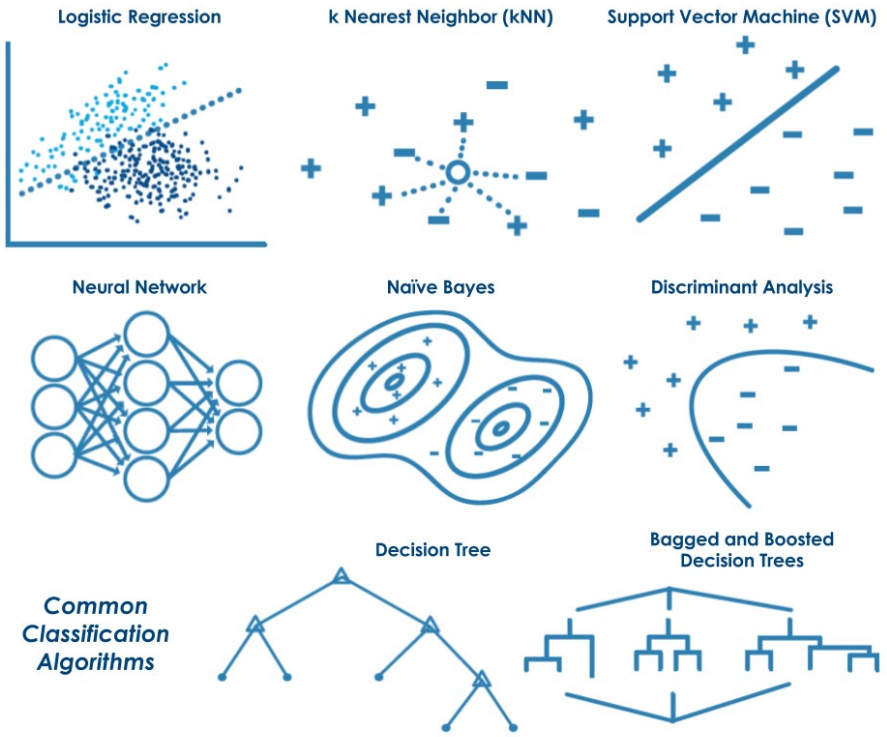


Clasificación I
Modelos lineales de clasificación

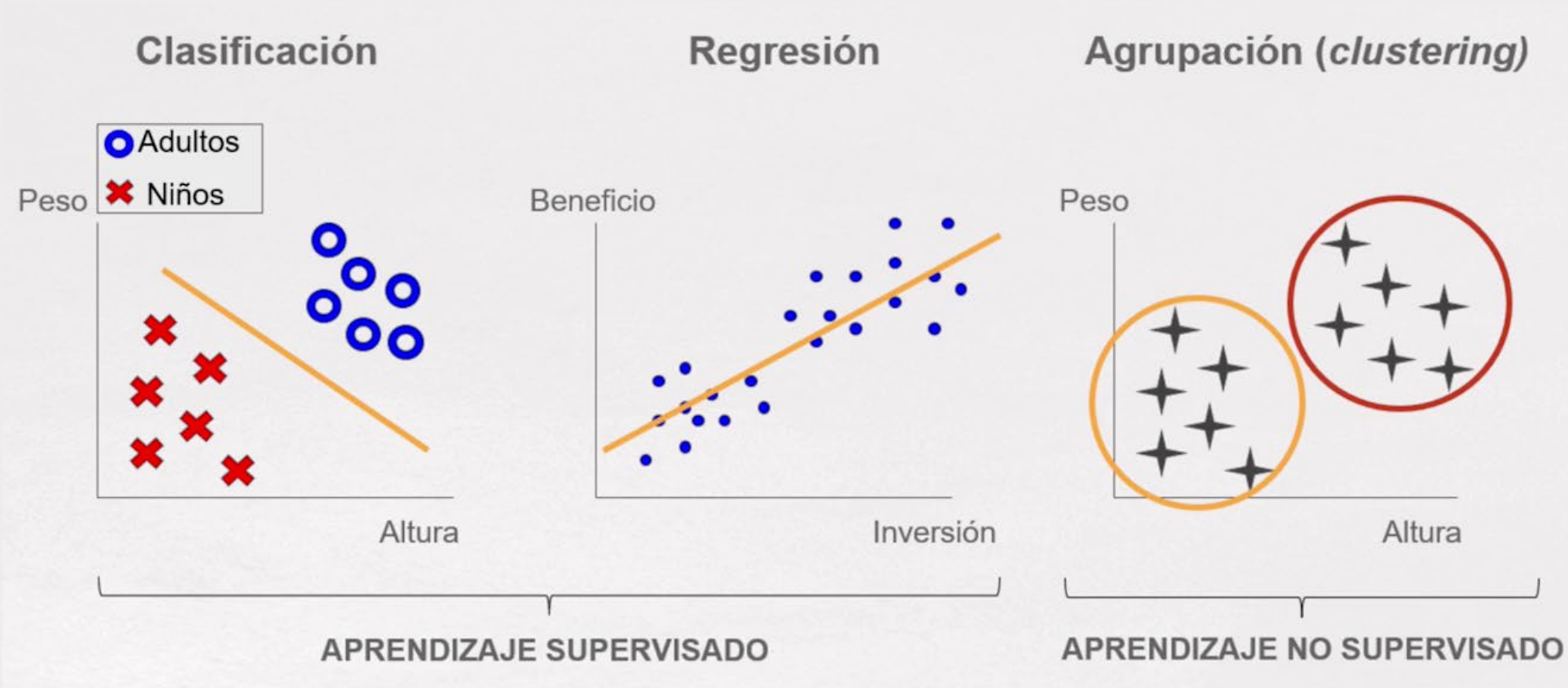
Jonnatan Arias Garcia
jonnatan.arias@utp.edu.co
jariasg@uniquindio.edu.co

David Cardenas peña - dcardenasp@utp.edu.co
Hernán Felipe Garcia - hernanf.garcia@udea.edu.co

Contenido

- Introducción
- Función discriminante
- Estimación de parámetros
- Modelos generativos probabilísticos

Técnicas de Machine Learning

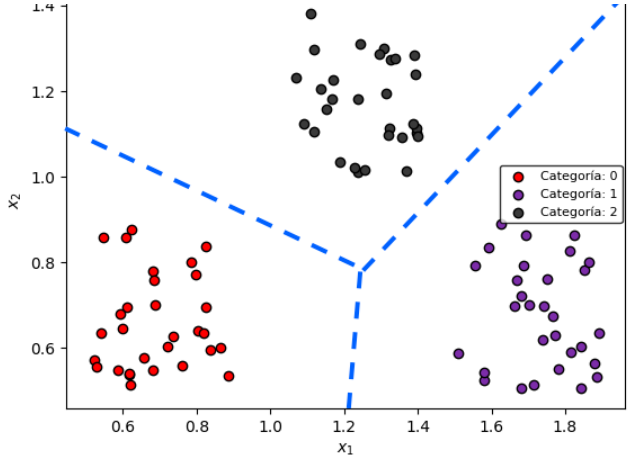


Definiciones (i)

- **Objetivo.** Tomar un vector de entrada, **x**, y asignarlo a una de K clases $\mathcal{C}_k$, para $k = 1, \ldots, K$.

- **Espacio de entrada.** Se divide en regiones de decisión $\mathcal{R}_k$.

- **Líneas o superficies de decisión.** Separan las regiones de decisión.



Definiciones (ii)

- **Modelo lineal.** Las superficies de decisión son funciones lineales de **w**.

- Las superficies están definidas por hiperplanos de dimensión $D - 1$, en un espacio de D dimensiones.

- Si los datos se pueden separar linealmente por una superficie de decisión lineal, se dice que los datos son *linealmente separables*.

Definiciones (iii)

- En regresión, t_n representaba un número real asociado a la entrada $\mathbf{x}_n$.

Etiqueta de clase

- En clasificación, $\mathbf{t}_n$ representa un vector, en codificación 1 de K.

One Hot encoding ó 1 de K: [0,0,0,1] ->4 clases 1 es la pertenencia

- Tres enfoques al problema de clasificación
  - Funciones discriminantes.
  - Modelos generativos, $p(\mathbf{x}, \mathcal{C}_k)$.
  - Modelos discriminativos, $p(\mathcal{C}_k|\mathbf{x})$.

