MODELOS COMPUTACIONALES: CUARTO CURSO DEL GRADO DE ING. INFORMÁTICA EN COMPUTACION

Introducción a las Máquinas de Vectores Soporte, SVM

César Hervás-Martínez
Pedro A. Gutiérrez Peña

Grupo de Investigación AYRNA

Departamento de Informática y Análisis Numérico
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es
pagutierrez@uco.es

2022-2023

INDICE

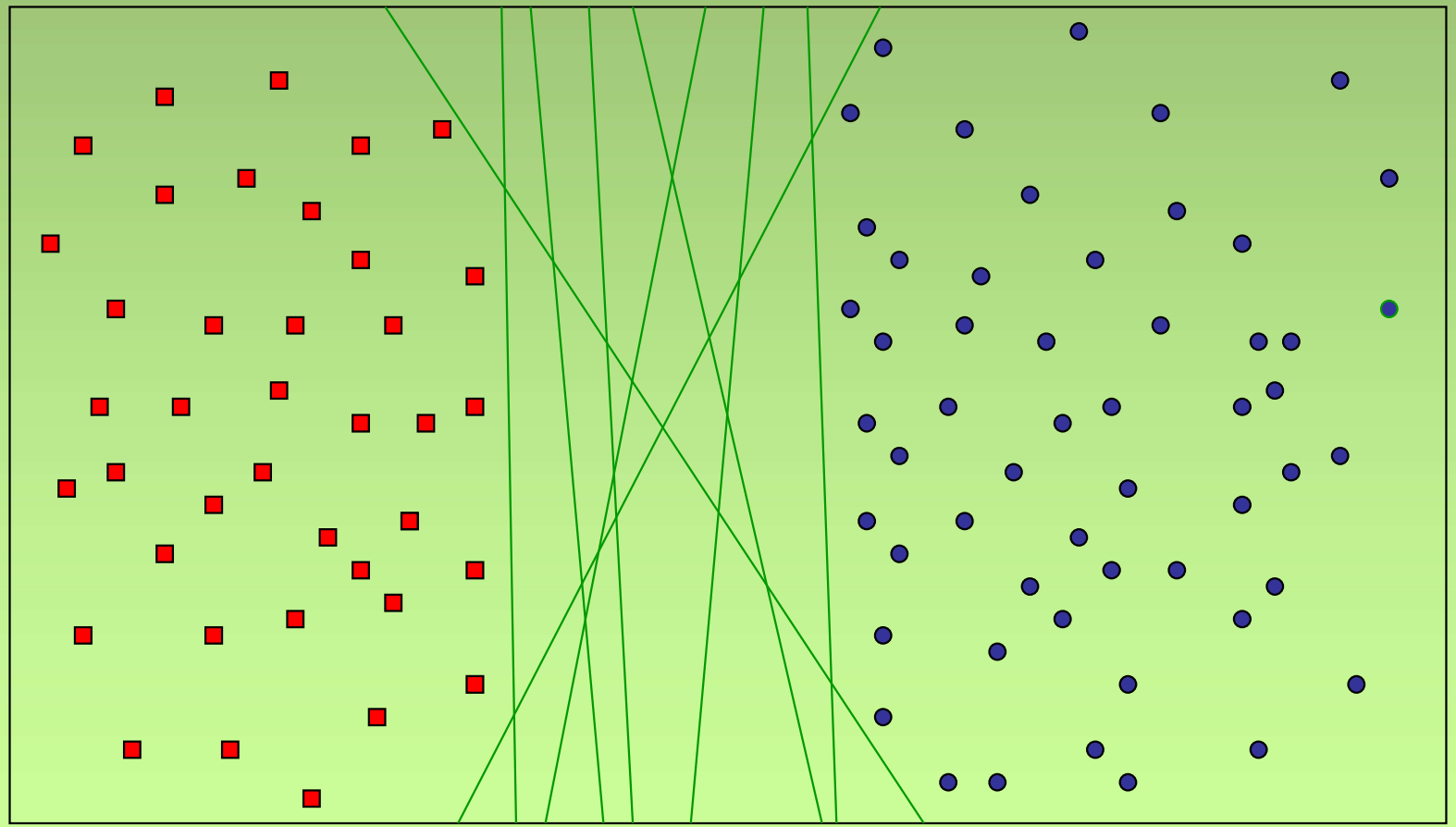
SVMs Lineales

SVMs No Lineales

Aplicación de las SVMs

Conclusiones y Futuro

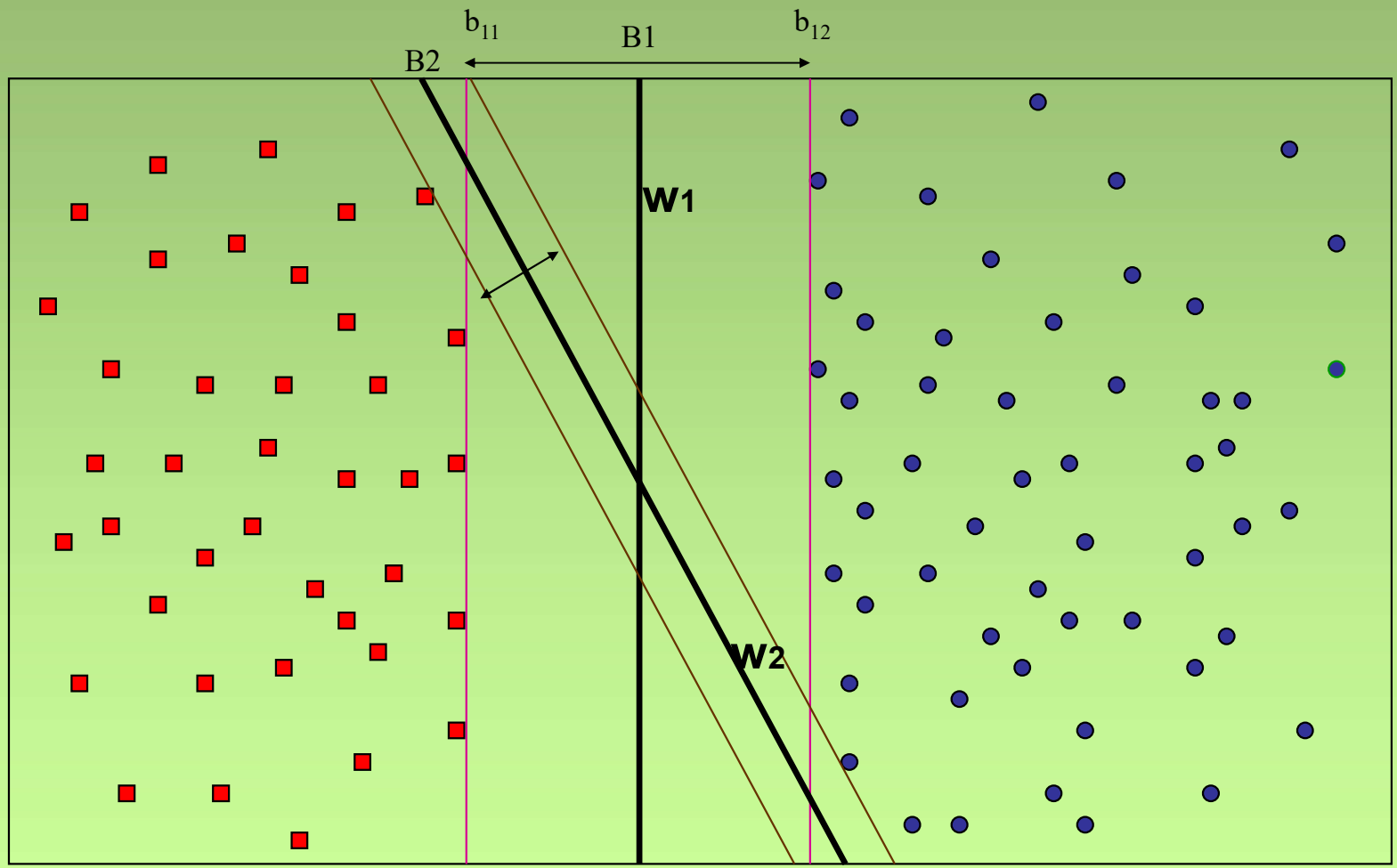
Posibles fronteras de decisión para datos linealmente separables



Objetivo

- Dado un conjunto de ejemplos de entrenamiento construir un hiperplano “w” como superficie de decisión. De tal forma que la separación de las dos clases sea máxima (principio de generalización)

Diferentes márgenes de decisión



Clasificadores lineales o cuasi-lineales

• Para aproximar la frontera de decisión óptima para un clasificador binario, desde un punto de vista probabilístico, se puede utilizar la razón de verosimilitudes. $$\mathrm{H}(\mathbf{x}) = \frac{P(\mathbf{x}/C_1)}{P(\mathbf{x}/C_2)}$$

También la aproximación se puede hacer utilizando:

– Una función lineal:

$$g(\mathbf{x}) = w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

– Una combinación de función lineal y logística:

$$h(\mathbf{x}) = \frac{1}{1+\exp(-g(\mathbf{x}))}$$

• En ambos casos la frontera de decisión generada es lineal o *cuasi* lineal.

REGLA DE DECISIÓN BAYES

A veces, desde un punto de vista analítico, es mas conveniente trabajar con la función, – logaritmo de la razón de verosimilitudes, y de esta forma la regla de decisión, para dos clases, es que

$$H(\mathbf{x}) = -L(\mathbf{x}) = -\ln P(\mathbf{x}/C_1) + \ln P(\mathbf{x}/C_2)$$

supere el umbral dado por $\ln \dfrac{P(C_2)}{P(C_1)}$

Al término H(x) se le llama la función discriminante. Si las probabilidades a priori de pertenencia a las dos clases son iguales, esto es

$$P(C_1) = P(C_2);\ \text{entonces}\ \ln \frac{P(C_2)}{P(C_1)} = 0$$

Las reglas de decisión son ahora

$$\text{Si}\ -\ln P(\mathbf{x}/C_1) + \ln P(\mathbf{x}/C_2) > 0\ \ \text{Entonces}\ \mathbf{x} \in C_1$$

$$\text{Si}\ -\ln P(\mathbf{x}/C_1) + \ln P(\mathbf{x}/C_2) < 0\ \ \text{Entonces}\ \mathbf{x} \in C_2$$

A estas reglas de decisión se les llama el contraste Bayes de mínimo error

Clasificadores lineales o cuasi-lineales

- En la práctica, cuando la frontera de decisión óptima no es lineal los resultados obtenidos con clasificadores lineales no son satisfactorios.

- En este tema se verán métodos para dividir el espacio de características en regiones de decisión cuya frontera no es lineal.

- Se presentarán dos tipos de clasificadores:

  – Clasificador polinomial.
  – Máquina de vectores soporte.

Clasificador Polinomial I

¿Cómo transformar el clasificador lineal para obtener fronteras de decisión no lineales?

– La forma del clasificador lineal es:

$$g(\mathbf{x}) = w_0 + \mathbf{w}^{\mathbf{T}}\mathbf{x} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

–Si introducimos los términos de grado 2:

$$g(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$
$$+ w_{11} x_1^2 + w_{12} x_1 x_2 + \ldots + w_{1d} x_1 x_d + w_{2d} x_2 x_d + \ldots + w_{dd} x_2^2 + \ldots\ldots\ldots\ldots\ldots$$

Este tipo de función es una función discriminante cuadrática.

– Podemos seguir con el grado 3:

$$g(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$
$$+ w_{11} x_1^2 + w_{12} x_1 x_2 + \ldots + w_{1d} x_1 x_d + w_{2d} x_2 x_d + \ldots + w_{dd} x_2^2 +$$
$$+ w_{111} x_1^3 + w_{112} x_1^2 x_2 + \ldots.$$

y seguir hasta un grado arbitrario, lo que nos lleva a tener un altísimo numero de coeficientes

Clasificador Polinomial II

• **Las funciones discriminantes polinomiales anteriores tienen una forma común:**

$$g(\mathbf{x}) = w_0 + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \ldots + w_d\phi_d(\mathbf{x})$$

• **Ejemplo:**
Si la función *g* es:

$$g(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + w_{11}x_1^2 + w_{12}x_1x_2 + w_{21}x_2x_1 + w_{22}x_2^2$$

Entonces definiendo:

$$\phi_1(\mathbf{x}) = x_1^2,\ \ \phi_2(\mathbf{x}) = x_1x_2,\ \phi_3(\mathbf{x}) = x_2x_1,\ \phi_4(\mathbf{x}) = x_2^2,\ \phi_5(\mathbf{x}) = x_1\ ,\ \phi_4(\mathbf{x}) = x_2$$

Podemos escribir:

$$g(\mathbf{x}) = w_0 + w_{11}\phi_1(\mathbf{x}) + w_{12}\phi_2(\mathbf{x}) + w_{21}\phi_3(\mathbf{x}) + w_{22}\phi_4(\mathbf{x}) + w_1\phi_5(\mathbf{x}) + w_2\phi_6(\mathbf{x})$$

Función Discriminante Lineal Generalizada (FDLG)

Llamaremos Función Discriminante Lineal Generalizada, FDLG, a toda función discriminante con la forma:

$$g(\mathbf{x}) = w_0\phi_0(\mathbf{x}) + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \ldots + w_M\phi_M(\mathbf{x})$$

$$g(\mathbf{x}) = \sum_{i=0}^{M} w_i\phi_i(\mathbf{x}), \text{ donde } \phi_0(\mathbf{x}) = 1$$

- **Las funciones ϕ pueden ser polinomiales o de otro tipo:**
  - **Gausianas, Splines, Sigmoides, producto de potencias, etc...**
- **La idea de descomponer una función compleja como suma de otras más simples es una idea recurrente en Matemáticas:**
  - **Series de Taylor (1715).**
  - **Series de Fourier (1822).**
  - **Series de *Wavelets "ondas"*. (1986)**
- **De hecho, también se usa con las ventanas de Parzen:**

$$\hat{p}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \partial(\mathbf{x} - \mathbf{x_i})$$

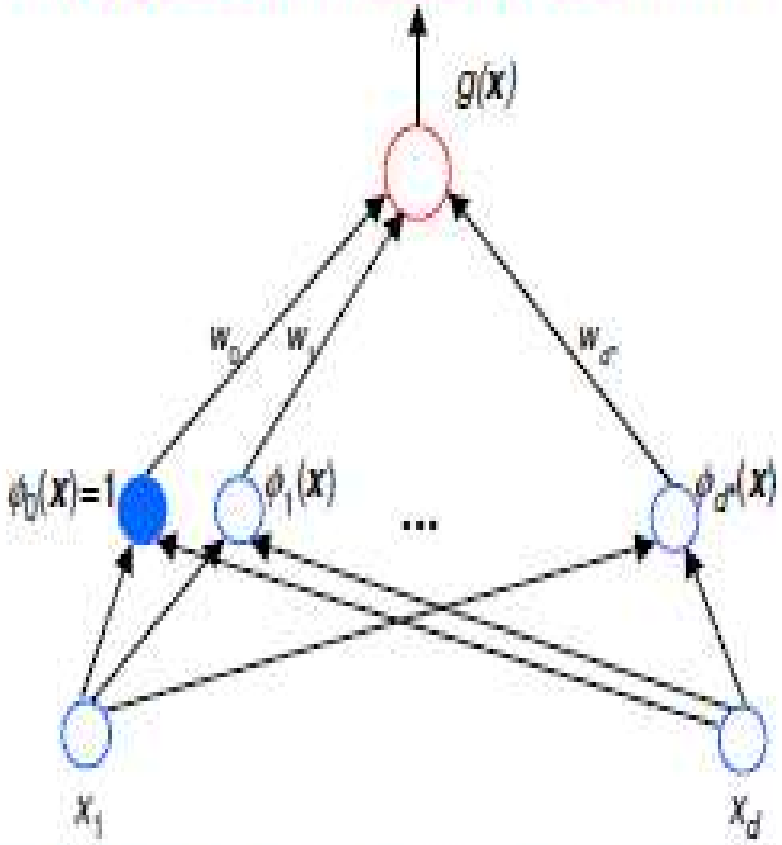
FDLG: Notación y Representación Gráfica

Entonces en notación matricial se puede definir en la forma

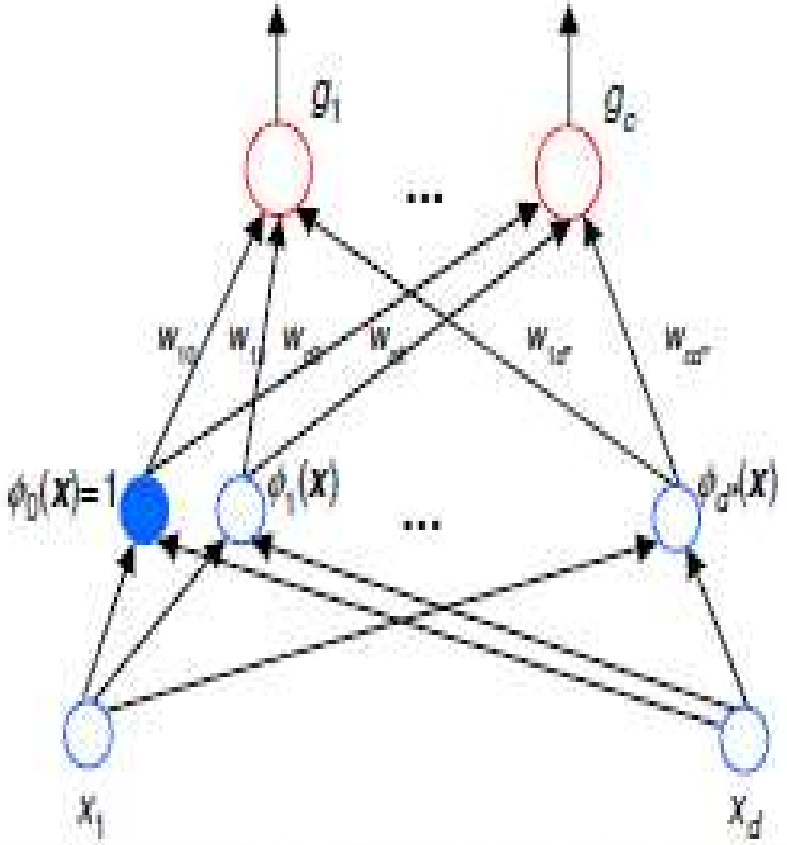
$$g(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}), \text{ donde } \mathbf{w}^T = (w_0, w_1, w_2, \ldots, w_M),$$

$$\text{y } \phi(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))^T$$

- **Representación gráfica:**



FDL generalizada: Representación gráfica (dos clases)



FDL generalizada: Representación gráfica (c clases)

FDLG: Entrenamiento

- **Una observación crucial:**
  - **Las funciones ϕ transforman el espacio original de variables independientes en un nuevo espacio de características.**

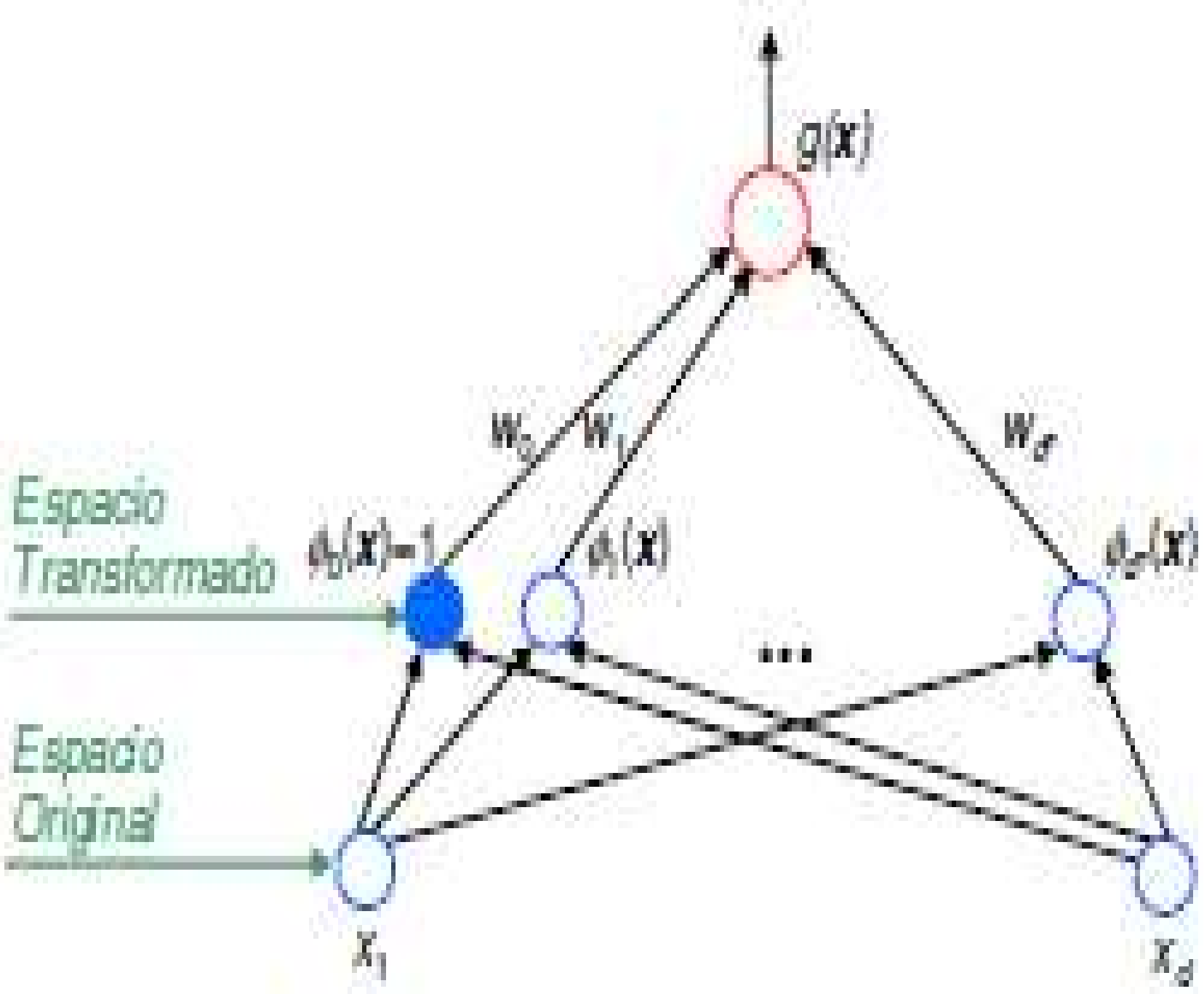
  - **En este nuevo espacio el problema es determinar una función discriminante lineal.**

- **Por tanto, el esquema de aprendizaje es:**

Paso 1: Transformar los datos de entrada con las funciones ϕ , a un espacio de características

Paso 2: Aplicar a los datos transformados alguno de los métodos del Análisis Discriminante Lineal

FDLG: Entrenamiento



Ejemplo de un clasificador cuadrático para el problema XOR

La función discriminante es:

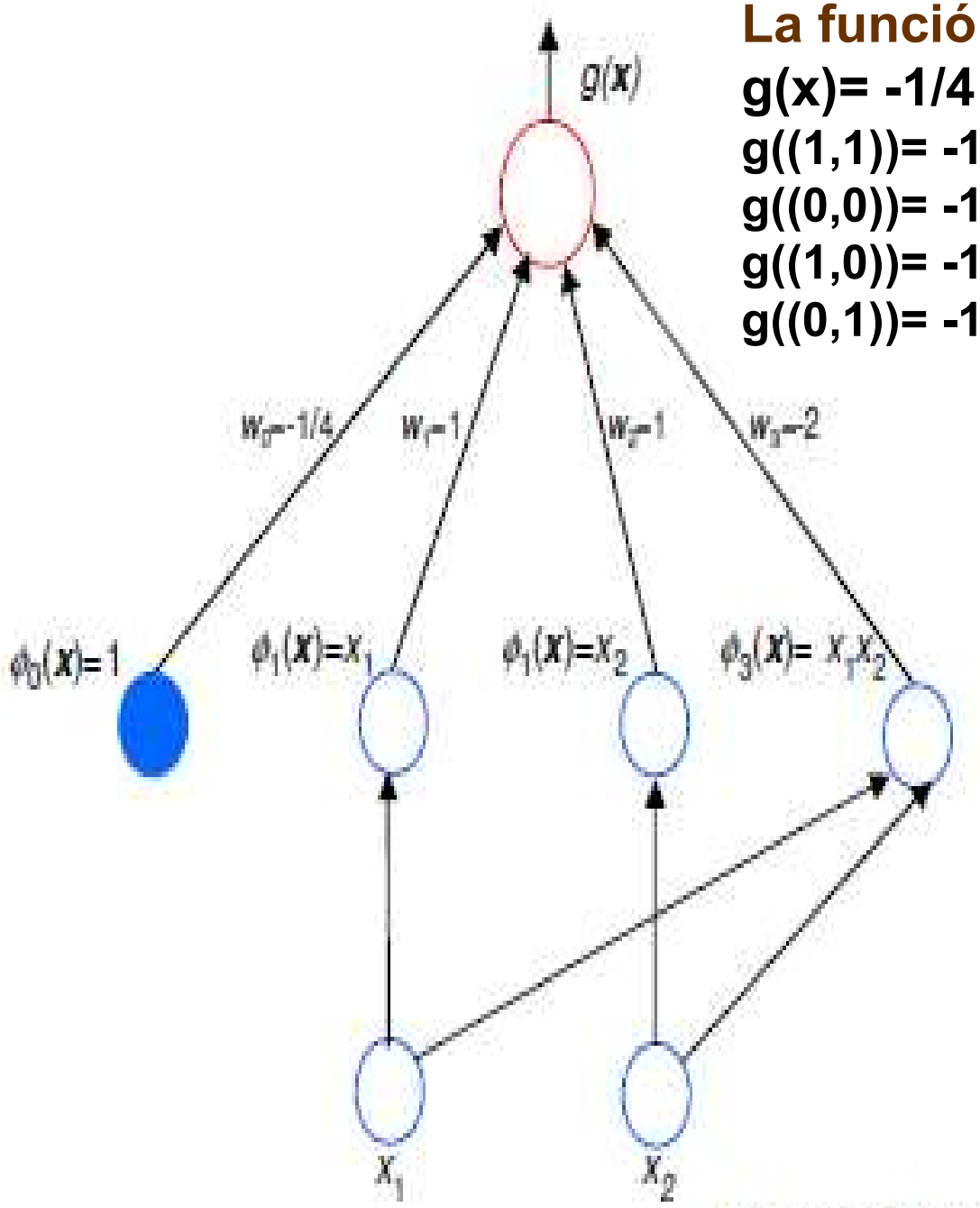
$g(x) = -1/4 - 2x_1x_2 + x_1 + x_2$

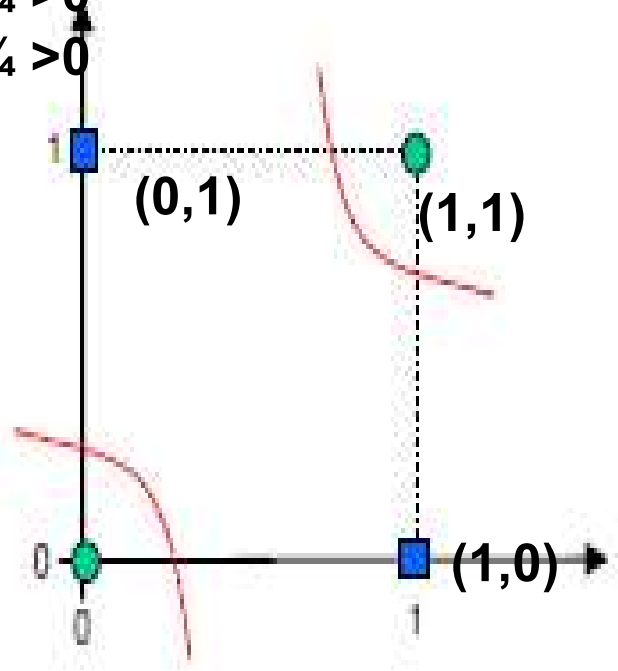
$g((1,1)) = -1/4 - 2 + 1 + 1 = -1/4 < 0$

$g((0,0)) = -1/4 - 0 + 0 + 0 = -1/4 < 0$

$g((1,0)) = -1/4 - 0 + 1 + 0 = 3/4 > 0$

$g((0,1)) = -1/4 - 0 + 0 + 1 = 3/4 > 0$





El Problema del XOR: Solución cuadrática

Ejemplo de un clasificador cuadrático para el problema XOR

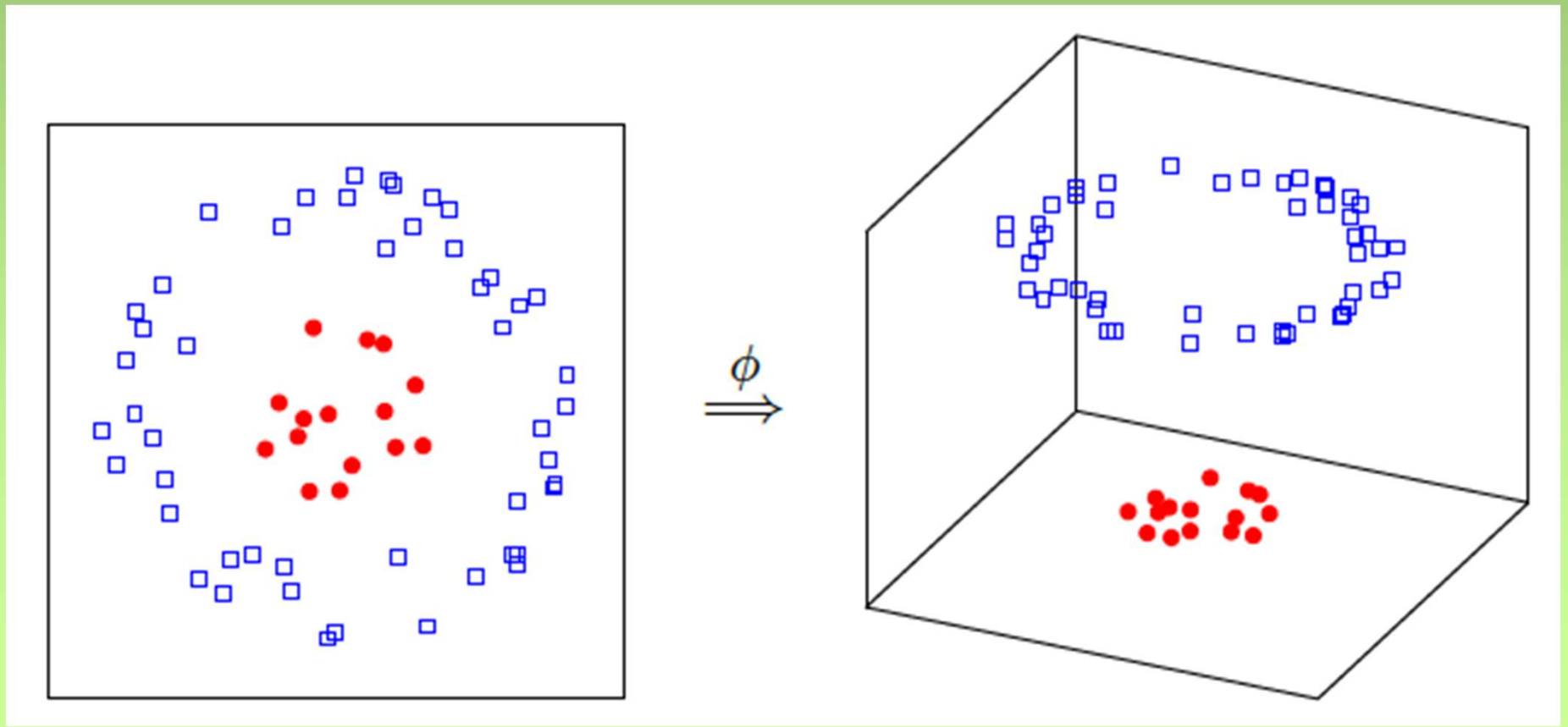
- **Observación:**

– El problema del XOR se resuelve porque hemos pasado del espacio de variables de entrada de dimensión 2, definido por $x = (x_1, x_2)$ a un espacio transformado de dimensión 4 definido por $f(x) = (1, x_1, x_2, x_1x_2)$ donde el problema sí es linealmente separable.

- Puede probarse que el aumento de la dimensión del espacio de entrada hace más fácil lograr la separabilidad lineal de los datos. Es el teorema de Cover

Separabilidad Lineal

Un ejemplo de lo que puede pasar al hacer una transformación a una dimensión superior



Separabilidad Lineal: Teorema de Cover

– Teorema de Cover

La probabilidad de que dos clases sean linealmente separables se aproxima a 1 cuando la dimensión del espacio de características *d* tiende a infinito y el número de muestras N crece delimitado por 2(*d*+1). Si d=1.000.000 N=2.000.000.

- **La aproximación tiene por tanto múltiples ventajas:**

– Aumentando el número de características es más probable que las clases sean linealmente separables (esto puede hacerse para clasificadores polinomiales incrementando el grado del polinomio)

– Se utilizan los algoritmos de entrenamiento lineales para determinar los pesos de los modelos.

SVMs Lineales

Las máquinas de vectores soporte, SVM, es uno de los métodos de aprendizaje supervisado más populares.

SVM es para clasificación binaria.
Es decir, 2 clases, como maligno vs benigno.

SVM se generaliza a múltiples clases.
Al igual que el análisis discriminante lineal y muchas otras técnicas.

