Técnicas de Análisis Multivariado 2- Práctica 4

1. 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 $G^\star$ con regiones de clasificación $\{\mathcal{G}_1, \mathcal{G}_2\}$ y expresarla como una función discriminante lineal.

   b) Calcular la probabilidad total de mala clasificación $P(G^\star) = \mathbb{P}(G^\star \neq G)$ cuando $\pi_1 = \pi_2 = 1/2$.

   c) Tomemos $\lambda_1 = 1$ y $\lambda_2 = \lambda > 1$. Estudiar el límite de $P(G^\star)$ cuando $\lambda \to \infty$. 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

   $$G_1^\star = \left\{ x : \frac{f_1(x)}{f_2(x)} > c \right\}$$

   donde c satisface que $p_{1|2} = p_{2|1}$.

4. Sea $\mathbf{x}$ un vector aleatorio, y sean $\boldsymbol{\mu}_1 = \mathbb{E}(\mathbf{x} \mid G_1)$, $\boldsymbol{\mu}_2 = \mathbb{E}(\mathbf{x} \mid G_2)$. Supongamos que la matriz de covarianza $\boldsymbol{\Sigma} = \mathbb{E}\left((\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}}\right)$, $i = 1, 2$ es la misma para ambas poblaciones, que los costos son iguales y que $\pi_1 = \pi_2 = 1/2$.

   a) Si $\mathbf{B}$ se define como $\mathbf{B} = c(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}$, para alguna constante c, verificar que $\mathbf{u} = c\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ es de hecho un autovector (sin escalar) de $\boldsymbol{\Sigma}^{-1}\mathbf{B}$.

   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 $\boldsymbol{\mu}_1 = \mathbb{E}(\mathbf{x} \mid G_1)$, $\boldsymbol{\mu}_2 = \mathbb{E}(\mathbf{x} \mid G_2)$. Supongamos que la matriz de covarianza $\boldsymbol{\Sigma} = \mathbb{E}\left((\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}}\right)$, $i = 1, 2$ es la misma para ambas poblaciones, que los costos son iguales y que $\pi_1 = \pi_2 = 1/2$. Más aún supongamos que

   $$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix} = (1 - \rho)\mathbf{I}_p + \rho\, \mathbf{1}_p\, \mathbf{1}_p^{\mathrm{T}}$$

Sea $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Muestre que la regla discriminante lineal puede escribirse como

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

donde

$$\mathbf{h} = \frac{p}{\mathbf{1}_p^{\mathrm{T}} \boldsymbol{\delta}} \, \boldsymbol{\delta} - \mathbf{1}_p \, .$$

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

6. Un test para el diagnóstico de gota se basa en el nivel de ácido úrico. Usando unidades apropiadas el nivel de ácido úrico tiene distribución $N(5,1)$ entre individuos sanos y distribución $N(8{,}5,1)$ entre personas con gota.

   a) Supongamos que el 1 % de la población tiene gota.

      1) Como clasifica a un paciente con nivel de ácido úrico igual a x?
      2) Calcule la probabilidad total de mala clasificación e_{opt} y las probabilidades a posteriori $q_i(x)$, $i = 1, 2$ cuando $x = 7$.

   b) Obtenga la regla minimax.

      1) Como clasifica en este caso a x.
      2) Calcule la probabilidad total de mala clasificación.

   c) Qué procedimiento utilizaría si se quiere que la probabilidad de diagnosticar como sana a una persona con gota no sea mayor que 0.1? Calcule la probabilidad de mala clasificación en este caso.

   d) Supongamos ahora que el nivel de ácido úrico tiene distribución $N(8{,}5,2)$ entre personas con gota.

      1) Como clasifica a un paciente con nivel de ácido úrico igual a x?
      2) Calcule la probabilidad total de mala clasificación e_{opt}
      3) Cómo clasifica ahora a $x = 7$?

7. Sea el siguiente modelo de regresión: $z = \mathbf{b}^{\mathrm{T}} \mathbf{x} + \mathbf{e}$, $\mathbf{e}$ con distribución logística. Supongamos que sólo observamos si $z > 0$ ó $z < 0$, (no su valor). Mostrar que el estimador de $\mathbf{b}$ de máxima verosimilitud coincide con encontrar el estimador de los parámetros del modelo de clasificación basada en el modelo de regresión logística en el caso $k = 2$.

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:

1) "Ingenuo" (calcular la proporción de datos mal clasificados) o error aparente.
2) 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 $\boldsymbol{\Sigma}$, testear a nivel $\alpha = 0{,}05$, la hipótesis

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 \quad \text{versus} \quad H_1 : \text{al menos una de las } \boldsymbol{\mu}_i \text{ es distinta de las otras}\,.$$

¿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 $\mathbf{x}_0 = (3{,}5, 1{,}75)^{\mathrm{T}}$ como perteneciente a alguno de los 3 grupos.

d) Supongamos que las matrices de covarianza $\boldsymbol{\Sigma}_i$ 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 $\mathbf{x}_0 = (3{,}5, 1{,}75)^{\mathrm{T}}$ 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.txt`:

X_1 = (flujo de caja)/(deuda total)
X_2 = (ingreso neto)/(total de activos)
X_3 = (activos corrientes)/(pasivos corrientes)
X_4 = (activos corrientes)/(ventas netas)

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 $\mathbf{x}_{1,j} = (X_{1,j}, X_{2,j})$, $1 \leq j \leq n_1$, de empresas en bancarrota y los $n_2 = 25$ pares de observaciones $\mathbf{x}_{2,j} = (X_{1,j}, X_{2,j})$, $1 \leq j \leq n_2$, de empresas sanas financieramente, calcular los vectores de medias muestrales $\overline{\mathbf{x}}_1$ y $\overline{\mathbf{x}}_2$ y las matrices de covarianza muestrales $\mathbf{S}_1$ y $\mathbf{S}_2$.

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 ahora las variables (X_1, X_3) y luego las variables (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 conjunto de datos del archivo `microtus-data.txt` 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:

        1) "Ingenuo" (calcular la proporción de datos mal clasificados) o error aparente.
        2) Validación cruzada.

    b) Idem a) pero con la clasificación cuadrática y comparar los resultados.

    c) Qué observa? Qué conclusión saca?

12. El conjunto de datos `arboles manzana.txt` 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:

1) "Ingenuo" (calcular la proporción de datos mal clasificados) o error aparente.
2) 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 `Datos Paises.xlx`, 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:

1) "Ingenuo" (calcular la proporción de datos mal clasificados) o error aparente.
2) 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?