Introducción

- **Existen dos aproximaciones para resolver el problema de clasificación:**
  - Aproximación Generativa (vista en el Tema 3)

    Basada en:
    - Modelar $p(\boldsymbol{x},w)=p(\boldsymbol{x}|\ w)\mathrm{P}(w)$
      - $p(\boldsymbol{x}|\ w)$ es la distribución condicional de las características en las clase.
      - $\mathrm{P}(w)$ es la probabilidad a priori de la clase
    - Aplicar la regla de clasificación óptima:
      - Dado un $\boldsymbol{x}$, su clase $\hat{w}$ se calcula a partir de:

        $\hat{w}=\mathrm{argmax}_w\ p(\boldsymbol{x},w)$, o de forma equivalente como

        $\hat{w}=\mathrm{argmax}_w\ \mathrm{P}(w\,|\,\boldsymbol{x})$ (probabilidad a posteriori)

    - Problema Práctico: Estimar $p(\boldsymbol{x}|w)$
      - Solución: Obtener un conjunto de entrenamiento y utilizar EMV, MAP, *k*-vecinos...

Fernando Pérez Nava

Introducción:Aproximación Discriminativa

- **Observación:**
  - Se pueden encontrar problemas de clasificación complejos para los que aparecen funciones discriminantes simples.
- **Ejemplo:**
  - Sabemos que en el caso gausiano con matrices de covarianzas iguales: $p(\boldsymbol{x}|\ w_i)=\mathrm{N}(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$, las funciones discriminantes son lineales.
    - Si tenemos *d* características y 2 clases tenemos que calcular $2(d+1)$ parámetros.
  - En realidad, para clasificar nos interesa la diferencia de las dos funciones discriminantes (para ver cual es la mayor).
    - La diferencia vuelve a ser lineal pero depende sólo de $(d+1)$ parámetros.
  - Para resolver el problema con la aproximación generativa hay que estimar primero:
    - Dos vectores de medias: $2d$ parámetros
    - Una matriz de covarianzas: $(d^2+d)/2$ parámetros
    - Total: $(d^2+5d)/2$ parámetros

    y después construir la diferencia de las funciones discriminantes.
  - Es decir, para resolver el problema de determinar $d+1$ parámetros tenemos como paso intermedio otro de determinar $(d^2+5d)/2$ parámetros.

Fernando Pérez Nava

Aproximación Discriminativa

- **Idea:**
  - Construir las funciones discriminantes estimando directamente los parámetros que las definen.

    "Si se posee una cantidad de información limitada para la resolución de un problema, intenta resolver el problema directamente y nunca intentes resolver un problema más general como paso intermedio. Es posible que la información disponible sea suficiente para la solución directa pero insuficiente para resolver un problema intermedio más general"

    V. Vapnik, *The nature of statistical learning theory* (1995)
- **Ventajas:**
  - No se necesita modelar la función de densidad condicional de las características
    - Simplicidad: Algunas funciones de densidad tienen una gran cantidad de parámetros
    - Robustez: Muchas funciones de densidad distintas dan lugar al mismo clasificador lineal
- **Desventajas:**
  - No se puede obtener un modelo de como se generaron los datos

El caso lineal

- **El caso más simple de construcción de funciones discriminantes ocurre cuando éstas son lineales**
  - Este es el caso que estudiaremos en este Tema
- **Llamaremos:**
  - Función discriminante lineal (FDL) a toda función discriminante que para una clase w_i es lineal. Por tanto tiene la forma: $g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$.
- **Clasificación con dos clases**
  - En este caso se tienen dos FDL:

    $g_1(\boldsymbol{x}) = \boldsymbol{w}_1^T\boldsymbol{x} + w_{10}$ para la clase w_1

    $g_2(\boldsymbol{x}) = \boldsymbol{w}_2^T\boldsymbol{x} + w_{20}$ para la clase w_2.
  - Para clasificar necesitamos saber cual es mayor (o ver si su diferencia es positiva o negativa).
  - Entonces si definimos su diferencia como:

