MAT-269: Análisis Discriminante Lineal

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

EX UMBRA IN SOLEM

Suponga observaciones multivariadas provenientes de g clases (o grupos) predefinidos[1] 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.

[1] Determinar los grupos será revisado más adelante, dentro de técnicas de agrupamiento (o clustering)

En esta clase consideraremos $(g = 2)$ clases o grupos y nuestro objetivo será contruir un único clasificador para diferenciar entre clases.

Suponga una población $\mathcal{P}$ particionada en 2 grupos denotados por Π_1 y Π_2. 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 2 clases.

Sea q_1 y q_2 las proporciones de individuos en Π_1 y Π_2, respectivamente.

Podemos tener 2 tipos de errores en la clasificación:

- Clasificar un individuo en Π_1, cuando pertenece a Π_2.
- Clasificar un individuo en Π_2, cuando pertenece a Π_1.

Esto lleva a los siguientes costos de clasificación:

- $C(2|1)$ costo de clasificar un individuo en Π_2 cuando realmente pertenece a Π_1.
- $C(1|2)$ costo de clasificar un individuo en Π_1 cuando realmente pertenece a Π_2.

Sea,

- $g_1(\boldsymbol{x})$ la densidad de $\boldsymbol{x}$ cuando un individuo pertenece a Π_1.
- $g_2(\boldsymbol{x})$ la densidad de $\boldsymbol{x}$ cuando un individuo pertenece a Π_2.

Suponga las regiones Ω_1 y $\Omega_2 \subset \mathbb{R}^p$ tal que $\Omega_1 \cup \Omega_2 = \mathbb{R}^p$ y $\Omega_1 \cap \Omega_2 = \varnothing$

De este modo la regla de decisión adopta la forma:

- Si $\boldsymbol{x} \in \Omega_1$, clasificamos $\boldsymbol{x}$ en Π_1.
- Si $\boldsymbol{x} \in \Omega_2$, clasificamos $\boldsymbol{x}$ en Π_2.

Objetivo:

Definir regiones Ω_1 y Ω_2 que minimicen los costos esperados de clasificación erronea.

Sea,

- $\mathsf{P}(2|1)$ probabilidad de clasificar erroneamente un individuo de Π_1 como perteneciente a Π_2.
- $\mathsf{P}(1|2)$ probabilidad de clasificar erroneamente un individuo de Π_2 como perteneciente a Π_1.

Tenemos,

$$\mathsf{P}(2|1) = \mathsf{P}\{\boldsymbol{x} \in \Omega_2 \text{ cuando } \boldsymbol{x} \sim g_1(\boldsymbol{x})\} = \int_{\Omega_2} g_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

y análogamente

$$\mathsf{P}(1|2) = \mathsf{P}\{\boldsymbol{x} \in \Omega_1 \text{ cuando } \boldsymbol{x} \sim g_2(\boldsymbol{x})\} = \int_{\Omega_1} g_2(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Suponga que la probabilidad de obtener una observación desde Π_1 es q_1 y análogamente la probabilidad de obtener una observación desde Π_2 es q_2. Además, $q_1 + q_2 = 1$.

De este modo, la probabilidad de que un individuo proveniente de Π_1 (Π_2) sea clasificado erroneamente es $q_1\,\mathsf{P}(2|1)$ $(q_2\,\mathsf{P}(1|2))$.

Esto lleva a la siguiente tabla:

Costo	$C(2\|1)$	$C(1\|2)$
Probabilidad	$q_1\,\mathsf{P}(2\|1)$	$q_2\,\mathsf{P}(1\|2)$

Luego

$$\begin{aligned}\mathsf{E}(\text{Clasificación}) &= C(2|1)q_1\,\mathsf{P}(2|1) + C(1|2)q_2\,\mathsf{P}(1|2)\\ &= C_1\int_{\Omega_2} g_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + C_2\int_{\Omega_1} g_2(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\end{aligned}$$

donde $C_1 = q_1C(2|1)$ y $C_2 = q_2C(1|2)$. Luego,

$$\begin{aligned}\mathsf{E}(\text{Clasificación}) &= C_1\left[1 - \int_{\Omega_2} g_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right] + C_2\int_{\Omega_1} g_2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= C_1 + \int_{\Omega_1}\{C_2g_2(\boldsymbol{x}) - C_1g_1(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Se desea determinar Ω_1 minimizando la integral:

$$\int_{\Omega_1}\{C_2g_2(\boldsymbol{x}) - C_1g_1(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x}$$

