Análisis Discriminante

Mathias Bourel

IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

October 12, 2019

Introducción

Problema: Se dispone de un conjunto amplio de individuos que pueden venir de dos o más poblaciones. Para cada individuo se observa una variable aleatoria p dimensional. Se desea clasificar un nuevo individuo, con valores de las variables conocidas, en una de las poblaciones.

Se puede considerar como un análisis de regresión donde la variable dependiente es categórica (etiqueta de cada grupo) y las variables independientes son continuas. Queremos encontrar una relación lineal entre estas variables que mejor discrimine a los individuos.

Aplicaciones
credit scoring (ingresos, antiguedad trabajo, patrimonio para predecir comportamiento futuro), reconocimiento de patrones, aplicaciones medicales (paciente con cierta enfermedad).

Introducción

$\mathcal{L} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ muestra de datos.

- **Discriminar:** usar $\mathcal{L}$ para construir un clasificador (función de las caracteristicas X_i) para separar lo mejor posibles los grupos dados.
- **Clasificar:** usar el clasificador para predecir la etiqueta Y_{new} de una nueva observación X_{new}.

Suponemos que hay dos grupos G_1 y G_2 y que cada individuo pertenece a un único grupo (por ejemplo sano/enfermo, spam/no spam).

Clasificador de Bayes

Sean $P_1 \sim f_1$ y $P_2 \sim f_2$ dos poblaciones. Queremos clasificar un nuevo elemento x_0 que proviene de una variable aleatoria X en una de estas dos poblaciones. Se sabe que $\pi_1 = \mathbb{P}(X \in P_1)$ y $\pi_2 = \mathbb{P}(X \in P_2)$ y que $\pi_1 + \pi_2 = 1$.
Si $\mathbb{P}(X = x | X \in P_i) = f_i(x)\Delta x$, entonces la distribución de X es

$$f_X(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$$

Entonces

$$\mathbb{P}(1|x_0) = \mathbb{P}(X \in P_1 | X = x_0) = \frac{\mathbb{P}(x_0|1)\pi_1}{\pi_1 \mathbb{P}(x_0|1) + \pi_2 \mathbb{P}(x_0|2)} = \frac{f_1(x_0)\pi_1}{\pi_1 f_1(x_0) + \pi_2 f_2(x_0)}$$

$$\mathbb{P}(2|x_0) = \mathbb{P}(X \in P_2 | X = x_0) = \frac{P(x_0|2)\pi_2}{\pi_1 \mathbb{P}(x_0|1) + \pi_2 \mathbb{P}(x_0|2)} = \frac{f_2(x_0)\pi_2}{\pi_1 f_1(x_0) + \pi_2 f_2(x_0)}$$

Clasificamos x_0 en la población más probable a posteriori (clasificador de Bayes), es decir en P_2 si

$$\pi_2 f_2(x_0) > \pi_1 f_1(x_0)$$

y si $\pi_1 = \pi_2$, clasificamos en P_2 si

$$f_2(x_0) > f_1(x_0)$$

Con costes

Suponemos que haya un coste por clasificar mal. Notamos por $c(i|j)$ el coste de clasificar en P_i cuando pertenece en realidad a P_j.
El coste esperado de la clasificación de x_0 en P_2 es:

$$c(2|1)\mathbb{P}(1|x_0) + 0\mathbb{P}(2|x_0) = c(2|1)\mathbb{P}(1|x_0)$$

El coste esperado de la clasificación de x_0 en P_1 es:

$$0\mathbb{P}(1|x_0) + c(1|2)\mathbb{P}(2|x_0) = c(1|2)\mathbb{P}(2|x_0)$$

Entonces asignamos x_0 a la población 2 si

$$\frac{f_2(x_0)\pi_2}{c(2|1)} > \frac{f_1(x_0)\pi_1}{c(1|2)}$$

A igualdad de los otros terminos, clasificamos en P_2 si:

- su probabilidad a priori es más alta.
- la verosimilitud de que x_0 provenga de P_2 es más alta.
- el coste de equivocarnos al clasificarlo en P_2 es más bajo.

