Técnicas multivariadas

Análisis discriminante

Mario J. P. López

Programa de Estadística
Universidad El Bosque

2020

Contexto

Un conjunto de observaciones multivariadas pertenecen a un conjunto de grupos (clases o categorías), mutuamente excluyentes, conocidos a priori.

Objetivos

- Estimación de la probabilidad de pertenecer a un grupo determinado en función de una o múltiples variables predictoras.
- Creación de una regla de discriminación/clasificación de los grupos conocidos a priori.
- Asignación de nuevos individuos a los grupos usando una regla de discriminación/clasificación

Métodos

los más comunes:

- Análisis discriminante lineal (ADL)
  Linear discriminant analysis (LDA)
- Análisis discriminante cuadrático (ADC)
  Quadratic discriminant analysis (QDA)
- Análisis discriminante por mixturas (ADM)
  Mixture discriminant analysis (MDA)
- Análisis discriminante flexible (ADF)
  Flexible Discriminant Analysis (FDA)
- Análisis discriminante regularizado (ADR)
  Regularized discriminant anlysis (RDA)

Análisis discriminante lineal (ADL)

- Generalización del análisis discriminante lineal de Fisher
- Busca una combinación lineal de variables que separan dos o más clases de individuos
- La combinación resultante se puede usar como un clasificador lineal
- Dentro de los grupos, las variables predictoras deben ser continuas y con distribución normal multivariada

ADL para dos grupos

Considere el v.a. $\boldsymbol{x}$ p-variado y la v.a. y, tal que

$$(\boldsymbol{x} \mid y = 1) \sim N_p(\mu_1, \Sigma_1)$$
$$(\boldsymbol{x} \mid y = 2) \sim N_p(\mu_2, \Sigma_2)$$

El criterio de discriminación lineal para el grupo $y = 1$ se basa en

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \left(\frac{C(1 \mid 2)}{C(2 \mid 1)} \frac{\pi_2}{\pi_1}\right)$$

donde $C(i \mid j)$ es el costo de clasificación al grupo i cuando realmente pertenece al grupo j y π_i es la probabilidad a priori de pertenencia al grupo i.

ADL para dos grupos

Asumiendo $C(1 \mid 2) = C(2 \mid 1)$ y $\pi_1 = \pi_2$, el criterio se transforma en

$$(\boldsymbol{x} - \mu_2)^\top \Sigma_2^{-1} (\boldsymbol{x} - \mu_2) + \ln \mid \Sigma_2 \mid > (\boldsymbol{x} - \mu_1)^\top \Sigma_1^{-1} (\boldsymbol{x} - \mu_1) + \ln \mid \Sigma_1 \mid$$

Si $\Sigma_1 = \Sigma_2 = \Sigma$, el criterio se reduce a

$$(\boldsymbol{x} - \mu_2)^\top \Sigma^{-1} (\boldsymbol{x} - \mu_2) > (\boldsymbol{x} - \mu_1)^\top \Sigma^{-1} (\boldsymbol{x} - \mu_1)$$

o de forma equivalente

$$(\mu_1 - \mu_2)^\top \Sigma^{-1} \boldsymbol{x} > \frac{1}{2} \left(\mu_1^\top \Sigma^{-1} \mu_1 - \mu_2^\top \Sigma^{-1} \mu_2 \right)$$

$$\boldsymbol{a}^\top \boldsymbol{x} > \boldsymbol{c}$$

ADL muestral

Si $\hat{\mu}_1 = \bar{\mathbf{x}}_1$, $\hat{\mu}_2 = \bar{\mathbf{x}}_2$ y

$$\hat{\Sigma} = S_p = \frac{(n_1 - 1)\, S_1 + (n_2 - 1)\, S_2}{n_1 + n_2 - 2}$$

la regla de discriminación consiste en

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top S_p^{-1} \mathbf{x} > \frac{1}{2} \left(\bar{\mathbf{x}}_1^\top S_p^{-1} \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^\top S_p^{-1} \bar{\mathbf{x}}_2 \right)$$

$$\hat{\mathbf{a}}^\top \mathbf{x} - \hat{c} > 0$$

ADL para más de dos grupos

La regla de discriminación para asignar un individuo con un vector de variables predictivas $\boldsymbol{x}$ al grupo i cuando se tienen G grupos a priori es

$$\left(\mu_i - \mu_g\right)^\top \Sigma^{-1} \boldsymbol{x} > \frac{1}{2}\left(\mu_i^\top \Sigma^{-1} \mu_i - \mu_g^\top \Sigma^{-1} \mu_g\right)$$

$$\boldsymbol{a}^\top \boldsymbol{x} - \boldsymbol{c} > 0$$

para todo $g \neq i$, $g = 1, \ldots, G$.

