

Análisis Multivariado 2 - Práctica 3

Discriminación

1. Sea $\mathbf{x} \sim Bi(n, \theta_i)$ en G_i , con $i = 1, 2$. Encontrar la regla óptima de clasificación y mostrar que puede llevarse a una función discriminante lineal.
2. Supongamos que $\mathbf{x} \sim Exp(\lambda_i)$ en G_i , con $i = 1, 2$.
 - (a) Encontrar la regla óptima de clasificación y expresarla como una función discriminante lineal.
 - (b) Calcular la probabilidad total de mala clasificación $P(\mathcal{R}, \mathbf{f})$ cuando $\pi_1 = \pi_2$.
 - (c) Tomemos $\lambda_1 = 1$ y $\lambda_2 = \lambda > 1$. Estudiar el límite de $P(\mathcal{R}, \mathbf{f})$ cuando $\lambda \rightarrow \infty$. Sacar conclusiones.
3. Una regla de clasificación es *minimax* si las regiones que definen el criterio de clasificación se buscan de modo que minimicen $\max\{P(1 | 2), P(2 | 1)\}$.

- (a) Dado $\alpha \in (0, 1)$, verificar que

$$\max\{P(1 | 2), P(2 | 1)\} \geq (1 - \alpha)P(1 | 2) + \alpha P(2 | 1)$$

- (b) Para cada α , encontrar la regla que minimiza el lado derecho de la ecuación anterior.
- (c) Probar que la regla minimax está dada por

$$R_1 = \left\{ x : \frac{f_1(x)}{f_2(x)} > c \right\}$$

donde c satisface que $P(1 | 2) = P(2 | 1)$.

4. Sea $\mathbf{x}$ un vector aleatorio, y sean $\mu_1 = E(\mathbf{x} | G_1)$, $\mu_2 = E(\mathbf{x} | G_2)$. Supongamos que la matriz de covarianza $\Sigma = E((\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)')$ ($i = 1, 2$) es la misma para ambas poblaciones, que los costos son iguales y que las π_i también lo son.
 - (a) Si B se define como $B = c(\mu_1 - \mu_2)(\mu_1 - \mu_2)'$ para alguna constante c , verificar que $\mathbf{e} = c\Sigma^{-1}(\mu_1 - \mu_2)$ es de hecho un autovector (sin escalar) de $\Sigma^{-1}B$.
 - (b) Deducir que cuando $k = 2$, el criterio de clasificación de Fisher (es decir el análisis discriminante) coincide con la regla óptima para la normal.
5. Supongamos que $\mathbf{x} \in \mathbb{R}^p$ un vector aleatorio con distribución normal en ambas poblaciones. Sean $\mu_1 = E(\mathbf{x} | G_1)$, $\mu_2 = E(\mathbf{x} | G_2)$. Supongamos que la matriz de covarianza $\Sigma = E((\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)')$ ($i = 1, 2$) es la misma para ambas poblaciones, que los costos son iguales y que las π_i también lo son. Más aún supongamos que

$$\Sigma = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix} = (1 - \rho)\mathbb{I}_p + \rho \mathbf{1}_p \mathbf{1}_p'$$

Sea $\delta = \mu_1 - \mu_2$. Muestre que la regla discriminante lineal puede escribirse como

$$L(\mathbf{x}) = \frac{1}{p(1-\rho)} \mathbf{1}_p' \delta \left[\mathbf{h}' \mathbf{x} + \frac{1-\rho}{1+(p-1)\rho} \mathbf{1}_p' \mathbf{x} \right]$$

donde

$$\mathbf{h} = \frac{p}{\mathbf{1}_p' \delta} \delta - \mathbf{1}_p$$

Luego $L(\mathbf{x})$ depende de dos factores $\mathbf{1}_p' \mathbf{x} = \sum_{j=1}^p x_j$ que se llama factor de tamaño y $\mathbf{h}' \mathbf{x}$ que Penrose llama factor de forma.