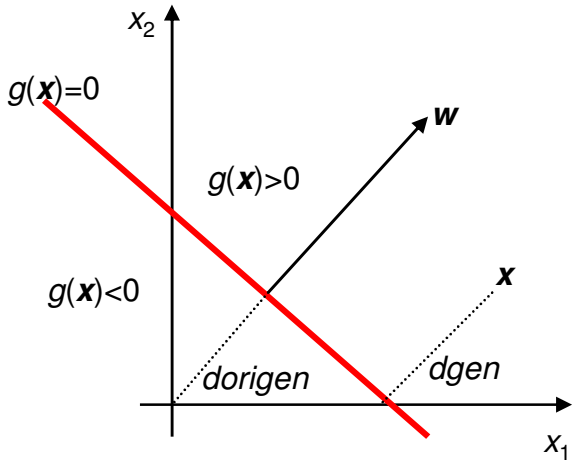
    $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) = (\boldsymbol{w}_1 - \boldsymbol{w}_2)^T\boldsymbol{x} + (w_{10} - w_{20}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$

    $\boldsymbol{w} = (\boldsymbol{w}_1 - \boldsymbol{w}_2)$, $w_0 = w_{10} - w_{20}$
  - La regla de decisión es: Elegir w_1 si $g(\boldsymbol{x}) \geq 0$; Elegir w_2 si $g(\boldsymbol{x}) \leq 0$
  - Llamaremos vector de pesos al vector $\boldsymbol{w}$ y término independiente al coeficiente w_0

Interpretación Geométrica de la Regla de Decisión: 2 Clases

- **Interpretación geométrica de $g(\boldsymbol{x})= \boldsymbol{w}^T\boldsymbol{x}+w_0$**
  - El vector $\boldsymbol{w}$ es perpendicular a la frontera de decisión
  - La distancia entre el origen y la frontera de decisión depende de w_0:

$$dorigen = \frac{|w_0|}{\sqrt{\boldsymbol{w}^T\boldsymbol{w}}}$$



Interpretación geométrica

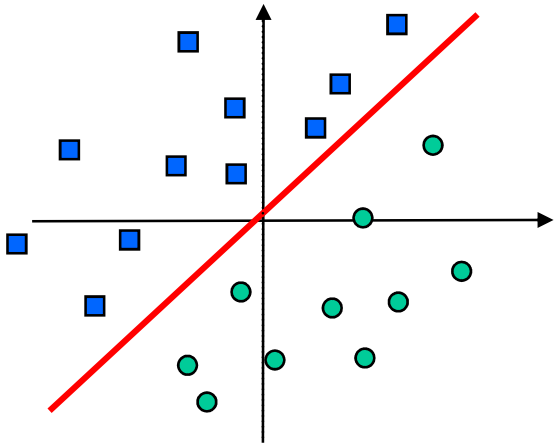
  - En general, la distancia de un punto $\boldsymbol{x}$ a la frontera de decisión es:

$$dgen = \frac{|g(\boldsymbol{x})|}{\sqrt{\boldsymbol{w}^T\boldsymbol{w}}}$$

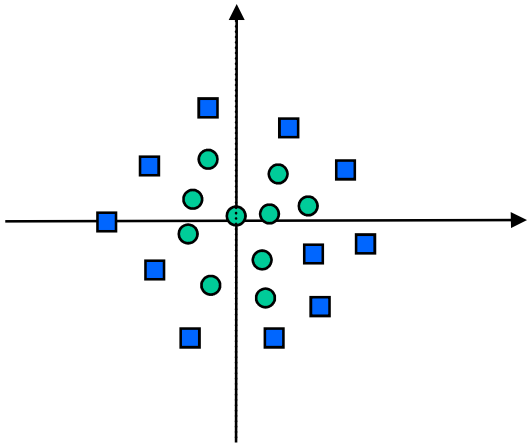
Fernando Pérez Nava

Separabilidad Lineal y Clasificación

- **Separabilidad Lineal**
  - Dos conjuntos de puntos son linealmente separables si existe una función lineal que los separa