Por notar que $\Omega_1 = \{\boldsymbol{x} : C_1 g_1(\boldsymbol{x}) \geq C_2 g_2(\boldsymbol{x})\}$ minimiza la integral

$$\int_{\Omega_1} \{C_2 g_2(\boldsymbol{x}) - C_1 g_1(\boldsymbol{x})\} \mathrm{d}\boldsymbol{x}$$

Recuerde que:

$$I_{\Omega_1}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in \Omega_1, \\ 0, & \text{en otro caso.} \end{cases}$$

Lleva a escribir

$$\int_{\Omega_1} \{C_2 g_2(\boldsymbol{x}) - C_1 g_1(\boldsymbol{x})\} \mathrm{d}\boldsymbol{x} = \int I_{\Omega_1}(\boldsymbol{x}) \{C_2 g_2(\boldsymbol{x}) - C_1 g_1(\boldsymbol{x})\} \mathrm{d}\boldsymbol{x}.$$

Sea Ω_1^* otra región. Se desea mostrar que

$$\int_{\Omega_1^*}\{C_2 g_2(\boldsymbol{x}) - C_1 g_1(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x} - \int_{\Omega_1}\{C_2 g_2(\boldsymbol{x}) - C_1 g_1(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x} \geq 0$$

Lo que es equivalente a mostrar que

$$\int\{I_{\Omega_1^*}(\boldsymbol{x}) - I_{\Omega_1}(\boldsymbol{x})\}\{C_2 g_2(\boldsymbol{x}) - C_1 g_1(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x} \geq 0$$

Es decir, basta probar que:

$$\{I_{\Omega_1^*}(\boldsymbol{x}) - I_{\Omega_1}(\boldsymbol{x})\}\{C_2 g_2(\boldsymbol{x}) - C_1 g_1(\boldsymbol{x})\} \geq 0$$

Supongamos que $\{I_{\Omega_1^*}(\boldsymbol{x}) - I_{\Omega_1}(\boldsymbol{x})\} > 0$. Entonces, $I_{\Omega_1}(\boldsymbol{x}) = 0$. Esto implica que $\boldsymbol{x} \notin \Omega_1$.

Si

$$\{I_{\Omega_1^*}(\boldsymbol{x}) - I_{\Omega_1}(\boldsymbol{x})\} < 0,$$

entonces $I_{\Omega_1}(\boldsymbol{x}) = 1 \Rightarrow \boldsymbol{x} \in \Omega_1$. Esto implica que

$$\{C_2 g_2(\boldsymbol{x}) - C_1 g_1(\boldsymbol{x})\} \leq 0.$$

Por simplicidad asuma que $C(1|2) = C(2|1)$ y que $q_2 = q_1$. Entonces la condición anterior resulta,

$$g_2(\boldsymbol{x}) \leq g_1(\boldsymbol{x}) \qquad \text{o bien,} \qquad \frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})} \geq 1.$$

Así, finalmente podemos escribir la región que minimiza la integral deseada como:

$$\Omega_1 = \Big\{\boldsymbol{x} : \frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})} \geq 1\Big\}.$$

Ejemplo:

Considere $g_1(\boldsymbol{x}) \stackrel{\mathsf{d}}{=} \mathsf{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ y $g_2(\boldsymbol{x}) \stackrel{\mathsf{d}}{=} \mathsf{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$.

Entonces

$$\frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})} = \frac{\exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1))}{\exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2))}$$

Luego, la región Ω_1 está determinada por la condición

$$\log\Big(\frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})}\Big) \geq \log K, \qquad K = \frac{C(1|2)}{C(2|1)}\frac{q_2}{q_1}.$$

Ahora

$$\log\Big(\frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})}\Big) = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Por lo tanto,

$$\Omega_1 = \Big\{\boldsymbol{x} : \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq \log K\Big\}.$$

Si $C(1|2) = C(2|1)$ y $q_1 = q_2$, obtenemos la condición:

$$\Omega_1 = \Big\{\boldsymbol{x} : \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}.$$

La función $\mathsf{U}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ se denomina función discriminante lineal de Fisher.