6. Sea el siguiente modelo de regresión: $z = b' \mathbf{x} + \mathbf{e}$, $\mathbf{e}$ con distribución logística. Supongamos que sólo observamos si $z > 0$ ó $z < 0$, (no su valor). Mostrar que estimar a b coincide con encontrar el estimador de los parámetros del modelo de clasificación basada en el modelo de regresión logística en el caso $k = 2$.
7. Derivar las ecuaciones de máxima verosimilitud para el vector de parámetros del modelo logístico multivariado con k grupos.
8. Consideremos los datos “iris” del R. Es un conjunto de datos analizados por Fisher que consisten en 4 mediciones realizadas en 50 flores iris de cada una de 3 especies distintas (Setosa, Versicolor y Virginica). Las 4 variables, medidas en centímetros, son

X_1 = Longitud de los sépalos (sepal length)
 X_2 = Ancho de los sépalos (sepal width)
 X_3 = Longitud de los pétalos (petal length)
 X_4 = Ancho de los pétalos (petal width)

- (a) Realizar un análisis discriminante y un scatterplot de las primeras 2 coordenadas discriminantes.
 - (b) Suponiendo normalidad en los datos, hacer una clasificación lineal y calcular el error de clasificación por los 2 métodos:
 - i. “Ingenuo” (calcular la proporción de datos mal clasificados) o error aparente.
 - ii. Validación cruzada.
 - (c) Idem b) pero con la clasificación cuadrática y comparar los resultados.
9. Del conjunto de datos “iris” consideremos las variables X_2 = Ancho de los sépalos y X_4 = Ancho de los pétalos para las 3 especies de flores.
 - (a) Graficar los pares de datos (X_2, X_4) en el plano. Para cada especie, estos datos ¿tienen aspecto de provenir de una distribución normal bivariada?
 - (b) Asumiendo que las muestras provienen de poblaciones con distribución normal bivariada con matriz de covarianza común Σ , testear a nivel $\alpha = 0.05$, la hipótesis $H_0 : \mu_1 = \mu_2 = \mu_3$, versus H_1 : al menos una de las μ_i es distinta de las otras. ¿Es razonable el supuesto de igualdad de matrices de covarianza en este caso?

- (c) Suponiendo que la distribución es normal bivariada para cada población, construir la regla de clasificación cuadrática, asumiendo costos de mala clasificación iguales y probabilidad a priori de pertenecer a cada grupo iguales. Usando esta regla de clasificación recién construida clasificar la nueva observación $\mathbf{x}_0 = (3.5, 1.75)'$ como perteneciente a alguno de los 3 grupos.
 - (d) Supongamos que las matrices de covarianza Σ_i son las mismas para las 3 poblaciones normales bivariadas. Construir la regla de clasificación lineal, asumiendo costos de mala clasificación iguales y probabilidad a priori de pertenecer a cada grupo iguales, y usarla para clasificar la nueva observación $\mathbf{x}_0 = (3.5, 1.75)'$ como perteneciente a alguno de los 3 grupos. Comparar los resultados obtenidos en b) y c). ¿Cuál enfoque es preferible en este caso?
 - (e) Graficar en el scatterplot realizado en a) las regiones halladas en d).
 - (f) Usando la clasificación lineal realizada en d), clasificar las observaciones de la muestra. Calcular el error aparente total y la estimación insesgada del error que se obtiene por validación cruzada.
10. Aproximadamente 2 años antes de la bancarrota de algunas empresas se recolectan datos financieros de las mismas, y también se recolectan datos de empresas sanas financieramente alrededor del mismo momento. A continuación figuran las 4 variables correspondientes a los datos que se encuentran en el archivo *finanzas*:

$$\begin{aligned}
X_1 &= (\text{flujo de caja})/(\text{deuda total}) \\
X_2 &= (\text{ingreso neto})/(\text{total de activos}) \\
X_3 &= (\text{activos corrientes})/(\text{pasivos corrientes}) \\
X_4 &= (\text{activos corrientes})/(\text{ventas netas})
\end{aligned}$$

