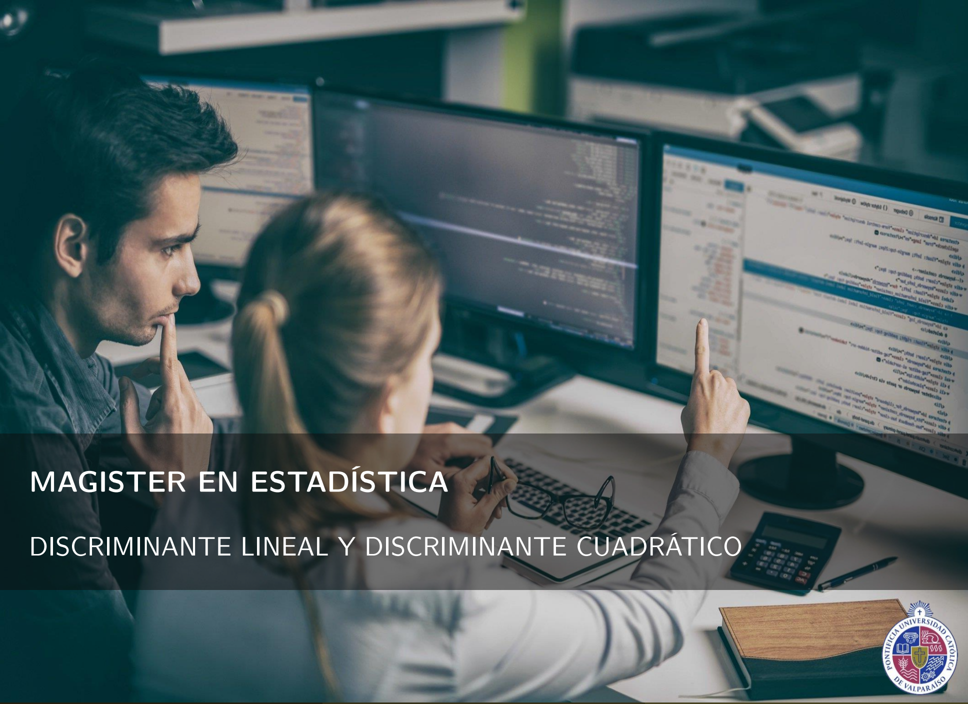

MAGISTER EN ESTADÍSTICA

DISCRIMINANTE LINEAL Y DISCRIMINANTE CUADRÁTICO

Agenda

1. Problema de Clasificación
2. Clasificación por densidades
3. Discriminante Lineal
4. Discriminante Cuadrático
5. Matriz de confusión

Problema de Clasificación

Considere una variable respuesta (Y), la cual tiene como recorrido un conjunto de K características.

Denote estas carácterísticas por G al grupo de las K clases. Es decir, $G = 1,2,3,\ldots, K$.

El problema de Clasificación consiste en asignar Y_0, dado un conjunto de covariables X, un valor de G.

Para tal fin se deben definir alguna medida de error.

Problema de Clasificación

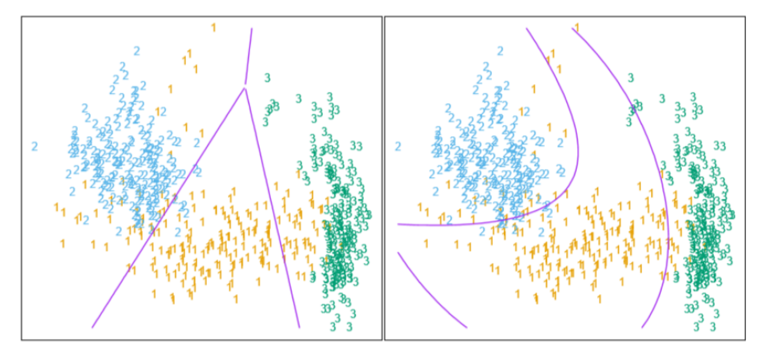

Figure: A la izquierda el clasificador x y a la derecha $\widetilde{x}$. Fuente Hastie and Tibshirani (2008)

.

Clasificación por densidades

Considere dos poblaciones p-variadas cada una, denotadas por P_1 y P_2.

Definición

Clasificar un individuo en algunas de estas dos poblaciones significa entregar una regla que permita discriminar dicha clasificación.

Clasificación por densidades

Existen dos errores de Clasificación:

- Clasificar un individuo proveniente de π_1 como si perteneciera a π_2.
- Clasificar un individuo proveniente de π_2 como si perteneciera a π_1.

