



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

Introducción a las Máquinas de Vectores Soporte, SVM

César Hervás-Martínez
Grupo de Investigación AYRNA

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

2019-2020



INDICE



Clasificadores Lineales

SVMs Lineales

SVMs No Lineales

Aplicación de las SVMs

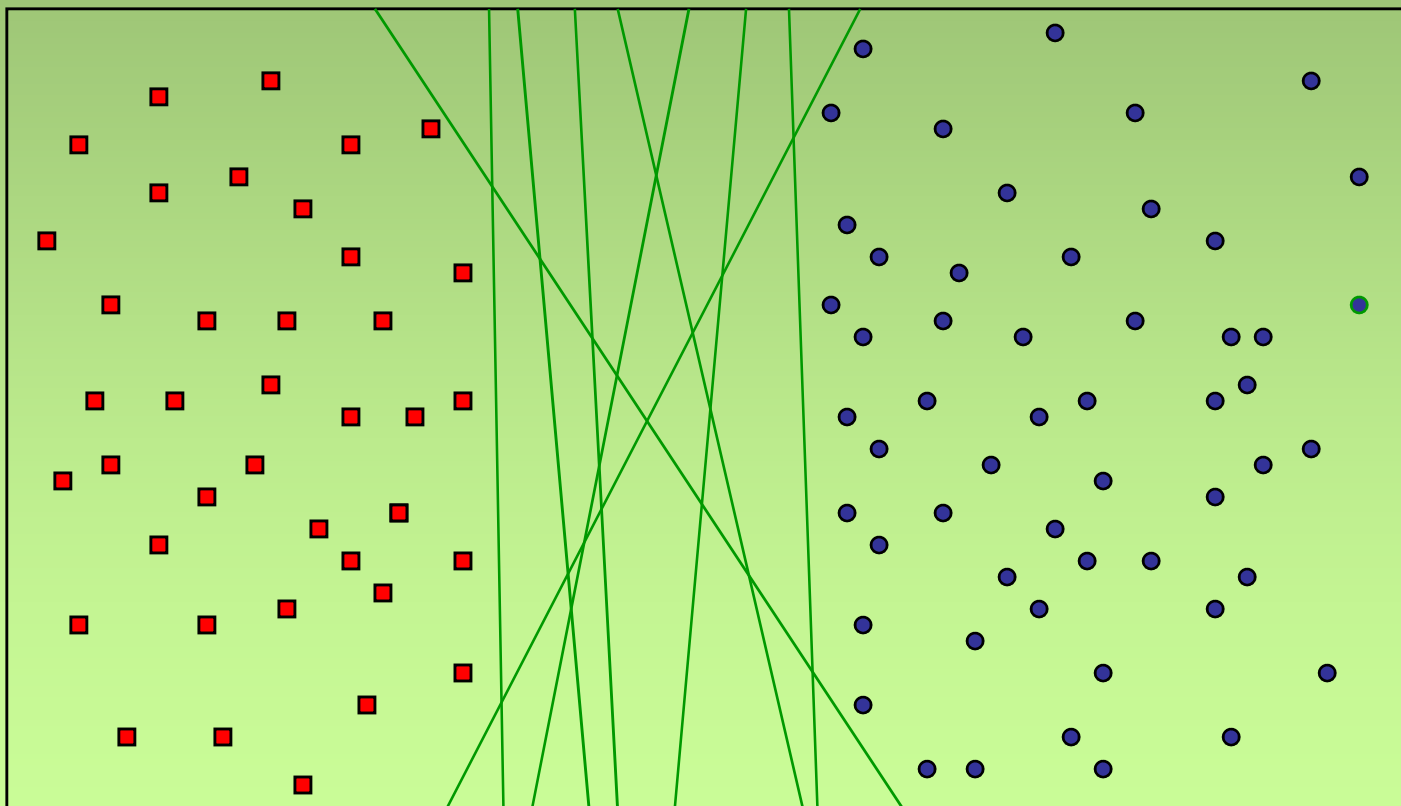
Conclusiones y Futuro



INTRODUCCION



Posibles fronteras de decisión para datos linealmente separables





INTRODUCCION



Objetivo

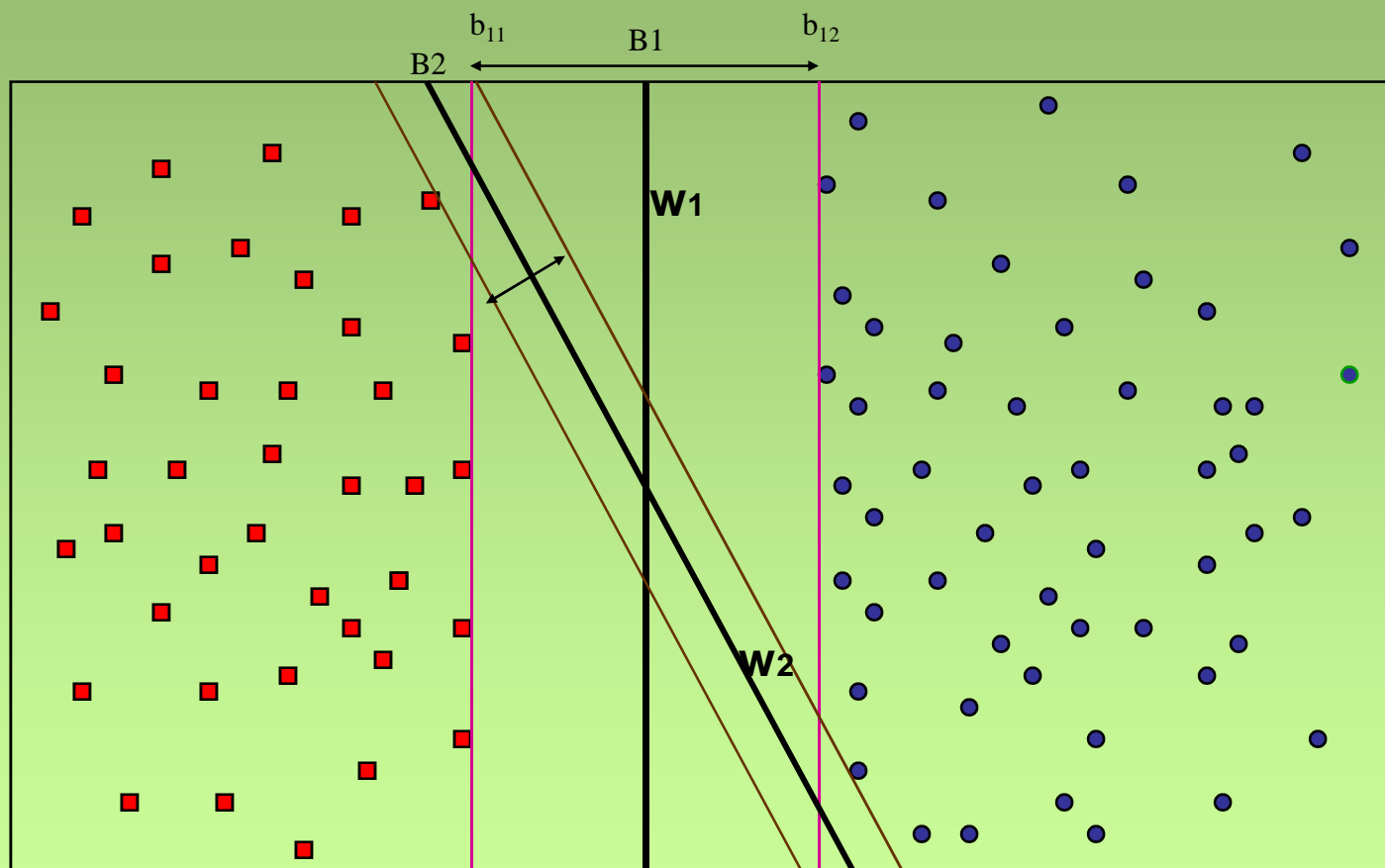
- Dado un conjunto de ejemplos de entrenamiento **construir un hiperplano “w” como superficie de decisión. De tal forma que la separación de las dos clases sea máxima (principio de generalización)**



INTRODUCCION



Diferentes márgenes de decisión





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 ahora es

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

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

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

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

Y las reglas de decisión son ahora

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

Si $-\ln P(\mathbf{x} / C_1) + \ln P(\mathbf{x} / C_2) < 0$ 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 temas anteriores se ha visto como aproximar la frontera de decisión óptima para un clasificador binario

$$h(\mathbf{x}) = P(\mathbf{w}_1, \mathbf{x}) - P(\mathbf{w}_2, \mathbf{x})$$

mediante:

- Una función lineal:

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

- 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.



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?

- Una idea simple:

- La forma del clasificador lineal es:

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

- Introduzcamos los términos de grado 2:

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

Este tipo de función es una función discriminante cuadrática.

- Podemos seguir con el grado 3:

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

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



Clasificador Polinomial II



- Las funciones discriminantes polinomiales anteriores tienen una forma común:

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

- **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:

$$\phi_1(\mathbf{x}) = x_1^2, \quad \phi_2(\mathbf{x}) = x_1x_2, \quad \phi_3(\mathbf{x}) = x_2x_1, \quad \phi_4(\mathbf{x}) = x_2^2, \quad \phi_5(\mathbf{x}) = x_1, \quad \phi_6(\mathbf{x}) = x_2$$

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:

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

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

- Las funciones ϕ pueden ser polinomiales o de otro tipo:
 - Gaussianas, 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

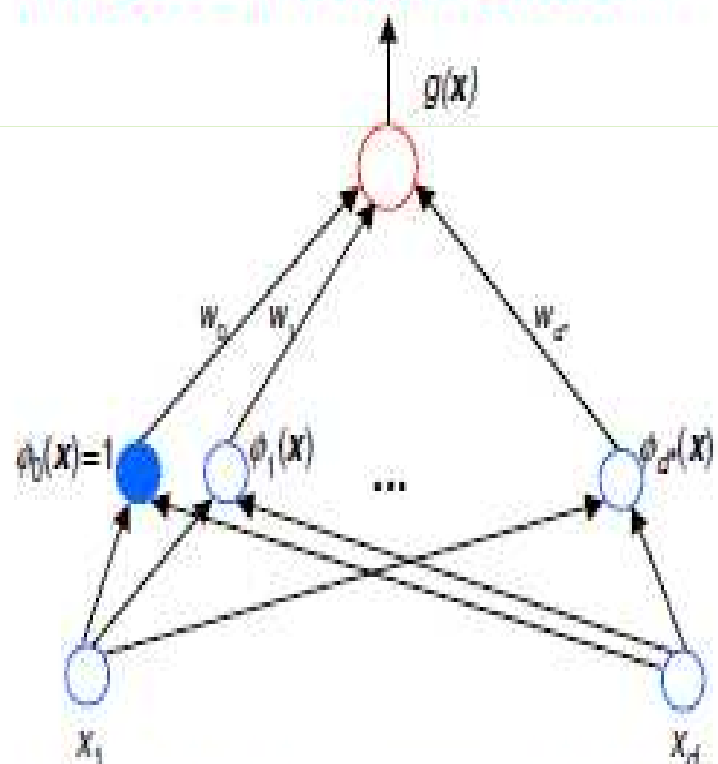


Entonces una notación matricial se puede definir en la forma

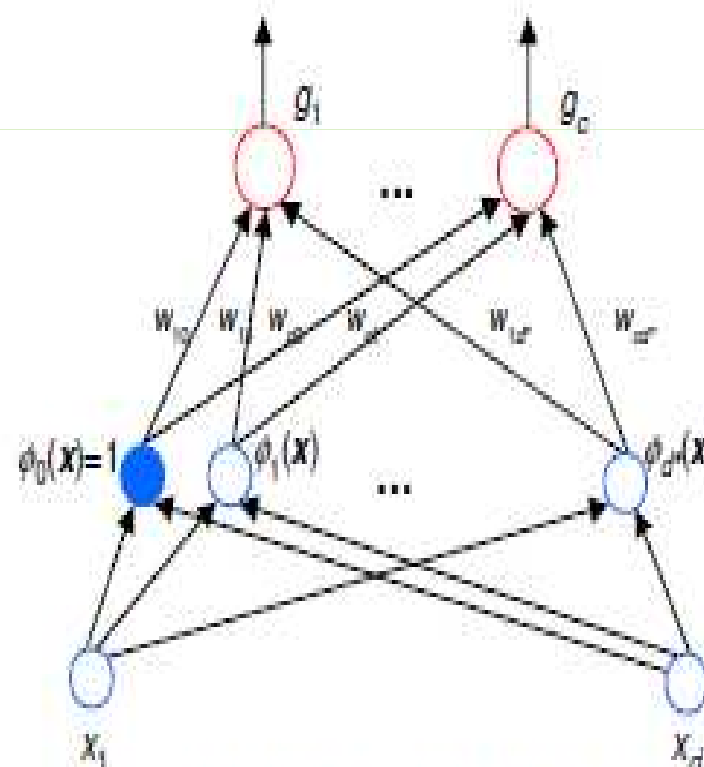
$$g(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}), \text{ donde } \mathbf{w}^T = (w_0, w_1, w_2, \dots, w_M),$$

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

• Representación gráfica:



FDLG generalizada: Representación gráfica (dos clases)



FDLG generalizada: Representación gráfica (c clases)



FDLG: Entrenamiento



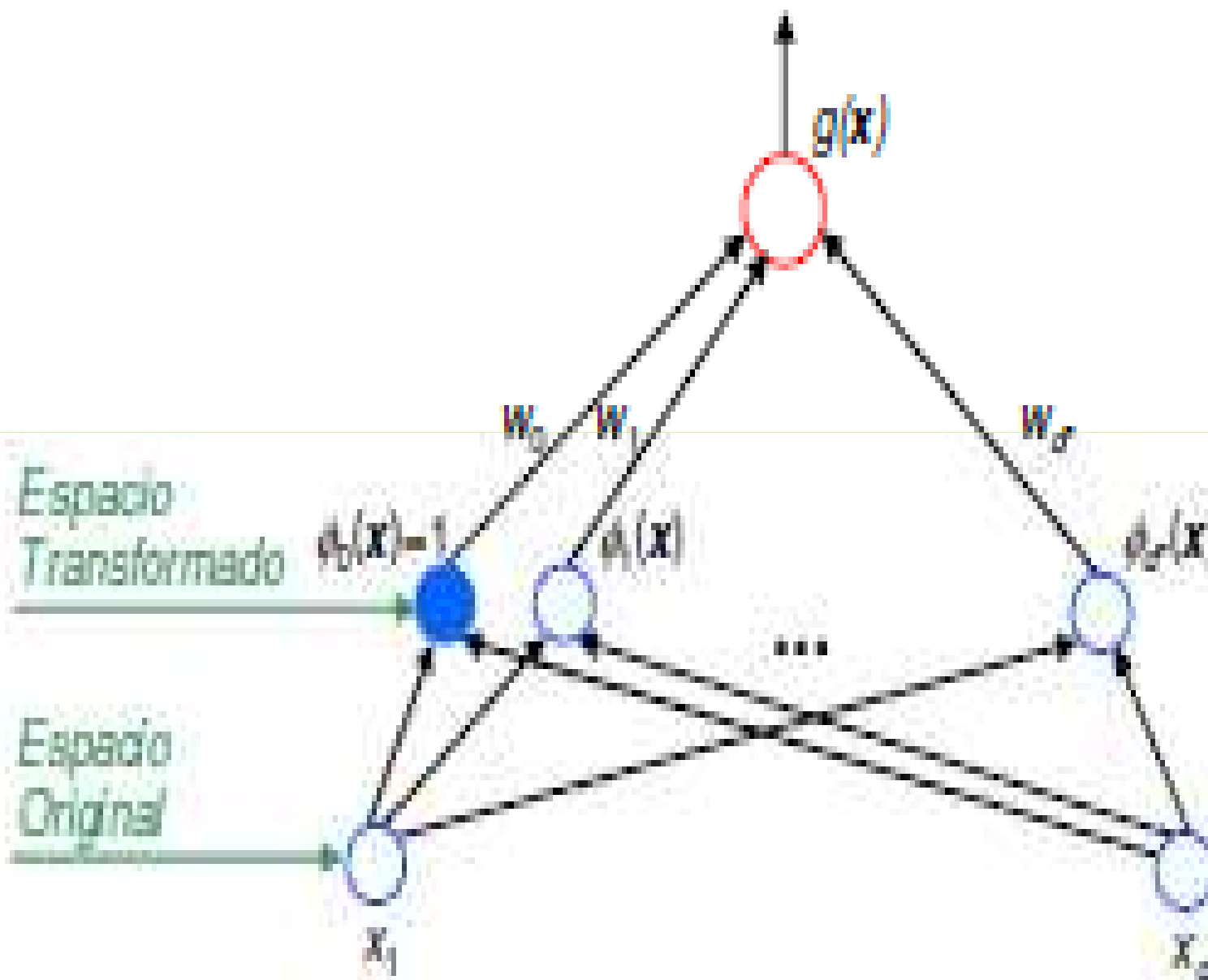
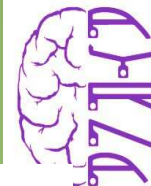
- Una observación crucial:
 - Las funciones ϕ transforman el espacio original de variables independientes en un nuevo espacio de características.
 - En este nuevo espacio el problema es determinar una función discriminante lineal.
- Por tanto, el **esquema de aprendizaje** es:

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:

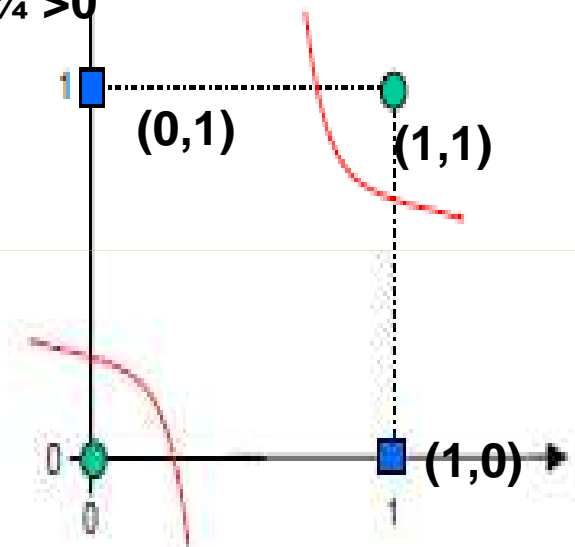
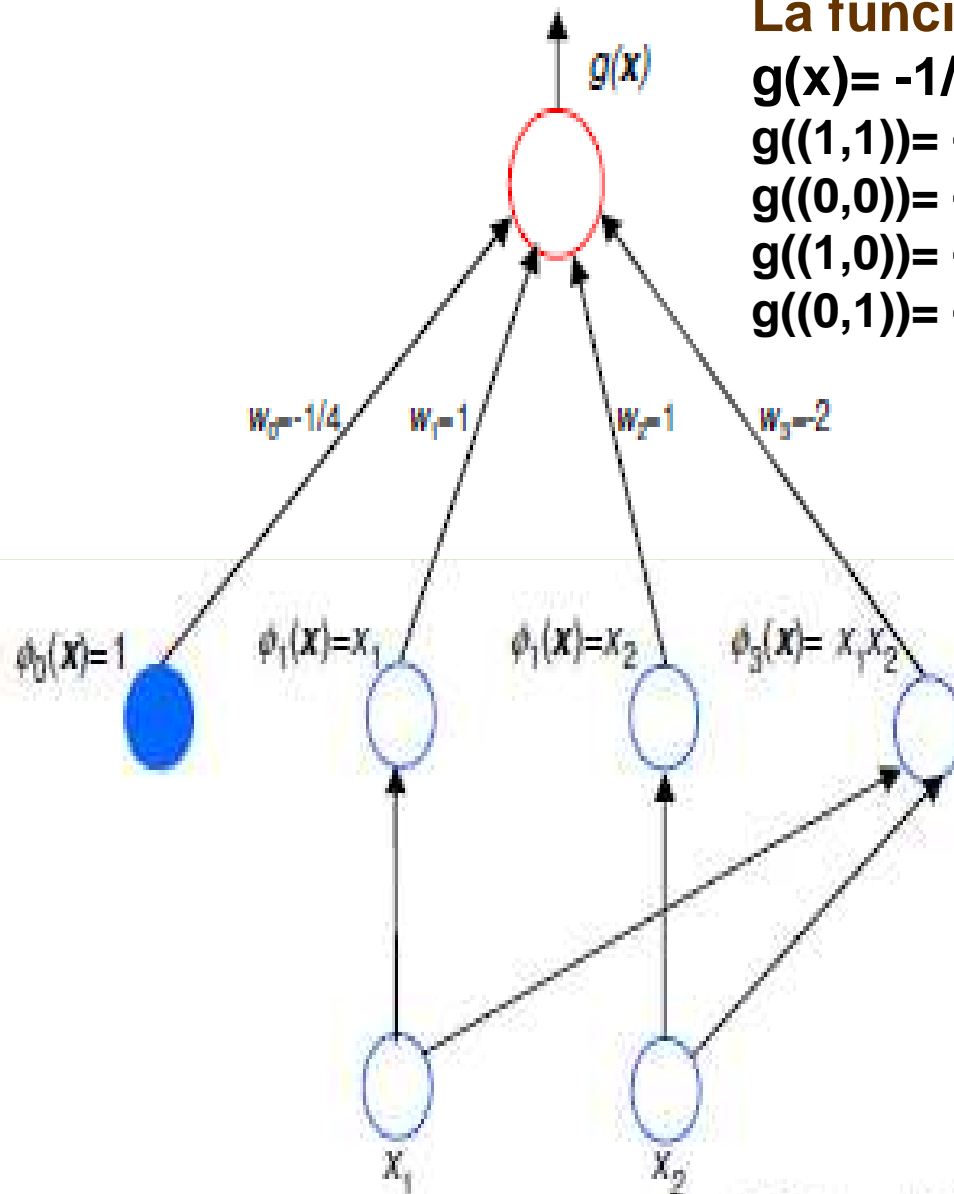
$$g(x) = -1/4 - 2x_1x_2 + x_1 + x_2$$

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

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

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

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



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: 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 tienen algoritmos de entrenamiento lineales para determinar los pesos de los modelos.



