Análisis discriminante

Daniel Czarnievicz

Descripción general

El análisis discriminante es una técnica supervisada con finalidades de descripción (analizar la existencia de diferencias entre grupos), predicción (clasificar nuevas observaciones) y re-clasificación. El problema consiste en construir un modelo que permita discriminar las observaciones según el grupo poblacional al que pertenecen. A la i-ésima observación se le miden p características, las cuales componen el vector $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Se asume que existen k grupos en la población.

Reglas de decisión

Existen distintas reglas de decisión para la asignación de observaciones a grupos.

Minimizar la probabilidad de error

La regla de decisión será aquella que minimize la probabilidad total de error. Supongamos que una población P está sub-dividida en k grupos excluyentes. Llamaremos $f_k(x)$ a la densidad de x, si x pertenece al k-ésimo grupo. El objetivo es encontrar una partición del espacio muestral R, tal que asigne x al grupo $k \Leftrightarrow x \in r_x$.

Llamaremos $\Pr(g'|g)$ al error de clasificar en el grupo g' una observación perteneciente al grupo g. Entonces:

$$\Pr(g'|g) = \int_{R_{g'}} f_g(x)dx$$

Por lo tanto, la probabilidad de clasificar erróneamente a todas las observaciones provenientes del grupo g está dada por:

$$\Pr(g) = \sum_{\substack{g'=1 \\ g' \neq g}}^{k} \Pr(g'|g) = 1 - \Pr(g|g)$$

De esta forma entonces, la probabilidad total de clasificación errónea está dada por:

$$\Pr(R,\, f) = \sum_{g=1}^{k} \pi_g \Pr(g)$$

donde π_g es la probabilidad a priori de que i pertenzca al grupo g.

Principio de máxima verosimilitud

El pricipio de clasificación por máxima verosimilitud consiste en asignar la observación i a la población donde el vector observado $\mathbf{x}_i'$ tenga mayor verosimilitud de ocurrir. Es decir, se asigna i al grupo g, sí y solo si:

$$f(\mathbf{x}_i|g) > f(\mathbf{x}_i|g')\ \ \forall g' \neq g \Leftrightarrow \Pr(\mathbf{x}_i|g) > \Pr(\mathbf{x}_i|g')\ \ \forall g' \neq g \Leftrightarrow \frac{f(\mathbf{x}_i|g)}{f(\mathbf{x}_i|g')} > 1$$

Principio de probabilidad a posteriori

La regla consiste en asignar la observación i a la población con mayor probabilidad a posteriori (la probailidad de que i pertenzca a g, dado $\mathbf{x}_i$). Utilizando el Teorema de Bayes, tenemos que la probabilidad a posteriori está dada por:

$$\Pr(i \in g|\mathbf{x} = \mathbf{x}_i) = \frac{\pi_g \Pr(\mathbf{x}_i|g)}{\Pr(\mathbf{x}_i)} = \frac{\pi_g \Pr(\mathbf{x}_i|g)}{\sum\limits_{g'=1}^{k} \pi_g'\ \Pr(\mathbf{x}_i|g')} = \frac{\pi_g f(\mathbf{x}_i|g)}{\sum\limits_{g'=1}^{k} \pi_{g'}\ f(\mathbf{x}_i|g')}$$

donde π_g es la probailidad previa de que $i \in g$. Salvo que información adicional sugiera lo contrario, π_g se estima como la proporción de observaciones en $\mathbf{X}$ que pertenecen a la clase g. Esto es, $\hat{\pi}_g = n_g/n$, siendo n la cantidad total de observaciones, y n_g la cantidad de observaciones con variable de respuesta igual a la etiqueta de la clase g.

