Estadística II
Grado en Matemáticas, UAM, 2020-2021

Hoja 4 (Clasificación)

---

1. Considera el conjunto de datos bidimensionales correspondientes a dos poblaciones π_0 y π_1:

X_1	X_2	Y
3	7	0
2	4	0
4	7	0
6	9	1
5	7	1
4	8	1

a) Estima, a partir de estos datos, la función lineal discriminante de Fisher.

b) Clasifica la observación $\mathbf{x} = (2, 7)^\intercal$ utilizando la regla obtenida en el apartado anterior.

2. *Sobre los coeficientes de la regresión logística y del probit.* En el modelo de regresión logística (*logit*), con observaciones $\mathbf{x} = (x_1, \ldots, x_k)^\intercal$ de dimensión k, se postula que $p(\mathbf{x}) := \mathbf{P}(Y = 1|\mathbf{x}) = h(\boldsymbol{\beta}^\intercal \cdot \widetilde{\mathbf{x}})$, donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\intercal$ es el vector de parámetros, $\widetilde{\mathbf{x}} = (1, x_1, \ldots, x_k)^\intercal$, y $h(x) = 1/(1 + e^{-x})$.

El modelo *probit* sigue el mismo planteamiento, salvo que $h(x) = \Phi(x)$, donde Φ es la función de distribución de la normal estándar.

En este ejercicio se analiza el significado de cada parámetro β_j, en ambos modelos, como respuesta a variaciones de la variable regresora j-ésima. Usaremos la siguiente notación: dada una observación $\mathbf{x} = (x_1, \ldots, x_k)$, le asociamos $\mathbf{x}^{(j)} = (x_1, \ldots, x_j + 1, \ldots, x_k)$, en la que la coordenada j ha subido una unidad.

a) En el modelo *logit*, halla una fórmula para β_j en función de $p(\mathbf{x})$ y $p(\mathbf{x}^{(j)})$.

b) Supongamos que $\beta_j = 2$. La observación $\mathbf{x}$ tiene probabilidad 30 %. ¿Cuánto vale $p(\mathbf{x}^{(j)})$?

c) Repite las dos cuestiones anteriores en el modelo *probit*.

3. a) Considera las dos siguientes funciones de densidad, correspondientes a la distribución de una variable X en dos poblaciones π_0 y π_1:

$$f_0(x) = 1 - |x|, \quad \text{para } |x| \leq 1,$$
$$f_1(x) = 1 - |x - 1/2|, \quad \text{para } -1/2 \leq x \leq 3/2.$$

Identifica las regiones R_0 (de clasificación en π_0) y R_1 (de clasificación en π_1) para los casos $p_0 = 50\,\%$ y $p_0 = 20\,\%$.

b) Calcula, en el caso $p_0 = p_1$, la probabilidad de mala clasificación.

c) Repite el apartado a) tomando $p_0 = p_1$, $f_0(x)$ como allí, y $f_1(x) = \frac{1}{4}(2 - |x - 1/2|)$, para $-3/2 \leq x \leq 5/2$.

4. Un vector $\mathbb{X} = (X_1, X_2)^\intercal$ se distribuye, en dos poblaciones π_0 y π_1, como se indica a continuación:

- En π_0, las variables X_1 y X_2 son dos normales estándar independientes.
- En π_1, el vector $\mathbb{X}$ se distribuye uniformemente en el rectángulo centrado en el origen cuyos lados vertical y horizontal miden 2 y $e\pi$, respectivamente.

Ponemos que las probabilidades a priori de cada población son iguales. Identifica (y dibuja con precisión) las regiones óptimas R_0 y R_1 de clasificación en π_0 y π_1, respectivamente.

5. Un vector $\mathbb{X} = \left(\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right)$ se distribuye, en dos poblaciones π_0 y π_1 (con probabilidades a priori iguales), como se indica a continuación:

- En la población π_0, las variables X_1 y X_2 son dos normales independientes tales que $\mathbf{E}(X_1) = \mathbf{E}(X_2) = a$ y $\mathbf{V}(X_1) = \mathbf{V}(X_2) = 1$.
- En la población π_1, el vector $\mathbb{X}$ se distribuye como una normal bidimensional de parámetros $\mathbf{m} = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$ y $V = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

Se sabe que a es un entero positivo, y que el punto $(5/2, 0)$ pertenece a la región R_1 de clasificación en π_1, mientras que el punto $(1, 5/2)$ pertenece a R_0, la región de clasificación en π_0.

¿Cuánto vale a? Representa gráficamente R_0 y R_1.