Conjunto Linealmente Separable

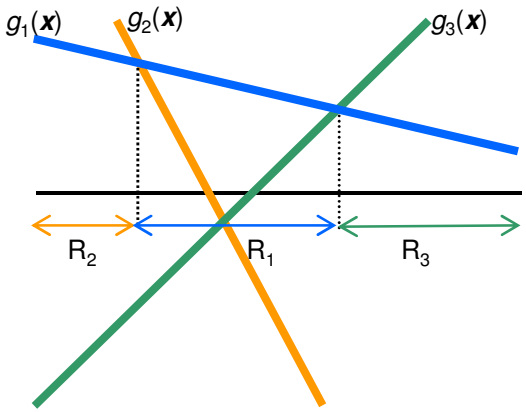


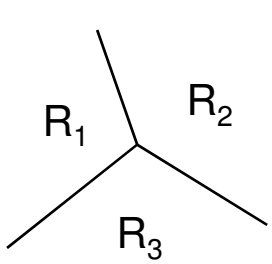
Conjunto no Linealmente Separable

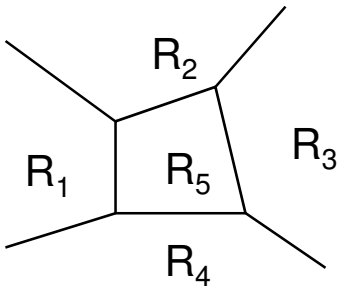
- **Un clasificador lineal no puede clasificar sin errores un conjunto no linealmente separable (al ser lineal su frontera de decisión).**

FDL: Varias clases

- **En este caso se tiene:**
  - Una FDL $g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}$ para cada clase w_i.
  - Entonces la regla de decisión es:
    Elegir w_i si $g_i(\boldsymbol{x}) \geq g_j(\boldsymbol{x})$ para todo $j \neq i$
- **Este esquema produce regiones de decisión:**
  - Convexas y simplemente conexas: dados dos puntos de en la región el segmento que los une está en la región.







Regiones de decisión para las FDL

Fernando Pérez Nava

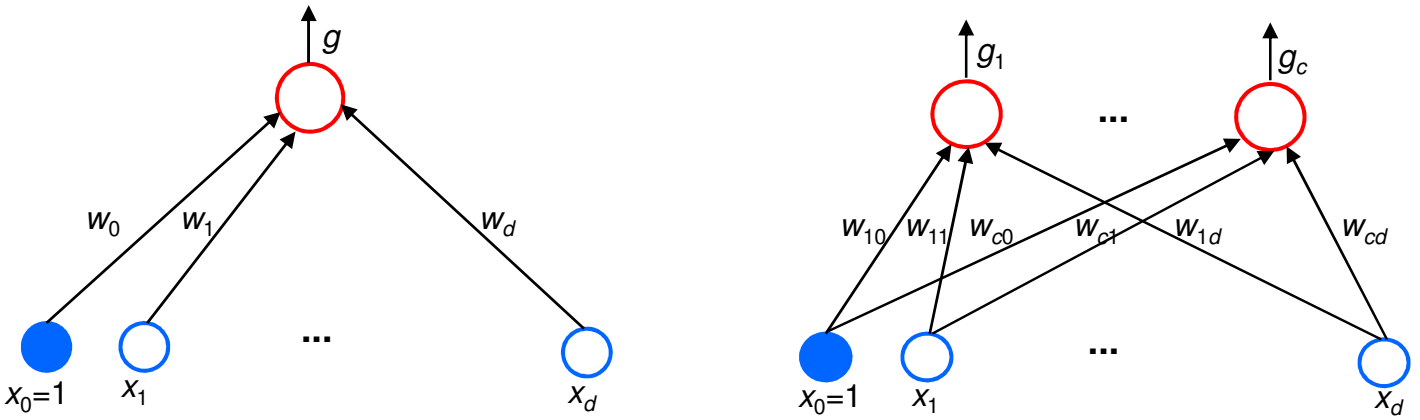
Representación Gráfica de las FDL

- **Notación**
  - A la hora de trabajar con la función discriminante $g_i(\boldsymbol{x})= \boldsymbol{w}_i^T \boldsymbol{x}+w_{i0}$ es a veces lioso tener por un lado el vector de pesos $\boldsymbol{w}_i$ y el término independiente w_{i0}.
  - Puesto que:

    $g_i(\boldsymbol{x})= \boldsymbol{w}_i^T \boldsymbol{x}+w_{i0} = w_{i1}\, x_1 + w_{i2}\, x_2 +\ldots+ w_{id}\, x_d + w_{i0}\, 1=$

