MAT-269: Sesión 20, Análisis Discriminante I

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Análisis discriminante

Suponga observaciones multivariadas provenientes de g clases (o grupos) predefinidos teniendo características similares.

Ejemplos:

Especies de plantas, Niveles de solvencia para clientes de un banco, presencia/ausencia de una condición médica, tipos de tumores, si un mensaje es SPAM o no, etc.

Se desea:

(a) Discriminar, esto es usar la información de aquellas observaciones similares para construir una regla de clasificación que permita separar tanto como sea posible las clases predefinidas.

(b) Clasificar, dada las mediciones de una **nueva** observación predecir a que clase pertenece.

Cuando existe sólo dos clases $(g = 2)$ se requiere de un único clasificador, mientras que cuando se dispone de más clases se necesita al menos dos (y a lo más $g - 1$) clasificadores para diferenciar entre clases.

Suponga una población $\mathcal{P}$ particionada en g grupos denotados por $\Pi_1, \dots, \Pi_g$. Además, cada elemento de $\mathcal{P}$ es clasificado sólo en una clase.

Las mediciones de una muestra son usadas para asignar observaciones futuras a las clases designadas.

El vector aleatorio $\boldsymbol{x} = (x_1, \dots, x_p)^\top$ representa las p mediciones de un ítem, las que son escogidas por su habilidad para distinguir entre las g clases.

Clasificación binaria

Considere $g = 2$, y deseamos discriminar entre las clases Π_1 y Π_2.

Regla de clasificación de Bayes

Sea

$$\mathsf{P}(\boldsymbol{x} \in \Pi_i) = \pi_i, \qquad i = 1, 2,$$

la probabilidad a priori de que una observación $\boldsymbol{x}$ seleccionada al azar, pertenezca a Π_1 o Π_2.

Suponga que asociado a la i-ésima clase, existe una densidad

$$\mathsf{P}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{X} \in \Pi_i) = f_i(\boldsymbol{x}), \qquad i = 1, 2$$

(no se realizan supuestos sobre f_i)

Considere $g = 2$, y deseamos discriminar entre las clases Π_1 y Π_2.

Regla de clasificación de Bayes

Sea

$$\mathsf{P}(\boldsymbol{x} \in \Pi_i) = \pi_i, \qquad i = 1, 2,$$

la probabilidad a priori de que una observación $\boldsymbol{x}$ seleccionada al azar, pertenezca a Π_1 o Π_2.

Suponga que asociado a la i-ésima clase, existe una densidad

$$\mathsf{P}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{X} \in \Pi_i) = f_i(\boldsymbol{x}), \qquad i = 1, 2$$

(no se realizan supuestos sobre f_i)

Clasificación binaria

El teorema de Bayes, permite obtener la probabilidad a posteriori

$$p(\Pi_i|\boldsymbol{x}) = \mathsf{P}(\boldsymbol{X} \in \Pi_i|\boldsymbol{X} = \boldsymbol{x}) = \frac{f_i(\boldsymbol{x})\pi_i}{f_1(\boldsymbol{x})\pi_1 + f_2(\boldsymbol{x})\pi_2},$$

esto es la probabilidad de que el $\boldsymbol{x}$ observado pertenezca a Π_i, $i = 1, 2$.

Para un $\boldsymbol{x}$ dado, una estratégia de clasificación razonable es asignar $\boldsymbol{x}$ a la clase con mayor densidad a posterior.

La regla clasificadora de Bayes es: asignar $\boldsymbol{x}$ a Π_1 si

$$\frac{p(\Pi_1|\boldsymbol{x})}{p(\Pi_2|\boldsymbol{x})} > 1,$$

y asignamos $\boldsymbol{x}$ a Π_2 en caso contrario.

Observación:

Note que el clasificador de Bayes asigna $\boldsymbol{x}$ a Π_1 si

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \frac{\pi_2}{\pi_1},$$

y a Π_2 en caso contrario.

Análisis discriminante lineal gaussiano

Suponiendo que ambas densidades siguen una distribución normal multivariada, la regla de Bayes adopta la forma propuesta por Fisher (1936).

Suponga que $f_1(\cdot)$ es una $\mathsf{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ y $f_2(\dot{})$ es $\mathsf{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ con el supuesto adicional de homogeneidad, es decir $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$.

La razón de las dos densidades está dada por

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \frac{\exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)\}}{\exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)\}},$$

tomando logaritmos

$$\begin{aligned} \log \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &= -\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) + \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{\mu}}), \qquad \overline{\boldsymbol{\mu}} = \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2). \end{aligned}$$

Note además que

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2,$$

luego sigue que

$$\mathsf{L}(\boldsymbol{x}) = \log\left(\frac{f_1(\boldsymbol{x})\pi_1}{f_2(\boldsymbol{x})\pi_2}\right) = b_0 + \boldsymbol{b}^\top \boldsymbol{x},$$

es una función lineal de $\boldsymbol{x}$, donde

$$\begin{aligned} \boldsymbol{b} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ b_0 &= -\tfrac{1}{2}(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2) + \log(\pi_2/\pi_1). \end{aligned}$$

Regla de clasificación:

Si $\mathsf{L}(\boldsymbol{x}) > 0$, asignar $\boldsymbol{x}$ a Π_1 y en caso contrario asignar $\boldsymbol{x}$ a Π_2.

