Introducción

- **En el tema anterior se ha visto como aproximar la frontera de decisión óptima $h(x)=P(w_1 \mid x)-P(w_2 \mid x)$ mediante:**
  - Una función lineal: $g(\boldsymbol{x})= \boldsymbol{w}^T\boldsymbol{x}+w_0 = w_1 x_1 + w_2 x_2 +...+ w_d x_d + w_0$
  - Una combinación de función lineal y logística: $1/(1+\exp(- g(\boldsymbol{x})))$
- **En ambos casos la frontera de decisión generada es lineal.**

- **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 del vector soporte.

Fernando Pérez Nava

Clasificador Polinomial (1)

- **¿Cómo transformar el clasificador lineal para obtener fronteras de decisión no lineales?**
- **Una idea simple:**
  - La forma del clasificador lineal es:

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

  - Introduzcamos los términos de grado 2:

    $g(\boldsymbol{x}) = w_{11} x_1^2 + w_{12} x_1 x_2 + \ldots + w_{1d} x_1 x_d + w_{21} x_2 x_1 + \ldots + w_{dd} x_d^2 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + w_0$

    Este tipo de función ya la encontramos en el Tema 2: es una función discriminante cuadrática.

  - Podemos seguir con el grado 3:

    $g(\boldsymbol{x}) = w_{111} x_1^3 + w_{112} x_1^2 x_2 + \ldots + w_0$

    y seguir hasta un grado arbitrario.

Fernando Pérez Nava

Clasificador Polinomial (2)

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

$$g(\boldsymbol{x}) = w_1 \phi_1(\boldsymbol{x}) + w_2 \phi_2(\boldsymbol{x}) + \ldots + w_d \phi_{d^*}(\boldsymbol{x}) + w_0$$

- **Ejemplo:**

Si la función g es:

$$g(\boldsymbol{x}) = w_{11} x_1^2 + w_{12} x_1 x_2 + w_{21} x_2 x_1 + w_{22} x_2^2 + w_1 x_1 + w_2 x_2 + w_0$$

Entonces definiendo:

$$\phi_1(\boldsymbol{x}) = x_1^2,\ \phi_2(\boldsymbol{x}) = x_1 x_2,\ \phi_3(\boldsymbol{x}) = x_2 x_1,\ \phi_4(\boldsymbol{x}) = x_2^2,\ \phi_5(\boldsymbol{x}) = x_1,\ \phi_6(\boldsymbol{x}) = x_2$$

Podemos escribir:

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

Fernando Pérez Nava

Función Discriminante Lineal Generalizada (FDLG)

- **Llamaremos Función Discriminante Lineal generalizada a toda función discriminante con la forma:**

$$g(\boldsymbol{x}) = w_1\,\phi_1(\boldsymbol{x}) + w_2\,\phi_2(\boldsymbol{x}) + \ldots + w_d\,\phi_{d^*}(\boldsymbol{x}) + w_0 = \sum_{i=1}^{d^*} w_i \phi_i(\boldsymbol{x}) + w_0$$

- **Las funciones ϕ pueden ser polinomiales o de otro tipo:**
  - Gausianas, Splines, 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. (1986)
- **De hecho, ya la hemos usado con las ventanas de Parzen:**

$$\hat{p}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\boldsymbol{x} - \boldsymbol{x}_i)$$

Fernando Pérez Nava

FDLG: Notación y Representación Gráfica

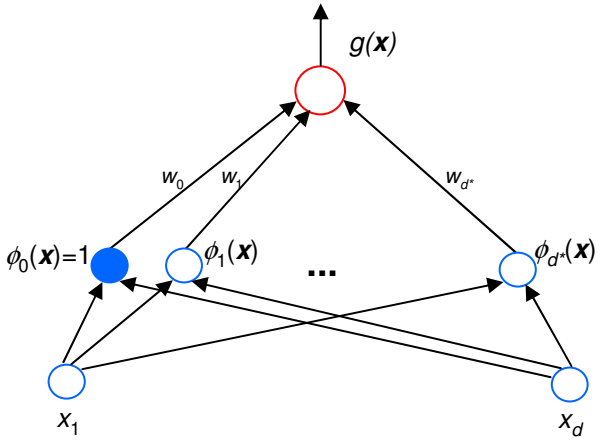
- **Notación:**
  - Tal y como ocurrió con las FDL escribiremos:

    $g(\boldsymbol{x})= w_1\ \phi_1(\boldsymbol{x})+ w_2\ \phi_2(\boldsymbol{x})+\ldots+ w_d\ \phi_{d^*}(\boldsymbol{x})+ w_0\ 1=$

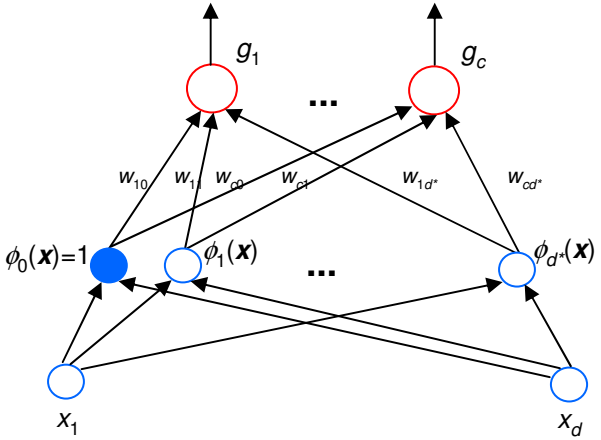
    $g(\boldsymbol{x})= w_1\ \phi_1(\boldsymbol{x})+ w_2\ \phi_2(\boldsymbol{x})+\ldots+ w_d\ \phi_{d^*}(\boldsymbol{x})+ w_0\ \phi_0(\boldsymbol{x})$

    donde $\phi_0(\boldsymbol{x})$ es la función que siempre vale uno.
  - Entonces: $g(\boldsymbol{x})= \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x})$

    $\boldsymbol{w}=(w_1,\ w_2,\ldots,w_{d^*},\ w_0)^{\mathsf{T}},\quad \phi(\boldsymbol{x})=(\phi_1(\boldsymbol{x}),\ \phi_2(\boldsymbol{x}),\ldots,\phi_{d^*}(\boldsymbol{x}),\phi_0(\boldsymbol{x}))^{\mathsf{T}}$
- **Representación gráfica:**



FDL generalizada: Representación gráfica (dos clases)