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

CASO: Clases linealmente separables

- ❑ Muestras de entrenamiento $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$
- ❑ Dos clases $y_i = +1, y_i = -1$
- ❑ Función de decisión

$$\mathbf{w}\mathbf{x} + w_0 = 0$$

$$\mathbf{w}\mathbf{x}_i + w_0 \geq 0 \quad \text{para } y_i = +1$$

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

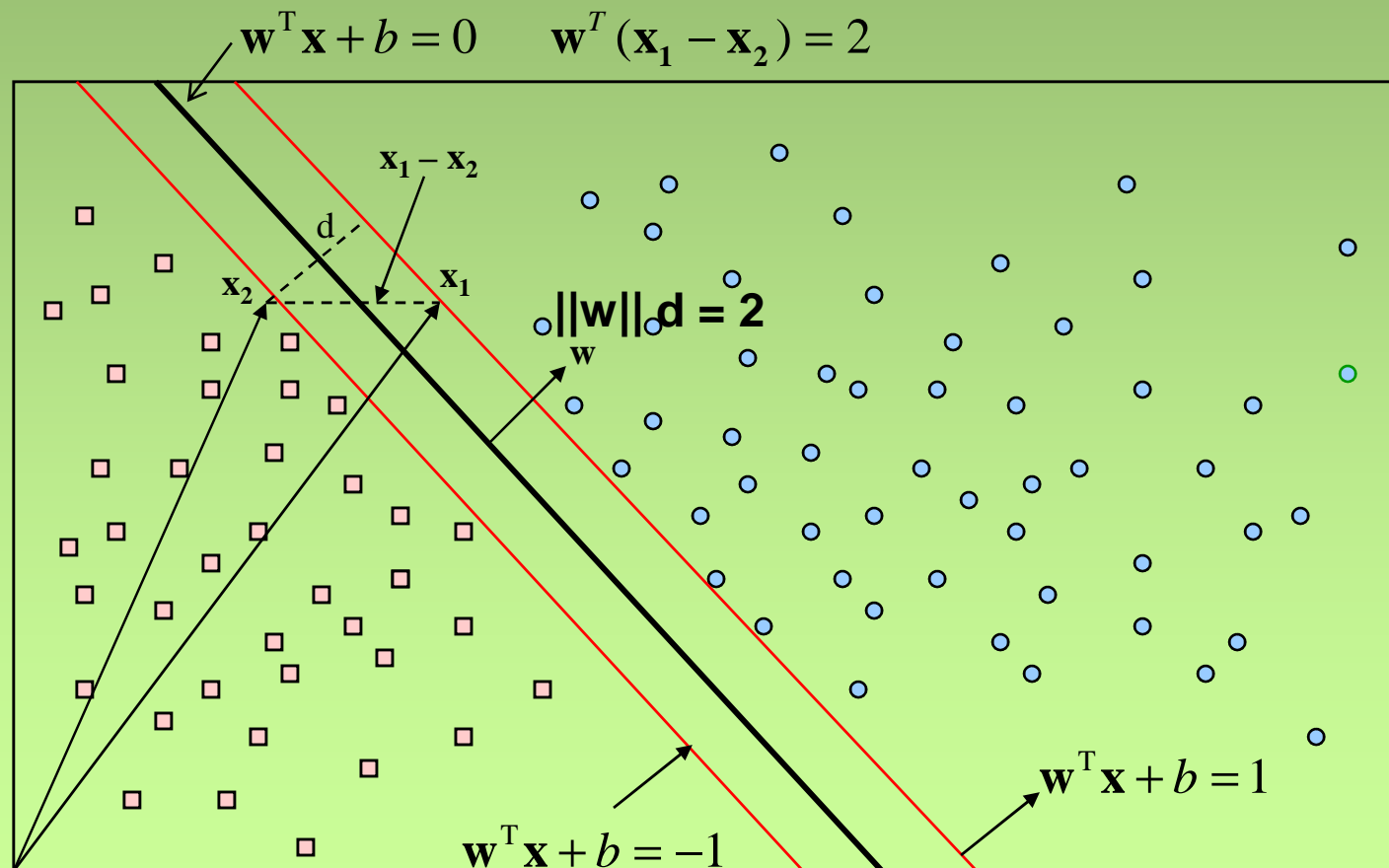
El entrenamiento busca estimar los parámetros o pesos \mathbf{w} , w_0



MARCO TEÓRICO



Frontera de decisión y margen de los modelos SVM





MARCO TEÓRICO



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 $\|\mathbf{w}\| d = 2$,

$$\text{luego } d = 2 / \|\mathbf{w}\|$$

Donde d es el margen de separación a maximizar

□ Entonces la función de optimización (minimizar en este caso) es una función monótona en \mathbf{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 $wx + w_0 = 0$ y los subespacios son

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

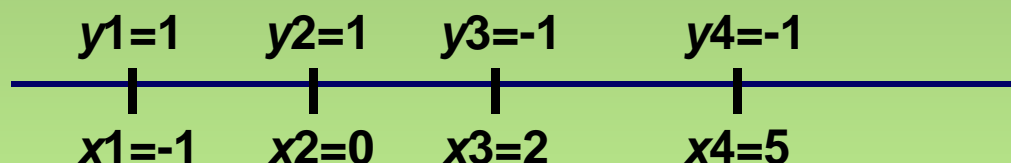
$$wx_i + w_0 < 0 \quad \text{para } y_i = -1$$

Clasificación unidimensional: $g(x) = wx + w_0$

– **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:

$$+ (wx_1 + w_0) > 0: \quad -w + w_0 > 0; \quad w_0 > w$$

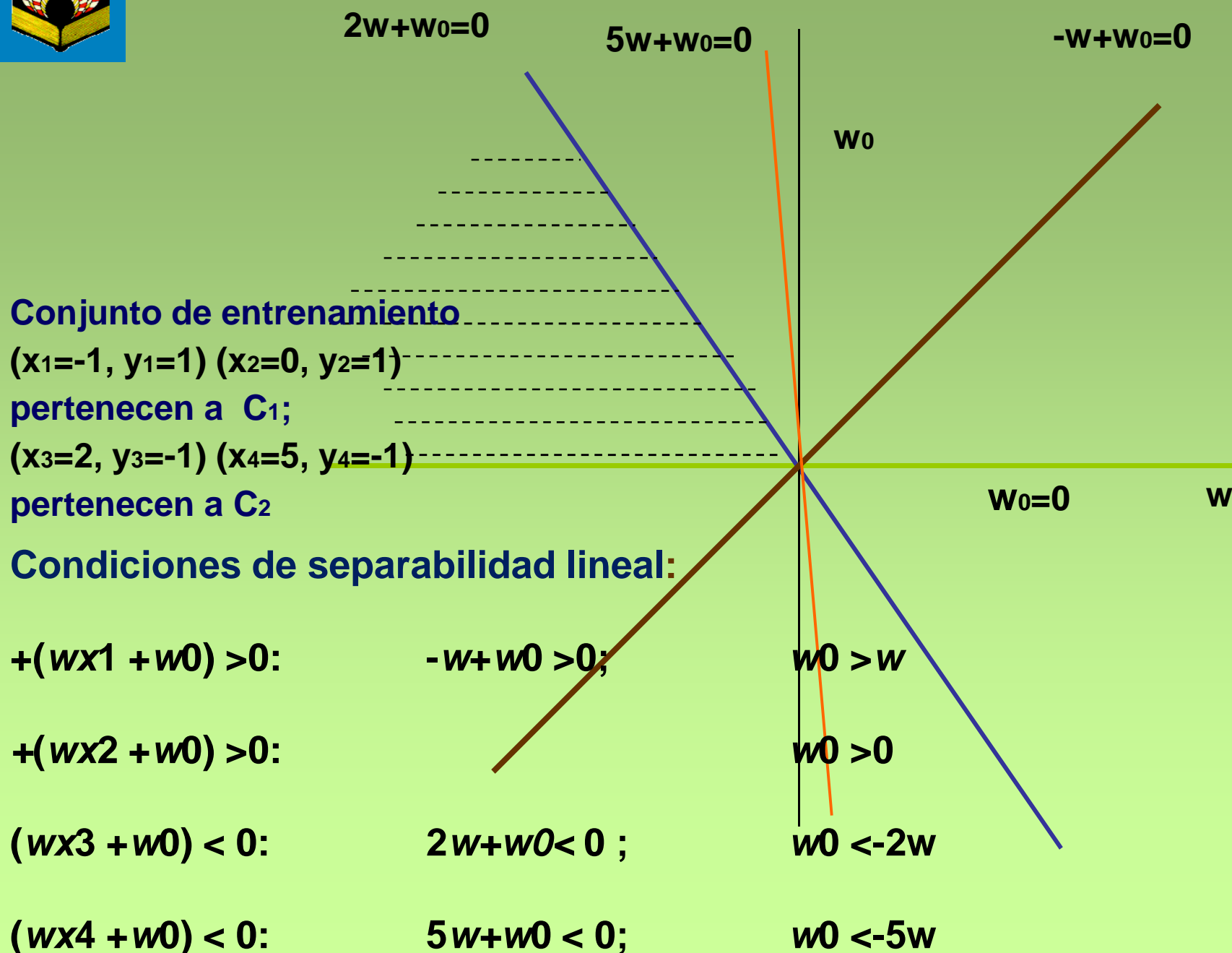
$$+ (wx_2 + w_0) > 0: \quad w_0 > 0$$

$$(wx_3 + w_0) < 0: \quad 2w + w_0 < 0; \quad w_0 < -2w$$

$$(wx_4 + w_0) < 0: \quad 5w + w_0 < 0; \quad w_0 < -5w$$



Representación Gráfica del Ejemplo Inicial





SVM: (CL). Formulación Inicial



- En el caso de separabilidad lineal:

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

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

- Por tanto para **calcular el margen** debemos calcular la menor distancia de los puntos del conjunto de entrenamiento a la frontera.

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



SVM: (CL).Formulación Inicial



Los valores de \mathbf{w} y w_0 representaran la frontera lineal que separe los puntos del conjunto de entrenamiento pertenecientes a las dos clases de patrones,

$$y_i (\mathbf{w}^T \mathbf{x}_i + w_0) > 0, i = 1 \dots n$$

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

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

$$s.a. \quad y_i (\mathbf{w}^T \mathbf{x}_i + w_0) > 0, \quad i = 1, \dots, n$$



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

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

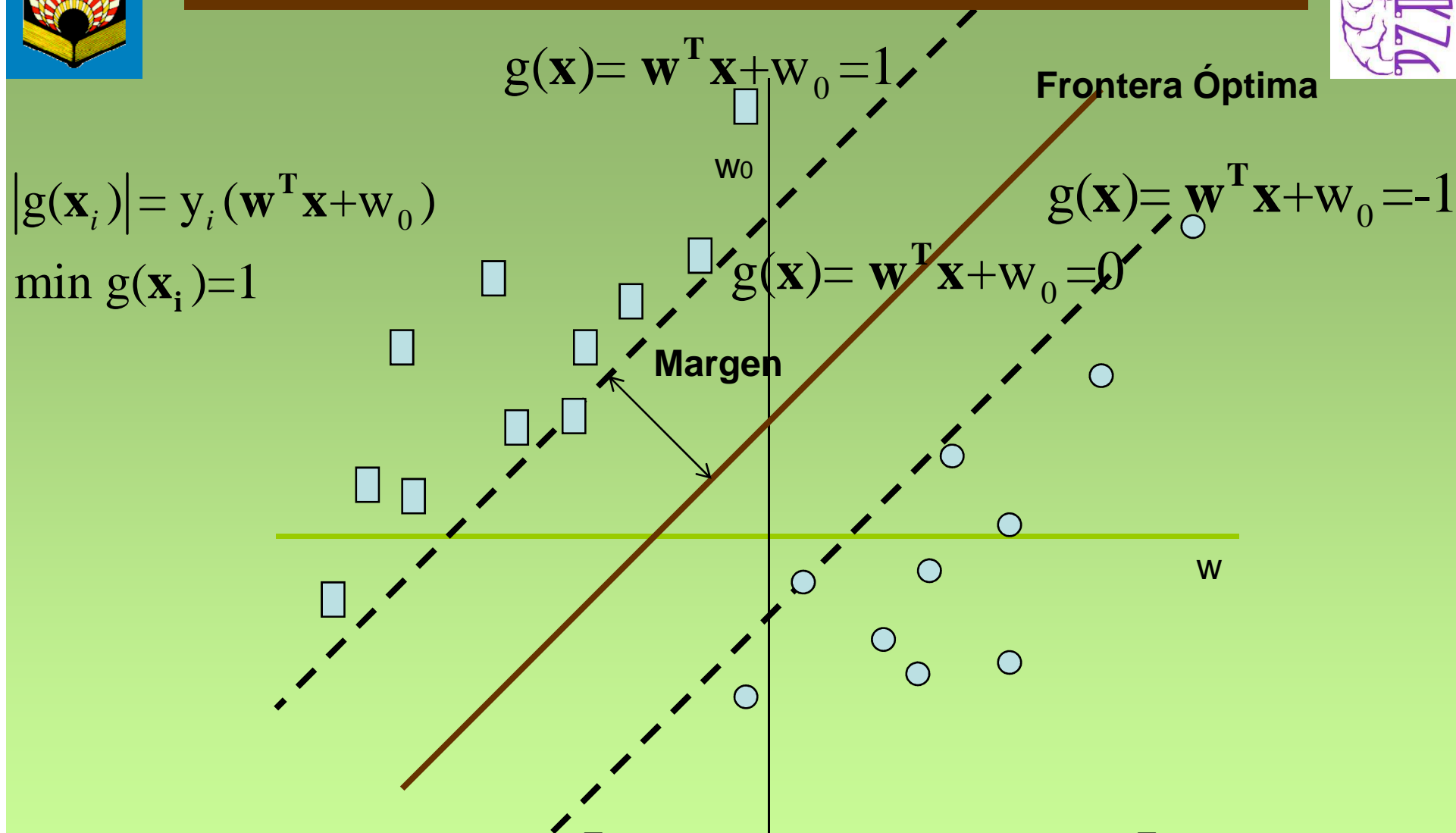
- Esto hace que el problema de optimización sea complicado de resolver.

- **Solución:**

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



Normalización



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:

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

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

Para maximizar el margen, 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

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w}, \\ \text{s.a.} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1, \quad i = 1 \dots n \end{aligned}$$

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



Normalización del Ejemplo Inicial



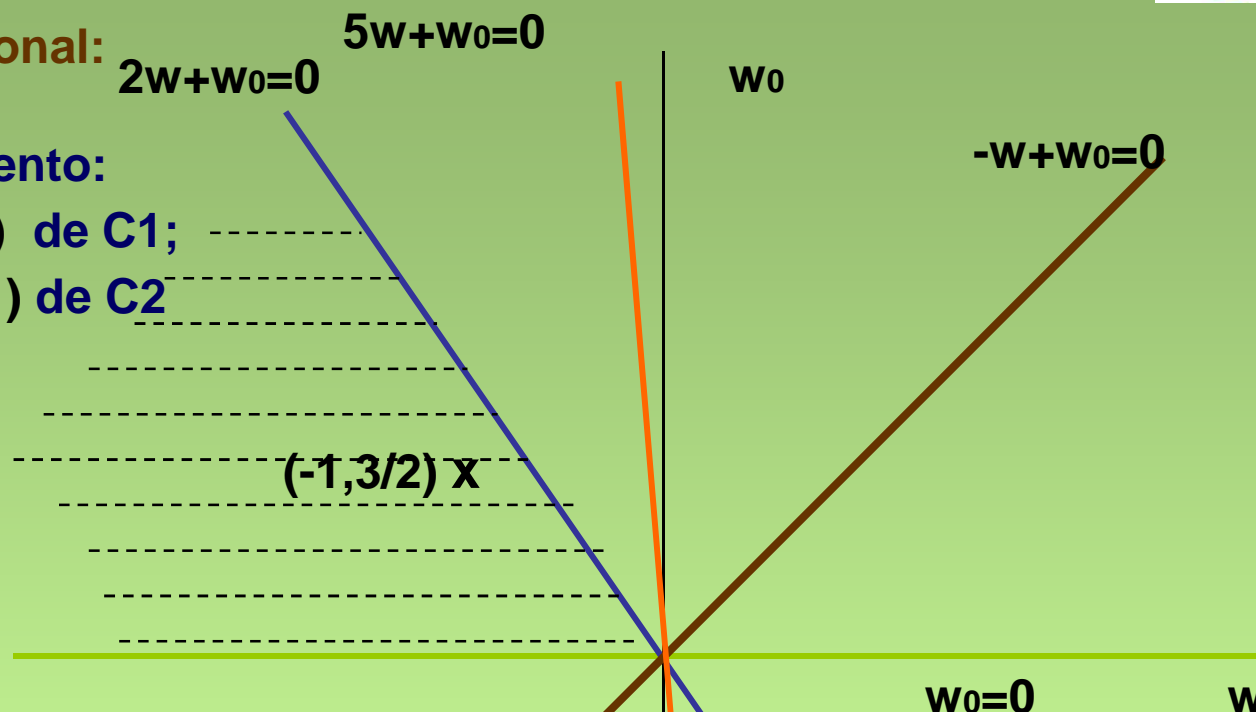
Clasificación unidimensional:

$$g(x) = wx + w_0$$

– **Conjunto de entrenamiento:**

$(x_1 = -1, y_1 = 1)$ $(x_2 = 0, y_2 = 1)$ de C_1 ;

$(x_3 = 2, y_3 = -1)$ $(x_4 = 5, y_4 = -1)$ de C_2



- **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 $(x_1=-1, y_1=1)$ $(x_2=0, y_2=1)$ $(x_3=2, y_3=-1)$ $(x_4=5, y_4=-1)$

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

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

El menor valor es 1/2 y ocurre para $x_3 = 2$.

– 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:

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

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

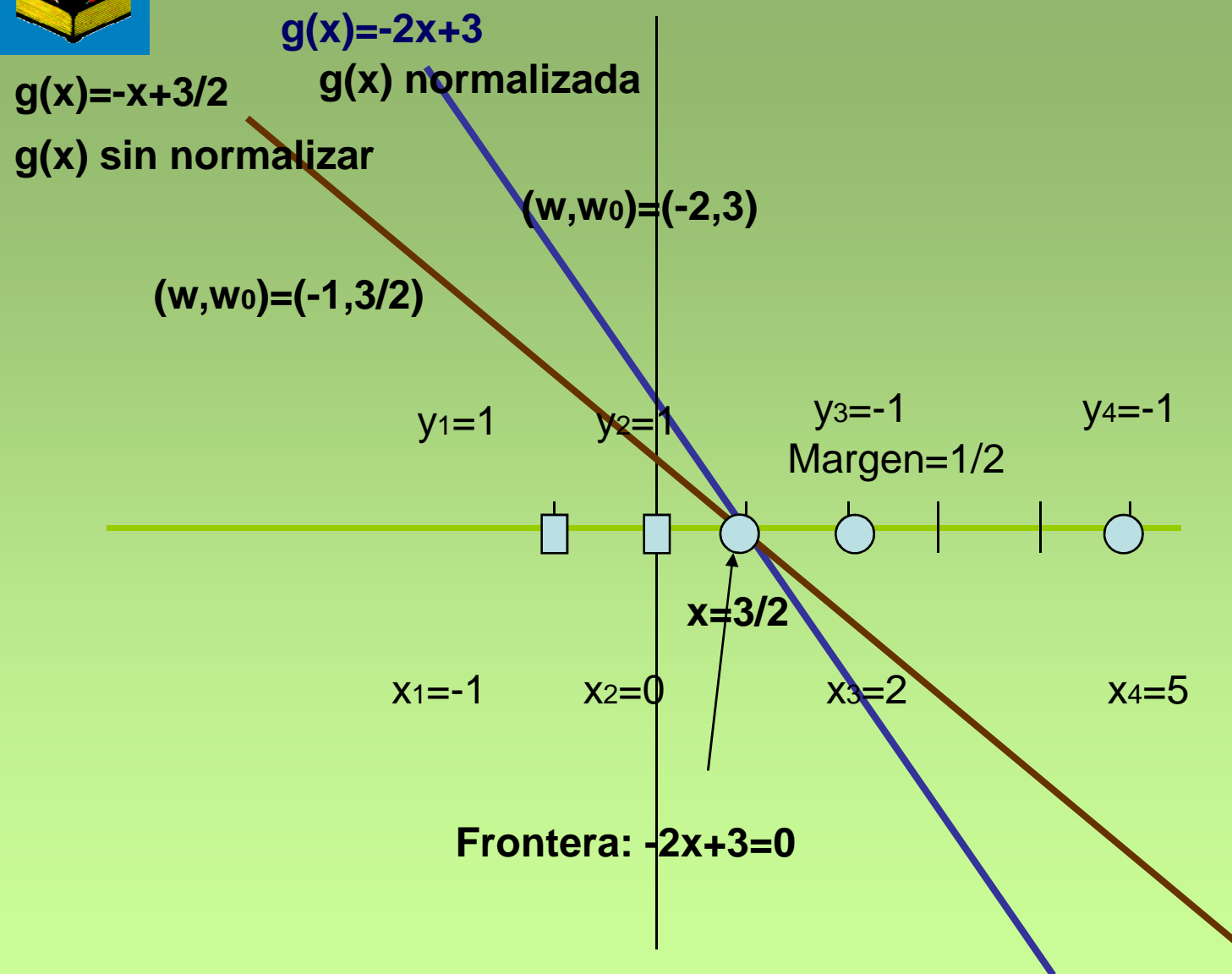
Además ahora, siempre es $y_i (wx_i + w_0) \geq 1$ y el margen es

en este caso

$$\text{margen}(w, w_0) = \min_{x_i \in H} \frac{y_i (w^T x_i + w_0)}{\sqrt{w^T w}} = \frac{1}{\sqrt{w^T w}} = \frac{1}{\sqrt{w^2}}$$



Gráfica Ejemplo Inicial





Ejemplo Inicial: Solución normalizada



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

– **Conjunto de entrenamiento:**

$(x_1 = -1, y_1 = 1)$ $(x_2 = 0, y_2 = 1)$ de C1;

$(x_3 = 2, y_3 = -1)$ $(x_4 = 5, y_4 = -1)$ de C2

Condiciones de separabilidad lineal NO normalizadas:

$$+ (wx_1 + w_0) > 0: -w + w_0 > 0;$$

$$w_0 > w$$

$$+ (wx_2 + w_0) > 0:$$

$$w_0 > 0$$

$$(wx_3 + w_0) < 0: 2w + w_0 < 0;$$

$$w_0 < -2w$$

$$(wx_4 + w_0) < 0: 5w + w_0 < 0;$$

$$w_0 < -5w$$

Condiciones de separabilidad lineal normalizadas:

$$+ (wx_1 + w_0) \geq 1: -w + w_0 \geq 1;$$

$$w_0 \geq w + 1, -w + w_0 \geq 1$$

$$+ (wx_2 + w_0) \geq 1:$$

$$w_0 \geq 1;$$

$$(wx_3 + w_0) \leq -1: 2w + w_0 \leq -1;$$

$$w_0 \leq -2w - 1, -2w - w_0 \geq 1$$

$$(wx_4 + w_0) \leq -1: 5w + w_0 \leq -1;$$

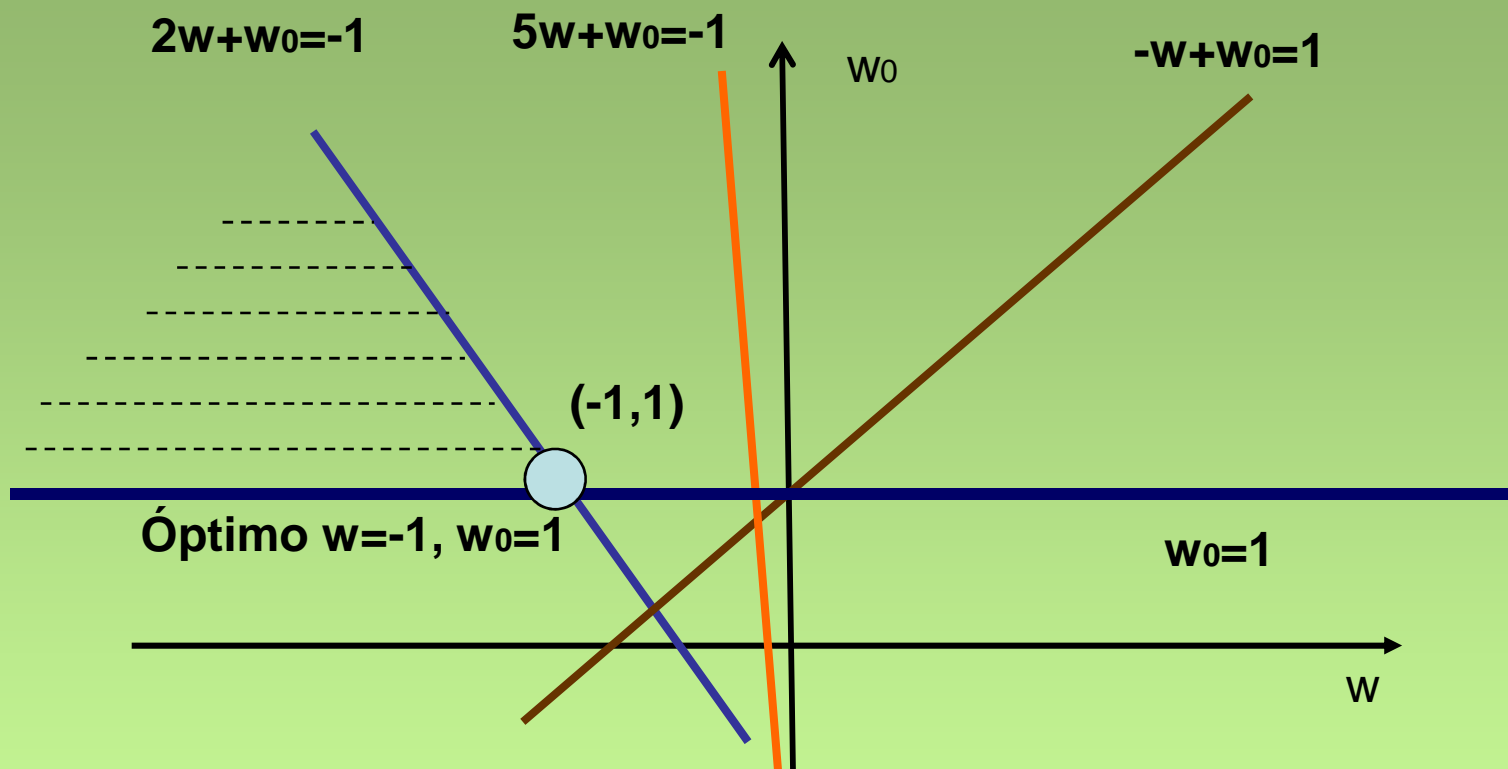
$$w_0 \leq -5w - 1, -5w - w_0 \geq 1$$



Ejemplo Inicial: Solución normalizada



Región rayada: Separabilidad lineal normalizada



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

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



Un Ejemplo Inicial: Solución



Clasificador unidimensional

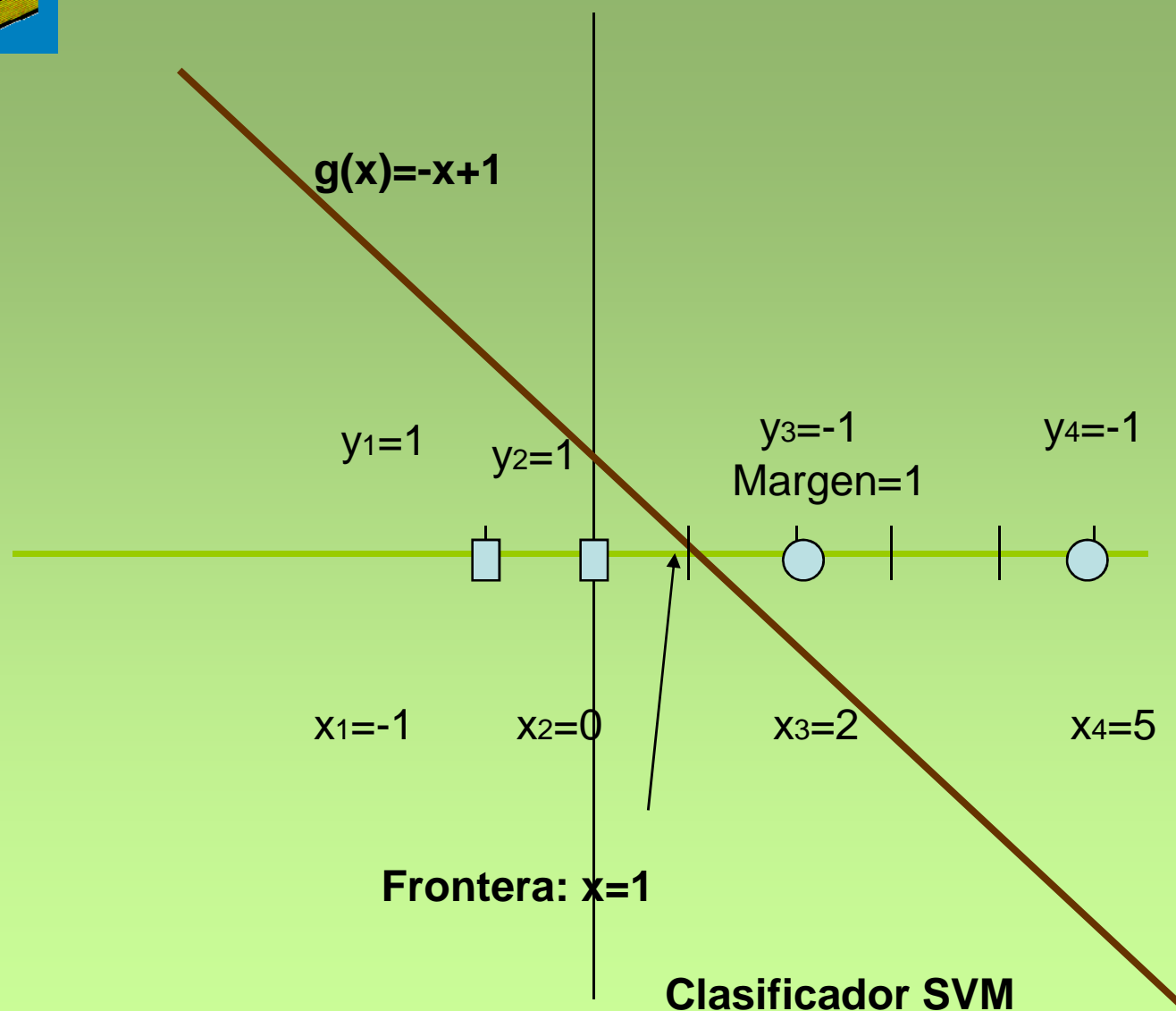
Función a minimizar: $\frac{1}{2} \mathbf{w}^T \mathbf{w} = \frac{1}{2} w^2$

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





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 programa primal.
- b) El problema *dual* tiene tantas restricciones como variables tiene el programa primal
- c) Los coeficientes de la función objetivo del problema *dual* son los términos independientes de las restricciones o RHS del programa *primal*.