    $(w_{i1}\ ,\ w_{i2}\ ,\ \ldots\ ,\ w_{id}\ \ w_{i0}\ )\ (x_1,\ x_2,\ \ ,x_d\ ,1)^T$

    introduciremos w_{i0} en el vector de pesos añadiendo una característica ficticia que toma siempre el valor 1.
- **Representación Gráfica**



FDL: Representación gráfica

Fernando Pérez Nava

Clasificación mediante Regresión Lineal (1)

- **Clasificación para dos clases**
  - Se tiene: $g(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}$ y la regla de decisión es:
    Elegir w_1 si $g(\boldsymbol{x}) \geq 0$ Elegir w_2 si $g(\boldsymbol{x}) \leq 0$
- **Problema:**
  - ¿Cómo determinar $\boldsymbol{w}$?
- **Observación:**
  - Al menos nos gustaría que $g(\boldsymbol{x}) \geq 0$ para los elementos del conjunto de entrenamiento que pertenecen a w_1 (H_1) y $g(\boldsymbol{x}) \leq 0$ en los que pertenecen a w_2 (H_2):
- **Idea:**
  - Intentaremos que $g(\boldsymbol{x})$ valga +1 para los elementos de H_1 y -1 para los de H_2 (de esta forma estarán bien clasificados).
  - Para ello se penalizan las diferencias entre los valores deseados (+1 ó –1) para cada elemento $\boldsymbol{x}_i$ de conjunto de entrenamiento y los valores obtenidos con la función discriminante $g(\boldsymbol{x}_i)$.

Fernando Pérez Nava

Clasificación mediante Regresión Lineal (2)

- **Notación**
  - Para tener presente la dependencia de la función discriminante del vector de pesos escribiremos $g(\boldsymbol{x};\boldsymbol{w})=\boldsymbol{w}^T\boldsymbol{x}$
- **Función de Penalización**
  - Se penaliza la diferencia entre los valores deseados y_i y los obtenidos $g(\boldsymbol{x}_i;\boldsymbol{w})$ mediante:

$$E_{ECM}(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - g(\boldsymbol{x}_i;\boldsymbol{w}))^2, \quad y_i = \begin{cases} +1 & \boldsymbol{x}_i \in H_1 \\ -1 & \boldsymbol{x}_i \in H_2 \end{cases}$$

- **Este es un problema de regresión lineal**
  - El vector óptimo de pesos es el que hace mínima la función E_{ECM}
  - Hay dos métodos para minimizar la función:
    - Método Analítico
    - Método Iterativo

Fernando Pérez Nava

Método Analítico

- **Cálculo del vector óptimo por el método analítico.**
  - Se basa en resolver el sistema de ecuaciones lineales: $\frac{\partial E_{ECM}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \mathbf{0}$
  - Definiendo:

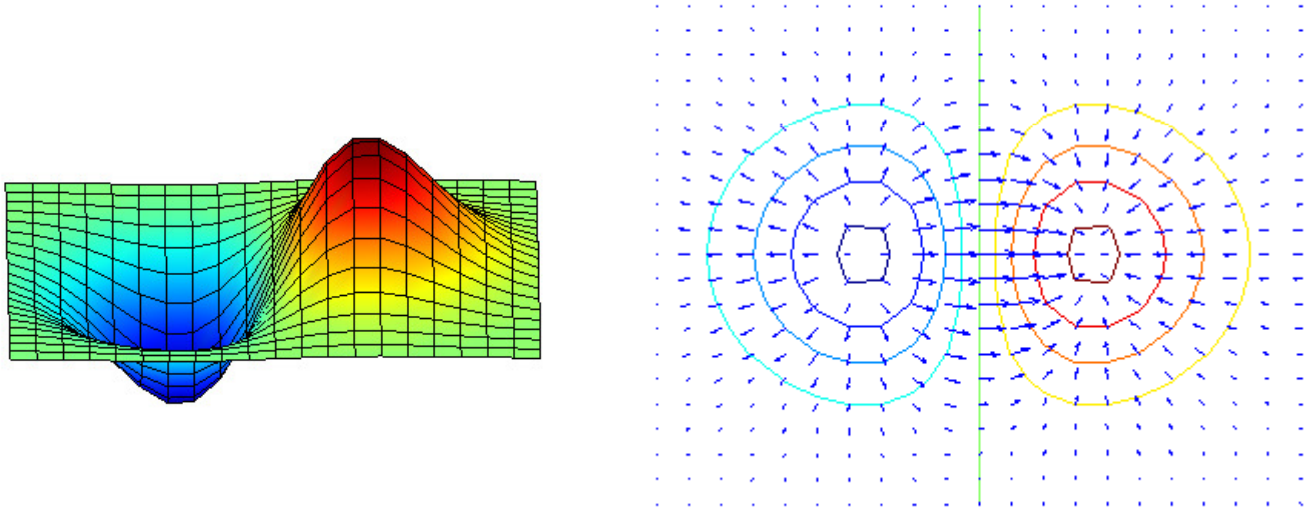
$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