De esta forma, la observación i se asignará al grupo g, sí y solo sí:

$$\Pr(i \in g|\mathbf{x} = \mathbf{x}_i) > \Pr(i \in g'|\mathbf{x} = \mathbf{x}_i)\ \ \forall g' \neq g$$

Normalidad

Si $\mathbf{x}_i \sim \mathrm{N}_p(\boldsymbol{\mu},\, \boldsymbol{\Sigma})$ su función de densidad viene dada por:

$$f(\mathbf{x}_i|g) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_g|^{1/2}}\ \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g)\right\}$$

La densidad puede estimarse utilizando los estimadores MV de $\boldsymbol{\mu}_g$ y $\boldsymbol{\Sigma}_g$, $\bar{\mathbf{x}}_g$ y $\mathbf{S}_g$ respectivamente, para obtener:

$$\hat{f}(\mathbf{x}_i|g) = \frac{1}{(2\pi)^{p/2}|\mathbf{S}_g|^{1/2}}\ \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \bar{\mathbf{x}}_g)'\mathbf{S}_g^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_g)\right\}$$

Si aplicamos el supuesto de normalidad a la probabilidad posteriori, obtenemos que:

$$\Pr(i \in g|\mathbf{x} = \mathbf{x}_i) = \frac{\pi_g \ |\mathbf{\Sigma}_g|^{-1/2} \exp\left\{(-1/2)\, D^2_{ig}\right\}}{\sum\limits_{g'=1}^{k} \pi_{g'} \ |\mathbf{\Sigma}_{g'}|^{-1/2} \exp\left\{(-1/2)\, D^2_{ig'}\right\}}$$

donde D^2_{ig} y $D^2_{ig'}$ son la distancia de Mahalanobis entre la observación i y los grupos g y g' respectivamente. Utilizando los estimadores de $\boldsymbol{\mu}_g$ y $\mathbf{\Sigma}_g$ mencionados anteriormente, obtenemos que:

$$\hat{\Pr}(i \in g|\mathbf{x} = \mathbf{x}_i) = \frac{\hat{\pi}_g \ |\mathbf{S}_g|^{-1/2} \exp\left\{(-1/2)\, \hat{D}^2_{ig}\right\}}{\sum\limits_{g'=1}^{k} \hat{\pi}_{g'} \ |\mathbf{S}_{g'}|^{-1/2} \exp\left\{(-1/2)\, \hat{D}^2_{ig'}\right\}}$$

y la observación i se asignará al grupo g, sí, y solo si se cumple que:

$$\hat{\pi}_g \ |\mathbf{S}_g|^{-1/2} \exp\left\{(-1/2)\, \hat{D}^2_{ig}\right\} > \hat{\pi}_{g'} \ |\mathbf{S}_{g'}|^{-1/2} \exp\left\{(-1/2)\, \hat{D}^2_{ig'}\right\} \quad \forall g' \neq g$$

Costos

Existen situaciones en las que el error de clasificación es más costoso para algunos grupos que para otros. La regla de decisión puede modificarse de forma tal de contemplar estas situaciones de la siguiente forma. Se define un costo para cada error de clasificación, $c(g|g')$. Luego, se asigna i al grupo g sí, y solo si, se cumple que:

$$\frac{f(\mathbf{x}_i|g)}{f(\mathbf{x}_i|g')} > \frac{\pi_{g'}\, c(g|g')}{\pi_g\, c(g'|g)} \quad \forall g' \neq g$$

Errores de clasificación

Tasa de error aparente

Luego de elegida una regla de clasificación, se utilizan los n datos para construir la función discriminante y clasificar las observaciones. Una vez clasificadas, se calcula la *tasa de error aparente*,

$$e_{i,\,app} = \frac{m_i}{n_i}$$

donde m_i es la cantidad de observaciones clasificadas erróneamente, de las n_i observaciones asignadas al grupo g_i.

LOOCV

Leave-one-out cross-validation consiste en:

- apartar una observación de la muestra.
- construir la función discriminante con las $n-1$ observaciones restantes.
- clasificar la observación apartada y registrar si dicha observación fue correcta o incorrectamente clasificada.
- repetir para cada una de las n observaciones.