DUALIDAD EN PROGRAMACION LINEAL



- d) Los términos independientes de las restricciones o RHS del PD 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	\leq	\geq	VARIABLES
	\geq	\leq	
	$=$	$><$	
VARIABLES	\geq	\geq	RESTRICCIONES
	\leq	\leq	
	$><$	$=$	

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:

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

s.a.

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

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

El problema dual (dual simétrico) es:

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

s.a.

$$\boldsymbol{\lambda}\mathbf{A} \geq \mathbf{c}$$

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



¿ 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 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

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

s.a:

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

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

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

Los términos independientes de las restricciones o RHS del PD 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 programa primal.

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

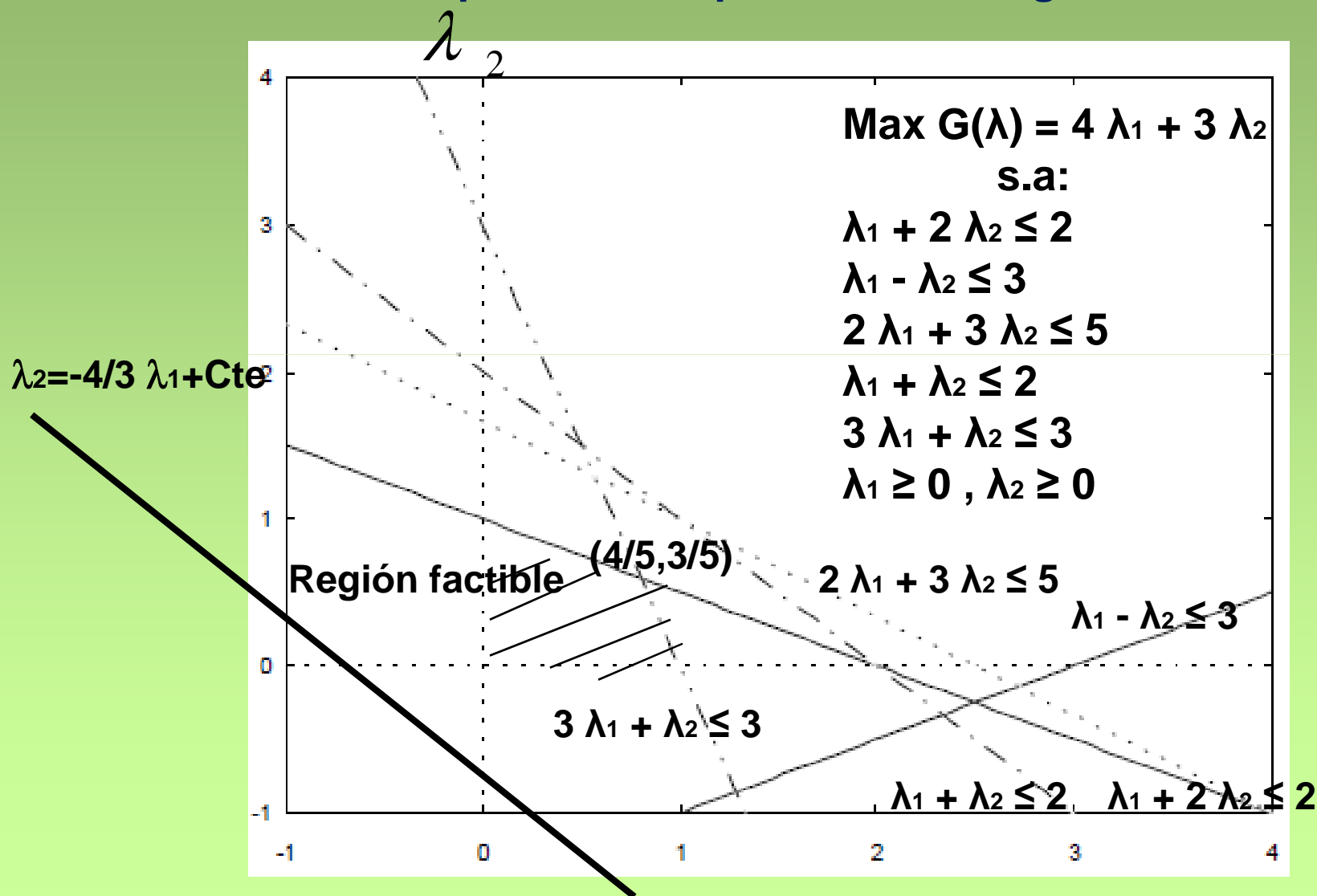
Los coeficientes de la función objetivo del problema dual son los términos independientes de las restricciones o RHS del problema primal.



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**:

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

s.a.

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

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

La **función lagrangiana** de este problema será:

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

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

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



Condiciones de Kuhn-Tucker)



Las condiciones de optimalidad de este problema respecto de las variables, son:

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

Respecto a los multiplicadores, son

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

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



- Un problema de programación lineal tiene la forma

$$\text{Min}_{x \geq 0} \left\{ c^T x \mid Ax \geq b \right\}$$

- La inferencia dual es

$$\text{Max}_{P \in \mathcal{P}} \left\{ v \mid Ax \geq b \xrightarrow{P} (c^T x \geq v) \right\}$$

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



Programación Lineal: Inferencia Dual



- Def 1.- Para un $u \geq 0$,
 $uAx \geq ub$ es una sustitución de $Ax \geq b$.

- Def 2.- $uAx \geq ub$ es una dominancia de $cx \geq v$
si $uAx \geq ub$ implica que $cx \geq v$ (para $x \geq 0$).
 - Esto es, $uA \leq c$ y $ub \geq v$.

- Esto es un método de inferencia completa.
 - Debido al Lema de Farkas.



- Así, la inferencia dual

$$\text{Max}_{P \in \mathcal{P}} \left\{ v \mid Ax \geq b \xrightarrow{P} (cx \geq v) \right\}$$

se convierte en

$$\text{Max}_{u \geq 0} \left\{ v \mid uA \leq c, ub \geq v \right\}$$

o

$$\text{Max}_{u \geq 0} \left\{ ub \mid uA \leq c \right\}$$

**cuando $Ax \geq b$, $x \geq 0$
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:

$$\text{Max}_{u \geq 0} \{ ub \mid uA \leq c \} = \text{Min}_{x \geq 0} \{ cx \mid Ax \geq b \}$$

Excepto cuando $uA \leq c$, $u \geq 0$ y $Ax \geq b$, $x \geq 0$ 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**

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

$$\text{Max}_{P \in \mathcal{P}} \left\{ v \mid C \xrightarrow{P} (f(x) \geq v) \right\}$$



Programación Lineal: Inferencia Dual



□ Primal:

$$\text{Min } 4x_1 + 7x_2$$

$$2x_1 + 3x_2 \geq 6 \quad (u_1)$$

$$2x_1 + x_2 \geq 4 \quad (u_2)$$

$$x_1, x_2 \geq 0$$

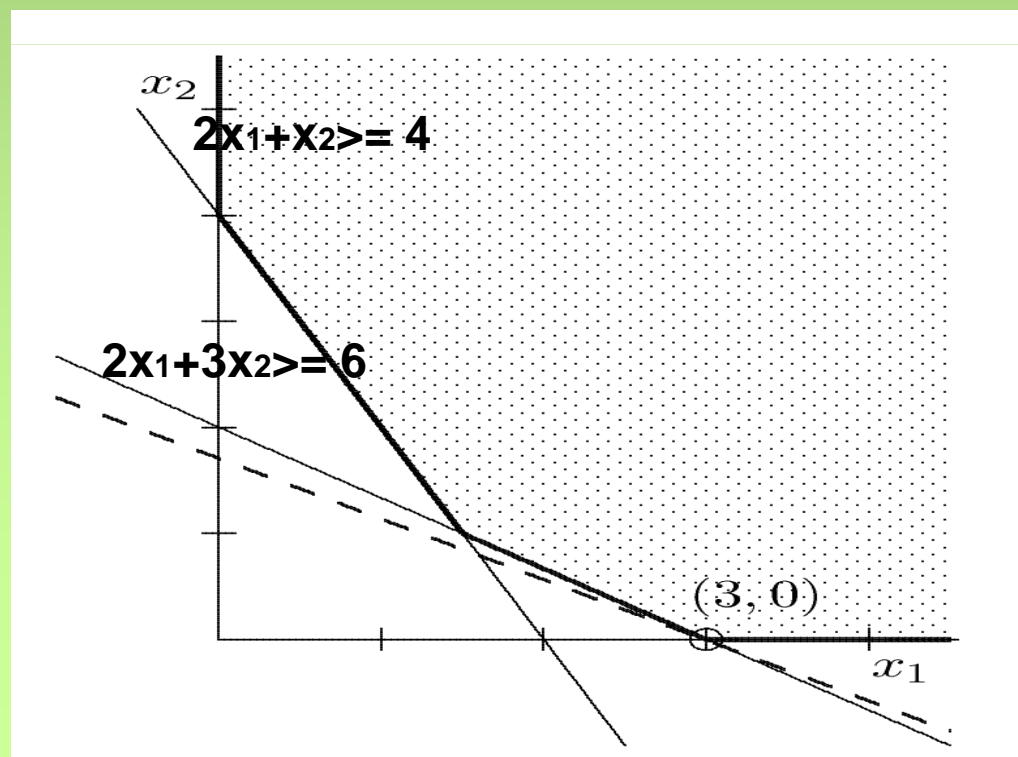
Dual:

$$\text{Max } 6u_1 + 4u_2$$

$$2u_1 + 2u_2 \leq 4 \quad (x_1)$$

$$3u_1 + u_2 \leq 7 \quad (x_2)$$

$$u_1, u_2 \geq 0$$





Programación Lineal: Inferencia Dual (una solución)



□ Primal:

Dual:

$$\text{Min } 4x_1 + 7x_2 = 12$$

$$2x_1 + 3x_2 \geq 6 \quad (u_1 = 2)$$

$$2x_1 + x_2 \geq 4 \quad (u_2 = 0)$$

$$x_1, x_2 \geq 0$$

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

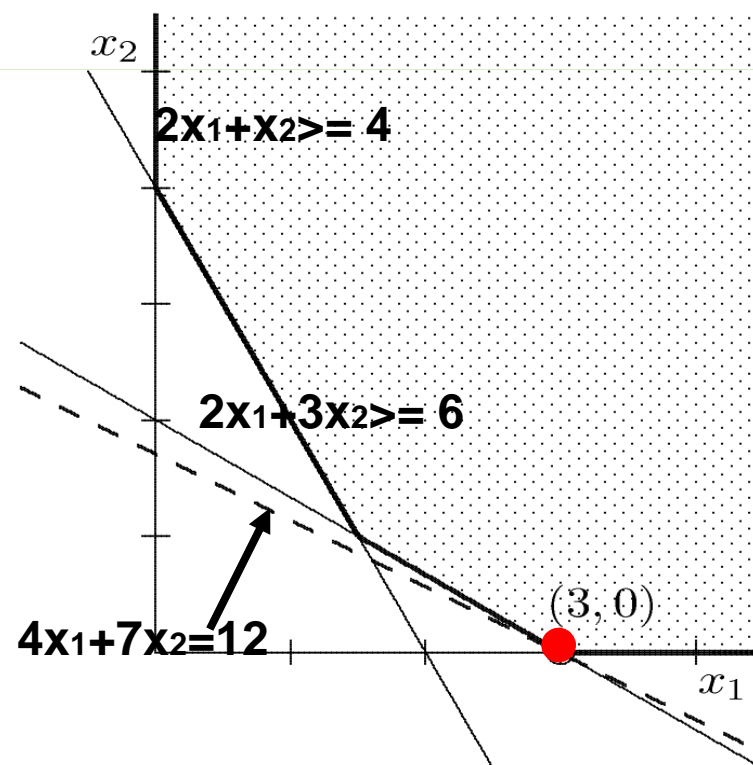
$$4x_1 + 7x_2 \geq 12$$

$$\text{Max } 6u_1 + 4u_2 = 12$$

$$2u_1 + 2u_2 \leq 4 \quad (x_1 = 3)$$

$$3u_1 + u_2 \leq 7 \quad (x_2 = 0)$$

$$u_1, u_2 \geq 0$$





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



□ Primal:

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

Dual:

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

Solución

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

Sustitución:
Multiplicamos la inecuación 2 por 2

$$\begin{aligned}2x_1 + 3x_2 &\geq 6 \quad (0) \\ 2x_1 + x_2 &\geq 4 \quad (2) \\ \hline 4x_1 + 2x_2 &\geq 8\end{aligned}$$

Dominancia

$$\begin{aligned}\Downarrow \\ 4x_1 + 7x_2 &\geq 8\end{aligned}$$



MVS:(CL).Resolución Práctica (1)



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

Se halla el problema dual asociado a maximizar la función Lagrangiana siguiente

$$\begin{aligned} \max \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j) &= \max \sum_{i=1} \alpha_i - \frac{1}{2} \mathbf{a}^T \mathbf{Q} \mathbf{a}, \quad \text{con } Q_{ij} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{s.a. } \sum_{i=1} \alpha_i y_i &= 0, \quad \alpha_i \geq 0, \quad i=1, \dots, n \end{aligned}$$

Una vez resuelto el problema dual

Si la solución óptima es $\hat{\mathbf{a}} = (\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_n)$, 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:

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



MVS:(CL).Resolución Práctica (2)



Vectores soporte

- En la solución óptima del problema dual $\hat{\alpha}=(\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_n)$ los únicos valores no nulos 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:

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

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

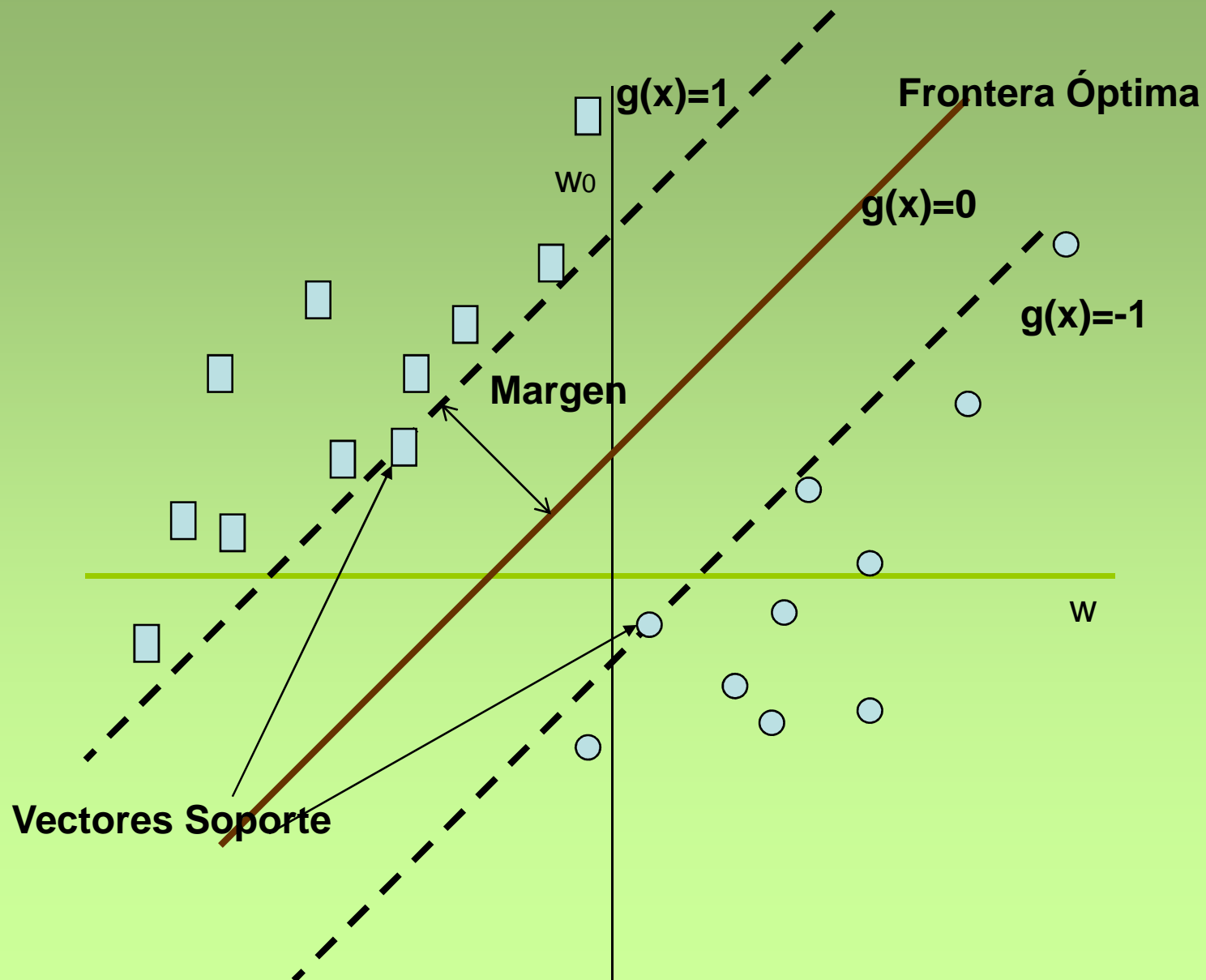
$$g(\mathbf{x}_j) = \hat{\mathbf{w}}^T \mathbf{x}_j + \hat{w}_0 = \left(\sum_{i \in \text{Sop}} \hat{\alpha}_i y_i \mathbf{x}_i \right)^T \mathbf{x}_j + \hat{w}_0 = \sum_{i \in \text{Sop}} \hat{\alpha}_i y_i (\mathbf{x}_i^T \mathbf{x}_j) + \hat{w}_0$$



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

$$\max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$
$$s.a. \sum_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i=1, \dots, n$$

Para calcular w_0 sólo se necesitan los productos escalares

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

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

Para calcular la clase a la que pertenece un vector \mathbf{x}_j

sólo se necesitan los productos escalares

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



Un Ejemplo Inicial: Solución Dual con vectores soporte



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

– Conjunto de entrenamiento: $(-1, 1)$ $(0, 1)$ de C1; $(2, -1)$ $(5, -1)$ de C2

$$\left. \begin{array}{l} \min_{\mathbf{w}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ s.a. \quad y_i (\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1, \quad i = 1 \dots n \end{array} \right\} \Rightarrow \begin{array}{l} -w + w_0 \geq 1 \\ w_0 \geq 1 \\ -2w - w_0 \geq 1 \\ -5w - w_0 \geq 1 \end{array} \quad \text{Primal}$$

Vectores soporte $(0,1)$ y $(2,-1)$

$$\left. \begin{array}{l} \max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j) \\ s.a. \quad \sum_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i=1, \dots, n \end{array} \right\} \Rightarrow \quad \text{Dual}$$
$$\max \alpha_1 + \alpha_2 - \frac{1}{2} (0\alpha_1^2 + 0\alpha_1\alpha_2 + 0\alpha_2\alpha_1 + 4\alpha_2^2)$$
$$s.a. \quad \alpha_1 - \alpha_2 = 0, \quad \alpha_1 \geq 0, \quad \alpha_2 \geq 0$$



Un Ejemplo Inicial: Solución Dual con vectores soporte



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

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

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

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

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



Un ejemplo inicial: Solución Dual con vectores soporte



Derivando con respecto a los α_i y λ tenemos

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

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

- Solución óptima:

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

- Vector óptimo:

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

- Constante óptima:

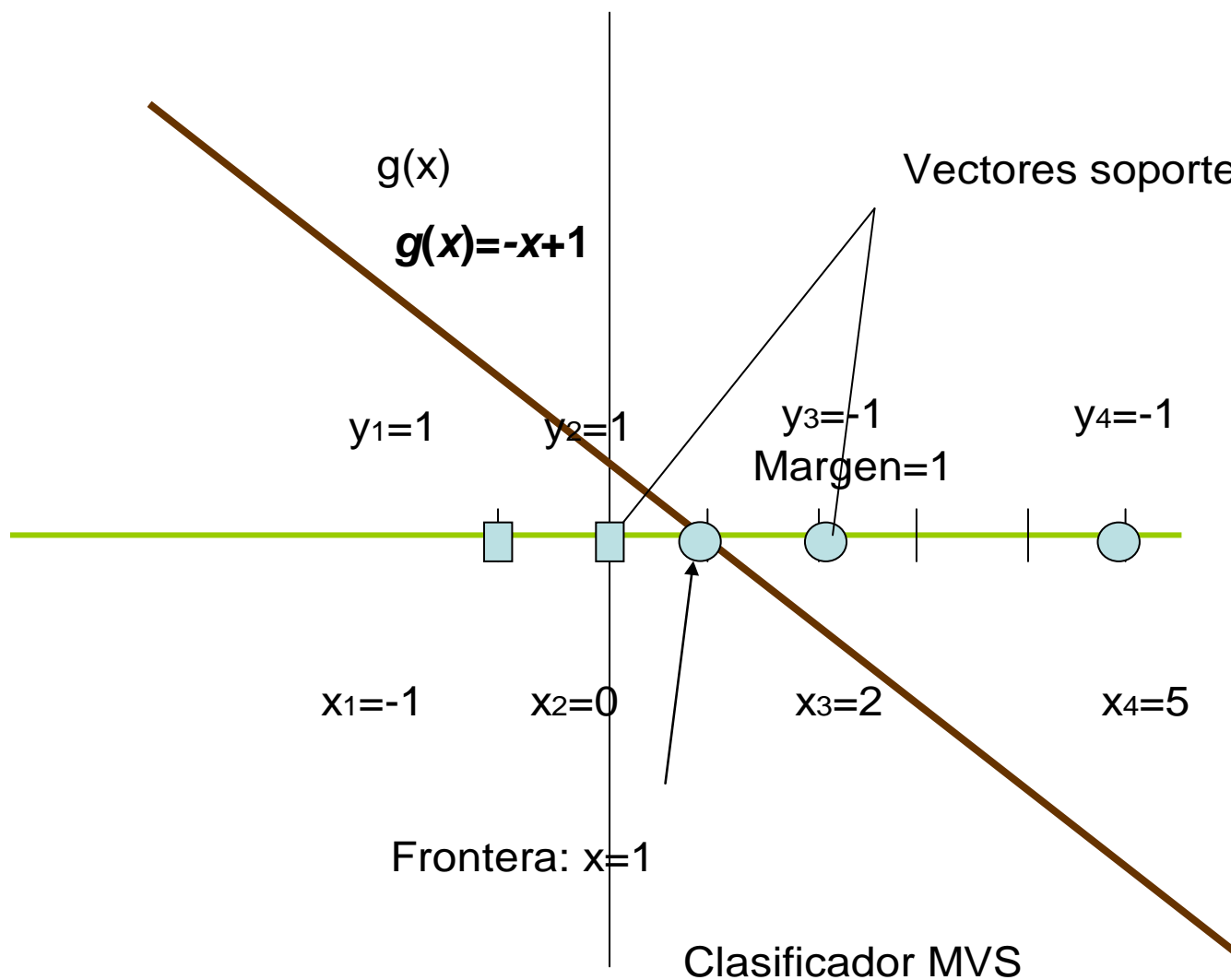
$$\hat{w}_0 = 1 - \hat{\mathbf{w}}x_1 = 1 - (-1) \times 0 = 1$$