  - La solución óptima es: $\hat{\boldsymbol{w}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{X}^{-}\boldsymbol{Y}$
    donde $\boldsymbol{X}^{-}$ recibe el nombre de *pseudoinversa* de $\boldsymbol{X}$
- **Problemas del método analítico**
  - La matriz $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ puede no tener inversa.
  - Es necesario realizar operaciones con matrices de gran tamaño.

Fernando Pérez Nava

Optimización iterativa

- **Muchos problemas de optimización (cálculo del máximo o mínimo de una función *f*(*x*)) carecen de una solución analítica**
- **Observación:**
  - El gradiente de una función en un punto ***x*** indica la dirección de máximo crecimiento de la función en ese punto.



Función bidimensional y gradiente

- **Idea**
  - Dado un valor inicial ***x*** , calcular el gradiente en ese punto $\nabla f(\boldsymbol{x})$ y avanzar un poco en esa dirección (si busco maximizar) o en la dirección contraria (si busco minimizar).

Fernando Péraz Nava

El Método del Gradiente

- **Es uno de los métodos de optimización iterativa más simples (y lentos)**
- **Requiere que la función $f(\boldsymbol{x})$ a optimizar sea diferenciable.**
- **Procedimiento (Minimización)**
  - Se define un proceso iterativo que modifica una aproximación inicial $\boldsymbol{x}^{(0)}$ mediante la regla:

    $\boldsymbol{x}^{(r+1)}=\boldsymbol{x}^{(r)} - \rho_r \nabla f(\boldsymbol{x}^{(r)})$.
  - Regla de Parada:
    - La distancia entre el nuevo punto y el anterior está por debajo de un umbral
    - Se alcanza el máximo número de iteraciones.
    - El módulo del gradiente está por debajo de un umbral.
  - La elección del *parámetro de aprendizaje* ρ_r es crítica. Un valor muy pequeño hace la convergencia muy lenta y uno muy grande provoca oscilaciones alrededor del mínimo.
  - Cuando el algoritmo converge lo hace a un punto para el que $\nabla f(\boldsymbol{x})=0$. Este es un mínimo local o un punto de inflexión

Método iterativo para Regresión Lineal

- **Se basa en aplicar el método del gradiente a la función $E_{ECM}(\boldsymbol{w})$**
- **El esquema iterativo basado en el método del gradiente es:**
  - Paso 0

    Elegir un valor inicial $\boldsymbol{w}^{(0)}$

    Hacer r=0
  - Paso 1
    - Calcular $\boldsymbol{w}^{(r+1)}=\boldsymbol{w}^{(r)}+\rho_r \sum_{k=1}^{n} \delta_k \boldsymbol{x}_k \qquad \delta_k=(y_k-g(\boldsymbol{x}_k;\boldsymbol{w}^{(r)}))$

      (δ_k es el error que se comente con la muestra $\boldsymbol{x}_k$)

    - Incrementar *r*
  - Paso 2
    - Si se cumple la regla de parada finalizar. En otro caso ir al Paso 1
  - A este esquema se le llama por Entrenamiento por época.

