Inspira Crea Transforma

ANÁLISIS DE CLASIFICACIÓN

Presentado por:
Mateo Restrepo S.
Juan S. Cárdenas R.
David Plazas E.

Prof.:
Francisco I. Zuluaga D.

Estadística II
Universidad EAFIT
2019

CONTENIDO

1. INTRODUCCIÓN

- Ingeniería y computación $\rightarrow$ *Reconocimiento de patrones.*
- *Análisis de Clusters.*
- Clasificar nuevas observaciones en grupos ya establecidos.

Se tienen k grupos, de donde se extraen muestras para obtener $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \ldots, \bar{\mathbf{y}}_k$. Se pretender acomodar una observación $\mathbf{y}$ en alguno de los k grupos; una aproximación intuitiva es compararla con las medias $\bar{\mathbf{y}}_i$, mirando cuál es la más cercana.

Ejemplos:

- Clasificación de aplicantes a una Universidad, entre los que desertarán y los que no.
- Orientación vocacional basada en tests de aptitudes.
- Clasificación de ciudades violentas (estudio previo de indicadores).

2. CLASIFICACIÓN EN DOS GRUPOS

2.1 Método de Fisher

- Se tienen dos grupos G_1 y G_2.
- Se requiere que $\Sigma_1 = \Sigma_2$.
- No requiere que $\mathbf{y}_i \sim N_p(\boldsymbol{\mu}, \Sigma)$.
- Se calcula $\bar{\mathbf{y}}_1$, $\bar{\mathbf{y}}_2$.
- Se basa en la función discriminante:

$$z = \mathbf{a}'\mathbf{y} = (\bar{\mathbf{y}}_\mathbf{1} - \bar{\mathbf{y}}_\mathbf{2})' \, \mathbf{S}_{\text{pl}}^{-1}\mathbf{y} \tag{1}$$

- Se calculan $\bar{\mathbf{z}}_1$ y $\bar{\mathbf{z}}_2$ para determinar si $\mathbf{y} \in G_1$ o $\mathbf{y} \in G_2$.

$$\left[z > \frac{1}{2}(\bar{\mathbf{z}}_1 + \bar{\mathbf{z}}_2)\right] \implies \mathbf{y} \in G_1$$
$$\left[z < \frac{1}{2}(\bar{\mathbf{z}}_1 + \bar{\mathbf{z}}_2)\right] \implies \mathbf{y} \in G_2$$

2. CLASIFICACIÓN EN DOS GRUPOS

2.2 Regla Óptima

Sean dos grupos G_1 y G_2 tal que $\Sigma_1 = \Sigma_2 = \Sigma$. Si se conocen las probabilidades p_1 y p_2 asociadas a las poblaciones y las respectivas funciones de densidad $f_{G_1}(\mathbf{y})$ y $f_{G_2}(\mathbf{y})$, se puede aprovechar esta información para una mejor clasificación.

- Minimizar el error de clasificación.
- El criterio de asignación óptima de $\mathbf{y}$ en G_1 es:

$$p_1 f_{G_1}(\mathbf{y}) > p_2 f_{G_2}(\mathbf{y}) \tag{2}$$

- Si se cumple que $f_{G_1}(\mathbf{y}) = N_p(\boldsymbol{\mu}_1, \Sigma)$ y $f_{G_2}(\mathbf{y}) = N_p(\boldsymbol{\mu}_2, \Sigma)$, entonces la regla se transforma en:

$$\left[z > \frac{1}{2}(\bar{\mathbf{z}}_1 + \bar{\mathbf{z}}_2) + \ln\left(\frac{p_2}{p_1}\right)\right] \implies \mathbf{y} \in G_1$$
$$\left[z < \frac{1}{2}(\bar{\mathbf{z}}_1 + \bar{\mathbf{z}}_2) + \ln\left(\frac{p_2}{p_1}\right)\right] \implies \mathbf{y} \in G_2$$

- Criterio asintóticamente óptimo.
- Si $p_1 = p_2 \rightarrow$ Fisher.