La proporción de observaciones mal clasificadas dentro de cada grupo se define como:

$$e_{i,\,c} = \frac{a_i}{n_i}$$

Funciones discriminantes

Existen distintas formas de construir una función discriminante. El AD busca:

- examinar la separación entre grupos.
- encontrar el subconjunto de las variables originales que separa los grupos tan bien como el conjunto original.
- determinar cuál variable es la que tiene mayor constribución a la discriminación.
- interpretar las nuevas dimensiones representadas por las funciones discriminantes.
- re-clasificar individuos.
- predecir (asignar nuevos individuos a un grupo).

AD factorial

El AD factorial consiste en encontrar las combinaciones lineales de los datos, $\mathbf{Z} = \mathbf{Xu}$, que tengan mayor poder discriminante para clasificar las observaciones en k grupos. Las nuevas variables, susceptibles de separar lo máximo posible los k grupos, representan un compromiso entre mínima inercia intra-clase (grupos homogéneos), y máxima incercia inter-clase (grupos separados). Por lo tanto, el objetivo es encontrar las variables que maximizen el cociente entre ambas inercias:

$$\frac{\mathbf{u'Bu}}{\mathbf{u'Wu}}$$

donde $\mathbf{u}$ son los ejes de inercia que maximizan dicho conciente, y $\mathbf{Z}$ son las coordenadas de los individuos en las nuevas variables (es decir, la proyección de $\mathbf{X}$ en los ejes de inercia). Por

lo tanto, el problema a resolver es:

$$\max \left\{ \frac{\mathbf{u'Bu}}{\mathbf{u'Wu}} \right\}$$

Condiciones de primer orden (CPO):

$$\frac{\partial(\cdot)}{\partial \mathbf{u'}} = \mathbf{0} \Rightarrow \frac{2\mathbf{Bu}(\mathbf{u'Wu}) - 2(\mathbf{u'Bu})\mathbf{Wu}}{(\mathbf{u'Wu})^2} = \mathbf{0}$$

Dado que $\mathbf{X}$ es una matriz de $n \times p$, $\mathbf{W}$ y $\mathbf{B}$ son matrices de $n \times n$, mientras que $\mathbf{u}$ es un vector de forma $n \times 1$, por lo que $\mathbf{u'}$ tiene forma $1 \times n$, tenemos entonces que $\mathbf{u'Bu}$ y $\mathbf{u'Wu}$ son escalares. Por lo tanto, las siguientes manipulaciones son válidas:

$$\frac{2\mathbf{Bu}(\mathbf{u'Wu}) - 2(\mathbf{u'Bu})\mathbf{Wu}}{(\mathbf{u'Wu})^2} = \mathbf{0} \Rightarrow \frac{\mathbf{Bu}(\mathbf{u'Wu})}{(\mathbf{u'Wu})^2} = \frac{(\mathbf{u'Bu})\mathbf{Wu}}{(\mathbf{u'Wu})^2} \Rightarrow$$

$$\Rightarrow \mathbf{Bu} = \mathbf{Wu}\,\frac{\mathbf{u'Bu}}{\mathbf{u'Wu}} \Rightarrow \mathbf{W}^{-1}\mathbf{Bu} = \mathbf{u}\,\frac{\mathbf{u'Bu}}{\mathbf{u'Wu}}$$

donde el último paso lo podemos hacer dado que sabemos que $\mathbf{W}$ es invertible. Luego, si definimos $\lambda = \dfrac{\mathbf{u'Bu}}{\mathbf{u'Wu}}$ obtenemos que:

$$\mathbf{W}^{-1}\mathbf{Bu} = \lambda\,\mathbf{u}$$

Hallamos entonces que $\mathbf{u}$ es el vector propio asociado al máximo valor propio de la matriz $\mathbf{W}^{-1}\mathbf{B}$, mientras que el valor propio λ representa la máxima varianza inter-clases $\mathbf{u'Bu}$ de la nueva variable Z.