SVMs Lineales

"Las SVM son un raro ejemplo de una metodología en la que se encuentran la intuición geométrica, las matemáticas elegantes, las garantías teóricas y los algoritmos prácticos"

"En las SVM se tiene algo que decir sobre cada aspecto de la ...geometría, las matemáticas, el estudio teórico y los algoritmos"

SVMs Lineales

SVM basado en cuatro grandes ideas

1.- Hiperplano separador

2.- Maximizar el "margen“

Maximizar la separación mínima entre clases

3.-Trabajar en un espacio de mayor dimensión

Más "espacio", por lo que es más fácil de separar

4.- El truco de kernel

Esto está íntimamente relacionado con el punto 3

5.-Tanto el punto 1 como 2 son bastante intuitivos

SVMs Lineales

Las SVM pueden aplicarse a cualquier conjunto de entrenamiento

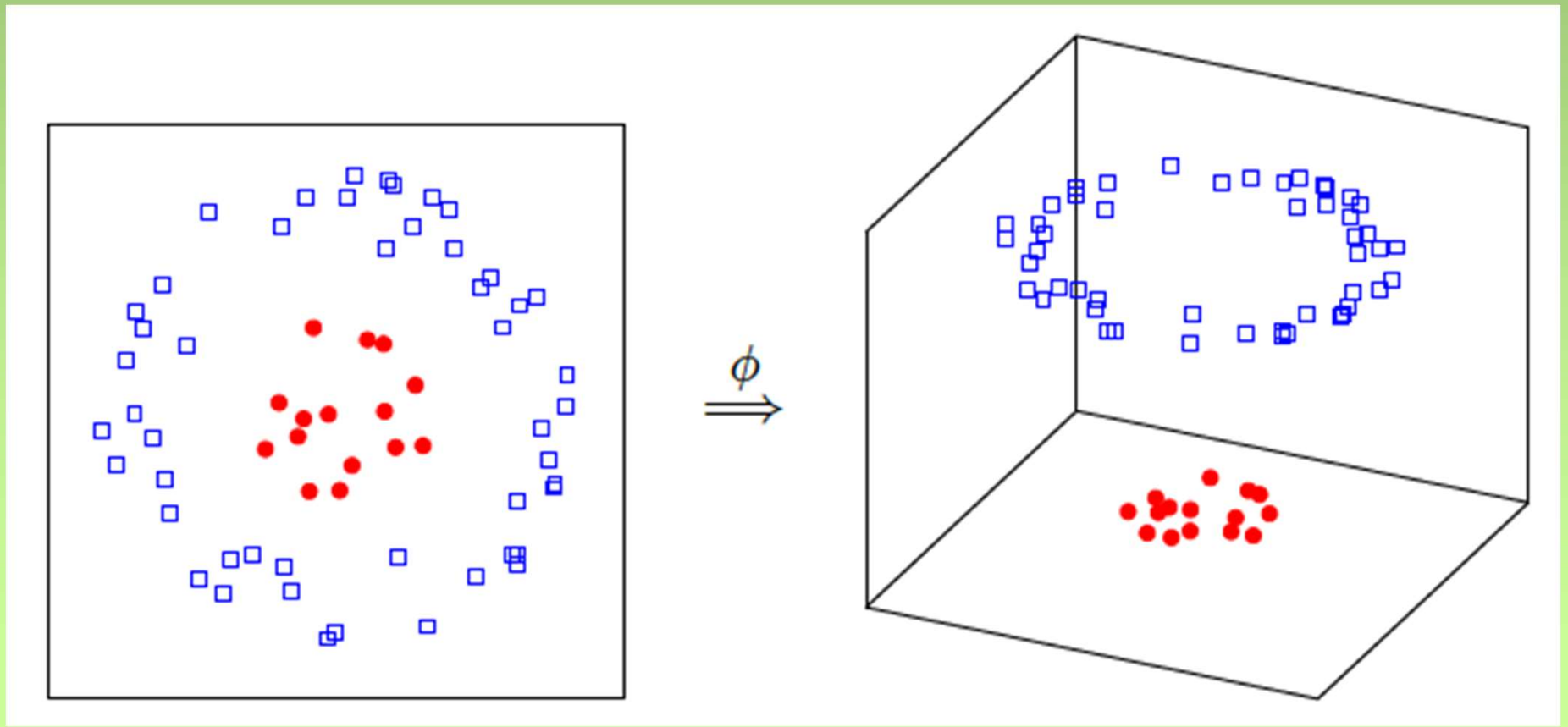
Hay que tener en cuenta que SVM produce un clasificador .....
no solo una función de puntuación

Con Regresión Logística, por ejemplo
Primero entrenamos a un modelo .....
luego generamos puntuaciones y establecemos un umbral

SVM da directamente la clasificación
Salta el paso intermedio

SVMs Lineales

Un ejemplo de lo que puede pasar al hacer una transformación a una dimensión superior



Frontera de decisión y margen de los modelos de las máquinas de vectores soporte, SVM

CASO: Clases linealmente separables

- **Muestras de entrenamiento** $\{(\mathbf{x_i}, y_i)\}^N_{i=1}$

- **Dos clases** $y_i = +1$, $y_i = -1$
- **Función de decisión**

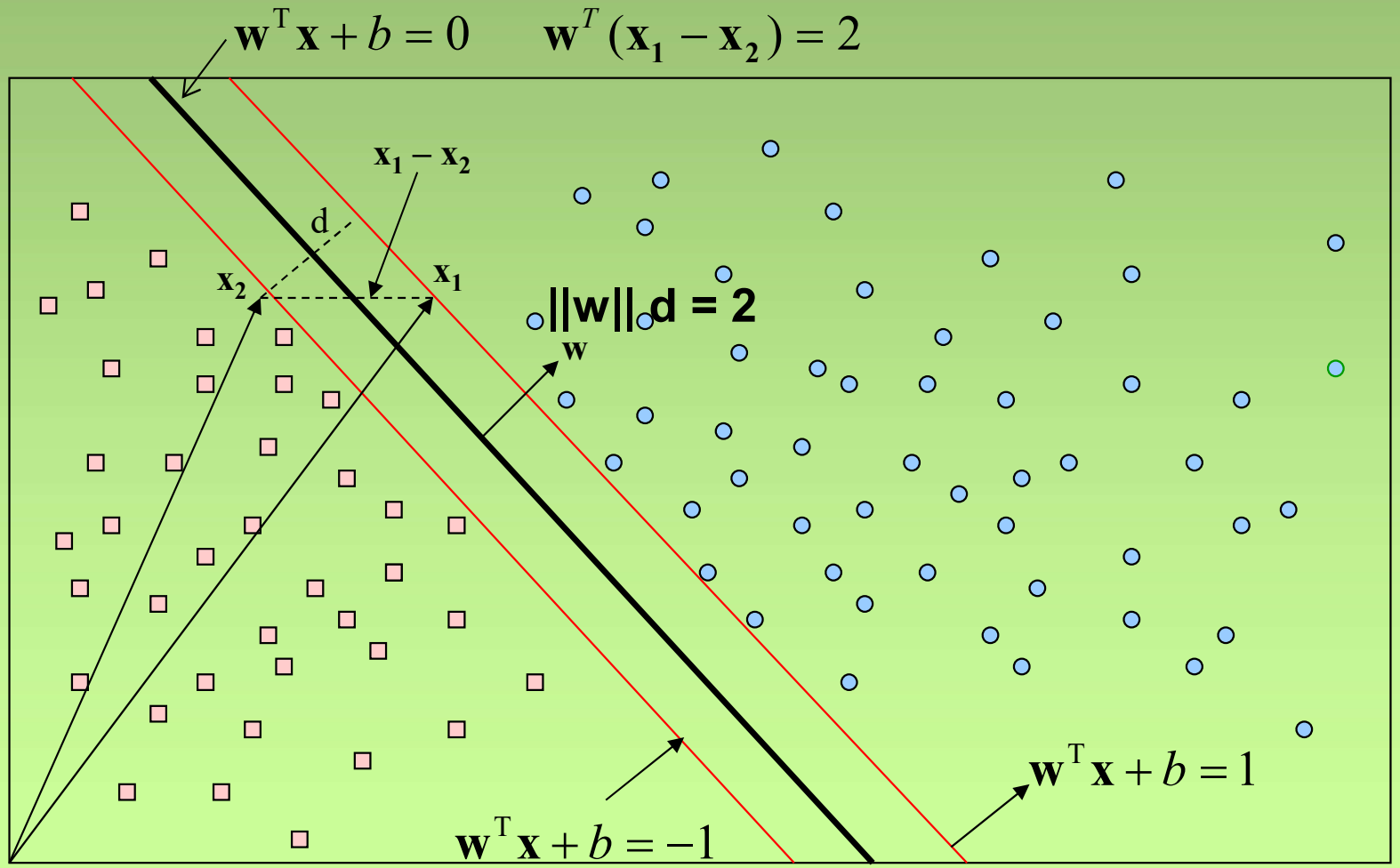
$$\mathbf{wx} + w_0 = 0$$

$$\mathbf{wx_i} + w_0 >= 0 \quad \text{para } y_i = +1$$

$$\mathbf{wx_i} + w_0 < 0 \quad \text{para } y_i = -1$$

El entrenamiento busca estimar los parámetros o pesos $\mathbf{w}$, w_0

Frontera de decisión y margen de los modelos SVM



Modelo de Vectores Soporte, SVM

- **De la gráfica anterior tenemos:**

$$\begin{cases} \mathbf{w}\mathbf{x_1} + w_0 = 1 \\ \mathbf{w}\mathbf{x_2} + w_0 = -1 \end{cases}, \text{ luego restando, tenemos } \mathbf{w}(\mathbf{x_1} - \mathbf{x_2}) = 2$$

- **A partir de esta ecuación podemos determinar un valor** d**, de forma tal que** ||**w**|| d = 2,

luego d = 2 / ||**w**||

Donde d **es el margen de separación a maximizar**

- **Entonces la función de optimización (minimizar en este caso) es una función monótona en w:**

$$f(\mathbf{w}) = ||\mathbf{w}||^2 / 2$$

Esta optimización es un problema de programación cuadrática PQ (Bertsekas 1995)

Un ejemplo inicial unidimensional

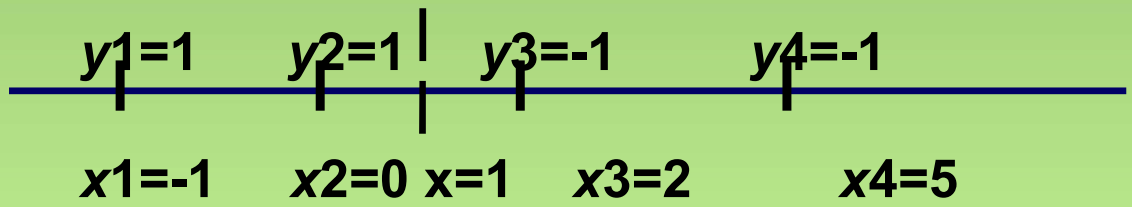
La función discriminante lineal es wx + w0 = 0 y los subespacios son
wxi + w0 > 0 para yi = +1
wxi + w0 < 0 para yi = -1

Clasificación unidimensional: g(x)=wx+w0

– **Conjunto de entrenamiento:**

(x_1=-1, y_1=1) (x_2=0, y_2=1) pertenecen a C_1;

(x_3=2, y_3=-1) (x_4=5, y_4=-1) pertenecen a C_2



Condiciones de separabilidad lineal:

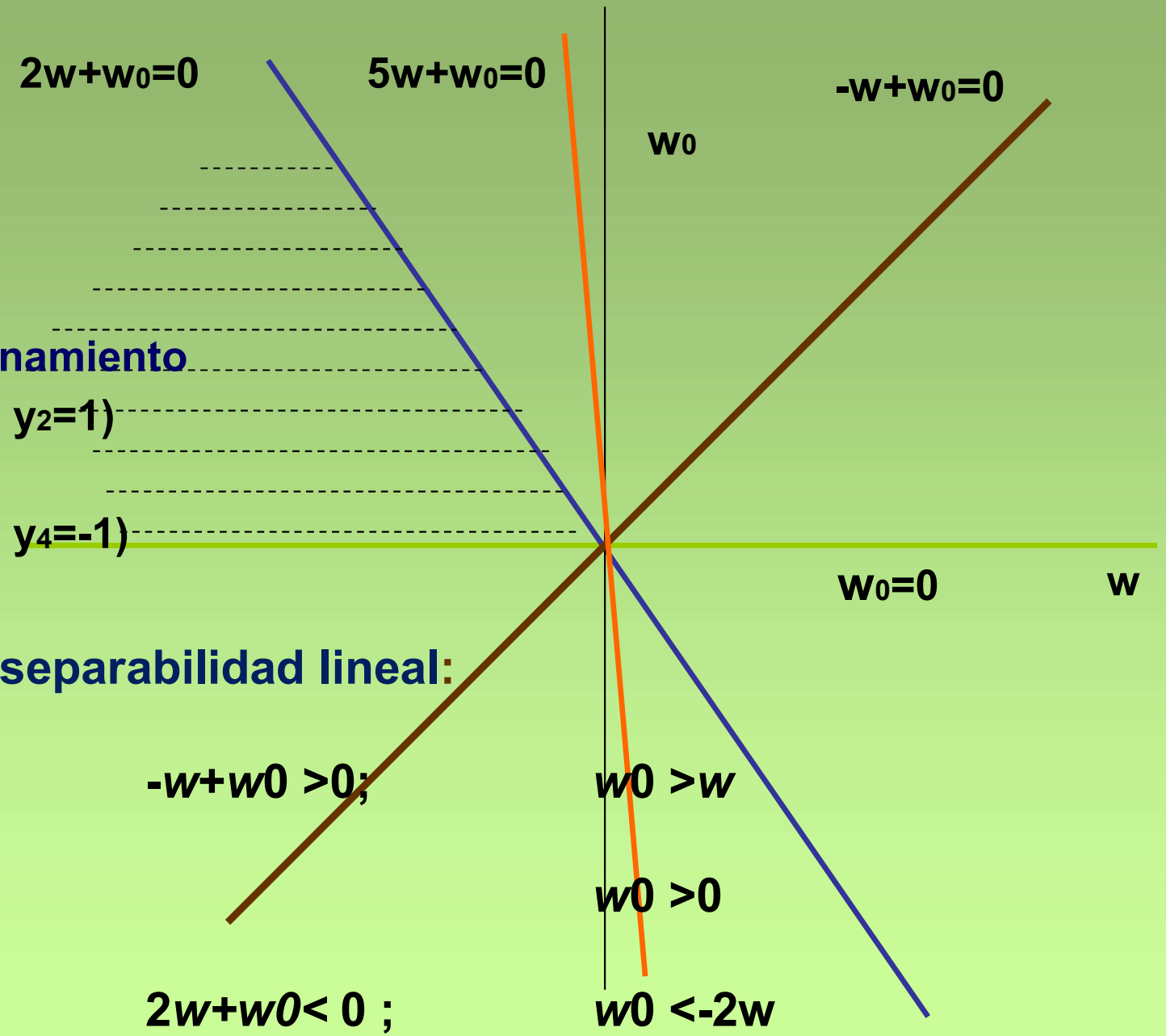
+1 (*wx*1 +*w*0) >0: **-1*w*+*w*0 >0;** ***w*0 >*w***

+1 (*wx*2 +*w*0) >0: **0*w*+*w*0 >0** ***w*0 >0**

-1(*wx*3 +*w*0) > 0: **2*w*+*w*0 < 0;** ***w*0 <-2w**

-1(*wx*4 +*w*0) > 0: **5*w*+*w*0 < 0;** ***w*0 <-5w**

Representación Gráfica del Ejemplo Inicial



Conjunto de entrenamiento

(x_1=-1, y_1=1) (x_2=0, y_2=1) pertenecen a C_1;

(x_3=2, y_3=-1) (x_4=5, y_4=-1) pertenecen a C_2

Condiciones de separabilidad lineal:

+(*wx*1 +*w*0) >0:	**-*w*+*w*0 >0;**	***w*0 >*w***
+(*wx*2 +*w*0) >0:		***w*0 >0**
(*wx*3 +*w*0) < 0:	**2*w*+*w0*< 0 ;**	***w*0 <-2*w***
(*wx*4 +*w*0) < 0:	**5*w*+*w*0 < 0;**	***w*0 <-5*w***

SVM: (CL). Formulación Inicial

- **En el caso de separabilidad lineal:**

– La distancia de un punto $\mathbf{x_i}$ a la frontera lineal dada por es: $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0$

$$d(\mathbf{x_i}, g(\mathbf{x})) = \frac{y_i(\mathbf{w}^T\mathbf{x}_i + w_0)}{\sqrt{\mathbf{w}^T\mathbf{w}}}$$

Esto es así porque la distancia de un punto P(x1,x2) a una recta de ecuación r≡Ax1+Bx2+C=0, es

$$d(P,r) = \frac{Ax_1 + Bx_2 + C}{\sqrt{A^2 + B^2}}$$

- **Por tanto para calcular el margen debemos calcular la menor distancia de los puntos del conjunto de entrenamiento a la frontera, siendo yi, +1 o -1**

$$\text{margen}(\mathbf{w}, w_0) = \min_{i=1\ldots n} \frac{y_i(\mathbf{w}^T\mathbf{x}_i + w_0)}{\sqrt{\mathbf{w}^T\mathbf{w}}}$$

SVM: (CL).Formulación Inicial

Los valores de *w* y w_0 representaran la frontera lineal que separe los puntos del conjunto de entrenamiento pertenecientes a las dos clases de patrones, $y_i=1$ e $y_i=-1$

$$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+\mathrm{w}_0)>0, i=1...n$$

Nos interesa que estén lo más alejadas posible; por lo tanto se obtiene un problema de optimización max min con restricciones:

$$\max_{\mathbf{w},w_0} \quad \min_{i=1...n} \frac{y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+\mathrm{w}_0)}{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}}$$

$$s.a.\ \ y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+\mathrm{w}_0)>0, \quad i=1,...,n$$

MVS:(CL). Normalización (1)

Problemas para buscar la frontera óptima:

– Una frontera lineal no tiene una representación única.

– Si se tiene $g(\mathbf{x})= \mathbf{w}^T\mathbf{x}+w_0 =0$ y se multiplica a ambos miembros de la ecuación por una constante no nula, la ecuación no cambia.

– Es decir, la frontera lineal representada por $g(\mathbf{x})= \mathbf{w}^T\mathbf{x}+w_0 = 0$ es equivalente a

$$\lambda(\mathbf{w}^T\mathbf{x}+w_0)= 0,\ \lambda \neq 0$$

– Esto hace que el problema de optimización sea complicado de resolver.

• **Solución:**

– Elegir de todas las representaciones de una frontera lineal dada, aquella para la que el menor valor de $\left|g(\mathbf{x}_i)\right|= y_i(\mathbf{w}^T\mathbf{x}_i+w_0)$ sea igual a 1.

Normalización

$g(\mathbf{x})= \mathbf{w}^T\mathbf{x}+w_0 =1$

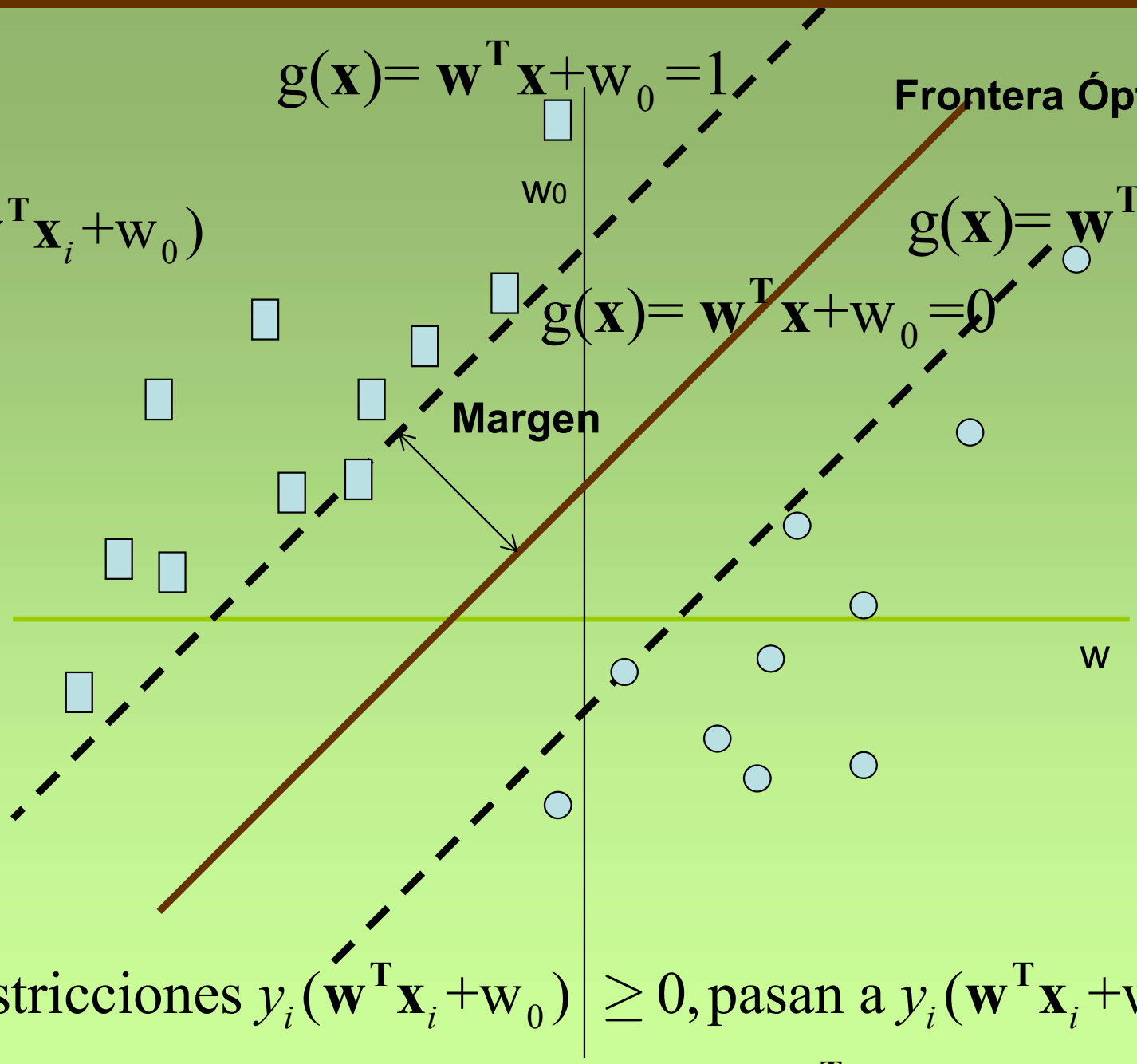
Frontera Óptima

$|g(\mathbf{x}_i)| = y_i(\mathbf{w}^T\mathbf{x}_i+w_0)$

$\min g(\mathbf{x}_i)=1$

$g(\mathbf{x})= \mathbf{w}^T\mathbf{x}+w_0 =-1$

$g(\mathbf{x})= \mathbf{w}^T\mathbf{x}+w_0 =0$



Las restricciones $y_i(\mathbf{w}^T\mathbf{x}_i+w_0) \geq 0$, pasan a $y_i(\mathbf{w}^T\mathbf{x}_i+w_0) \geq 1$

puesto que el menor valor que toma $y_i(\mathbf{w}^T\mathbf{x}_i+w_0)$ es 1

SVM: (CL). Normalización (2)

– **El margen se calcula de forma simple como:**

$$\text{margen}(\mathbf{w}, w_0) = \min_{x_i \in H} \frac{y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + \mathrm{w}_0)}{\sqrt{\mathbf{w}^{\mathbf{T}}\mathbf{w}}} = \frac{1}{\sqrt{\mathbf{w}^{\mathbf{T}}\mathbf{w}}}$$

Las restricciones $y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + \mathrm{w}_0) \geq 0$, pasan a $y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + \mathrm{w}_0) \geq 1$

puesto que el menor valor que toma $y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + \mathrm{w}_0)$ es 1

Para maximizar el margen, esto es $\frac{1}{\sqrt{\mathbf{w}^{\mathbf{T}}\mathbf{w}}}$, hay que maximizar el cociente, o lo que es lo mismo minimizar el denominador o una función monótona del mismo, por lo tanto ahora el problema max min con restricciones es de la forma

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w},$$

$$s.a. \;\; y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + \mathrm{w}_0) \geq 1, \quad i = 1...n$$

Este un problema de programación cuadrática, tiene un único óptimo y puede resolverse con complejidad computacional polinomial.

Normalización del Ejemplo Inicial

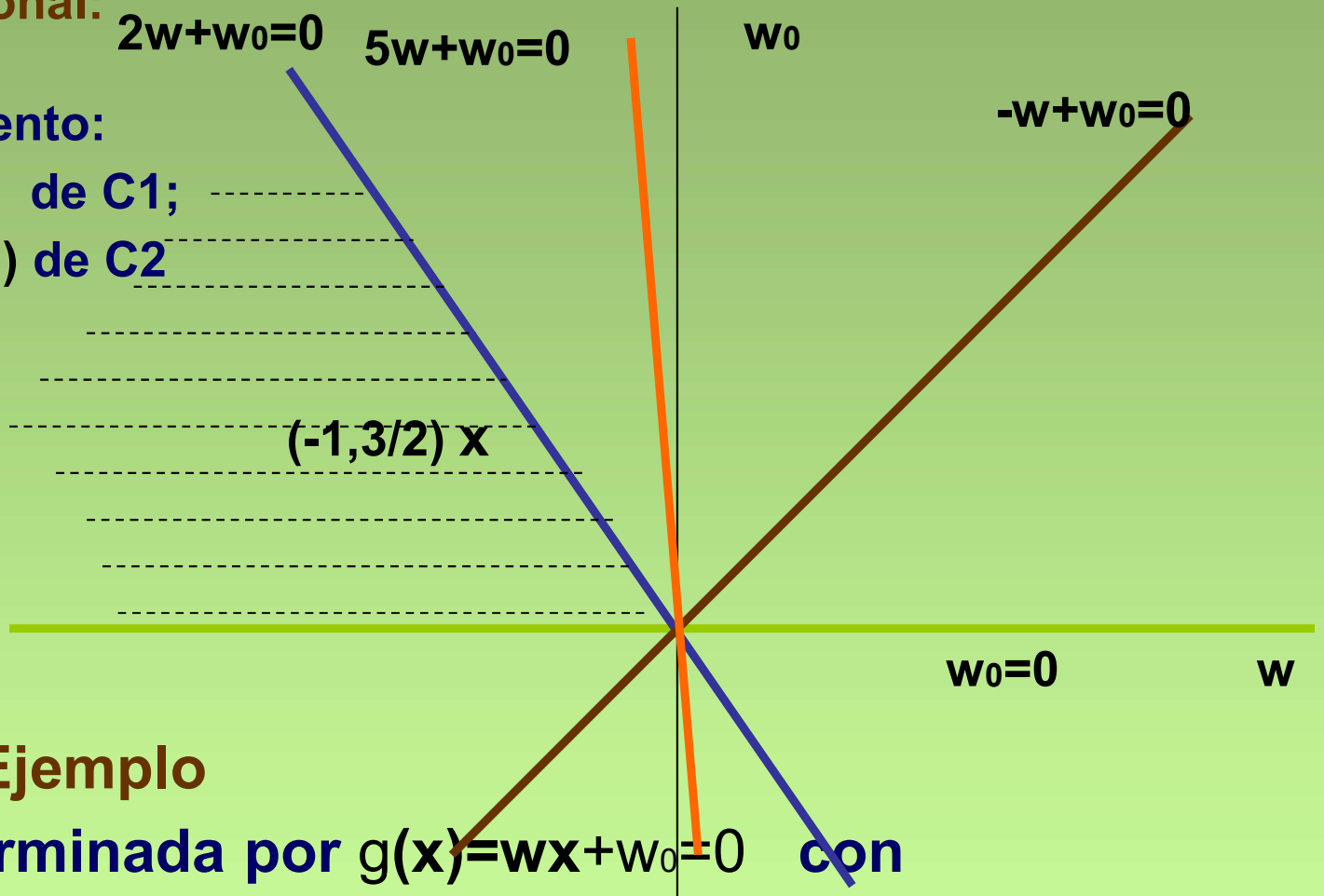
Clasificación unidimensional:

$g(x)=wx+w_0$

– Conjunto de entrenamiento:

(x1=-1, y1=1) (x2=0, y2=1) de C1;

(x3=2, y3=-1) (x4=5, y4=-1) de C2



- **Normalización: Ejemplo**

– La frontera determinada por $g(x)=wx+w_0=0$ con $(w,w_0)= (-1,3/2)$, esto es, $-x+3/2= 0$ o lo que es igual $-2x+3=0$ cumple las restricciones de separación lineal, (ver gráfica, está en la zona factible). Para normalizarla, el menor valor de $|g(x)|= y_i (wx_i+w_0)$ debe ser igual a 1.

Normalización del Ejemplo Inicial

– Calculamos los valores de la función discriminante $-x+3/2=0$ de los puntos del conjunto de entrenamiento, para $(w,w_0)= (-1,3/2)$ y para los puntos de entrenamiento $(x_1=-1, y_1=1)$ $(x_2=0, y_2=1)(x_3=2, y_3=-1)$ $(x_4=5, y_4=-1)$

$y_1 (wx_1+w_0)=1((-1)(-1)+3/2)=5/2$; $y_2 (wx_2+w_0)=1((-1)0+3/2)=3/2$;

$y_3 (wx_3+w_0) =-1((-1)2+3/2)=1/2$; $y_4 (wx_4+w_0)=-1((-1)5+3/2)=7/2$

El menor valor es 1/2 y ocurre para x3= 2.

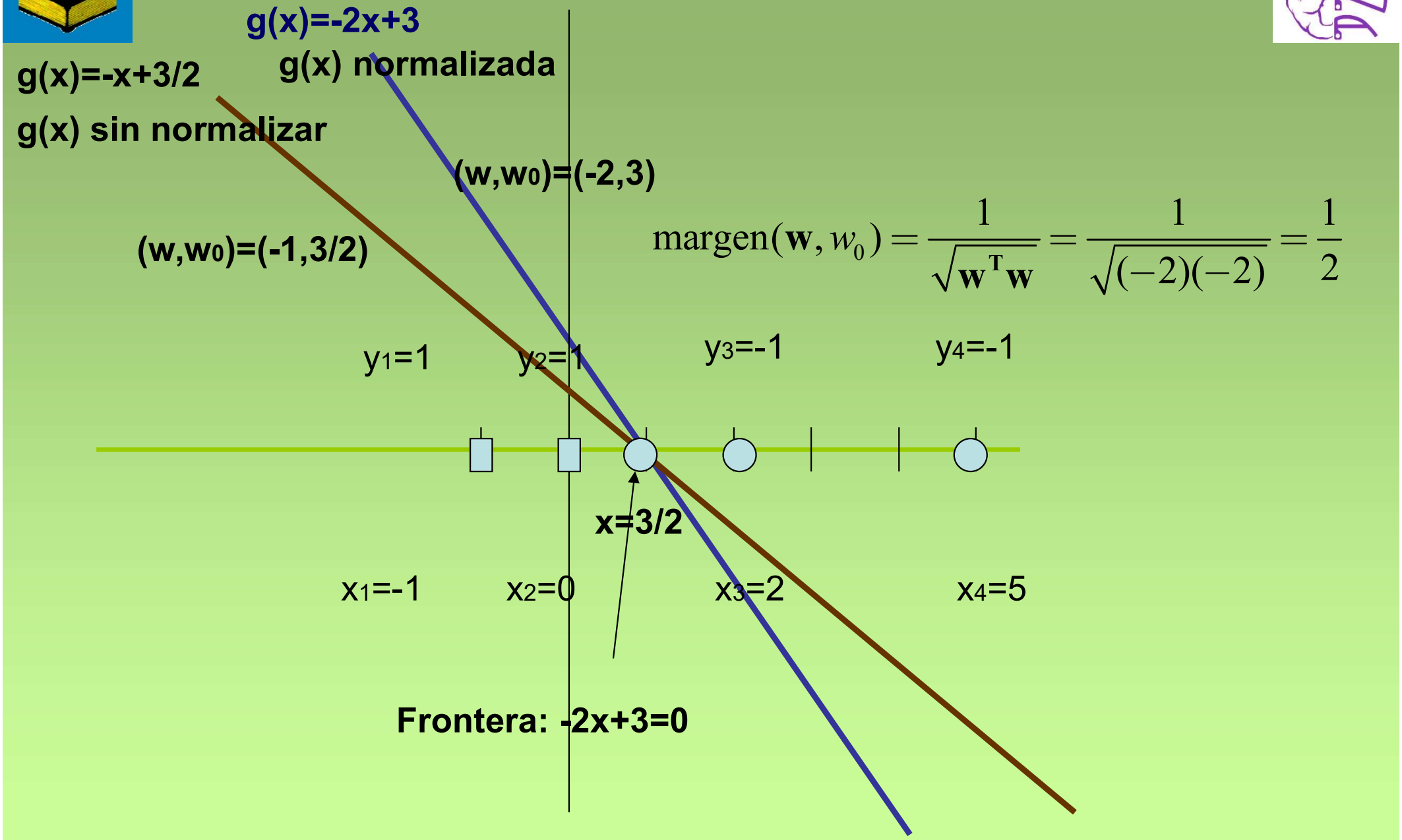
– Entonces dividiendo por 1/2 el par $(w,w_0)=(-1,3/2)$ obtenemos el nuevo vector $(w,w_0)= (-2,3)$ que define exactamente la misma frontera que antes, esto es $-2x+3=0$ pero ahora:

$y_1 (wx_1+w_0)=2+3=5$; $y_2 (wx_2+w_0)=0+3=3$;

$y_3 (wx_3+w_0)=-(-4+3)=1$; $y_4 (wx_4+w_0)=-(-10+3)=7$;

Además ahora, siempre es $y_i (wx_i +w_0) \geq 1$

Gráfica Ejemplo Inicial



Ejemplo Inicial: Solución normalizada

Clasificación unidimensional: g(**x**)=**-2x+3**

– Conjunto de entrenamiento:

(x1=-1, y1=1) (x2=0, y2=1) de C1;

(x3=2, y3=-1) (x4=5, y4=-1) de C2

Condiciones de separabilidad lineal NO normalizadas:

+ (*w*x1 +*w*0) > 0: -*w*+*w*0 >0; *w*0 >*w*

+ (*w*x2 +*w*0) > 0: *w*0 >0

(*w*x3 +*w*0) < 0: 2*w*+*w*0 <0; *w*0 <-2w

(*w*x4 +*w*0) < 0: 5*w*+*w*0 <0; *w*0 <-5w

Condiciones de separabilidad lineal normalizadas:

+(*w*x1 +*w*0)>=1: -*w*+*w*0>=1; *w*0>=*w*+1, -w+w0≥1

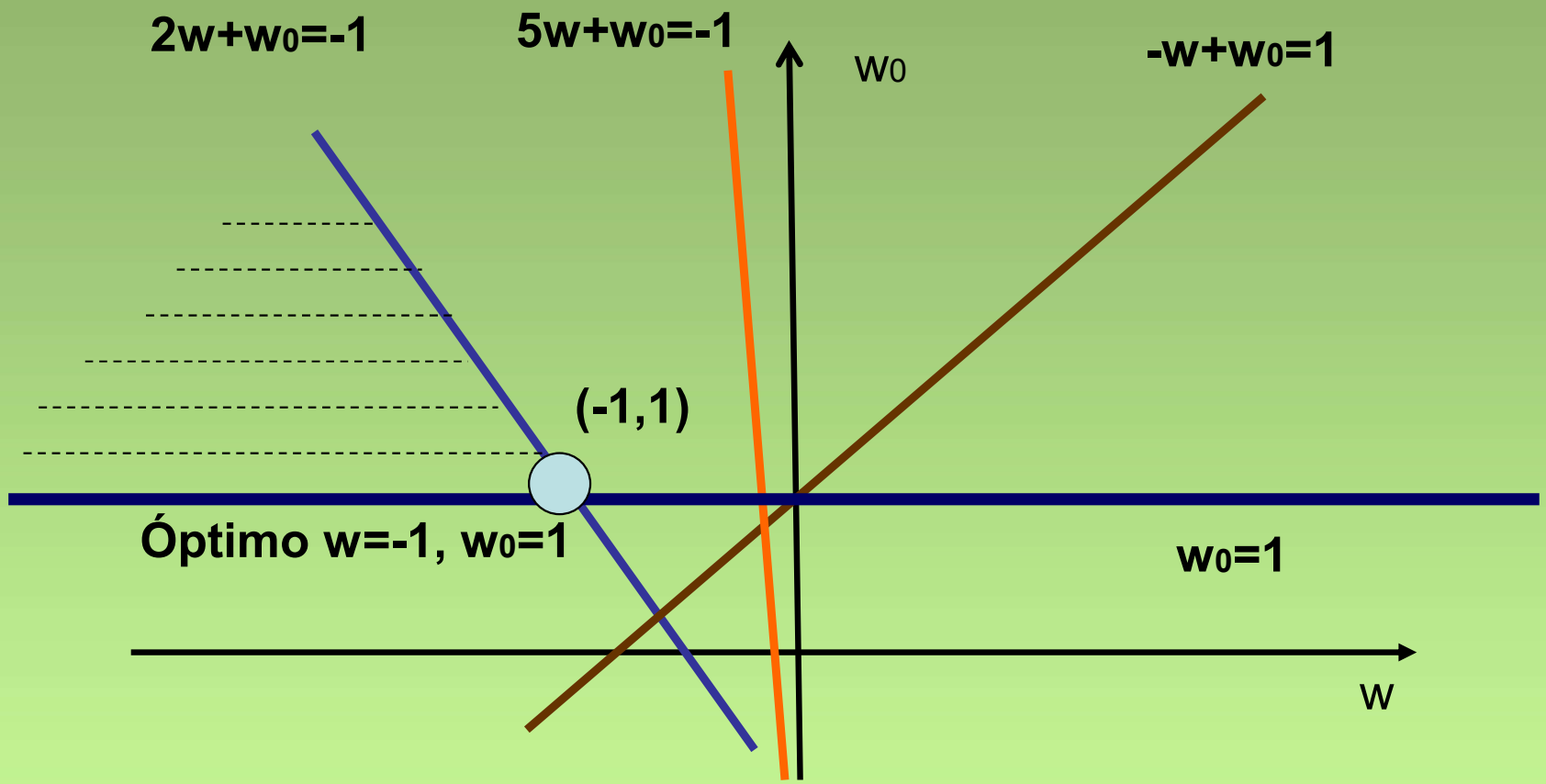
+(*w*x2 +*w*0)>=1: *w*0>=1;

(*w*x3 +*w*0)<=-1: 2*w*+*w*0<=-1; *w*0 <=-2w-1, -2w-w0≥1

(*w*x4 +*w*0)<=-1: 5*w*+*w*0<=-1; *w*0 <=-5w-1, -5w-w0 ≥1

Ejemplo Inicial: Solución normalizada

Región rayada: Separabilidad lineal normalizada



función a maximizar $\frac{1}{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}}$ o lo que es igual

minimizar $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} = \frac{1}{2}w^2$

Un Ejemplo Inicial: Solución

Clasificador unidimensional

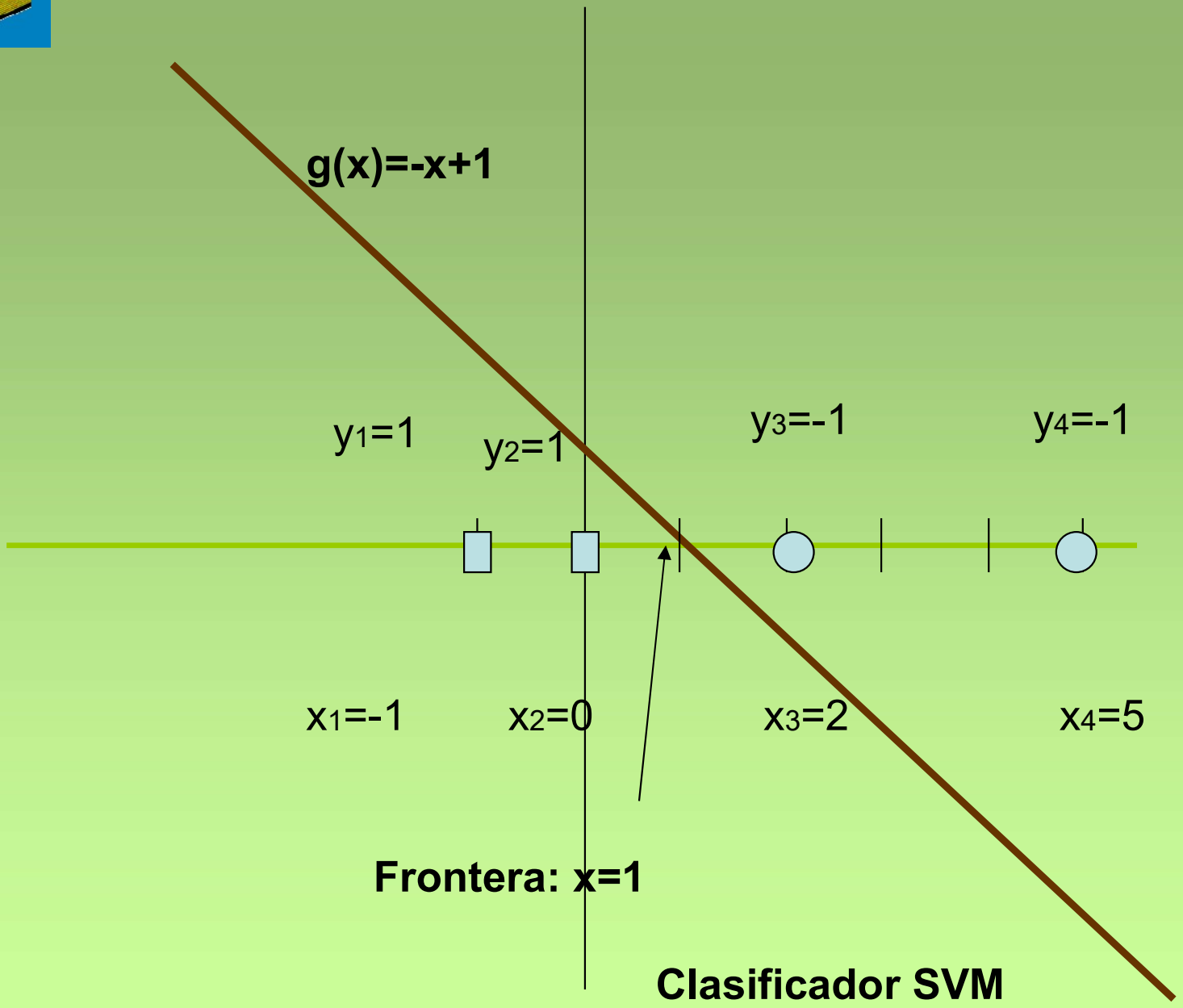
Función a minimizar: $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} = \frac{1}{2}w^2$

Valor óptimo:

– El menor valor que puede tomar $\frac{1}{2}w^2$ **en la región sombreada se obtiene con** w**=-1** w_0**=1.**

– Por tanto el clasificador SVM es: $g(x)=-x+1$ **y la frontera** $g(x)=0$ **se obtiene para x=1**

Un Ejemplo Inicial: Frontera



Vectores soporte

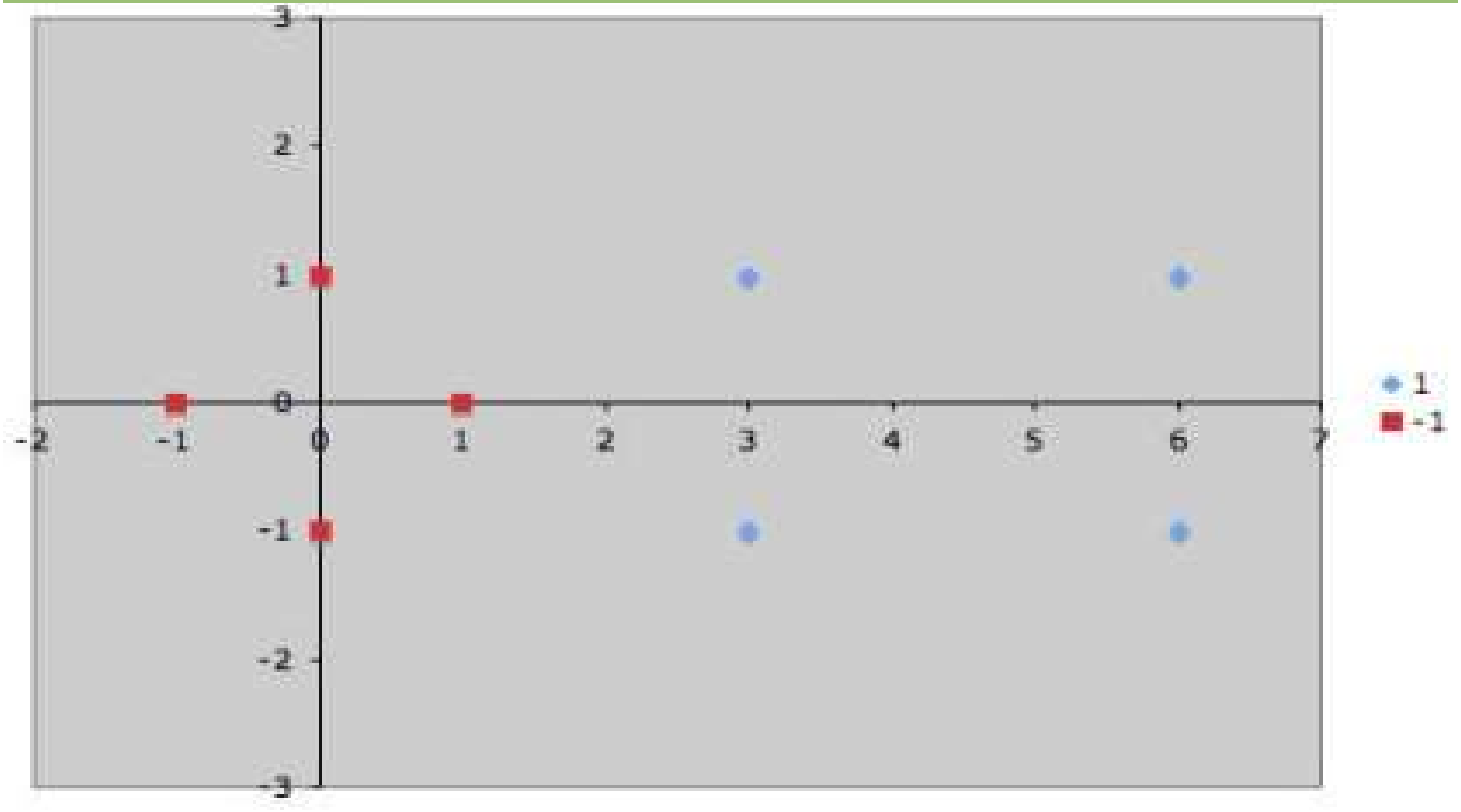
–Son los patrones correspondientes a los puntos que están sobre el margen. Estos patrones se llaman vectores soporte.