Función Discriminante

F. Discriminante (i)

- Función $y(\boldsymbol{x})$: $\boldsymbol{x} \rightarrow k, k \in \{1, \ldots, K\}$, tales que la decisión de clasificación se toma como:

$$C = argmax_{C_k} y_k(\boldsymbol{x})$$

- La clase asignada será aquella cuya función discriminante sea la mayor.

Caso Binario o biclase

- El modelo lineal esta dado por $y(\boldsymbol{x}) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + w_0$
  - **w** es el vector de pesos
  - $w_0$es el intercepto (Bias) o tendencia

- $\mathbf{x} \in \mathcal{C}_1$, si $y(\mathbf{x}) > 0$. De lo contrario, $\mathbf{x} \in \mathcal{C}_2$.

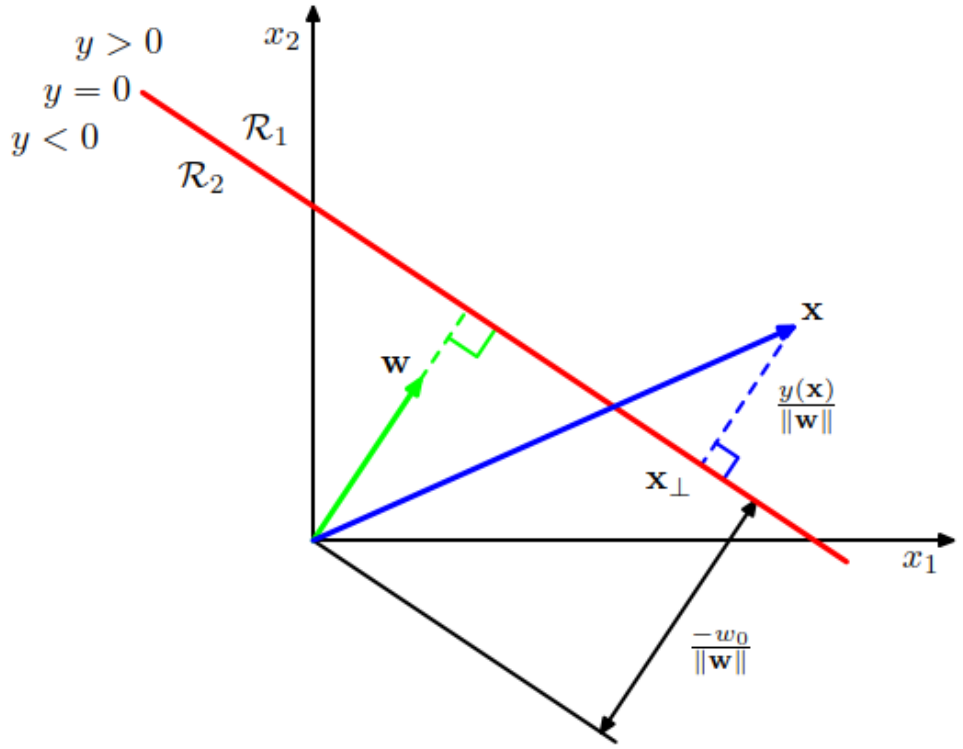
- La línea de decisión o superficie de decisión es $y(\mathbf{x}) = 0$.

F. Discriminante (ii)

El vector **w** es ortogonal a la superficie de decisión.

Distancia de $y(\mathbf{x})$ al origen: $-w_0/\|\mathbf{w}\|$.

Distancia de un punto **x** a $y(\mathbf{x})$: $y(\mathbf{x})/\|\mathbf{w}\|$.

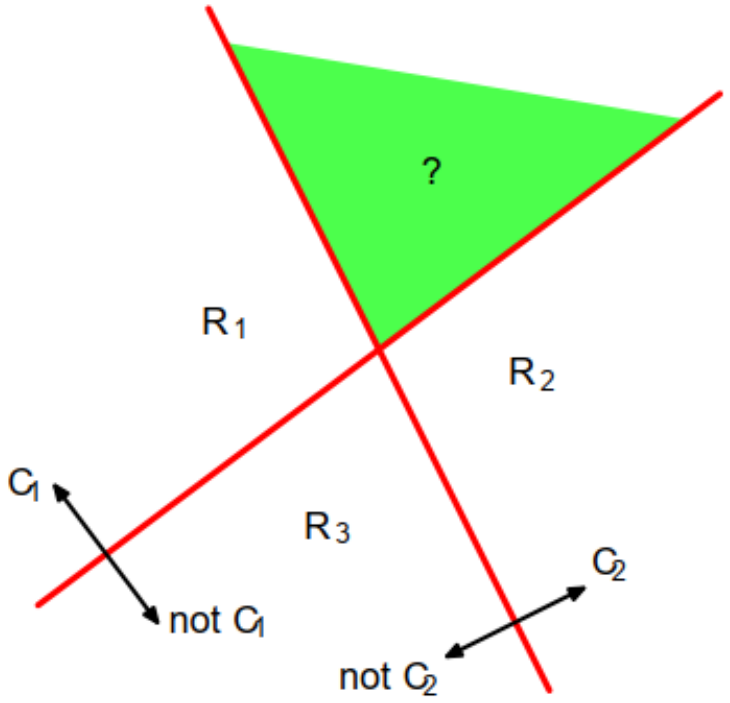


Se puede hacer $\widetilde{\mathbf{x}} = (1, \mathbf{x})$, y $\widetilde{\mathbf{w}} = (w_0, \mathbf{w})$, $y(\mathbf{x}) = \widetilde{\mathbf{w}}^\top \widetilde{\mathbf{x}}$.

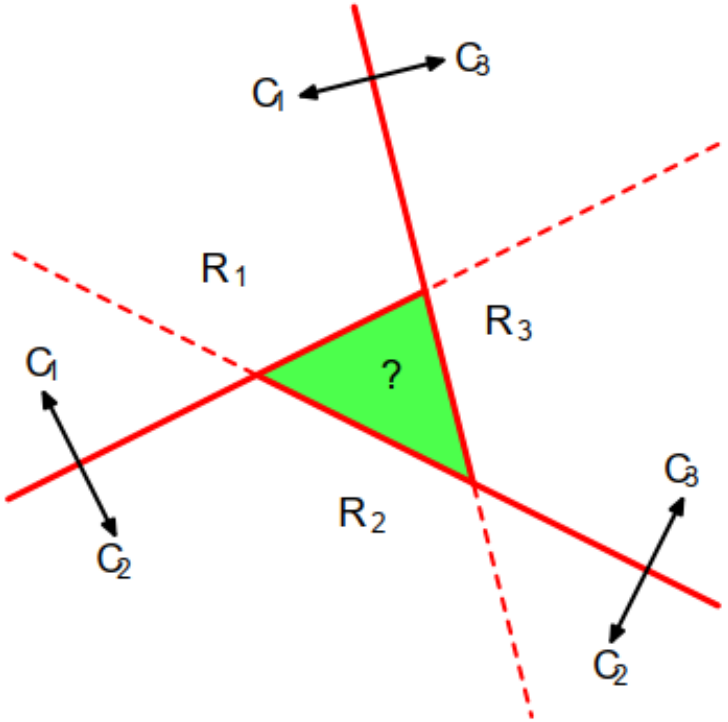
F. Discriminante (iii)

Múltiples clases (I)

Construir un clasificador de *K* clases a partir de clasificadores de 2 clases.



1 vs. todos



1 vs 1

Errores?