En total, pueden hallarse $r = \min(k-1,\, p)$ valores y vectores propios no nulos. Llamamos $\lambda_1,\, \lambda_2,\, \ldots,\, \lambda_r$ a los valores propios, y $\mathbf{u}_1,\, \mathbf{u}_2, \ldots,\, \mathbf{u}_r$ a los vectores propios asociados, tales que $\lambda_1 > \lambda_2 > \ldots > \lambda_r$. Las variables $\mathbf{Z}_j = \mathbf{X}\mathbf{u}_j$ proporcionan la máxima separación para discriminar entre los k grupos. Estas variables son incorreladas, dado que se construyen de forma secuencial y de manera ortogonal. Es decir, $\mathbf{u}_1$ es tal que la proyección de los grupos sobre si misma tiene máxima separación relativa. La segunda dirección, $\mathbf{u}_2$, se construye de forma tal de que la separación entre grupos sea máxima, y sea ortogonal a la dirección determinada por $\mathbf{u}_1$ (esto es, $\langle \mathbf{u}_1,\, \mathbf{u}_2 \rangle = 0$).

Para determinar con cuántas funciones discriminantes trabajar se calcula la variación explicada por ellas, siendo $\dfrac{\lambda_1}{\sum_{j=1}^{r} \lambda_j}$ la de la primera, $\dfrac{\lambda_1 + \lambda_2}{\sum_{j=1}^{r} \lambda_j}$ la primera y la segunda juntas, y así sucesivamente. La correlación entre las variables orginales y las combinaciones lineales establece la importancia de cada una de las variables originales para discriminar.

AD probabilístico (distribución normal)

Se asume que cada grupo tiene una distribución normal p-variada.

$$\begin{aligned} \mathbf{x}_1 &\sim \mathrm{N}_p(\boldsymbol{\mu}_1,\, \boldsymbol{\Sigma}_1) \\ \mathbf{x}_2 &\sim \mathrm{N}_p(\boldsymbol{\mu}_2,\, \boldsymbol{\Sigma}_2) \\ &\vdots \\ \mathbf{x}_k &\sim \mathrm{N}_p(\boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}_k) \end{aligned}$$

AD lineal

Adicionalmente, se asume que las matrices de covarianzas son iguales en todos los grupos:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_k$$

Sabemos que para aplicar la regla de probabilidad a posteriori, se debe maximizar la probabilidad condicional $\Pr(i \in g|\mathbf{x} = \mathbf{x}_i)$. Sabemos tambien que maximizar dicha probabilidad condicional es equivalente a maximizar $\pi_g \exp\left\{-\frac{1}{2}D_{ig}^2\right\}$. Definimos la función L_{ig} como el logaritmo de nuestra función objetivo, esto es:

$$\begin{aligned} L_{ig} &= \log \pi_g - \frac{1}{2} D_{ig}^2 \\ &= \log \pi_g - \frac{1}{2}\left[(\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}^{-1}((\mathbf{x}_i - \boldsymbol{\mu}_g))\right] \\ &= \log \pi_g - \frac{1}{2}\mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \mathbf{x}_i + \frac{1}{2}\boldsymbol{\mu}_g' \boldsymbol{\Sigma}^{-1} \mathbf{x}_i + \frac{1}{2}\mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_g - \frac{1}{2}\boldsymbol{\mu}_g' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_g \end{aligned}$$

dado que $\boldsymbol{\mu}_g$ y $\boldsymbol{\Sigma}$ son desconocidos, utilizamos sus estimadores MV. Adicionalmente, dado que $\boldsymbol{\Sigma}$ es una matriz simétrica, tenemos que $\frac{1}{2}\boldsymbol{\mu}_g' \boldsymbol{\Sigma}^{-1} \mathbf{x}_i + \frac{1}{2}\mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_g = \boldsymbol{\mu}_g' \boldsymbol{\Sigma}^{-1} \mathbf{x}_i$. Por útimo, dado que el vector $\mathbf{x}_i$ es el mismo para todos los grupo, el término $\mathbf{x}_i' \mathbf{S}^{-1} \mathbf{x}_i$ puede descartarse ya que es constante. De esta forma, obtenemos que la función L_{ig} está dada por

$$L_{ig} = \log \hat{\pi}_g + \bar{\mathbf{x}}_g' \mathbf{S}^{-1} \mathbf{x}_i - \frac{1}{2}\bar{\mathbf{x}}_g' \mathbf{S}^{-1} \bar{\mathbf{x}}_g$$

La función L_{ig} es la conocida como **función discriminante lineal**. El valor que L_{ig} toma es conocido como el **score** de la observación i en el grupo g. La misma es lineal en las x.