– En la práctica son pocos los elementos que están sobre el margen. Así, son los que determinan los márgenes de la función discriminante lineal

Nuevo ejemplo linealmente separable

Clase negativa (-1, 0; -1) (0, -1; -1) (0, 1; -1) (1, 0; -1)

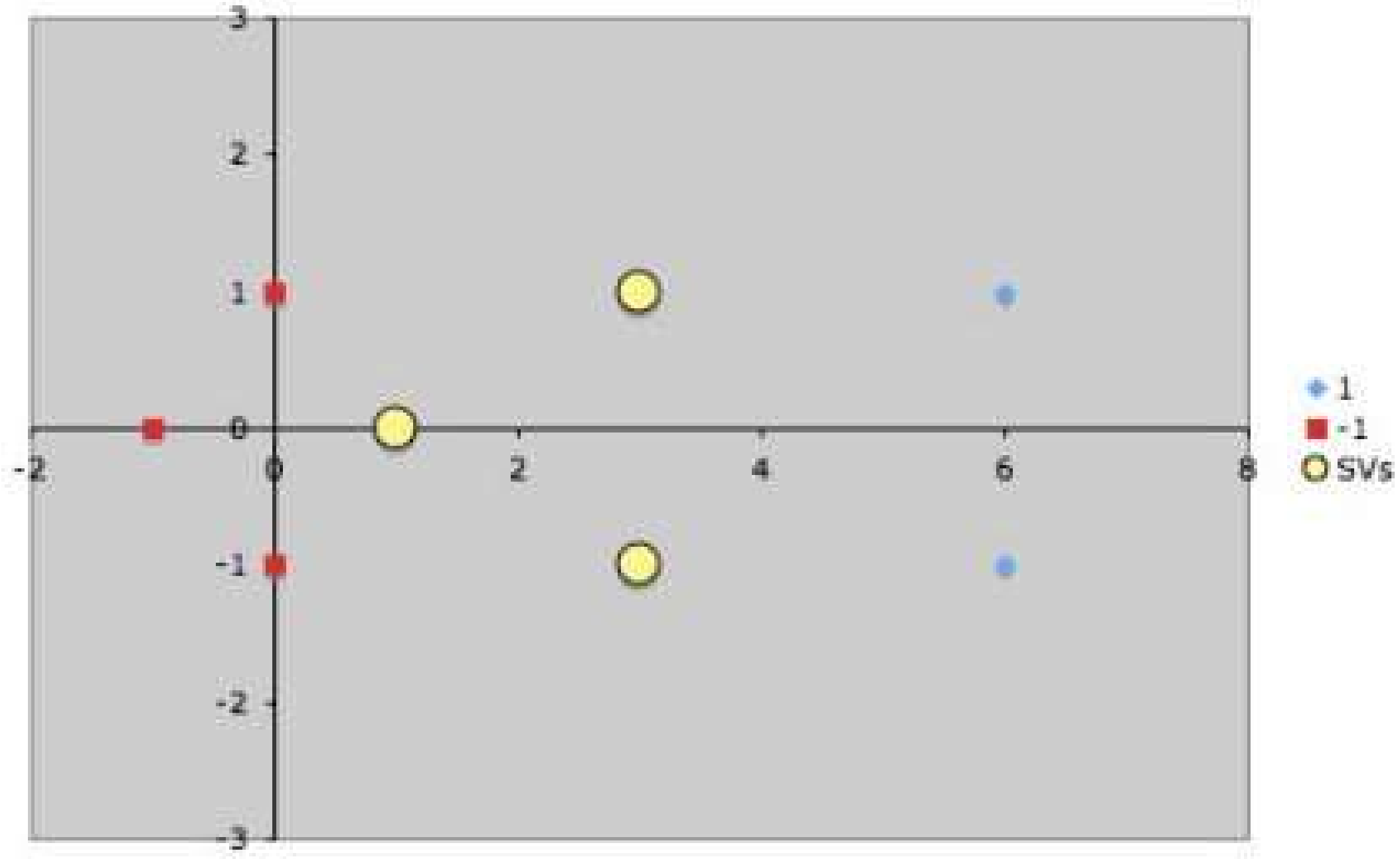
Clase positiva (3, -1; +1) (3, 1; +1) (6, -1; +1) (6, 1; +1)



Ejemplo: Vectores soporte

Clase negativa (-1, 0; -1) (0, -1; -1) (0, 1; -1) (1, 0; -1)

Clase positiva (3, -1; +1) (3, 1; +1) (6, -1; +1) (6, 1; +1)



Ejemplo

Clase positiva (3, -1, +1) (3, 1, +1) (6, -1, +1) (6, 1, +1)

Clase negativa (-1, 0, -1) (0, -1, -1) (0, 1, -1) (1, 0, -1)

H_1: $\mathbf{w} \cdot \mathbf{x} + w_0 = 1$

H_2: $\mathbf{w} \cdot \mathbf{x} + w_0 = -1$

$w_1x_1 + w_2x_2 + w_0 = -1$

$1w_1 + 0w_2 + w_0 = -1$ **(1)**

$\rightarrow w_0 = -1 - w_1$

$w_1x_1 + w_2x_2 + w_0 = 1$

$3w_1 - 1w_2 + w_0 = 1$ **(2)**

$\rightarrow w_2 = 3w_1 - 1 - w_1 - 1$

$\rightarrow w_2 = 2w_1 - 2$

$3w_1 + 1w_2 + w_0 = 1$ **(3)**

$\rightarrow 3w_1 + 2w_1 - 2 - 1 - w_1 = 1$

$\rightarrow w_1 = 1$

$\rightarrow w_0 = -2$

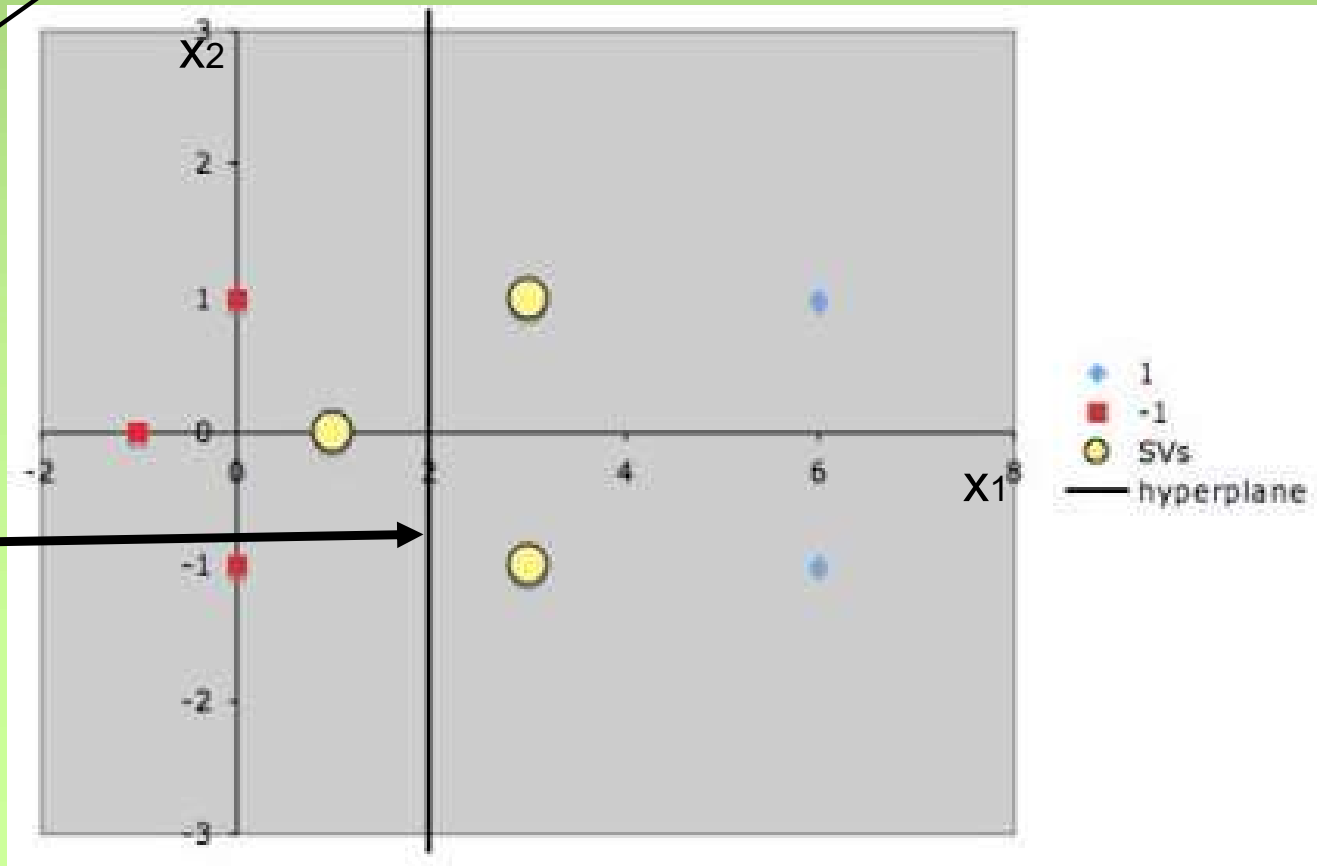
$\rightarrow w_2 = 0$

(1, 0) → -1

(3, -1) → +1

(3, 1) → +1

Vectores soporte



Función discriminante

$$(1,0)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 2 = 0$$

Recta: $x_1 = 2$

DUALIDAD EN PROGRAMACION LINEAL

Relaciones primal-dual

Asociado a cada problema lineal existe otro problema de programación lineal denominado *problema dual, que posee* importantes propiedades y relaciones notables con respecto al problema lineal original, problema que para diferencia del dual se denomina entonces como *problema primal.*

Las relaciones las podemos enumerar como siguen:

a) El problema *dual tiene tantas variables como restricciones tiene el problema primal.*

b) El problema *dual tiene tantas restricciones como variables tiene el problema primal*

c) Los coeficientes de la función objetivo del problema *dual* son los términos independientes de las restricciones del problema *primal.*

DUALIDAD EN PROGRAMACION LINEAL

d) Los términos independientes de las restricciones del problema dual *son los coeficientes de la función objetivo del problema primal.*

e) La matriz de coeficientes técnicos del problema *dual es la* traspuesta de la matriz de coeficientes técnicos del problema *primal.*

f) El sentido de las desigualdades de las restricciones del problema *dual y el signo de las variables del mismo problema,* dependen de la forma que tenga el signo de las variables del problema *primal y del sentido de las restricciones del mismo* problema.

g) Si el problema *primal es un problema de maximización,* el problema *dual es un problema de minimización.*

h) El problema *dual de un problema dual es el problema primal original*

TABLA DE TUKER

MAXIMIZACIÓN			MINIMIZACIÓN
RESTRICCIONES	≤	≥	VARIABLES
	≥	≤	
	=	><	
VARIABLES	≥	≥	RESTRICCIONES
	≤	≤	
	><	=	

Los problemas duales simétricos son los que se obtienen de un problema primal en forma canónica y 'normalizada', es decir, cuando llevan asociadas desigualdades de la forma mayor o igual en los problemas de minimización, y desigualdades menor o igual para los problemas de maximización

Optimización lineal con restricciones lineales

Si el problema primal es:

$$Max\ Z(\mathbf{x}) = \mathbf{c}^T \mathbf{x}$$

$$s.a.$$

$$\mathbf{Ax} \leq \mathbf{b}$$

$$\mathbf{x} \geq \mathbf{0}$$

Ejemplo. Dado el modelo lineal

$$\max z = 2x1 - x2 + 3x3$$

sujeto a

$$x1 - x2 + x3 \leq 2$$
$$3x1 - x2 + 2x3 \leq 1$$
$$x1, x2, x3 \geq 0$$

3 variables, dos restricciones

El problema dual (dual simétrico) es:

$$Min\ G(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathbf{T}}\mathbf{b}$$

$$s.a.$$

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{\lambda} \geq \mathbf{c}$$

$$\boldsymbol{\lambda} \geq \mathbf{0}$$

$$\min G = 2\lambda1 + \lambda2$$

sujeto a

$$\lambda1 + 3\lambda2 \geq 2$$
$$-\lambda1 - \lambda2 \geq -1$$
$$\lambda1 + 2\lambda2 \geq 3$$
$$\lambda1, \lambda2 \geq 0$$

Dos variables, tres restricciones

¿ Porqué se plantea el programa dual?.

a) Por una parte permite resolver problemas lineales donde el numero de restricciones es mayor que el numero de variables.

Gracias a los teoremas que no expondremos (ver apuntes adicionales) la solución de uno de los problemas (primal o dual) nos proporciona de forma automática la solución del otro problema.

b) Otra de las ventajas de la dualidad, es la posibilidad de resolver gráficamente algunos problemas.

Ejemplo

Consideremos el siguiente problema lineal

$$\text{Min } Z(x) = 2x_1 + 3x_2 + 5x_3 + 2x_4 + 3x_5$$

s.a:

$$x_1 + x_2 + 2x_3 + x_4 + 3x_5 \geq 4$$

$$2x_1 - x_2 + 3x_3 + x_4 + x_5 \geq 3$$

$$x_1 \geq 0 \, , \, x_2 \geq 0 \, , \, x_3 \geq 0 \, , \, x_4 \geq 0 \, , \, x_5 \geq 0$$

Los términos independientes de las restricciones del problema dual *son los coeficientes de la función objetivo del problema primal.*

Dado que se trata de un programa lineal en forma canónica, ello nos proporciona un dual en forma simétrica como el siguiente:

$$\text{Max } G(\lambda) = 4\lambda_1 + 3\lambda_2$$

s.a:

$$\lambda_1 + 2\lambda_2 \leq 2$$

$$\lambda_1 - \lambda_2 \leq 3$$

$$2\lambda_1 + 3\lambda_2 \leq 5$$

$$\lambda_1 + \lambda_2 \leq 2$$

$$3\lambda_1 + \lambda_2 \leq 3$$

$$\lambda_1 \geq 0 \, , \, \lambda_2 \geq 0$$

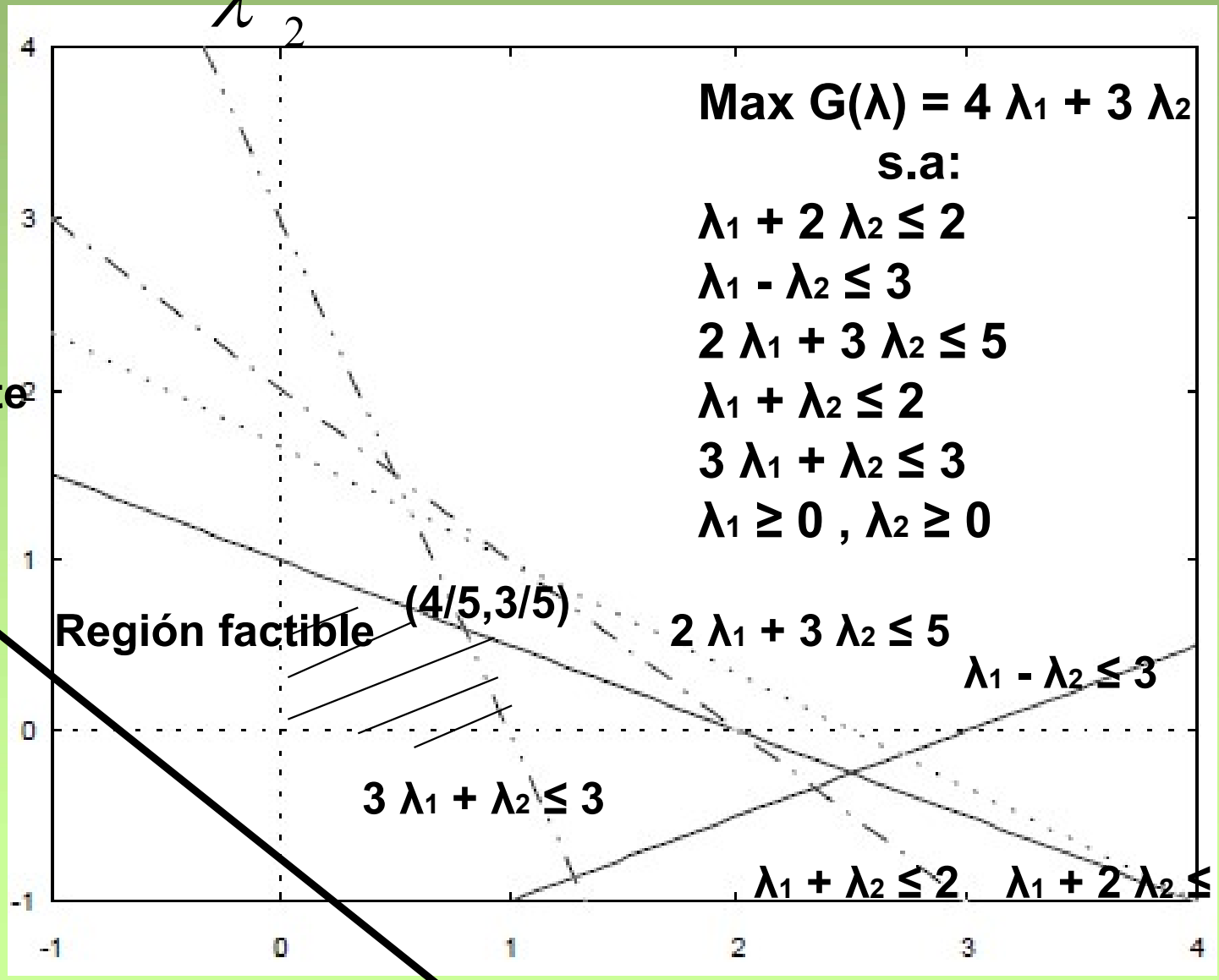
El problema *dual tiene tantas variables como restricciones tiene el problema primal.*

El problema *dual tiene tantas restricciones como variables tiene el problema primal*

Los coeficientes de la función objetivo del problema *dual* son los términos independientes de las restricciones del problema *primal.*

Ejemplo

Este problema solo tiene dos variables y cinco restricciones por tanto se puede resolver gráficamente:



Ejemplo

El vértice solución es el punto (4/5,3/5) con un valor de la función objetivo de

$$G(\lambda) = 4\lambda_1 + 3\lambda_2 = 16/5 + 9/5 = 5$$

Si lo resolvemos por el primal, para minimizar Z(x) hacemos

x_2, x_3 y x_4 =0, luego

$$x_1 + 3x_5 = 4$$
$$2x_1 + x_5 = 3$$

La solución de este sistema es: $x_1 = 1$ y $x_5 = 1$, lo cual nos proporciona un valor de la función objetivo de

$$Z(x) = 2x_1+3x_2+5x_3+2x_4+3x_5 = 2+3+0+0+0=5,$$

idéntico a la solución del dual

Condiciones de Kuhn-Tucker en los problemas lineales (primales y duales).

Consideremos el siguiente problema de optimización lineal, que denominaremos **PRIMAL**:

$$Max\ Z(\mathbf{x}) = \mathbf{c}^T \mathbf{x}$$

$$s.a.$$

$$\mathbf{A}\mathbf{x} \leq \mathbf{b}$$

$$\mathbf{x} \geq \mathbf{0}$$

La función lagrangiana a optimizar de este problema será:

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{c}^T \mathbf{x} + \boldsymbol{\lambda}(\mathbf{b} - \mathbf{A}\mathbf{x})$$

Donde $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)$

representa el vector de los multiplicadores de Lagrange asociados a las **m** restricciones

Condiciones de Kuhn-Tucker)

Las condiciones de optimalidad de este problema

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{c}^T \mathbf{x} + \boldsymbol{\lambda}(\mathbf{b} - \mathbf{A}\mathbf{x})$$

respecto de las variables, son:

$$\frac{\partial L}{\partial \mathbf{x}} = \mathbf{c} - \boldsymbol{\lambda}\mathbf{A} \leq \mathbf{0} \quad \text{o} \quad \mathbf{x}\frac{\partial L}{\partial \mathbf{x}} = \mathbf{x}(\mathbf{c} - \boldsymbol{\lambda}\mathbf{A}) = \mathbf{0}, \text{ para } \mathbf{x} \geq \mathbf{0}$$

Respecto a los multiplicadores, son

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = \mathbf{b} - \mathbf{A}\mathbf{x} \geq \mathbf{0} \quad \text{o} \quad \boldsymbol{\lambda}\frac{\partial L}{\partial \boldsymbol{\lambda}} = \boldsymbol{\lambda}(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{0}, \text{ para } \boldsymbol{\lambda} \geq \mathbf{0}$$

Como puede observarse, ambas condiciones de optimalidad son las mismas para los dos problemas. Por lo tanto, asociado a todo problema de programación lineal existe otro problema de programación lineal denominado programa dual que tiene importantes relaciones con el problema original denominado programa primal

Programación Lineal: Inferencia Dual

- Si un problema de programación lineal tiene la forma

$$\underset{\mathbf{x}\geq\mathbf{0}}{\text{Min}}\left\{ \mathbf{c}^T\mathbf{x} \,\middle|\, \mathbf{A}\mathbf{x} \geq b \right\}$$

- La inferencia dual es

$$\underset{P\in\mathcal{P}}{\text{Max}}\left\{ v \,\middle|\, \mathbf{A}\mathbf{x} \geq b \underset{P}{\rightarrow} \left(\mathbf{c}^T\mathbf{x} \geq v\right) \right\}$$

  - La familia P de pruebas consta de la dominancia por sustituciones (combinaciones lineales no negativas).

Programación Lineal: Inferencia Dual

De $\min_{\mathbf{x}\geq\mathbf{0}} \left\{ \mathbf{c}^T\mathbf{x} \mid \mathbf{Ax} \geq b \right\}$ pasamos a $\max_{P\in\mathcal{P}} \left\{ v \mid \mathbf{Ax} \geq b \underset{P}{\rightarrow} \left(\mathbf{c}^T\mathbf{x} \geq v \right) \right\}$

- **Def 1.- Para un $\mathbf{u} \geq \mathbf{0}$,**

  $\mathbf{uAx} \geq \mathbf{ub}$ es una sustitución de $\mathbf{Ax} \geq \mathbf{b}$.

- **Def 2.- $\mathbf{uAx} \geq \mathbf{ub}$ es una dominancia de $\mathbf{cx} \geq \mathbf{v}$**

  si $\mathbf{uAx} \geq \mathbf{ub}$ implica que $\mathbf{cx} \geq \mathbf{v}$ (para $\mathbf{x} \geq \mathbf{0}$).

  - **Esto es, $\mathbf{uA} \leq \mathbf{c}$ y $\mathbf{ub} \geq \mathbf{v}$.**

- **Esto es un método de inferencia completa.**

  - **Debido al Lema de Farkas.**

Programación Lineal: Inferencia Dual

- **Así, la inferencia dual**

$$\underset{P\in\mathcal{P}}{\text{Max}}\left\{ v \,\middle|\, \mathbf{Ax} \geq b \underset{P}{\rightarrow} \left(\mathbf{cx} \geq v\right) \right\}$$

se convierte en

$$\underset{u\geq 0}{\text{Max}}\left\{ v \,\middle|\, \mathbf{uA} \leq c, \mathbf{ub} \geq v \right\}$$

o

$$\underset{u\geq 0}{\text{Max}}\left\{ ub \,\middle|\, \mathbf{uA} \leq c \right\}$$

cuando $\mathbf{Ax} \geq \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$ es factible

  - **Este es el modelo clásico dual de programación lineal.**

Programación Lineal: Inferencia Dual

- Puesto que la familia de pruebas es completa, tenemos una dualidad fuerte:

$$\max_{\mathbf{u}\geq\mathbf{0}}\left\{\mathbf{u}b \mid \mathbf{uA}\leq c\right\} = \min_{\mathbf{x}\geq\mathbf{0}}\left\{c\mathbf{x} \mid \mathbf{Ax}\geq b\right\}$$

Excepto cuando uA ≤ c, u ≥ 0 y Ax ≥ b, x ≥ 0 son ambas no factibles.

En este caso, la dualidad es simétrica

- **El dual del dual es el primal.**
- **De esta forma el dual pertenece a la clase NP y el primal a la co-NP.**

Inferencia Dual

- A veces se verifica una **dualidad fuerte**, esto es:

Máximo valor del problema dual = **Mínimo valor del problema primal**

Si P es una familia de pruebas completa $\Rightarrow$ **Dualidad fuerte**

$$\underset{P\in\mathcal{P}}{\text{Max}}\left\{ v \,\middle|\, C \underset{P}{\rightarrow} \left(f(\mathbf{x}) \geq v \right) \right\}$$

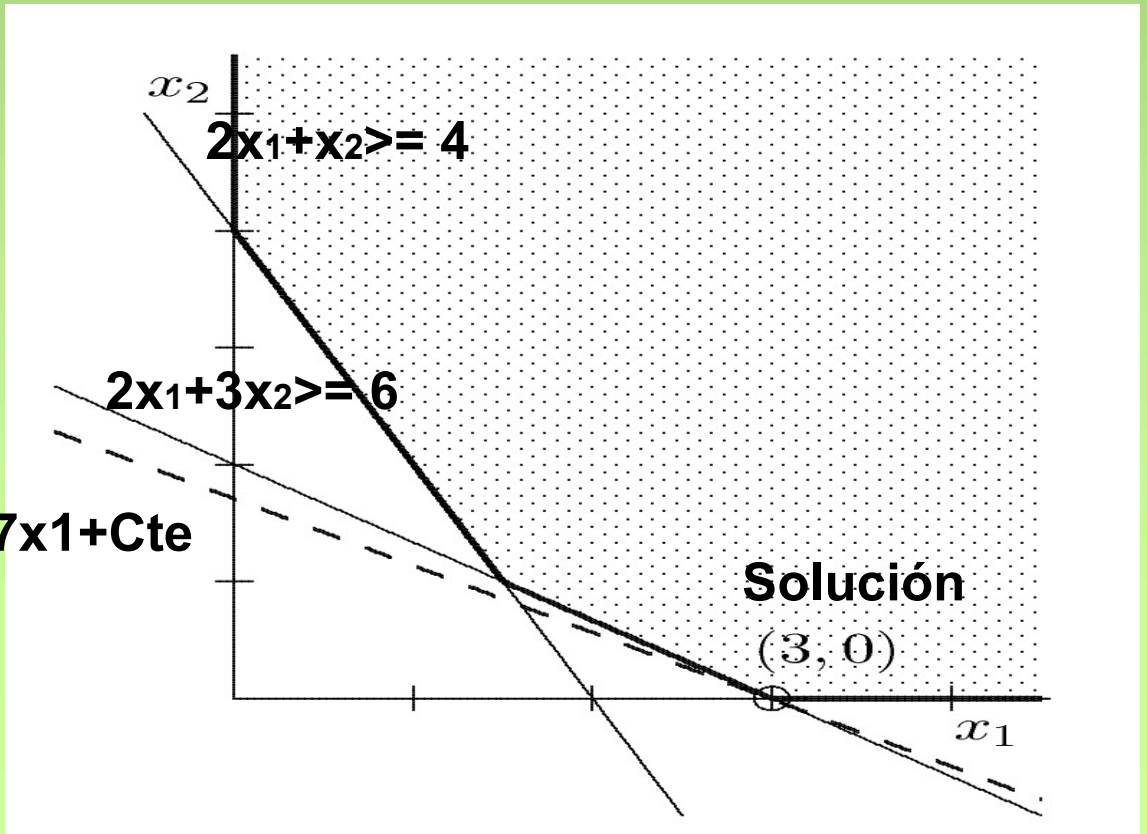
Programación Lineal: Inferencia Dual, ejemplo

- **Primal:**

$$\text{Min } 4x_1 + 7x_2$$
$$2x_1 + 3x_2 \geq 6 \quad (u_1)$$
$$2x_1 + x_2 \geq 4 \quad (u_2)$$
$$x_1, x_2 \geq 0$$

Dual:

$$\text{Max } 6u_1 + 4u_2$$
$$2u_1 + 2u_2 \leq 4 \quad (x_1)$$
$$3u_1 + u_2 \leq 7 \quad (x_2)$$
$$u_1, u_2 \geq 0$$



Programación Lineal: Inferencia Dual (una solución)

Primal:

$$\text{Min } 4x_1 + 7x_2 \quad = 12$$
$$2x_1 + 3x_2 \geq 6 \quad (u_1 = 2)$$
$$2x_1 + x_2 \geq 4 \quad (u_2 = 0)$$
$$x_1, x_2 \geq 0$$

Dual:

$$\text{Max } 6u_1 + 4u_2 \quad = 12$$
$$2u_1 + 2u_2 \leq 4 \quad (x_1 = 3)$$
$$3u_1 + u_2 \leq 7 \quad (x_2 = 0)$$
$$u_1, u_2 \geq 0$$

Solución

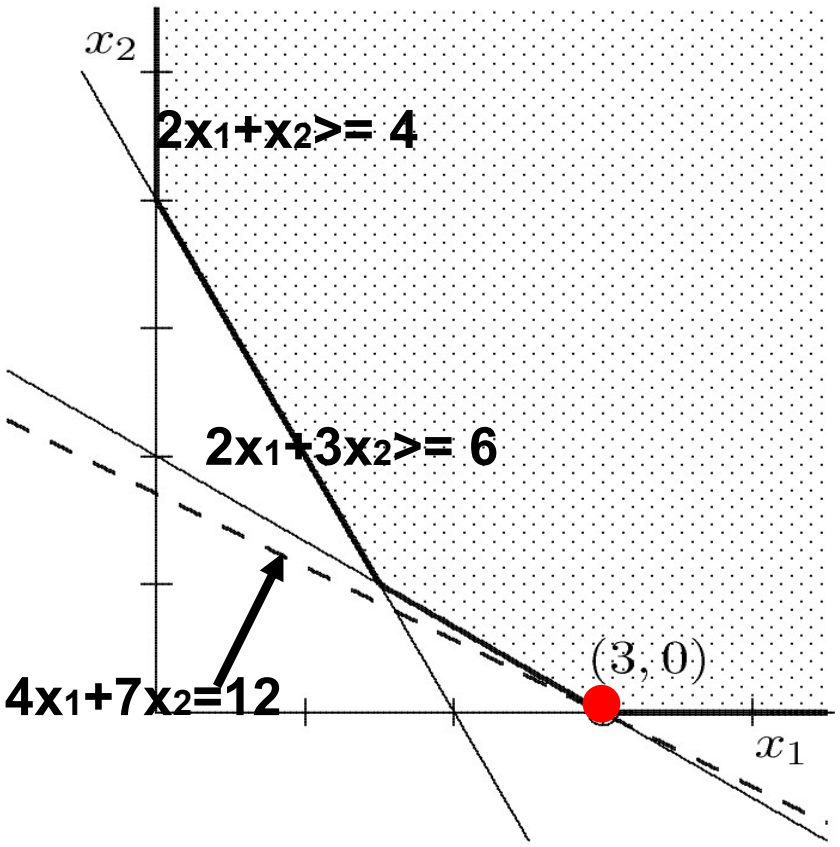
Sustitución:

Multiplicamos la inecuación 1 por 2 →

$$2x_1 + 3x_2 \geq 6 \quad (2)$$
$$2x_1 + x_2 \geq 4 \quad (0)$$
$$4x_1 + 6x_2 \geq 12$$

Dominancia → ⇓

$$4x_1 + 7x_2 \geq 12$$



Programación Lineal: Inferencia Dual (otra posible solución, que no es factible)

- **Primal:**

$$\text{Min } 4x_1 + 7x_2 \quad = 8$$
$$2x_1 + 3x_2 \geq 6 \quad (u_1 = 0)$$
$$2x_1 + x_2 \geq 4 \quad (u_2 = 2)$$
$$x_1, x_2 \geq 0$$

Dual:

$$\text{Max } 6u_1 + 4u_2 \quad = 8$$
$$2u_1 + 2u_2 \leq 4 \quad (x_1 = 2)$$
$$3u_1 + u_2 \leq 7 \quad (x_2 = 0)$$
$$u_1, u_2 \geq 0$$

Solución

Sustitución:

Multiplicamos la inecuación 2 por 2

$$2x_1 + 3x_2 \geq 6 \quad (0)$$
$$2x_1 + x_2 \geq 4 \quad (2)$$
$$4x_1 + 2x_2 \geq 8$$

Dominancia

$$\Downarrow$$

$$4x_1 + 7x_2 \geq 8$$

Pero la solución (2,0) no verifica la inecuación primera, pues 4 no es ≥ 6, luego no es solución factible

MVS:(CL).Resolución Práctica (1)

Para resolver el problema primal $\begin{cases} \min\limits_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} \\ s.a. \ y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + w_0) \geq 1, \quad i = 1...n \end{cases}$

Se halla el problema dual asociado a maximizar

la función Lagrangiana siguiente

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j) = \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\boldsymbol{\alpha}^{\mathbf{T}}\mathbf{Q}\boldsymbol{\alpha}, \ \text{ con } Q_{ij} = y_i y_j \mathbf{x}_i^{\mathbf{T}}\mathbf{x}_j$$

$$s.a. \ \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \ i=1,...,n$$

Una vez resuelto el problema dual

Si la solución óptima es $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_n)$, el valor óptimo de $\mathbf{w}$ es

$$\hat{\mathbf{w}} = \sum_{i=1} \hat{\alpha}_i y_i \mathbf{x}_i$$

El valor óptimo de w_0 se obtiene como:

$$\hat{w}_0 = \pm 1 - \hat{\mathbf{w}}^{\mathbf{T}}\mathbf{x}_i, \quad \text{con } \mathbf{x_i} \in \omega_1 \ y \ \ \hat{\alpha}_i > 0$$

Vectores soporte

– En la solución óptima del problema dual los únicos valores no nulos $\hat{\boldsymbol{\alpha}}=(\hat{\alpha}_1,\hat{\alpha}_2,...,\hat{\alpha}_n)$ son los correspondientes a las muestras sobre el margen. Estas muestras se llaman vectores soporte.