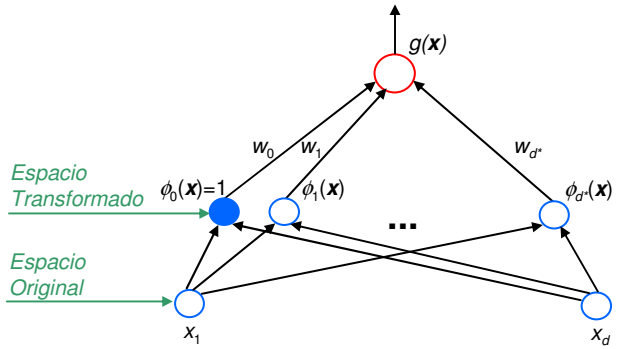


FDL generalizada: Representación gráfica (*c* clases)

Fernando Pérez Nava

FDLG: Entrenamiento

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

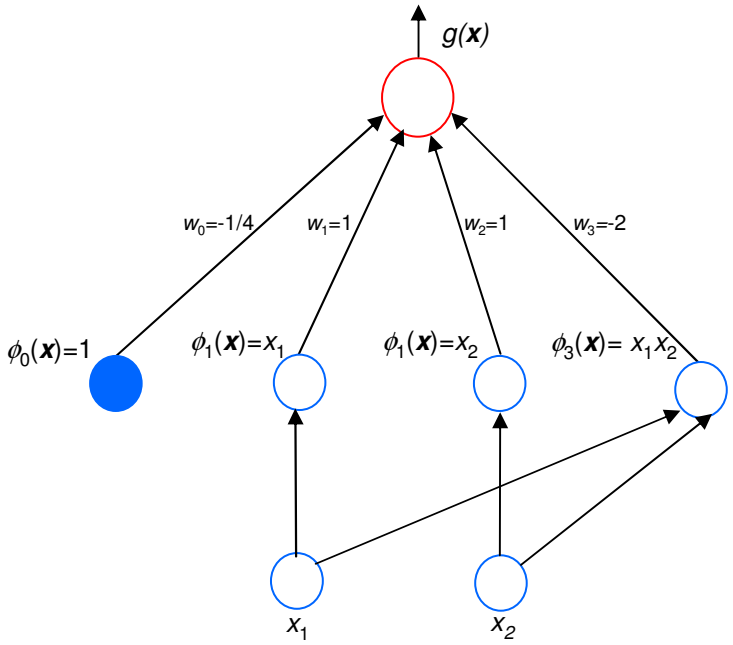


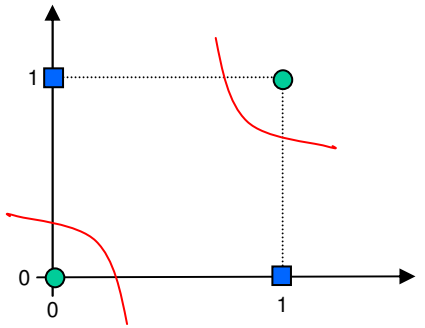
- **Por tanto, el esquema de aprendizaje es:**
  - Paso 1:
    Transformar los datos de entrada con las funciones ϕ
  - Paso 2:
    Aplicar a los datos transformados alguno de los métodos del tema anterior

Fernando Pérez Nava

Clasificador Cuadrático:Ejemplo

- **Un clasificador cuadrático capaz de resolver el problema del XOR:**





El Problema del XOR: Solución cuadrática

- **La función discriminante es:**

$$g(\boldsymbol{x}) = -1/4 - 2\,x_1 x_2 + x_1 + x_2$$

Fernando Pérez Nava

Separabilidad Lineal: Teorema de Cover

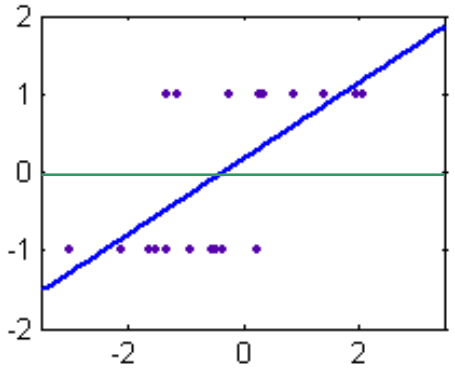
- **Observación:**
  - El problema del XOR se resuelve porque hemos pasado del espacio de características original de dimensión 2 definido por $\boldsymbol{x}=(x_1, x_2)$ a un espacio de transformado de dimensión 4 definido por $\phi(\boldsymbol{x})=(1, x_1, x_2, x_1x_2)$ donde el problema sí es linealmente separable.
- **Puede probarse que el incremento de la dimensionalidad hace más fácil lograr la separabilidad lineal**
  - Teorema de Cover

    La probabilidad de que dos clases sean linealmente separables se aproxima a 1 cuando la dimensionalidad del espacio de características d tiende a infinito y el número de muestras crece de limitado por $2(d+1)$.
- **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 para determinar los pesos.

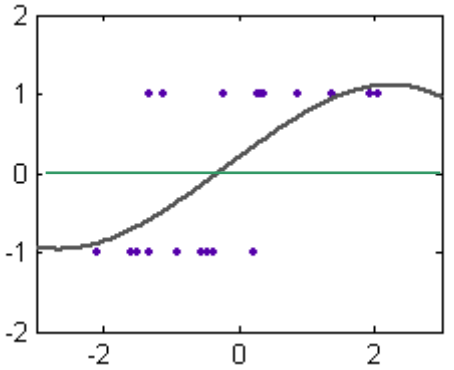
Fernando Pérez Nava

Problemas...

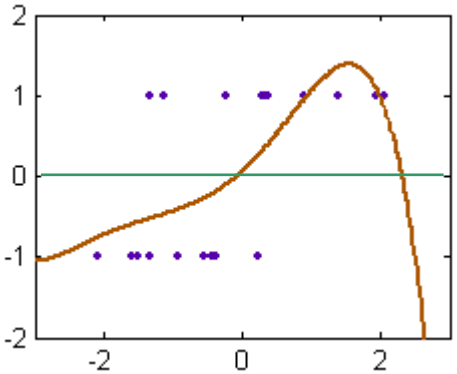
- **El número de parámetros a estimar es inmenso.**
  - Por ejemplo para diez características y un polinomio de grado 10 el número de parámetros es: 184.756
- **La presencia de un número tan grande de parámetros hace que el clasificador sufra de sobreajuste:**



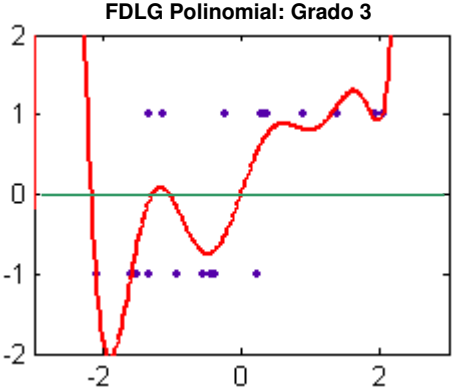
FDLG Polinomial: Grado 1



FDLG Polinomial: Grado 3



FDLG Polinomial: Grado 5



FDLG Polinomial: Grado 9

Fernando Pérez Nava

Sobreajuste: Regularización

- **Observación:**
  - En general, cuando se produce sobreajuste los pesos son cada vez más grandes.
- **Podemos introducir como información adicional que pensamos que el vector de pesos debe ser pequeño.**
  - La inserción de información adicional en el esquema estadístico bayesiano se hace mediante la probabilidad a priori.
  - Por tanto definiremos: $p(\boldsymbol{w})=N(\mathbf{0};\sigma^2\mathbf{I})$
- **Aprendizaje:**
  - Puesto que hay información a priori se utiliza la estimación MAP (ya se ha visto como se hace en el tema anterior para regresión logística).
  - ¿Cuál es el valor de σ^2 ?