F. Discriminante (iv)

Múltiples clases (II)

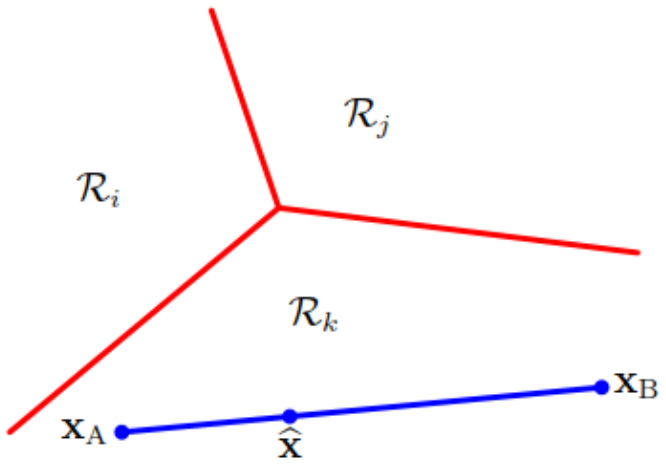
- **Solución:** discriminante de K clases con K funciones lineales

$$y_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0}.$$

- $\mathbf{x} \in C_k$, si $y_k(\mathbf{x}) > y_j(\mathbf{x})$, $k \neq j$.
- La superficie de decisión entre las clases C_k y C_j está dada por $y_k(\mathbf{x}) = y_j(\mathbf{x})$.
- Corresponde a un plano $(D-1)$ dimensional definido por

$$(\mathbf{w}_k - \mathbf{w}_j)^\top \mathbf{x} + (w_{k0} - w_{j0}) = 0.$$

- Regiones de decisión: conectadas de manera simple, y convexas.



Estimación de parámetros funciones discriminantes (entrenamiento)

- Por Mínimos cuadrados
- Por Fisher LDA
- Por perceptrón

Por Mínimos Cuadrado (i)

- Cada clase $\mathcal{C}_k$ se describe por su propio modelo lineal

$$y_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0},$$

para $k \in \{1, \ldots, K\}$.

- Agrupando,

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^\top \widetilde{\mathbf{x}},$$

donde $\widetilde{\mathbf{W}}$ tiene columnas $\widetilde{\mathbf{w}}_k = (w_{k0}, \mathbf{w}_k)^\top$, y $\widetilde{\mathbf{x}} = (1, \mathbf{x}^\top)^\top$.

Por Mínimos Cuadrado (ii)

- Sea un conjunto de entrenamiento $\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^{N}$.
- La matrix $\mathbf{T}$ tiene vectores fila $\mathbf{t}_n^\top$, y la matriz $\widetilde{\mathbf{X}}$ vectores fila $\widetilde{\mathbf{x}}_n$.
- La función de error cuadrático está dada como

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\,\mathrm{tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^\top(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\}.$$

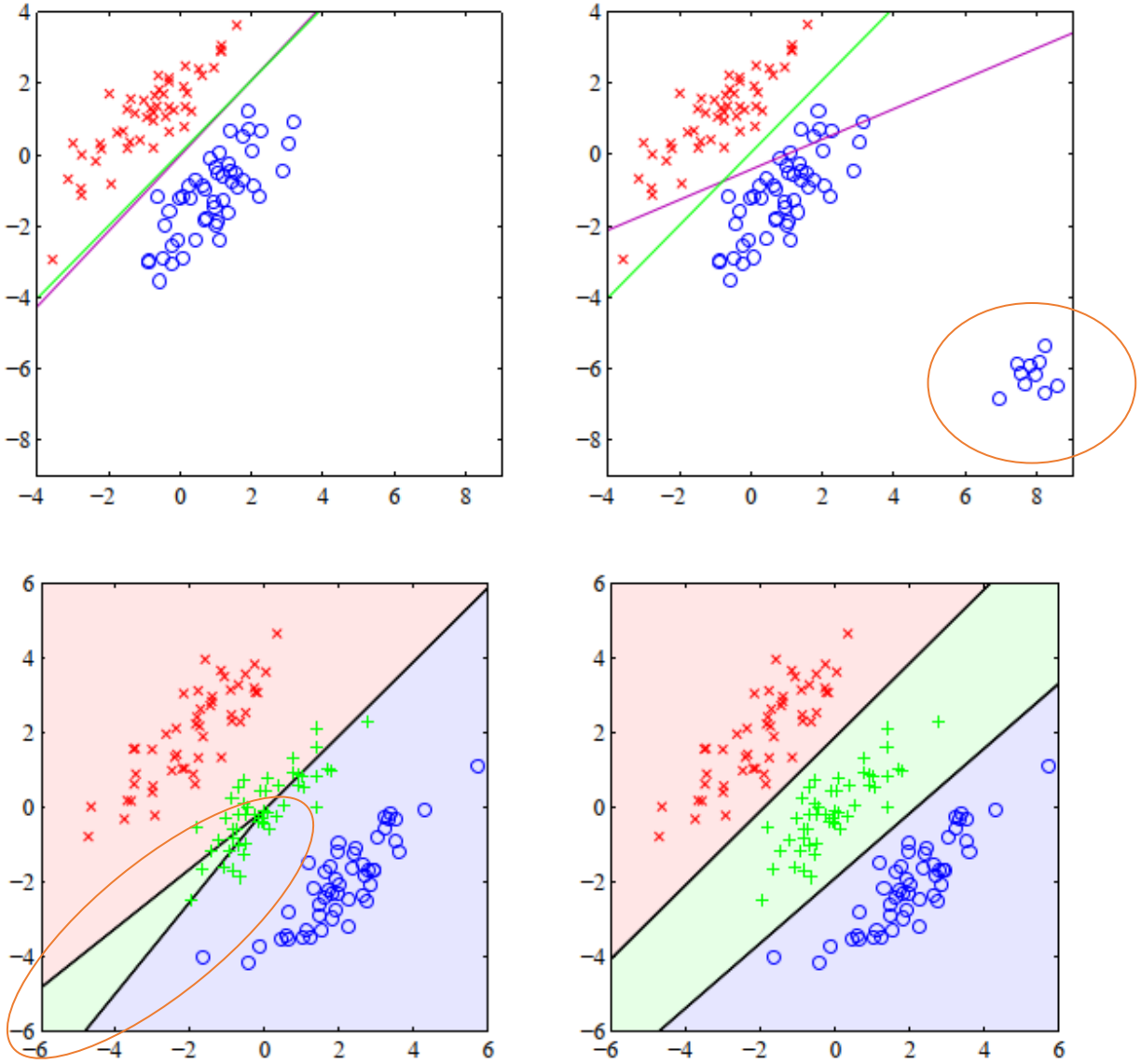
- Minimizando e igualando a cero se obtiene

$$\widetilde{\mathbf{W}}_{MSE} = \left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^\top\mathbf{T} = \widetilde{\mathbf{X}}^\dagger\mathbf{T},$$

  donde $\widetilde{\mathbf{X}}^\dagger$ es la seudo inversa de $\widetilde{\mathbf{X}}$.

- La función discriminante está dada por $\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}_{MSE}^\top\widetilde{\mathbf{x}} = \mathbf{T}^\top\left(\widetilde{\mathbf{X}}^\dagger\right)^\top\widetilde{\mathbf{x}}$.

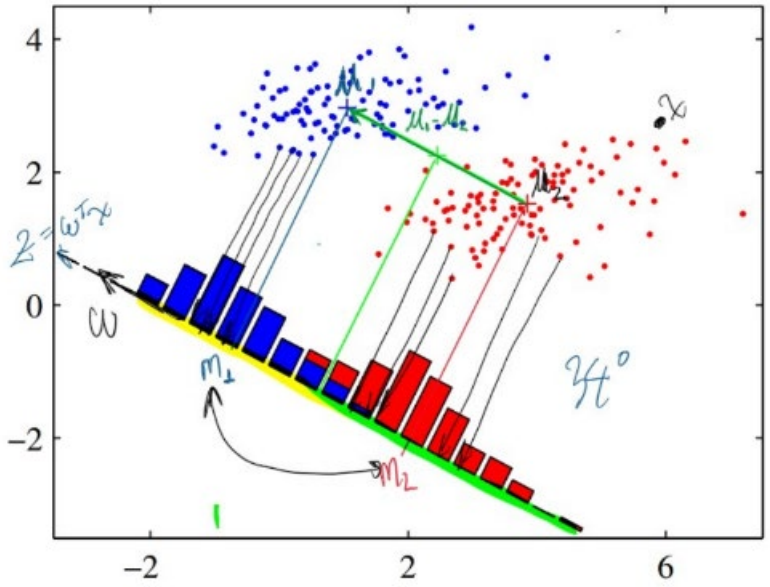
Por Mínimos Cuadrado (iii)



A. Sufre en la presencia de valores atípicos
B. Se diseño para variables aleatoria gaussiana y las etiquetas no necesariamente lo son.
C. La pérdida asociada a una clase puede ser asimétrica. Por ejemplo, los errores en una clase minoritaria pueden ser más costosos que los errores en una clase mayoritaria. Los mínimos cuadrados no capturan directamente esta asimetría en la pérdida.