– En la práctica son pocos los elementos que están sobre el margen. Por tanto la mayoría de los α_i son nulos.

– Llamaremos *Sop* a los índices correspondientes a los vectores soporte. Entonces:

MVS:(CL).Resolución Práctica (2)

Así, una vez resuelto el problema dual, el valor óptimo de w es:

$$\hat{\mathbf{w}} = \sum_{i=1} \hat{\alpha}_i y_i \mathbf{x_i} = \sum_{i \in Sop} \hat{\alpha}_i y_i \mathbf{x_i}$$

mientras que el valor óptimo del término independiente w_0 es:

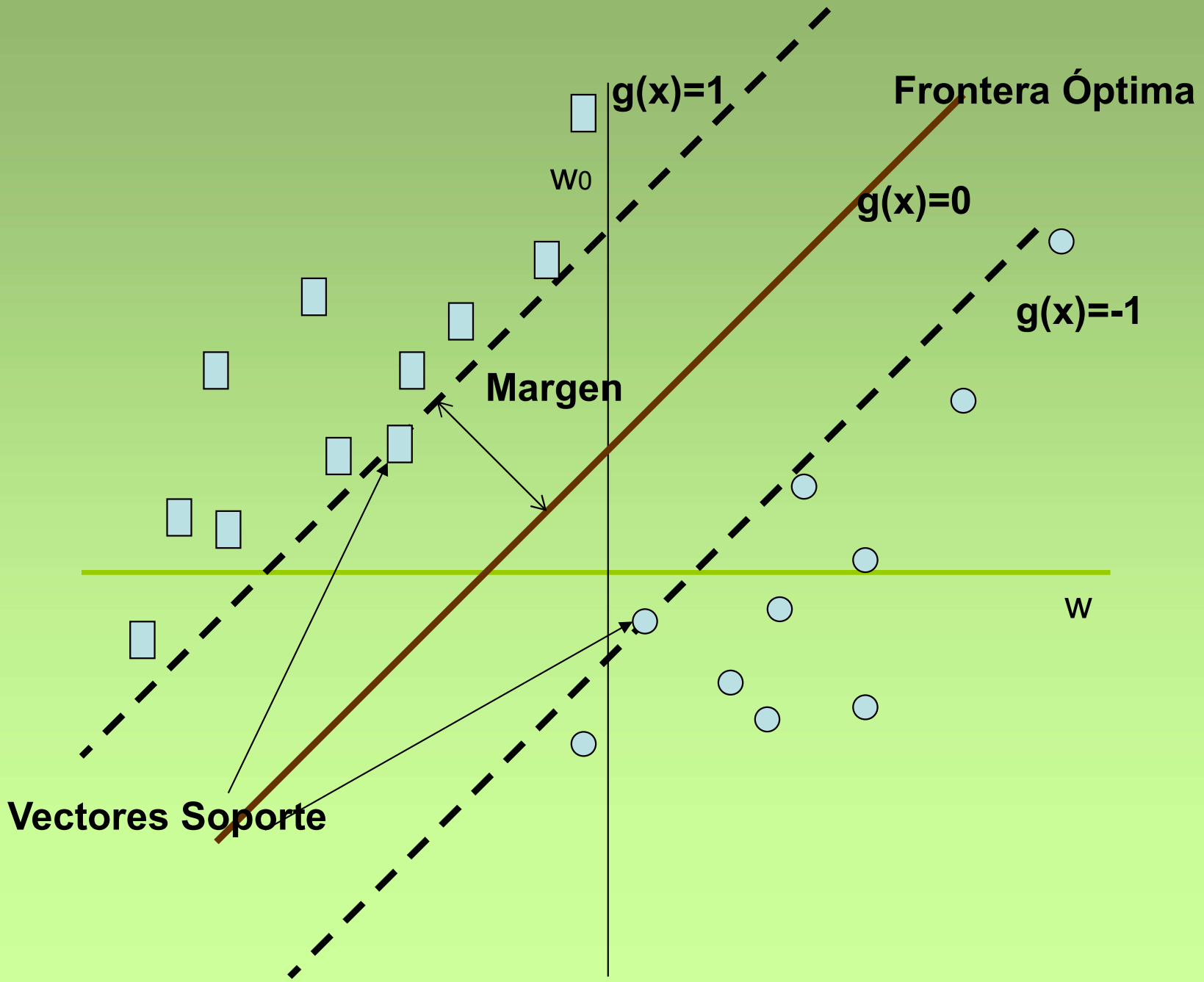
$$\hat{w}_0 = \pm 1 - \sum_{j \in Sop} \hat{\alpha}_j y_j (\mathbf{x_j^T x_i}), \quad \text{con } \mathbf{x_i} \in \omega_1 \; y \;\; \hat{\boldsymbol{\alpha}}_\mathbf{i} > 0$$

Por lo que la función discriminante lineal es:

$$g(\mathbf{x}_j) = \hat{\mathbf{w}}^{\mathbf{T}} \mathbf{x}_j + \hat{w}_0 = \left(\sum_{i \in Sop} \hat{\alpha}_i y_i \mathbf{x}_i \right)^T \mathbf{x}_j + \hat{w}_0 = \sum_{i \in Sop} \hat{\alpha}_i y_i (\mathbf{x_i^T x}_j) + \hat{w}_0$$

Vectores Soporte

Vectores Soporte: Representación



SVM: Resolución Práctica

– En la formulación del problema dual los elementos del conjunto de entrenamiento solo intervienen a través de sus productos escalares $(\mathbf{x}_i^T \mathbf{x}_j)$tanto en la función a optimizar

$$\max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

$$s.a. \sum_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \text{ i=1,...,n}$$

Como para calcular el valor de w_0

$$\hat{w}_0 = \pm 1 - \sum_{j \in Sop} \hat{\alpha}_j y_j (\mathbf{x}_j^T \mathbf{x}_i), \quad \text{con } \mathbf{x}_i \in \omega_1 \; y \;\; \hat{\boldsymbol{\alpha}}_i > 0$$

Siendo w el conjunto de los vectores soporte pertenecientes o a la clase positiva o a la clase negativa

Para calcular la clase a la que pertenece un nuevo vector x, calculamos la función $g(\mathbf{x}) = \sum_{i \in Sop} \hat{\alpha}_i y_i (\mathbf{x}_i^T \mathbf{x}) + \hat{w}_0$

si su valor es mayor que 0 pertenece a la clase positiva, en otro caso a la negativa

Un Ejemplo Inicial: Solución Dual con vectores soporte

Clasificación unidimensional: $g(\mathbf{x})=\mathbf{wx}+w_0$

– Conjunto de entrenamiento, OJO estamos en un espacio de dimensión 1

Los puntos (-1) (0) son de C1; los puntos (2) (5) son de C2

$$\left.\begin{aligned} &\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &s.a. \;\; y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+w_0) \geq 1, \quad i=1...n \end{aligned}\right\} \Rightarrow \begin{aligned} -w+w_0 &\geq 1 \\ w_0 &\geq 1 \\ -2w-w_0 &\geq 1 \\ -5w-w_0 &\geq 1 \end{aligned}$$

Primal

Vectores soporte (0) + y (2) -

$$\left.\begin{aligned} &\max_{\alpha \in Sop} \sum_{i=1} \alpha_i - \frac{1}{2}\sum_{i,j=1} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j) \\ &s.a. \;\; \sum_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \; i=1,...,n \end{aligned}\right\} \Rightarrow$$

Dual

$$L(\alpha_1,\alpha_2,\lambda) = \alpha_1+\alpha_2-\frac{1}{2}\left(0\alpha_1^2-0\alpha_1\alpha_2-0\alpha_2\alpha_1+4\alpha_2^2\right)-\lambda(\alpha_1-\alpha_2)$$

Un Ejemplo Inicial: Solución Dual con vectores soporte

Conjunto de entrenamiento: (-1), (0) de C1; (2), (5) de C2. Vectores soporte (0) para la clase positiva y (2) para la negativa. Así $x_1=0$, $x_2= 2$, $y_1=1$ e $y_2=-1$

$$\left.\begin{array}{l} \max\limits_{\alpha \in Sop} \sum\limits_{i=1} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j) \\ s.a. \ \sum\limits_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \ \text{i=1,...,n} \end{array}\right\} \Rightarrow$$

$$\max \alpha_1 + \alpha_2 - \frac{1}{2}\left(0\alpha_1^2 - 0\alpha_1\alpha_2 - 0\alpha_2\alpha_1 + 4\alpha_2^2\right)$$

$$s.a. \quad \alpha_1 - \alpha_2 = 0, \ \alpha_1 \geq 0, \ \alpha_2 \geq 0$$

$$L = \alpha_1 + \alpha_2 - 2(\alpha_2^2) - \lambda(\alpha_1 - \alpha_2)$$

Un ejemplo inicial: Solución Dual con vectores soporte

Derivando con respecto a los α_i y λ tenemos

$$1)\ \frac{\partial L}{\partial \alpha_1} = 1-\lambda = 0 \qquad 2)\ \frac{\partial L}{\partial \alpha_2} = 1-4\alpha_2+\lambda = 0 \qquad 3)\ \frac{\partial L}{\partial \lambda} = -\alpha_1+\alpha_2 = 0$$

$$Solución \equiv\ \lambda = 1,\ \alpha_1 = 1/2,\ \alpha_2 = 1/2,$$

- **Solución óptima:**

$$\hat{\boldsymbol{\alpha}} = (1/2,\ 1/2)$$

- **Vector óptimo:**

$$\hat{\mathrm{w}} = \sum_{i \in Sop} \hat{\alpha}_i y_i \mathrm{x}_\mathrm{i} = (1/2)\times 1\times 0 + (1/2)\times(-1)\times 2 = -1,$$

- **Constante óptima:** $\hat{\mathrm{w}}_0 = 1\text{-}\hat{w}x_1 = 1-(-1)\times 0 = 1$

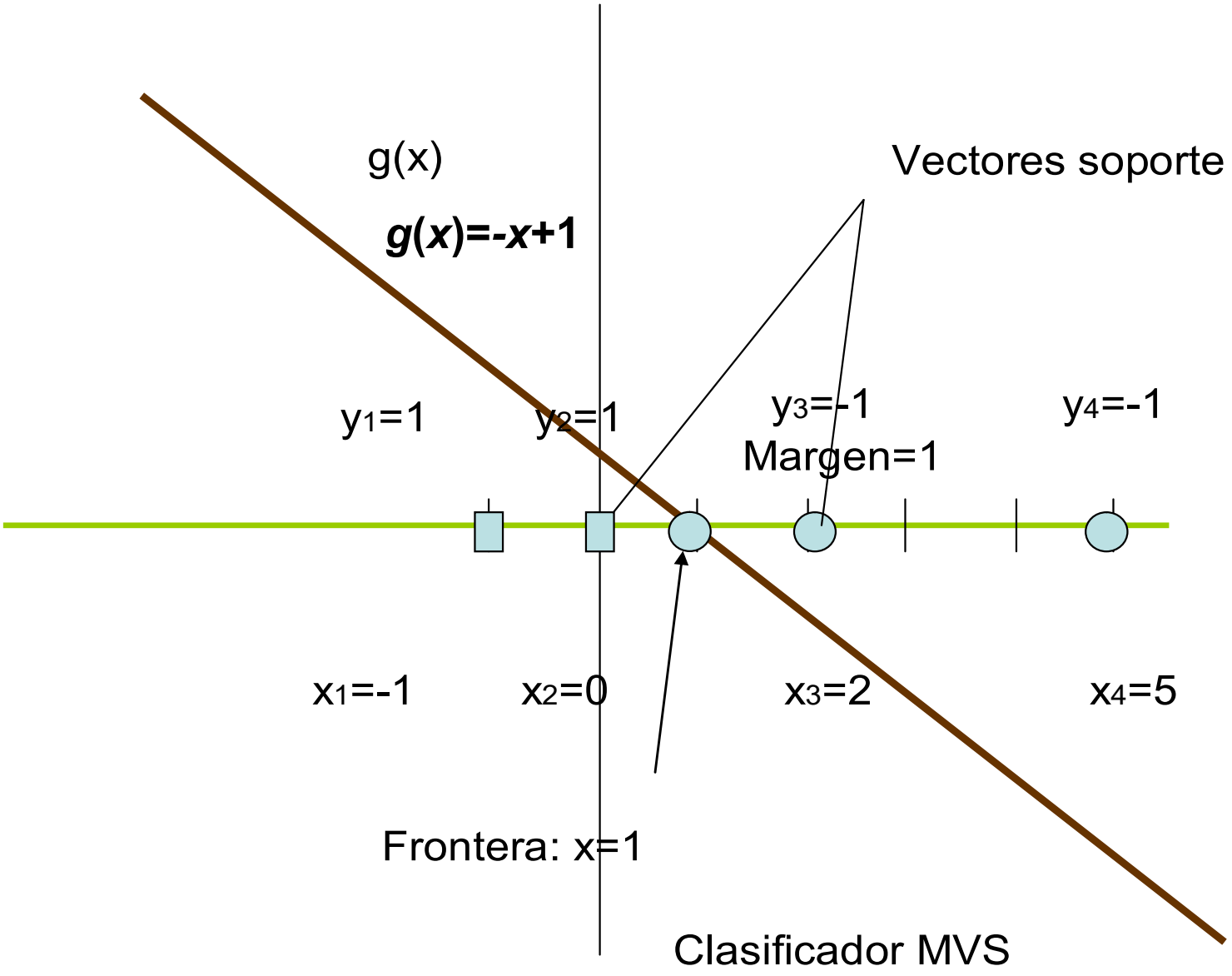
O también

$$\hat{\mathrm{w}}_0 = -1\text{-}\hat{w}x_2 = -1-(-1)\times 2 = 1$$

- **Clasificador:**

– **Por tanto el clasificador SVM es *g*(*x*)=-*x*+1 y la frontera *g*(*x*)=0 es *x*=1**

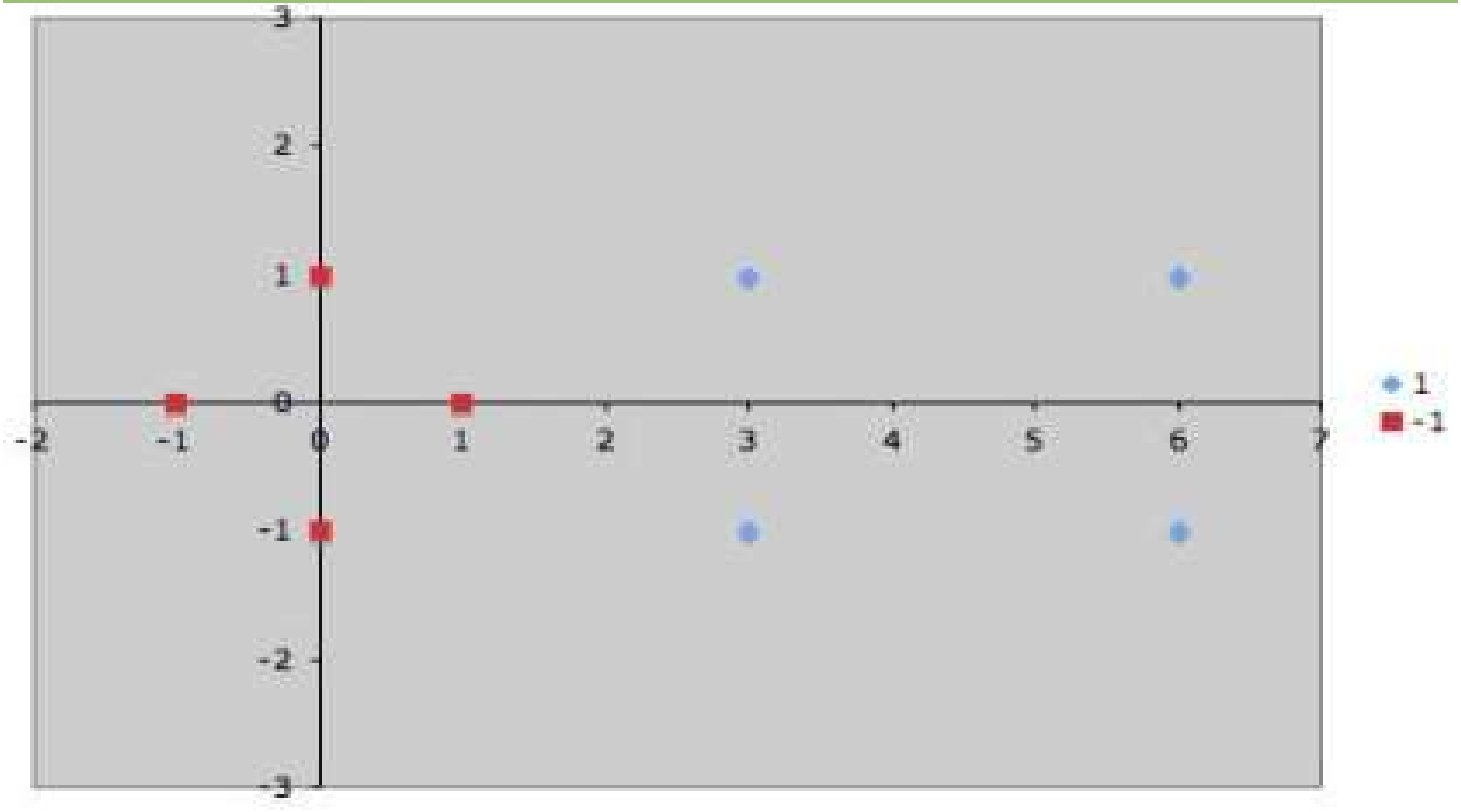
Un Ejemplo Inicial: Solución Dual con vectores soporte



Nuevo ejemplo linealmente separable

Clase negativa (-1, 0, -1) (0, -1, -1) (0, 1, -1) (1, 0, -1)

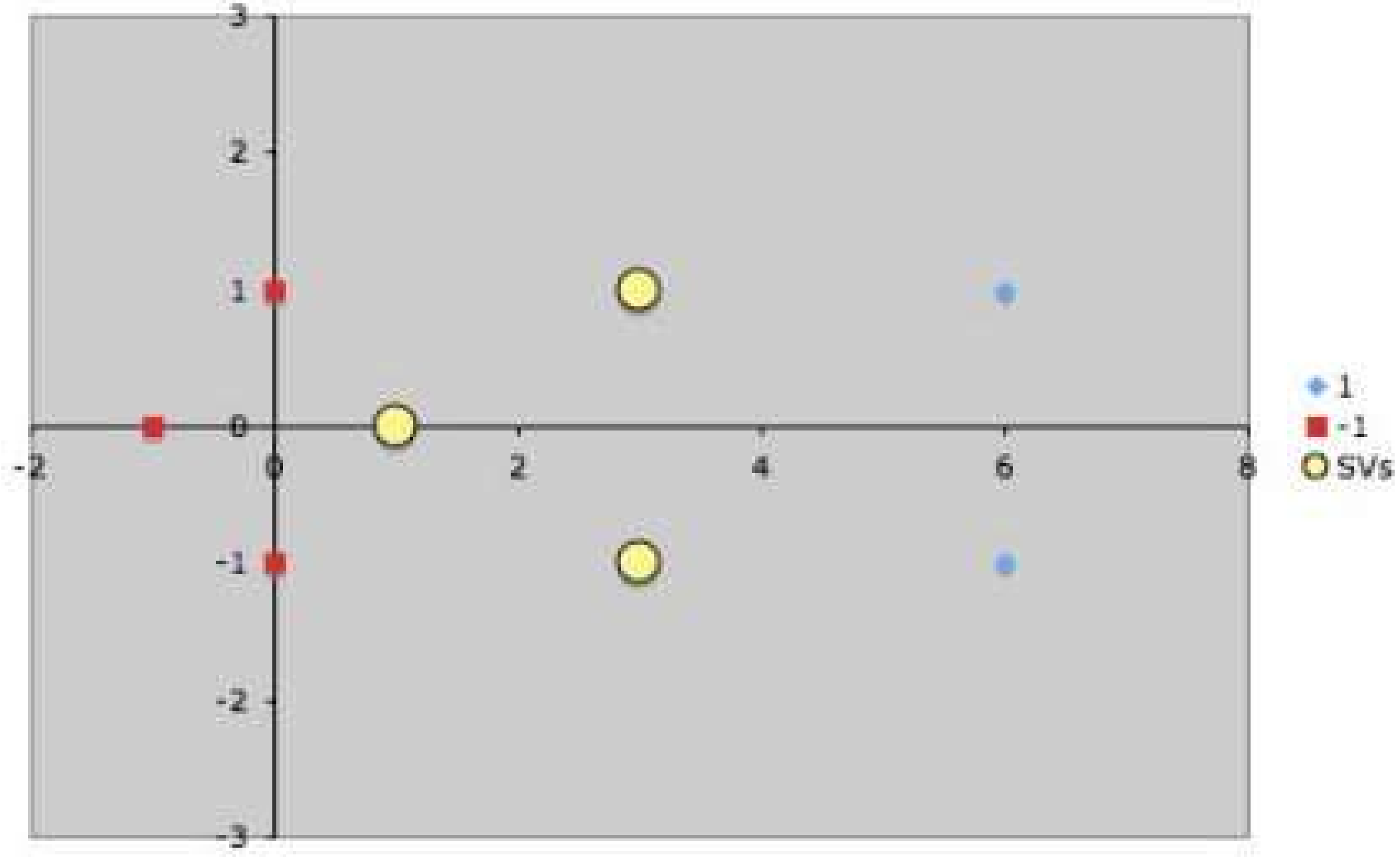
Clase positiva (3, -1, +1) (3, 1, +1) (6, -1, +1) (6, 1, +1)



Ejemplo: Vectores soporte

Clase negativa (-1, 0, -1) (0, -1, -1) (0, 1, -1) (1, 0, -1)

Clase positiva (3, -1, +1) (3, 1, +1) (6, -1, +1) (6, 1, +1)



Ejemplo

Clase positiva (3, -1, +1) (3, 1, +1) (6, -1, +1) (6, 1, +1)

Clase negativa (-1, 0, -1) (0, -1, -1) (0, 1, -1) (1, 0, -1)

H_1: $\mathbf{w} \cdot \mathbf{x} + w_0 = 1$

H_2: $\mathbf{w} \cdot \mathbf{x} + w_0 = -1$

$w_1x_1 + w_2x_2 + w_0 = -1$

$1w_1 + 0w_2 + w_0 = -1$ **(1)**

$\rightarrow w_0 = -1 - w_1$

$w_1x_1 + w_2x_2 + w_0 = 1$

$3w_1 - 1w_2 + w_0 = 1$ **(2)**

$\rightarrow w_2 = 3w_1 - 1 - w_1 - 1$

$\rightarrow w_2 = 2w_1 - 2$

$3w_1 + 1w_2 + w_0 = 1$ **(3)**

$\rightarrow 3w_1 + 2w_1 - 2 - 1 - w_1 = 1$

$\rightarrow w_1 = 1$

$\rightarrow w_0 = -2$

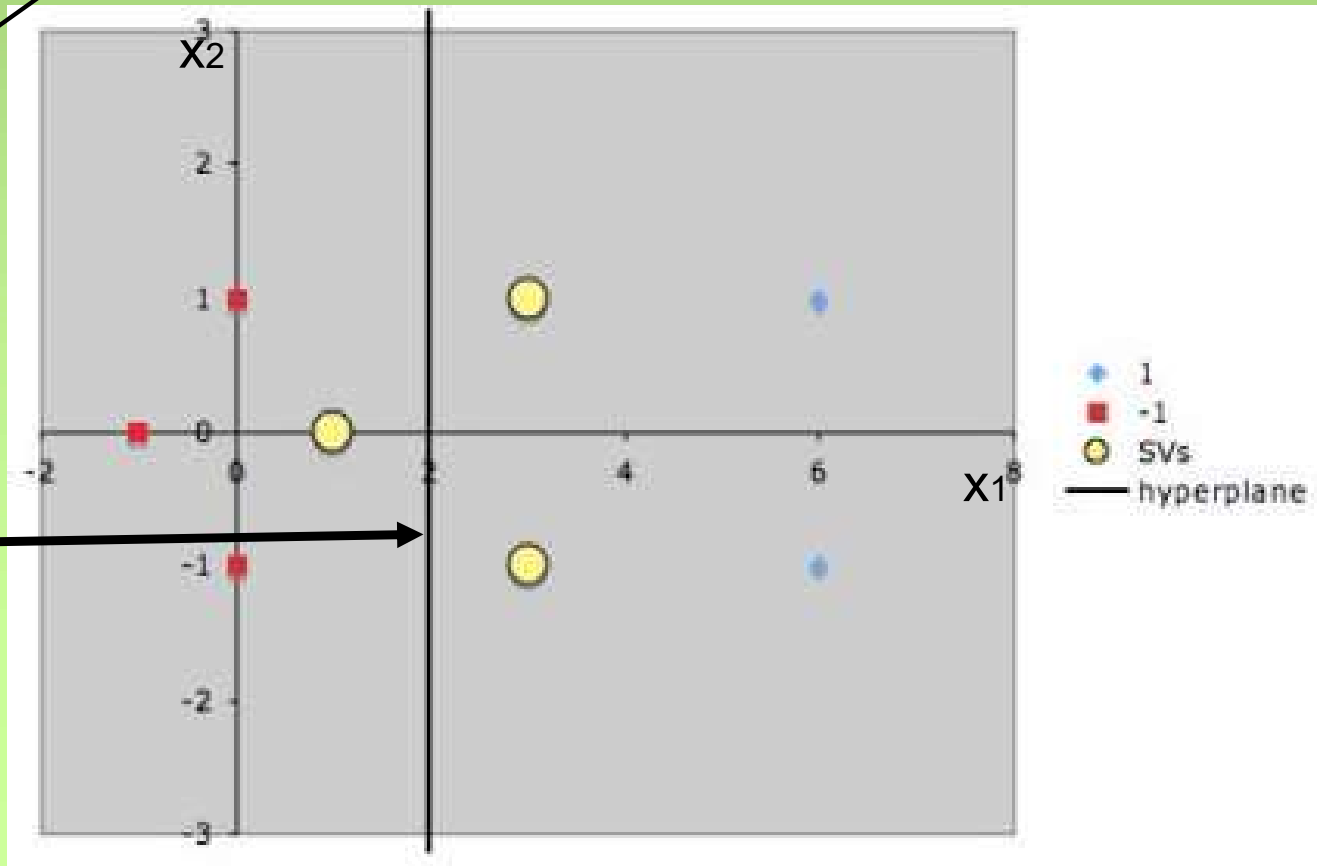
$\rightarrow w_2 = 0$

(1, 0) → -1

(3, -1) → +1

(3, 1) → +1

Vectores soporte



Función discriminante

$$(1,0)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 2 = 0$$

Recta: $x_1 = 2$

Un ejemplo inicial (4): Solución Dual

Vectores soporte

$$\left.\begin{aligned}&\max_{\alpha\in Sop}\sum_{i=1}\alpha_i-\frac{1}{2}\sum_{i,j=1}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i^T\mathbf{x}_j)\\&s.a.\ \sum_{i=1}\alpha_i y_i=0,\ \ \alpha_i\geq 0,\ \text{i=1,...,n}\end{aligned}\right\}\Rightarrow$$

$\mathbf{x_1}=(x_{11},x_{12})=(1,\ 0)\rightarrow y_1=-1$
$\mathbf{x_2}=(x_{21},x_{22})=(3,\ -1)\rightarrow y_2=+1$
$\mathbf{x_3}=(x_{31},x_{32})=(3,\ 1)\rightarrow y_3=+1$

$$\max \alpha_1+\alpha_2+\alpha_3-\frac{1}{2}\left(\begin{aligned}&\alpha_1^2(x_1^T.x_1)-\alpha_1\alpha_2(x_1^T.x_2)-\alpha_1\alpha_3(x_1^T.x_3)-\alpha_2\alpha_1(x_2^T.x_1)+\alpha_2^2(x_2^T.x_2)+\alpha_2\alpha_3(x_2^T.x_3)+\\&-\alpha_3\alpha_1(x_3^T.x_1)+\alpha_3\alpha_2(x_3^T.x_2)+\alpha_3^2(x_3^T.x_3)\end{aligned}\right)$$

$$s.a.\quad -\alpha_1+\alpha_2+\alpha_3=0,\ \alpha_1\geq 0,\ \alpha_2\geq 0,\ \alpha_3\geq 0$$

Operando y simplificando utilizando los multiplicadores de Lagrange para la restricción de igualdad, dado que al ser maximización las restricciones de que los α_i sean positivos se cumplen

por lo que ahora la función a optimizar (maximizar) es

$$F(\alpha_1,\alpha_2,\alpha_3,\lambda;\ \mathbf{x_1},\mathbf{x_2},\mathbf{x_3})=$$

$$=\alpha_1+\alpha_2+\alpha_3-\frac{1}{2}\left(\begin{aligned}&\alpha_1^2(\mathbf{x_1^T}.\mathbf{x_1})-\alpha_1\alpha_2(\mathbf{x_1^T}.\mathbf{x_2})-\alpha_1\alpha_3(\mathbf{x_1^T}.\mathbf{x_3})-\alpha_2\alpha_1(\mathbf{x_2^T}.\mathbf{x_1})+\alpha_2^2(\mathbf{x_2^T}.\mathbf{x_2})+\alpha_2\alpha_3(\mathbf{x_2^T}.\mathbf{x_3})+\\&-\alpha_3\alpha_1(\mathbf{x_3^T}.\mathbf{x_1})+\alpha_3\alpha_2(\mathbf{x_3^T}.\mathbf{x_2})+\alpha_3^2(\mathbf{x_3^T}.\mathbf{x_3})\end{aligned}\right)$$

$$-\lambda(-\alpha_1+\alpha_2+\alpha_3)$$

Un ejemplo inicial (4): Solución Dual

Derivando con respecto a los α_i y λ tenemos

$$\frac{\partial F}{\partial \alpha_1} = 1 - \frac{1}{2}\left(2\alpha_1(\mathbf{x_1^T}.\mathbf{x_1}) - \alpha_2(\mathbf{x_1^T}.\mathbf{x_2}) - \alpha_3(\mathbf{x_1^T}.\mathbf{x_3}) - \alpha_2(\mathbf{x_2^T}.\mathbf{x_1}) - \alpha_3(\mathbf{x_3^T}.\mathbf{x_1})\right) - \lambda = 0$$

esto es

$$1 - \alpha_1 + 3\alpha_2 + 3\alpha_3 - \lambda = 0$$

$$\frac{\partial F}{\partial \alpha_2} = 1 - \frac{1}{2}\left(-\alpha_1(\mathbf{x_1^T}.\mathbf{x_2}) - \alpha_1(\mathbf{x_2^T}.\mathbf{x_1}) + 2\alpha_2(\mathbf{x_2^T}.\mathbf{x_2}) + \alpha_3(\mathbf{x_2^T}.\mathbf{x_3}) + \alpha_3(\mathbf{x_3^T}.\mathbf{x_2})\right) + \lambda = 0$$

esto es

$$1 + 3\alpha_1 - 10\alpha_2 - 8\alpha_3 + \lambda = 0$$

$$\frac{\partial F}{\partial \alpha_3} = 1 - \frac{1}{2}\left(-\alpha_1(\mathbf{x_1^T}.\mathbf{x_3}) + \alpha_2(\mathbf{x_2^T}.\mathbf{x_3}) - \alpha_1(\mathbf{x_3^T}.\mathbf{x_1}) + \alpha_2(\mathbf{x_3^T}.\mathbf{x_2}) + 2\alpha_3(\mathbf{x_3^T}.\mathbf{x_3})\right) - \lambda = 0$$

esto es

$$1 + 3\alpha_1 - 8\alpha_2 - 10\alpha_3 + \lambda = 0$$

$$\frac{\partial F}{\partial \lambda} = \alpha_1 - \alpha_2 - \alpha_3 = 0$$

Un ejemplo inicial (4): Solución Dual

De la última ecuación tenemos $\alpha_1 = \alpha_2 + \alpha_3$ **y sustituyendo en las otras tres ecuaciones, planteamos un sistema de ecuaciones en** α_2, α_3 y λ

$$\begin{cases} 1+2\alpha_2+2\alpha_3+\lambda=0 \\ 1-7\alpha_2-5\alpha_3-\lambda=0 \\ 1-5\alpha_2-7\alpha_3-\lambda=0 \end{cases}$$

Cuya solución es $\alpha_1 = 1/2\ ;\ \alpha_2 = 1/4;\ \alpha_3 = 1/4$

Vector óptimo:

$$\hat{\mathbf{w}} = \sum_{i \in Sop} \hat{\alpha}_i y_i \mathbf{x_i} = \frac{1}{2}(-1)\begin{pmatrix}1\\0\end{pmatrix} + \frac{1}{4}(1)\begin{pmatrix}3\\-1\end{pmatrix} + \frac{1}{4}(1)\begin{pmatrix}3\\1\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}$$

Constante óptima:

$$\hat{w}_0 = 1 - \sum_{j \in Sop} \hat{\alpha}_j y_j (\mathbf{x_j^T x_i}), \quad \text{con } \mathbf{x_i} \in \omega_1 \ y \ \ \hat{\alpha}_i > 0$$

siendo ω_1 el conjunto de los vectores soporte pertenecientes a la clase positiva

$$\hat{w}_0 = -1 - (\mathbf{x}_i^T \hat{\mathbf{w}}) = -1 - (1,0)\begin{pmatrix}1\\0\end{pmatrix} = -2, \text{ o también}$$

$$\hat{w}_0 = 1 - (\mathbf{x}_i^T \hat{\mathbf{w}}) = 1 - (3,1)\begin{pmatrix}1\\0\end{pmatrix} = -2, \text{ o también}$$

$$\hat{w}_0 = 1 - (\mathbf{x}_i^T \hat{\mathbf{w}}) = 1 - (3,-1)\begin{pmatrix}1\\0\end{pmatrix} = -2,$$

Vectores soporte
(1, 0) → -1
(3, -1) → +1
(3, 1) → +1

Función discriminante

H: (1, 0) · x - 2 = 0; H: x_1 - 2 = 0

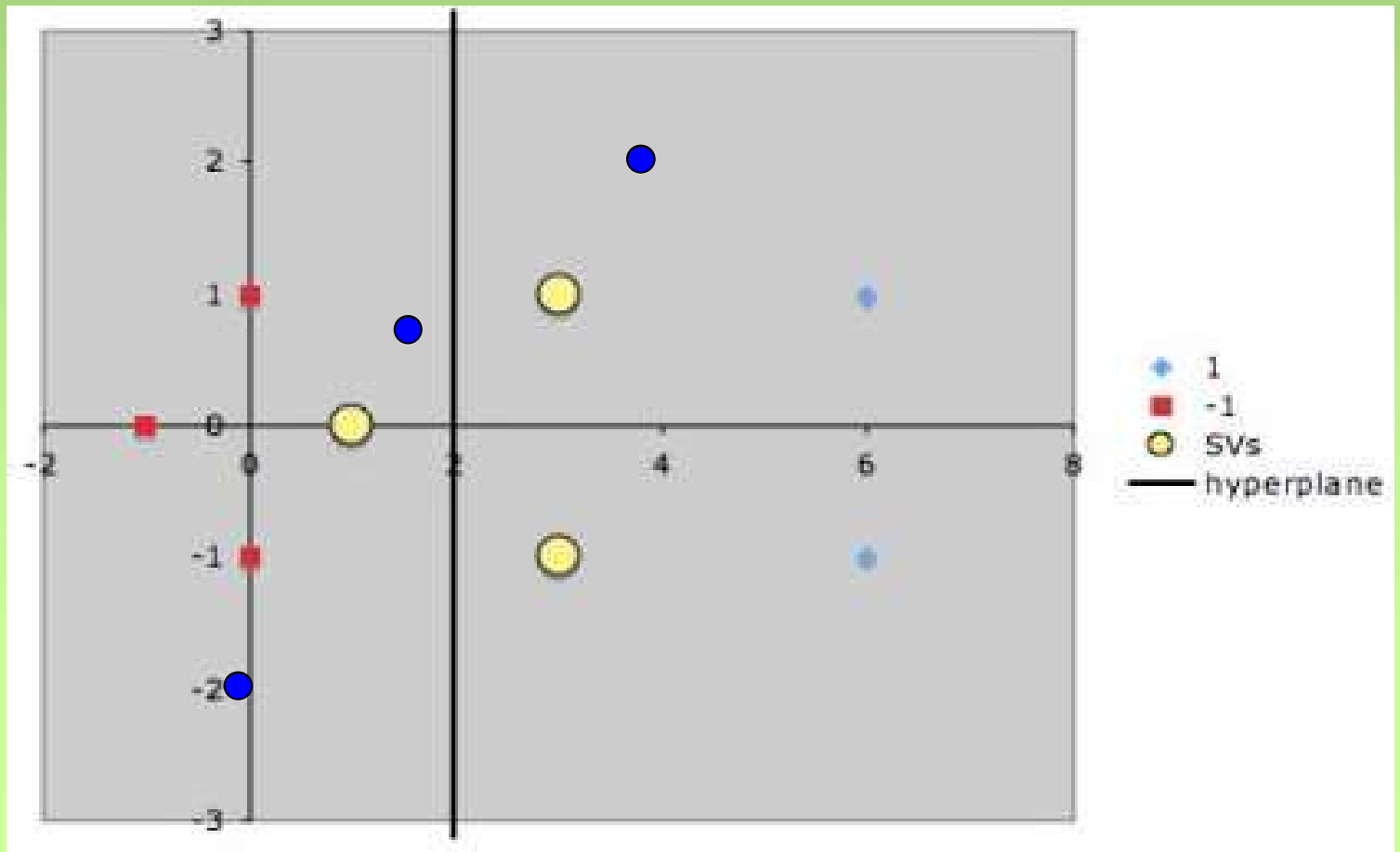
Datos del conjunto de test

(4, 2, 1), (1.5, 0.5, -1), (0, -2, -1)

4-2 = 2 se asigna a la clase [+1], exito

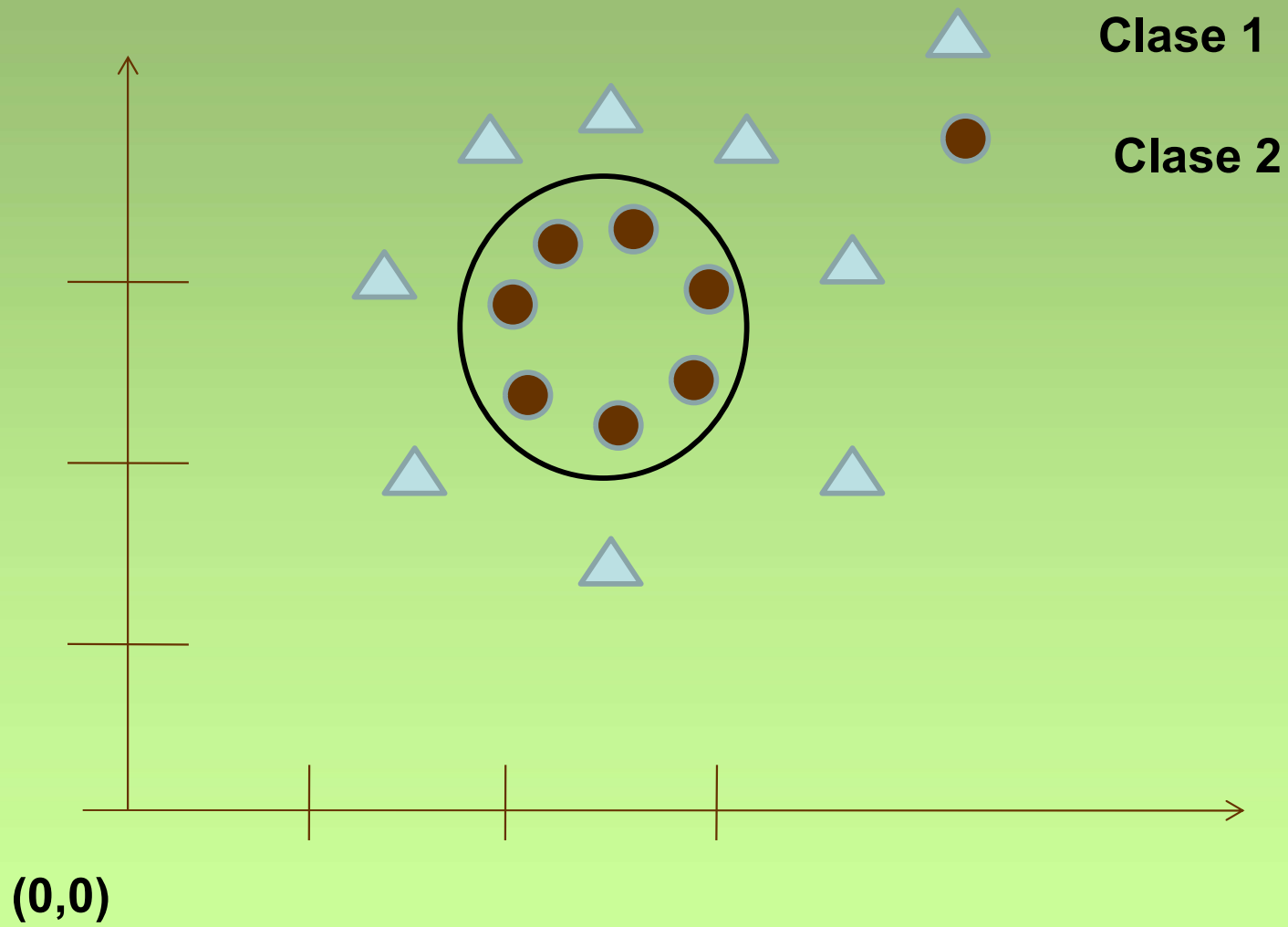
1.5-2 = -0.5 se asigna a la clase [-1], exito

0-2 = -2 se asigna a la clase [-1], exito



Como funciona para datos linealmente no separables?