    En cada iteración se utiliza todo el conjunto de entrenamiento *H*
- **Entrenamiento por muestra**
  - En cada paso *r* se utiliza la muestra $\boldsymbol{x}_{k(r)}$ de *H*

    $\boldsymbol{w}^{(r+1)}=\boldsymbol{w}^{(r)}+\rho_r \delta_{k(r)} \boldsymbol{x}_{k(r)} \qquad \delta_{k(r)}=(y_{k(r)}-g(x_{\boldsymbol{k(r)}};\boldsymbol{w}^{(r)}))$

Fernando Pérez Nava

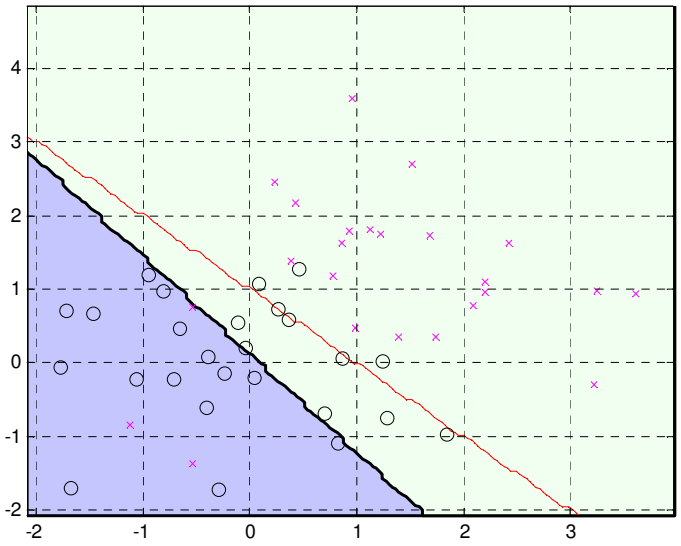
Clasificación mediante Regresión Lineal con más de dos clases

- **Cuando hay más de dos clases se tiene que estimar un vector de pesos por clase**
- **Estimación de los vectores de pesos:**
  - Se resuelve un problema de regresión lineal por clase.
    - Dada una clase w_i
      - El valor de y_k para un elemento del conjunto de entrenamiento $\boldsymbol{x}_k$ que pertenece a w_i es +1. Si no pertenece a w_i es –1.
      - Entonces se aplica el método analítico o el iterativo y se obtiene el vector de pesos de esa clase
- **Clasificación:**
  - Se calcula el máximo de las funciones discriminantes. La clase para la que se obtiene el máximo es la asignada al patrón que se quiere clasificar

Fernando Péraz Nava

Clasificación por Regresión Lineal: 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
  - Error de clasificación estimado:
    - 0.36
  - Error bayesiano:
    - 0.23



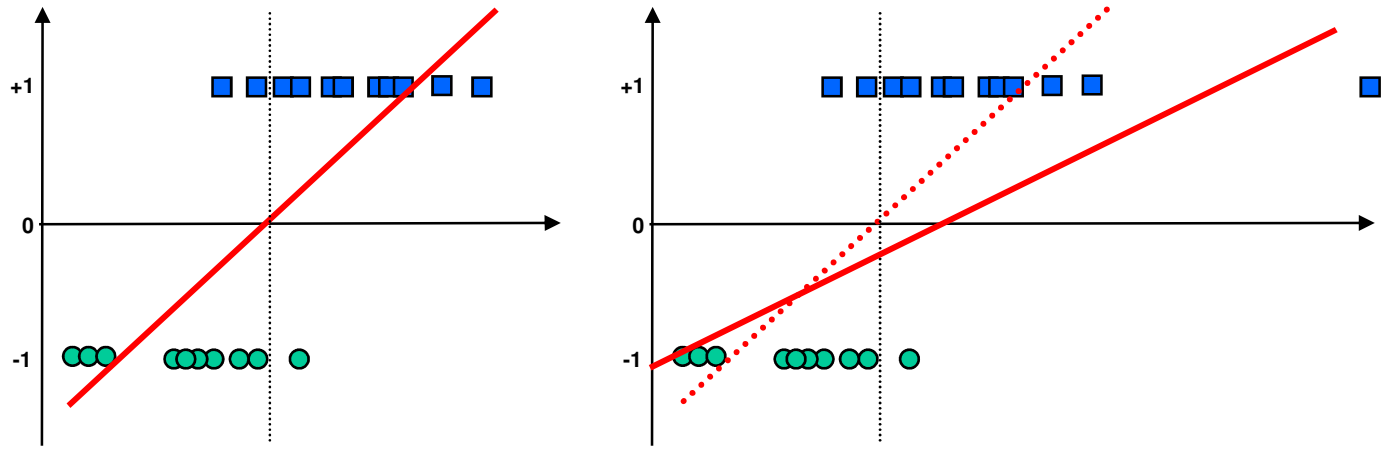
Ejemplo de clasificación tras estimación por regresión lineal
Circulos: muestras de la clase 1
Aspas: muestras de la clase 2
Linea negra: Frontera de decisión a partir de la estimación
Linea roja: Frontera de decisión bayesiana