Por discriminante de Fisher (i)

- La idea es proyectar los datos a un espacio de menor dimensionalidad donde la clasificación sea más sencilla.



- Sea $\mathbf{x} \in \mathbb{R}^D$.

- Se proyecta a una dimensión usando

$$y = \mathbf{w}^\top \mathbf{x}.$$

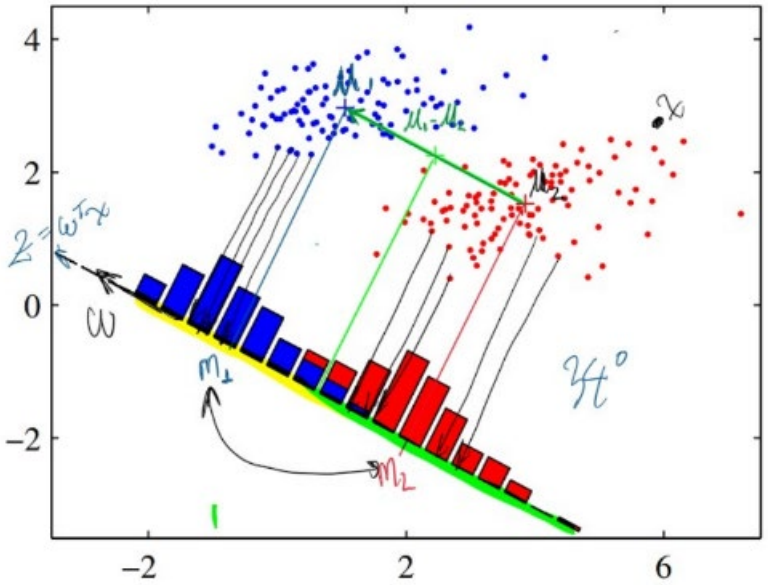
- Se establece un umbral y_0, y se clasifica un nuevo punto como de la clase $\mathcal{C}_1$ si $y \geq y_0$, o de la clase $\mathcal{C}_2$, si pasa lo contrario.

- La idea es escoger **w** de manera que maximice la separabilidad de las clases.

Por discriminante de Fisher (ii)

- Sea un problema biclase, con vectores de media

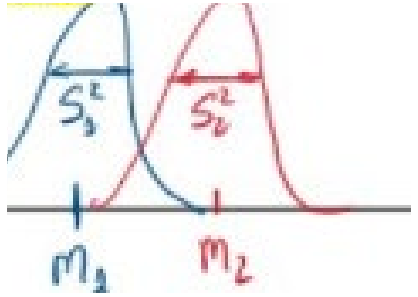
$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n, \qquad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n.$$



- Una medida de la separación entre las clases es

$$m_2 - m_1 = \mathbf{w}^\top (\mathbf{m}_2 - \mathbf{m}_1).$$

- En la expresión anterior $\mathbf{w}$ se puede hacer muy grande, pero se limita para que tenga longitud unitaria ($\mathbf{w}^\top \mathbf{w} = 1$).



- Usando multiplicadores de Lagrange, se puede demostrar que $\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1)$.

Por discriminante de Fisher (ii-A)

Ecuación para la media_k

$$m_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \boldsymbol{w}^\top \boldsymbol{x}_n = \boldsymbol{w}^\top \left(\frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \boldsymbol{x}_n \right) = \boldsymbol{w}^\top \boldsymbol{\mu}_k$$

Distancia de las medias

$$(m_2 - m_1)^2 = (m_1 - m_2)(m_1 - m_2)$$
$$(m_2 - m_1)^2 = \left(\boldsymbol{w^T \mu_1} - \boldsymbol{w^T \mu_2}\right)\left(\boldsymbol{\mu_1^T w} - \boldsymbol{\mu_2^T w}\right)$$
$$(m_2 - m_1)^2 = \boldsymbol{w^T}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\boldsymbol{w}$$
$$(m_2 - m_1)^2 = \boldsymbol{w^T S_B w}$$

Con $\boldsymbol{S}_B = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^{\boldsymbol{T}}$ (Covarianza interclases)

Por discriminante de Fisher (ii-B)

La dispersión dentro de cada clase se estima como la varianza intraclases de los vectores transformados de la clase C_k como:

$$s_k^2 = \sum_{n \in \mathcal{C}_k} \underbrace{(y(\boldsymbol{x}_n)}_{\omega^\top \mathcal{X}_n} - m_k)^2 = \sum_{n \in \mathcal{C}_k} \left(\boldsymbol{w}^\top \boldsymbol{x}_n - m_k\right)^2$$

$$s_k^2 = \sum_{n \in \mathcal{C}_k} \left(\boldsymbol{w}^\top \boldsymbol{x}_n - m_k\right)\left(\boldsymbol{x}_n^\top \boldsymbol{w} - m_k\right)$$

$$s_k^2 = \sum_{n \in \mathcal{C}_k} \left(\boldsymbol{w}^\top \boldsymbol{x}_n - \boldsymbol{w}^\top \boldsymbol{\mu}_k\right)\left(\boldsymbol{x}_n^\top \boldsymbol{w} - \boldsymbol{\mu}_k^\top \boldsymbol{w}\right)$$

$$s_k^2 = \sum_{n \in \mathcal{C}_k} \boldsymbol{w}^T (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{w}$$

$$s_k^2 = \boldsymbol{w}^T \left(\sum_{n \in \mathcal{C}_k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \right) \boldsymbol{w}$$

$$s_k^2 = \boldsymbol{w}^\top \boldsymbol{S}_k \boldsymbol{w}$$

$\boldsymbol{S_k} \in R^{DxD}$ es la covarianza de la clase k y la varianza total dentro de las clases, será:

$$s_1^2 + s_2^2 = \boldsymbol{w}^\top \boldsymbol{S}_1 \boldsymbol{w} + \boldsymbol{w}^\top \boldsymbol{S}_2 \boldsymbol{w}$$

$$s_1^2 + s_2^2 = \boldsymbol{w}^\top (\boldsymbol{S}_1 + \boldsymbol{S}_2) \boldsymbol{w}$$

$$s_1^2 + s_2^2 = \boldsymbol{w}^\top \boldsymbol{S}_w \boldsymbol{w}$$

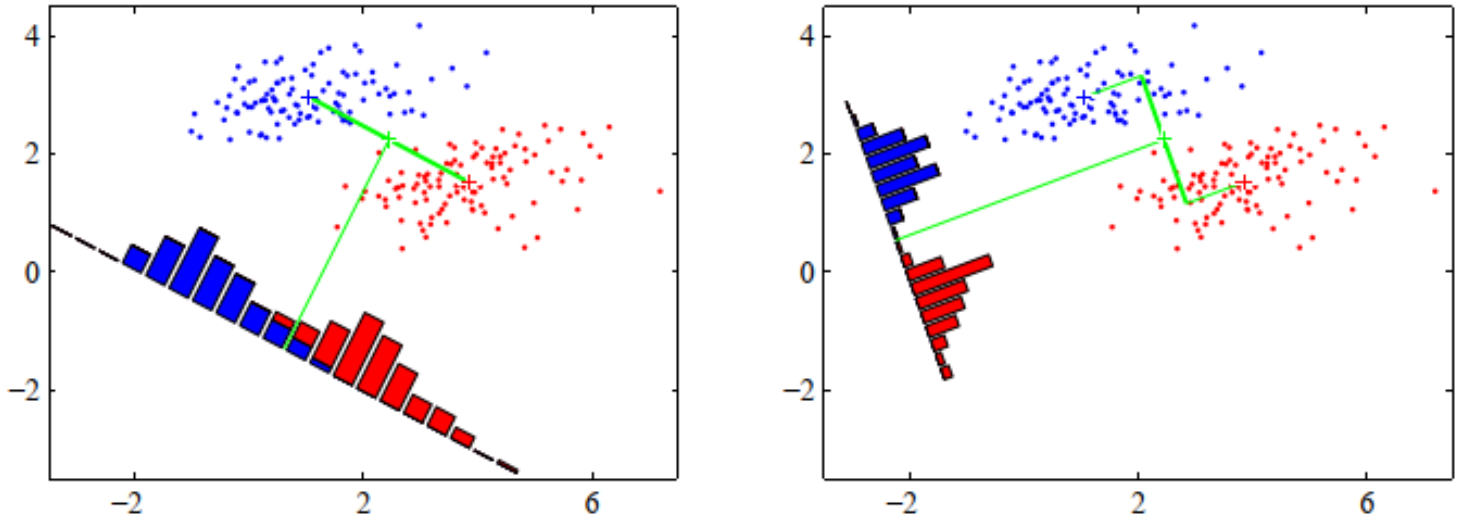
$\boldsymbol{S_w} \in \boldsymbol{R^{DxD}}$ es la covarianza intraclase.