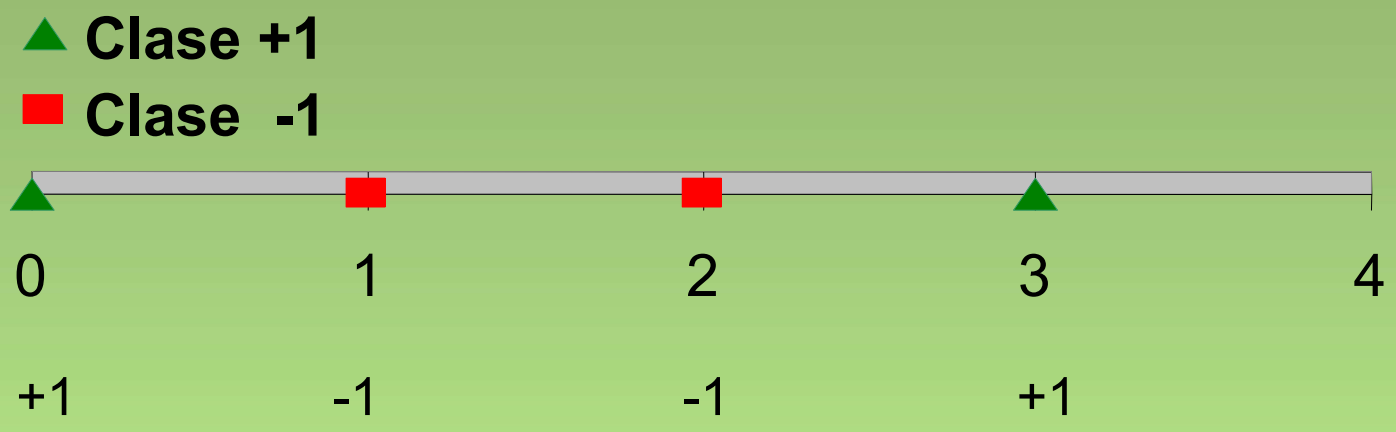
- Transformación de un espacio de variables de entrada de dimensión D a un espacio de características de dimensión HD siendo HD > D
- Los vectores de entrada se transforman de forma no lineal
- Después de la transformación el nuevo espacio de características debe de ser linealmente separable


Clase 1
Clase 2
(0,0)

Ejemplo

Como separar las dos clases mediante un punto

X_1	Clase
0	+1
1	-1
2	-1
3	+1



SVM utiliza una función no lineal sobre los atributos del espacio inicial de variables

$$\Phi(x_1) = (x_1, x_1^2)$$

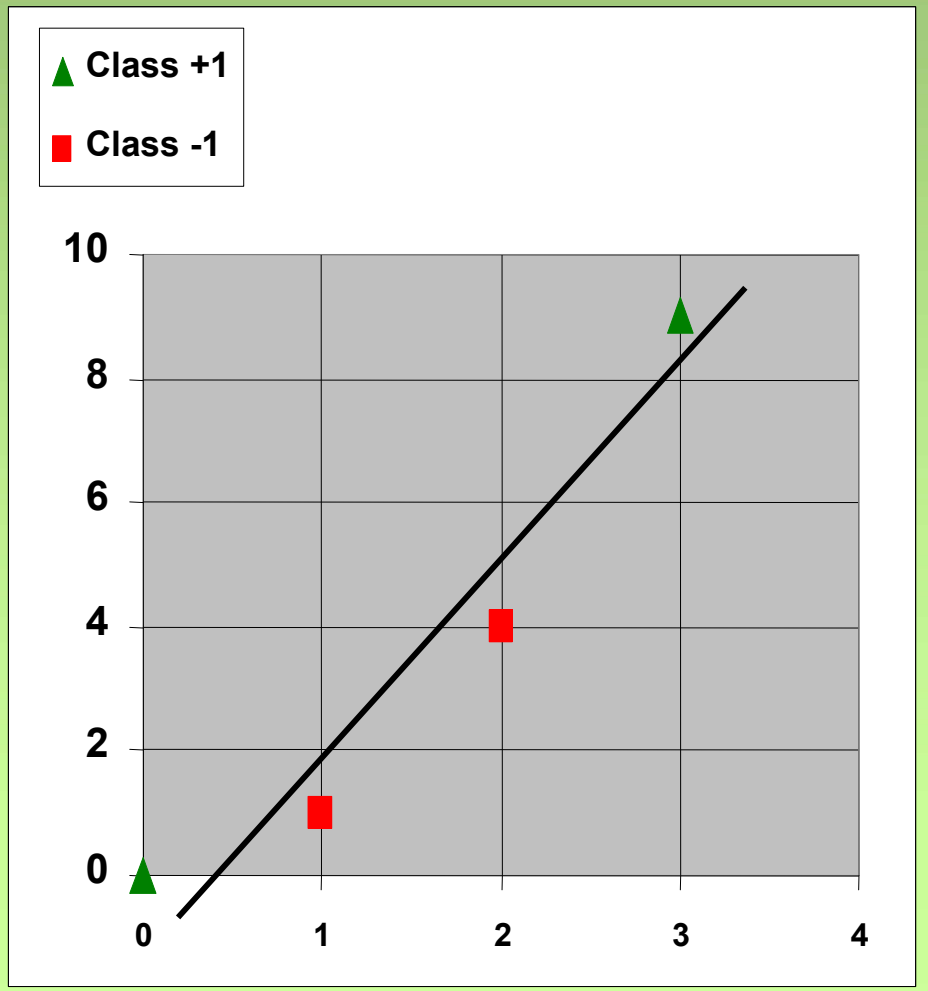
Esta función pasa de un espacio unidimensional a otro bidimensional

Ejemplo

$$\Phi(x_1) = (x_1, x_1^2)$$

Esta función pasa a un problema bidimensional, donde los datos son separables

X_1	X_1^2	Class
0	0	+1
1	1	-1
2	4	-1
3	9	+1



Ejemplo

Los cuatro son vectores soporte

x_1	x_1^2	Clase
0	0	+1
1	1	-1
2	4	-1
3	9	+1

- **w·x** + w_0 = +1 (para la clase positiva)

  $w_1x_1 + w_2x_2 + w_0 = +1$

  1) $0w_1 + 0w_2 + w_0 = +1$ → **$w_0 = 1$**

  2) $3w_1 + 9w_2 + w_0 = +1$

- **w·x** + w_0 = -1 (para la clase negativa)

  $w_1x_1 + w_2x_2 + w_0 = -1$

  3) $1w_1 + 1w_2 + w_0 = -1$

  4) $2w_1 + 4w_2 + w_0 = -1$

sustituyendo w_0 en las dos ecuaciones

→ $w_1 = -2 - w_2$

→ $-4 - 2w_2 + 4w_2 + 1 = -1$

de donde w_2 =1 y w_1 = -3, valores que verifican la ecuación 2)

- **Función discriminante**

  wT.x + w_0 = 0

  $w_1x_1 + w_2x_2 + w_0 = 0$ → **$-3x_1 + x_2 + 1 = 0$, o lo que es lo mismo $x_2 = 3x_1 - 1$**

Ejemplo

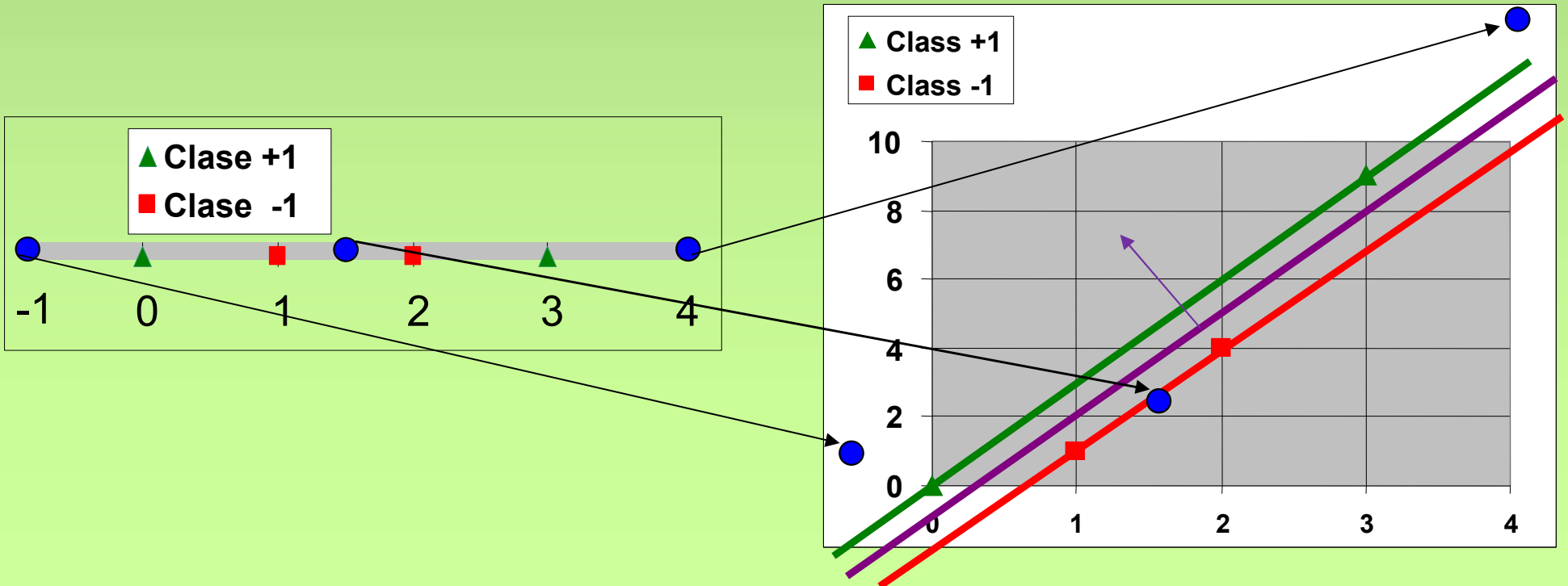
H: $-3x_1 + x_2 + 1 = 0$

Datos de test (1,5), (-1), (4)

(1,5) se transforma en (1,5, 2,25) como -3(1,5) + 2,25 + 1 = - 1,15 pertenece a la clase [-1]

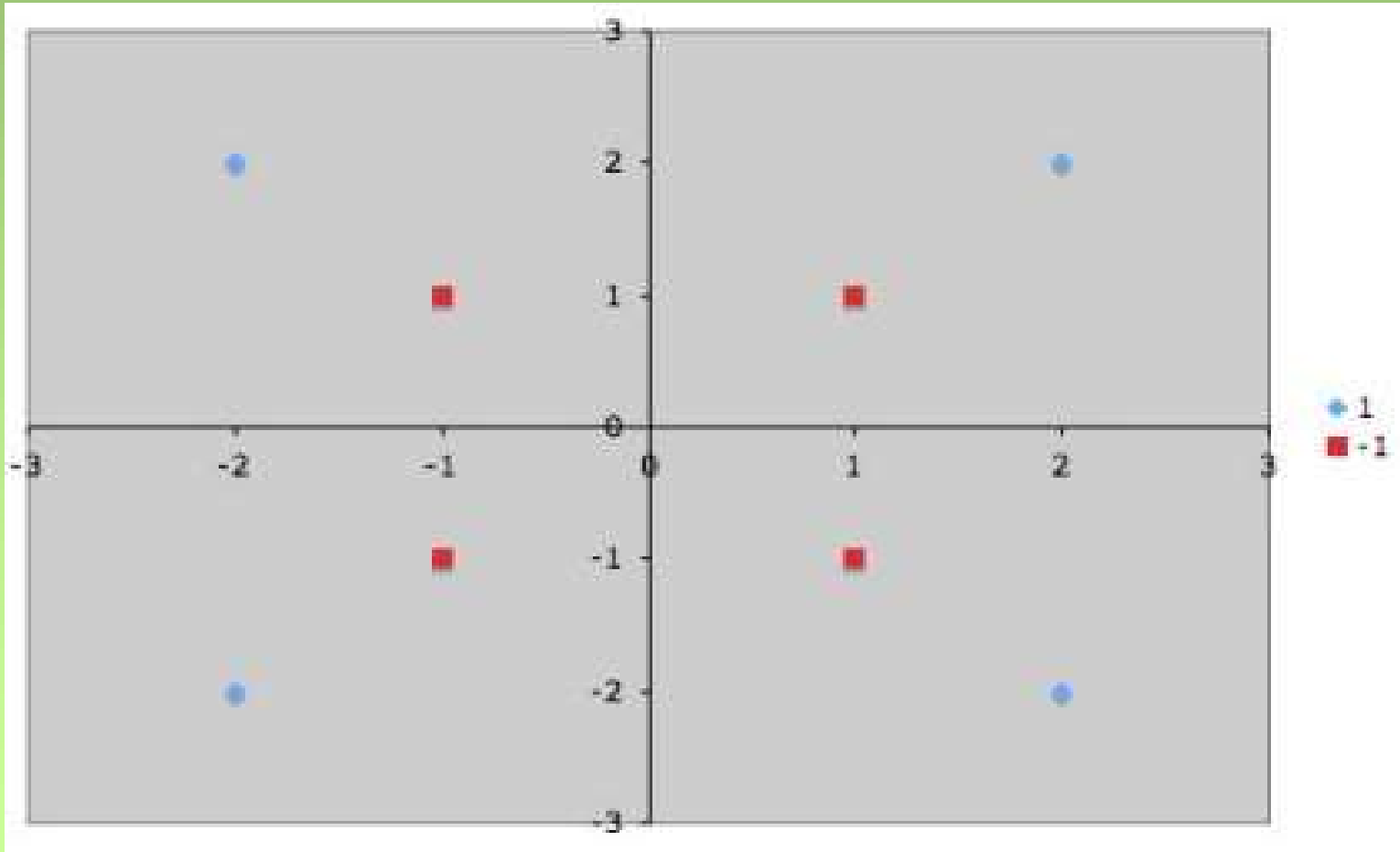
(-1) se transforma en (-1,1); como -3(-1) + 1 + 1 = 5 pertenece a la clase [+1]

(4) se transforma en (4,16); como -3(4) + 16 + 1 = 5 pertenece a la clase [+1]



Como separar estas dos clases mediante una función lineal?

x_1	x_2	Clase
1	1	-1
-1	1	-1
1	-1	-1
-1	-1	-1
2	2	+1
-2	2	+1
2	-2	+1
-2	-2	+1



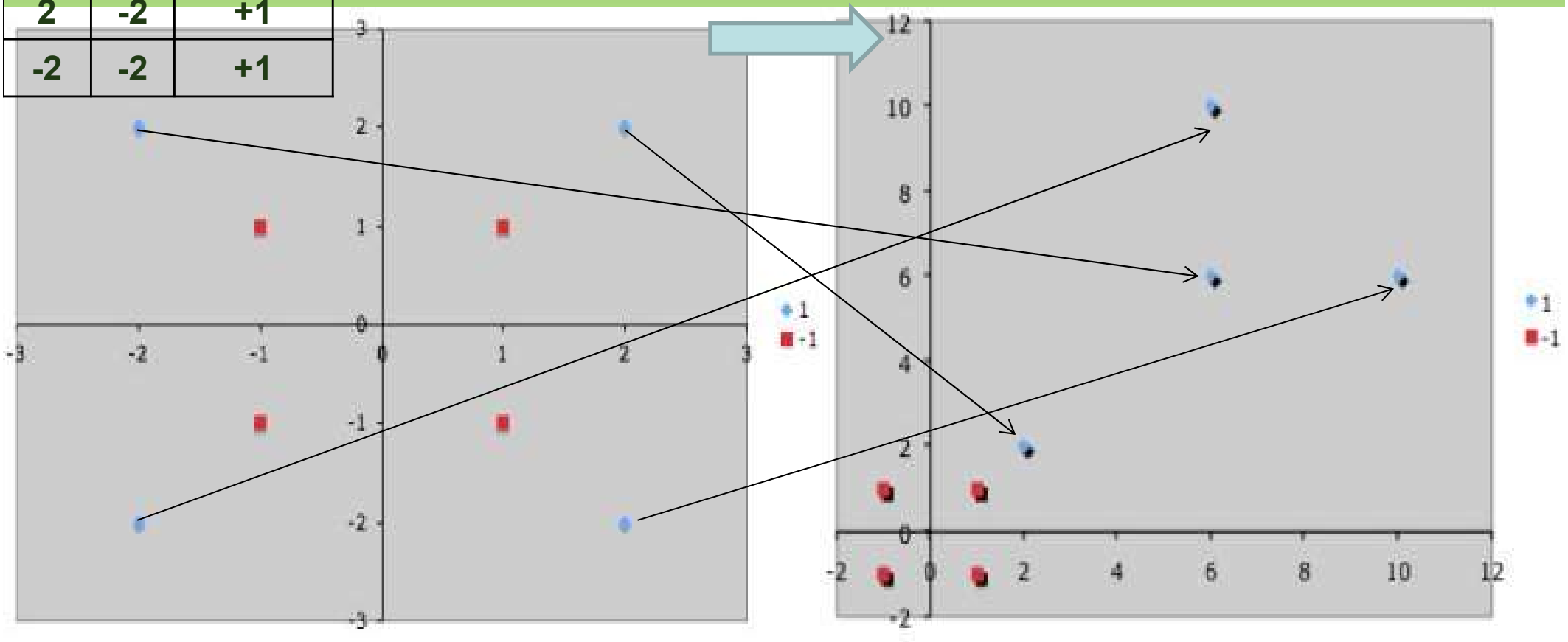
Nuevo Ejemplo

x_1	x_2	Clase
1	1	-1
-1	1	-1
1	-1	-1
-1	-1	-1
2	2	+1
-2	2	+1
2	-2	+1
-2	-2	+1

Esta función mantiene la dimensión bidimensional del espacio de entrada

$$\Phi(x_1,x_2)=\begin{cases}(4-x_2+|x_1-x_2|,4-x_1+|x_1-x_2|) & \text{si } \sqrt{x_1^2+x_2^2}>2\\ (x_1,x_2) & \text{en otro caso}\end{cases}$$

El (2,2) Pasa al (2,2). El (-2,2) al (6,10). El (2,-2) al (10,6). El (-2,-2) al (6,6)



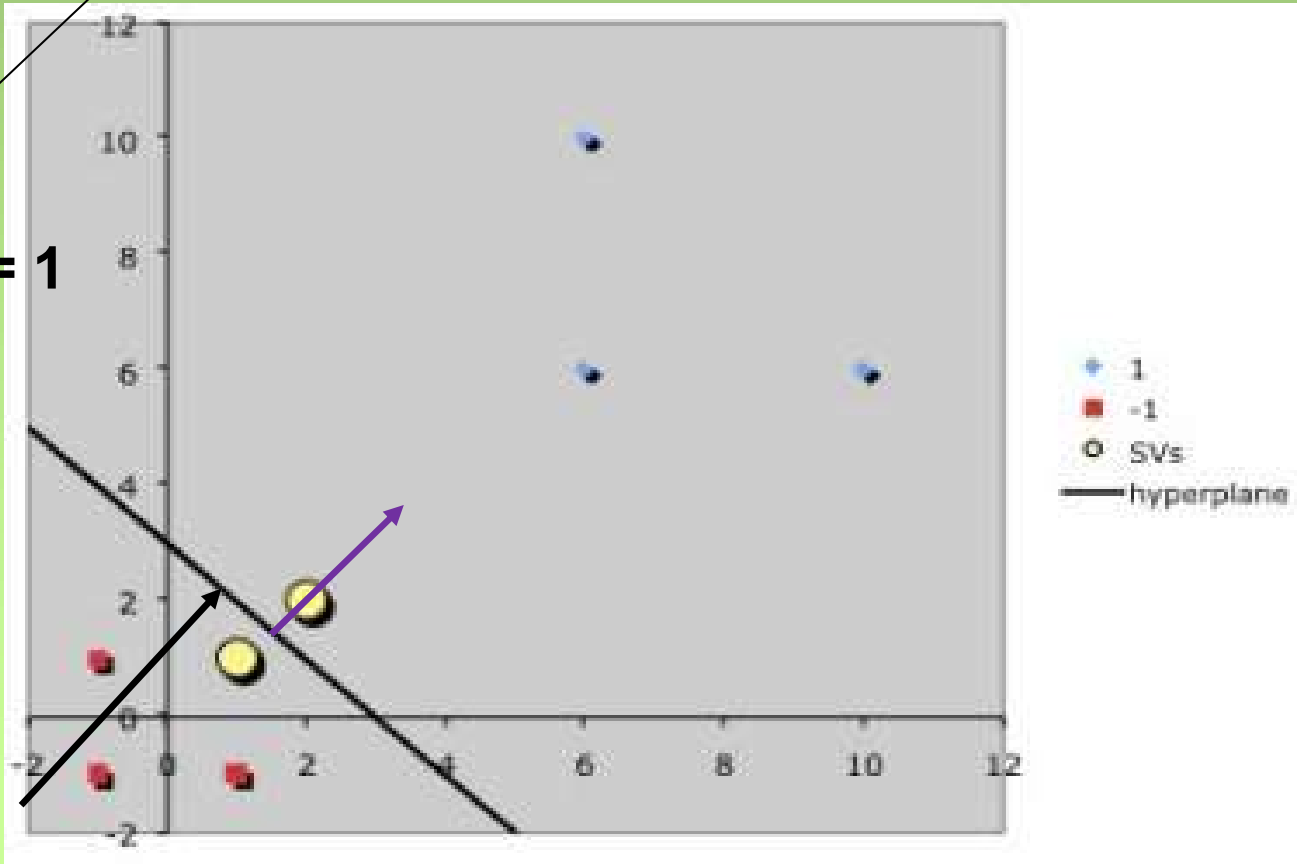
Vectores Soporte en el espacio de características

H_0: **w.x** + w_0 = 1
H_1: **w.x** + w_0 = -1

x_1	x_2	Clase
1	1	-1
2	2	+1

$w_1x_1 + w_2x_2 + w_0 = -1$
1) $1w_1 + 1w_2 + w_0 = -1$
→ $w_1 = -1 - w_0 - w_2$

$w_1x_1 + w_2x_2 + w_0 = 1$
2) $2(-1 - w_0 - w_2) + 2w_2 + w_0 = 1$
$-2 - 2w_0 - 2w_2 + 2w_2 + w_0 = 1$
→ $w_0 = -3$
Solución $w_1+w_2=2$

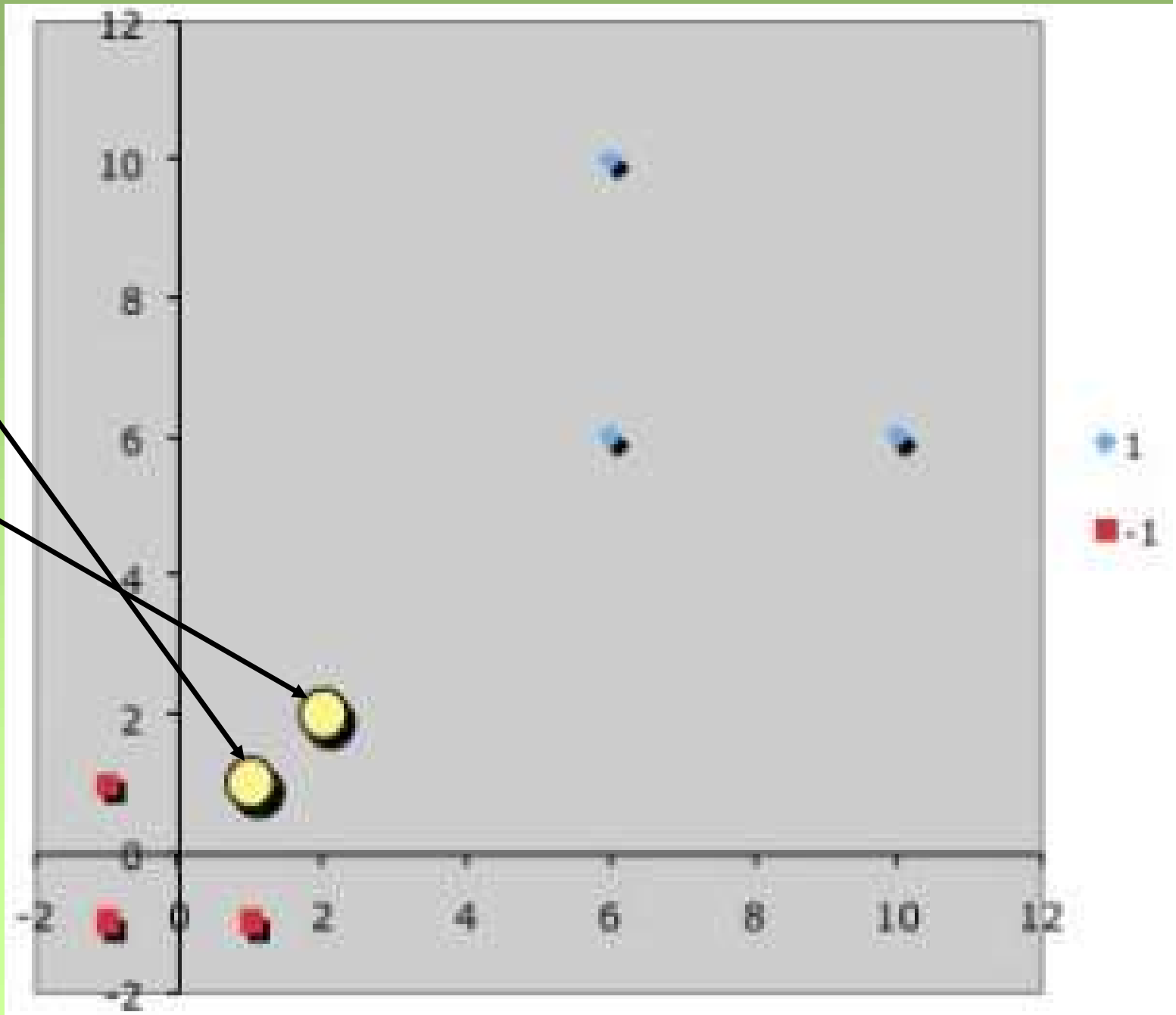


Una posible solución

$w_1 = w_2 = 1$
H: $(1,1)^T \cdot (x_1,x_2) - 3 = 0$
$x_1 + x_2 - 3 = 0$; **Ecuación de la recta, $x_2=-x_1+3$**

Vectores Soporte

x_1	x_2	Clase
1	**1**	**-1**
-1	1	-1
1	-1	-1
-1	-1	-1
2	**2**	**+1**
6	6	+1
10	6	+1
6	10	+1



Un ejemplo inicial (4): Solución Dual

$$\left.\begin{aligned} \max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2}\sum_{i,j=1} \alpha_i \alpha_j y_i y_j (\Phi(s_i)^T . \Phi(s_j)) \\ s.a.\ \sum_{i=1} \alpha_i y_i = 0,\ \ \alpha_i \geq 0,\ \text{i=1,...,n} \end{aligned}\right\} \Rightarrow \begin{aligned} \Phi(s_1)^T &= (1,1),\ \ y_1 = -1 \\ \Phi(s_2)^T &= (2,2),\ \ y_1 = 1 \end{aligned}$$

Vectores soporte en el espacio transformado

$$\max \alpha_1 + \alpha_2 - \frac{1}{2}\left(\alpha_1^2 (\Phi(s_1)^T . \Phi(s_1)) - \alpha_1\alpha_2 (\Phi(s_1)^T . \Phi(s_2)) - \alpha_2\alpha_1 (\Phi(s_2)^T . \Phi(s_1)) + \alpha_2^2 (\Phi(s_2)^T . \Phi(s_2))\right)$$

$$L(\alpha_1, \alpha_2, \lambda) = \alpha_1 + \alpha_2 - \frac{1}{2}\left(\alpha_1^2 (2) - \alpha_1\alpha_2 (4) - \alpha_2\alpha_1 (4) + \alpha_2^2 (8)\right) - \lambda(-\alpha_1 + \alpha_2)$$

$$\Phi\left((x_1, x_2)^T\right) = \left\{(2,2)^T,\ (10,6)^T,\ (6,6)^T,\ (6,10)^T\right\}\ \text{Clase +, } y_i = 1$$

$$\Phi\left((x_1, x_2)^T\right) = \left\{(1,1)^T,\ (1,-1)^T,\ (-1,-1)^T,\ (-1,1)^T\right\}\ \text{Clase -, } y_i = -1$$

Derivando con respecto a los α_i y λ tenemos

$$\begin{cases} 1 - 2\alpha_1 + 4\alpha_2 + \lambda = 0 \\ 1 + 4\alpha_1 - 8\alpha_2 - \lambda = 0 \\ \alpha_1 - \alpha_2 = 0 \end{cases} \qquad \lambda = -3;\ \alpha_1 = 1;\ \alpha_2 = 1$$

Nuevo Ejemplo: Resolución mediante el dual

Vector óptimo:

$$\hat{\mathbf{w}} = \sum_{i \in Sop} \hat{\alpha}_i y_i \Phi(s_i) = 1(-1)\begin{pmatrix}1\\1\end{pmatrix} + 1(1)\begin{pmatrix}2\\2\end{pmatrix} = \begin{pmatrix}1\\1\end{pmatrix} \qquad \begin{aligned}\Phi(s_1)^T &= (1,1), \;\; y_1 = -1\\ \Phi(s_2)^T &= (2,2), \;\; y_1 = 1\end{aligned}$$

Constante óptima:

$$\hat{w}_0 = \pm 1 - \sum_{j \in Sop} \hat{\alpha}_j y_j (\Phi(s_j)^T \Phi(s_i)), \quad \text{con } s_i \in \omega_1 \; y \;\; \hat{\alpha}_i > 0$$

$$\hat{w}_0 = -1 - (\Phi(s_1)^T \hat{\mathbf{w}}) = -1 - (1,1)\begin{pmatrix}1\\1\end{pmatrix} = -3, \text{ o también}$$

$$\hat{w}_0 = 1 - (\Phi(s_2)^T \hat{\mathbf{w}}) = 1 - (2,2)\begin{pmatrix}1\\1\end{pmatrix} = -3,$$

La función discriminante lineal es

H: (1,1)·(x1,x2)T -3 = 0, y la ecuación de la recta es

x1 + x2 -3 = 0 , x2=-x1+3

Capacidad de Generalización

Pasamos ahora a ver como podemos usar el modelo SVM para clasificar datos del conjunto de generalización, dado un patrón x, la clasificación f(x) se obtiene mediante la ecuación

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{2} \alpha_i y_i (\Phi(\mathbf{s_i}).\Phi(\mathbf{x})) - 3\right), \text{ donde sign(z) devuelve el signo de z, (1)}$$

Por ejemplo si queremos clasificar el patrón de características (4, 5, +1) usando la función de transformación dada por la ecuación (1), tenemos

$$f((4,5)^T) = sign(1\times(-1)\times(\Phi(1,1).\Phi(4,5)^T)) + 1\times 1\times(\Phi(2,2).\Phi(4,5)^T) - 3) =$$

$$= sign(-(1,1).(0,1)^T + (2,2).(0,1)^T - 3) = sign(-2),$$

donde sign devuelve el signo de -2

Y entonces deberíamos clasificar al patrón (4,5) como negativo. Si observamos el espacio de características de entrada vemos que es una clasificación errónea, pues el (4,5) tiene etiqueta +1

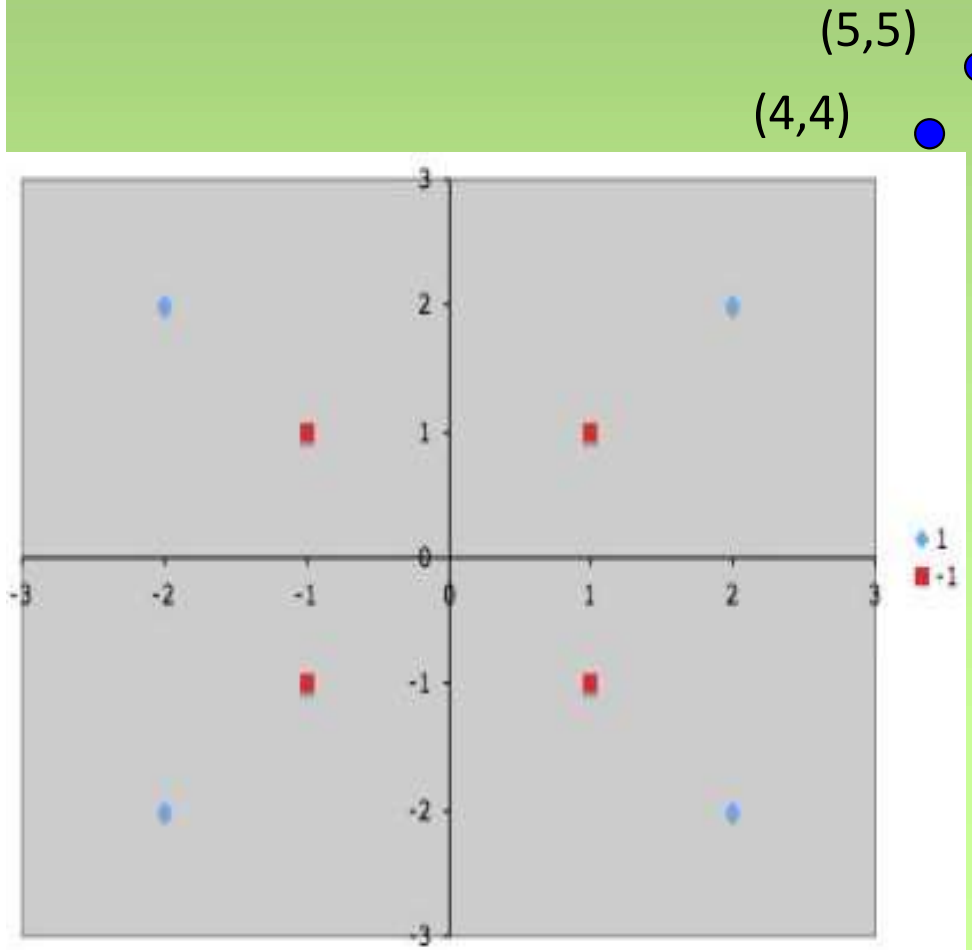
Capacidad de Generalización

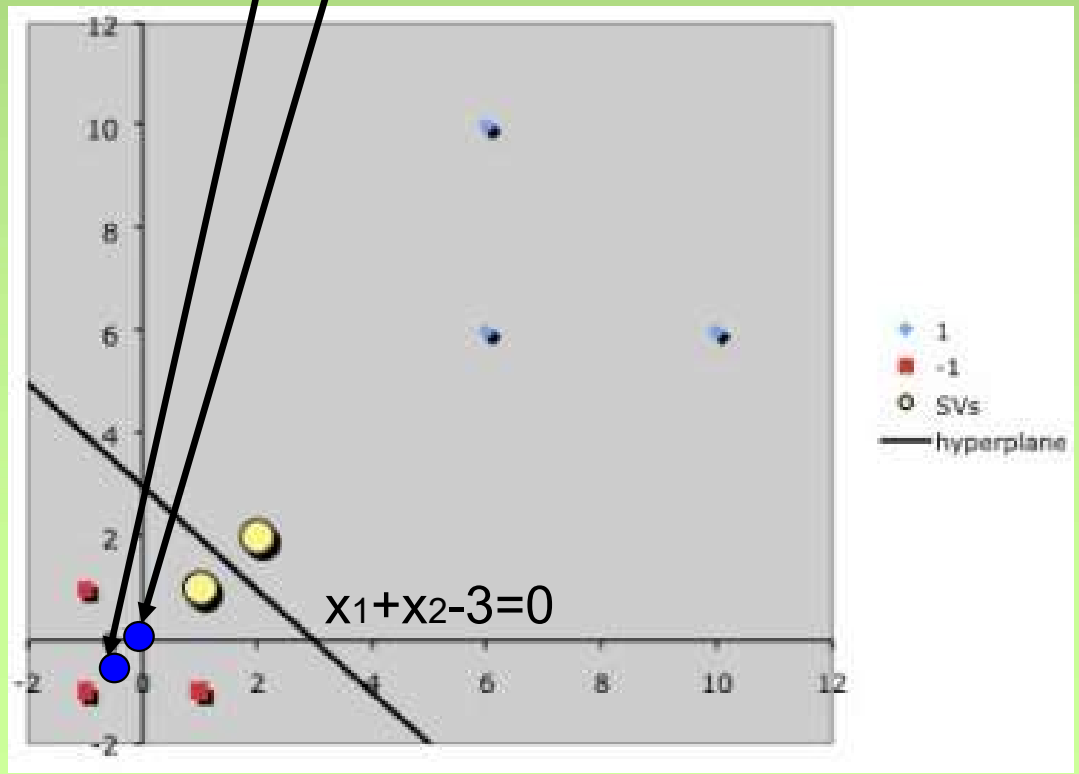
$$\Phi(x_1, x_2) = \begin{cases} (4 - x_2 + |x_1 - x_2|, 4 - x_1 + |x_1 - x_2|) & \text{si } \sqrt{x_1^2 + x_2^2} > 2 \\ (x_1, x_2) & \text{en otro caso} \end{cases}$$

Esta función separa realmente al espacio original de forma lineal?