Análisis Discriminante Lineal Gaussiano (LDA)

Supongamos que $f_1 \sim N(\mu_1, \Sigma)$ y $f_2 \sim N(\mu_2, \Sigma)$ - misma matriz de covarianzas-. De

$$\frac{f_2(x)\pi_2}{c(2|1)} > \frac{f_1(x)\pi_1}{c(1|2)}$$

tomando logaritmo tenemos que

$$-\frac{1}{2}\underbrace{(x-\mu_2)'\Sigma^{-1}(x-\mu_2)}_{D_2^2} + \log\left(\frac{\pi_2}{c(2|1)}\right) > -\frac{1}{2}\underbrace{(x-\mu_1)'\Sigma^{-1}(x-\mu_1)}_{D_1^2} + \log\left(\frac{\pi_1}{c(1|2)}\right) \quad (*)$$

donde D_i^2 es la distancia de Mahalanobis entre el punto observado x y la media de la población i (recordar los slides sobre normal multivariada). Entonces:

$$D_1^2 - 2\log\left(\frac{\pi_1}{c(1|2)}\right) > D_2^2 - 2\log\left(\frac{\pi_2}{c(2|1)}\right)$$

y si suponemos que $\pi_1 = \pi_2$ y los costes iguales, clasificamos en la población 2 si

$$D_1^2 > D_2^2$$

Obs: si $\Sigma = \sigma^2 I$ entonces la regla equivale en usar la distancia euclidea.

Análisis Discriminante Lineal Gaussiano

Volviendo a $(*)$, si desarrollamos, al tener la misma matriz de varianzas-covarianzas Σ, se elimina el termino cuadrático $x'\Sigma^{-1}x$. Entonces

$$-\mu_1'\Sigma^{-1}x + \frac{1}{2}\mu_1'\Sigma^{-1}\mu_1 > -\mu_2'\Sigma^{-1}x + \frac{1}{2}\mu_2'\Sigma^{-1}\mu_2 - \log\left(\frac{c(1|2)\pi_2}{c(2|1)\pi_1}\right)$$

$$(\mu_2-\mu_1)'\Sigma^{-1}x > (\mu_2-\mu_1)'\Sigma^{-1}\left(\frac{\mu_1+\mu_2}{2}\right) - \log\left(\frac{c(1|2)\pi_2}{c(2|1)\pi_1}\right)$$

Si $w = \Sigma^{-1}(\mu_2-\mu_1)$, entonces

$$w'x > \underbrace{w'\left(\frac{\mu_1+\mu_2}{2}\right) - \log\left(\frac{c(1|2)\pi_2}{c(2|1)\pi_1}\right)}_{-w_0}$$

clasificamos en la población 2 si

$$w'x > -w_0 \Rightarrow L(x) = w'x + w_0 > 0$$

Análisis Discriminante Lineal Gaussiano

Suponiendo costes y probabilidades a priori iguales, volviendo a $w'x > \underbrace{w'\left(\frac{\mu_1+\mu_2}{2}\right)}_{-w_0}$ (es decir $L(x) > 0$) entonces:

$$w'x - w'\mu_1 > w'\mu_2 - w'x$$

Entonces el procedimiento para clasificar el individuo x_0 en P_1 o en P_2 según este método es el siguiente:

1. Calcular el vector $w = \Sigma^{-1}(\mu_2 - \mu_1)$.
2. Construir la variables indicadora discriminante $z = w'x$
3. Clasificar en la población donde la distancia $|z_0 - m_i|$ es mínima siendo $z_0 = w'x_0$ y $m_i = w'\mu_i$.