Por discriminante de Fisher (iii)

- No sólo se quiere maximizar la distancia entre las medias, si no también minimizar la variabilidad de las muestras en cada clase.
- La varianza intraclase se obtiene de los vectores transformados de la clase $\mathcal{C}_k$ como

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2,$$

donde $y_n = \mathbf{w}^\top \mathbf{x}_n$.



Por discriminante de Fisher (iv)

- El criterio de Fisher se define como el radio de la varianza entre clases sobre la varianza intraclase

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}.$$

- Haciendo los cambios apropiados se tiene

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}},$$

La idea ahora es maximizar para encontrar el **w** optimo.

donde

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top,$$

$$\mathbf{S}_W = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top.$$

- $\mathbf{S}_B$ es la matriz de covarianza *entre clases*, y $\mathbf{S}_W$ es la matriz de covarianza *intra clases*.

Por discriminante de Fisher (v)

- Derivando $J(\mathbf{w})$ con respecto a $\mathbf{w}$ e igualando a cero se tiene

$$\frac{d}{dw}J(w) = \frac{(2S_B w)(w^\top S_w w) - (w^\top S_B w)(2S_w w)}{(w^\top S_w w)^2} = 0 \qquad (\mathbf{w}^\top \mathbf{S}_B \mathbf{w})\mathbf{S}_W \mathbf{w} = (\mathbf{w}^\top \mathbf{S}_W \mathbf{w})\mathbf{S}_B \mathbf{w}.$$

- Lo que importa de $\mathbf{w}$ es su dirección, no su magnitud.
- Por lo tanto los escalares $\mathbf{w}^\top \mathbf{S}_B \mathbf{w}$ y $\mathbf{w}^\top \mathbf{S}_W \mathbf{w}$ se pueden omitir.
- Además $\mathbf{S}_B \mathbf{w}$ está siempre en la dirección de $\mathbf{m}_2 - \mathbf{m}_1$.
- Premultiplicando por $\mathbf{S}_W^{-1}$ se encuentra que

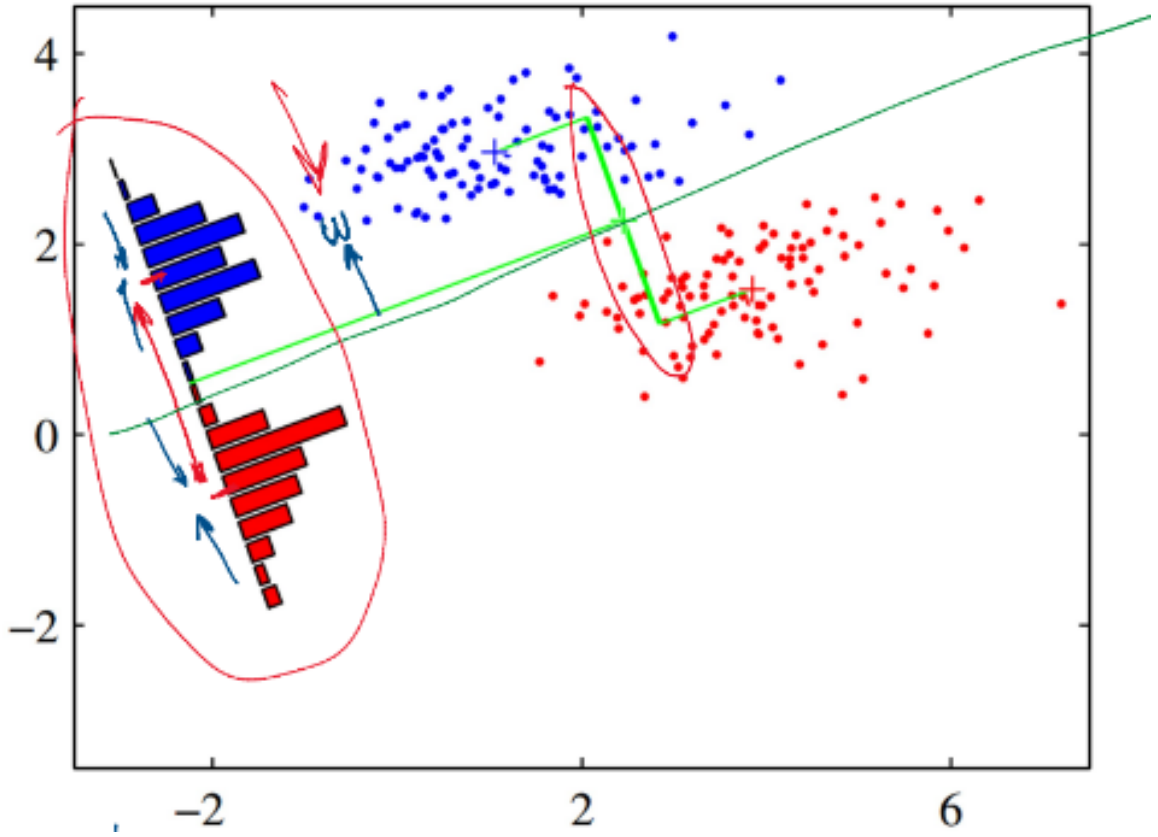
$$\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1).$$

- El resultado se conoce como el *discriminante lineal de Fisher*.

Por discriminante de Fisher (vi)

$$\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1).$$

- El resultado se conoce como discriminante lineal de Fisher (Linear Discriminant Analysis - LDA).



Por Algoritmo del perceptrón (i)

Es el modelo matemático de una neurona

- Dos clases. Los datos de entrada se transforman como $\phi(\mathbf{x})$.

- El modelo lineal tiene la forma

$$y(\mathbf{x}) = f(\mathbf{w}^\top \phi(\mathbf{x})),$$

donde $f(\cdot)$ es la función escalón unitario.

Se conoce como función de activación

- En el perceptrón se asume $t = +1$, para $\mathcal{C}_1$, y $t = -1$ para $\mathcal{C}_2$.

Por Algoritmo del perceptrón (ii)

- La función a minimizar se conoce como el criterio del perceptrón.

$$E_P(\mathbf{w}) = -\sum_{n\in\mathcal{M}} \mathbf{w}^\top \phi(\mathbf{x}_n) t_n,$$

donde $\mathcal{M}$ denota el conjunto de patrones incorrectamente clasificados.

- Aplicando el algoritmo de gradiente descendente estocástico a esta función, se tiene

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta\nabla E_P(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta\phi(\mathbf{x}_n)t_n,$$

donde η se conoce como la razón de aprendizaje, y τ indexa los pasos del algoritmo.