Dato de test (5,5, +1) **Φ(5,5) = (-1, -1)** **Error de clasificación!**

Dato de test (4,4, +1) **Φ(4,4) = (0, 0)** **Error de clasificación!**



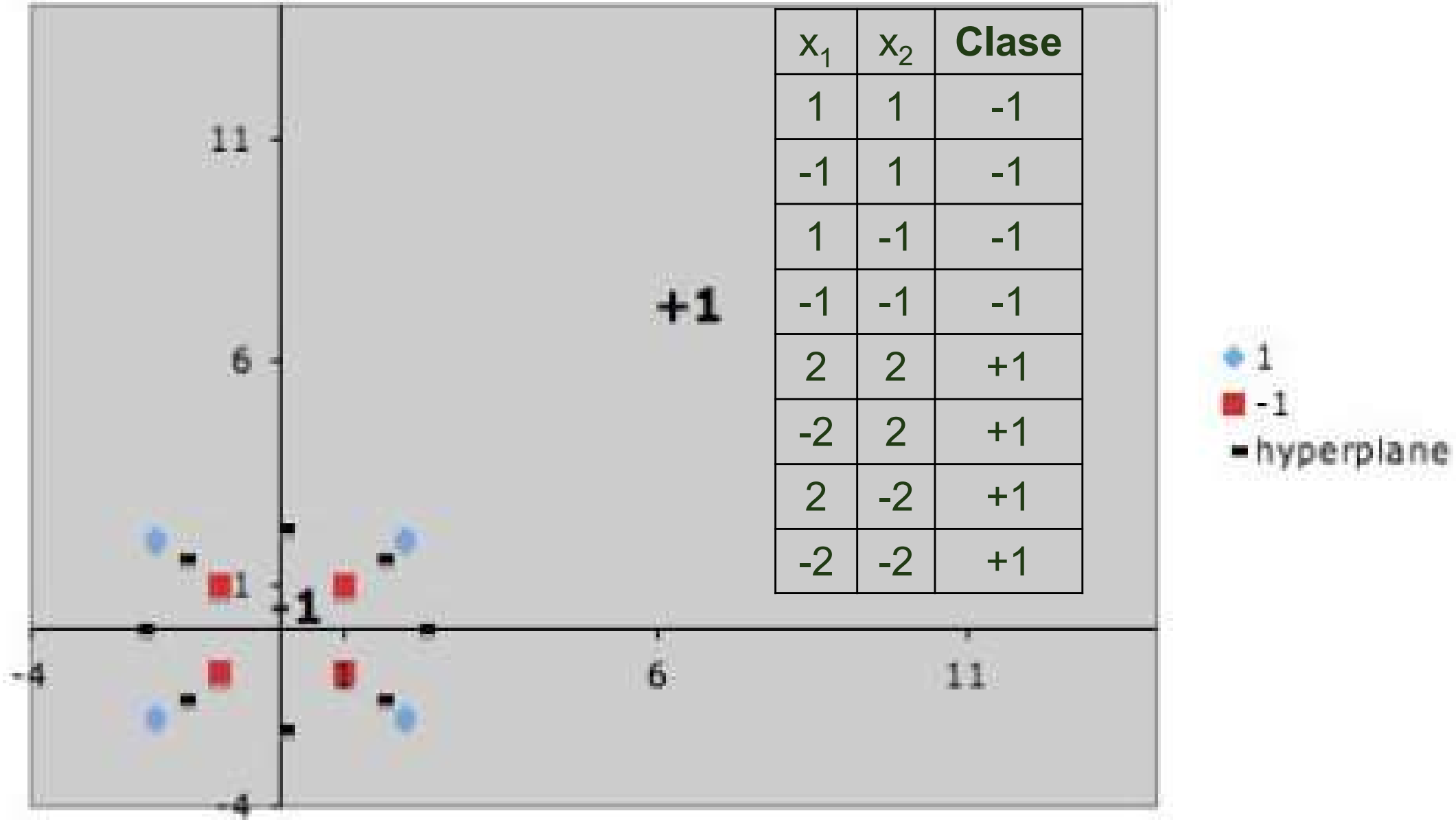


Nuevo Ejemplo

$$\Phi(x_1, x_2) = \left[x_1, x_2, \left((x_1^2 + x_2^2) - 5\right)/3\right]$$

Con esta transformación la capacidad de generalización aumenta

x_1	x_2	Clase
1	1	-1
-1	1	-1
1	-1	-1
-1	-1	-1
2	2	+1
-2	2	+1
2	-2	+1
-2	-2	+1



Nuevo Ejemplo

$$\Phi(x_1, x_2) = \left[x_1, x_2, \left((x_1^2 + x_2^2) - 5\right)/3\right]$$

Con esta transformación la capacidad de generalización aumenta

Todos son vectores soporte

x_1	x_2	$x_3 = ((x_1^2 + x_2^2) - 5)/3$	Clase
1	1	-1	-1
-1	1	-1	-1
1	-1	-1	-1
-1	-1	-1	-1
2	2	1	+1
-2	2	1	+1
2	-2	1	+1
-2	-2	1	+1

Vectores Soporte

H_0: **w.x** + $w_0 = 1$

H_1: **w.x** + $w_0 = -1$

1	1	-1	-1
-1	1	-1	-1
1	-1	-1	-1
-1	-1	-1	-1
2	2	1	+1
-2	2	1	+1
2	-2	1	+1
-2	-2	1	+1

$w_1x_1 + w_2x_2 + w_3x_3 + w_0 = -1$

$\left.\begin{aligned} 1w_1 + 1w_2 - 1w_3 + w_0 &= -1 \\ -1w_1 + 1w_2 - 1w_3 + w_0 &= -1 \end{aligned}\right\}$ $2w_2 - 2w_3 + 2w_0 = -2$

$\left.\begin{aligned} 1w_1 - 1w_2 - 1w_3 + w_0 &= -1 \\ -1w_1 - 1w_2 - 1w_3 + w_0 &= -1 \end{aligned}\right\}$ $-2w_2 - 2w_3 + 2w_0 = -2$

$\left.\begin{aligned} 2w_1x_1 + 2w_2x_2 + w_3x_3 + w_0 &= 1 \\ -2w_1x_1 + 2w_2x_2 + w_3x_3 + w_0 &= 1 \end{aligned}\right\}$ $4w_2 + 2w_3 + 2w_0 = 2$

$\left.\begin{aligned} 2w_1x_1 - 2w_2x_2 + w_3x_3 + w_0 &= 1 \\ -2w_1x_1 - 2w_2x_2 + w_3x_3 + w_0 &= 1 \end{aligned}\right\}$ $-4w_2 + 2w_3 + 2w_0 = 2$

Sustituyendo y despejando tenemos $w_0=0$, $w_3=1$, $w_2=0$ y $w_1=0$

$$(0,0,1)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + 0 = 0; \quad x_3 = 0$$ es la función discriminante y los

hiperplanos separadores son $x_3 = 1 \;\; y \;\; x_3 = -1$

Nuevo Ejemplo

$$\Phi(x_1, x_2) = \left[x_1, x_2, ((x_1^2 + x_2^2) - 5)/3\right]$$

X3 = 0; **Ecuación del plano**

¿Esta función separa realmente al espacio original de forma lineal?

Dato de test (5, 5, +1) **Φ(5,5) = (5, 5, 45/3) , al ser 45/3 positivo la clase es +1 Acierto**

Dato de test (4, 4, +1) **Φ(4,4) = (4, 4, 27/3) al ser 27/3 positivo la clase es +1 Acierto**

Dato de test (0,5, 0,5, -1) **Φ(0.5,0.5) = (0.5, 0.5, -4.5/3) al ser -4.5/3 negativo la clase es -1 Acierto**

$$x_3 = ((x_1^2 + x_2^2) - 5)/3$$

x_1	x_2		**Clase**
5	5	45/3	1
4	4	27/3	1
0,5	0,5	-4.5/3	-1

SVMs Lineales,
Caso no separable

SVM: Caso no Separable (1)

- En el caso no separable no es posible clasificar sin errores todo el conjunto de entrenamiento mediante un clasificador lineal.
- Ya no será posible cumplir todas las condiciones

– Por tanto será $y_i(\mathbf{w}^T\mathbf{x}_i+w_0) \geq 1-\xi_i, \quad i=1...n, \text{ con } \xi_i \geq 0$

– Entonces:

- Si $\xi_i = 0$ el patrón está en la zona de su clase.
- Si $0 \leq \xi_i \leq 1$ la muestra se mete en la zona del margen.
- Si $\xi_i > 1$ se mete en la zona de la otra clase.

- Ahora tenemos dos criterios u objetivos:

– Obtener la frontera de mayor margen.

– Que esta frontera tenga pocas “equivocaciones” (es decir que ξ_i sea lo más cercana a 0).

- Lo que se hace es combinar los dos criterios u objetivos.

SVM: Caso no Separable (2)

En este caso (w, w_0) no satisface la restricción

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1, \quad i = 1 \ldots n,$$

Por lo que decimos que la solución no es factible. En 1995 (Cortes y Vapnik) introducen variables artificiales ξi, i = 1, . . . , n en las restricciones

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i$$

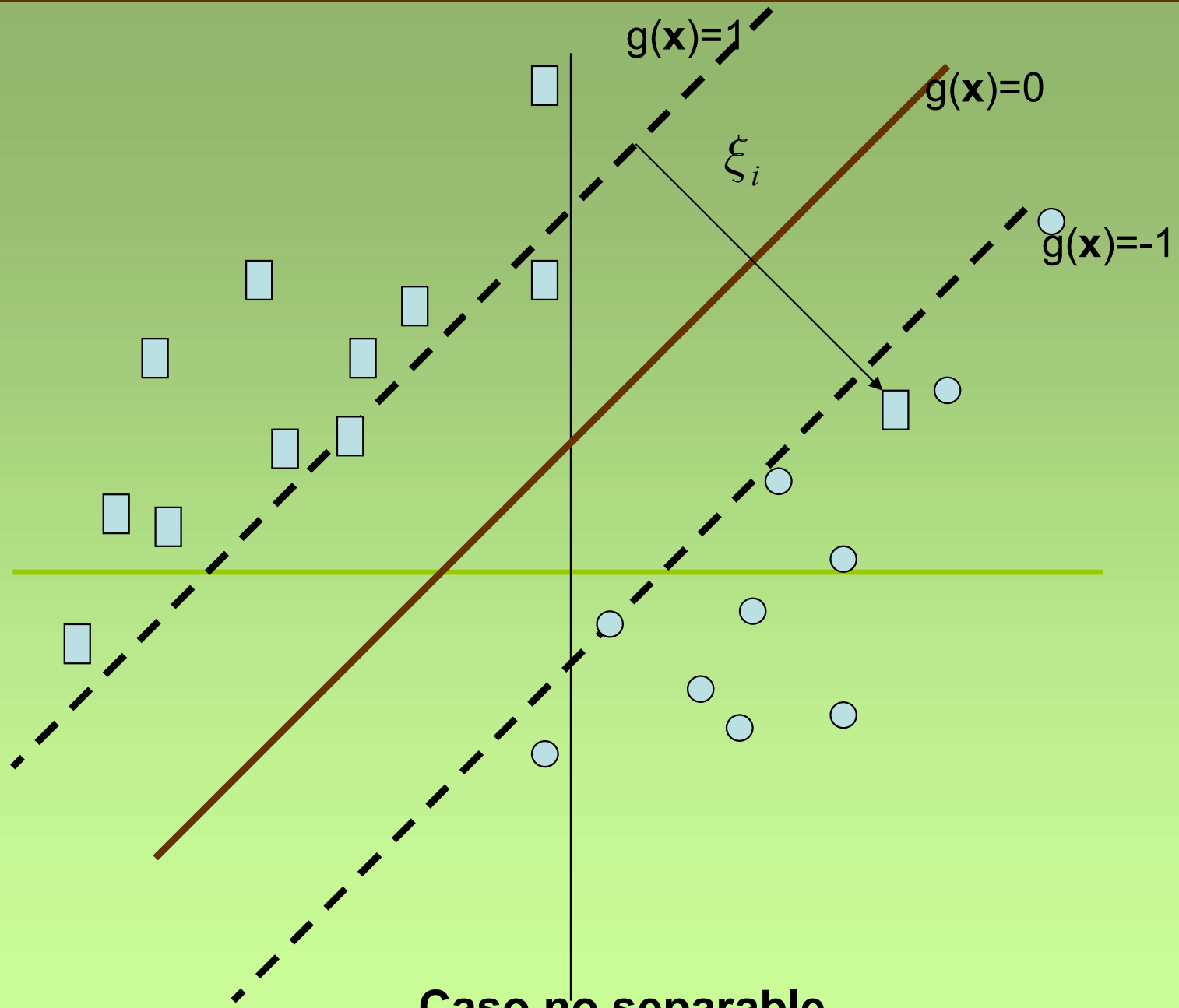
$$s.a. \; y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1-\xi_i, \quad i = 1 \ldots n, \text{ con } \xi_i \geq 0 \quad (3)$$

Es decir, las restricciones (3) permiten que los datos de entrenamiento puedan no estar en el lado correcto del hiperplano de separación $\mathbf{w}^T\mathbf{x} + w_0 = 0$

Esta situación ocurre cuando ξi > 1 . Un ejemplo se muestra en la siguiente figura

El nuevo problema es siempre factible para todo par (w, w_0), dado que el valor $\xi_i = \max(0, 1 - y_i(\mathbf{w}^T\mathbf{x} + w_0)), \; i=1,\ldots,n$ **permite que (w, w_0), sea una solución factible**

SVM: Caso no Separable (4)



Caso no separable

SVM: Caso no Separable (5)

Combinación de criterios u optimización multiobjetivo:

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{n}\xi_i$$

$$s.a.\ \ y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \mathrm{w}_0) \geq 1-\xi_i, \quad i = 1...n, \ \text{con}\ \xi_i \geq 0$$

– El primer término de la suma a optimizar busca el mayor margen, el segundo el menor número de errores. La importancia de cada uno se expresa a través de la constante *C* que esta definida por el usuario y se define como un parámetro de penalización. Si pasamos al dual

- **Problema dual:**

$$\max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

$$s.a.\ \ \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \mathrm{C} \geq \alpha_i \geq 0,\ \mathrm{i}=1,\ldots,\mathrm{n}$$

- **Vector óptimo:**

$$\hat{\mathbf{w}} = \sum_{i \in Sop} \hat{\alpha}_i y_i \mathbf{x}_\mathbf{i}$$

SVM: Caso no Separable (6)

Parámetro C en SVM lineal

El parámetro C en (SVM lineal):

Establece el peso de la suma de las variables de holgura.

Sirve como un parámetro de regularización.

Controla el número de vectores soporte.

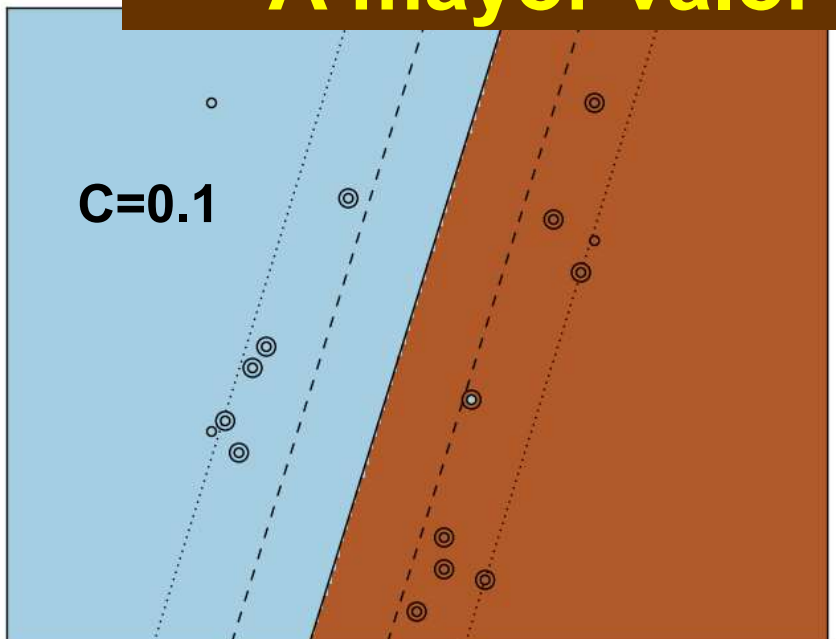
Si C es grande se penaliza mucho los errores, por lo que el algoritmo tiende a hacer el margen más pequeño.

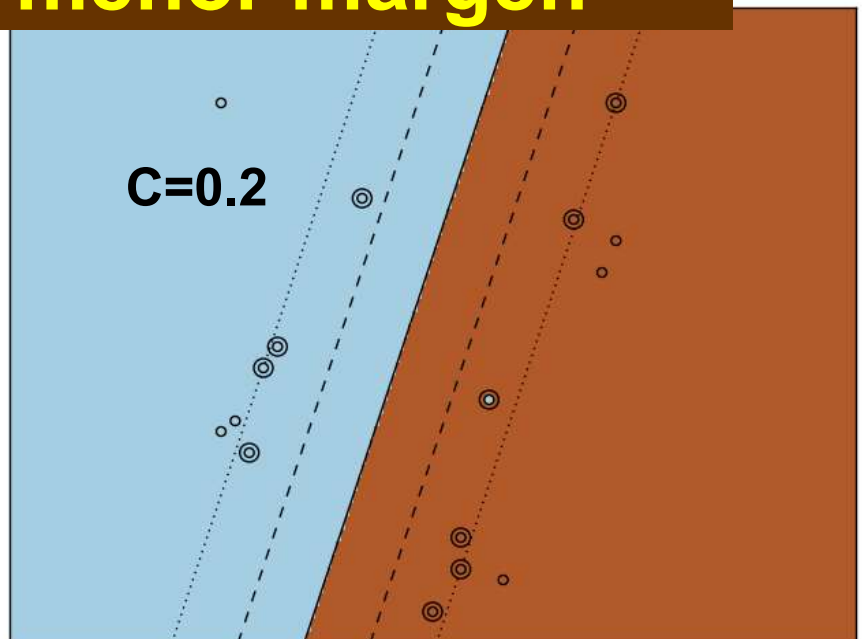
Si C es pequeño se penaliza poco los errores, por lo que el algoritmo tiende a hacer el margen mayor.

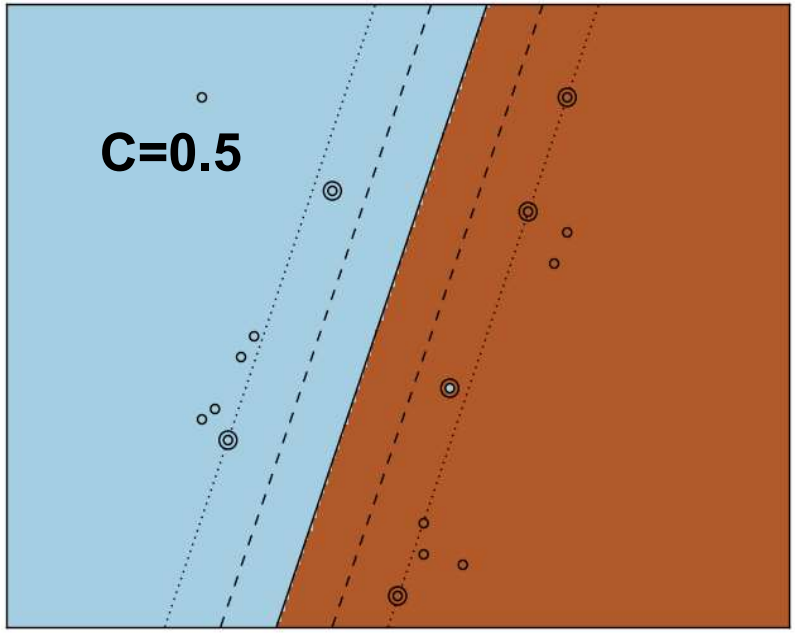
C	Penalización por Error	Nº de puntos considerados	Margen	Sesgo	Varianza
Pequeño	Baja	Alto	Alto	Alto	Grande
Grande	Alta	Alto	Bajo	Bajo	Pequeña

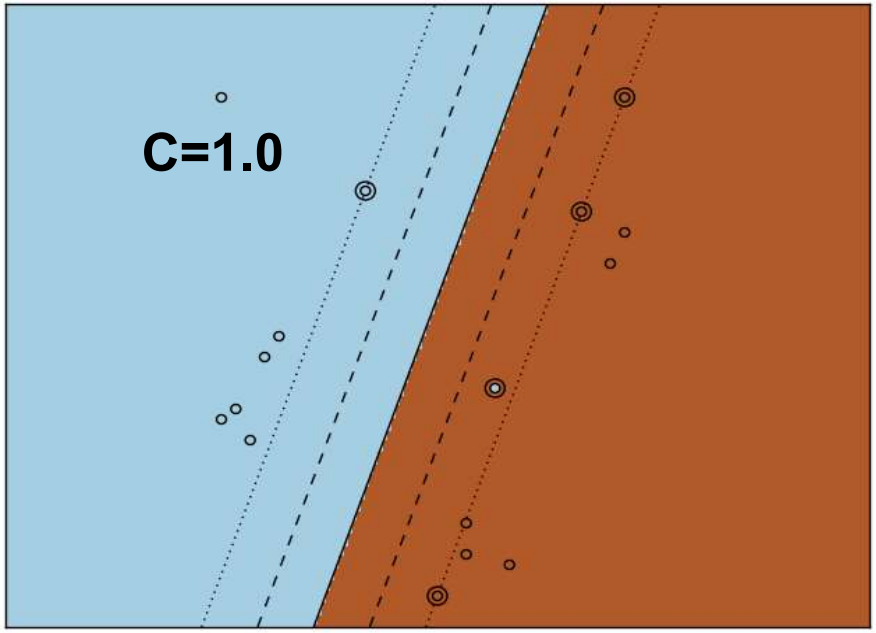
PARAMETRO C EN SVM LINEAL (1)

A mayor valor de C menor margen



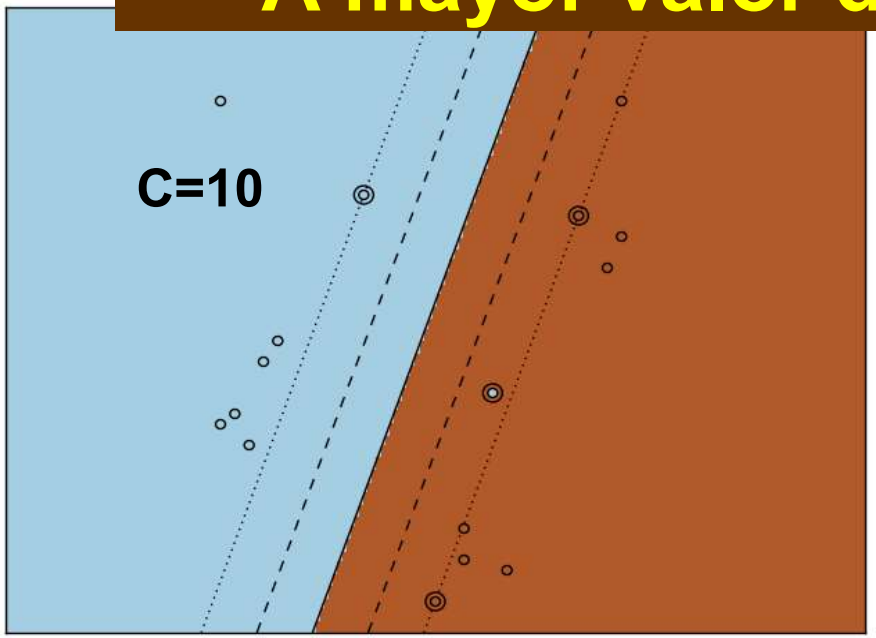


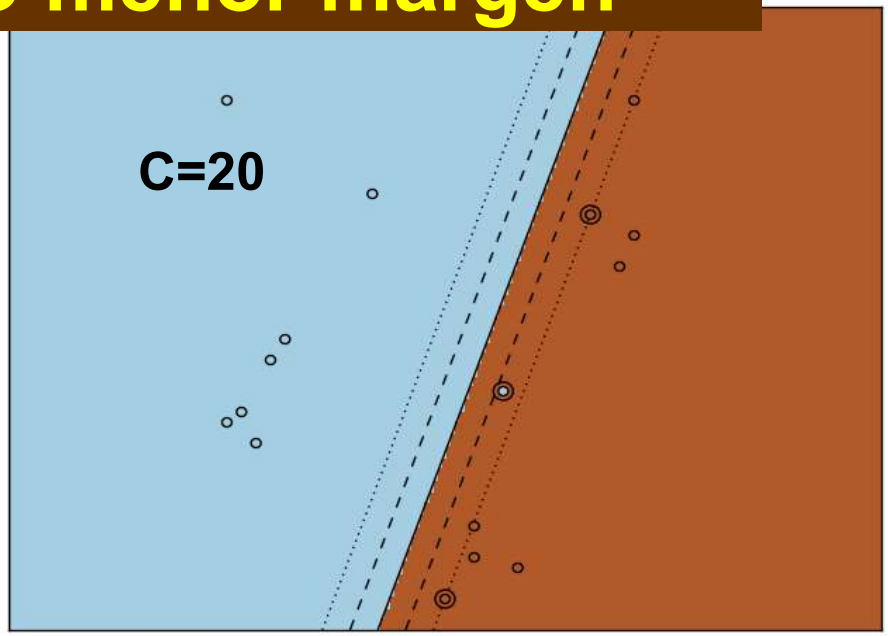


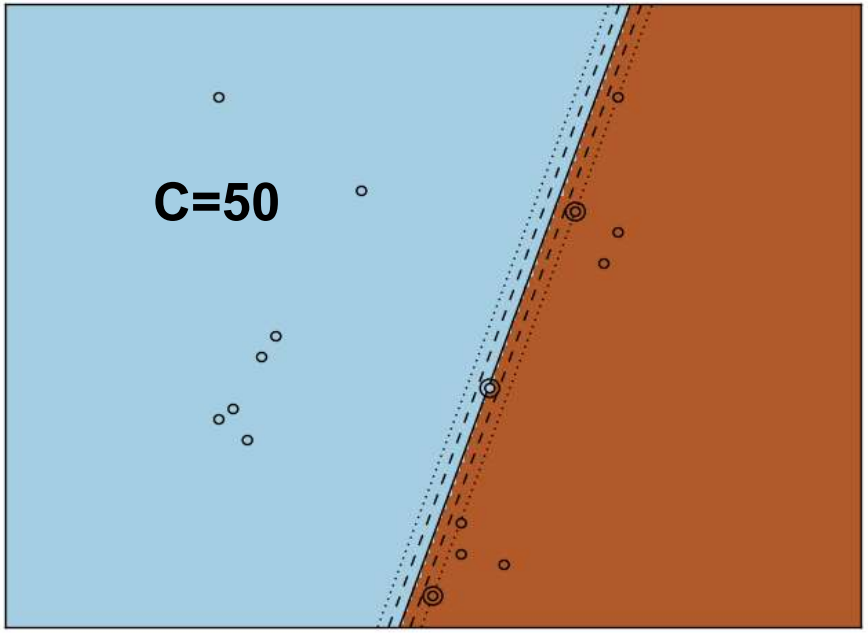


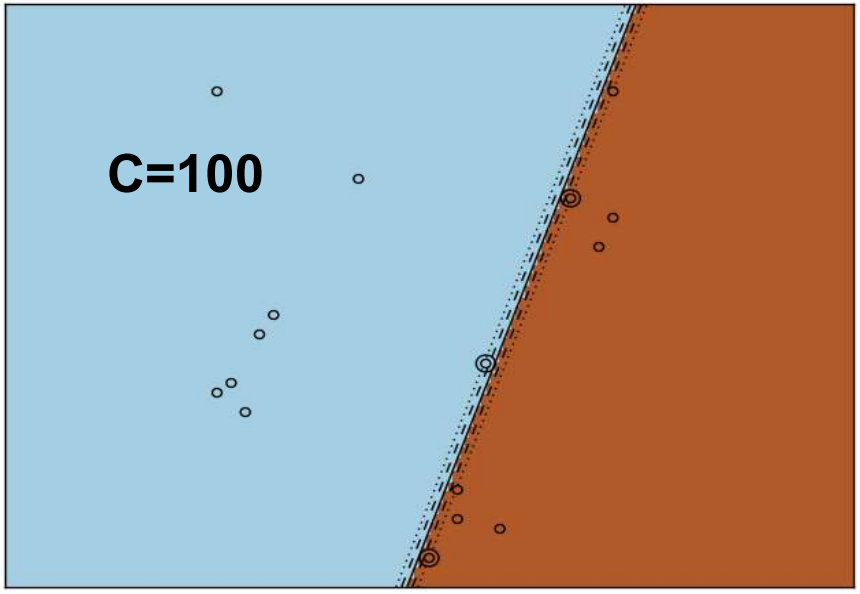
PARAMETRO C EN SVM LINEAL (2)

A mayor valor de C menor margen









SVMs No Lineales
Truco del Kernel

Teorema de Cover:

Las SVM ofrecen un método elegante y eficaz de resolver problemas de clasificación lineal

Pero, son estos problemas interesantes o frecuentes?

Teorema de Cover:

si N > d + 1, el número de problemas de tamaño N linealmente separables es

$$2\sum_{i=0}^{d}\binom{N-1}{i} \le 2(d+1)\binom{N-1}{d} \le 2\frac{d+1}{d!}N^{d}$$

Siendo N el tamaño muestral y d la dimensión del espacio de entrada. Como el número total de problemas de dos clases es 2N, si N » d, la inmensa mayoría de los problemas NO van a ser linealmente separables

Problema

- Como escoger una función $\Phi(x)$ tal que el espacio de características transformado sea eficiente para clasificación sin un costo computacional demasiado alto?
- Truco del Kernel

- Funciones de Núcleo (kernel)
  - Polinomial
  - Gaussiana
  - Sigmoidal

  - Siempre aumentan el número de dimensiones del espacio de entrada

Maquina de Vectores Soporte, SVM, NO lineal (1)

Con una metodología SVM lineal solo pueden obtenerse fronteras lineales

• La forma de obtener una metodología SVM para fronteras no lineales se basa en la misma idea que las Funciones Discriminantes Generalizadas.

– Esta consiste en transformar los datos de entrada mediante un conjunto de funciones no lineales

$$\phi(\mathbf{x})=(\phi_0(\mathbf{x}),\phi_1(\mathbf{x}),\phi_2(\mathbf{x}),\ldots,\phi_M(\mathbf{x}))^T$$

y luego aplicar la metodología SVM lineal.

• Tras esta transformación se tiene (por ej. para el caso hard-linealmente separable):

Maquina de Vectores Soporte, SVM, NO lineal (2)

– Problema dual:

$$\max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j (\Phi(\mathbf{x}_i)^T . \Phi(\mathbf{x}_j))$$

$$s.a. \sum_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \text{ i=1,...,n}$$

w óptimo:

$$\hat{\mathbf{w}} = \sum_{i \in Sop} \hat{\alpha}_i y_i \Phi(\mathbf{x}_i),$$

donde $\Phi(\mathbf{x}_i)$ son los vectores soporte en el espacio transformado

w_0 óptimo:

$$\hat{w}_0 = 1\text{-}\hat{\mathbf{w}}\Phi(\mathbf{x_i}), \text{ con } \mathbf{x_i} \in \omega_1 \text{ y } \hat{\alpha}_i > 0$$

solo un vector soporte positivo

Clasificador óptimo:

$$g(\mathbf{x}) = \hat{\mathbf{w}}^{\mathbf{T}} \Phi(\mathbf{x}) + \hat{w}_0$$

El truco del *kernel* (1)

Problema:

– Ya vimos que uno de los problemas de trabajar con funciones $\phi(\mathbf{x})=(\phi_0(\mathbf{x}),\phi_1(\mathbf{x}),\phi_2(\mathbf{x}),\ldots,\phi_M(\mathbf{x}))^T$ no lineales consiste en que generalmente el número de ellas, *M,* es muy grande.

– El coste computacional de calcular los productos escalares $\Phi(\mathbf{x}_i)^T.\Phi(\mathbf{x}_j)$ es también muy grande.

• **Observación:**

– En determinadas circunstancias es posible evaluar los productos escalares $\Phi(\mathbf{x}_i)^T.\Phi(\mathbf{x}_j)$ sin calcular las funciones transformadas.

El Truco del *kernel* (2) : Ejemplos

Truco del núcleo:

– **El truco del núcleo consiste en calcular**

$$\Phi(\mathbf{x}_i)^T.\Phi(\mathbf{x}_j)$$

mediante alguna función de las muestras originales k(x_i,x_j). A esta función se le llama función núcleo o *kernel*.