- 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





Un Ejemplo Inicial: Solución Dual Transformada



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

– Conjunto de entrenamiento: (-1, 1) (0, 1) de C1; (2, -1) (5, -1) de C2

$$\left. \begin{array}{l} \min_{\mathbf{w}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ s.a. \quad y_i (\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1, \quad i = 1 \dots n \end{array} \right\} \Rightarrow \begin{array}{l} -w + w_0 \geq 1 \\ w_0 \geq 1 \\ -2w - w_0 \geq 1 \\ -5w - w_0 \geq 1 \end{array} \quad \text{Primal}$$

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

$$\max \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2} \left(\begin{array}{l} 2\alpha_1^2 + 1\alpha_1\alpha_2 + 3\alpha_1\alpha_3 + 6\alpha_1\alpha_4 + \\ 1\alpha_2\alpha_1 + 1\alpha_2^2 + 1\alpha_2\alpha_3 + 1\alpha_2\alpha_4 + \\ 3\alpha_3\alpha_1 + 1\alpha_3\alpha_2 + 5\alpha_3^2 + 11\alpha_3\alpha_4 + \\ 6\alpha_4\alpha_1 + 1\alpha_4\alpha_2 + 11\alpha_4\alpha_3 + 26\alpha_4^2 \end{array} \right)$$

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



Un Ejemplo Inicial: Solución Dual Transformada



$$\max \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2} \begin{pmatrix} 2\alpha_1^2 + 1\alpha_1\alpha_2 + 3\alpha_1\alpha_3 + 6\alpha_1\alpha_4 + \\ 1\alpha_2\alpha_1 + 1\alpha_2^2 + 1\alpha_2\alpha_3 + 1\alpha_2\alpha_4 + \\ 3\alpha_3\alpha_1 + 1\alpha_3\alpha_2 + 5\alpha_3^2 + 11\alpha_3\alpha_4 + \\ 6\alpha_4\alpha_1 + 1\alpha_4\alpha_2 + 11\alpha_4\alpha_3 + 26\alpha_4^2 \end{pmatrix}$$

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

$$L = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - (\alpha_1)^2 - \alpha_1\alpha_2 - 3\alpha_1\alpha_3 - 6\alpha_1\alpha_4 - (1/2)(\alpha_2)^2 - \alpha_2\alpha_3 - \alpha_2\alpha_4 - (5/2)(\alpha_3)^2 - 11\alpha_3\alpha_4 - 13(\alpha_4)^2 - \lambda(\alpha_1 + \alpha_2 - \alpha_3 - \alpha_4)$$

•Derivando parcialmente con respecto a los α_i y a λ

tenemos la solución óptima: $\hat{\alpha} = (-1, 3/2, 1/2, 0)$



Un ejemplo inicial: Solución Dual Transformada



Derivando con respecto a los α_i y λ tenemos

$$1) \frac{\partial L}{\partial \alpha_1} = 1 - 2\alpha_1 - \alpha_2 - 3\alpha_3 - 6\alpha_4 - \lambda = 0$$

$$2) \frac{\partial L}{\partial \alpha_2} = 1 - \alpha_1 - \alpha_2 - \alpha_3 - \alpha_4 - \lambda = 0$$

$$3) \frac{\partial L}{\partial \alpha_3} = 1 - 3\alpha_1 - \alpha_2 - 5\alpha_3 - 11\alpha_4 + \lambda = 0$$

$$4) \frac{\partial L}{\partial \alpha_4} = 1 - 6\alpha_1 - \alpha_2 - 11\alpha_3 - 26\alpha_4 + \lambda = 0$$

$$5) \frac{\partial L}{\partial \lambda} = -\alpha_1 - \alpha_2 + \alpha_3 + \alpha_4 = 0 \quad \text{Solución} \equiv \lambda = 0, \alpha_1 = -1, \alpha_2 = 3/2, \alpha_3 = 1/2 \text{ y } \alpha_4 = 0$$

Tenemos un sistema de cinco ecuaciones con cinco incógnitas y resolviéndolo de forma matricial o por reducción y sustitución tenemos que si restamos la Eq 1 de la Eq 2 $\alpha_1 = -2\alpha_3 - 5\alpha_4$

Si restamos la Eq 1 de la Eq 3, tenemos que junto con el resultado anterior $\lambda = 0$

Sustituyendo el valor de α_1 en la Eq 5, tenemos que $\alpha_2 = 3\alpha_3 + 6\alpha_4$

Por último sustituyendo en las Eq 2 y Eq 4 los valores de α_1 y α_2

Tenemos las ecuaciones $2\alpha_3 + 2\alpha_4 = 1$, y $2\alpha_3 + 10\alpha_4 = 1$,

de donde se deduce que $\alpha_4 = 0$ y $\alpha_3 = 1/2$



Un Ejemplo Inicial: Solución Dual Transformada



- **Solución óptima:**

$$\hat{\alpha} = (-1, 3/2, 1/2, 0)$$

- **Vector óptimo:**

$$\begin{aligned}\hat{\mathbf{w}} &= \sum_{i \in \text{Sop}} \hat{\alpha}_i y_i \mathbf{x}_i = \\ &= (3/2) \times 1 \times 0 + (1/2) \times (-1) \times 2 = -1,\end{aligned}$$

- **Constante óptima:**

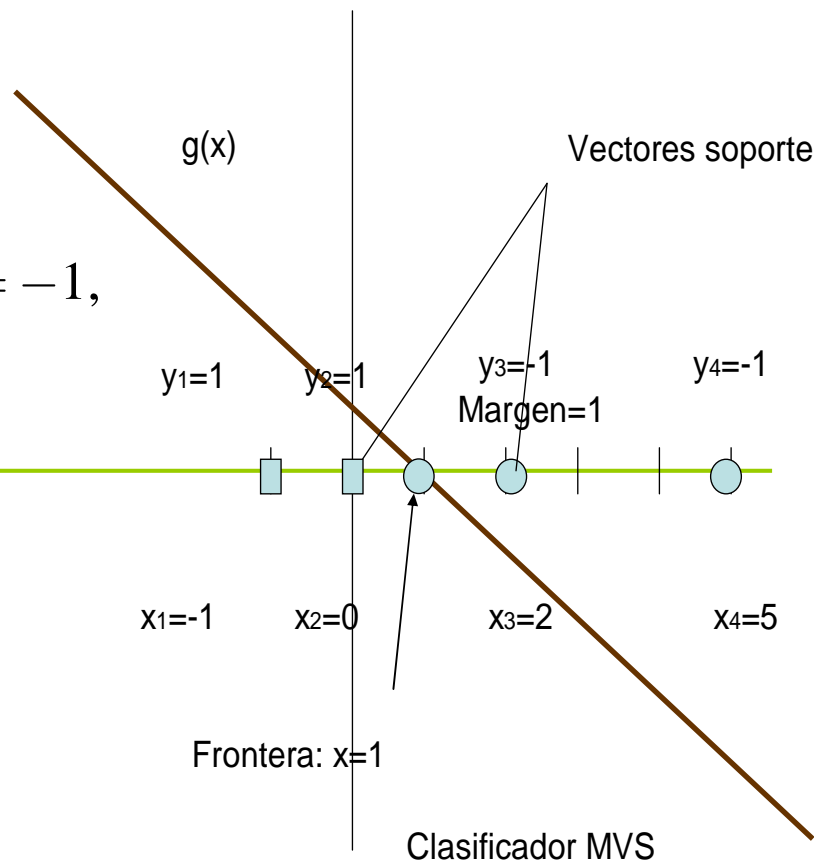
$$\hat{w}_0 = 1 - \hat{\mathbf{w}}x_2 = 1 - (-1) \times 0 = 1$$

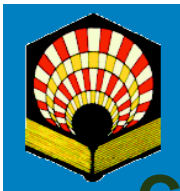
- **Clasificador:**

– **Por tanto el clasificador SVM**

es $g(x) = -x + 1$ y

la frontera $g(x) = 0$ es $x = 1$



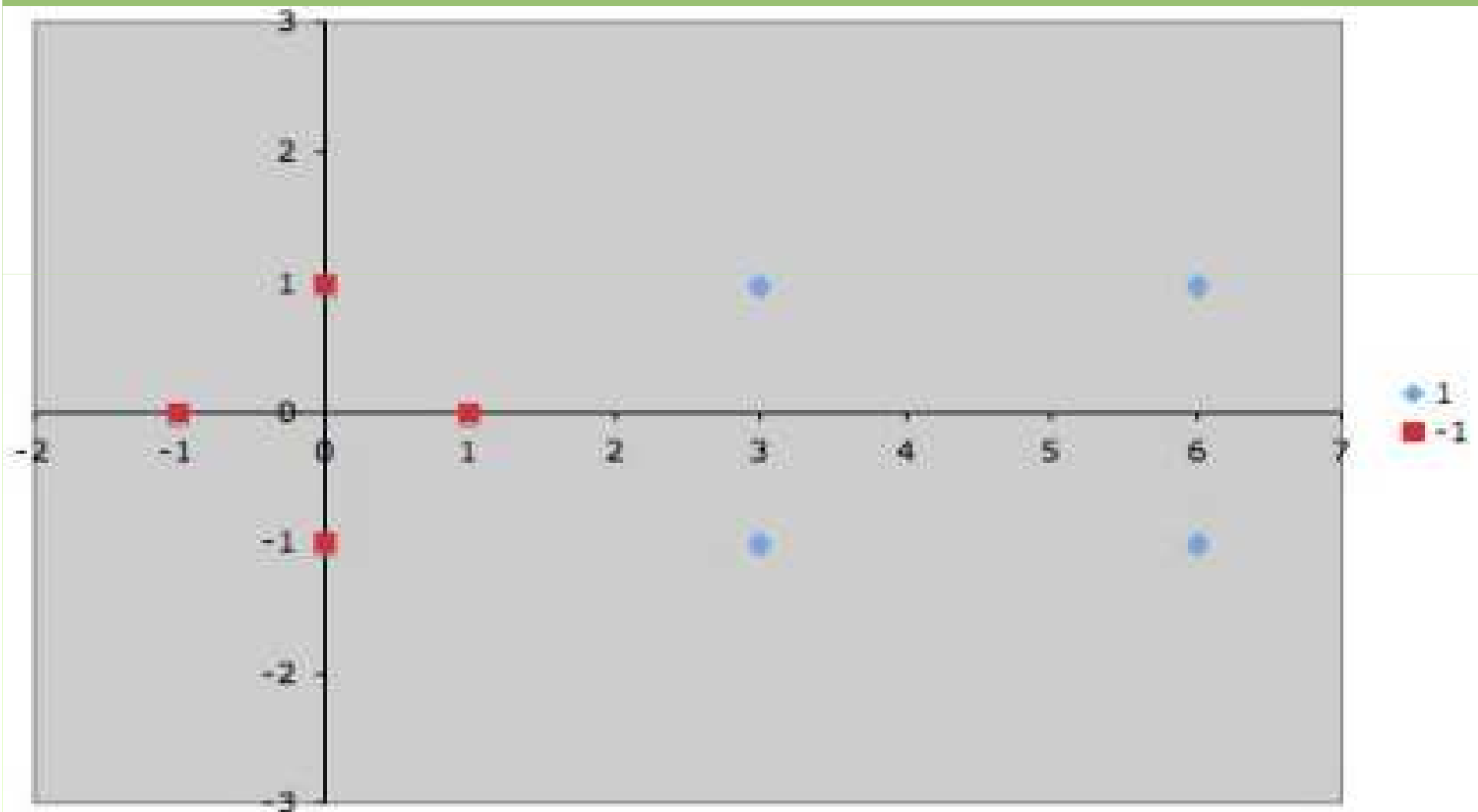


Nuevo ejemplo linealmente separable



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

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



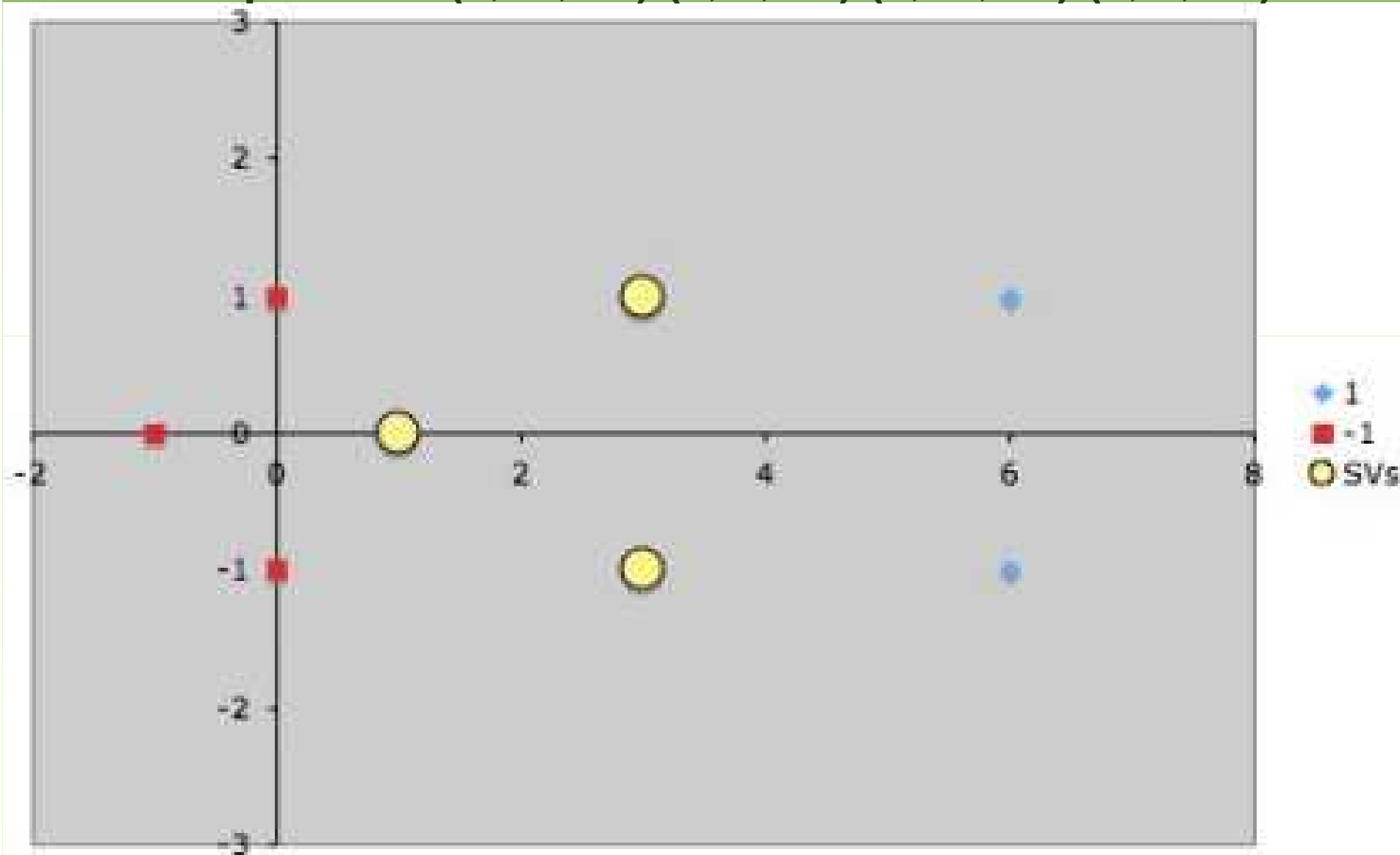


Ejemplo: Vectores soporte



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) $H_1: \mathbf{w} \cdot \mathbf{x} + w_0 = 1$
 Clase negativa (-1, 0, -1) (0, -1, -1) (0, 1, -1) (1, 0, -1) $H_2: \mathbf{w} \cdot \mathbf{x} + w_0 = -1$

$$w_1 x_1 + w_2 x_2 + w_0 = -1$$

$$1w_1 + 0w_2 + w_0 = -1 \quad (1)$$

$$\rightarrow w_0 = -1 - w_1$$

$$(1, 0) \rightarrow -1$$

$$(3, -1) \rightarrow +1$$

$$(3, 1) \rightarrow +1$$

Vectores soporte

$$w_1 x_1 + w_2 x_2 + w_0 = 1$$

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

$$\rightarrow w_2 = 3w_1 - 1 - w_0 - 1$$

$$\rightarrow w_2 = 2w_1 - 2$$

$$3w_1 + 1w_2 + w_0 = 1 \quad (3)$$

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

$$\rightarrow w_1 = 1$$

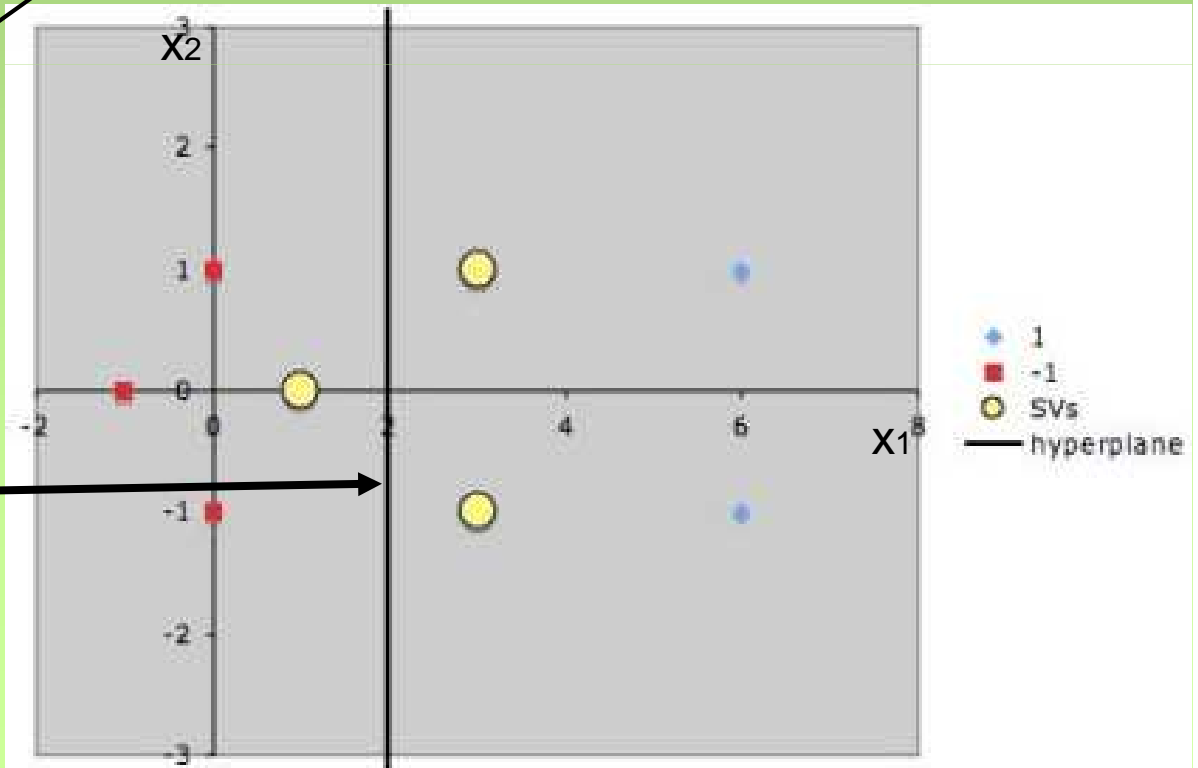
$$\rightarrow w_0 = -2$$

$$\rightarrow w_2 = 0$$

Función discriminante

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

$$\text{Recta: } x_1 = 2$$





Un ejemplo inicial (4): Solución Dual



Vectores soporte

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

$$\begin{aligned} \mathbf{x}_1 &= (x_{11}, x_{12}) = (1, 0) \rightarrow y_1 = -1 \\ \mathbf{x}_2 &= (x_{21}, x_{22}) = (3, -1) \rightarrow y_2 = +1 \\ \mathbf{x}_3 &= (x_{31}, x_{32}) = (3, 1) \rightarrow y_3 = +1 \end{aligned}$$

$$\max_{\alpha} \alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2} \left(\alpha_1^2 (\mathbf{x}_1^T \cdot \mathbf{x}_1) - \alpha_1 \alpha_2 (\mathbf{x}_1^T \cdot \mathbf{x}_2) - \alpha_1 \alpha_3 (\mathbf{x}_1^T \cdot \mathbf{x}_3) - \alpha_2 \alpha_1 (\mathbf{x}_2^T \cdot \mathbf{x}_1) + \alpha_2^2 (\mathbf{x}_2^T \cdot \mathbf{x}_2) + \alpha_2 \alpha_3 (\mathbf{x}_2^T \cdot \mathbf{x}_3) + \right. \\ \left. - \alpha_3 \alpha_1 (\mathbf{x}_3^T \cdot \mathbf{x}_1) + \alpha_3 \alpha_2 (\mathbf{x}_3^T \cdot \mathbf{x}_2) + \alpha_3^2 (\mathbf{x}_3^T \cdot \mathbf{x}_3) \right)$$

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

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

$$F(\alpha_1, \alpha_2, \alpha_3, \lambda; \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) =$$

$$= \alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2} \left(\alpha_1^2 (\mathbf{x}_1^T \cdot \mathbf{x}_1) - \alpha_1 \alpha_2 (\mathbf{x}_1^T \cdot \mathbf{x}_2) - \alpha_1 \alpha_3 (\mathbf{x}_1^T \cdot \mathbf{x}_3) - \alpha_2 \alpha_1 (\mathbf{x}_2^T \cdot \mathbf{x}_1) + \alpha_2^2 (\mathbf{x}_2^T \cdot \mathbf{x}_2) + \alpha_2 \alpha_3 (\mathbf{x}_2^T \cdot \mathbf{x}_3) + \right. \\ \left. - \alpha_3 \alpha_1 (\mathbf{x}_3^T \cdot \mathbf{x}_1) + \alpha_3 \alpha_2 (\mathbf{x}_3^T \cdot \mathbf{x}_2) + \alpha_3^2 (\mathbf{x}_3^T \cdot \mathbf{x}_3) \right)$$

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



Un ejemplo inicial (4): Solución Dual



Derivando con respecto a los α_i y λ tenemos

$$\frac{\partial F}{\partial \alpha_1} = 1 - \frac{1}{2} \left(2\alpha_1 (\mathbf{x}_1^T \cdot \mathbf{x}_1) - \alpha_2 (\mathbf{x}_1^T \cdot \mathbf{x}_2) - \alpha_3 (\mathbf{x}_1^T \cdot \mathbf{x}_3) - \alpha_2 (\mathbf{x}_2^T \cdot \mathbf{x}_1) - \alpha_3 (\mathbf{x}_3^T \cdot \mathbf{x}_1) \right) + \lambda = 0$$

esto es

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

$$\frac{\partial F}{\partial \alpha_2} = 1 - \frac{1}{2} \left(-\alpha_1 (\mathbf{x}_1^T \cdot \mathbf{x}_2) - \alpha_1 (\mathbf{x}_2^T \cdot \mathbf{x}_1) + 2\alpha_2 (\mathbf{x}_2^T \cdot \mathbf{x}_2) + \alpha_3 (\mathbf{x}_2^T \cdot \mathbf{x}_3) + \alpha_3 (\mathbf{x}_3^T \cdot \mathbf{x}_2) \right) - \lambda = 0$$

esto es

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

$$\frac{\partial F}{\partial \alpha_3} = 1 - \frac{1}{2} \left(-\alpha_1 (\mathbf{x}_1^T \cdot \mathbf{x}_3) + \alpha_2 (\mathbf{x}_2^T \cdot \mathbf{x}_3) - \alpha_1 (\mathbf{x}_3^T \cdot \mathbf{x}_1) + \alpha_2 (\mathbf{x}_3^T \cdot \mathbf{x}_2) + 2\alpha_3 (\mathbf{x}_3^T \cdot \mathbf{x}_3) \right) - \lambda = 0$$

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

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

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

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

Constante óptima:

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

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

$(1, 0) \rightarrow -1$

$(3, -1) \rightarrow +1$

$(3, 1) \rightarrow +1$



Ejemplo



Función discriminante

$$H: (1, 0) \cdot \mathbf{x} - 2 = 0; H: x_1 - 2 = 0$$

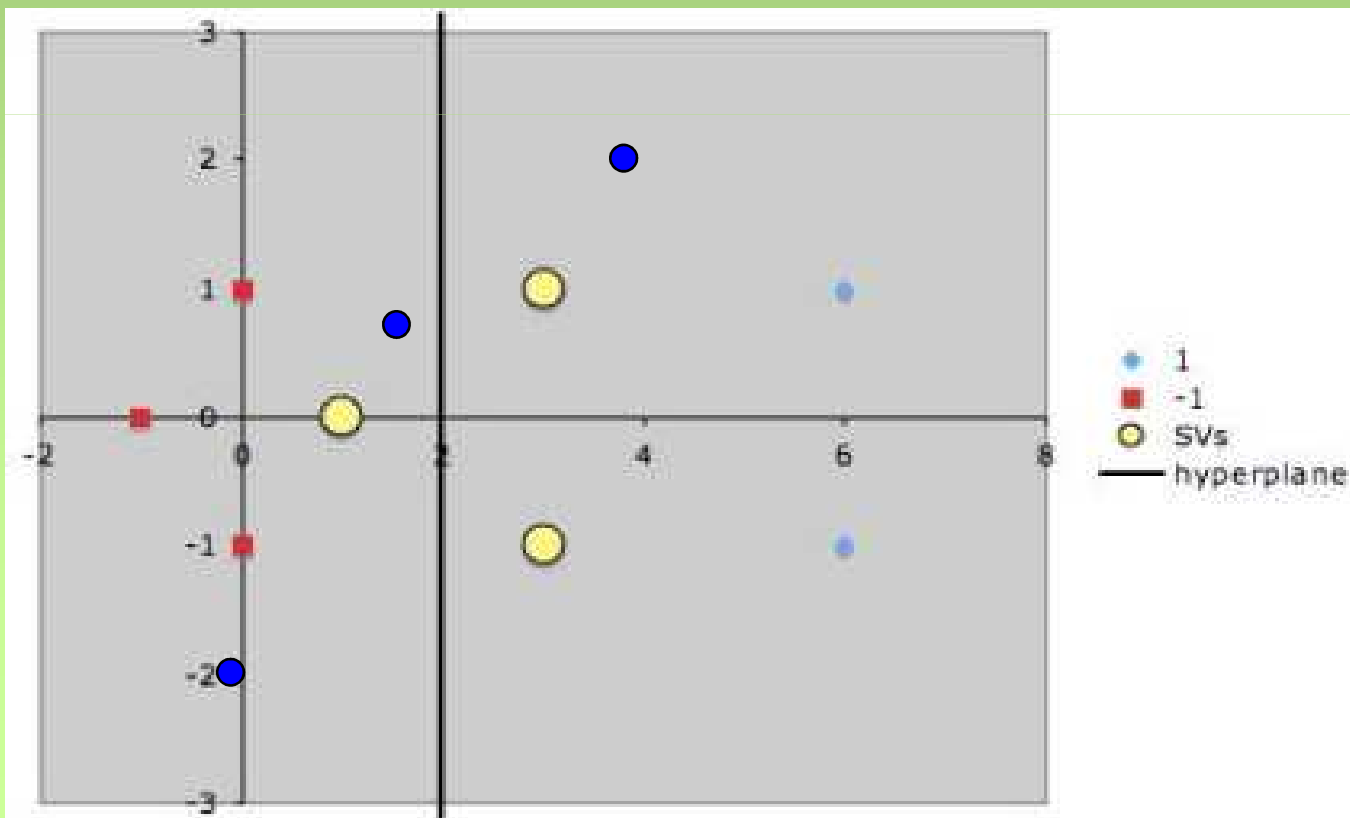
Datos del conjunto de test

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

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

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

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

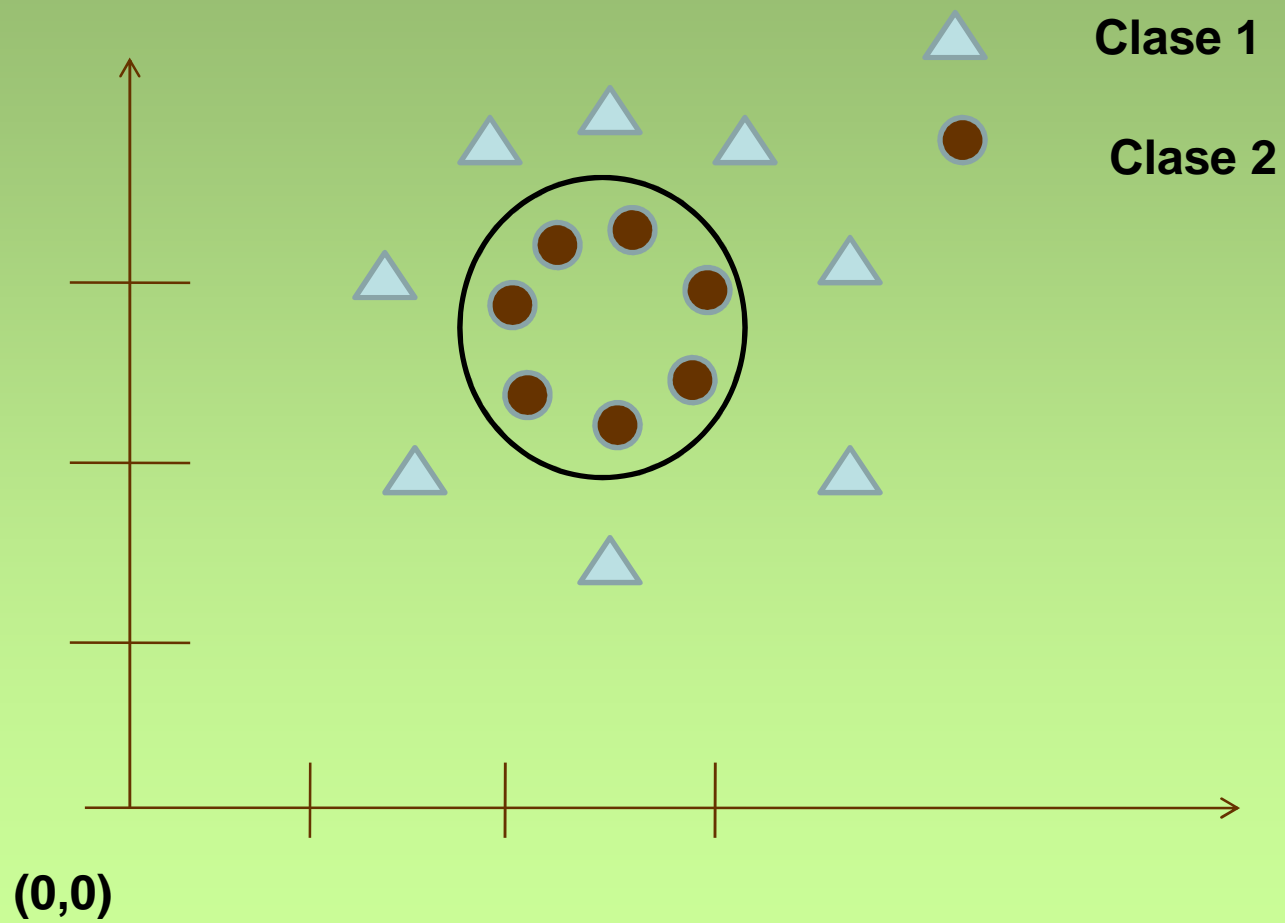




Como funciona para datos linealmente no separables?



- ❑ 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





Ejemplo

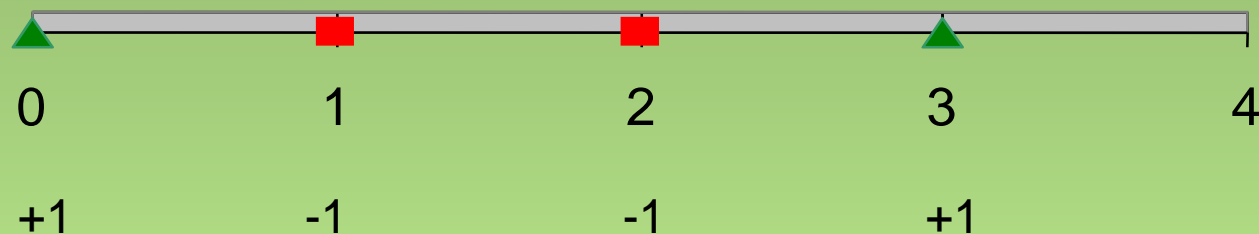


Como separar las dos clases mediante un punto

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

▲ Clase +1

■ Clase -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



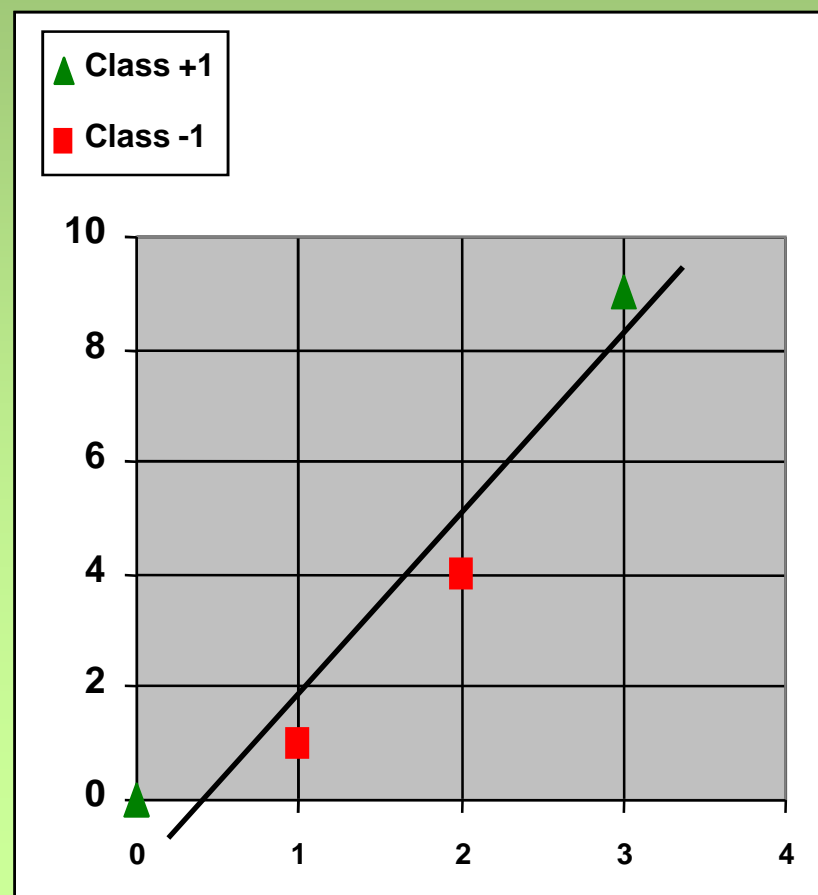
Ejemplo



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

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

- $w \cdot x + w_0 = +1$ (para la clase positiva)

$$w_1 x_1 + w_2 x_2 + w_0 = +1$$

1) $0w_1 + 0w_2 + w_0 = +1 \rightarrow w_0 = 1$

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

- $w \cdot x + w_0 = -1$ (para la clase negativa)

$$w_1 x_1 + w_2 x_2 + w_0 = -1$$

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

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

sustituyendo w_0 en las dos ecuaciones

$$\rightarrow w_1 = -2 - w_2$$

$$\rightarrow -4 - 2w_2 + 4w_2 + 1 = -1$$

de donde $w_2 = 1$ y $w_1 = -3$, valores

que verifican la ecuación 2)

- **Función discriminante**

$$w^T \cdot x + w_0 = 0$$

$$w_1 x_1 + w_2 x_2 + w_0 = 0$$

$$\rightarrow -3x_1 + x_2 + 1 = 0, \text{ o lo que es lo}$$

$$\text{mismo } x_2 = 3x_1 - 1$$

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



Ejemplo



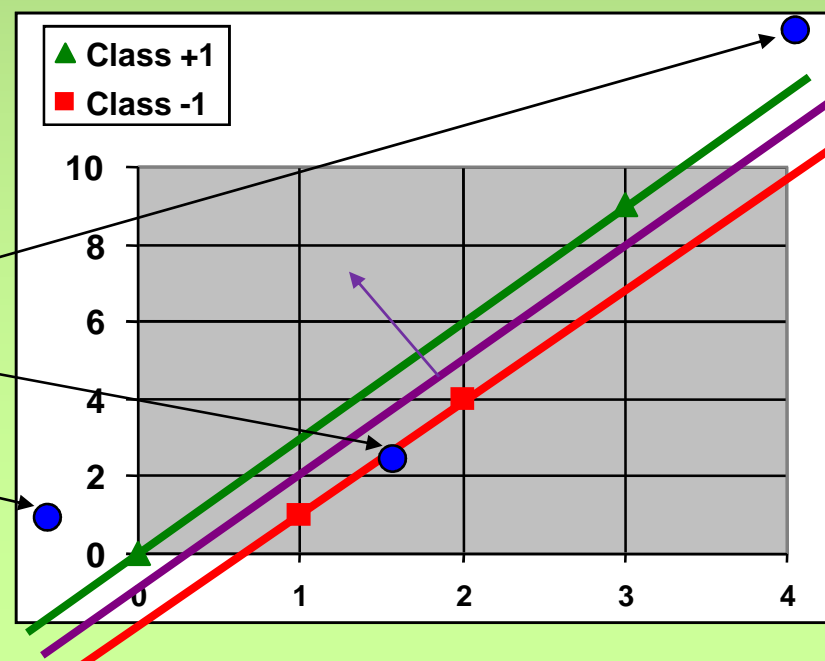
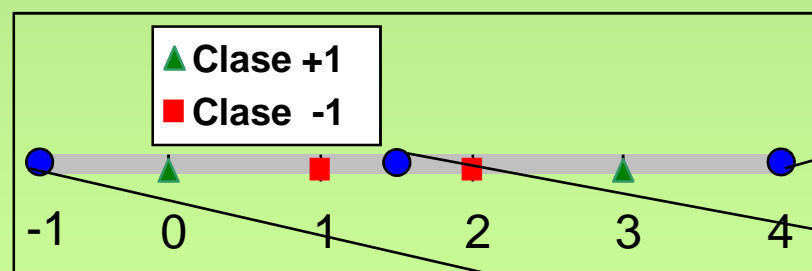
$$H: -3x_1 + x_2 + 1 = 0$$

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]



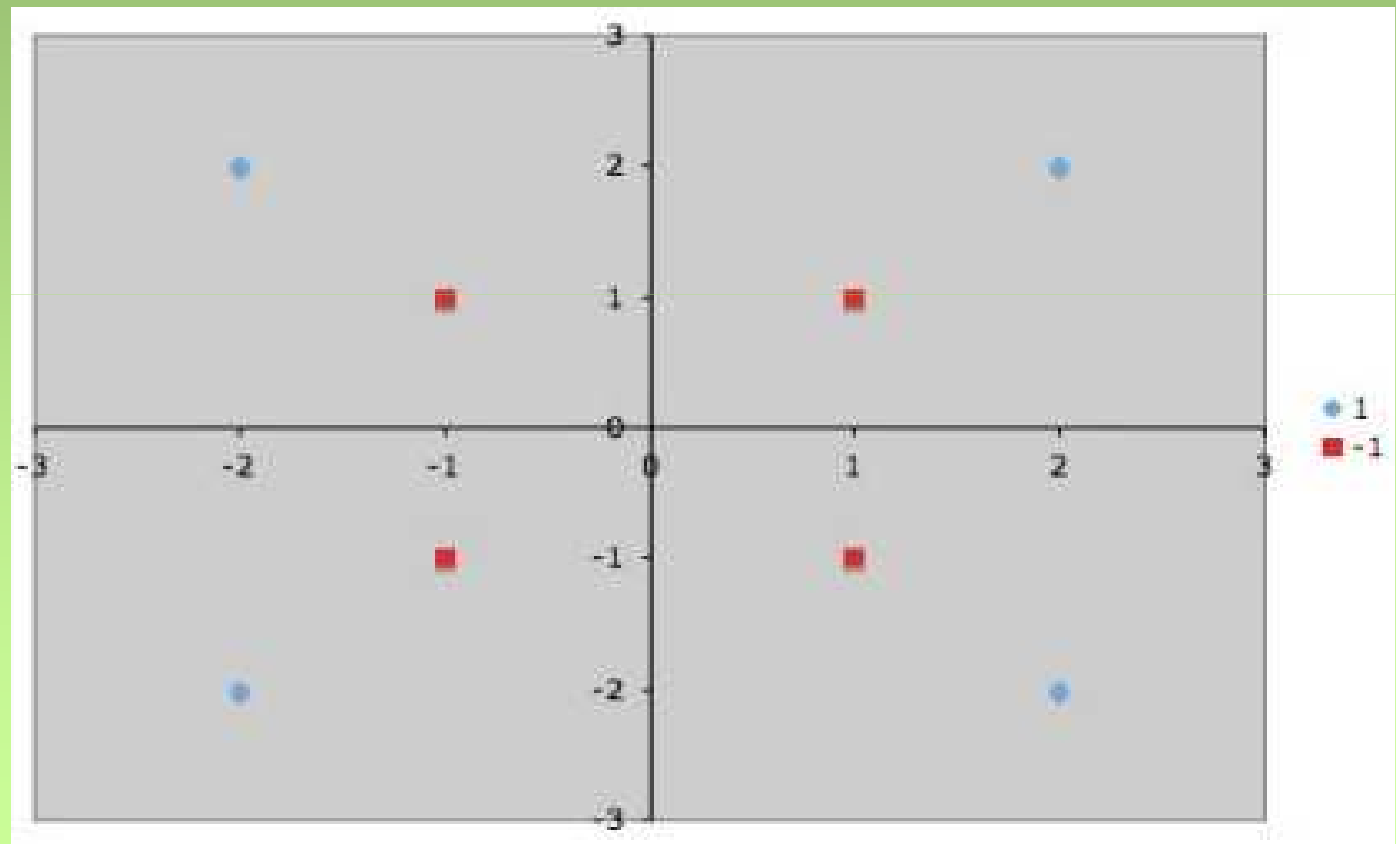


Nuevo Ejemplo



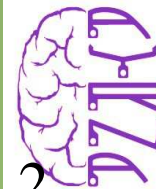
Como separar estas dos clases mediante una función lineal?

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



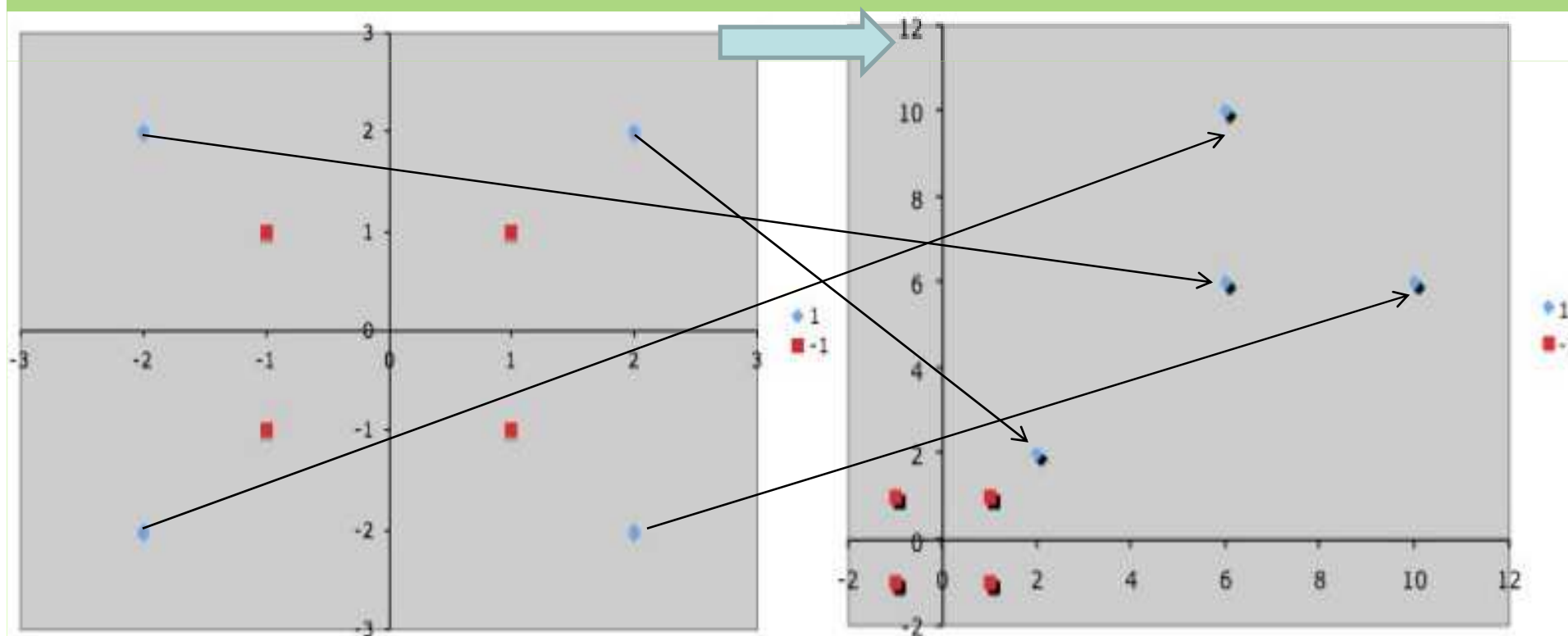


Nuevo Ejemplo



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

Esta función mantiene la dimensión bidimensional del espacio de entrada





Nuevo Ejemplo



Vectores Soporte en el espacio de características

$$H_0: \mathbf{w} \cdot \mathbf{x} + w_0 = 1$$

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

$$w_1 x_1 + w_2 x_2 + w_0 = -1$$

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

$$\rightarrow w_1 = -1 - w_0 - w_2$$

$$w_1 x_1 + w_2 x_2 + w_0 = 1$$

$$2) 2(-1 - w_0 - w_2) + 2w_2 + w_0 = 1$$

$$-2 - 2w_0 - 2w_2 + 2w_2 + w_0 = 1$$

$$\rightarrow w_0 = -3$$

$$(2-1) x_2 = (2-1) x_1$$

$$2x_2 - x_2 = 2x_1 - x_1$$

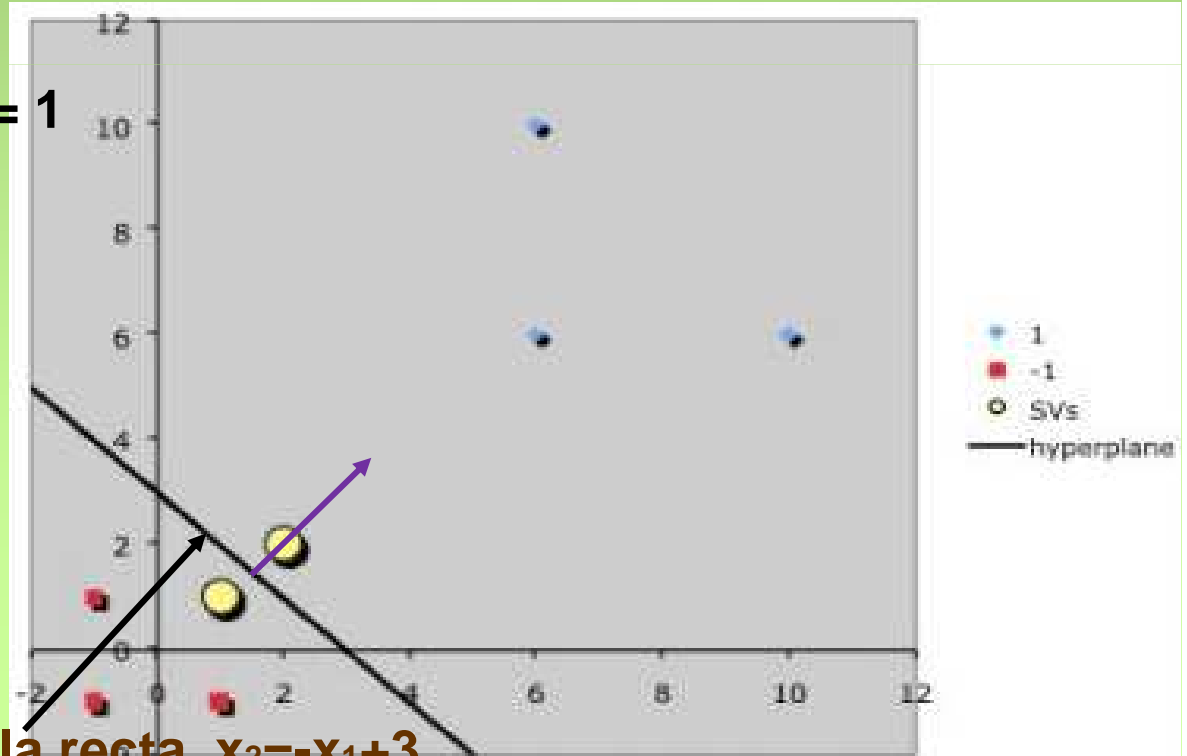
$$x_2 = x_1$$

$$w_1 = w_2 = 1$$

$$H: (1,1)^T \cdot (x_1, x_2) - 3 = 0$$

$$x_1 + x_2 - 3 = 0; \text{ Ecuación de la recta, } x_2 = -x_1 + 3$$

x_1	x_2	Clase
1	1	-1
2	2	+1



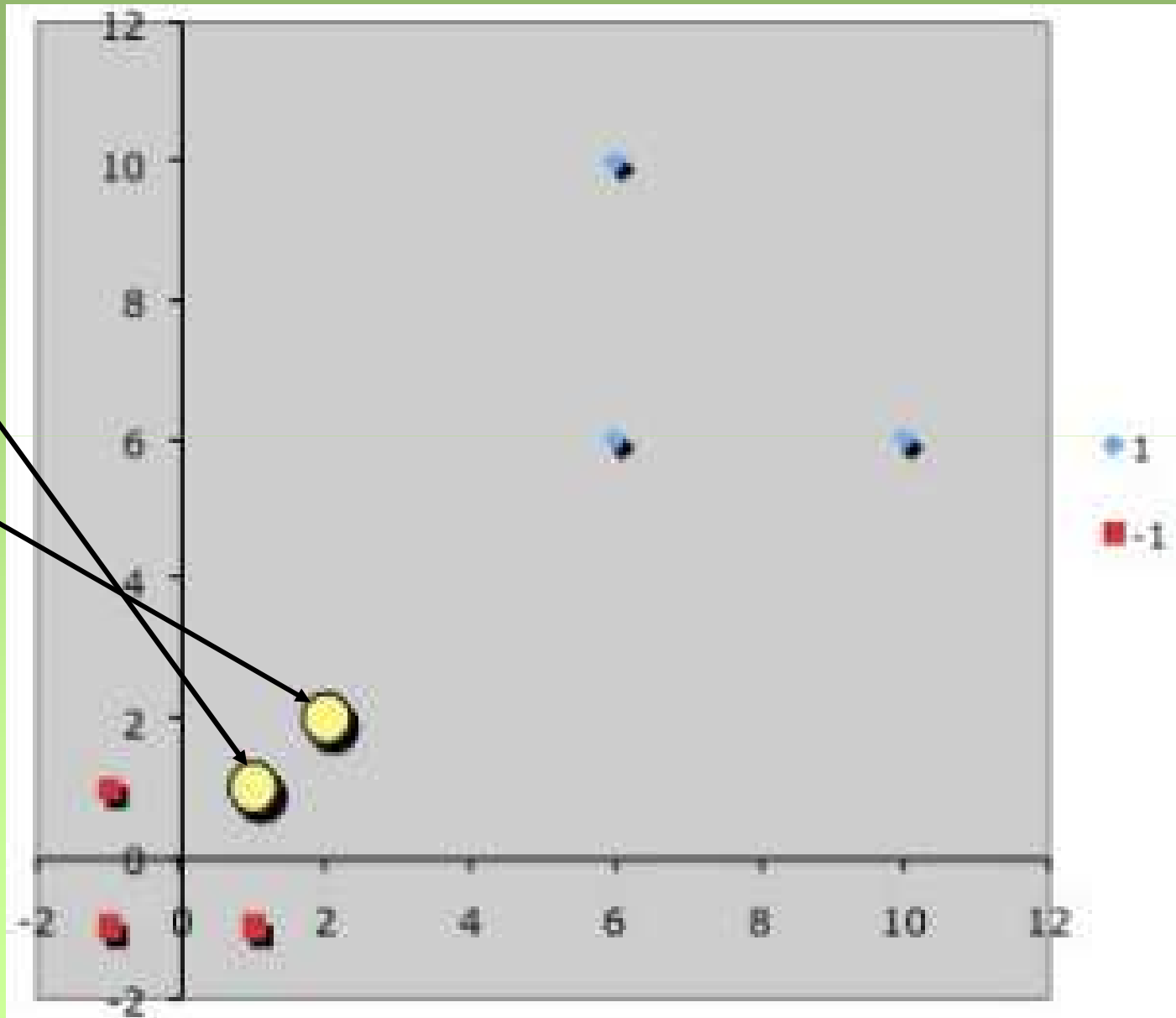


Nuevo Ejemplo



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



$$\left. \begin{aligned} \max_{\alpha} \quad & \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j (\Phi(s_i)^T \cdot \Phi(s_j)) \\ \text{s.a.} \quad & \sum_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, i=1, \dots, n \end{aligned} \right\} \Rightarrow \begin{aligned} & \Phi(s_1)^T = (1, 1), \quad y_1 = -1 \\ & \Phi(s_2)^T = (2, 2), \quad y_2 = 1 \end{aligned}$$