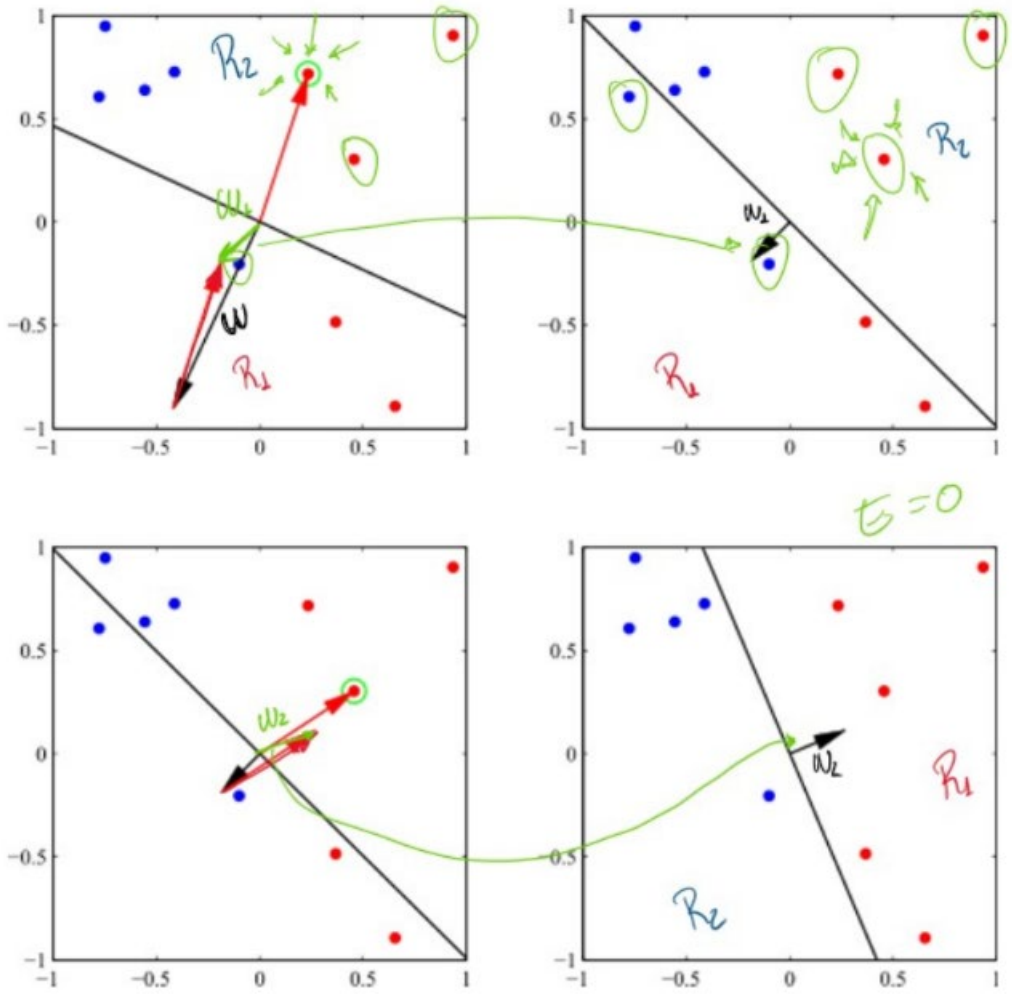
Por Algoritmo del perceptrón (iii)

Pasos:

1. Inicializar el modelo con w_0 y tasa de aprendizaje η,
2. Realizar las predicciones con los parámetros actuales $\boldsymbol{y} = f(\boldsymbol{\phi w_k})$
3. Determinar el conjunto de muestras mal etiquetadas: $\mathcal{M} = \{x_n : \boldsymbol{t_n} \neq y_n\}$
4. Calcular la función de costo total para los parámetros actuales $E_p(\boldsymbol{w_k})$
5. Actualizar los parámetros para la siguiente iteración usando algunas de las muestras mal etiquetadas

$$\boldsymbol{w_{k+1}} = \boldsymbol{w_k} + \boldsymbol{\eta t_{\mathcal{M}}^T \phi_{\mathcal{M}}}$$

1. Hasta la convergencia volver al paso 2.



Modelos generativos probabilísticos

Introducción

- En los modelos generativos se modelan las densidades de clase condicional $p(\mathbf{x}|\mathcal{C}_k)$, y las funciones de probabilidad a priori, $p(\mathcal{C}_k)$.

verosimilitud

prior

- Ambas probabilidades se usan para calcular el posterior $p(\mathcal{C}_k|\mathbf{x})$.

Modelo generativo (i)

Para el caso biclase,

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \frac{1}{1 + \frac{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}}$$

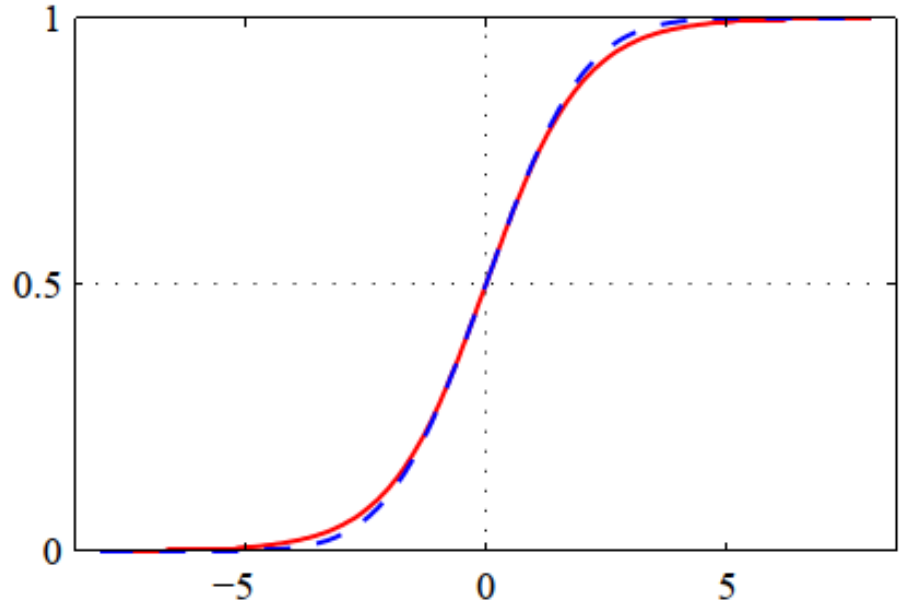
$$= \frac{1}{1 + \exp\left\{\ln\left[\frac{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}\right]\right\}} = \frac{1}{1 + \exp\left\{-\ln\left[\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}\right]\right\}}$$

$$= \frac{1}{1 + \exp(-a)} = \sigma(a),$$

donde $a = \ln\left[\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}\right]$, y $\sigma(a) = 1/(1 + \exp(-a))$ es la función logística sigmoidal .

Modelo generativo (ii)

donde $\sigma(a)$ se denomina función sigmoide:

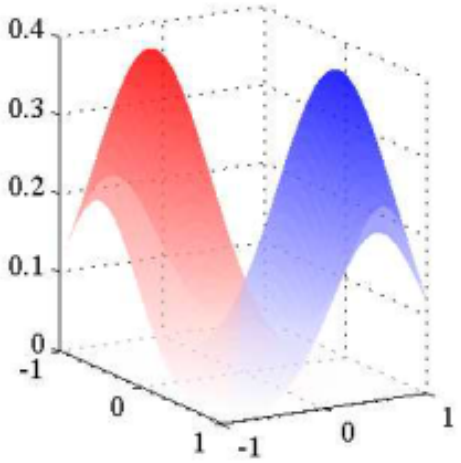


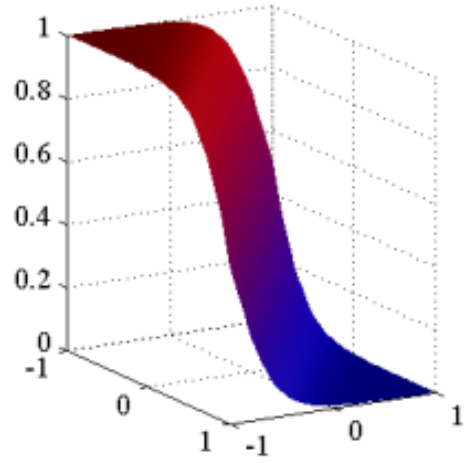
Para el caso $K > 2$ se tiene

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)},$$

con $a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$. (función soft-max).

Modelo generativo: entradas continuas (ii)





Para el caso $K > 2$ se tiene

$$a_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + w_{k0},$$

donde se ha definido $\mathbf{w}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$, $w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k)$.