Ejemplo 1.-

Si consideramos $\Phi(\mathbf{x}) = \Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$, se tiene que

$$\Phi(\mathbf{x})^T.\Phi(\mathbf{x}') = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T.(x'^2_1, \sqrt{2}x'_1x'_2, x'^2_2) =$$

$$= x_1^2x'^2_1 + 2x_1x_2x'_1x'_2 + x_2^2x'^2_2 =$$

$$= (x_1x'_1 + x_2x'_2)^2 = (\mathbf{x}^T.\mathbf{x}')^2 = k(\mathbf{x}, \mathbf{x}')$$

El Truco del *kernel* (3) : Ejemplos

Ejemplo 2.-

Si, $\mathbf{x}=(x_1, x_2)$ y $\mathbf{x}'=(x'_1, x'_2)$

Si definimos $\Phi(\mathbf{x}) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2)$, se tiene que

$$\Phi(\mathbf{x})^T.\Phi(\mathbf{x}') = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2)\begin{pmatrix} 1 \\ \sqrt{2}x_1' \\ \sqrt{2}x_2' \\ x_1'^2 \\ \sqrt{2}x_1'x_2' \\ x_2'^2 \end{pmatrix} =$$

$$= \left(1 + 2x_1x_1' + 2x_2x_2' + x_1^2x_1'^2 + 2x_1x_2x_1'x_2' + x_2^2x_2'^2\right) =$$

$$= (1 + x_1x_1' + x_2x_2')^2 = (1 + \mathbf{x}^T.\mathbf{x}')^2 = k(\mathbf{x}, \mathbf{x}')$$

El Truco del *kernel* (4)

- **Todo el problema SVM no lineal se puede expresar con funciones de tipo núcleo.**
- **Por ejemplo para el caso linealmente separable en el espacio de características:**
  - **Problema dual:**

$$\max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j k(\mathbf{x_i}, \mathbf{x_j})$$

$$s.a. \sum_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \; \text{i=1,...,n}$$

  - ***w* óptimo:**

$$\mathbf{w} = \sum_{i \in \text{Sop}} \hat{\alpha}_i y_i K(\mathbf{x}_i, \mathbf{x})$$

  - ***w*$_0$ óptimo:**

$$\hat{\text{w}}_0 = 1 - \sum_{i \in Sop} \hat{\alpha}_i y_i k(\mathbf{x_i}, \mathbf{x_j}), \; \text{con } \mathbf{x_i} \in \omega_1 \; \text{y} \; \hat{\alpha}_i > 0$$

solo un vector soporte positivo

  - **Clasificador óptimo:**

$$\text{g}(\mathbf{x}) = \sum_{i \in Sop} \hat{\alpha}_i y_i k(\mathbf{x_i}, \mathbf{x}) + \hat{w}_0,$$

Teorema de Mercer

- Si el núcleo $K(x,z)$ es definido positivo, hay una aplicación no lineal $\Phi(x)$ de los patrones originales en un espacio de Hilbert (de dimensión posiblemente infinita) tal que para cada x, x_0,

$$\Phi(x)\,\Phi(x_0) = K(x, x_0)$$

- Consecuencia: truco del núcleo (kernel trick)
  no hace falta conocer $\Phi(x)$, $\Phi(x_0)$ para calcular

$$\Phi(x).\Phi(x_0)$$

Teorema de Mercer

- El principal requerimiento para definir una función de tipo *kernel* (nucleo) es que exista su correspondiente transformación, tal que la función *kernel*, calculada para un par de vectores, sea equivalente a su producto escalar en el espacio transformado.

- Algunas funciones *kernel* para clasificación son:

$$k(\mathbf{x}, \mathbf{x}_j) = (\mathbf{x}^T \mathbf{x}_j + 1)^p$$

$$k(\mathbf{x}, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}-\mathbf{x}_j\|^2}{2\sigma^2}}$$

$$k(\mathbf{x}, \mathbf{x}_j) = \tanh(c\mathbf{x}^T \mathbf{x}_j - d)$$

El Truco del *kernel* (5)

- La función núcleo ***k*(*x*,*y*)** mide la similitud entre las muestras ***x*** e ***y***.

- No toda función ***k*** **(*x*,*y*)** puede ser utilizada como función núcleo. Debe satisfacer la denominada condición de Mercer.

- Algunas funciones núcleo:

– Lineal: $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{y}$

– Polinomial: $k(\mathbf{x},\mathbf{y}) = (\lambda(\mathbf{x}^{\mathrm{T}}\mathbf{y}) + \theta)^{p}, p = 1,2,\ldots$

– Tangente hiperbólica: $k(\mathbf{x},\mathbf{y}) = \tanh(\lambda(\mathbf{x}^{\mathrm{T}}\mathbf{y}) - \theta)$

El Truco del *kernel* (6)

- **Función de base radial:** Núcleo con características de dimensión infinita

$$K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) = \exp(-\lambda(\mathbf{x}-\mathbf{y})^T.(\mathbf{x}-\mathbf{y})),\ \lambda > 0$$

- **Observación:**

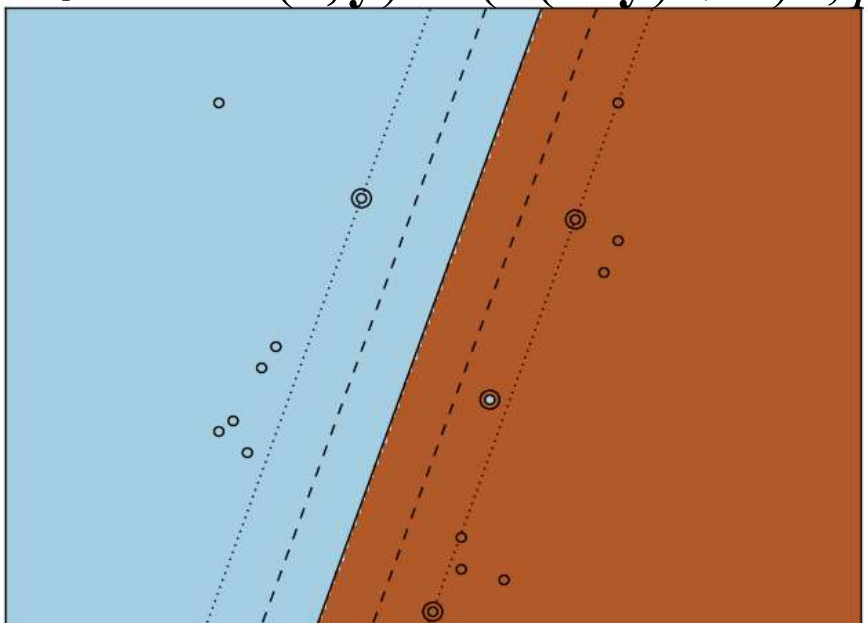
– **La forma final del clasificador lineal:**

$$g(\mathbf{x}) = \sum_{i \in Sop} \hat{\boldsymbol{\alpha}}_i y_i k(\mathbf{x_i}, \mathbf{x}) + \hat{w}_0,$$

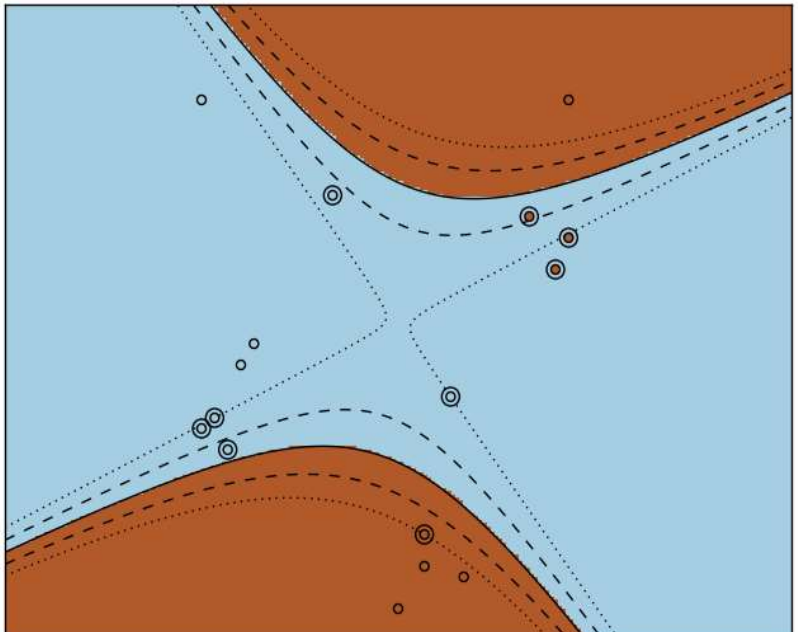
indica que el clasificador *g*(*x*) está siendo aproximado por las funciones *k*(x_i, *x*) correspondientes a los vectores soporte.

kernel polinómico, θ=1

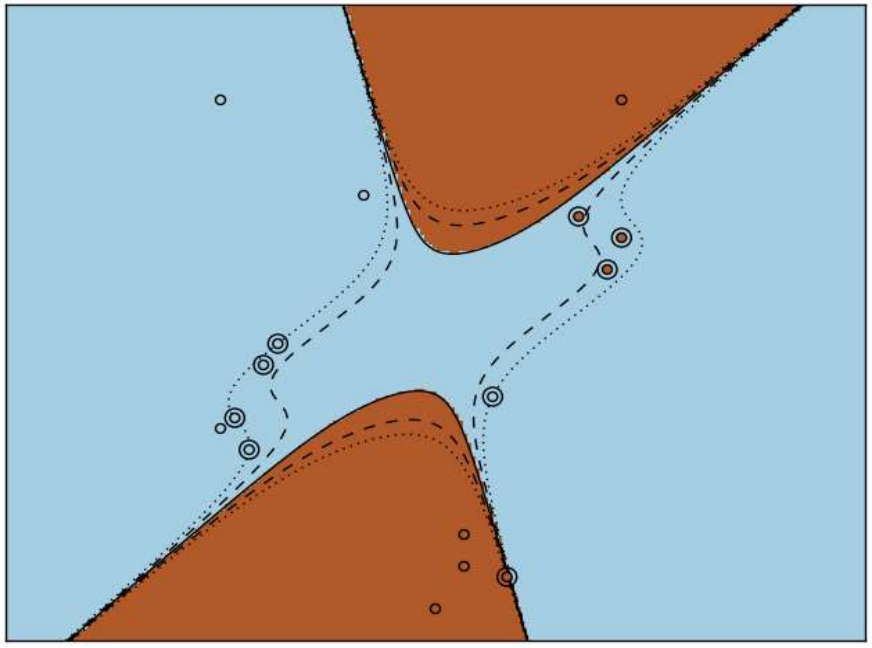
Grado p=1 $k(\mathbf{x},\mathbf{y}) = (\lambda(\mathbf{x}^{\mathrm{T}}\mathbf{y}) + \theta)^p, p = 1,2,...$



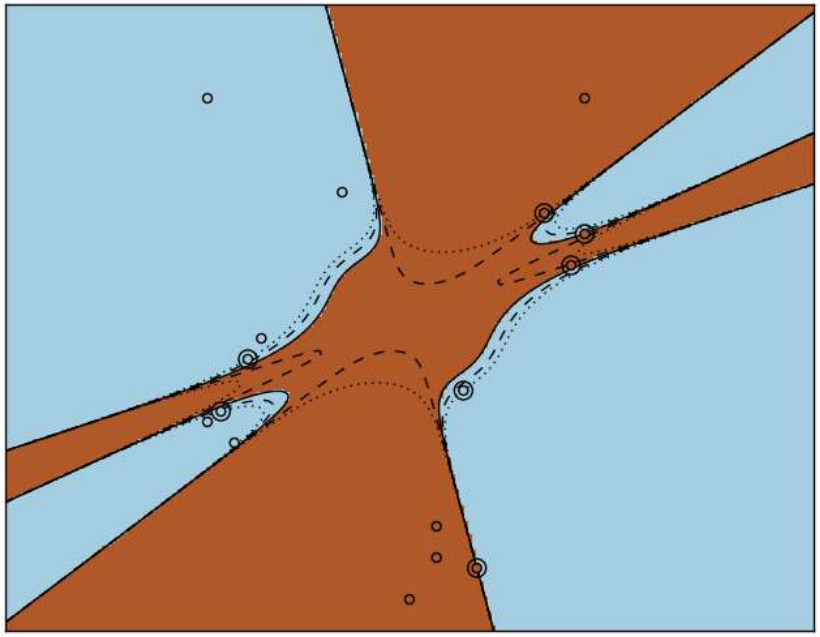
Grado p=2



Grado p=4



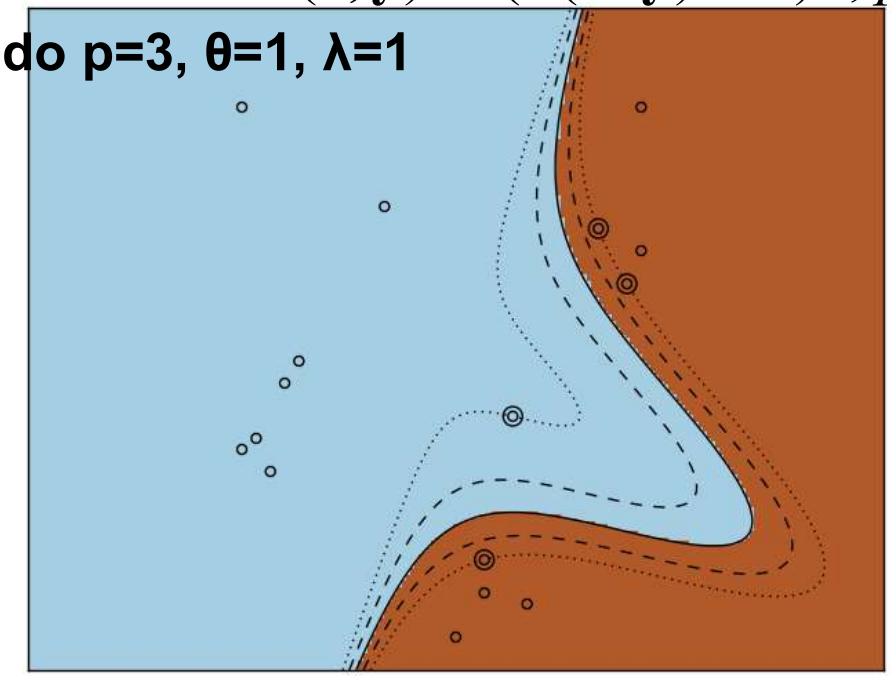
Grado p=6



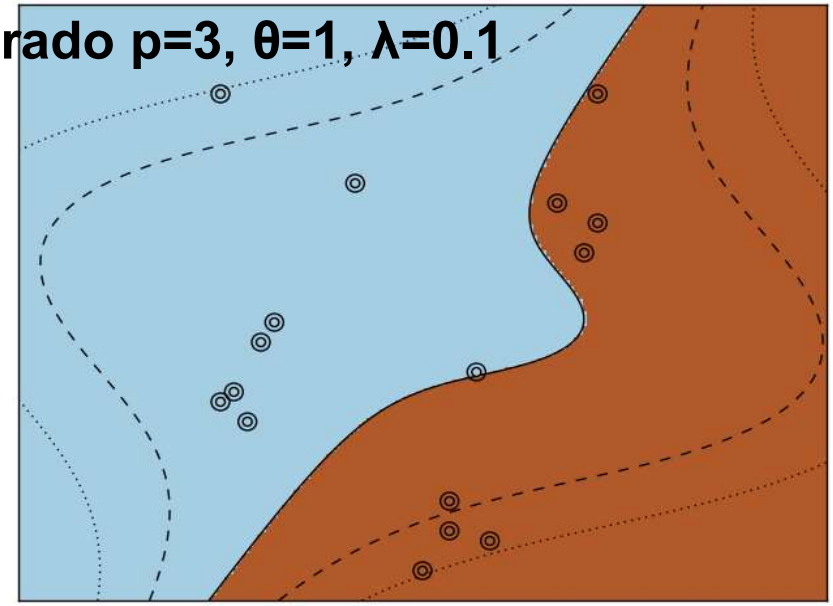
kernel polinómico, θ=1

$$k(\mathbf{x},\mathbf{y}) = (\lambda(\mathbf{x}^{\mathrm{T}}\mathbf{y}) + \theta)^{p}, p = 1,2,...$$

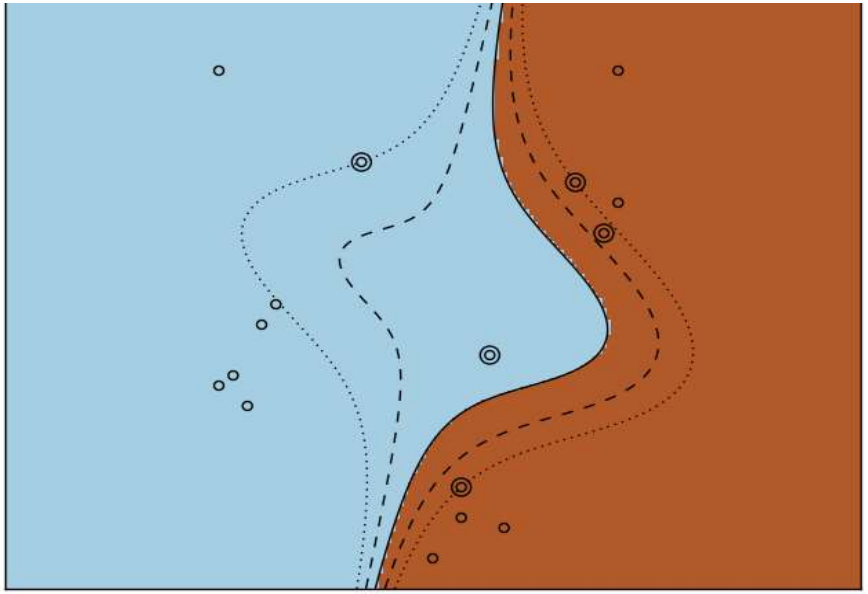
Grado p=3, θ=1, λ=1



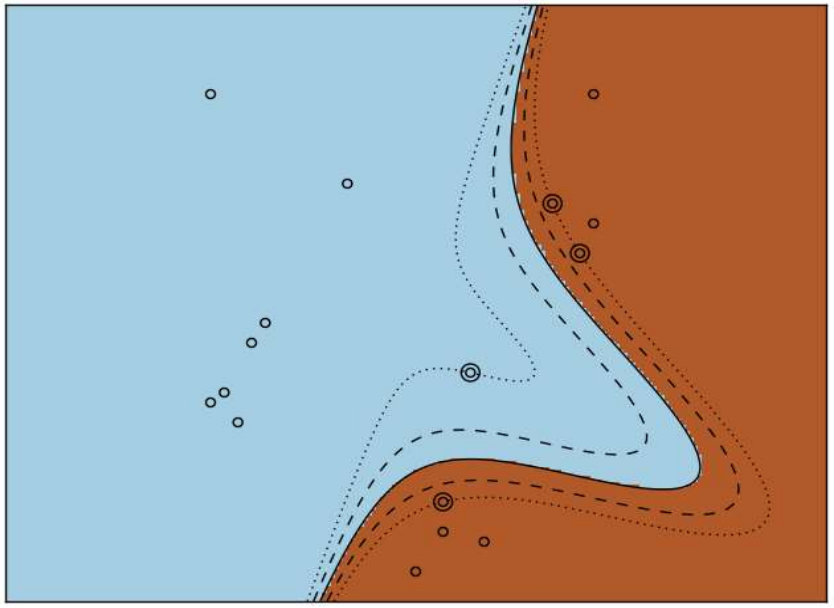
Grado p=3, θ=1, λ=0.1



Grado p=3, θ=1, λ=0.5



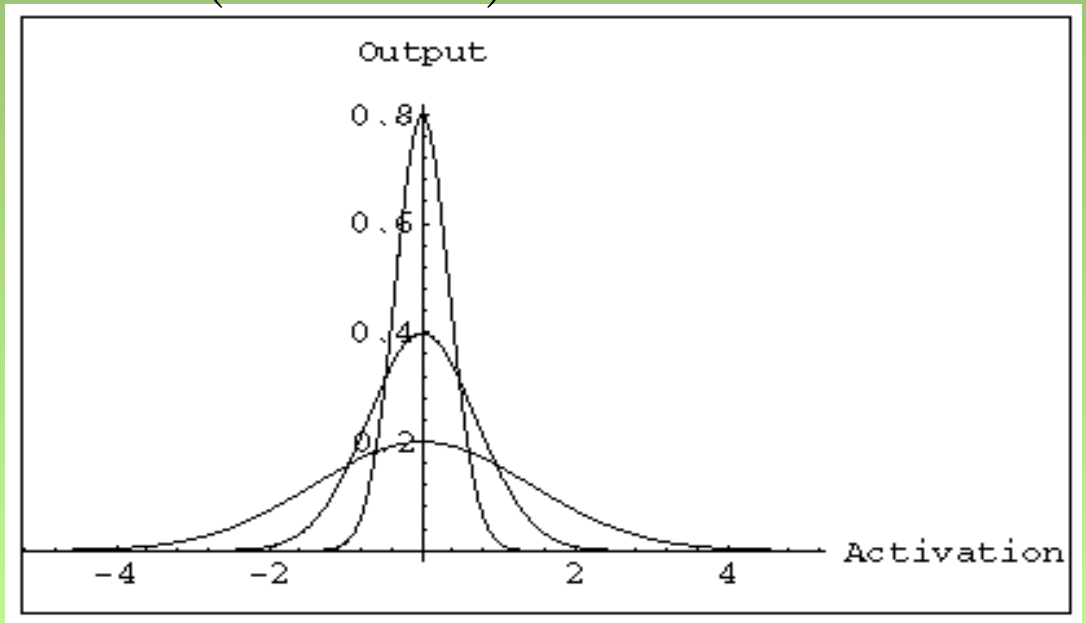
Grado p=3, θ=1, λ=2



kernel RBF

– Función de base radial: Núcleo con características de dimensión infinita

$$K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) = \exp(-\lambda(\mathbf{x}-\mathbf{y})^T.(\mathbf{x}-\mathbf{y})),\ \lambda > 0$$



Cada punto del conjunto de entrenamiento crea su campana de Gauss. La forma completa es la suma de las campanas de Gauss.

Es de la clase de “todos los vecinos más cercanos ".

Parámetros del *kernel* RBF

C	Superficie de decisión	Modelo	Sesgo	Varianza
Pequeño	Alisada	Sencillo	Alto	Grande
Grande	Picuda	Complejo	Bajo	Pequeña

Valor de Lambda λ	Puntos afectados
Bajo, sigma alto	Pueden estar lejos de los ejemplos de entrenamiento
Alto, sigma bajo	Deben estar cerca de los ejemplos de entrenamiento

Un valor de lambda alto hace una menor varianza

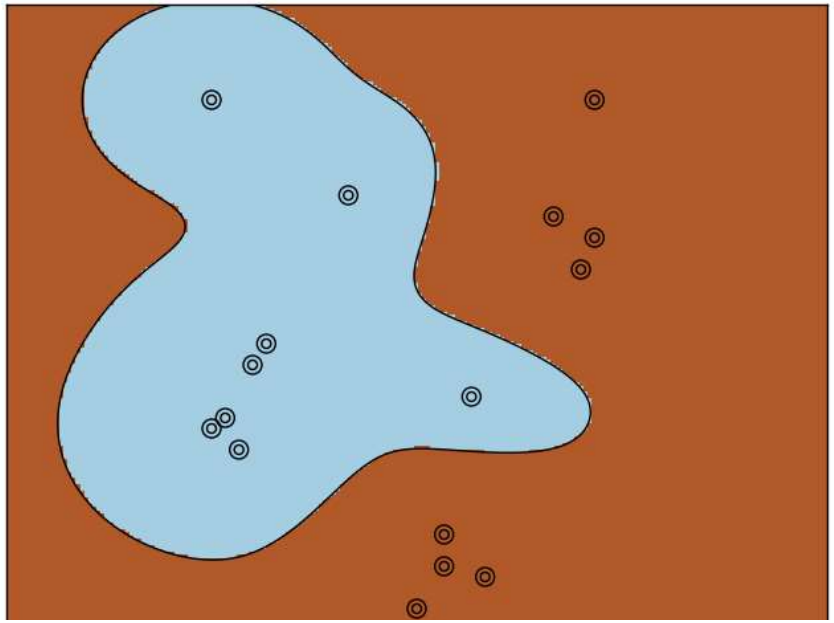
$$K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) = \exp(-\lambda(\mathbf{x}-\mathbf{y})^T.(\mathbf{x}-\mathbf{y})),\ \lambda > 0$$

- La elección de lambda es crítica para el rendimiento de SVM.
- El consejo para la elección de estos dos parámetros C y lambda es hacer una búsqueda de tipo malla por validación cruzada

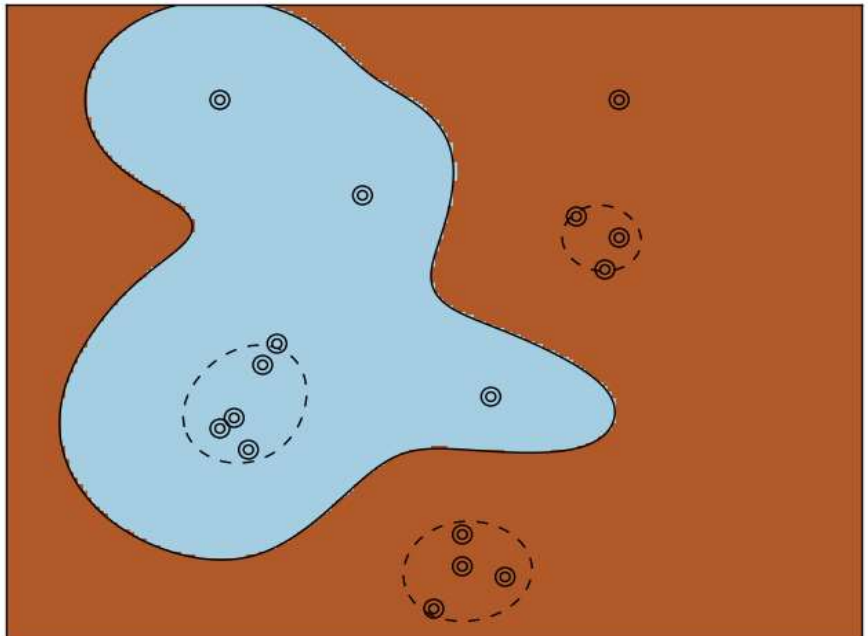
- Hacer pruebas en espacios exponenciales
- Utilizar un rango amplio