Observar que:

- $Var(z) = Var(w'x) = w'Var(x)w = w'\Sigma w = \underbrace{(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1)}_{D^2}$
- Por otro lado:
  $m_2 - m_1 = w'(\mu_2 - \mu_1) = (\Sigma^{-1}(\mu_2 - \mu_1))'(\mu_2 - \mu_1) = (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1) = D^2$

Entonces

$$Var(z) = m_2 - m_1$$

Análisis Discriminante Lineal Gaussiano

Podemos interpretar a la variable z de la siguiente manera:
si dividimos la relación $w'x - w'\mu_1 > w'\mu_2 - w'x$ por $||w||$ y $u = \frac{w}{||w||}$ entonces

$$u'x - u'\mu_1 > u'\mu_2 - u'x$$

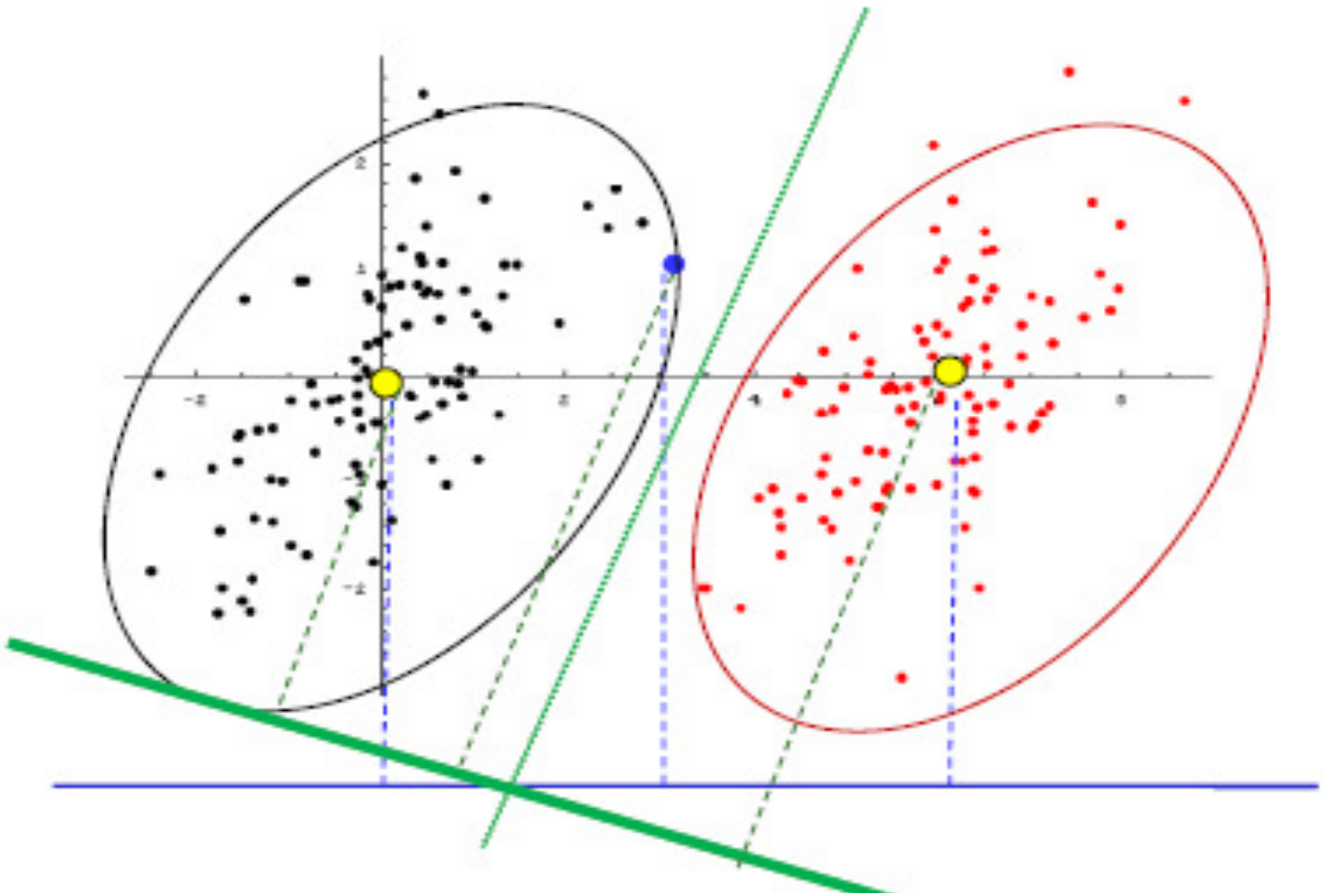
y $\widehat{P}_u(x) = u'x$ es la proyección (el escalar) de x en la dirección u, y $u'\mu_i$ es la proyección de μ_i en la dirección u para $i = 1, 2$. Entonces elegiremos la población 2 si