    Como siempre se determina mediante un conjunto de entrenamiento y otro de testeo
- **Este esquema se suele llamar regularización.**
- **Aún así queda el problema de estimar un enorme número de parámetros.**

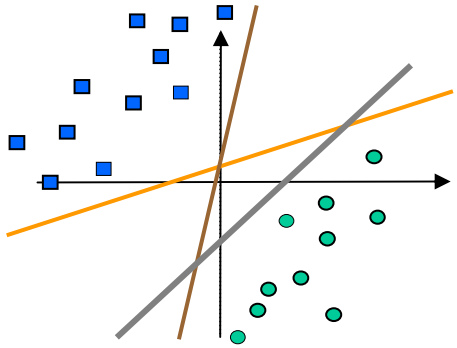
Máquina del Vector Soporte (MVS)

- **La MVS "*Support Vector Machine (SVM)*" es un clasificador desarrollado por Vapnik (1995).**
- **Está basada en la teoría de aprendizaje computacional de Vapnik-Chernovenkis (VC).**
- **Además de utilizarse en problemas de clasificación puede extenderse a:**
  - Regresión
  - Estimación de funciones de densidad.
- **Como clasificador está diseñado tanto para proporcionar una alta capacidad de generalización como para trabajar en espacios de alta dimensionalidad.**
- **Combina una sólida fundamentación teórica con buenos resultados en problemas reales.**
- **Representa actualmente el "estado del arte" en clasificación.**

Fernando Pérez Nava

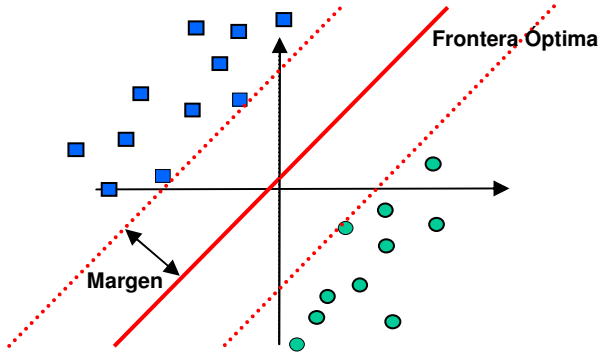
MVS: Caso Lineal (CL)

- **Caso Linealmente Separable:**
  - Dados conjuntos linealmente separables ¿Qué frontera lineal proporciona una mejor capacidad de generalización?



Fronteras de separación lineal

  - Al abordar este problema desde la teoría de aprendizaje VC se demuestra que la frontera óptima es aquella que proporciona un mayor margen



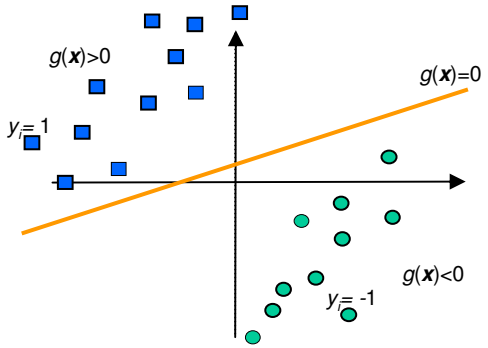
Frontera óptima según la teoría VC

MVS:(CL).Separación Lineal

- **Recordamos:**
  - Toda frontera lineal se escribe como: $g(\boldsymbol{x})= \boldsymbol{w}^T\boldsymbol{x}+w_0 =0$
  - La regla de decisión es:

    Elegir w_1 si $g(\boldsymbol{x}) \geq 0$; Elegir w_2 si $g(\boldsymbol{x}) \leq 0$

    (Si $g(\boldsymbol{x})$ es positivo elegir w_1 , si $g(\boldsymbol{x})$ es negativo elegir w_2).
- **Representaremos el signo deseado para cada elemento del conjunto de entrenamiento como:**
  - $y_i = 1$ si $\boldsymbol{x}_i \in w_1$, $y_i = -1$ si $\boldsymbol{x}_i \in w_2$
- **Para que todo el conjunto de entrenamiento esté bien clasificado es necesario que:**
  - El signo deseado y el obtenido coincidan: $y_i(\boldsymbol{w}^T\boldsymbol{x}_i+w_0)>0$, $i=1...n$

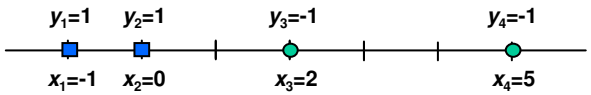


Signos deseados y obtenidos

Fernando Pérez Nava

Un Ejemplo Inicial (1)

- **Clasificación unidimensional: $g(x)=wx+w_0$**
  - Conjunto de entrenamiento: $x_1=-1$, $x_2=0 \in w_1$; $x_3=2$, $x_4=5 \in w_2$
  - Signos deseados: $y_1=1$, $y_2=1$, $y_3=-1$, $y_4=-1$



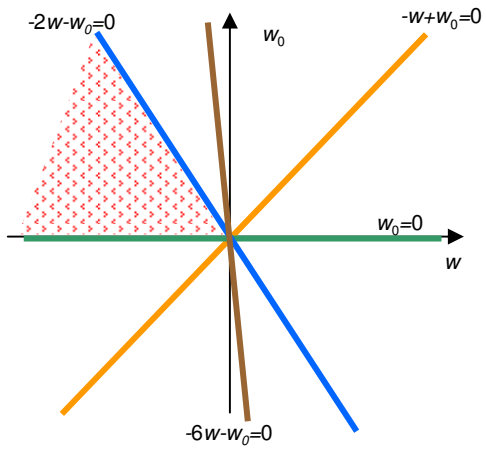
  - Condiciones de separabilidad lineal:

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

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

$-(wx_3+w_0)>0$: $-2w-w_0>0$

$-(wx_4+w_0)>0$: $-6w-w_0>0$



Región sombreada: valores de w y w_0 que producen separación lineal

Fernando Pérez Nava

MVS:(CL).Formulación Inicial

- **En el caso de separabilidad lineal:**
  - La distancia de un punto $\boldsymbol{x}_i$ a la frontera lineal $g(\boldsymbol{x})= \boldsymbol{w}^T\boldsymbol{x}+w_0 =0$ es:

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

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

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

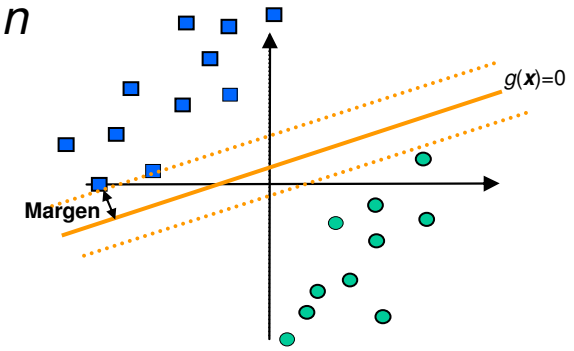
para los valores de $\boldsymbol{w}$ y w_0 que representan fronteras que separan los puntos del conjunto de entrenamiento:

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

- **Por lo tanto se obtiene un problema de optimización con restricciones:**

$$\max_{\boldsymbol{w}} \quad \text{margen}(\boldsymbol{w}, w_0)$$

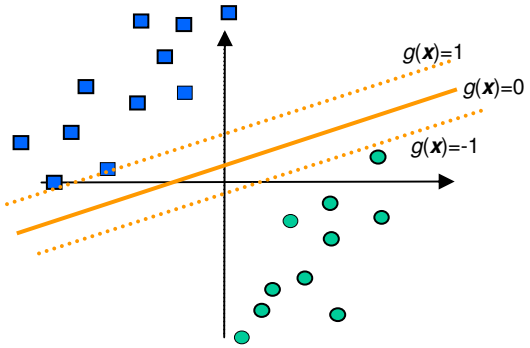
$$\text{s.a} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) > 0 \quad i = 1\ldots n$$



Margen y frontera de decisión

MVS:(CL).Normalización (1)

- **Un problema para buscar la frontera óptima:**
  - Una frontera lineal no tiene una representación única.
  - Si se tiene $g(\boldsymbol{x})= \boldsymbol{w}^T\boldsymbol{x}+w_0 =0$ y se multiplica en ambos miembros de la ecuación por una constante no nula la ecuación no cambia.
  - Es decir, la frontera lineal representada por $g(\boldsymbol{x})= \boldsymbol{w}^T\boldsymbol{x}+w_0 =0$ es equivalente a $\lambda(\boldsymbol{w}^T\boldsymbol{x}+w_0)=0$, $\lambda \neq 0$.
  - 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 $|g(\boldsymbol{x}_i)|= y_i\,(\boldsymbol{w}^T\boldsymbol{x}+w_0\,)$ sea igual a 1.



Normalización

MVS:(CL).Normalización (2)

- **De esta forma se tiene que:**
  - El margen se calcula de forma simple como:

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

  - Las ecuaciones $y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + w_0) \geq 0$ pasan a $y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + w_0) \geq 1$ pues el menor valor que toma $y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + w_0)$ es uno.
- **Por tanto se obtiene el problema:**

$$\max_{\boldsymbol{w}} \quad 1/(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w})^{1/2}$$

$$\text{s.a} \quad y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + w_0) \geq 1 \quad i = 1 \ldots n$$

- **Para hacer máximo el cociente podemos hacer mínimo el denominador con lo que se obtiene el problema equivalente:**

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}$$

$$\text{s.a} \quad y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + w_0) \geq 1 \quad i = 1 \ldots n$$

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

Un Ejemplo Inicial (2)

- **Clasificación unidimensional: $g(x)=wx+w_0$**
  - Conjunto de Entrenamiento :x_1=-1, x_2=0 $\in w_1$; x_3=2, x_4=5 $\in w_2$
  - Signos deseados: y_1=1, y_2=1, y_3=-1, y_4=-1
- **Normalización: Ejemplo**
  - La frontera determinada por $g(x)=wx+w_0=0$ con $(w,w_0)=(-1,3/2)$ cumple las restricciones de separación lineal.

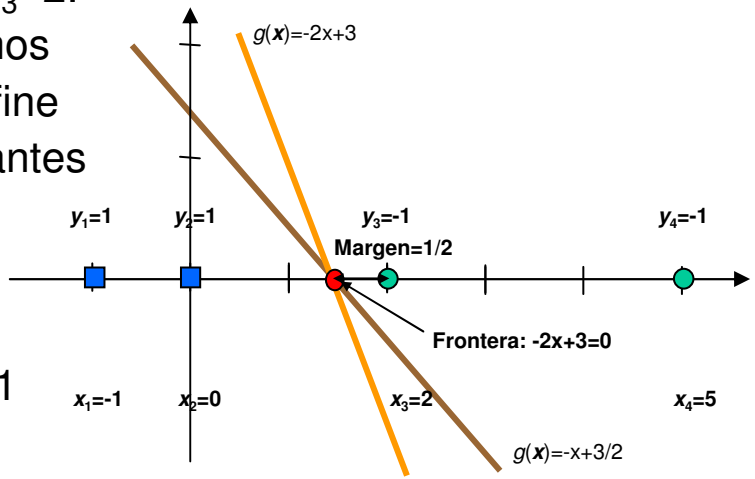
    Para normalizarla, el menor valor de $|g(x)|= y_i\,(wx_i+w_0\,)$ debe ser igual a 1.
  - Calculamos los valores:

    $y_1\,(wx_1+w_0\,)=5/2$; $y_2\,(wx_2+w_0\,)=3/2$; $y_3\,(wx_3+w_0\,)=1/2$; $(wx_4+w_0\,)=7/2$

    El menor valor es 1/2 y ocurre para x_3=2.
  - Entonces dividiendo por 1/2 obtenemos el nuevo vector $(w,w_0)=(-2,3)$ que define exactamente la misma frontera que antes pero ahora:

    $y_1\,(wx_1+w_0\,)=5$; $y_2\,(wx_2+w_0\,)=3$;

    $y_3\,(wx_3+w_0\,)=1$; $(wx_4+w_0\,)=7$
  - Además ahora siempre $y_i\,(wx+w_0\,)\geq 1$ y el margen es $1/\sqrt{w^2}=1/2$