Parámetros del *Kernel* RBF

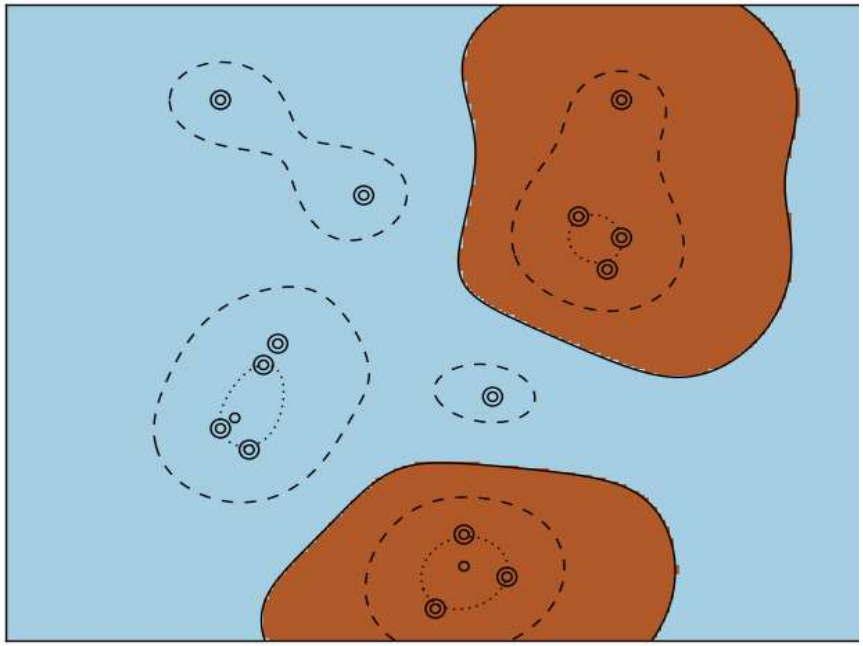
C=0.05, λ=2



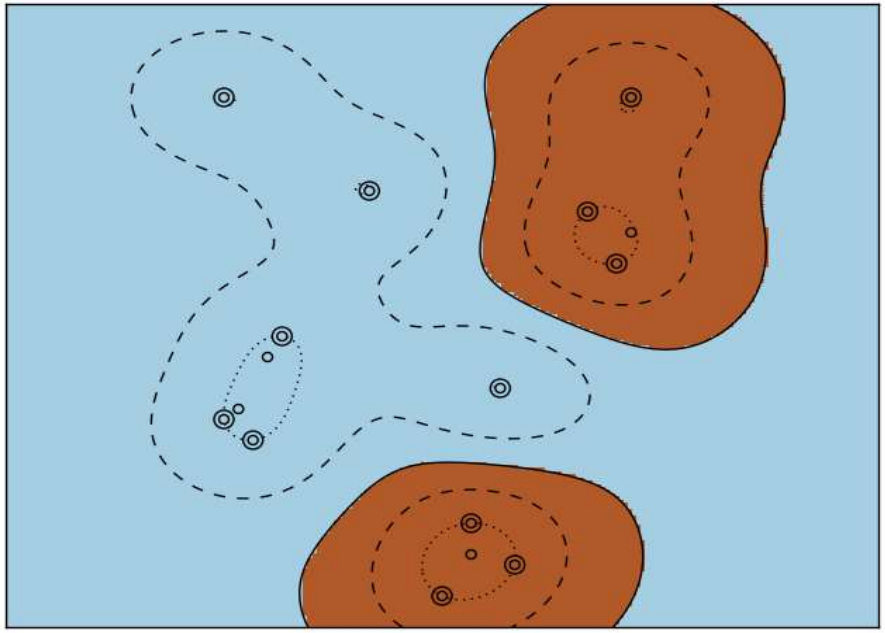
C=0.2, λ=2



C=0.6, λ=2

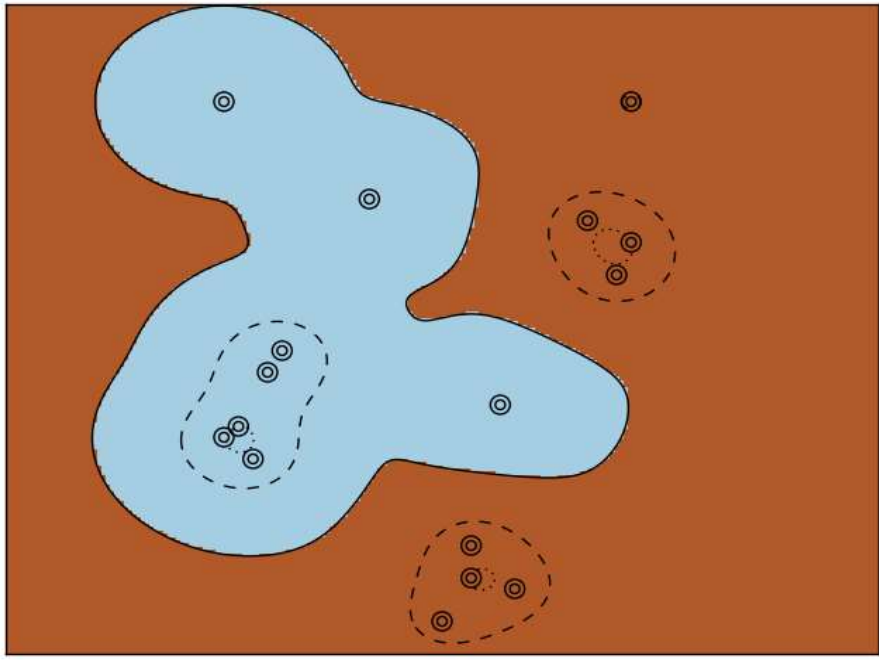


C=2, λ=2

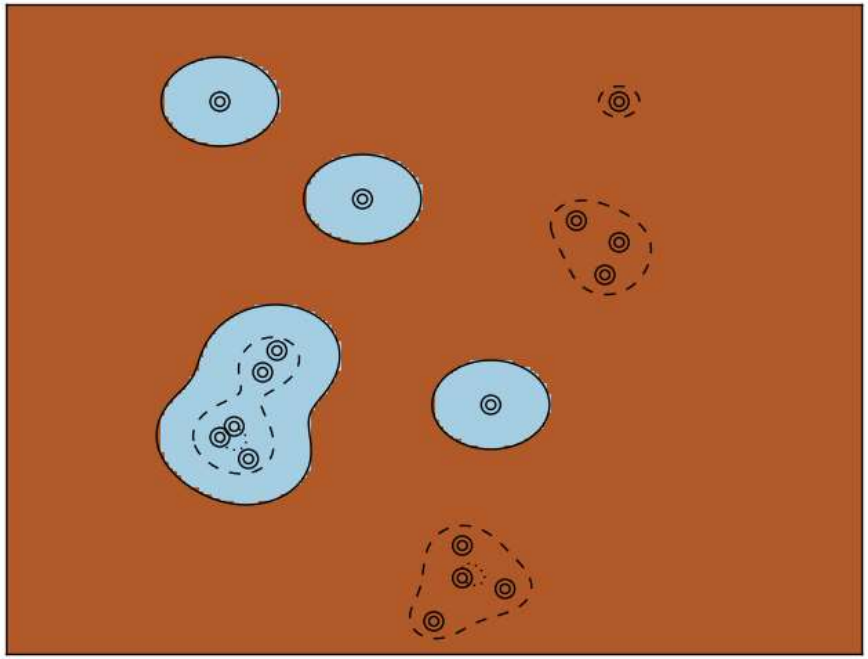


Parámetros del *Kernel* RBF

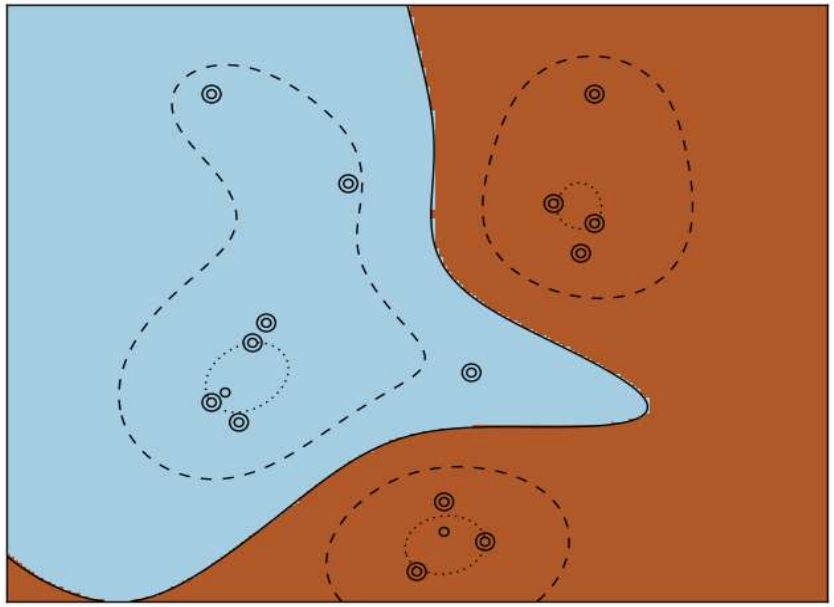
C=0.5, λ=5



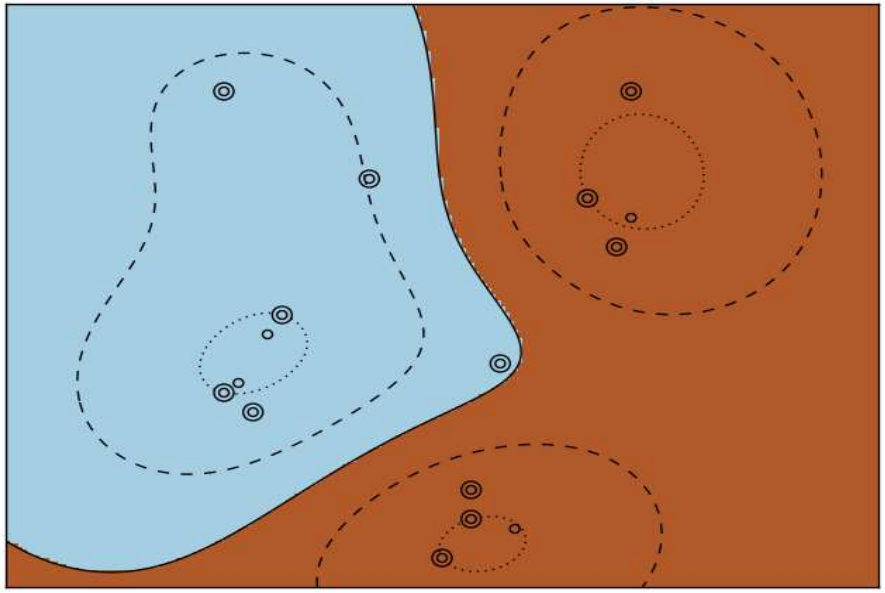
C=0.5, λ=10



C=0.5, λ=1

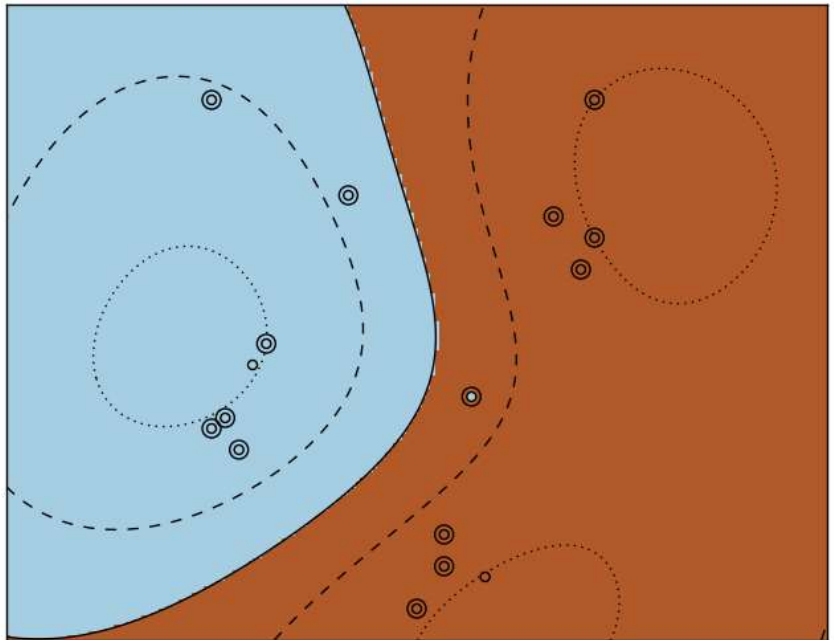


C=0.5, λ=0.5

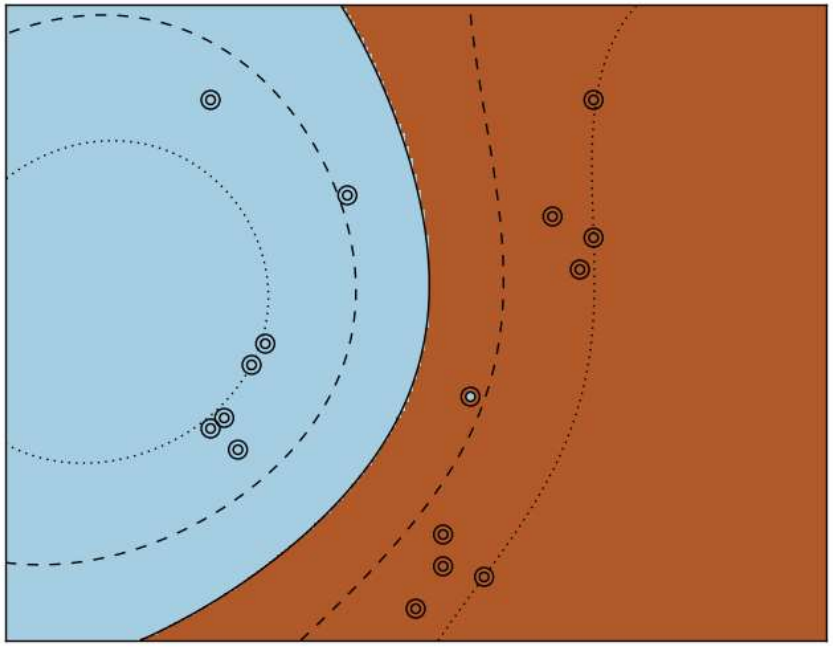


Parámetros del *Kernel* RBF

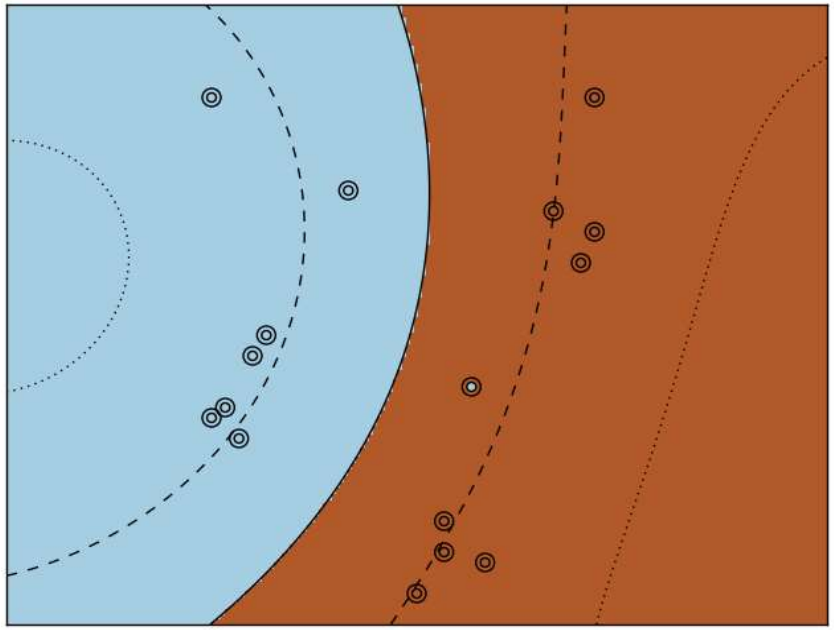
C=0.5, λ=0.2



C=0.5, λ=0.1

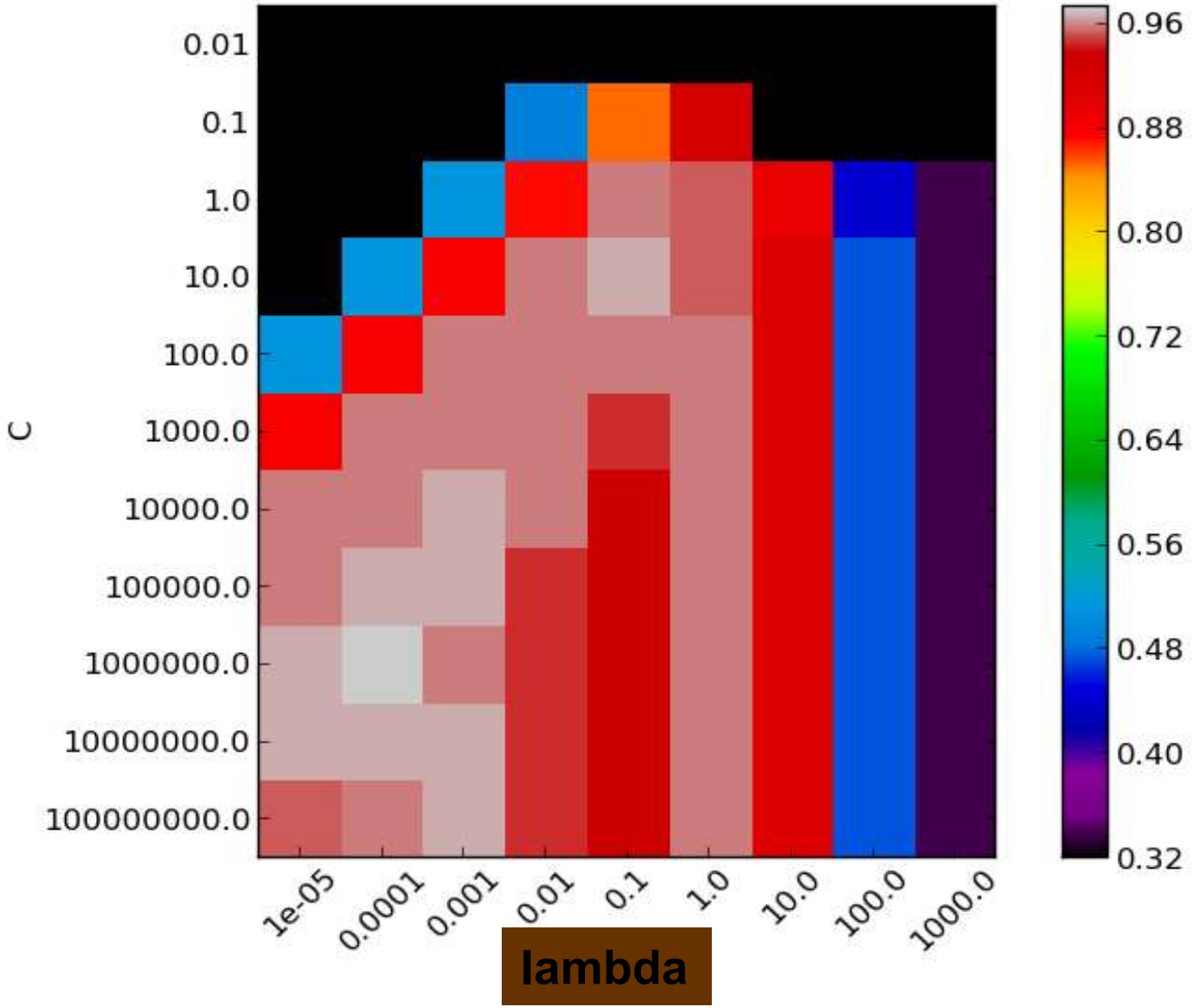


C=0.5, λ=0.05



Parámetros del *Kernel* RBF

Mapa de validación cruzada



http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

SVM: Aspectos Prácticos

- A la hora de clasificar es conveniente seguir esta guía:

– Escalar:

› Es recomendable escalar todas las características al rango [-1,1] o [0,1]

› De esta forma se evita que una característica domine a las demás y se evitan dificultades numéricas

– Selección del modelo

› En general se suele utilizar la función de base radial como una primera elección. La razón fundamental es que la función núcleo tiene un único parámetro y además el proceso de optimización tiene menos dificultades numéricas

SVM: Aspectos Prácticos

– **Selección del modelo**

› **La selección de los parámetros de las funciones núcleo y el parámetro *C* del caso no lineal se hace mediante una validación cruzada.**

› **En el caso de funciones de base radial se suele seleccionar los conjuntos de prueba:**

$$C \in \left\{10^{-3}, 10^{-2}, ..., 10^{3}\right\} \text{ y } \lambda \in \left\{10^{-3}, 10^{-2}, ..., 10^{3}\right\}$$

SVM: Clasificación Multiclase

- Clasificación multiclase:

– Se construye una SVM por clase.

› El problema a resolver es el de esa clase contra el resto.

– Se elige la clase a partir del máximo valor de los clasificadores

Dos implementaciones en scikit-learn:

SVC: uno-contra-uno, OAO

n(n- 1)/2 clasificadores construídos

SVCLineal: uno-contra-el-resto, OAA

n clasificadores entrenados y construidos

Ejemplo NO lineal (1): XOR

El problema del XOR resuelto con una SVM no lineal (sin núcleo).

– **Cjto. entrenamiento:** función XOR $x_1=(0,1)$, $x_2=(1,0)$ ambos pertenecen a C_1

$x_3=(0,0)$, $x_4=(1,1)$, ambos pertenecen a C_2

– **Signos deseados:** $y_1=1$, $y_2=1$, $y_3=-1$, $y_4=-1$

– **Funciones**

$$\phi : \phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$$

› **Puntos transformados**

$$(0,1) \rightarrow (0,1,0,0,\sqrt{2},1); \qquad (1,0) \rightarrow (1,0,0,\sqrt{2},0,1)$$

$$(0,0) \rightarrow (0,0,0,0,0,1); \qquad (1,1) \rightarrow (1,1,\sqrt{2},\sqrt{2},\sqrt{2},1)$$

› **Productos Escalares** $\Phi(\mathbf{x}_i)^T.\Phi(\mathbf{x}_j)$

$$(\Phi(\mathbf{x}_1)^T.\Phi(\mathbf{x}_1)) = (0,1,0,0,\sqrt{2},1)^T(0,1,0,0,\sqrt{2},1) = 4$$

$$(\Phi(\mathbf{x}_1)^T.\Phi(\mathbf{x}_2)) = (0,1,0,0,\sqrt{2},1)^T(1,0,0,\sqrt{2},0,1) = 1$$

.....

.....

$$(\Phi(\mathbf{x}_4)^T.\Phi(\mathbf{x}_4)) = (1,1,\sqrt{2},\sqrt{2},\sqrt{2},1)^T(1,1,\sqrt{2},\sqrt{2},\sqrt{2},1) = 9$$

Ejemplo NO lineal (2): XOR

$$(\Phi(\mathbf{x}_1)^T.\Phi(\mathbf{x}_1)) = (0,1,0,0,\sqrt{2},1)^T(0,1,0,0,\sqrt{2},1) = 4$$

$$(\Phi(\mathbf{x}_1)^T.\Phi(\mathbf{x}_2)) = (0,1,0,0,\sqrt{2},1)^T(1,0,0,\sqrt{2},0,1) = 1$$

.....

.....

$$(\Phi(\mathbf{x}_4)^T.\Phi(\mathbf{x}_4)) = (1,1,\sqrt{2},\sqrt{2},\sqrt{2},1)^T(1,1,\sqrt{2},\sqrt{2},\sqrt{2},1) = 9$$

El problema es ahora:

$$\left.\begin{array}{l} \max\limits_{\alpha} \sum\limits_{i=1} \alpha_i - \frac{1}{2}\sum\limits_{i,j=1} \alpha_i\alpha_j y_i y_j (\phi(\mathbf{x}_i)^T.\phi(\mathbf{x}_j)) \\ s.a. \ \sum\limits_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \ \text{i=1,...,n} \end{array}\right\} \Rightarrow$$

$$\max \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\begin{pmatrix} 4\alpha_1^2 + \alpha_1\alpha_2 - \alpha_1\alpha_3 - 4\alpha_1\alpha_4 + \\ \alpha_1\alpha_2 + 4\alpha_2^2 - \alpha_2\alpha_3 - 4\alpha_2\alpha_4 - \\ \alpha_3\alpha_1 - \alpha_3\alpha_2 + \alpha_3^2 + \alpha_3\alpha_4 - \\ 4\alpha_4\alpha_1 - 4\alpha_4\alpha_2 + \alpha_4\alpha_3 + 9\alpha_4^2 \end{pmatrix}$$

$$s.a. \quad \alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 = 0, \ \alpha_1 \geq 0, \ \alpha_2 \geq 0, \ \alpha_3 \geq 0, \ \alpha_4 \geq 0$$

Ejemplo no lineal (2): XOR

$$\left.\begin{aligned} \max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j (\phi(\mathbf{x}_i)^T . \phi(\mathbf{x}_j)) \\ s.a. \ \sum_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \ \text{i=1,...,n} \end{aligned}\right\} \Rightarrow$$

$$L(\alpha_1,\alpha_2,\alpha_3,\alpha_4,\lambda) = \alpha_1+\alpha_2+\alpha_3+\alpha_4 - \frac{1}{2}(4\alpha_1^2+\alpha_1\alpha_2-\alpha_1\alpha_3-4\alpha_1\alpha_4+$$

$$\alpha_1\alpha_2+4\alpha_2^2-\alpha_2\alpha_3-4\alpha_2\alpha_4-\alpha_3\alpha_1-\alpha_3\alpha_2+\alpha_3^2+\alpha_3\alpha_4$$

$$-4\alpha_4\alpha_1-4\alpha_4\alpha_2+\alpha_4\alpha_3+9\alpha_4^2)-\lambda(\alpha_1+\alpha_2-\alpha_3-\alpha_4)$$

Las cinco ecuaciones se obtienen derivando la función lagrangiana

$$1\text{-}4\alpha_1-\alpha_2-\alpha_3+4\alpha_4-\lambda=0$$

$$1\text{-}\alpha_1-4\alpha_2+\alpha_3+4\alpha_4-\lambda=0$$

$$1\text{+}\alpha_1+\alpha_2-\alpha_3-\alpha_4+\lambda=0$$

$$1\text{+}4\alpha_1+4\alpha_2-\alpha_3-9\alpha_4+\lambda=0$$

$$\alpha_1+\alpha_2=\alpha_3+\alpha_4$$

Solución: $\hat{\alpha} = (8/3,\ 8/3,\ 10/3,\ 6/3)$

Ejemplo no lineal (3): XOR

– ***w*** **óptimo:** $\hat{\mathbf{w}} = \sum_{i \in Sop} \hat{\boldsymbol{\alpha}}_i y_i \Phi(\mathbf{x}_i)$

$$\hat{\mathbf{w}} = (8/3)(0,1,0,0,\sqrt{2},1) + (8/3)(1,0,0,\sqrt{2},0,1)$$
$$-(10/3)(0,0,0,0,0,1) - 2(1,1,\sqrt{2},\sqrt{2},\sqrt{2},1) =$$
$$=(2/3,\ 2/3,\ -2\sqrt{2},\ (2/3)\sqrt{2},\ (2/3)\sqrt{2},\ 0)$$

– **Constante *w*0 óptima:**

$$\hat{w}_0 = 1\text{-}\hat{w}\Phi(\mathbf{x_i}),\ \text{con}\ \mathbf{x_i} \in \omega_1\ \text{y}\ \ \hat{\alpha}_i > 0$$

$$\hat{w}_0 = 1\text{-}(2/3,\ 2/3,\ -2\sqrt{2},\ (2/3)\sqrt{2},\ (2/3)\sqrt{2},\ 0)^{\mathrm{T}}(0,\ 1,\ 0,\ 0,\ \sqrt{2},\ 1) = -1$$

– **Clasificador óptimo:**

$$g(\mathbf{x}) = \hat{w}^T \Phi(\mathbf{x}) + \hat{w}_0$$

$$g(\mathbf{x}) = (2/3, 2/3, -2\sqrt{2}, (2/3)\sqrt{2}, (2/3)\sqrt{2}, 0)^T (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1) - 1 =$$
$$= 2/3x_1^2 + 2/3x_2^2 - 4x_1x_2 + 2/3x_1 + 2/3x_2 - 1$$

Ejemplo no lineal (4): XOR

El problema del XOR resuelto con una MVS no lineal (con núcleo).

– Cjto. entrenamiento: función XOR $x_1=(0,1)$, $x_2=(1,0)$ ambos pertenecen a C_1

$x_3=(0,0), x_4=(1,1)$, ambos pertenecen a C_2

– Signos deseados: $y_1=1$, $y_2=1$, $y_3=-1$, $y_4=-1$

– Funciones

$$\phi : \phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$$

-Productos escalares

$$\Phi(\mathbf{x}_i)^T.\Phi(\mathbf{x}_j) = (1+\mathbf{x}_i^T\mathbf{x}_j)^2 = k(\mathbf{x}_i,\mathbf{x}_j) =$$

$$= \left(1+2x_1x_1^{'}+2x_2x_2^{'}+x_1^2x_1^{'2}+2x_1x_2x_1^{'}x_2^{'}+x_2^2x_2^{'2}\right)$$

-La función de núcleo es entonces

$$k(\mathbf{x}_1,\mathbf{x}_1) = ((0,1)^T(0,1)+1)^2 = 4;\ k(\mathbf{x}_3,\mathbf{x}_1) = ((0,0)^T(0,1)+1)^2 = 1;$$

$$k(\mathbf{x}_1,\mathbf{x}_2) = ((0,1)^T(1,0)+1)^2 = 1;.....$$

$$k(\mathbf{x}_1,\mathbf{x}_3) = ((0,1)^T(0,0)+1)^2 = 1;.....$$

$$k(\mathbf{x}_1,\mathbf{x}_4) = ((0,1)^T(1,1)+1)^2 = 4;.....$$

$$k(\mathbf{x}_2,\mathbf{x}_1) = ((1,0)^T(0,1)+1)^2 = 1;$$

$$k(\mathbf{x}_2,\mathbf{x}_2) = ((1,0)^T(1,0)+1)^2 = 4;$$

$$k(\mathbf{x}_2,\mathbf{x}_3) = ((1,0)^T(0,0)+1)^2 = 1;$$

$$k(\mathbf{x}_2,\mathbf{x}_4) = ((1,0)^T(1,1)+1)^2 = 4; k(\mathbf{x}_4\mathbf{x}_4)) = ((1,1)^T(1,1)+1)^2 = 9$$

Ejemplo no lineal (5): XOR

$$\left.\begin{array}{l}\max\limits_{\alpha} \sum\limits_{i=1} \alpha_i - \frac{1}{2}\sum\limits_{i,j=1} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i \mathbf{x}_j) \\ s.a. \ \sum\limits_{i=1} \alpha_i y_i = 0, \ \ \alpha_i \geq 0, \ \text{i=1,...,n}\end{array}\right\} \Rightarrow$$

El problema es ahora:

$$\max \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\begin{pmatrix} 4\alpha_1^2 + \alpha_1\alpha_2 - \alpha_1\alpha_3 - 4\alpha_1\alpha_4 + \\ \alpha_1\alpha_2 + 4\alpha_2^2 - \alpha_2\alpha_3 - 4\alpha_2\alpha_4 - \\ \alpha_3\alpha_1 - \alpha_3\alpha_2 + \alpha_3^2 + \alpha_3\alpha_4 - \\ 4\alpha_4\alpha_1 - 4\alpha_4\alpha_2 + \alpha_4\alpha_3 + 9\alpha_4^2 \end{pmatrix}$$

$$s.a. \quad \alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 = 0, \ \alpha_1 \geq 0, \ \alpha_2 \geq 0, \ \alpha_3 \geq 0, \ \alpha_4 \geq 0$$

Las cinco ecuaciones son

$$1\text{-}4\alpha_1 - \alpha_2 - \alpha_3 + 4\alpha_4 - \lambda = 0$$

$$1\text{-}\alpha_1 - 4\alpha_2 + \alpha_3 + 4\alpha_4 - \lambda = 0$$

$$1+\alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 + \lambda = 0$$

$$1+4\alpha_1 + 4\alpha_2 - \alpha_3 - 9\alpha_4 + \lambda = 0$$

$$\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4$$

Con la misma solución: $\hat{\alpha} = (8/3, \ 8/3, \ 10/3, \ 6/3)$

Ejemplo no lineal (6): XOR

$$k(\mathbf{x}_1,\mathbf{x}_1) = ((0,1)^T(0,1)+1)^2 = 4$$
$$k(\mathbf{x}_1,\mathbf{x}_2) = ((0,1)^T(1,0)+1)^2 = 1$$
$$k(\mathbf{x}_1,\mathbf{x}_3) = ((0,1)^T(0,0)+1)^2 = 1$$
$$k(\mathbf{x}_1,\mathbf{x}_4) = ((0,1)^T(1,1)+1)^2 = 4$$
.....
$$k(\mathbf{x}_4\mathbf{x}_4)) = ((1,1)^T(1,1)+1)^2 = 9$$

$$\hat{w}_0 = 1-(8/3,\ 8/3,\ -10/3,\ -6/3)\begin{pmatrix}4\\1\\1\\4\end{pmatrix} = -1$$

para obtener el sesgo, hemos utilizado un sólo vector soporte positivo en el espacio transformado, el $(0,1,0,0,\sqrt{2},1)$

– **Constante w_0 óptima:**

$$\hat{w}_0 = 1-\sum_{i\in Sop}\hat{\boldsymbol{\alpha}}_i y_i k(\mathbf{x_i},\mathbf{x}_j)\ \ ,\ \text{con}\ \mathbf{x_i}\in\omega_1\ \text{y}\ \hat{\alpha}_i > 0$$

– **Clasificador óptimo:**

$$g(\mathbf{x}) = \sum_{i\in Sop}\hat{\boldsymbol{\alpha}}_i y_i k(\mathbf{x_i},\mathbf{x})\ + \hat{w}_0$$

donde $k(\mathbf{x_i},\mathbf{x}) = (1+\mathbf{x}_i^T\mathbf{x})^2$

Ejemplo no lineal (7): XOR

– **Clasificador óptimo:**

$$g(\mathbf{x}) = \sum_{i \in Sop} \hat{\boldsymbol{\alpha}}_i y_i k(\mathbf{x_i}, \mathbf{x'}_j) + \hat{w}_0$$

$$k(\mathbf{x}_i, \mathbf{x'}_j) = (1 + \mathbf{x}_i^T \mathbf{x'}_j)^2 = \left(1 + 2x_1 x_1^{'} + 2x_2 x_2^{'} + x_1^2 x_1^{'2} + 2x_1 x_2 x_1^{'} x_2^{'} + x_2^2 x_2^{'2}\right)$$

$$k((0,1), \mathbf{x'}_j) = (1 + (0,1)\begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix})^2 = (1 + x'_2)^2$$

$$k((1,0), \mathbf{x'}_j) = (1 + (1,0)\begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix})^2 = (1 + x'_1)^2$$

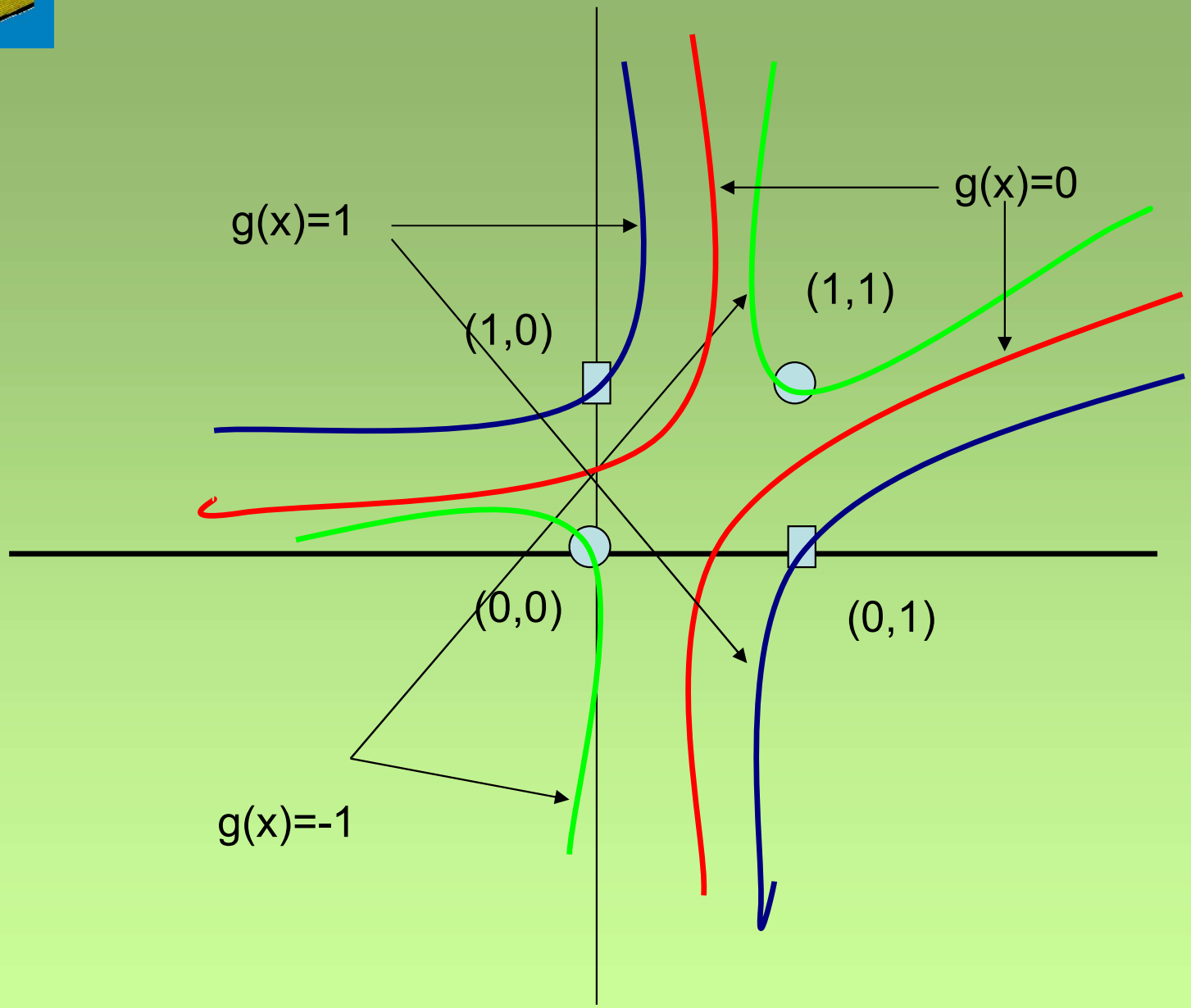
$$k((0,0), \mathbf{x'}_j) = (1 + (0,0)\begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix})^2 = 1$$

$$k((1,1), \mathbf{x'}_j) = (1 + (1,1)\begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix})^2 = (1 + x'_1 + x'_2)^2$$

$$g(\mathbf{x}) = (8/3)(x'_2 + 1)^2 + (8/3)((x'_1 + 1)^2 - (10/3) - (6/3)(x'_1 + x'_2 + 1)^2 - 1 =$$

$$= (2/3)(x'_1 - 1/2)^2 + (x'_2 - 1/2)^2 - 6(x'_1 - 1/2)(x'_2 - 1/2) - 1/2$$

Ejemplo no lineal (7): XOR



Frontera de decisión de la metodología No lineal SVM

Resumen: hard-margen en SVM lineal

Hard-margen en SVM lineal

Primal

$$\min_{\mathbf{w},w_0} \frac{\langle \mathbf{w},\mathbf{w} \rangle}{2}$$

$$\text{s.a. } y_i\left(\langle \mathbf{w},\mathbf{x}_i \rangle + w_0\right) \geq 1,\ i \in \{1,2,\ldots,n\}$$

Dual

$$\max_{\boldsymbol{\alpha},\lambda} L(\boldsymbol{\alpha},\lambda) = \sum_{i\in S}\alpha_i - \frac{1}{2}\sum_{i,j\in S}\alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i,\mathbf{x}_j \rangle - \lambda\left(\sum_{i\in S}\alpha_i y_i\right)$$

$$\text{s.a. } \lambda \geq 0,\ \ \alpha_i \geq 0,\ i = 1,\ldots,n$$

$$\mathbf{w} = \sum_{i\in Sop}\alpha_i y_i \mathbf{x}_i \quad \hat{w}_0 = 1 - \sum_{j\in Sop}\hat{\alpha}_j y_j (\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_i), \quad \text{con } \mathbf{x}_i \in \omega_1 \text{ y } \ \hat{\boldsymbol{\alpha}}_i > 0$$

Resumen: soft-margen en SVM lineal

Soft-margen en SVM lineal

Primal

$$\min_{\mathbf{w},w_0} \frac{\langle \mathbf{w},\mathbf{w} \rangle}{2} + C\sum_{i=1}^{n} \varepsilon_i$$

$$\text{s.a. } y_i\left(\langle \mathbf{w},\mathbf{x}_i \rangle + w_0\right) \geq 1-\varepsilon_i,\ i \in \{1,2,\ldots,n\}$$

Dual

$$\max_{\boldsymbol{\alpha},\lambda} L(\boldsymbol{\alpha},\lambda) = \sum_{i\in S}\alpha_i - \frac{1}{2}\sum_{i,j\in S}\alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i,\mathbf{x}_j \rangle - \lambda\left(\sum_{i\in S}\alpha_i y_i\right)$$

$$\text{s.a. } \lambda \geq 0,\ \ C \geq \alpha_i \geq 0,\ i=1,\ldots,n$$

$$\mathbf{w} = \sum_{i\in Sop}\alpha_i y_i \mathbf{x}_i \quad \hat{w}_0 = 1 - \sum_{j\in Sop}\hat{\alpha}_j y_j (\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_i),\quad \text{con } \mathbf{x}_i \in \omega_1 \text{ y } \hat{\boldsymbol{\alpha}}_i > 0$$

Resumen: hard-margen SVM no-lineal

- Hard-margen SVM no-lineal
- En general se trabaja en el dual

Primal

$$\min_{\mathbf{w},w_0} \frac{\langle \mathbf{w},\mathbf{w} \rangle}{2}$$

$$\text{s.a. } y_i\left(\langle \mathbf{w},\Phi(\mathbf{x}_i)\rangle + w_0\right) \geq 1,\ i \in \{1,2,\ldots,n\}$$

Dual

$$\max_{\boldsymbol{\alpha},\lambda} L(\boldsymbol{\alpha},\lambda) = \sum_{i\in Sop} \alpha_i - \frac{1}{2}\sum_{i,j\in Sop} \alpha_i\alpha_j y_i y_j \, \mathrm{k}(\mathbf{x}_i,\mathbf{x}_j) - \lambda\left(\sum_{i\in Sop} \alpha_i y_i\right)$$

$$\text{s.a. } \lambda \geq 0,\ \ \alpha_i \geq 0,\ i = 1,\ldots,n$$

$$\mathrm{g}(\mathbf{x}) = \left(\sum_{i\in \mathrm{Sop}} \alpha_i y_i K(\mathbf{x}_i,\mathbf{x})\right) + w_0 \qquad \hat{\mathrm{w}}_0 = 1 - \sum_{i\in Sop} \hat{\alpha}_i y_i k(\mathbf{x_i},\mathbf{x_j}),\ \text{con } \mathbf{x_i} \in \omega_1 \text{ y } \hat{\alpha}_i > 0$$

Resumen: soft-margen SVM no-lineal

- Soft-margen SVM no-lineal
- En general se trabaja en el dual

Primal

$$\min_{\mathbf{w},w_0} \frac{\langle \mathbf{w},\mathbf{w} \rangle}{2} + C\sum_{i=1}^{n} \varepsilon_i$$

$$\text{s.a. } y_i\left(\langle \mathbf{w}, \Phi(\mathbf{x}_i)\rangle + w_0\right) \geq 1-\varepsilon_i,\ i \in \{1,2,\ldots,n\}$$

Dual

$$\max_{\boldsymbol{\alpha},\lambda} L(\boldsymbol{\alpha},\lambda) = \sum_{i\in Sop} \alpha_i - \frac{1}{2}\sum_{i,j\in Sop} \alpha_i\alpha_j y_i y_j \, \mathrm{k}(\mathbf{x}_i,\mathbf{x}_j) - \lambda\left(\sum_{i\in Sop} \alpha_i y_i\right)$$

$$\text{s.a. } \lambda \geq 0,\ \ \mathrm{C} \geq \alpha_i \geq 0,\ i = 1,\ldots,n$$

$$g(\mathbf{x}) = \left(\sum_{i\in \mathrm{Sop}} \alpha_i y_i K(\mathbf{x}_i,\mathbf{x})\right) + w_0 \qquad \hat{\mathrm{w}}_0 = 1 - \sum_{i\in Sop} \hat{\alpha}_i y_i k(\mathbf{x_i},\mathbf{x_j}),\ \text{con } \mathbf{x_i} \in \omega_1 \text{ y } \hat{\alpha}_i > 0$$

SVM: Resumen

- **Los clasificadores no lineales:**
  - **Permiten trabajar con fronteras de decisión no lineales**
- **Los clasificadores presentados en este tema se basan en:**

1. Realizar transformaciones no lineales de las características.

2. Aplicar a los datos transformados un clasificador lineal

- **Un primer problema:**
  - **Si el número de transformaciones es muy grande el clasificador sufre de sobreajuste.**
- **Solución:**
  - **Clasificador cuadrático: Regularización**
  - **SVM: Máximo margen.**
- **Un segundo problema:**
  - **¿Cuál debe ser la transformación no lineal de los datos?**
- **Una Solución:**
  - **Elegir una de las conocidas (polinomial, funciones de base radial, etc)**

Selección de parámetros: Resumen

Siempre hay que fijar un valor para C

Si $C \ll 1$, el ajuste en la muestra puede ser malo

Si $C \gg 1$, hay un riesgo de sobreajuste en la muestra; también malo!!

También hay que escoger σ si además se usa un núcleo Gaussiano

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

Si $\sigma \gg 1$, se tienen gaussianas muy extendidas, que no diferencian entre datos.

Si $\sigma \ll 1$, las Gaussianas aproximan deltas de Dirac, y hay un riesgo de sobreajuste en la muestra de entrenamiento

Selección de parámetros: Resumen II

Posibles enfoques para obtener $C,\ \sigma$:

i) Fijar conjuntos de validación

ii) Generar diferentes modelos y evaluarlos sobre dichos conjuntos

Opción 1: búsqueda en rejilla sobre rectángulo

$$[C_{\min}, C_{\max}] \times [\sigma_{\min}, \sigma_{\max}]$$

Opción 2: búsqueda evolutiva a partir de $C_0,\ \sigma_0$ iniciales

Alternativa: Usar técnicas analíticas

CONCLUSIONES

- Las SVMs son una de las metodologías más eficaces en clasificación y modelización.

- Ofrecen una teoría sólida y elegante con métodos de construcción eficaces, buen software de acceso público (LivSVM)

-Juntan ideas diversas e importantes: márgenes en clasificación, optimización Lagrangiana, espacios de Hilbert reproductores.

- No están exentas de problemas:

-. Explotación costosa en test:

$$F(\mathbf{x}; w^*) = \sum_{\{\alpha_p^* > 0\}} \alpha_p^* y_p \, \mathrm{k}(\mathbf{x_p}, \mathbf{x})$$

$$\Rightarrow \text{coste } \left|\{\alpha_p^* > 0\}\right| = \Theta(N) \text{ en general}$$

-. Construcción imposible si N muy grande

-. Entrenamiento esencialmente por lotes, en modo *batch*

AREA DE INVESTIGACIÓN

Área de gran actividad investigadora

Métodos de entrenamiento que produzcan modelos con parsimonia (sparsity)

Si d=1, trabajo con núcleo lineal: $K(\mathbf{x}, \mathbf{x}_0) = \mathbf{x}^T \mathbf{x}_0$

Ejemplos: Microarrays de ADN,

Recuperación de información en la web

Recomendación:

Cuando N es enorme: usar núcleos lineales (si se puede), trabajar en el primal (por ejemplo, mediante descenso por gradiente), aprovechar la tecnología (máquinas multinúcleos, paralelización)

Referencias

-N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, 2000.

-B. Schölkopf and A. J. Smola. Learning with Kernels. MIT Press, 2002

-Ch. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2 (1998), 121–167

C. Bishop, Pattern Recognition and Machine Learning, Springer,

Berwick, An idiot's guide to support vector machines

E. Kim, Everything you wanted to know about the kernel trick (but were too afraid to ask)

M. Law, A simple introduction to support vector machines

W.S. Noble, What is a support vector machine?, Nature Biotechnology, 24(12):1565-1567, 2006

Referencias: Multiplicadores de Lagrange

- C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006. Appendix E. Lagrange Multipliers (p. 707-710).
- D. Klein, Lagrange multipliers without permanent scarring
- Wikipedia, Lagrange multiplier

MODELOS COMPUTACIONALES: CUARTO CURSO DEL GRADO DE ING. INFORMÁTICA EN COMPUTACION

Introducción a las Máquinas de Vectores Soporte, SVM

GRACIAS POR SU ATENCIÓN 

César Hervás-Martínez
Pedro A. Gutiérrez Peña

Grupo de Investigación AYRNA

Departamento de Informática y Análisis Numérico
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es
pagutierrez@uco.es

2022-2023