La función que permite determinar en qué grupo clasificar una observación es llamada **función de la clasificación**. Para hallarla, partimos de la desigualdad derivada en la sección de reglas de decisión: la observación i se asigna al grupo g, sí, y solo si,

$$\Pr(i \in g|\mathbf{x} = \mathbf{x}_i) > \Pr(i \in g'|\mathbf{x} = \mathbf{x}_i) \;\; \forall g \neq g'$$

Para el caso de la normalidad con igualdad de matrices de covarianzas, sabemos que esto es equivalente a asignar la observación i al grupo g si se cumple que:

$$\hat{\pi}_g \, \hat{D}^2_{ig} > \hat{\pi}_{g'} \, \hat{D}^2_{ig'} \;\; \forall g \neq g'$$

Por lo tanto, para hallar la función de clasificación, debemos comparar las funciones discriminantes para los grupo g y g'. Esto es, i se asignará al grupo g si se cumple:

$$\begin{aligned}
L_{ig} &> L_{ig'} \\
\log \hat{\pi}_g + \bar{\mathbf{x}}'_g \, \mathbf{S}^{-1} \, \mathbf{x}_i - \frac{1}{2} \bar{\mathbf{x}}'_g \, \mathbf{S}^{-1} \, \bar{\mathbf{x}}_g &> \log \hat{\pi}_{g'} + \bar{\mathbf{x}}'_{g'} \, \mathbf{S}^{-1} \, \mathbf{x}_i - \frac{1}{2} \bar{\mathbf{x}}'_{g'} \, \mathbf{S}^{-1} \, \bar{\mathbf{x}}_{g'} \\
\bar{\mathbf{x}}'_g \, \mathbf{S}^{-1} \, \mathbf{x}_i - \bar{\mathbf{x}}'_{g'} \, \mathbf{S}^{-1} \, \mathbf{x}_i - \frac{1}{2} \bar{\mathbf{x}}'_g \, \mathbf{S}^{-1} \, \bar{\mathbf{x}}_g + \frac{1}{2} \bar{\mathbf{x}}'_{g'} \, \mathbf{S}^{-1} \, \bar{\mathbf{x}}_{g'} &> \log \hat{\pi}_{g'} - \log \hat{\pi}_g \\
\left(\bar{\mathbf{x}}'_g - \bar{\mathbf{x}}'_{g'} \right) \mathbf{S}^{-1} \, \mathbf{x}_i - \frac{1}{2} \bar{\mathbf{x}}'_g \, \mathbf{S}^{-1} \, \bar{\mathbf{x}}_g + \frac{1}{2} \bar{\mathbf{x}}'_{g'} \, \mathbf{S}^{-1} \, \bar{\mathbf{x}}_{g'} &> \log \left(\frac{\hat{\pi}_{g'}}{\hat{\pi}_g} \right) \\
\left(\bar{\mathbf{x}}'_g - \bar{\mathbf{x}}'_{g'} \right) \mathbf{S}^{-1} \, \mathbf{x}_i - \frac{1}{2} \left[\bar{\mathbf{x}}'_g \, \mathbf{S}^{-1} \, \bar{\mathbf{x}}_g - \bar{\mathbf{x}}'_{g'} \, \mathbf{S}^{-1} \, \bar{\mathbf{x}}_{g'} \right] &> \log \left(\frac{\hat{\pi}_{g'}}{\hat{\pi}_g} \right) \\
\left(\bar{\mathbf{x}}'_g - \bar{\mathbf{x}}'_{g'} \right) \mathbf{S}^{-1} \, \mathbf{x}_i - \frac{1}{2} \left[\left(\bar{\mathbf{x}}'_g - \bar{\mathbf{x}}'_{g'} \right) \mathbf{S}^{-1} \left(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_{g'} \right) \right] &> \log \left(\frac{\hat{\pi}_{g'}}{\hat{\pi}_g} \right) \\
\left(\bar{\mathbf{x}}'_g - \bar{\mathbf{x}}'_{g'} \right) \mathbf{S}^{-1} \left[\mathbf{x}_i - \frac{1}{2} \left(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_{g'} \right) \right] &> \log \left(\frac{\hat{\pi}_{g'}}{\hat{\pi}_g} \right) \\
\underbrace{\left(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_{g'} \right)' \mathbf{S}^{-1} \left[\mathbf{x}_i - \frac{1}{2} \left(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_{g'} \right) \right]}_{L_{igg'}} &> \log \left(\frac{\hat{\pi}_{g'}}{\hat{\pi}_g} \right)
\end{aligned}$$