$$\widehat{P}_u(x) > \widehat{P}_u\left(\frac{\mu_1 + \mu_2}{2}\right)$$

(el hiperplano perpendicular a u por $u'\left(\frac{\mu_1+\mu_2}{2}\right)$ divide el espacio muestral en dos regiones)

Interpretación geométrica del Análisis Discriminante Lineal

En la figura siguiente representamos la situación establecida en la transparencia anterior: proyectando el punto medio de las medias sobre u (el punto medio de los dos puntos amarillos), y proyectando x (el punto azul) sobre u sabremos cuál de las dos poblaciones atribuirle.



Cálculo de probabilidades de error

Recordamos que la variable $z = w'x$ tiene esperanza $\mathbb{E}(z) = m_i = w'\mu_i$ y varianza $D^2 = m_2 - m_1$. Entonces

$$\mathbb{P}(2|1) = \mathbb{P}\left(z \geq \frac{m_1 + m_2}{2} | z \sim N(m_1, D)\right) = \mathbb{P}\left(y \geq \frac{\frac{m_1+m_2}{2} - m_1}{D} | y \sim N(0,1)\right)$$

$$\mathbb{P}(2|1) = 1 - \Phi\left(\frac{D}{2}\right)$$

$$\mathbb{P}(1|2) = \mathbb{P}\left(z \leq \frac{m_1 + m_2}{2} | z \sim N(m_2, D)\right) = \mathbb{P}\left(y \leq \frac{\frac{m_1+m_2}{2} - m_2}{D} | y \sim N(0,1)\right)$$

$$\mathbb{P}(1|2) = \Phi\left(-\frac{D}{2}\right)$$

Las probabilidades de error son iguales, el error de clasificación sólo depende de la distancia de Mahalanobis entre las medias.

Probabilidads a posteriori

Volviendo a la cuenta del principio:

$$\mathbb{P}(1|x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}$$

$$\mathbb{P}(1|x) = \frac{\pi_1 \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right)}{\pi_1 \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right) + \pi_2 \exp\left(-\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right)}$$

$$\mathbb{P}(1|x) = \frac{1}{1 + \frac{\pi_2}{\pi_1} \exp\left(-\frac{1}{2}(D_2^2 - D_1^2)\right)}$$

En el caso que las probabilidades a priori sean iguales, cuanto más alejado está el punto de la población 1, $(D_1^2 > D_2^2)$, el denominador es más grande y menor será $\mathbb{P}(1|x)$ y al contrario.

Ejemplo (Peña, pág 406)

Retrato entre dos posibles pintores. Se miden dos variables: X_1 profundidad del trazo y X_2 proporción que ocupa el retrato sobre la superficie del lienzo.

Retratos del pintor A $\sim N\left(\mu_A = \begin{pmatrix} 2 \\ 0.8 \end{pmatrix}, \Sigma\right)$, Restratos de pintor B $\sim N\left(\mu_B = \begin{pmatrix} 2.3 \\ 0.7 \end{pmatrix}, \Sigma\right)$

Covarianzas $\Sigma = \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}$, nueva obra a clasificar $x_0 = \begin{pmatrix} 2.1 \\ 0.75 \end{pmatrix}$

$$D_A^2 = \begin{pmatrix} 2.1-2 & 0.75-0.8 \end{pmatrix} \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}^{-1} \begin{pmatrix} 2.1-2 \\ 0.75-0.8 \end{pmatrix} = 0.52$$

$$D_B^2 = \begin{pmatrix} 2.1-2.3 & 0.75-0.7 \end{pmatrix} \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}^{-1} \begin{pmatrix} 2.1-2.3 \\ 0.75-0.7 \end{pmatrix} = 0.8133$$