$C(i|j)$: Costo de clasificar erroneamente un individuo procedente de π_j como si perteneciera a π_i, $i = 1, 2$.

		Decisión	
		π_1	π_2
Verdadero	π_1	0	C(2|1)
	π_2	C(1|2)	0

Clasificación por densidades

Notación:

- q_i: Proporción de individuos en π_i.
- $f_i(x)$: densidad de X cuando el individuo pertenece a π_i.

Supuesto: $x \in \mathbb{R}^p = R_1 \cup R_2$, tal que R_1 y R_2 son conjuntos disjuntos. **Decisión**:

- si $x \in R_1$ entonces x se clasifica como π_1.
- si $x \in R_2$ entonces x se clasifica como π_2.

Criterio: R_1 y R_2 serán determinados de modo que el costo de clasificar sea mínimo.

Clasificación por densidades

Sean $c_1 = q_1 \cdot C(2|1)$ y $c_2 = q_2 \cdot C(1|2)$ costos medios de clasificar erroneamente. Note que basta con encontrar sólo R_1, luego la función objetivo viene dada por:

$$\text{minimizar} \int_{R_{\mathbf{1}}} [c_2 f_2(x) - c_1 f_1(x) dx] \tag{1}$$

Clasificación por densidades

Proposición

La solución al problema (1) es:

$$R_1 = \{x \in \mathbb{R}^p : \; c_1 f_1(x) \geq c_2 f_2(x)\}$$
$$= \left\{x \in \mathbb{R}^p : \; \frac{f_1(x)}{f_2(x)} \geq k\right\},$$

donde $k = \frac{c_2}{c_1} = \frac{q_2 \cdot C(1|2)}{q_1 \cdot C(2|1)}$.

Clasificación: Discriminante lineal

Si $C(1|2) = C(2|1)$ y $q_1 = q_2$, entonces:

$$R_1 = \left\{ x \in \mathbb{R}^p : \frac{f_1(x)}{f_2(x)} \geq 1 \right\}$$
$$= \left\{ x \in \mathbb{R}^p : ln\left(\frac{f_1(x)}{f_2(x)}\right) \right\}.$$

Si además, $\pi_i = \mathcal{N}_p(\mu^{(i)}, \Sigma)$, entonces $ln\left(\frac{f_1(x)}{f_2(x)}\right) = xb - a \geq 0$, donde

$$b = \Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$$
$$a = \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)^T \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

Este método se conoce como **Discriminante Lineal**.

Clasificación: Discriminante Lineal

Considere $U = x^T\ a - b \geq 0$, entonces U tiene una distribución normal univariada, dónde:

- Si $x \in R_1$, entonces $U \sim \mathcal{N}\left(\frac{\Delta^2}{2}, \Delta^2\right)$.
- Si $x \in R_2$, entonces $U \sim \mathcal{N}\left(-\frac{\Delta^2}{2}, \Delta^2\right)$.

Donde:

$$\Delta^2 = (\mu^{(1)} - \mu^{(2)})^T \Sigma^{-1} (\mu^{(1)} - \mu^{(2)})^T$$

$P(1|2) = \mathbb{P}[U \geq 0]$: La probabilidad de clasificar erróneamente en la población π_1. La probabilidad se obtiene con $U \sim \mathcal{N}\left(\frac{\Delta^2}{2} \Delta^2\right)$

Clasificación: Discriminante Lineal

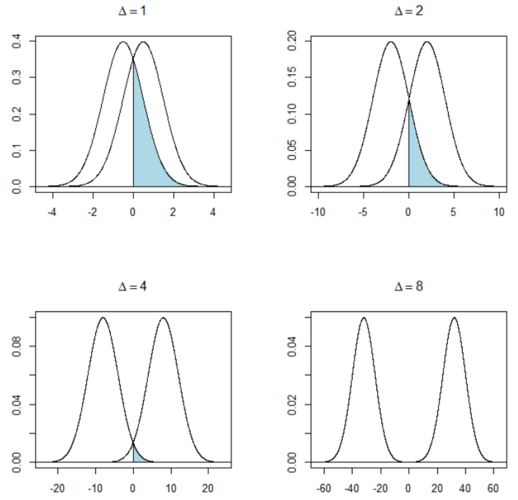

Figure: Comparación de LDA para diferentes Δ

Clasificación: Discriminante Lineal

Observación

Si $\mu^{(1)}, \mu(2)$ y Σ son desconocidos entonces deben ser estimados a partir de los datos. En base a un $m.a(n_i)$ desde π_i, $i = 1, 2$, se tiene que:

- $\widehat{\mu}^(i) = \bar{X}^i$, $i = 1, 2$.
- $\widehat{\Sigma} = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2} = S_{pooled}$, donde S_i, corresponde al estimador de Σ desde la muestra π_i.