Vectores soporte en el espacio transformado

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

$$\max(\alpha_1 + \alpha_2 - \frac{1}{2} (\alpha_1^2 (2) - \alpha_1 \alpha_2 (4) - \alpha_2 \alpha_1 (4) + \alpha_2^2 (8))) = \max(\alpha_1 + \alpha_2 - \frac{1}{2} (2\alpha_1^2 - 8\alpha_1 \alpha_2 + 8\alpha_2^2))$$

$$\text{s.a.} \quad -\alpha_1 + \alpha_2 = 0, \quad \alpha_1 \geq 0, \quad \alpha_2 \geq 0$$

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

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

Derivando con respecto a los α_i y λ tenemos

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



Nuevo Ejemplo: Resolución mediante el dual



Vector óptimo:

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

Constante óptima:

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

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

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

La función discriminante lineal es

H: $(1,1) \cdot (x_1, x_2)^T - 3 = 0$, y la ecuación de la recta es

$$\mathbf{x}_1 + \mathbf{x}_2 - 3 = 0, \quad \mathbf{x}_2 = -\mathbf{x}_1 + 3$$



Capacidad de Generalización



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

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

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

$$\begin{aligned} f((4,5)^T) &= \sigma(1 \times (-1) \times (\Phi(1,1) \cdot \Phi(4,5)^T) + 1 \times 1 \times (\Phi(2,2) \cdot \Phi(4,5)^T) - 3) = \\ &= \sigma(-(1,1) \cdot (0,1)^T + (2,2) \cdot (0,1)^T - 3) = \sigma(-2), \text{ donde } \sigma \text{ devuelve el signo de } -2 \end{aligned}$$

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?

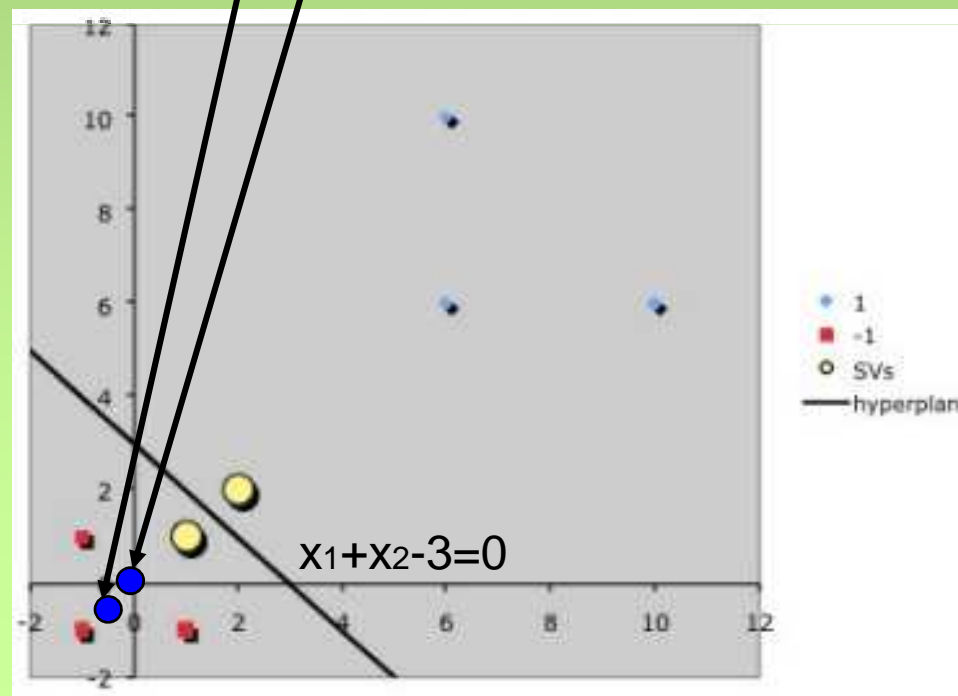
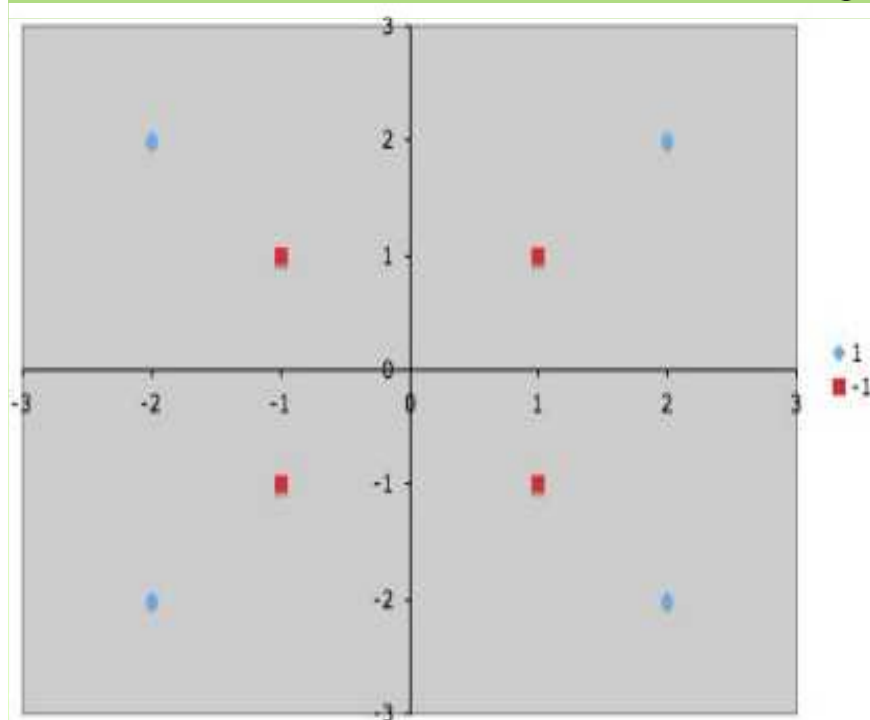
Dato de test (5,5, +1)

$\Phi(5,5) = (-1, -1)$ Error de clasificación!

Dato de test (4,4, +1)

$\Phi(4,4) = (0, 0)$ Error de clasificación!

(5,5)
(4,4)



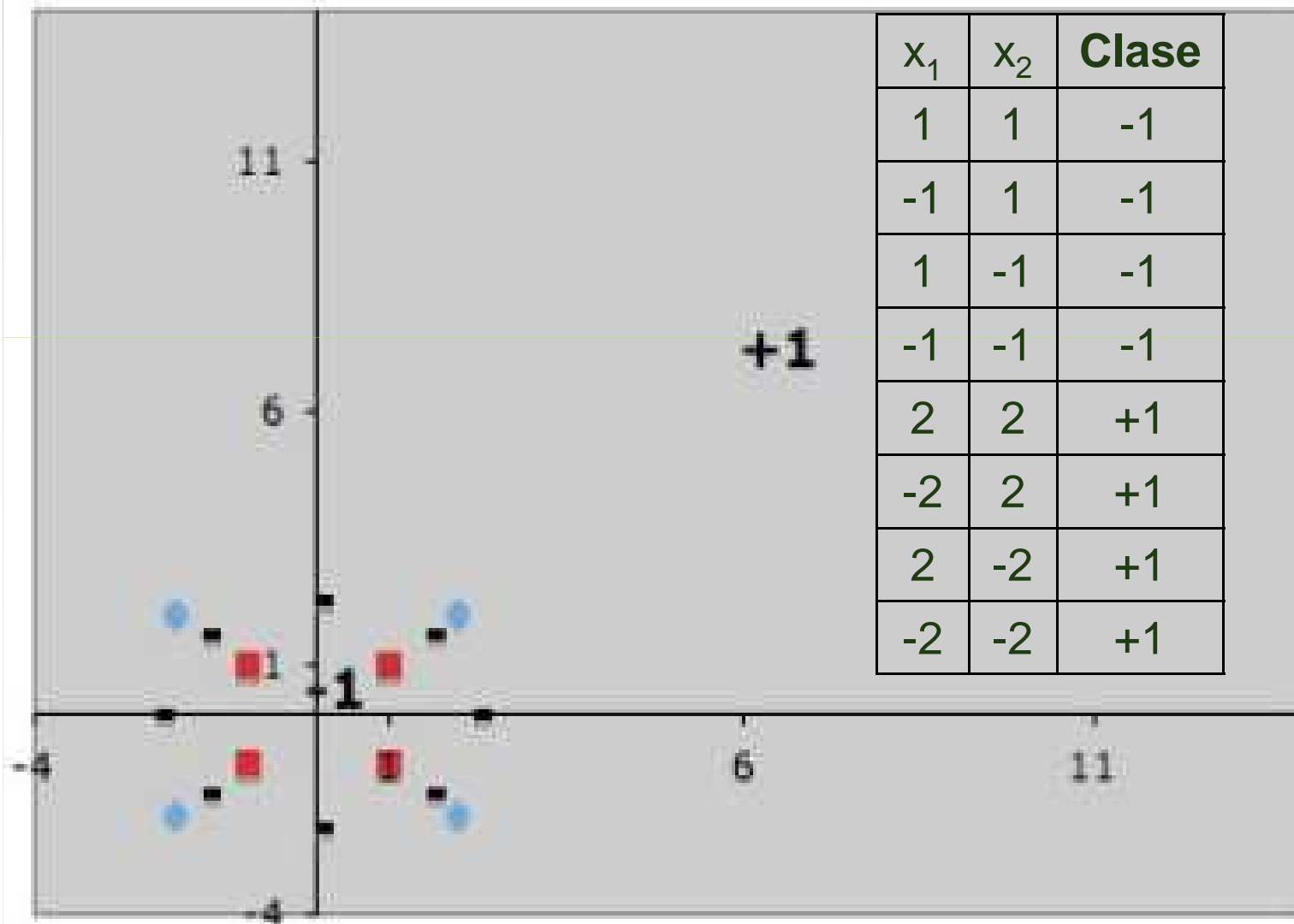


Nuevo Ejemplo



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

Con esta transformación la capacidad de generalización aumenta



● +1
■ -1
— hyperplane



Nuevo Ejemplo



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

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



Nuevo Ejemplo



Vectores Soporte

$$H_0: \mathbf{w} \cdot \mathbf{x} + w_0 = 1$$

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

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0 = -1$$

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

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

$$\left. \begin{aligned} 2w_1 x_1 + 2w_2 x_2 + w_3 x_3 + w_0 &= 1 \\ -2w_1 x_1 + 2w_2 x_2 + w_3 x_3 + w_0 &= 1 \end{aligned} \right\} 4w_2 + 2w_3 + 2w_0 = 2$$

$$\left. \begin{aligned} 2w_1 x_1 - 2w_2 x_2 + w_3 x_3 + w_0 &= 1 \\ -2w_1 x_1 - 2w_2 x_2 + w_3 x_3 + w_0 &= 1 \end{aligned} \right\} -4w_2 + 2w_3 + 2w_0 = 2$$

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

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

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

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



Nuevo Ejemplo



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

$x_3 = 0$; Ecuación del plano

Esta función separa realmente al espacio original de forma lineal?

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

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

Dato de test (0,5,0,5, -1) $\Phi(0.5,0.5) = (0.5, 0.5, -4.5/3)$ al ser -4,5/3 negativo la clase es -1 Acierto

x_1	x_2	$x_3 = ((x_1^2 + x_2^2) - 5) / 3$	Clase
5	5	45/3	1
4	4	27/3	1
0,5	0,5	-4.5/3	-1



Problema



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

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

- Siempre aumentan el número de dimensiones del espacio de entrada



MVS: 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 $y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1 - \xi_i$, $i = 1 \dots n$, con $\xi_i \geq 0$

– Entonces:

- Si $\xi_i = 0$ la muestra 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 (1)



En este caso (\mathbf{w}, w_0) no satisface la restricción

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

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

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i \\ \text{s.a.} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1 - \xi_i, \quad i = 1 \dots n, \text{ con } \xi_i \geq 0 \quad (3) \end{aligned}$$

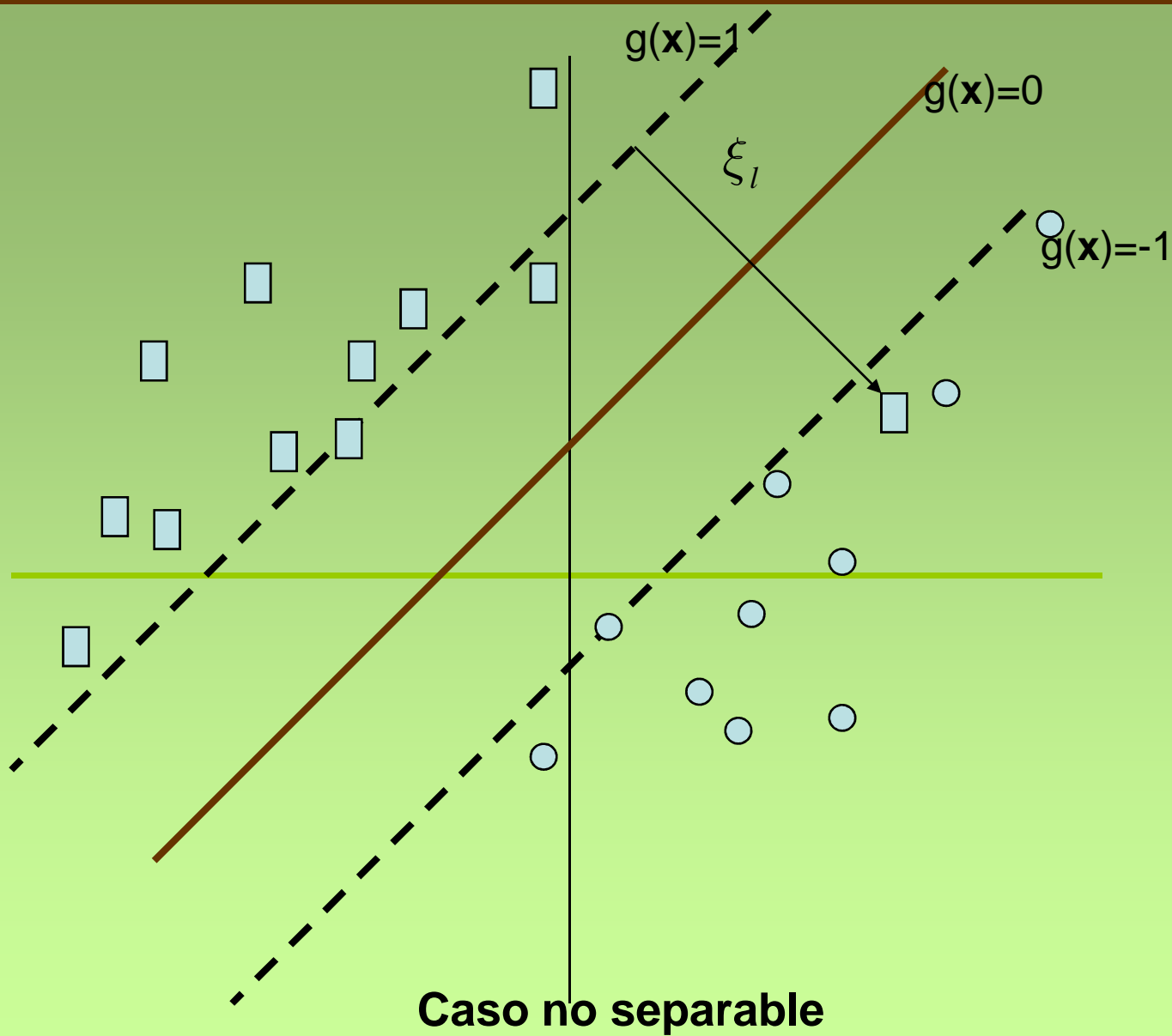
Es decir, las restricciones (3) permiten que los datos de entrenamiento puedan no estar en el lado correcto del hiperplano de separación $\mathbf{w}^T \mathbf{x} + w_0 = 0$

Esta situación ocurre cuando $\xi_i > 1$. Un ejemplo se muestra en la siguiente figura

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



SVM: Caso no Separable (1)





SVM: Caso no Separable (2)



Combinación de criterios u optimización multiobjetivo:

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

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

– El primer término de la suma a optimizar busca el mayor margen, el segundo el menor número de equivocaciones. 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:**
$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

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

• **Vector óptimo:**
$$\hat{\mathbf{w}} = \sum_{i \in S_{op}} \hat{\alpha}_i y_i \mathbf{x}_i$$



SVM: Caso no Separable (1)



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.

Pensemos en C como una varianza.

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.

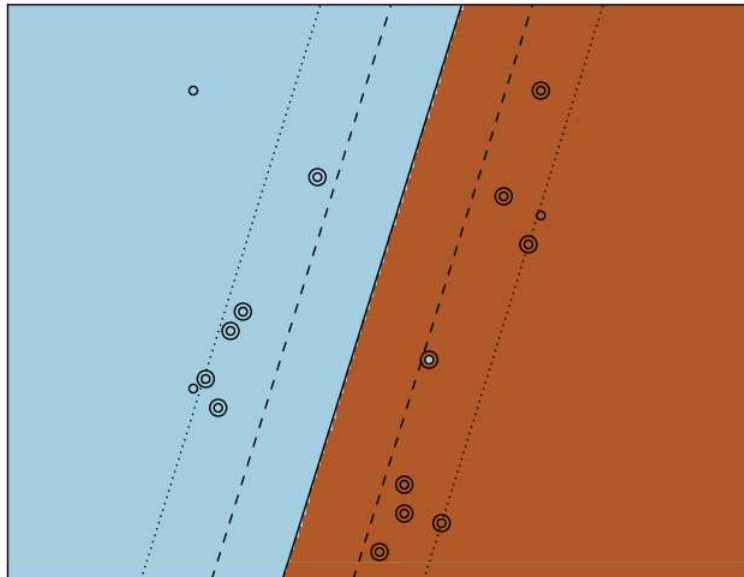
C	Penalización por Error	Nº de puntos considerados	Margen	Sesgo	Varianza
Bajo	Bajo	Alto	Bajo	Alto	Grande
Alto	Alto	Alto	Bajo	Bajo	Pequeña



