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

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

La densidad puede estimarse utilizando los estimadores MV de μ_g y $\mathbf{\Sigma}_g$, $\bar{\mathbf{x}}_g$ 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 μ_g y $\mathbf{\Sigma}_g$ mencionadas 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 costo 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 laa 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 rrepresentadas 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 contrar 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):

$$\frac{\partial(\bullet)}{\partial \mathbf{u}'} = \mathbf{0} \Rightarrow \frac{2\mathbf{B}\mathbf{u}(\mathbf{u}'\mathbf{W}\mathbf{u}) - 2(\mathbf{u}'\mathbf{B}\mathbf{u})\mathbf{W}\mathbf{u}}{(\mathbf{u}'\mathbf{W}\mathbf{u})^2} = \mathbf{0}$$

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 $\mathbf{u}'\mathbf{B}\mathbf{u}$ y $\mathbf{u}'\mathbf{W}\mathbf{u}$ son escalares. Por lo tanto, las siguientes manipulaciones son válidas:

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

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

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

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

Hallamos entonces que $\mathbf{u}$ es entonces 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 $\mathbf{u}'\mathbf{B}\mathbf{u}$ 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. Esta 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.

$$\begin{aligned} \mathbf{x}_1 &\sim \mathrm{N}_p(\mu_1,\, \mathbf{\Sigma}_1) \\ \mathbf{x}_2 &\sim \mathrm{N}_p(\mu_2,\, \mathbf{\Sigma}_2) \\ &\vdots \\ \mathbf{x}_k &\sim \mathrm{N}_p(\mu_k,\, \mathbf{\Sigma}_k) \end{aligned}$$

AD lineal

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

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

Sabemos que para aplicar la regla de probabilidad a posteriori, se debe maximizar $\Pr(i \in g|\mathbf{x} = \mathbf{x}_i)$, lo cual requiere maximizar $\pi_g \exp\left\{-\frac{1}{2}D_{ig}^2\right\}$. Tomando logaritmos tenemos que la función objetivo es entonces:

$$\begin{aligned}
\log \pi_g - \frac{1}{2}D_{ig}^2 &= \log \pi_g - \frac{1}{2}\left[(\mathbf{x}_i - \mu_g)'\mathbf{\Sigma}^{-1}((\mathbf{x}_i - \mu_g))\right] \\
&= \log \pi_g - \frac{1}{2}\mathbf{x}_i' \mathbf{\Sigma}^{-1} \mathbf{x}_i + \frac{1}{2}\mu_g' \mathbf{\Sigma}^{-1} \mathbf{x}_i + \frac{1}{2}\mathbf{x}_i' \mathbf{\Sigma}^{-1} \mu_g - \frac{1}{2}\mu_g' \mathbf{\Sigma}^{-1} \mu_g \\
&= \log \hat{\pi}_g - \frac{1}{2}\mathbf{x}_i' \mathbf{S}^{-1} \mathbf{x}_i + \frac{1}{2}\bar{\mathbf{x}}_g' \mathbf{S}^{-1} \mathbf{x}_i + \frac{1}{2}\mathbf{x}_i' \mathbf{S}^{-1} \bar{\mathbf{x}}_g - \frac{1}{2}\bar{\mathbf{x}}_g' \mathbf{S}^{-1} \bar{\mathbf{x}}_g
\end{aligned}$$

AD probabilístico (distribución desconocida)

Referencias

Beygelzimer, Alina, Sham Kakadet, John Langford, Sunil Arya, David Mount, and Shengqiao Li. 2018. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. https://CRAN.R-project.org/package=FNN.

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

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.

Wasserman, Larry. 2007. *All of Nonparametric Statistics*. Springer, New York.

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