Entonces

$$\mathbb{P}(A|x) = \frac{1}{1+\exp\left(-\frac{1}{2}(D_B^2 - D_A^2)\right)} = \frac{1}{1+\exp\left(-\frac{1}{2}(0,8133 - 0.52)\right)} = 0,5376$$

$$\mathbb{P}(A|B) = 1 - \Phi\left(\frac{D^2}{2}\right) = 1 - \Phi\left(\frac{\begin{pmatrix} 2-2.3 & 0.8-0.7 \end{pmatrix}\begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}^{-1}\begin{pmatrix} 2-2.3 \\ 0.8-0.7 \end{pmatrix}}{2}\right) =$$

$$1 - \Phi(0.808) = 0,189$$

Generalización para varias poblaciones desconocidas

Ahora vamos a generalizar lo anterior suponiendo que tenemos G poblaciones y que no conocemos la distribución de las que provienen (no es normal, hay que estimar media y matriz de varianzas-covarianzas).

Generalización para varias poblaciones desconocidas

Ahora vamos a generalizar lo anterior suponiendo que tenemos G poblaciones y que no conocemos la distribución de las que provienen (no es normal, hay que estimar media y matriz de varianzas-covarianzas). Supongamos que tenemos G poblaciones. Consideramos la matriz $X \in \mathcal{M}_{n\times p}$ y notamos por x_{ijg} donde i es el individuo, j la característica y g la población. Sea n_g la cantidad de elementos en el grupo g entonces la cantidad global de individuos es $n = \sum_{g=1}^{G} n_g$.

Notaciones:

- $\mathbf{x_{ig}}' = (x_{i1g}, \ldots, x_{ipg}) \in \mathbb{R}^p$ (p variables del individuo i en la pop. g)

Generalización para varias poblaciones desconocidas

Ahora vamos a generalizar lo anterior suponiendo que tenemos G poblaciones y que no conocemos la distribución de las que provienen (no es normal, hay que estimar media y matriz de varianzas-covarianzas). Supongamos que tenemos G poblaciones. Consideramos la matriz $X \in \mathcal{M}_{n\times p}$ y notamos por x_{ijg} donde i es el individuo, j la característica y g la población. Sea n_g la cantidad de elementos en el grupo g entonces la cantidad global de individuos es $n = \sum_{g=1}^{G} n_g$.

Notaciones:

- $\mathbf{x_{ig}}' = (x_{i1g}, \ldots, x_{ipg}) \in \mathbb{R}^p$ (p variables del individuo i en la pop. g)
- Vector de medias de los individuos de la población g: $\bar{\mathbf{x}}_\mathbf{g} = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{x_{ig}} \in \mathbb{R}^p$.

Generalización para varias poblaciones desconocidas

Ahora vamos a generalizar lo anterior suponiendo que tenemos G poblaciones y que no conocemos la distribución de las que provienen (no es normal, hay que estimar media y matriz de varianzas-covarianzas). Supongamos que tenemos G poblaciones. Consideramos la matriz $X \in \mathcal{M}_{n\times p}$ y notamos por x_{ijg} donde i es el individuo, j la característica y g la población. Sea n_g la cantidad de elementos en el grupo g entonces la cantidad global de individuos es $n = \sum_{g=1}^{G} n_g$.

Notaciones:

- $\mathbf{x_{ig}}' = (x_{i1g}, \ldots, x_{ipg}) \in \mathbb{R}^p$ (p variables del individuo i en la pop. g)
- Vector de medias de los individuos de la población g: $\bar{\mathbf{x}}_\mathbf{g} = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{x_{ig}} \in \mathbb{R}^p$.

  Utilizaremos $\bar{\mathbf{x}}_\mathbf{g}$ como estimación de μ_g.