Por lo tanto, la región de clasificación queda determinada por los $k-1$ hiperplanos definidos por las inecuaciones de la forma

$$R_{gg'} : \; L_{igg'} > \log \left(\frac{\hat{\pi}_{g'}}{\hat{\pi}_g} \right)$$

Los bordes de estas regiones están determinadas por las ecuaciones $L_{igg'} = \log \left(\frac{\hat{\pi}_{g'}}{\hat{\pi}_g} \right)$, y en ellos, no se cuenta con regla de decisión.

Si lo costos de error de clasificación no son iguales entre grupos, la regla de decisión será:

$$R_{gg'} : \; L_{igg'} > \log \left(\frac{\hat{\pi}_{g'} \, c(g|g')}{\hat{\pi}_g \, c(g'|g)} \right)$$

AD cuadrático

A diferencia del ADL, el cuadrático no asume la igualdad de matrices de covarianzas. No obstante ello, el planteo del problema se mantiene incambiado:

$$\max\Big\{\Pr(i \in g|\mathbf{x}=\mathbf{x}_i)\Big\} = \max\left\{\pi_g \exp\left\{-\frac{1}{2}D^2_{ig}\right\}|\mathbf{\Sigma}_g|^{-1/2}\right\}$$

Definimos la **función discriminante cuadrática** aplicando logarítmos y utilizando los estimadores MV para $\boldsymbol{\mu}_g$ y $\mathbf{\Sigma}_g$ en la función objetivo. De esta forma, tenemos que:

$$\begin{aligned}
Q_{ig} &= \pi_g \exp\left\{-\frac{1}{2}D^2_{ig}\right\}|\mathbf{\Sigma}_g|^{-1/2} \\
&= \log\hat{\pi}_g - \frac{1}{2}\hat{D}^2_{ig} - \frac{1}{2}\log|\mathbf{S}_g| \\
&= \log\hat{\pi}_g - \frac{1}{2}\log|\mathbf{S}_g| - \frac{1}{2}(\mathbf{x}_i-\bar{\mathbf{x}}_g)'\mathbf{S}_g^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}_g) \\
&= \log\hat{\pi}_g - \frac{1}{2}\log|\mathbf{S}_g| - \frac{1}{2}\mathbf{x}_i'\mathbf{S}_g^{-1}\mathbf{x}_i + \frac{1}{2}\mathbf{x}_i'\mathbf{S}_g^{-1}\bar{\mathbf{x}}_g + \frac{1}{2}\bar{\mathbf{x}}_g'\mathbf{S}_g^{-1}\mathbf{x}_i - \frac{1}{2}\bar{\mathbf{x}}_g'\mathbf{S}_g^{-1}\bar{\mathbf{x}}_g
\end{aligned}$$

La función Q_{ig} se utiliza para calcular el **score** de la observación i en el grupo g, para cada grupo $g = 1, \ldots, k$. Así entonces, i se asigna al grupo g, sí, y solo si, $Q_{ig} > Q_{ig'} \ \ \forall g \neq g'$.

Comparación entre LDA y QDA

La diferencia estadística entre LDA y QDA radica en el bias-variance trade-off. Cuando se cuenta con p variables, la estimación de la matriz de covarianzas requiere estimar $p(p+1)/2$ parámetros para LDA, pero $Kp(p+1)/2$ parámetros para QDA.