Grupo 1: Empresas en bancarrota

Grupo 2: Empresas sanas financieramente

- (a) Graficar los datos para los pares de observaciones (X_1, X_2) , (X_1, X_3) y (X_1, X_4) . Para alguno de estos pares de variables, ¿tienen aspecto de provenir de una distribución normal bivariada?
- (b) Usando los $n_1 = 21$ pares de observaciones (X_1, X_2) de empresas en bancarrota y los $n_2 = 25$ pares de observaciones (X_1, X_2) de empresas sanas financieramente, calcular los vectores de medias muestrales $\bar{X}_1$ y $\bar{X}_2$ y las matrices de covarianza muestrales S_1 y S_2 .
- (c) Usando los resultados de b) y asumiendo que las dos muestras aleatorias provienen de dos poblaciones normales, construir la regla de clasificación cuadrática asumiendo $\pi_1 = \pi_2$ y $c(1 | 2) = c(2 | 1)$.
- (d) Evaluar la performance de la regla de clasificación desarrollada en c) calculando el error aparente total y la estimación del error actual esperado que se obtiene por validación cruzada.
- (e) Repetir los items c) y d) tomando $\pi_1 = 0.05$ y $\pi_2 = 0.95$ y $c(1 | 2) = c(2 | 1)$. ¿Es razonable esta elección de probabilidades a priori?

- (f) Usando los resultados de b), construir la matriz de covarianza ponderada y realizar el análisis de coordenadas discriminantes. Usar esta función para clasificar las observaciones muestrales y evaluar el error aparente. ¿Es apropiada la elección del método de coordenadas discriminantes para clasificar las observaciones en este caso?
 - (g) Repetir los items b) a e) usando las variables (X_1, X_3) y (X_1, X_4) . ¿Parecen ser algunas variables mejores clasificadoras que otras?
 - (h) Repetir los items b) a e) usando las 4 variables.
11. En el archivo *microtus* de la práctica 1 se encuentran datos correspondientes a 8 variables medidas en dos tipos de ratas, las *multiplex* y las *subterráneas*.
- (a) Suponiendo normalidad en los datos, hacer una clasificación lineal y calcular el error de clasificación por los 2 métodos:
 - i. “Ingenuo” (calcular la proporción de datos mal clasificados) o error aparente.
 - ii. Validación cruzada.
 - (b) Idem a) pero con la clasificación cuadrática y comparar los resultados.
 - (c) ¿Qué observa? ¿Qué conclusión saca?
12. En el archivo *apples* se encuentran datos correspondientes a árboles de manzanas con 6 diferentes injertos. En cada uno de los 6 tipos de injertos hay 8 árboles de manzanas.

Las 4 mediciones corresponden a:

Y_1 : circunferencia del tronco a los 4 años en unidades de 10 cm

Y_2 : altura a los 4 años en metros

Y_3 : circunferencia del tronco a los 15 años en unidades de 10 cm

Y_4 : peso del árbol sobre la tierra a los 15 años, en unidades de 1000 libras

Utilice solamente los datos correspondientes a los tipos de injerto 1, 3 y 5

- (a) Suponiendo normalidad en los datos, hacer una clasificación lineal y calcular el error de clasificación por los 2 métodos:
 - i. “Ingenuo” (calcular la proporción de datos mal clasificados) o error aparente.
 - ii. Validación cruzada.
 - (b) Idem a) pero con la clasificación cuadrática y comparar los resultados.
 - (c) Hacer la clasificación con el método de vecinos cercanos.
13. Con los datos del archivo *mundodes*, utilizados en la práctica 1, se quiere hacer una clasificación de acuerdo al producto bruto nacional per cápita. Se considera un primer grupo de países con producto bruto bajo a aquellos cuyo PNB es menor a 2000, un segundo grupo de países con producto bruto medio a aquellos cuyo PNB está entre 2000 y 10000 y un tercer grupo de países con alto producto bruto a aquellos cuyo PNB es mayor a 10000.
- (a) Suponiendo normalidad en los datos, hacer una clasificación lineal y calcular el error de clasificación por los 2 métodos:

- i. “Ingenuo” (calcular la proporción de datos mal clasificados) o error aparente.
 - ii. Validación cruzada.
- (b) Idem a) pero con la clasificación cuadrática y comparar los resultados.
- (c) Hacer la clasificación con el método de vecinos cercanos.
- (d) Hacer la clasificación usando regresión logística
- (e) ¿Qué método recomienda basándose en los errores de clasificación?