Para el caso muestral basta con calcular

$$\hat{\Sigma} = S_p = \frac{\sum_{g=1}^{G} \left(n_g - 1\right) S_g}{\sum_{g=1}^{G} \left(n_g - 1\right)}$$

Análisis discriminante cuadrático (ADC)

Si $\Sigma_1 \neq \Sigma_2$, el criterio de discriminación es

$$(\boldsymbol{x} - \mu_2)^\top \Sigma_2^{-1} (\boldsymbol{x} - \mu_2) + \ln \mid \Sigma_2 \mid > (\boldsymbol{x} - \mu_1)^\top \Sigma_1^{-1} (\boldsymbol{x} - \mu_1) + \ln \mid \Sigma_1 \mid$$

el cual puede ser reescrito como

$$\boldsymbol{x}^\top \left(\Sigma_2^{-1} - \Sigma_1^{-1}\right) \boldsymbol{x} + 2\left(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2\right)\boldsymbol{x} + \left(\mu_2^\top \Sigma_2^{-1}\mu_2 - \mu_1^\top \Sigma_1^{-1}\mu_1\right) + \ln\left(\frac{\mid \Sigma_1 \mid}{\mid \Sigma_2 \mid}\right) > 0$$

esto es

$$\boldsymbol{x}^\top A \boldsymbol{x} + 2\boldsymbol{b}^\top \boldsymbol{x} + \boldsymbol{c} > 0$$

ADC muestral

Si $\hat{\mu}_1 = \bar{\boldsymbol{x}}_1$, $\hat{\mu}_2 = \bar{\boldsymbol{x}}_2$, $\hat{\Sigma}_1 = S_1$ y $\hat{\Sigma}_2 = S_2$, la regla de discriminación consiste en

$$\boldsymbol{x}^\top \left(S_2^{-1} - S_1^{-1}\right)\boldsymbol{x} + 2\left(\Sigma_1^{-1}\bar{\boldsymbol{x}}_1 - \Sigma_2^{-1}\bar{\boldsymbol{x}}_2\right)\boldsymbol{x} +$$
$$\left(\bar{\boldsymbol{x}}_2^\top S_2^{-1}\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1^\top S_1^{-1}\bar{\boldsymbol{x}}_1\right) + \ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) > 0$$

esto es

$$\boldsymbol{x}^\top \widehat{A}\boldsymbol{x} + 2\hat{\boldsymbol{b}}^\top \boldsymbol{x} + \hat{c} > 0$$

ADC para más de dos grupos

La regla de discriminación para asignar un individuo con un vector de variables predictivas $\boldsymbol{x}$ al grupo i cuando se tienen G grupos a priori es

$$\boldsymbol{x}^\top\left(\Sigma_g^{-1}-\Sigma_i^{-1}\right)\boldsymbol{x}+2\left(\Sigma_i^{-1}\mu_i-\Sigma_g^{-1}\mu_g\right)\boldsymbol{x}+ \left(\mu_g^\top\Sigma_g^{-1}\mu_g-\mu_i^\top\Sigma_i^{-1}\mu_i\right)+\ln\left(\frac{|\,\Sigma_i\,|}{|\,\Sigma_g\,|}\right)>0$$

esto es

$$\boldsymbol{x}^\top A\boldsymbol{x}+2\boldsymbol{b}^\top\boldsymbol{x}+\boldsymbol{c}>0$$

para todo $g\neq i$, $g=1,\ldots,G$.
Para el caso muestral basta con reemplazar $\hat{\mu}_g=\bar{\boldsymbol{x}}_g$ y $\hat{\Sigma}_g=S_g$, $g=1,\ldots,G$.

Análisis discriminante por mixturas (ADM)

- El ADL asume que cada grupo proviene de una distribución normal multivariada pero esto es muy restrictivo
- En ADM se asume que cada grupo proviene de una mixtura de distribuciones normales
- Se asume igualdad de matrices de covarianza entre clases

Análisis discriminante flexible (ADF)

- Es una extensión del LDA que utiliza combinaciones no lineales, tipo splines, como predictores
- Construye una regla de discriminación sin los supuestos de normalidad multivariada y de relaciones lineales entre las variables dentro de cada grupo

Análisis discriminante regularizado (ADR)

- Construye una regla de discriminación (como en ADC) regularizando las matrices de covarianza de los grupos como

$$\hat{\Sigma}_g = (1-\gamma)\,\hat{\Sigma}_g(\lambda) + \gamma \frac{1}{p} \mathrm{tr}\left(\hat{\Sigma}_g(\lambda)\right) I$$

  donde

$$\hat{\Sigma}_g(\lambda) = (1-\lambda)\, S_g + \lambda S_p$$

- Permite un modelo más robusto en presencia de multicolinealidad en las variables predictoras
- Es un punto intermedio entre ADL y ADC
- Mejora la estimación de las matrices de covarianza en situaciones donde el número de predictores es mayor que el número de individuos.

Referencias

- Johnson, R., Wichern, D.
  Applied Multivariate Statistical Analysis.
  6th ed. Pearson, 2007.
- Mardia, K., Kent, J., Bibby, J.
  Multivariate Analysis.
  Academic Press, 1995.
- Rencher, A.
  Methods of Multivariate Analysis.
  2nd ed. Willey, 2002.