Marrón: $g(x)$ sin normalizar.Naranja: $g(x)$ normalizada

Fernando Pérez Nava

Un Ejemplo Inicial (3):Solución

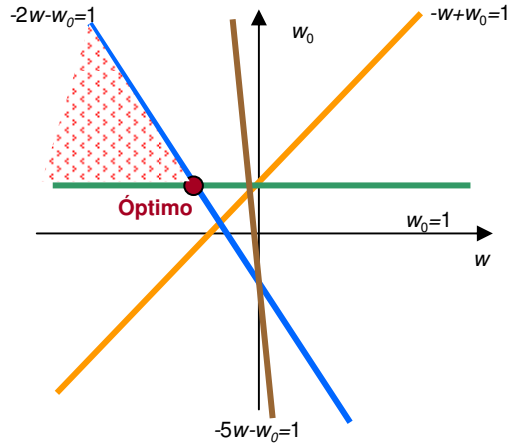
- **Clasificación unidimensional: $g(x)=wx+w_0$**
  - Cjto. Entrenamiento H: $x_1=-1$, $x_2=0 \in w_1$; $x_3=2$, $x_4=5 \in w_2$
  - Signos deseados: $y_1=1$, $y_2=1$, $y_3=-1$, $y_4=-1$
- **Condiciones de separabilidad lineal normalizadas:**

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

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

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

$-(wx_4+w_0) \geq 1$: $-6w-w_0 \geq 1$



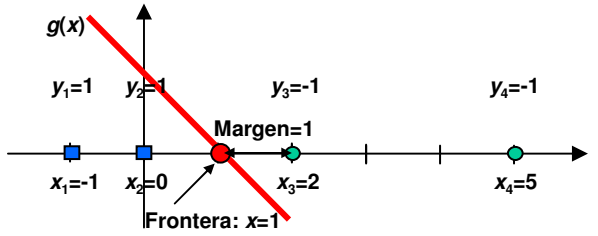
Región sombreada: separabilidad lineal normalizada

- **Función a minimizar:**

$1/2\ (\mathbf{w}^T\mathbf{w})=1/2\ w^2$

- **Valor óptimo:**
  - El menor valor que puede tomar $1/2\ w^2$ en la región sombreada se obtiene con $w=-1$ $w_0=1$.
  - Por tanto el clasificador MVS es:

    $g(x)=-x+1$ y la frontera $g(x)=0$ es $x=1$



Clasificador MVS

MVS:(CL).Resolución Práctica (1)

- **Para resolver el problema:**

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}$$

$$\text{s.a} \quad y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + w_0) \geq 1 \quad i = 1...n$$

- **Se halla su problema dual:**

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

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

- **Una vez resuelto el problema dual:**
  - Si la solución óptima es $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_n)$ el valor óptimo de $\boldsymbol{w}$ es:

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{n}\hat{\alpha}_i y_i \boldsymbol{x}_i$$

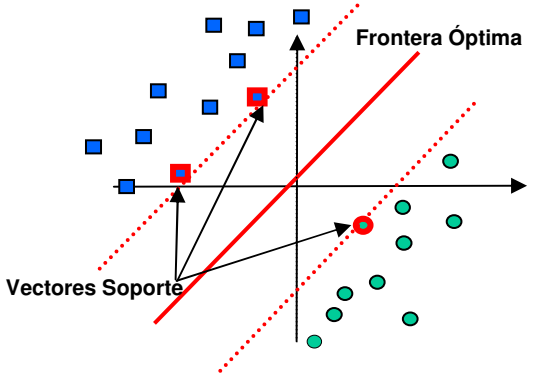
  - El valor óptimo de w_0 se obtiene como:

$$\hat{w}_0 = 1 - \hat{\boldsymbol{w}}^{\mathsf{T}}\boldsymbol{x}_i, \quad \text{con } \boldsymbol{x}_i \in w_1 \text{ y } \hat{\alpha}_i > 0$$

Fernando Pérez Nava

MVS:(CL).Resolución Práctica (2)

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



Vectores Soporte: Representación

  - 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:

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i = \sum_{i \in Sop} \hat{\alpha}_i y_i \boldsymbol{x}_i$$

$$\hat{w}_0 = 1 - \sum_{j \in Sop} \hat{\alpha}_j y_j (\boldsymbol{x}_j^{\mathsf{T}} \boldsymbol{x}_i), \quad \text{con } \boldsymbol{x}_i \in w_1 \text{ y } \hat{\alpha}_i > 0$$

$$g(\boldsymbol{x}) = \hat{\boldsymbol{w}}^{\mathsf{T}} \boldsymbol{x}_i + w_0 = \left(\sum_{i \in Sop} \hat{\alpha}_i y_i \boldsymbol{x}_i \right)^{\mathsf{T}} \boldsymbol{x} + w_0 = \sum_{i \in Sop} \hat{\alpha}_i y_i (\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}) + w_0$$

Fernando Pérez Nava

MVS:(CL).Resolución Práctica (3)

- **Productos escalares**
  - En la formulación del problema dual los elementos del conjunto de entrenamiento solo intervienen a través de sus productos escalares ($\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_j$).

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

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

  - Para calcular w_0 solo aparecen los productos escalares

$$\hat{w}_0 = 1 - \sum_{j \in Sop}\hat{\alpha}_j y_j (\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{x}_i), \quad \text{con } \boldsymbol{x}_i \in w_1 \text{ y } \hat{\alpha}_i > 0$$

  - Para calcular la clase de un vector $\boldsymbol{x}$ de nuevo solo aparecen productos escalares:

$$g(\boldsymbol{x}) = \sum_{i \in Sop}\hat{\alpha}_i y_i (\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}) + w_0$$

Fernando Pérez Nava

Un Ejemplo Inicial (4):Solución Dual

- **Clasificación unidimensional: $g(x)=wx+w_0$**
  - Cjto. Entrenamiento H:x_1=-1, x_2=0 $\in w_1$; x_3=2, x_4=5 $\in w_2$
  - Signos deseados: y_1=1, y_2=1, y_3=-1, y_4=-1
- **Problema a resolver:**