Modelo generativo: entradas continuas (ii) demostración (A)

CASO BINARIO

- La frontera la componen todos $\boldsymbol{x}$ los tales que

$$p(\mathcal{C}_1|\boldsymbol{x}) = p(\mathcal{C}_2|\boldsymbol{x})$$

$$p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1) = p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2) \rightarrow a = 0 = \ln\left\{\frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}\right\}$$

$$a = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + C$$

$$= -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{x} + \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{x} - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{x} + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + C$$

$$a = -\frac{1}{2}\boldsymbol{x}^T\left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right)\boldsymbol{x} + \left(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\right)\boldsymbol{x} + C$$

$$a = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + C = 0 \quad \text{: Frontera es cuadrática}$$

$$\boldsymbol{A} = 0 \quad a = 0 = \boldsymbol{b}^T\boldsymbol{x} + C \quad \text{: Frontera lineal}$$

$$\boldsymbol{A} = \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1} = 0 \rightarrow \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$$

$$\boldsymbol{b}^T = \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1} = (\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\boldsymbol{\Sigma}^{-1} \rightarrow \boldsymbol{b} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

- Por lo tanto, la probabilidad a posteriori de clase cuando se asumen covarianzas iguales se reduce a:

$$p(\mathcal{C}_1|\boldsymbol{x}) = \sigma(a)$$

$$a = \ln\left(\frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}\right)$$

$$a = \ln(p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)) - \ln(p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2))$$

$$a = \boldsymbol{w}^T\boldsymbol{x} + w_0$$

donde

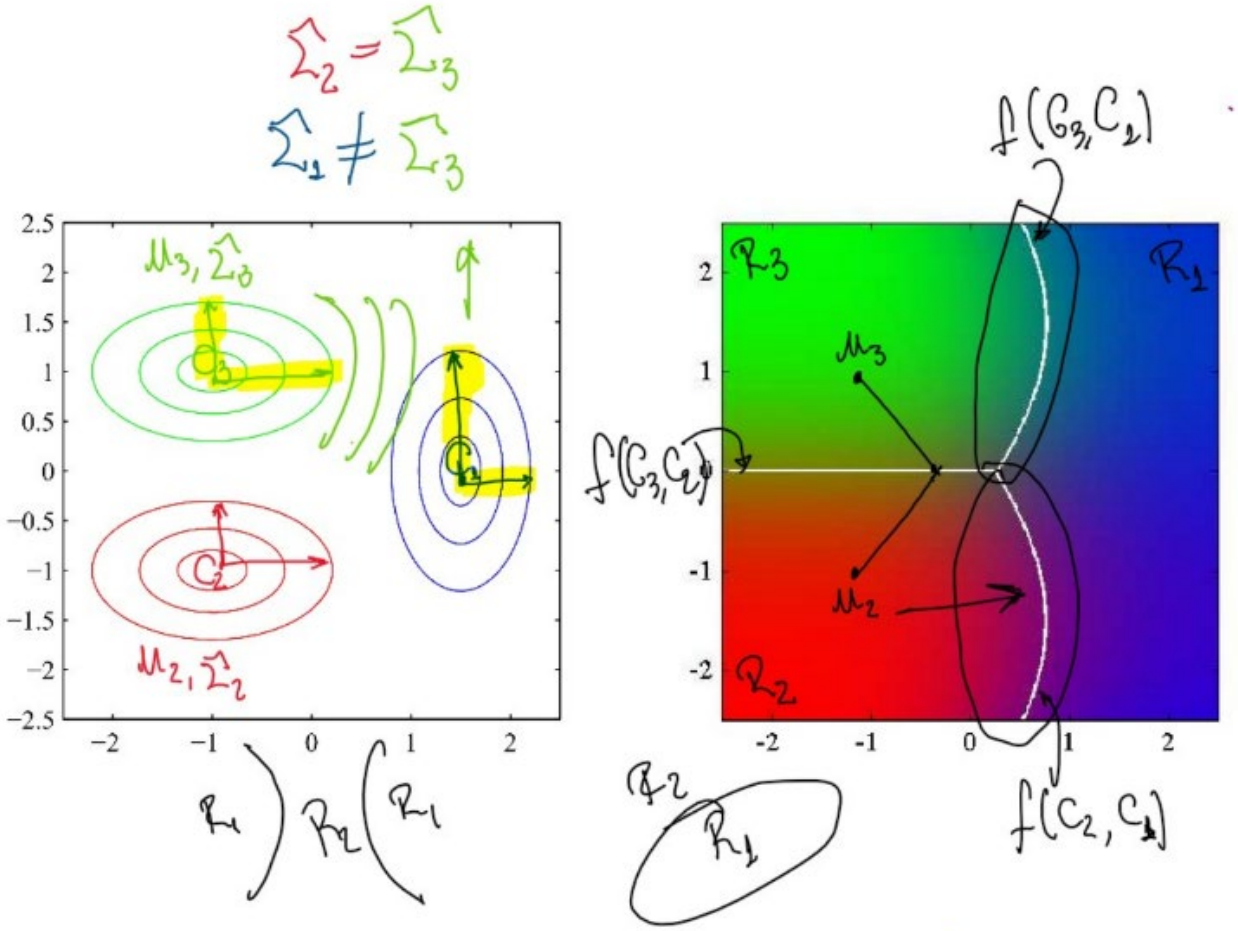
$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad \text{: LDA}$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\left(\frac{\pi}{1-\pi}\right)$$

$$\pi = p(\mathcal{C}_1)$$

Modelo generativo: entradas continuas (iii)

Si las clases no comparten la misma matriz de covarianza, las fronteras de decisión son cuadráticas



Máxima verosimilitud (i)

- Los valores de $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$, $p(\mathcal{C}_1)$ y $p(\mathcal{C}_2)$ se determinan usando máxima verosimilitud.
- Sean dos clases y un conjunto de datos $\{\mathbf{x}_n, t_n\}_{n=1}^{N}$.
- $t_n = 1$ denota $\mathcal{C}_1$, y $t_n = 0$ denota $\mathcal{C}_2$.
- Las probabilidades a priori se escriben como $p(\mathcal{C}_1) = \pi$ y $p(\mathcal{C}_2) = 1 - \pi$.

Máxima verosimilitud (ii)

- Para un punto $\mathbf{x}_n$ de la clase $\mathcal{C}_1$ se tiene $t_n = 1$ y así

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathbf{x}_n|\mathcal{C}_1)p(\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}).$$

- De forma similar, para la clase $\mathcal{C}_2$

$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathbf{x}_n|\mathcal{C}_2)p(\mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}).$$

- Luego,

$$p(t_n, \mathbf{x}_n|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \begin{cases} p(\mathbf{x}_n, \mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), & \text{si } t_n = 1 \\ p(\mathbf{x}_n, \mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), & \text{si } t_n = 0. \end{cases}$$

- Lo anterior se puede escribir de forma resumida como

$$p(t_n, \mathbf{x}_n|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = [\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n}[(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

Máxima verosimilitud (iii)

- Asumiendo que los datos son iid,

$$p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}.$$

- Se realiza la maximización de $\ln p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1 \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, el logaritmo de la verosimilitud, con respecto a $\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$,

Máxima verosimilitud: Solución para $\pi, \mu_1 y\ \mu_2$

- Se puede demostrar que

$$\pi_{ML} = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2},$$

donde N_1 denota el número de puntos de la clase $\mathcal{C}_1$, y N_2 el número de puntos de la clase $\mathcal{C}_2$.