Generalización para varias poblaciones desconocidas

Ahora vamos a generalizar lo anterior suponiendo que tenemos G poblaciones y que no conocemos la distribución de las que provienen (no es normal, hay que estimar media y matriz de varianzas-covarianzas). Supongamos que tenemos G poblaciones. Consideramos la matriz $X \in \mathcal{M}_{n\times p}$ y notamos por x_{ijg} donde i es el individuo, j la característica y g la población. Sea n_g la cantidad de elementos en el grupo g entonces la cantidad global de individuos es $n = \sum_{g=1}^{G} n_g$.

Notaciones:

- $\mathbf{x_{ig}}' = (x_{i1g}, \ldots, x_{ipg}) \in \mathbb{R}^p$ (p variables del individuo i en la pop. g)
- Vector de medias de los individuos de la población g: $\bar{\mathbf{x}}_\mathbf{g} = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{x_{ig}} \in \mathbb{R}^p$.

  Utilizaremos $\bar{\mathbf{x}}_\mathbf{g}$ como estimación de μ_g.
- Matriz de covarianzas para la clase g:

$$\widehat{S}_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (\mathbf{x_{ig}} - \bar{\mathbf{x}}_\mathbf{g})(\mathbf{x_{ig}} - \bar{\mathbf{x}}_\mathbf{g})'$$

  Supondremos que las G poblaciones tienen la misma matriz de varianzas y covarianzas Σ (en la práctica se debe realizar la prueba M de Box: ¡buscar información!).

Generalización para varias poblaciones desconocidas

Ahora vamos a generalizar lo anterior suponiendo que tenemos G poblaciones y que no conocemos la distribución de las que provienen (no es normal, hay que estimar media y matriz de varianzas-covarianzas). Supongamos que tenemos G poblaciones. Consideramos la matriz $X \in \mathcal{M}_{n\times p}$ y notamos por x_{ijg} donde i es el individuo, j la característica y g la población. Sea n_g la cantidad de elementos en el grupo g entonces la cantidad global de individuos es $n = \sum_{g=1}^{G} n_g$.

Notaciones:

- $\mathbf{x_{ig}}' = (x_{i1g}, \ldots, x_{ipg}) \in \mathbb{R}^p$ (p variables del individuo i en la pop. g)
- Vector de medias de los individuos de la población g: $\bar{\mathbf{x}}_\mathbf{g} = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{x_{ig}} \in \mathbb{R}^p$.

  Utilizaremos $\bar{\mathbf{x}}_\mathbf{g}$ como estimación de μ_g.
- Matriz de covarianzas para la clase g:

$$\widehat{S}_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (\mathbf{x_{ig}} - \bar{\mathbf{x}}_\mathbf{g})(\mathbf{x_{ig}} - \bar{\mathbf{x}}_\mathbf{g})'$$

  Supondremos que las G poblaciones tienen la misma matriz de varianzas y covarianzas Σ (en la práctica se debe realizar la prueba M de Box: ¡buscar información!).
- La matriz de varianzas y covarianzas de la población global es (para estimar Σ):

$$\widehat{S}_w = \sum_{g=1}^{G} \frac{n_g - 1}{n - G} \widehat{S}_g$$

Generalización para varias poblaciones desconocidas

Para obtener las funciones discriminantes entre las clases usaremos:

- $\bar{\mathbf{x}}_{\mathbf{g}}$ como estimación de μ_g
- $\widehat{S}_w$ como estimación de Σ.

Generalización para varias poblaciones desconocidas

Para obtener las funciones discriminantes entre las clases usaremos:

- $\bar{\mathbf{x}}_\mathbf{g}$ como estimación de μ_g
- $\widehat{S}_w$ como estimación de Σ.

Clasificaremos entonces un nuevo individuo x_0 en aquella clase g que haga mínima la distancia de Mahalanobis entre x_0 y la media $\bar{\mathbf{x}}_\mathbf{g}$ del grupo g,

Generalización para varias poblaciones desconocidas

Para obtener las funciones discriminantes entre las clases usaremos:

- $\bar{\mathbf{x}}_\mathbf{g}$ como estimación de μ_g
- $\widehat{S}_w$ como estimación de Σ.