Para evaluar $\mathsf{P}(2|1)$ y $\mathsf{P}(1|2)$ definamos la función:

$$\mathsf{L}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Entonces, de acuerdo al desarrollo anterior, la región Ω_1 queda determinada por la condición $\mathsf{U}(\boldsymbol{x}) \geq 0$.

Asumamos que $\mathsf{U}(\boldsymbol{x})$ tiene una distribución normal univariada. Entonces, si $\boldsymbol{x}$ pertenece a Π_1,

$$\begin{aligned}\mathsf{E}\{\mathsf{U}(\boldsymbol{x})\} &= \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}\Delta^2.\end{aligned}$$

Note que

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

es el cuadrado de la distancia de Mahalanobis entre las dos medias poblacionales.

Ahora note que

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

Luego, si $\boldsymbol{x}$ proviene de Π_1, entonces

$$\mathsf{U}(\boldsymbol{x}) \sim \mathsf{N}(\tfrac{1}{2}\Delta^2, \Delta^2).$$

Similarmente se puede probar que si $\boldsymbol{x}$ proviene de Π_2, entonces

$$\mathsf{U}(\boldsymbol{x}) \sim \mathsf{N}(-\tfrac{1}{2}\Delta^2, \Delta^2).$$

Finalmente para evaluar $\mathsf{P}(1|2)$ calculamos

$$\mathsf{P}(1|2) = \mathsf{P}\{\mathsf{U}(\boldsymbol{x}) \geq 0 \text{ cuando } \mathsf{U}(\boldsymbol{x}) \sim \mathsf{N}(-\tfrac{1}{2}\Delta^2, \Delta^2)\}.$$

- Si q_1 y q_2 no están disponibles, es posible derivar una regla de clasificación poniendo condiciones sobre $C(1|2), C(2|1), P(1|2)$, y $P(2|1)$.

- Por ejemplo, si ambas poblaciones son normales, digamos, $\mathsf{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ y $\mathsf{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ y la región Ω_1 es determinada por la condición $\mathsf{U}(\boldsymbol{x}) \geq c$.

- Podemos calcular

$$\begin{aligned}
\mathsf{P}(2|1) &= \mathsf{P}\{\mathsf{U}(\boldsymbol{x}) < c \text{ cuando } \boldsymbol{x} \sim \mathsf{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\} \\
&= \mathsf{P}\{\mathsf{U}(\boldsymbol{x}) < c \text{ cuando } \mathsf{U}(\boldsymbol{x}) \sim \mathsf{N}(\tfrac{1}{2}\Delta^2, \Delta^2)\} \\
&= \int_{-\infty}^{\frac{c-\Delta^2/2}{\Delta}} \frac{2}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}y^2) \mathrm{d}y.
\end{aligned}$$

  Luego podemos determinar una constante c tal que, por ejemplo, $\mathsf{P}(2|1) = 0.05$.

- Otro tipo de restricciones que pueden considerarse son, por ejemplo,

$$C(1|2)\,\mathsf{P}(1|2) = C(2|1)\,\mathsf{P}(2|1).$$

Resultado:

En el contexto del problema de minimización de la integral

$$\int_{\Omega_1} \{C_2 g_2(\boldsymbol{x}) - C_1 g_1(\boldsymbol{x})\} \mathrm{d}\boldsymbol{x},$$

supongamos que para otra región Ω_1^* que satisface

$$(I_{\Omega_1^*}(\boldsymbol{x}) - I_{\Omega_1}(\boldsymbol{x}))\{C_2 g_2(\boldsymbol{x}) - C_1 g_1(\boldsymbol{x})\} \geq 0$$

se tiene que Ω_1^* también minimiza el costo esperado. Entonces

$$\Omega_1^* = \Omega_1.$$

Demostración:

La demostración consiste en probar que $I_{\Omega_1^*}(\boldsymbol{x}) = I_{\Omega_1}(\boldsymbol{x})$.