3. CLASIFICACIÓN EN VARIOS GRUPOS

3.1 Grupos con Igual Covarianza Poblacional

- Se tienen k grupos tales que $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$, con vectores de medias $\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_k$.
- Se utiliza una función de distancia para encontrar el vector de medias más cercano y asignarlo a su población.
- Se estima la matriz Σ con:

$$\mathbf{S}_{\text{pl}} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\mathbf{S}_i \tag{3}$$

- Utiliza la función lineal de clasificación:

$$L_i(\mathbf{y}) = \bar{\mathbf{y}}_i'\mathbf{S}_{\text{pl}}^{-1}\left(\mathbf{y} - \frac{1}{2}\bar{\mathbf{y}}_i\right) \tag{4}$$

- Se asigna $\mathbf{y}$ al grupo G_j tal que j cumple

$$L_j = \max_i\{L_i(\mathbf{y})\} \tag{5}$$

3. CLASIFICACIÓN EN VARIOS GRUPOS

3.1 Grupos con Igual Covarianza Poblacional

De igual forma que en clasificación de dos grupos, si se conocen las probabilidades p_i y las funciones de densidad $f_{G_i}(\mathbf{y})$ para cada uno de los k grupos, se tiene la siguiente regla óptima:

Asigne $\mathbf{y}$ al grupo en el cual $p_i\, f_{G_i}(\mathbf{y})$ es máximo.

Este criterio minimiza el error de asignación. Asumiendo normalidad, $f_{G_i}(\mathbf{y}) = N_p(\boldsymbol{\mu}_i, \Sigma)$, la función lineal de clasificación se transforma en:

$$L_i^*(\mathbf{y}) = \ln p_i + L_i(\mathbf{y}) \tag{6}$$

Si $p_1 = p_2 = \cdots = p_k$

$$\max_i\{L_i(\mathbf{y})\} \equiv \max_i\{L_i^*(\mathbf{y})\} \tag{7}$$

3.2 Grupos con Diferente Covarianza Poblacional

- En general es difícil que $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$.
- En este caso, no es posible reducir a una función lineal de clasificación $\rightarrow$ función de clasificación cuadrática.
- Este método sí requiere normalidad.
- Función de clasificación cuadrática:

$$Q_i(\mathbf{y}) = \ln p_i - \frac{1}{2}\ln|\mathbf{S}_i| - \frac{1}{2}(\mathbf{y} - \bar{\mathbf{y}}_i)'\mathbf{S}_i^{-1}(\mathbf{y} - \bar{\mathbf{y}}_i) \tag{8}$$

- Criterio de clasificación:

  Asigne $\mathbf{y}$ al grupo que proporcione el máximo $Q_i(\mathbf{y})$

- Este método requiere que $(\forall i = 1, \ldots, k)(n_i > p) \rightarrow$ existencia $\mathbf{S}_i^{-1}$.

4. ERROR DE ESTIMACIÓN

4.1 Matriz de Confusión

- El error de estimación ϵ se define como la probabilidad que una función de clasificación se equivoque asignando **y**. Este se puede estimar con la matriz de confusión.
- La matriz de confusión M, es una matriz de orden $(k \times k)$. Esta se calcula usando la observación de cada uno de los grupos utilizados y la función de clasificación, contando a qué grupo asigna las observaciones.
- Si n_{ij} es el numero de veces que la función de clasificación asignó una observación del grupo i al grupo j y N_i es la cantidad de observaciones hechas en el grupo i, entonces el error de estimación se calcula como:

$$\epsilon = 1 - \frac{\sum_{i=1}^{k} n_{ii}}{\sum_{i=1}^{k} N_i} = 1 - \frac{\mathrm{tr}(M)}{\sum_{i=1}^{k} N_i} \tag{9}$$

5. EJEMPLOS

Clasificación entre hombres y mujeres midiendo algunos rasgos psicológicos.

Se tienen 32 observaciones de cuatro diferentes factores psicológicos medidos para hombres y mujeres. Diga si la siguiente observación pertenece al grupo masculino o femenino:

$$\mathbf{y} = \begin{bmatrix} 11 \\ 17 \\ 15 \\ 23 \end{bmatrix} \tag{10}$$

Implementación en R.

Medidas de la cabeza de jugadores de fútbol americano.

Se tienen 30 medidas de jugadores de fútbol americano de bachillerato, de universidad y de personas que no juegan fútbol americano sobre 6 aspectos de la cabeza tales como ancho, longitud y más. Construya una función que clasifique a que grupo pertenece una observación nueva y halle su matriz de clasificación.
Implementación en R.

REFERENCIAS

- A. C. Rencher, *Methods of multivariate analysis*. John Wiley & Sons, 2003, vol. 492.

Gracias