Clasificaremos entonces un nuevo individuo x_0 en aquella clase g que haga mínima la distancia de Mahalanobis entre x_0 y la media $\bar{\mathbf{x}}_\mathbf{g}$ del grupo g, es decir en aquella clase g tal que:

$$\min_{g\in\{1,\ldots,G\}} (x_0-\bar{\mathbf{x}}_\mathbf{g})'\widehat{S}_w^{-1}(x_0-\bar{\mathbf{x}}_\mathbf{g}) = \min_{g\in\{1,\ldots,G\}} \widehat{w}_g'(\bar{\mathbf{x}}_\mathbf{g}-x_0)$$

siendo

$$\widehat{w}_g = \widehat{S}_w^{-1}(\bar{\mathbf{x}}_\mathbf{g}-x_0)$$

Generalización para varias poblaciones desconocidas

Si notamos por

$$\widehat{w}_{g,g+1} = \widehat{S}_w^{-1}(\overline{\mathbf{x}}_{\mathbf{g}} - \overline{\mathbf{x}}_{\mathbf{g+1}}) = \widehat{w}_g - \widehat{w}_{g+1}$$

las variables discriminantes necesarias son

$$z_{g,g+1} = \widehat{w}'_{g,g+1} x_0, \quad g = 1, \ldots, G$$

Generalización para varias poblaciones desconocidas

Si notamos por

$$\widehat{w}_{g,g+1} = \widehat{S}_w^{-1}(\overline{\mathbf{x}}_{\mathbf{g}} - \overline{\mathbf{x}}_{\mathbf{g+1}}) = \widehat{w}_g - \widehat{w}_{g+1}$$

las variables discriminantes necesarias son

$$z_{g,g+1} = \widehat{w}'_{g,g+1}x_0, \quad g = 1, \ldots, G$$

En efecto, observar que $w_{j,j+2} = w_{j,j+1} + w_{j+1,j+2}\ \forall j = 1, \ldots, G-1$, así que necesito $G-1$ ejes discriminantes si $p > G-1$, ya que todos los demás se deducen de ellos. Por ejemplo si $G = 4$ y conozco $w_{1,2}, w_{2,3}$ y $w_{3,4}$ entonces puedo deducir de la igualdad anterior $w_{1,3}$, $w_{2,4}$ y $w_{1,4}$.

Generalización para varias poblaciones desconocidas

Si notamos por

$$\widehat{w}_{g,g+1} = \widehat{S}_w^{-1}(\overline{\mathbf{x}}_{\mathbf{g}} - \overline{\mathbf{x}}_{\mathbf{g+1}}) = \widehat{w}_g - \widehat{w}_{g+1}$$

las variables discriminantes necesarias son

$$z_{g,g+1} = \widehat{w}'_{g,g+1}x_0, \quad g = 1, \ldots, G$$

En efecto, observar que $w_{j,j+2} = w_{j,j+1} + w_{j+1,j+2}\ \forall j = 1, \ldots, G-1$, así que necesito $G-1$ ejes discriminantes si $p > G-1$, ya que todos los demás se deducen de ellos. Por ejemplo si $G = 4$ y conozco $w_{1,2}, w_{2,3}$ y $w_{3,4}$ entonces puedo deducir de la igualdad anterior $w_{1,3}$, $w_{2,4}$ y $w_{1,4}$.

Cuando $p \geq G-1$, como estos vectores pertenecen a $\mathbb{R}^p$ la cantidad máxima de vectores linealmente independientes es p.