- Igualmente, se puede demostrar que

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n, \qquad \boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1 - t_n)\mathbf{x}_n.$$

Máxima verosimilitud: Solución para Σ

Finalmente, para Σ

$$\Sigma = \mathbf{S} = \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2$$

$$\mathbf{S}_1 = \frac{1}{N_1}\sum_{n \in \mathcal{C}_1}(\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^{\top},$$

$$\mathbf{S}_2 = \frac{1}{N_2}\sum_{n \in \mathcal{C}_2}(\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^{\top}.$$

Demostración solución de $\pi, \mu_1 y\ \mu_2$

$$\max_{\mu_1,\mu_2,\Sigma,\pi} \ln p(\boldsymbol{t}, \mathbf{X}|\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \pi)$$

$$L = \sum_{n=1}^{N} t_n \left[-\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_1| - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) + \ln \pi_1 \right]$$
$$+ \sum_{n=1}^{N} (1 - t_n) \left[-\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_2| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) + \ln(1 - \pi_1) \right]$$

$$L = \sum_{n=1}^{N} t_n \left[-\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_1| - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) + \ln \pi_1 \right]$$
$$+ \sum_{n=1}^{N} (1 - t_n) \left[-\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_2| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) + \ln(1 - \pi_1) \right]$$

$$\frac{\partial L}{\partial \pi_1} = \sum_n \frac{t_n}{\pi_1} + \sum_n (1 - t_n) \cdot \frac{-1}{1 - \pi_1} = 0$$

$$\frac{1}{\pi_1} \sum_n t_n - \frac{1}{1 - \pi_1} \sum_n (1 - t_n) = 0$$

N_1 $\qquad$ N_2

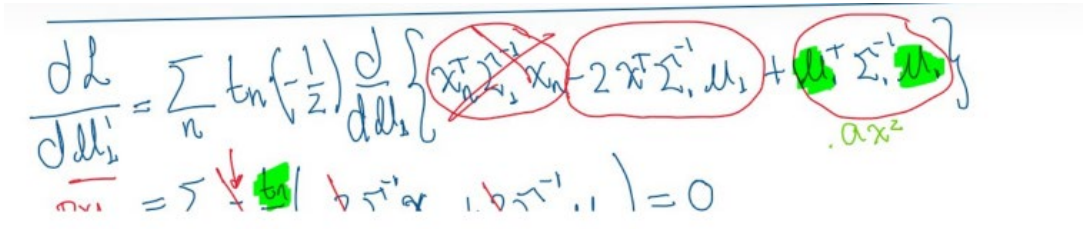
$$\frac{N_1}{\pi_1} = \frac{N_2}{1 - \pi_1} \rightarrow (1 - \pi_1) N_1 = \pi_1 N_2$$

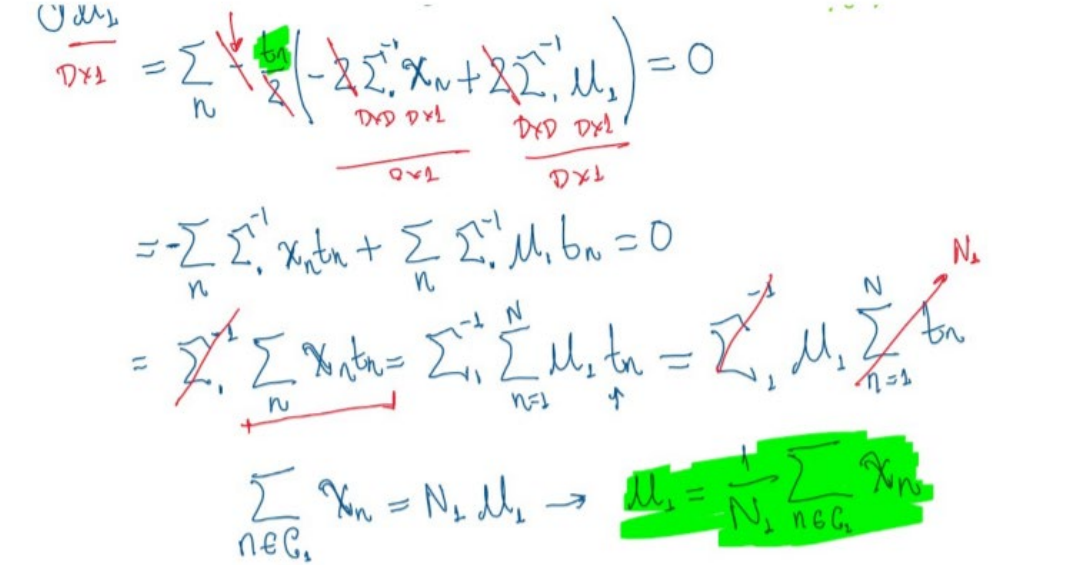
$$\hookrightarrow N_1 = \pi_1 N_1 + \pi_1 N_2 \rightarrow \pi_1 = \frac{N_1}{N_1 + N_2} = \frac{N_1}{N}$$

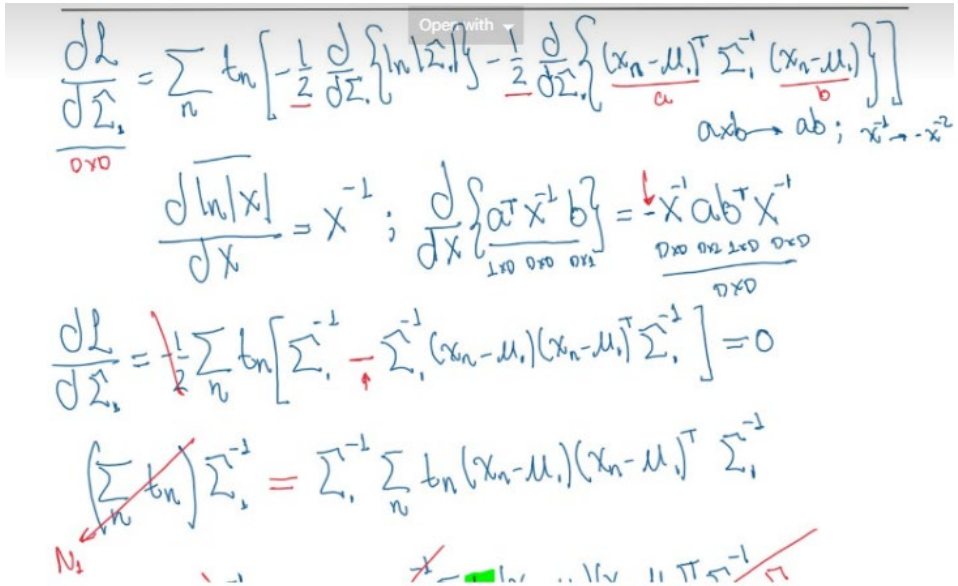
$N = 6$

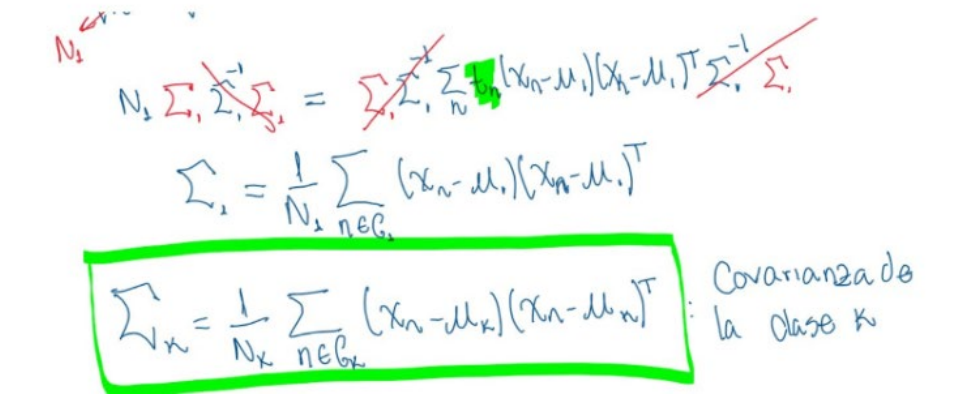
n	t_n	$1 - t_n$
1	0 ✗	1 ✓
2	1 ✓	0 ✓
3	0 ✗	1 ✓
4	0 ✗	1 ✓
5	1 ✓	0 ✓
6	0 ✗	1 ✓
	$N_1 = 2$	$4 = N_2$

Demostración solución de $\pi, \mu_1 y\ \mu_2$









Demostración solución de $\pi, \mu_1 y\ \mu_2$

- Por lo tanto:

$$\pi = \frac{N_1}{N}$$

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \boldsymbol{x}_n$$

$$\boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1 - t_n)\boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n (\boldsymbol{x}_n - \boldsymbol{\mu}_1)(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^{\mathsf{T}}$$

$$\boldsymbol{\Sigma}_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1 - t_n)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)^{\mathsf{T}}$$