6. Considera las dos siguientes funciones de densidad, correspondientes a la distribución de un par de variables (X_1, X_2) en dos poblaciones π_0 y π_1:

$$f_0(x_1, x_2) = e^{-x_1 - x_2}, \quad \text{para } x_1, x_2 > 0,$$
$$f_1(x_1, x_2) = \frac{4}{\pi}\, e^{-(x_1^2 + x_2^2)}, \quad \text{para } x_1, x_2 > 0.$$

Ponemos $p_1 = p_0$. Identifica la región R_1 (clasificación en π_1) y dibújala (con cierto detalle).

7. Considera las dos siguientes funciones de densidad, correspondientes a la distribución de un trío de variables (X_1, X_2, X_3) en dos poblaciones π_0 y π_1:

$$f_0(x_1, x_2, x_3) \text{ es la función indicadora del cubo unidad (tridimensional),}$$
$$f_1(x_1, x_2, x_3) = 12\, x_1^2\, x_2\, x_3, \quad \text{para } 0 \le x_1, x_2, x_3 \le 1.$$

Supongamos que $p_1 = p_0$. Identifica la región R_1 (de clasificación en π_1) y, ups, dibújala.

8. Supongamos que la distribución del vector k-dimensional $\mathbb{X}$ en la población π_0 es normal con vector de medias $\boldsymbol{\mu}_0$ y matriz de covarianzas Σ, mientras que la distribución de $\mathbb{X}$ en π_1 es normal con vector de medias $\boldsymbol{\mu}_1$ y matriz de covarianzas Σ (caso homocedástico). Suponemos que las probabilidades a priori de ambas poblaciones son iguales, $p_0 = p_1 = 1/2$.

Recuerda que, en este caso, las regiones de clasificación vienen dadas por

$$R_1 \quad \longrightarrow \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \cdot \Sigma^{-1} \cdot \mathbf{x} \ge \frac{1}{2}\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \cdot \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0),$$
$$R_0 \quad \longrightarrow \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \cdot \Sigma^{-1} \cdot \mathbf{x} < \frac{1}{2}\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \cdot \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

Analiza la variable aleatoria $Y = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \cdot \Sigma^{-1} \cdot \mathbb{X}$ (en las dos poblaciones) y concluye que la probabilidad de clasificación errónea es $1 - \Phi(\Delta/2)$, donde Φ denota la función de distribución de una normal estándar, y

$$\Delta^2 := (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\intercal \cdot \Sigma^{-1} \cdot (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

9. *El caso de más de dos poblaciones.* Fijamos $m \geq 2$. Supongamos que tenemos las poblaciones $\pi_1, \ldots, \pi_m$, con probabilidades a priori $p_1, \ldots, p_m$. El vector k-dimensional $\mathbb{X}$ se distribuye con funciones de densidad $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$ en las respectivas poblaciones.

Digamos que una cierta regla de clasificación divide el espacio Ω en regiones $R_1, \ldots, R_m$.

a) Comprueba que la probabilidad de mala clasificación viene dada por

$$\sum_{i=1}^{m} p_i \Big(\sum_{1 \leq j \leq m,\, j \neq i} \int_{R_j} f_i(\mathbf{x})\, d\mathbf{x} \Big).$$

b) Dada una observación $\mathbf{x}$, calcula las distintas probabilidades a posteriori $\mathbf{P}(\pi_i|\mathbf{x})$.

c) Deduce que la regla de clasificación de la observación $\mathbf{x}$ dada por "clasificamos $\mathbf{x}$ como proveniente de π_i si $\mathbf{P}(\pi_i|\mathbf{x})$ es máxima" se traduce en

clasificamos $\mathbf{x}$ como proveniente de π_i si $p_i f_i(\mathbf{x}) > p_j f_j(\mathbf{x})$ para todo $j \neq i$.

Si hay "empates", se puede clasificar $\mathbf{x}$ en cualquiera de las poblaciones empatadas.

(La regla es equivalente a $\ln(p_i\, f_i(\mathbf{x})) > \ln(p_j\, f_j(\mathbf{x}))$ para todo $j \neq i$, siempre que las funciones de densidad no se anulen en $\mathbf{x}$).

d) Supongamos que cada $f_i(\mathbf{x})$ es una normal con media $\boldsymbol{\mu}_i$ y matriz de covarianzas Σ_i. Comprueba que si definimos, para cada $i = 1, \ldots, m$,

$$d_i(\mathbf{x}) = -\frac{1}{2}\, \ln(|\Sigma_i|) - \frac{1}{2}\, (\mathbf{x} - \boldsymbol{\mu}_i)^{\intercal} \cdot \Sigma_i^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu}_i) + \ln(p_i),$$

entonces la regla de clasificación es: "clasificamos $\mathbf{x}$ en π_i si $d_i(\mathbf{x})$ es el mayor de $d_1(\mathbf{x}), \ldots, d_m(\mathbf{x})$".

e) Escribe la (simplificada) expresión de $d_i(\mathbf{x})$ en el caso en el que $\Sigma_1 = \cdots = \Sigma_m = \Sigma$.