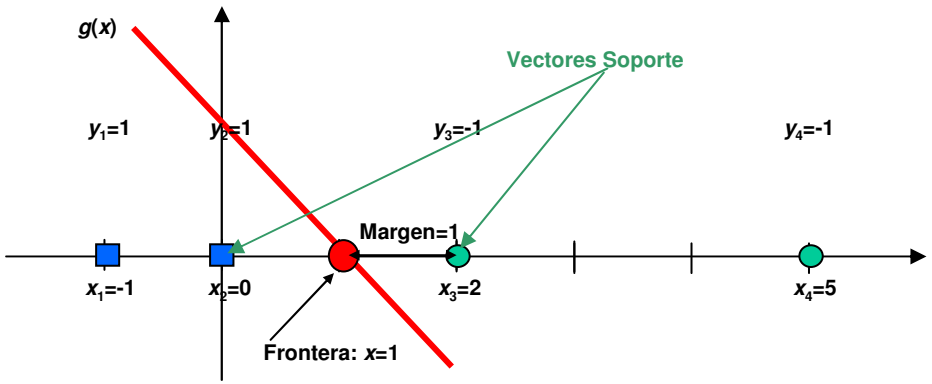
$$\left.\begin{array}{ll}\min\limits_{\boldsymbol{w}} & \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} \\ \text{s.a} & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + w_0) \geq 1 \quad i=1...n\end{array}\right\} \Rightarrow \begin{array}{ll}\min\limits_{w} & \frac{1}{2}w^2 \\ \text{s.a} & -w + w_0 \geq 1 \\ & w_0 \geq 1 \\ & -2w - w_0 \geq 1 \\ & -5w - w_0 \geq 1\end{array}$$

- **Problema dual:**

$$\left.\begin{array}{ll}\max\limits_{\boldsymbol{\alpha}} & \sum\limits_{i=1}^{n}\alpha_i - \frac{1}{2}\sum\limits_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j (\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{x}_j) \\ \text{s.a} & \sum\limits_{i=1}^{n}\alpha_i y_i = 0 \quad \alpha_i \geq 0 \quad i=1,...n\end{array}\right\} \Rightarrow \begin{array}{ll}\max\limits_{\boldsymbol{\alpha}} & \alpha_1+\alpha_2+\alpha_3+\alpha_4 - 1/2\begin{pmatrix}\alpha_1^2 + 0\alpha_1\alpha_2 + 2\alpha_1\alpha_3 + 5\alpha_1\alpha_4 + \\ 0\alpha_2\alpha_1 + 0\alpha_2^2 + 0\alpha_2\alpha_3 + 0\alpha_2\alpha_4 + \\ 2\alpha_3\alpha_1 + 0\alpha_3\alpha_2 + 4\alpha_3^2 + 10\alpha_3\alpha_4 + \\ 5\alpha_4\alpha_1 + 0\alpha_4\alpha_2 + 10\alpha_4\alpha_3 + 25\alpha_4^2\end{pmatrix} \\ \text{s.a} & \alpha_1+\alpha_2-\alpha_3-\alpha_4 = 0 \\ & \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0, \alpha_4 \geq 0\end{array}$$

Un Ejemplo Inicial (5):Solución Dual

- **Solución óptima:**
  - $\hat{\alpha}=(0,1/2,1/2,0)$
- **Vector óptimo:**
  - $\hat{\boldsymbol{w}} = \sum_{i \in Sop} \hat{\alpha}_i y_i \boldsymbol{x}_i = (1/2)\cdot 1\cdot 0 + (1/2)\cdot(-1)\cdot 2 = -1$
- **Constante óptima:**
  - $\hat{w}_0=1-\hat{w}x_2=1-(-1)\cdot 0=1$
- **Clasificador:**
  - Por tanto el clasificador MVS es $g(x)=-x+1$ y la frontera $g(x)=0$ es $x=1$

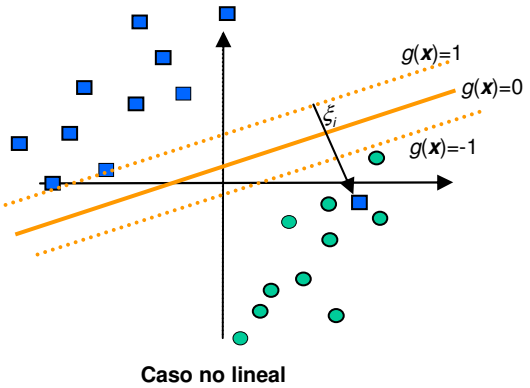


Clasificador MVS

Fernando Pérez Nava

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** $y_i(\boldsymbol{w}^T\boldsymbol{x}_i+w_0)\geq 1$
  - Por tanto será $y_i(\boldsymbol{w}^T\boldsymbol{x}_i+w_0)\geq 1-\xi_l$ con $\xi_l\geq 0$
  - Entonces:
    - Si $\xi_l = 0$ la muestra está en la zona de su clase.
    - Si $1\geq\xi_l\geq 0$ la muestra se mete en la zona
    - del margen.
    - Si $\xi_l > 1$ se mete en la zona de la otra clase.



Caso no lineal

- **Ahora tenemos dos criterios:**
  - Obtener la frontera de mayor margen.
  - Que esta frontera tenga pocas "equivocaciones" (es decir que ξ_l sea lo más cercano a 0).
- **Lo que se hace es combinar los dos criterios.**

Fernando Pérez Nava

MVS: Caso no Separable (2)

- **Combinación de criterios:**

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

$$\text{s.a} \quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + w_0) \geq 1 - \xi_i \quad i = 1...n$$

$$\xi_i \geq 0$$

  - 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 pone el usuario.

- **Problema dual:**

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

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

- **Vector óptimo:**

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

MVS no lineal

- **Con una MVS lineal solo pueden obtenerse fronteras lineales**
- **La forma de obtener una MVS 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 $\phi(\boldsymbol{x})=(\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}),...,\phi_{d^*}(\boldsymbol{x}))^{\mathsf{T}}$ y luego aplicar la MVS lineal.
- **Tras esta transformación se tiene (por ej. para el caso separable):**
  - Problema dual:

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

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

  - $\boldsymbol{w}$ óptimo: $\hat{\boldsymbol{w}} = \sum_{i\in Sop}\hat{\alpha}_i y_i \phi(\boldsymbol{x}_i)$
  - w_0 óptimo: $\hat{w}_0 = 1 - \hat{\boldsymbol{w}}^{\mathsf{T}}\phi(\boldsymbol{x}_i)$, con $\boldsymbol{x}_i \in w_1$ y $\hat{\alpha}_i > 0$
  - Clasificador óptimo: $g(\boldsymbol{x})=\hat{\boldsymbol{w}}^{\mathsf{T}}\phi(\boldsymbol{x})+\hat{w}_0$

Fernando Pérez Nava

El Truco del Núcleo (1)

- **Problema:**
  - Ya vimos que uno de los problemas de trabajar con funciones no lineales $\phi(\boldsymbol{x})=(\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}),...,\phi_{d^*}(\boldsymbol{x}))^T$ consiste en que generalmente el número de ellas d^* es muy grande.
  - El costo computacional de calcular los productos escalares $(\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j))$ es también d^* y por tanto muy grande.
- **Observación:**
  - En determinadas circunstancias en posible evaluar los productos escalares $(\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j))$ sin calcular las funciones ϕ.
- **Ejemplo:**
  - Si $\phi(\boldsymbol{x})=(x_1^2, x_2^2, \sqrt{2}\, x_1 x_2)$ entonces:
  - $(\phi(\boldsymbol{x})^T \phi(\boldsymbol{y})) = (x_1^2, x_2^2, \sqrt{2}\, x_1 x_2)^T (y_1^2, y_2^2, \sqrt{2}\, y_1 y_2) = (\boldsymbol{x}^T \boldsymbol{y})^2$
- **Truco:**
  - El truco del núcleo consiste en calcular $(\phi(\boldsymbol{x})^T \phi(\boldsymbol{y}))$ mediante alguna función de las muestras originales $k(\boldsymbol{x},\boldsymbol{y})$. A esta función se la llama función núcleo.