LDA es un método menos flexible que QDA y, consequentemente, tiene menor varianza en la estimación de sus parámetros. Si bien esto puede derivar en una mejor capacidad predictiva en la clasificación, también puede generar grandes sesgos en los casos en que el supuesto de homocedasticidad no sea creible.

AD probabilístico (distribución desconocida)

Logit con dos grupos

Supongamos que se cuenta con una población en la que existen dos grupos, a los que se suele denominar como "éxito" y "fracaso". Por lo tanto, podemos modelar la variable de respuesta según el modelo:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

Por lo que sabemos que:

$$\Pr(Y_i=1)=\pi_i \qquad \Pr(Y_i=0)=1-\pi_i \qquad \mathbf{E}(Y_i|X)=\pi_i$$

$$\Pr(Y_i = 1|\mathbf{x} = \mathbf{x}_i) = \pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

Este modelo se puede linealizar utilizando las *transformación Logit*, y encontramos que:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}$$

Al cociente $\frac{\pi_i}{1-\pi_i} = e^{\mathbf{x}_i'\beta}$ se lo denomina *odds*, y se lo puede conciderar como una medida de riesgo. Es el cociente entre la probabilidad de éxito π_i y la probabilidad de fracaso $1 - \pi_i$.

Para hallar estimaciones de los parámetros se utiliza el método de máxima verosimilitud. Dado que $Y_i \sim$ Bernoulli(π_i), su función de densidad está dada por:

$$f(Y_i) = \pi_i^{Y_i}\left(1-\pi_i\right)^{1-Y_i}$$

Si asumimos que las variables Y_i son independientes, tenemos que su función de verosimilitud está dada por:

$$\mathscr{L}(\beta) = \prod_{i=1}^{n} f(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}\left(1-\pi_i\right)^{1-Y_i}$$

y su log-verosimilitud será:

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n}\log\left(1-Y_i\right)\\ &= \sum_{i=1}^{n} Y_i\mathbf{x}_i'\boldsymbol{\beta} + \sum_{i=1}^{n}\log\left(1-Y_i\right)\end{aligned}$$

Las estimaciones de los coeficientes se obtienen resolviendo las CPO:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$$

utilizando algorítmos numéricos, por ejemplo Newton-Raphson. Los valores estimados de Y_i se obtienen utilizando las estimaciones de los parámetros:

$$\hat{\pi}_i = \frac{e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}}{1 + e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}}$$

Debido a que el AD logístico se trata de un modelo, se pueden realizar pruebas de significación del modelo, de los parámetros y de la bondad de ajuste. Las pruebas de significación tienen la forma habitual:

$$\begin{aligned}&H_0)\ \ \beta_1 = \beta_2 = \ldots = \beta_p = 0\\ &H_1)\ \ \exists\beta_j \neq 0\end{aligned}$$

$$H_0)\ \ \beta_q = \beta_{q+1} = \ldots = \beta_p = 0$$
$$H_1)\ \ \exists \beta_j \neq 0$$

Un estadístico que suele utilizarse para dichas pruebas es el de ratio de verosimilitud. Primero se define:

$$\lambda = \frac{\mathscr{L}_M}{\mathscr{L}_R}$$

donde $\mathscr{L}_M$ es la verosimilitud del modelo ajustado, mientras que $\mathscr{L}_R$ es la verosimilitud del modelo reducido. El estadístico de ratio de verosimilitud se define como:

$$-2\log\lambda = -2\log\left(\frac{\mathscr{L}_M}{\mathscr{L}_R}\right) \sim \chi^2_{p+1-q}$$

Para las pruebas de significación indiviidual de los parámetros se puede utilizar el estadístico de Wald. La prueba es de la forma:

$$H_0)\ \ \beta_k = 0$$
$$H_1)\ \ \beta_k \neq 0$$

y el estadístico se define como:

$$W = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \overset{a}{\sim} \mathrm{N}(0,\, 1)$$