Fernando Pérez Nava

–

Resultados

- **En general los resultados de la regresión lineal para clasificación no son muy buenos.**
  - La fundamentación no es sólida.
    - Se intenta predecir una variable discreta (que toma los valores +1 y –1) mediante una continua (función lineal)
    - Los valores +1 y –1 son arbitrarios
    - Los resultados son muy sensibles a la aparición de datos "extraños" (por ejemplo como resultados de errores al recolectar el conjunto de entrenamiento)



Regresión Lineal: Problemas con la Robustez

La visión Bayesiana

- **El método generativo:**
  - Se modela $p(\boldsymbol{x},w_i)$ a través de la descomposición $p(\boldsymbol{x},w_i)=p(\boldsymbol{x}|\ w_i)\mathrm{P}(w_i)$
    - En el caso paramétrico $p(\boldsymbol{x}|\ w_i)$ depende de un vector de parámetros $\boldsymbol{\theta}_i$, es decir, tenemos $p(\boldsymbol{x}|\ w_i,\boldsymbol{\theta}_i)$
    - Se obtiene $\hat{\boldsymbol{\theta}}_i$, que estima $\boldsymbol{\theta}_i$
    - Se halla la clase óptima: $\hat{w}=\mathrm{argmax}_w\ p(\boldsymbol{x}|\ w_i,\hat{\boldsymbol{\theta}}_i)$
- **Interpretación bayesiana de la aproximación discriminativa:**
  - Se modela $p(\boldsymbol{x},w)$ a través de la descomposición:
    - $p(\boldsymbol{x},w_i)=p(w_i|\ \boldsymbol{x})\mathrm{P}(\boldsymbol{x})$
    - Se asume que $\mathrm{P}(w_i|\ \boldsymbol{x})$depende de un vector de parámetros $\boldsymbol{\theta}_i$, es decir, tenemos $\mathrm{P}(w_i|\ \boldsymbol{x},\boldsymbol{\theta}_i)$
    - Se obtiene $\hat{\boldsymbol{\theta}}_i$, que estima $\boldsymbol{\theta}_i$
    - Hallar la clase óptima (para la que no hace falta conocer $\mathrm{P}(\boldsymbol{x})$): $\hat{w}=\mathrm{argmax}_w\ \mathrm{P}(w|\ \boldsymbol{x})$

Fernando Pérez Nava

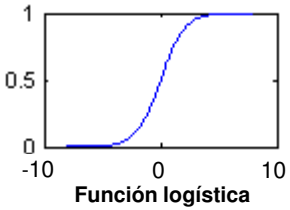
Regresión Logística: 2 clases

- **Modelo:**

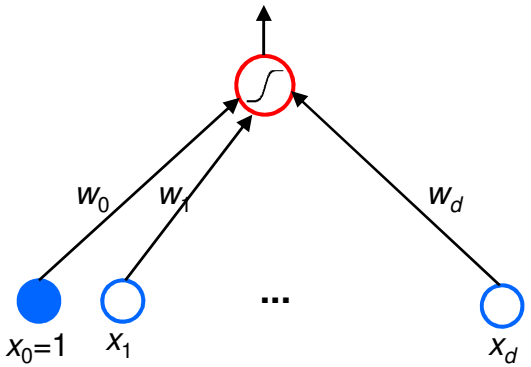
$$P(w_1 \mid \boldsymbol{x}) = \frac{1}{1+\exp(-g(\boldsymbol{x}))}, \quad g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} \quad \text{(Integrando } w_0 \text{ en } \boldsymbol{w}\text{)}, \quad P(\omega_2|\boldsymbol{x})=1-P(\omega_1|\boldsymbol{x})$$

- **La función logística:**
  - Tras calcular la FDL $g(\boldsymbol{x})$ el resultado se pasa por la función logística $\tau(x)$:

$$\tau(x) = \frac{1}{1+\exp(-x)}$$



- **Representación gráfica:**



- **Frontera de decisión:**
  - Sigue siendo $g(\boldsymbol{x})=0$ (lineal) debido a que la función logística es monótona creciente.