Generalización para varias poblaciones desconocidas

Si notamos por

$$\widehat{w}_{g,g+1} = \widehat{S}_w^{-1}(\overline{\mathbf{x}}_{\mathbf{g}} - \overline{\mathbf{x}}_{\mathbf{g+1}}) = \widehat{w}_g - \widehat{w}_{g+1}$$

las variables discriminantes necesarias son

$$z_{g,g+1} = \widehat{w}'_{g,g+1}x_0, \quad g = 1, \ldots, G$$

En efecto, observar que $w_{j,j+2} = w_{j,j+1} + w_{j+1,j+2}\ \forall j = 1, \ldots, G-1$, así que necesito $G-1$ ejes discriminantes si $p > G-1$, ya que todos los demás se deducen de ellos. Por ejemplo si $G = 4$ y conozco $w_{1,2}, w_{2,3}$ y $w_{3,4}$ entonces puedo deducir de la igualdad anterior $w_{1,3}$, $w_{2,4}$ y $w_{1,4}$.

Cuando $p \geq G-1$, como estos vectores pertenecen a $\mathbb{R}^p$ la cantidad máxima de vectores linealmente independientes es p.

Por todo eso podemos suponer que la cantidad de ejes discriminantes necesarios es

$$r = \min(p, G-1)$$

Generalización para varias poblaciones desconocidas

Si notamos por

$$\widehat{w}_{g,g+1} = \widehat{S}_w^{-1}(\overline{\mathbf{x}}_{\mathbf{g}} - \overline{\mathbf{x}}_{\mathbf{g+1}}) = \widehat{w}_g - \widehat{w}_{g+1}$$

las variables discriminantes necesarias son

$$z_{g,g+1} = \widehat{w}'_{g,g+1}x_0, \quad g = 1, \ldots, G$$

En efecto, observar que $w_{j,j+2} = w_{j,j+1} + w_{j+1,j+2}\ \forall j = 1, \ldots, G-1$, así que necesito $G-1$ ejes discriminantes si $p > G-1$, ya que todos los demás se deducen de ellos. Por ejemplo si $G = 4$ y conozco $w_{1,2}, w_{2,3}$ y $w_{3,4}$ entonces puedo deducir de la igualdad anterior $w_{1,3}$, $w_{2,4}$ y $w_{1,4}$.

Cuando $p \geq G-1$, como estos vectores pertenecen a $\mathbb{R}^p$ la cantidad máxima de vectores linealmente independientes es p.

Por todo eso podemos suponer que la cantidad de ejes discriminantes necesarios es

$$r = \min(p, G-1)$$

Como en el caso de dos clases, clasifico en la clase g en vez de la clase $g+1$ si

$$|z_{g,g+1} - \widehat{m}_g| < |z_{g,g+1} - \widehat{m}_{g+1}|$$

siendo $\widehat{m}_g = \widehat{w}'_{g,g+1}\overline{\mathbf{x}}_{\mathbf{g}}$

Análisis Discriminate Cuadrático (QDA)

Si suponemos que las matrices de covarianzas no son iguales, el clasificador de Bayes asigna la observación x a la clase para la cual:

$$-\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)+\log(\pi_k)=-\frac{1}{2}x'\Sigma_k^{-1}x+x'\Sigma_k^{-1}\mu_k-\frac{1}{2}\mu_k'\Sigma_k^{-1}\mu_k+\log(\pi_k)$$

es mayor (observe que esto es una función cuadrática).

Discriminación logística

Recordamos la función *logit*:

$$\mathbb{P}(X) = \frac{1}{1 + e^{-(\beta_0 + \beta' X)}}$$

Entonces

$$\frac{\mathbb{P}(X)}{1 - \mathbb{P}(X)} = e^{\beta_0 + \beta' X} \in [0, +\infty)$$

y

$$\log\left(\frac{\mathbb{P}(X)}{1 - \mathbb{P}(X)}\right) = \beta_0 + \beta' x$$

1. La discriminación logística es más robusta que LDA a la no normalidad. Si hay normalidad, LDA en general es mejor.
2. La estimación de los paramétros en la discriminación logística se hace por logverosimilitud.

Referencias

1. D. Peña. Análisis de datos multivariantes, Mac Graw Hill, 2002
2. James, Witten, Hastie and Tibshirani. An introduction to Statistical Learning with application in R, Springer, 2013.