Fernando Pérez Nava

El Truco del Núcleo (2)

- **Todo el problema MVS no lineal se puede poner con funciones núcleo.**

- **Por ejemplo para el caso linealmente separable:**
  - Problema dual:

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

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

  - w_0 óptimo:

$$\hat{w}_0 = 1 - \sum_{j \in Sop} \hat{\alpha}_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad \text{con } \boldsymbol{x}_i \in w_1 \text{ y } \hat{\alpha}_i > 0$$

  - Clasificador óptimo:

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

Fernando Pérez Nava

El Truco del Núcleo (3)

- **La función núcleo $k(\boldsymbol{x}, \boldsymbol{y})$ mide la similitud entre las muestras $\boldsymbol{x}$ e $\boldsymbol{y}$.**
- **No toda función $k(\boldsymbol{x}, \boldsymbol{y})$ puede ser utilizada como función núcleo. Debe satisfacer la denominada condición de Mercer.**
- **Algunas funciones núcleo:**
  - Lineal: $k(\boldsymbol{x}, \boldsymbol{y})=\boldsymbol{x}^T \boldsymbol{y}$
  - Polinomial: $k(\boldsymbol{x}, \boldsymbol{y})=(\lambda\ (\boldsymbol{x}^T \boldsymbol{y}) + \theta)^p$, p=1,2,...
  - Función de base radial: $k(\boldsymbol{x}, \boldsymbol{y})= \exp(- \lambda\ (\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y}))$, $\lambda>0$
  - Tangente hiperbólica: $k(\boldsymbol{x}, \boldsymbol{y})= \tanh(\lambda\ (\boldsymbol{x}^T \boldsymbol{y}) - \theta)$
- **Observación:**
  - La forma final del clasificador lineal:

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

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

Fernando Pérez Nava

Ejemplo no Lineal (1)

- **El problema del XOR resuelto con una MVS no lineal (sin núcleo).**
  - Cjto. entrenamiento: $\boldsymbol{x}_1=(0,1)$, $\boldsymbol{x}_2=(1,0) \in w_1$ $\boldsymbol{x}_3=(0,0)$, $\boldsymbol{x}_4=(1,1) \in w_2$
  - Signos deseados: $y_1=1$, $y_2=1$, $y_3=-1$, $y_4=-1$

  - Funciones ϕ: $\phi(\boldsymbol{x})=(x_1^2, x_2^2, \sqrt{2}\,x_1x_2, \sqrt{2}x_1, \sqrt{2}\,x_2, 1)$
    - 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 $(\phi(\boldsymbol{x}_i)^T\,\phi(\boldsymbol{x}_j))$

      $(\phi(\boldsymbol{x}_1)^T\,\phi(\boldsymbol{x}_1))= (0,1,0,0,\sqrt{2},1)^T\,(0,1,0,0,\sqrt{2},1)=4$

      $(\phi(\boldsymbol{x}_1)^T\,\phi(\boldsymbol{x}_2))= (0,1,0,0,\sqrt{2},1)^T\,(1,0,0,\sqrt{2},0,1)=1$

      ...

      ...

      $(\phi(\boldsymbol{x}_4)^T\,\phi(\boldsymbol{x}_4))= (1,1,\sqrt{2},\sqrt{2},\sqrt{2},1)^T\,(1,1,\sqrt{2},\sqrt{2},\sqrt{2},1)=9$

Fernando Pérez Nava

Ejemplo no Lineal (2)

- **Problema a resolver:**

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

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

$$\max_{\boldsymbol{\alpha}} \quad \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\begin{pmatrix} 4\alpha_1^2 + \alpha_1\alpha_2 - \alpha_1\alpha_3 - 4\alpha_1\alpha_4 + \\ \alpha_2\alpha_1 + 4\alpha_2^2 - \alpha_2\alpha_3 - 4\alpha_2\alpha_4 + \\ -\alpha_3\alpha_1 - \alpha_3\alpha_2 + \alpha_3^2 + \alpha_3\alpha_4 + \\ -4\alpha_4\alpha_1 - 4\alpha_4\alpha_2 + \alpha_4\alpha_3 + 9\alpha_4^2 \end{pmatrix}$$

$$\text{s.a} \quad \alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 = 0 \quad \alpha_i \geq 0 \quad i = 1...4$$

- **Solución:**
  - $\hat{\alpha}$=(8/3,8/3,10/3,2)

Fernando Pérez Nava

Ejemplo no Lineal (3)

- **Vector *w*:**
  - $\hat{\boldsymbol{w}} = \sum_{i \in Sop} \hat{\alpha}_i y_i \boldsymbol{\phi}(\boldsymbol{x}_i)$
  - $\hat{\boldsymbol{w}}=(8/3)(0,1,0,0,\sqrt{2},1)+(8/3)(1,0,0,\sqrt{2},0,1)-(10/3)(0,0,0,0,0,1)-$
    $-2\ (1,1,\sqrt{2},\sqrt{2},\sqrt{2},1)=(2/3,2/3,-2\sqrt{2},2/3\sqrt{2},2/3\sqrt{2},0)$
- **Constante w_0:**
  - $\hat{w}_0 = 1 - \hat{\boldsymbol{w}}^T \boldsymbol{\phi}(\boldsymbol{x}_i)$, con $\boldsymbol{x}_i \in w_1$ 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:**
  - $g(\boldsymbol{x})=\hat{\boldsymbol{w}}^T\phi(\boldsymbol{x})+\hat{w}_0$
  - $g(\boldsymbol{x})=(2/3,2/3,-2\sqrt{2},2/3\sqrt{2},2/3\sqrt{2},0)^T(x_1^2,\ x_2^2,\sqrt{2}\ x_1x_2,\sqrt{2}\ x_1,\sqrt{2}\ x_2,\ 1)-1=$
    $=2/3\ x_1^2+\ 2/3\ x_2^2-4\ x_1x_2+2/3\ x_1+\ 2/3\ x_2-1$

Fernando Pérez Nava

Ejemplo no Lineal (4)