Fernando Pérez Nava

Regresión Logística: Estimación

- **Estimación EMV**

  – Tenemos un conjunto de entrenamiento:

  $H=\{ (\boldsymbol{x}_1,y_1), (\boldsymbol{x}_2,y_2), \ldots (\boldsymbol{x}_n,y_n)\}$, $y_i = 1$ si $\boldsymbol{x}_i \in w_1$, $y_i = -1$ si $\boldsymbol{x}_i \in w_2$

  – Construimos la función de verosimilitud:

$$L = p(H \mid \boldsymbol{w}) = \prod_{k=1}^{n} \mathrm{P}(y_k \mid \boldsymbol{x}_k, \boldsymbol{w})$$

  y la optimizamos de forma iterativa:

$$\boldsymbol{w}^{(r+1)} = \boldsymbol{w}^{(r)} + \rho_r \sum_{k=1}^{n} \delta_k \boldsymbol{x}_k ,$$

$$\delta_k = (1 - \tau (y_i \, g(\boldsymbol{x}_k ; \boldsymbol{w}^{(r)\mathrm{T}})) \, y_i$$

Fernando Pérez Nava

Clasificación mediante Regresión Logística con más de dos clases

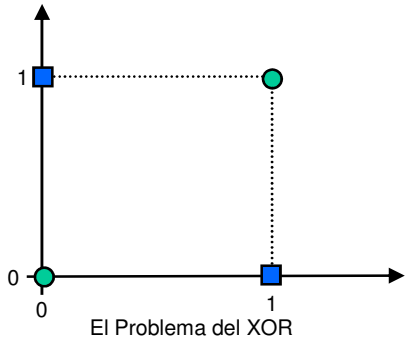
- **Modelo:**

$$P(w_i \mid \boldsymbol{x}) = \frac{\exp(g_i(\boldsymbol{x}))}{\sum_{j=1}^{c} \exp(g_i(\boldsymbol{x}))}, \quad g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x}$$

- **Estimación del vector de pesos:**
  - Se resuelve un problema de regresión logística por clase.
    - Dada una clase w_i
      - El valor de y_k para un elemento del conjunto de entrenamiento $\boldsymbol{x}_k$ que pertenece a w_i es +1. Si no pertenece a w_i es –1.
      - Entonces se aplica EMV
  - Como resultado obtengo una estimación de la probabilidad a posteriori de cada clase
- **Clasificación**
  - Se calcula el máximo de las estimaciones de las probabilidades a posteriori. La clase para la que se obtiene el máximo es la asignada al patrón que se quiere clasificar

Limitaciones de los Clasificadores Lineales

- **Hay problemas de clasificación muy sencillos que no se pueden resolver con un clasificador lineal**
- **El problema del XOR**
  - Elementos de la primera clase (clase 1) (0,0) y (1,1)
  - Elementos de la segunda clase (clase 0) (0,1) y (1,0)



El Problema del XOR

  - No hay ningún clasificador lineal que no cometa errores al clasificar este conjunto.
  - Por tanto, es necesario buscar clasificadores más complejos.

Fernando Pérez Nava

En Resumen...

- **Se ha presentado una nueva aproximación al problema de clasificación: la aproximación discriminativa**
  - Se basa en calcular directamente las funciones discriminantes
  - Tiene las ventajas de mayor simplicidad y robustez
  - La principal desventaja es que no se obtiene un modelo de la generación de los datos
- **Se han presentado dos métodos para la aproximación discriminativa en el caso más simple: el lineal**
  - Regresión Lineal
    - Ventajas: Método Sencillo
    - Desventajas: Falta de fundamentación. No muy buenos resultados
  - Regresión Logística
    - Ventajas: Se estiman las probabilidades a posteriori de las clases
    - Desventajas. Método ligeramente más complejo.
- **Hay problemas simples para los que los clasificadores lineales son inapropiados.**
  - Es necesario por tanto buscar clasificadores más complejos (no lineales)