cálculo de $f(x')$ seguirá siendo:

$$f(x') = b_0 + \sum_{i=1}^{n} \alpha_i K(x', x_i)$$

- A diferencia del núcleo polinómico, el núcleo radial no expande el espacio con una cantidad de términos de mayor grado (aumentando la dimensión del espacio ampliado).

Máquina basada en vectores de soporte

SVM - Support vector machine multiclase

¿Qué ocurre cuando nuestro conjunto de entrenamiento tiene más de 2 clases?

Hay dos abordajes:

- Uno contra uno
- Uno contra el resto

Máquina basada en vectores de soporte

SVM - Support vector machine multiclase

- En el abordaje *uno contra uno* se construye un SVM para cada par distinto de clases (una clase se le asignará $+1$ y a la otra -1). Los ejemplos de las clases restantes se ignorará.
- En este caso, de cada SMV construido se obtendrá una respuesta cuando se presente una nueva observación y se sopesará dichas respuestas para obtener una única salida (por ejemplo, método de votación)

Máquina basada en vectores de soporte

SVM - Support vector machine multiclase

- En el abordaje *uno contra todos* se construye un SVM para cada clase (a los ejemplos de clase se le asignará $+1$ y al resto de los ejemplos del conjunto de entrenamiento se le asignará -1).
- En este caso, de cada SMV construido se obtendrá una respuesta cuando se presente una nueva observación y la clase resultante será la que corresponda al SVM cuyo $f(x')$ sea mayor.