- **El problema del XOR resuelto con una MVS no lineal (con núcleo).**
  - Cjto. entrenamiento: $\boldsymbol{x}_1=(0,1)$, $\boldsymbol{x}_2=(1,0) \in w_1$ $\boldsymbol{x}_3=(0,0)$, $\boldsymbol{x}_4=(1,1) \in w_2$
  - Signos deseados: $y_1=1$, $y_2=1$, $y_3=-1$, $y_4=-1$
  - Funciones ϕ: $\phi(\boldsymbol{x})=(x_1^2, x_2^2, \sqrt{2}\,x_1x_2, \sqrt{2}\,x_1, \sqrt{2}\,x_2, 1)$
  - Productos escalares: $(\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)) = ((\boldsymbol{x}_i^T \boldsymbol{x}_j) + 1)^2$
  - Por tanto la función núcleo es $k(\boldsymbol{x},\boldsymbol{y})=((\boldsymbol{x}^T \boldsymbol{y}) + 1)^2$

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

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

    ...

    $k(\boldsymbol{x}_4, \boldsymbol{x}_4)= ((\boldsymbol{x}_4^T \boldsymbol{x}_4) + 1)^2= ((1,1)^T (1,1) +1)^2 =9$
- **Problema dual (el mismo de antes):**

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

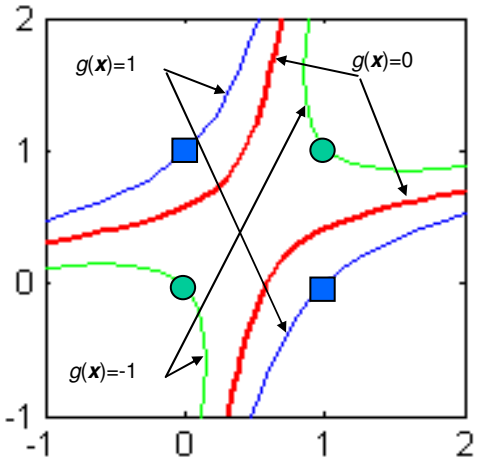
$$\max_{\boldsymbol{\alpha}} \quad \alpha_1+\alpha_2+\alpha_3+\alpha_4 - \frac{1}{2}\begin{pmatrix} 4\alpha_1^2+\alpha_1\alpha_2-\alpha_1\alpha_3-4\alpha_1\alpha_4+ \\ \alpha_2\alpha_1+4\alpha_2^2-\alpha_2\alpha_3-4\alpha_2\alpha_4+ \\ -\alpha_3\alpha_1-\alpha_3\alpha_2+\alpha_3^2+\alpha_3\alpha_4+ \\ -4\alpha_4\alpha_1-4\alpha_4\alpha_2+\alpha_4\alpha_3+9\alpha_4^2 \end{pmatrix}$$

$$\text{s.a} \quad \alpha_1+\alpha_2-\alpha_3-\alpha_4=0 \quad \alpha_i \geq 0 \quad i=1...4$$

Fernando Pérez Nava

Ejemplo no Lineal (5)

- **Solución:**
  - $\hat{\alpha}=(8/3,8/3,10/3,2)$
- **Constante w_0 (la misma de antes):**
  - $\hat{w}_0 = 1 - \sum_{j \in Sop} \hat{\alpha}_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, con $\boldsymbol{x}_i \in w_1$ y $\hat{\alpha}_i > 0$
  - $\hat{w}_0=1-((8/3)4+8/3-(10/3)-(6/3)4)=-1$
- **Clasificador (el mismo de antes):**
  - $g(\boldsymbol{x}) = \sum_{i \in Sop} \hat{\alpha}_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + \hat{w}_0$
  - $g(\boldsymbol{x})=(8/3)\ (x_2+1)^2 + (8/3)(x_1+1)^2 - (10/3) -(6/3)(x_1+ x_2+1)^2 -1= (2/3)(\ (x_1 -1/2)^2 + (x_2 -1/2)^2 - 6(x_1 -1/2)(x_2 -1/2) -1/2)$



Frontera de decisión de la MVS (rojo)

Fernando Pérez Nava

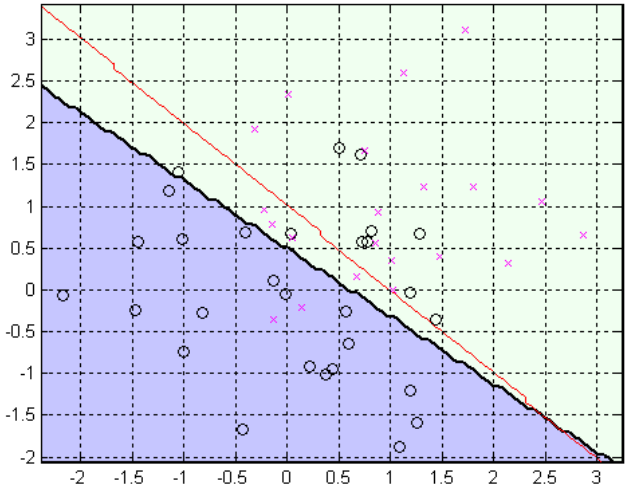
MVS: Aspectos Prácticos

- **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 los 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 tiene el proceso de optimización tiene menos dificultades numéricas
    - La selección de los parámetros de las funciones núcleo y el parámetro C del caso no lineal se hace mediante un conjunto de testeo.
    - En el caso de funciones de base radial se suele seleccionar las sucesiones: $C=2^{-5}, 2^{-3}, ..., 2^{15}$, $\lambda = 2^{-15}, 2^{-13}, ..., 2^{3}$
- **Clasificación multiclase:**
  - Se construye una MVS 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

Clasificación con MVS: Ejemplo

- **Distribuciones verdaderas:**
  - $p(\boldsymbol{x} \mid w_1, \boldsymbol{\theta}_1) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$, $p(\boldsymbol{x} \mid w_2, \boldsymbol{\theta}_2) \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$
  - $P(w_1)=0.5$, $P(w_2)=0.5$

- **Clasificación:**
  - Conjunto de testeo:
    - 50 muestras por clase
  - Conjunto de entrenamiento:
    - 50 muestras por clase
  - Núcleo: lineal
  - Valor óptimo calculado para *C*:
    - 0.0005
  - Error de clasificación estimado:
    - 0.28
  - Error bayesiano:
    - 0.23



Ejemplo de clasificación tras estimación por MVS
Circulos: muestras de la clase 1
Aspas: muestras de la clase 2
Linea negra: Frontera de decisión a partir de la estimación
Linea roja: Frontera de decisión bayesiana

Fernando Pérez Nava

Resumiendo...

- **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
  - MVS: 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...)