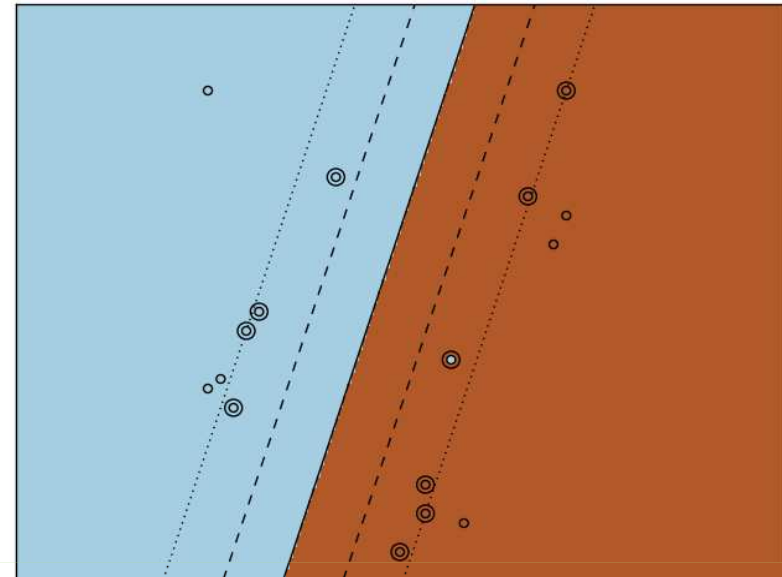
PARAMETRO C EN SVM LINEAL



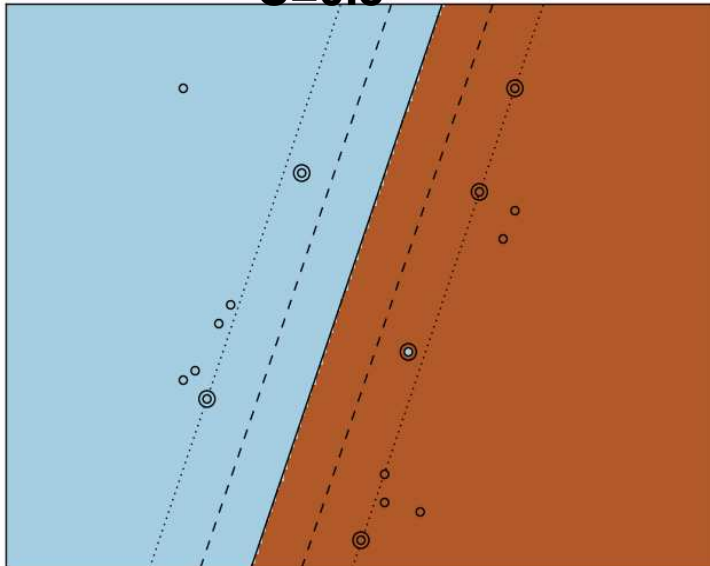
C=0.1



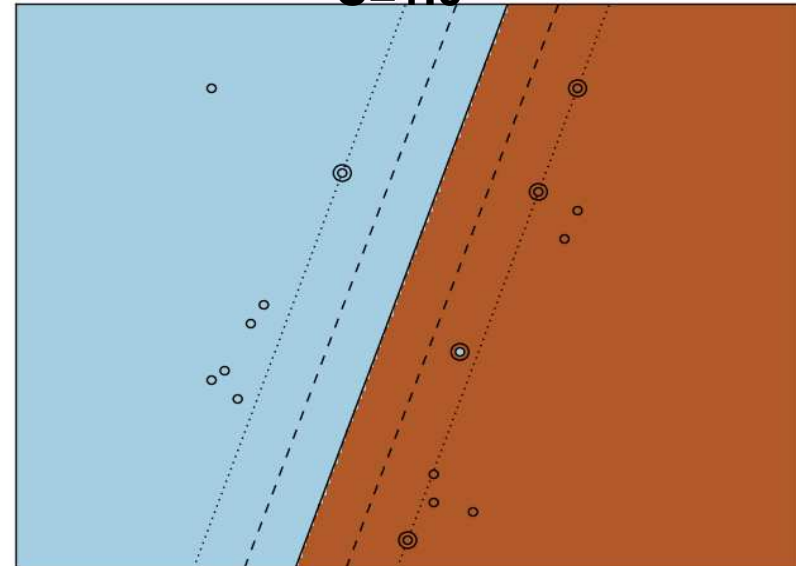
C=0.2



C=0.5



C=1.0

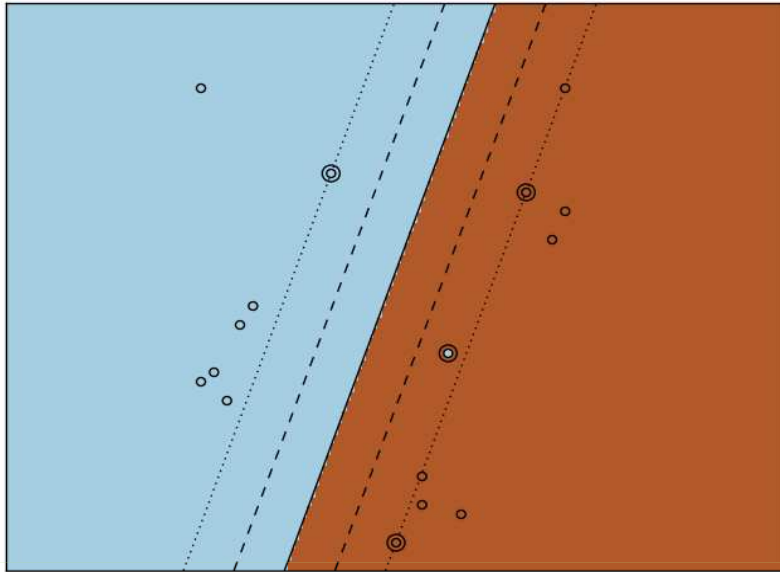




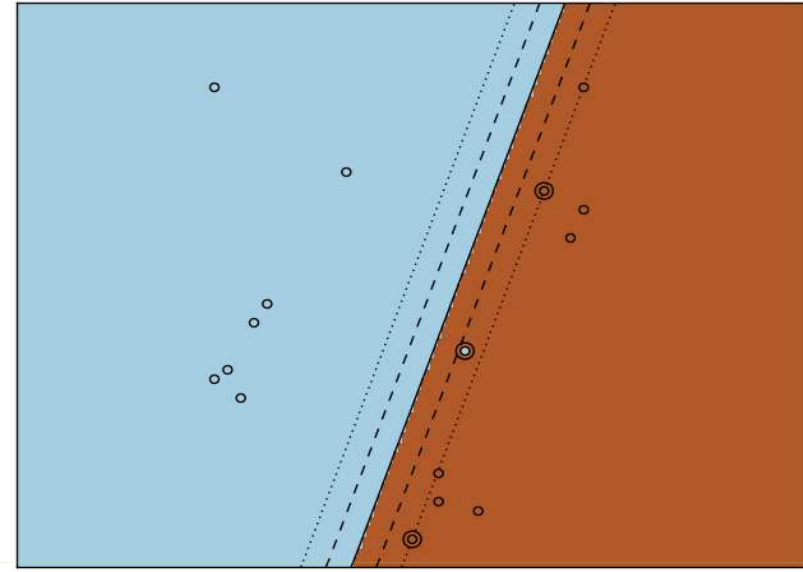
PARAMETRO C EN SVM LINEAL



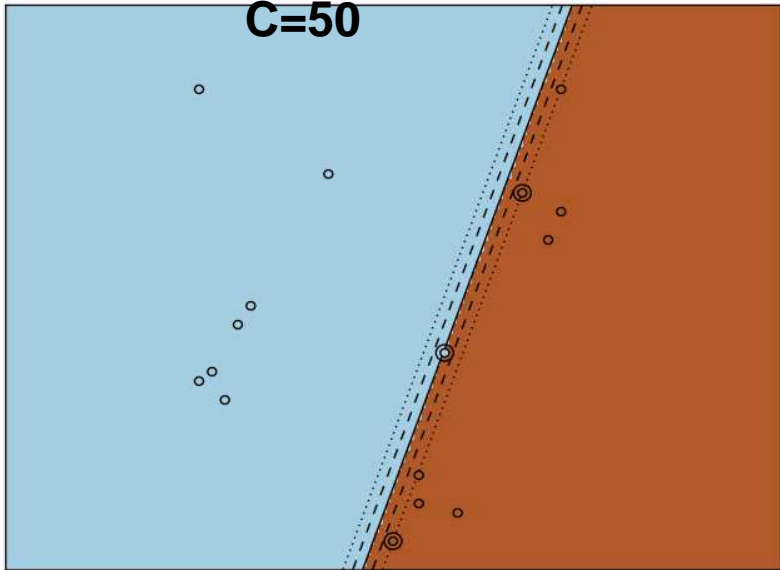
C=10



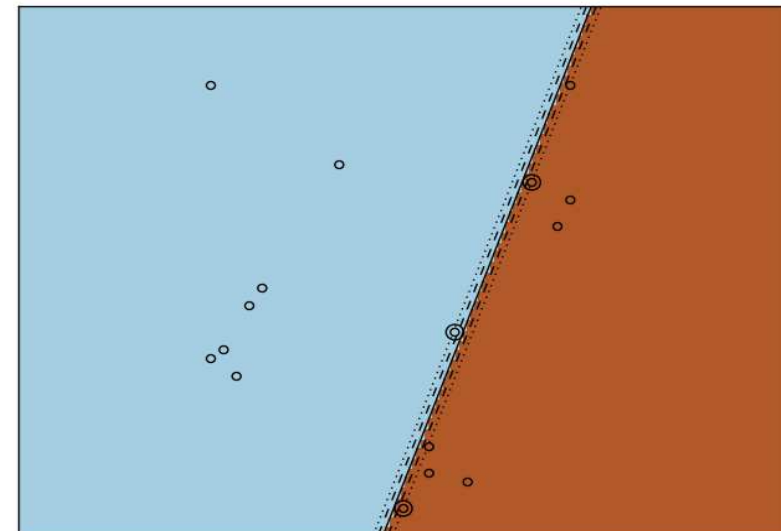
C=20



C=50



C=100





Maquina de Vectores Soporte, SVM, no lineal



Con una SVM lineal solo pueden obtenerse fronteras lineales

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

- Transformar los datos de entrada mediante un conjunto de funciones no lineales $\phi(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \dots, \phi_M(\mathbf{x}))^T$ y luego aplicar la SVM lineal.

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

- **Problema dual:**

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

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

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

- **w₀ óptimo:**

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

- **Clasificador óptimo:**

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



El truco del núcleo (1)



Problema:

- Ya vimos que uno de los problemas de trabajar con **funciones no lineales** $\phi(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \dots, \phi_M(\mathbf{x}))^T$ consiste en que generalmente el número de ellas, M , es muy grande.
- El coste computacional de calcular los productos escalares $\Phi(\mathbf{x}_i)^T \cdot \Phi(\mathbf{x}_j)$ es también M y por tanto muy grande.

• Observación:

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

• Truco del núcleo:

- El truco del núcleo consiste en calcular $\Phi(\mathbf{x}_i)^T \cdot \Phi(\mathbf{x}_j)$ mediante alguna función de las muestras originales $k(\mathbf{x}_i, \mathbf{x}_j)$. A esta función se le llama **función núcleo**.



El Truco del Núcleo (2) : Ejemplos



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

$$\begin{aligned}\Phi(\mathbf{x})^T \cdot \Phi(\mathbf{x}') &= (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T \cdot (x_1'^2, \sqrt{2}x_1'x_2', x_2'^2) = \\ &= x_1^2x_1'^2 + 2x_1x_2x_1'x_2' + x_2^2x_2'^2 = \\ &= (x_1x_1' + x_2x_2')^2 = (\mathbf{x}^T \cdot \mathbf{x}')^2 = k(\mathbf{x}, \mathbf{x}')\end{aligned}$$

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

siendo $\mathbf{x}=(x_1, x_2)$ y $\mathbf{x}'=(x_1', x_2')$

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

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

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



El Truco del Núcleo (3)



- 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:**

$$\max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$
$$s.a. \sum_{i=1} \alpha_i y_i = 0, \quad \alpha_i \geq 0, i=1, \dots, n$$

– **w_0 óptimo:**

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

– **Clasificador óptimo:**

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



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:

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + 1)^p$$

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

$$k(\mathbf{x}, \mathbf{x}') = \tanh(c\mathbf{x}^T \mathbf{x}' - d)$$



El Truco del Núcleo (4)



- La función núcleo $k(\mathbf{x}, \mathbf{y})$ mide la similitud entre las muestras \mathbf{x} e \mathbf{y} .
- No toda función $k(\mathbf{x}, \mathbf{y})$ puede ser utilizada como función núcleo.

Debe satisfacer la denominada condición de Mercer.

- Algunas funciones núcleo:

- Lineal: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$
- Polinomial: $k(\mathbf{x}, \mathbf{y}) = (\lambda(\mathbf{x}^T \mathbf{y}) + \theta)^p, p = 1, 2, \dots$
- Tangente hiperbólica: $k(\mathbf{x}, \mathbf{y}) = \tanh(\lambda(\mathbf{x}^T \mathbf{y}) - \theta)$
- Función de base radial: Núcleo con características de dimensión infinita

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

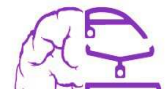
- Observación:
- La forma final del clasificador lineal:

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

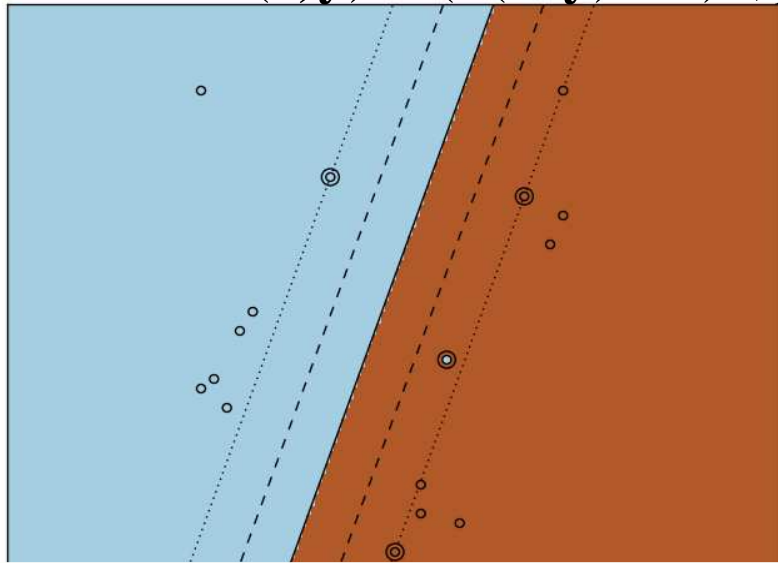
indica que el clasificador $g(\mathbf{x})$ está siendo aproximado por las funciones $k(\mathbf{x}_i, \mathbf{x})$ correspondientes a los vectores soporte.



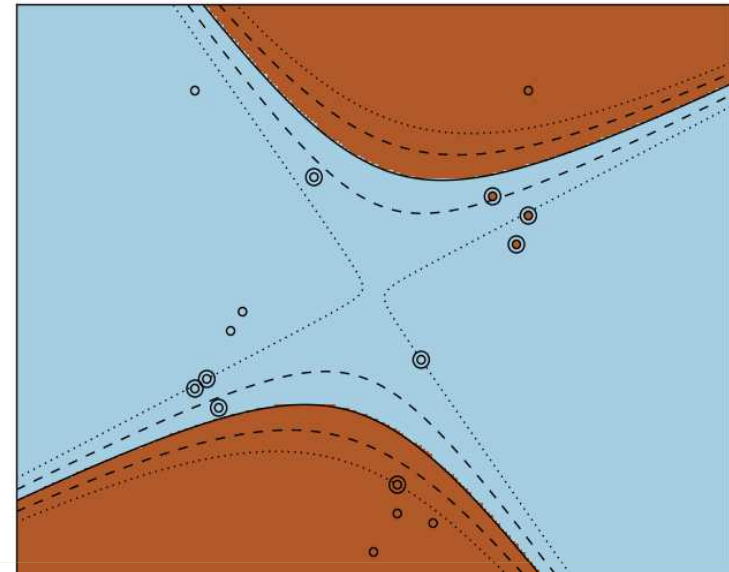
Kernel polinómico, $\theta=1$



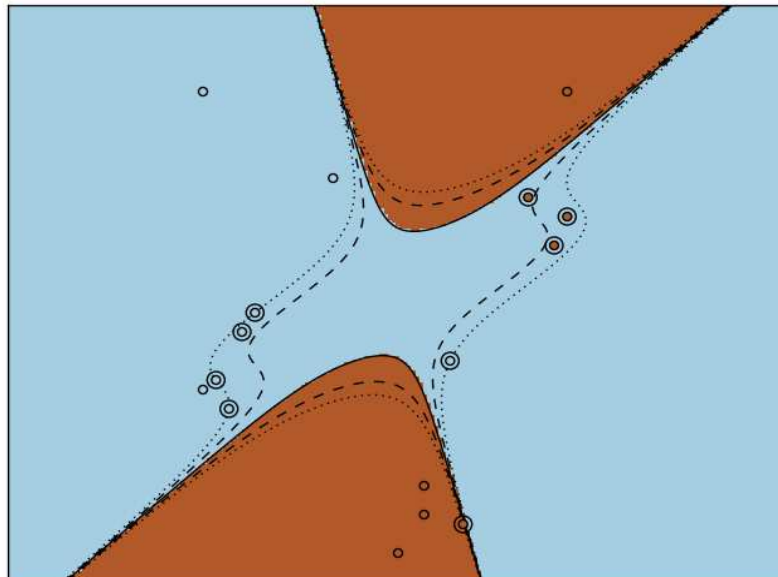
Grado $p=1$ $k(\mathbf{x}, \mathbf{y}) = (\lambda(\mathbf{x}^T \mathbf{y}) + \theta)^p, p = 1, 2, \dots$



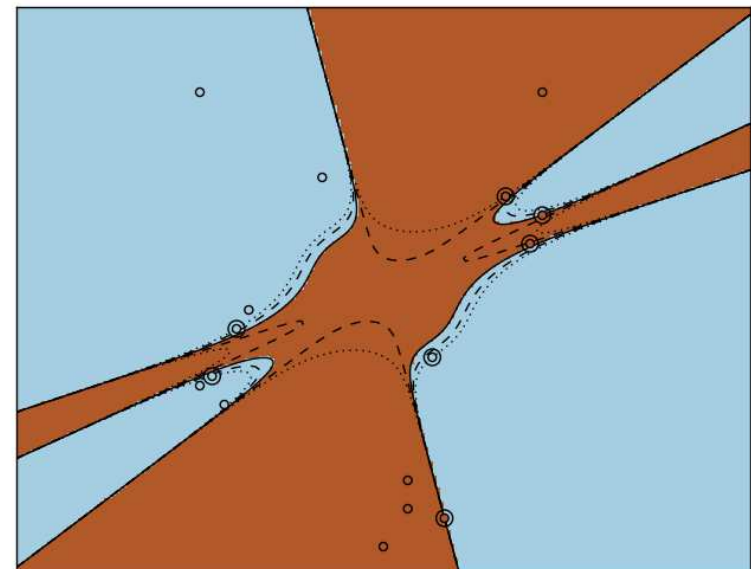
Grado $p=2$



Grado $p=4$



Grado $p=6$



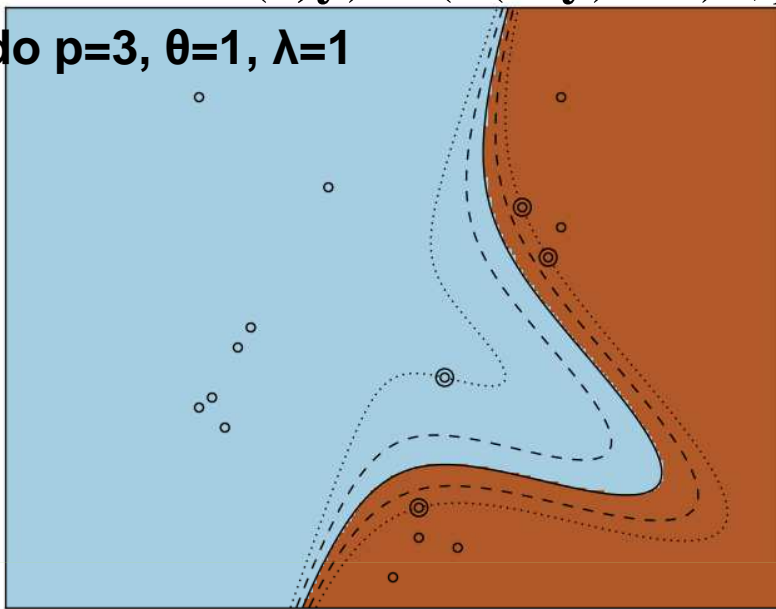


Kernel polinómico, $\theta=1$

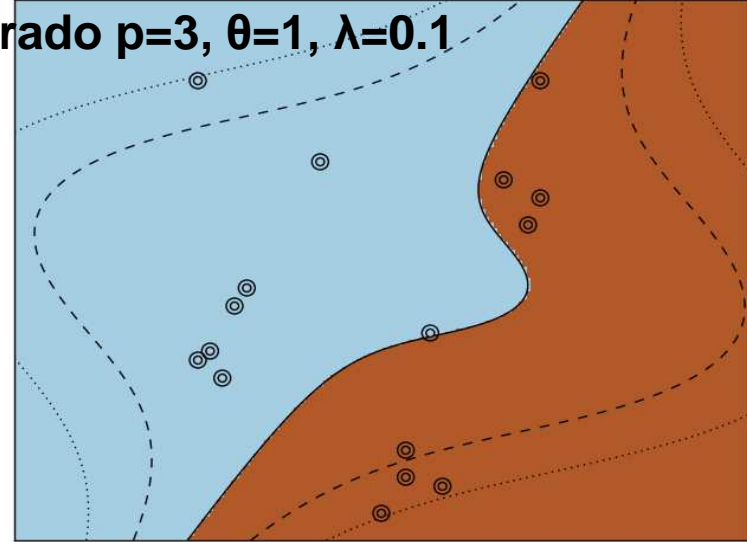


$$k(\mathbf{x}, \mathbf{y}) = (\lambda(\mathbf{x}^T \mathbf{y}) + \theta)^p, p = 1, 2, \dots$$

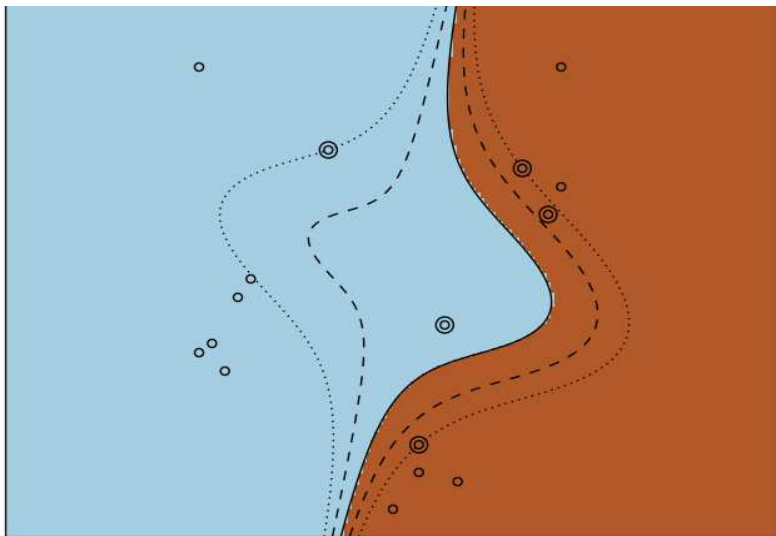
Grado $p=3$, $\theta=1$, $\lambda=1$



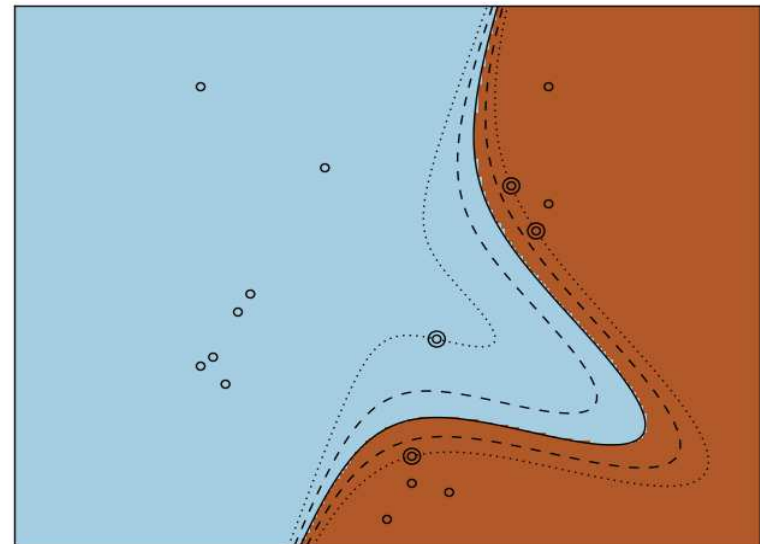
Grado $p=3$, $\theta=1$, $\lambda=0.1$



Grado $p=3$, $\theta=1$, $\lambda=0.5$



Grado $p=3$, $\theta=1$, $\lambda=2$



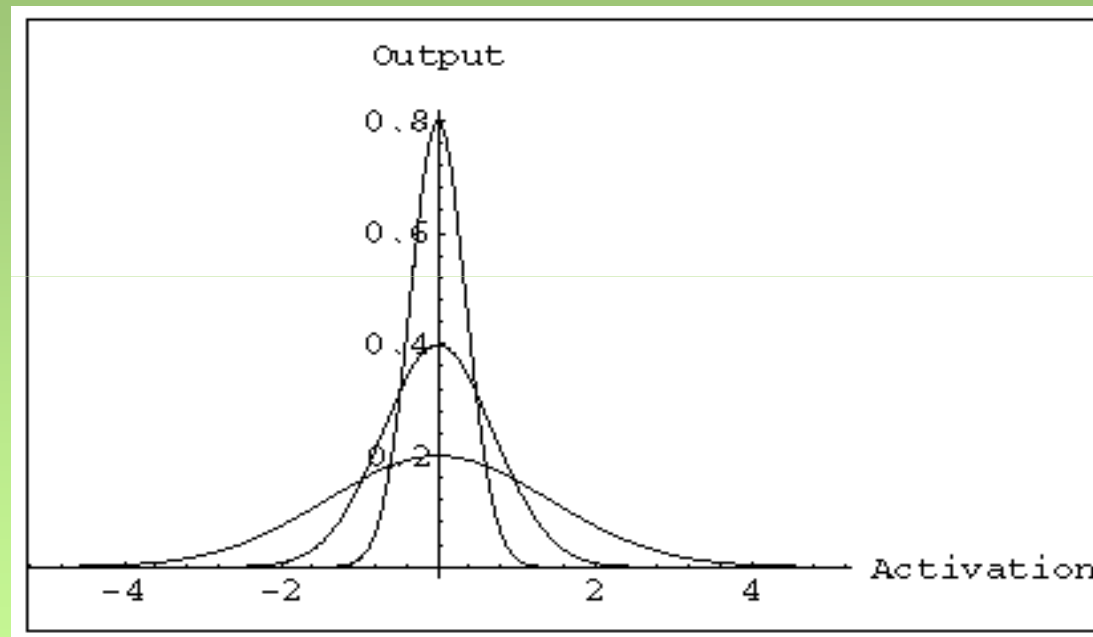


Kernel RBF



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

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



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	Baja
Grande	Picuda	Complejo	Bajo	Alta

Valor de Lambda	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 mayor valor de lambda hace una menor varianza

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

**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**

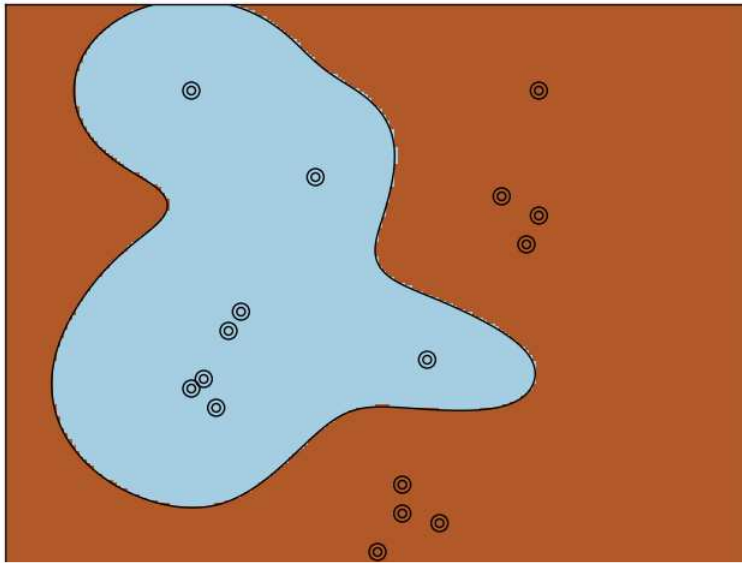
**Hacer pruebas en espacios exponenciales
Utilizar un rango amplio**