$\mathsf{L}(\boldsymbol{x})$ define un hiperplano que divide las dos clases. La regla anterior se conoce como Análisis discriminante lineal (LDA) gaussiano.

La función

$$\mathsf{U}(\boldsymbol{x}) = \boldsymbol{b}^\top \boldsymbol{x} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x},$$

se conoce como la función discriminante lineal de Fisher (LDF).

Probabilidad total de mala clasificación

La LDF particiona la población $\mathcal{P}$ en regiones de clasificación disjuntas R_1 y R_2. Si $\boldsymbol{x}$ se asigna a la región R_1 la misma es clasificada como perteneciente a Π_1 en otro caso $\boldsymbol{x}$ se asigna a R_2 y es clasificada en Π_2

Se puede cometer los siguientes errores (mala clasificación):

(a) Asignar $\boldsymbol{x}$ a Π_2 cuando pertenece a Π_1.

(b) Asignar $\boldsymbol{x}$ a Π_1 cuando pertenece a Π_2.

Defina

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

la distancia de Mahalanobis entre Π_1 y Π_2. Entonces

$$\mathsf{E}(\mathsf{U}|\boldsymbol{x} \in \Pi_i) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i,$$

y

$$\begin{aligned} \mathsf{var}(\mathsf{U}|\boldsymbol{x} \in \Pi_i) &= \boldsymbol{b}^\top \boldsymbol{\Sigma} \boldsymbol{b} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \Delta^2, \end{aligned}$$

para $i = 1, 2$.

Probabilidad total de mala clasificación

La probabilidad de mala clasificación total es:

$$\mathsf{P}(\Delta) = \mathsf{P}(\boldsymbol{x} \in R_2 | \boldsymbol{x} \in \Pi_1)\pi_1 + \mathsf{P}(\boldsymbol{x} \in R_1 | \boldsymbol{x} \in \Pi_2)\pi_2,$$

donde

$$\begin{aligned}
\mathsf{P}(\boldsymbol{x} \in R_2 | \boldsymbol{x} \in \Pi_1) &= \mathsf{P}(\mathsf{L}(\boldsymbol{x}) < 0 | \boldsymbol{x} \in \Pi_1) \\
&= \mathsf{P}\left(Z < -\frac{\Delta}{2} - \frac{1}{\Delta} \log \frac{\pi_2}{\pi_1}\right) \\
&= \Phi\left(-\frac{\Delta}{2} - \frac{1}{\Delta} \log \frac{\pi_2}{\pi_1}\right)
\end{aligned}$$

y análogamente

$$\begin{aligned}
\mathsf{P}(\boldsymbol{x} \in R_1 | \boldsymbol{x} \in \Pi_2) &= \mathsf{P}(\mathsf{L}(\boldsymbol{x}) > 0 | \boldsymbol{x} \in \Pi_2) \\
&= \mathsf{P}\left(Z > \frac{\Delta}{2} - \frac{1}{\Delta} \log \frac{\pi_2}{\pi_1}\right) \\
&= \Phi\left(-\frac{\Delta}{2} + \frac{1}{\Delta} \log \frac{\pi_2}{\pi_1}\right)
\end{aligned}$$

En el cálculo de estas probabilidades se usa el hecho que $\mathsf{L}(\boldsymbol{x}) = b_0 + \mathsf{U}(\boldsymbol{x})$ y se estandariza U como

$$Z = \frac{\mathsf{U} - \mathsf{E}(\mathsf{U}|\boldsymbol{x} \in \Pi_i)}{\sqrt{\mathsf{var}(\mathsf{U}|\boldsymbol{x} \in \Pi_i)}}.$$

Si $\pi_1 = \pi_2 = \frac{1}{2}$, entonces

$$\mathsf{P}(\boldsymbol{x} \in R_2|\boldsymbol{x} \in \Pi_1) = \mathsf{P}(\boldsymbol{x} \in R_1|\boldsymbol{x} \in \Pi_2) = \Phi(-\Delta/2),$$

y de ahí que

$$\mathsf{P}(\Delta) = 2\Phi(-\Delta/2).$$

Observación:

Note que $\mathsf{P}(\Delta) = 1$ cuando $\Delta = 0$ es decir, las dos poblaciones son idénticas y tiende a cero conforme Δ crece. Es decir, mientras mayor sea la distancia entre las dos medias poblacionales menos probable es clasificar erroneamente $\boldsymbol{x}$.