Una vez estimado el modelo y calculadas las predicciones $\hat{\pi}_i$, se debe decidir cuáles clasificar como éxito ($Y_i = 1$) y cuáles como fracaso ($Y_i = 0$). Salvo casos en que otra información esté disponible, lo usual es tomar 0,5 como punto de corte. De esta forma entonces:

$$\hat{Y}_i = \begin{cases} 1 & \text{si} \quad \hat{\pi}_i \geq 0.5 \\ 0 & \text{si} \quad \hat{\pi}_i < 0.5 \end{cases}$$

La acuracidad de predicción de un modelo de respuesta discreta suele medirse mediante la *sensibilidad* y la *especificidad*, donde:

$$\text{sensibilidad} = \frac{\#(\text{éxitos observados clasificados como éxitos})}{\#(\text{éxitos observados})}$$

$$\text{especificidad} = \frac{\#(\text{fracasos observados clasificados como fracasos})}{\#(\text{fracasos observados})}$$

Estos resultados suelen mostrarse en la matriz de confusión (observados Vs. predichos) y/o en la curva ROC (sensibilidad Vs. 1 - especificidad). Estos indicadores también se utilizar para el ADL y el ADC.

Logit con múltiples grupos

Si la variable de respuesta tiene más de dos categorías, existen varias formas de plantear el modelo.

Logits acumulados (modelos proporcionales)

El ajuste se basa en que la influencia de las variables explicativas en la variable de respuesta es independienete del punto de corte del logit acumulado. El modelo tiene la siguiente especificación:

$$logit(\theta_{ik}) = \alpha_k + \mathbf{x}_i'\boldsymbol{\beta}$$

donde i indica la observación y k la sub-población o grupo. Los modelos tienen entonces la siguiente forma:

$$\begin{aligned}
logit(\theta_{i1}) &= \log\left(\frac{\pi_{i1}}{\pi_{i2}+\pi_{i3}+\ldots+\pi_{ik}}\right) &&= \mathbf{x}_i'\boldsymbol{\beta} \\
logit(\theta_{i2}) &= \log\left(\frac{\pi_{i1}+\pi_{i2}}{\pi_{i3}+\pi_{i4}+\ldots+\pi_{ik}}\right) &&= \mathbf{x}_i'\boldsymbol{\beta} \\
\vdots &\qquad\vdots &&\qquad\vdots \\
logit(\theta_{ik-1}) &= \log\left(\frac{\pi_{i1}+\pi_{i2}+\ldots+\pi_{ik-1}}{\pi_{ik}}\right) &&= \mathbf{x}_i'\boldsymbol{\beta}
\end{aligned}$$

donde π_{ig} es la probabilidad de que la observación i pertenezca al grupo g.

Logits generalizados

Se utiliza una categoría como referencia. Para cada categoría se ajusta un modelo logit. Existen por lo tanto $K-1$ juegos de parámetros para cada uno de los $K-1$ modelos de la forma:

$$logit_{ig} = \log\left(\frac{\pi_{ig}}{\pi_{iK}}\right) = \mathbf{x}_{ig}'\boldsymbol{\beta}_g$$

donde π_{ig} es la probabilidad de que la observación i pertenezca al grupo $g = 1, \ldots, K$.

Logits de categorías adyacentes

También existen $K-1$ modelos, pero cada modelo compara su categoría, con la categoría anterior.

$$logit_{i,\,g} = \log\left(\frac{\pi_{i,\,g-1}}{\pi_{i,\,g}}\right) = \mathbf{x}_{i,\,g-1}'\boldsymbol{\beta}_{g-1}$$

donde $\pi_{i,\,g}$ es la probabilidad de que la observación i pertenezca al grupo $g = 2, \ldots, K$.

Referencias

Blanco, Jorge, and others. 2006. "Introducción Al análisis Multivariado." *IESTA. Montevideo*.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.

Peña, Daniel. 2013. *Análisis de Datos Multivariantes*. McGraw-Hill España.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rencher, Alvin C. 1998. *Multivariate Statistical Inference and Applications*. Wiley New York.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.