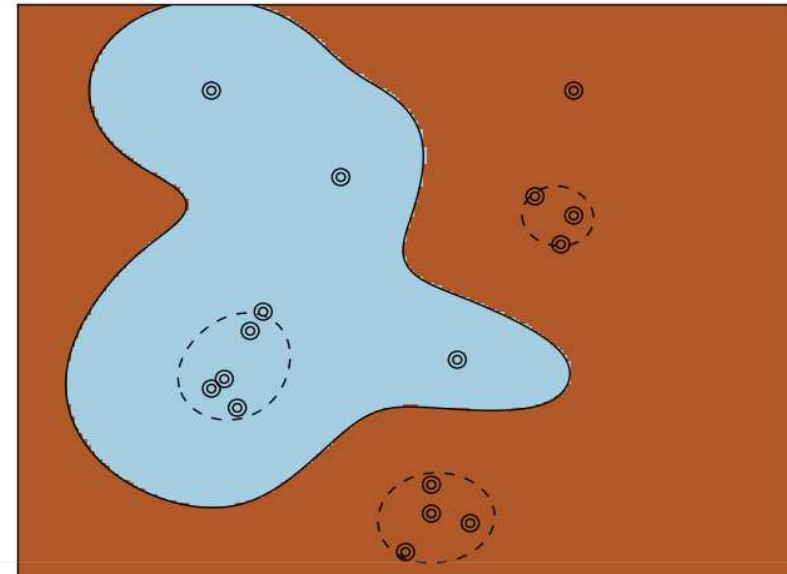
Parámetros del *Kernel* RBF



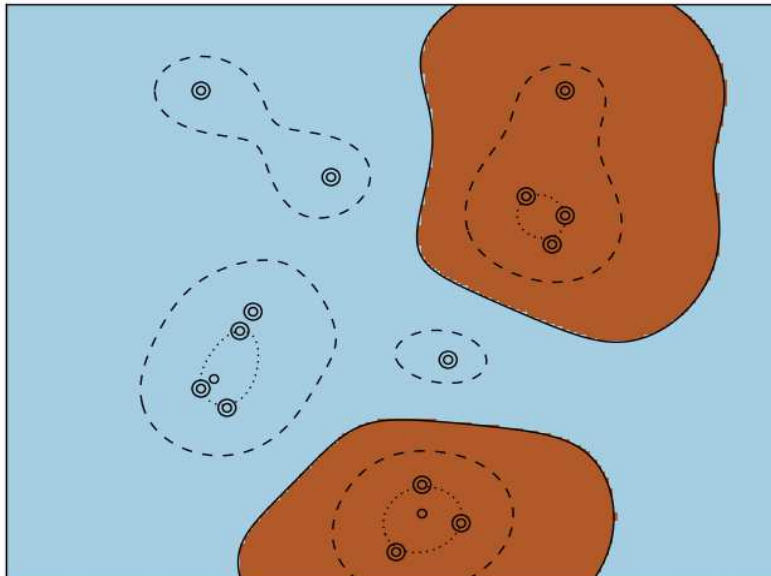
$C=0.05, \lambda=2$



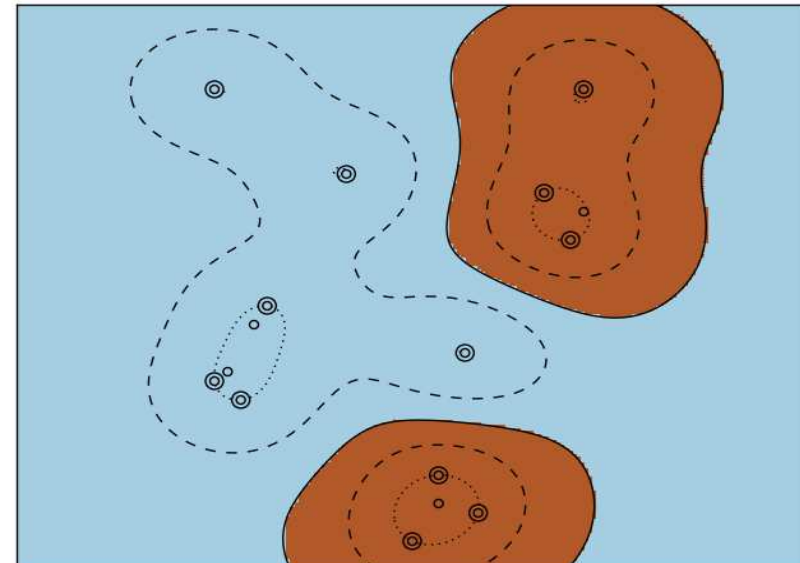
$C=0.2, \lambda=2$



$C=0.6, \lambda=2$

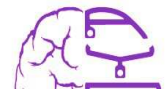


$C=2, \lambda=2$

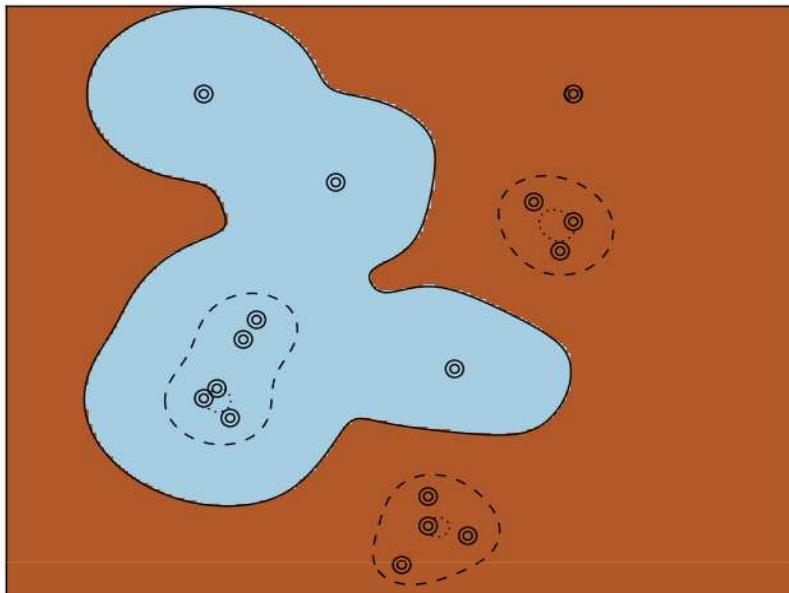




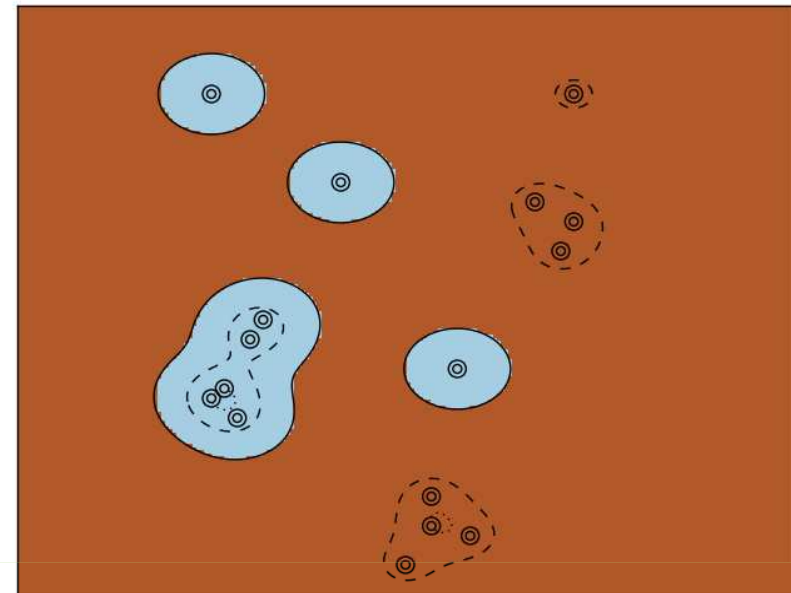
Parámetros del *Kernel* RBF



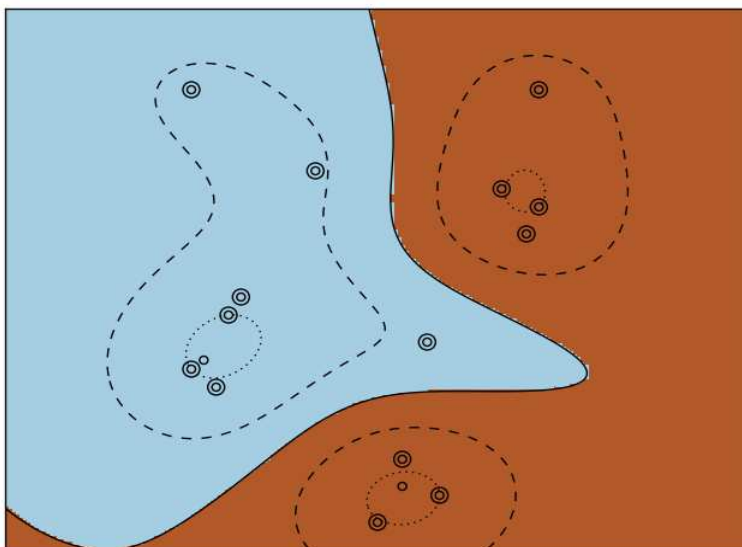
$C=0.5, \lambda=5$



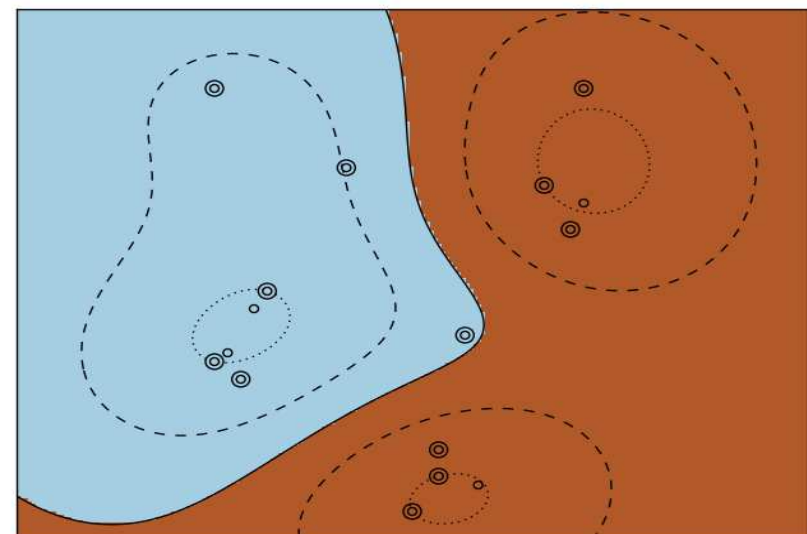
$C=0.5, \lambda=10$



$C=0.5, \lambda=1$



$C=0.5, \lambda=0.5$

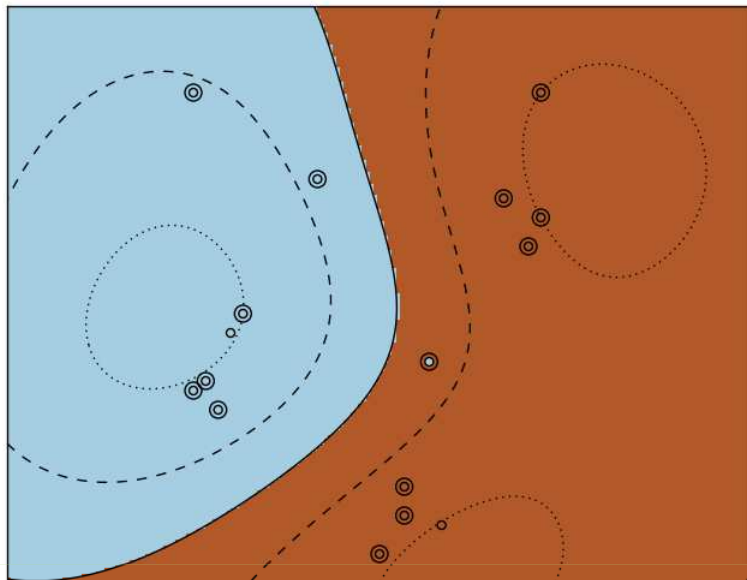




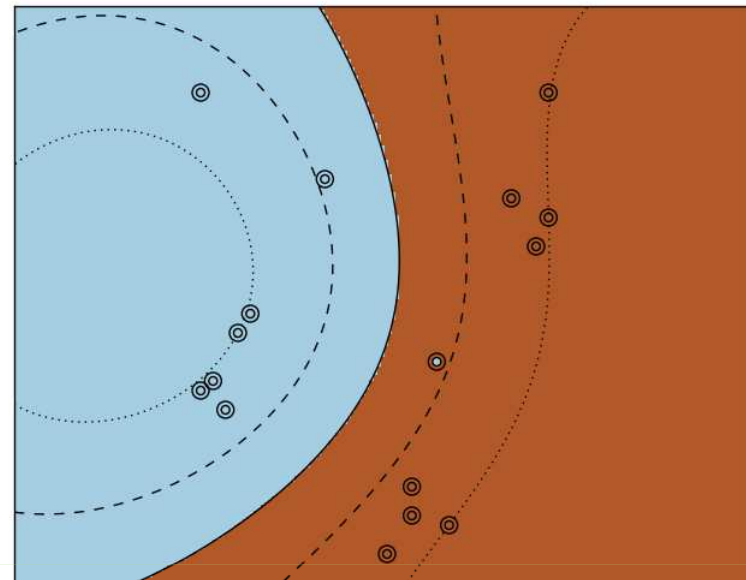
Parámetros del *Kernel* RBF



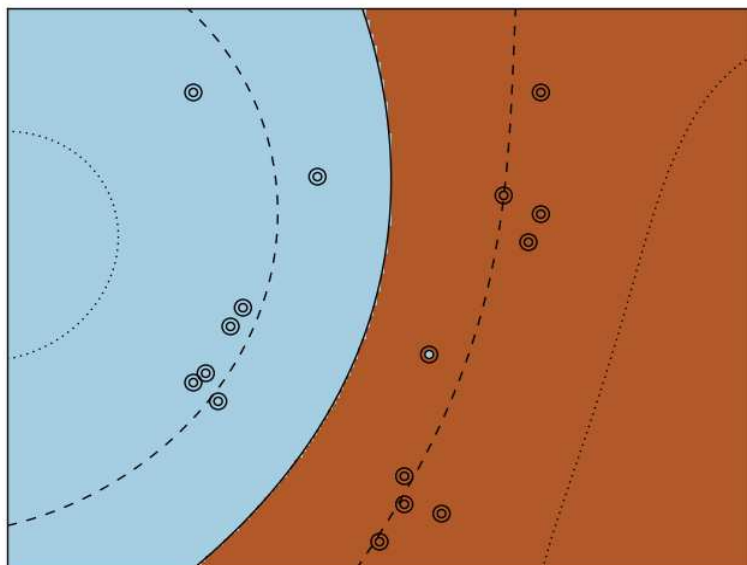
$C=0.5, \lambda=0.2$



$C=0.5, \lambda=0.1$



$C=0.5, \lambda=0.05$

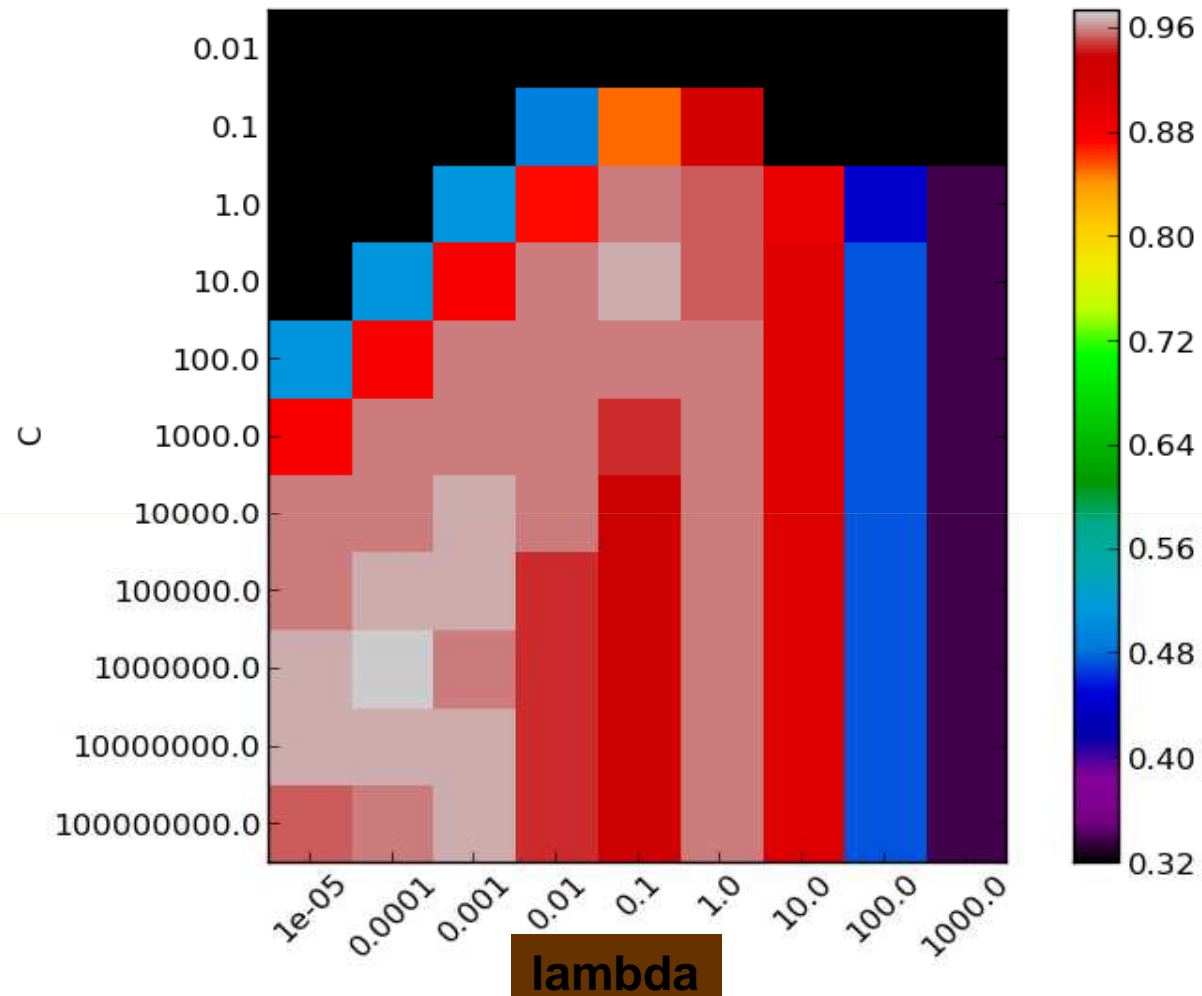




Parámetros del *Kernel* RBF



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 las sucesiones:

$$C \in \{10^{-3}, 10^{-2}, \dots, 10^3\} \text{ y } \lambda \in \{10^{-3}, 10^{-2}, \dots, 10^3\}$$



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 contruídos

Soporta varios núcleos, incluidos los personalizados

SVCLineal: uno-contra-el-resto, OAA

n clasificadores entrenados



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

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

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

› 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:

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

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

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



Ejemplo no lineal (2): XOR



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

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

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

Las cinco ecuaciones se obtienen derivando la función lagrangiana

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

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

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

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

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

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



Ejemplo no lineal (3): XOR



– **w óptimo:**

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

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

– **Constante w0 óptima:**

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

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

– **Clasificador óptimo:**

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

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



Ejemplo no lineal (4): XOR



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

– **Cjto. entrenamiento:** función XOR $\mathbf{x}_1=(0,1)$, $\mathbf{x}_2=(1,0)$ ambos pertenecen a C_1

$\mathbf{x}_3=(0,0)$, $\mathbf{x}_4=(1,1)$, 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**

$$\begin{aligned}\Phi(\mathbf{x}_i)^T \cdot \Phi(\mathbf{x}_j) &= (1 + \mathbf{x}_i^T \mathbf{x}_j)^2 = k(\mathbf{x}_i, \mathbf{x}_j) = \\ &= (1 + 2x_1x_1' + 2x_2x_2' + x_1^2x_1'^2 + 2x_1x_2x_1'x_2' + x_2^2x_2'^2)\end{aligned}$$

– **La función de núcleo es entonces**

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

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

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

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

.....

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



Ejemplo no lineal (5): XOR



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

El problema es ahora:

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

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

Las cinco ecuaciones son

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

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

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

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

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

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



Ejemplo no lineal (6): XOR



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

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

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

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

.....

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

– **Constante w_0 óptima:**

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

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

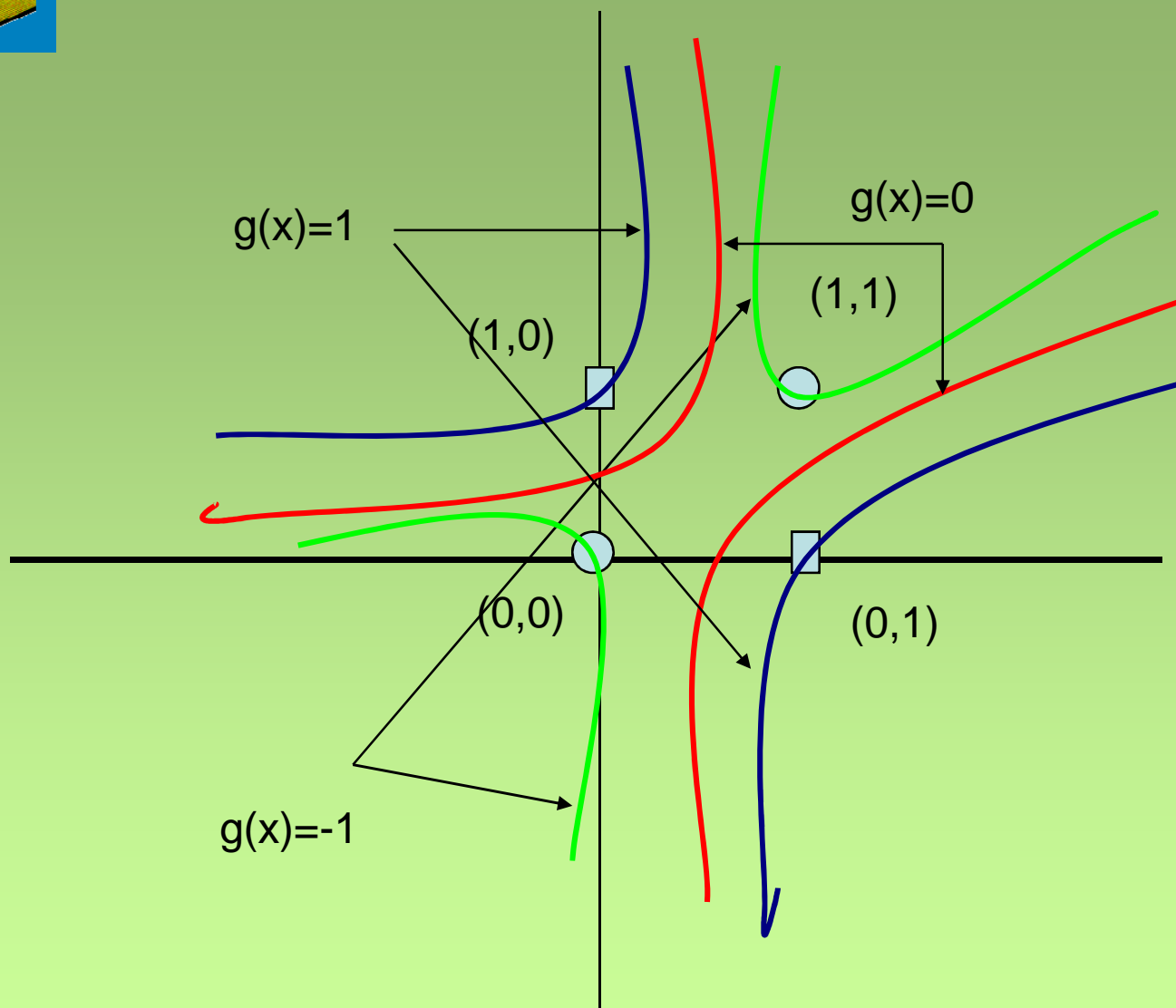
– **Clasificador óptimo:**

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

$$\begin{aligned} g(\mathbf{x}) &= (8/3)(x_2 + 1)^2 + (8/3)((x_1 + 1)^2 - (10/3) - (6/3)(x_1 + x_2 + 1)^2 - 1 = \\ &= (2/3)(x_1 - 1/2)^2 + (x_2 - 1/2)^2 - 6(x_1 - 1/2)(x_2 - 1/2) - 1/2 \end{aligned}$$



Ejemplo no lineal (7): XOR



Frontera de decisión de la SVM



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 sobre-ajuste en la muestra; también malo!!

También hay que escoger 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 sobre-ajuste en la muestra



Selección de parámetros: Resumen II



Posibles enfoques para obtener C, σ :
Fijar conjuntos de validación
Generar diferentes modelos y evaluarlos sobre dichos conjuntos

Opción 1: búsqueda en rejilla sobre rectángulo

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

Opción 2: búsqueda evolutiva a partir de C_0, σ_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, métodos de construcción eficaces, buen software de acceso público
- Juntan ideas diversas e importantes: márgenes en clasificación, optimización lagrangiana, espacios de Hilbert reproductores

- **No están exentas de problemas:**

- **Explotación costosa en test:**

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

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

- **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: $K(x, x_0) = x^T \cdot x_0$

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 multicore, paralelización)



BIBLIOGRAFIA Y SOFTWARE



Bibliografía:

- 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

Software:

Exploración: WEKA, Matlab

Explotación: LIBSVM

Big Data más abierta: Pegasos, LibLinear, ...



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
Grupo de Investigación AYRNA

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

2019-2020