Observación

Si q_i son desconocidos entonces pueden ser estimados a partir de los datos como:

$$\widehat{q}_i = \frac{n_i}{n}, \quad \text{donde } n = n_1 + n_2$$

Clasificación: Discriminante Cuadrático

Si $C(1|2) = C(2|1)$ y $q_1 = q_2$, entonces

$$R_1 = \left\{ x \in \mathbb{R}^p : ln\left(\frac{f_1(x)}{f_2(x)} \geq 0\right)\right\}$$

Si además, $\pi_i = \mathcal{N}_p(\mu^{(i)}, \Sigma^{(i)})$, entonces:

$$ln\left(\frac{f_1(x)}{f_2(x)}\right) = xCx^T + xb - a \geq 0,$$

donde C, b y a se encuentran simplifiacando el cuociente anterior. Este método se conoce como **Discriminante Cuadrático**.

Clasificación: Discriminante Cuadrático

Observación

La extensión para K poblaciones es natural. Si $\mathbb{R}^p = R_1 \cup R_2 \cup \cdots \cup R_K$ disjuntos.

Entonces, si los $C(i|j)$ son todos iguales, la región que minimiza el costo de clasificación erróneo es:

$$
\begin{aligned}
R_k &= \left\{ x \in \mathbb{R}^p : \frac{f_k(x)}{f_j(x)} \geq \frac{q_j}{q_k} \right\} \\
&= \left\{ x \in \mathbb{R}^p : ln\left(\frac{f_k(x)}{f_j(x)} \right) \geq ln\left(\frac{q_j}{q_k} \right) \right\}
\end{aligned}
$$

Costo Total $= \sum_{i=1}^{K} q_i \sum_{j=1, j\neq i}^{K} C(i|j) \int_{R_j} f_i(x)dx$

Matriz de confusión

		Predicho	
		A	B
Verdadero	A	n_{1c}	n_{1m}
	B	n_{2m}	n_{2c}

donde:

- n_{1c}: Observaciones de la población 1 clasificadas correctamente.
- n_{2c}: Observaciones de la población 2 clasificadas correctamente.
- n_{1m}: Observaciones de la población 1 clasificadas erróneamente en la población 2.
- n_{2m}: Observaciones de la población 2 clasificadas erróneamente en la población 1.

Matriz de confusión

- En general, esta se construye para validar nuevas reglas de clasificación usando muestras de entrenamiento.
- A partir de la tabla de confusión se pueden medir diferentes tipos de indicadores, el más común es:

$$APER = \frac{n_{1m} + n_{2m}}{n_1 + n_2}, \quad n_i = n_{ic} + n_{im}, \quad i = 1, 2.$$

  Representa la proporción de individuos de la muestra de entrenamiento que fueron erróneamente clasificados.
- Problema: APER es sesgado y usualmente subestima el verdadero valor. Además no mejora cuando crece n.

Matriz de confusión

	Predicho	
Observación	Verdadero positivo (VP)	Falso negativo (FN)
	Falso positivo (FP)	Verdadero negativo (VN)

dónde:

- VP es la cantidad de positivos que fueron clasificados correctamente como positivos por el modelo.
- VN es la cantidad de negativos que fueron clasificados correctamente como negativos por el modelo.
- FN es la cantidad de positivos que fueron clasificados incorrectamente como negativos.
- FP es la cantidad de negativos que fueron clasificados incorrectamente como positivos.

Matriz de confusión

$$\text{Exactitud} = \frac{VP + VN}{\text{Total}}$$

$$\text{Sensibilidad} = \frac{VP}{\text{Total Positivos}}$$

$$\text{Especificidad} = \frac{VN}{\text{Total Negativos}}$$

$$\text{Precisión} = \frac{VP}{\text{Total clasificados positivos}}$$

Matriz de confusión

- Alta Precisión y alta Sensibilidad: El modelo maneja perfectamente la clase.
- Alta Precisión y baja Sensibilidad: El modelo no detecta la clase muy bien, pero cuando lo hace es altamente confiable.
- Baja Precisión y alta Sensibilidad: El modelo detecta bien la clase, pero también incluye muestras de otras clases.
- Baja Precisión y baja Sensibilidad: El modelo no logra clasificar la clase correctamente.

Cuando tenemos un conjunto de datos con desequilibrio, suele ocurrir que obtenemos un alto valor de precisión en la clase **mayoritaria** y una baja sensibilidad en la clase **minoritaria**.