Tema 8: Análisis Discriminante y Clasificación

Aurea Grané
Departamento de Estadística
Universidad Carlos III de Madrid

Análisis discriminante

Supongamos que tenemos g poblaciones conocidas $\Omega_1, \ldots, \Omega_g$ y en cada una de ellas observamos una muestra de cierto vector de interés $\mathbf{X} = (X_1, \ldots, X_p)'$.

El **análisis discriminante** se ocupa de describir, mediante las variables X_i, los rasgos diferenciales entre las poblaciones.

Se trata de encontrar *funciones discriminantes* o *reglas de decisión* $h = h(x_1, \ldots, x_p)$ cuyos valores en los distintos grupos (o poblaciones) estén lo más separados posible. Es decir, buscamos funciones h sencillas que permitan asignar cada uno de los individuos a una población concreta Ω_α, $\alpha = 1, \ldots, g$, minimizando la tasa de error en dicha asignación.

La más conocida es la regla discriminante lineal de Fisher, donde h es una función lineal de $\mathbf{x} = (x_1, \ldots, x_p)'$.

Clasificación

Dado un nuevo individuo ω, cuya población de procedencia se desconoce, y sobre el cual se pueden medir las variables, $X_1, \ldots, X_p$, es decir, $\mathbf{x} = (x_1, \ldots, x_p)'$, donde $x_i = X_i(\omega)$, para $i = 1, \ldots, p$, el **problema de clasificación** trata de asignar éste individuo a alguna de las poblaciones Ω_α, para $\alpha = 1, \ldots, g$. Para ello se utilizan las funciones discriminantes construidas a partir de la muestra.

Clasificación en dos poblaciones

1. Discriminador lineal

Sean $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ los vectores de medias de las poblaciones Ω_1, Ω_2, respectivamente. Sea $\boldsymbol{\Sigma}$ la matriz de covarianzas común para ambas poblaciones. Sea ω el individuo a clasificar, para el cual se ha observado $\mathbf{x} = (x_1, \ldots, x_p)'$

El *criterio geométrico*, consiste en asignar el individuo ω a la población más próxima, utilizando la distancia de Mahalanobis:

$$\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_i) = (\mathbf{x} - \boldsymbol{\mu}_i)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i), \quad i = 1, 2.$$

La regla de decisión es la siguiente:

ω se asigna a Ω_1 si $\delta_M^2(\mathbf{x}, \boldsymbol{\mu}_1) < \delta_M^2(\mathbf{x}, \boldsymbol{\mu}_2)$,

ω se asigna a Ω_2 en caso contrario.

A partir de la diferencia $\delta_M^2(\mathbf{x},\boldsymbol{\mu}_2)-\delta_M^2(\mathbf{x},\boldsymbol{\mu}_1)$, se construye la función discriminante lineal

$$L(\mathbf{x})=\left(\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right)'\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

y se expresa la regla de decisión en función de ésta:

ω se asigna a Ω_1 si $L(\mathbf{x})>0$,

en caso contrario, se asigna ω a Ω_2.

Esta función discriminante es el **discriminador lineal de Fisher**.

Ejercicio: Expresar $L(\mathbf{x})$ como una diferencia entre $\delta_M^2(\mathbf{x},\boldsymbol{\mu}_2)$ y $\delta_M^2(\mathbf{x},\boldsymbol{\mu}_1)$.

Clasificación cuando los parámetros son estimados

En las aplicaciones prácticas, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ y $\boldsymbol{\Sigma}$ son desconocidas y se deberán estimar a partir de muestras de tamaños n_1, n_2 de las dos poblaciones Ω_1 y Ω_2.

Sean $\overline{\mathbf{x}}_1$, $\overline{\mathbf{x}}_2$ y $\mathbf{S}_1$, $\mathbf{S}_2$ los vectores de medias y las matrices de covarianzas muestrales.

La versión muestral del discriminador lineal de Fisher es

$$\hat{L}(\mathbf{x})=\left(\mathbf{x}-\frac{1}{2}(\overline{\mathbf{x}}_1+\overline{\mathbf{x}}_2)\right)'\mathbf{S}_P^{-1}\,(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2),$$

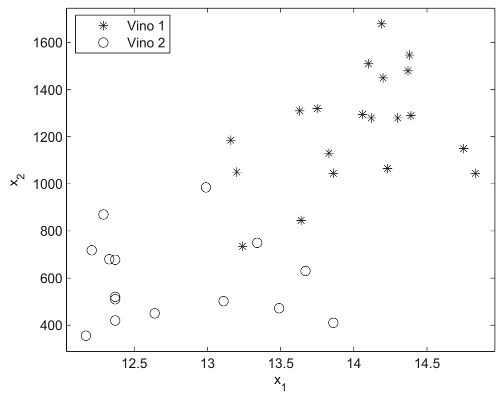
donde $\mathbf{S}_P=(n_1\,\mathbf{S}_1+n_2\,\mathbf{S}_2)/(n_1+n_2-2)$, es la *pooled within matrix*.

De forma análoga, ω se asigna a Ω_1 si $\hat{L}(\mathbf{x})>0$.

Ejemplo 1: Problema 9.4

Un enólogo analiza dos componentes X_1 y X_2 en sendas muestras de dos tipos de vinos (Datos de Newman *et al.*, 1998).

Vino 1		Vino 2	
X_1	X_2	X_1	X_2
14.23	1065	12.37	520
13.20	1050	12.33	680
13.16	1185	12.64	450
14.37	1480	13.67	630
13.24	735	12.37	420
14.20	1450	12.17	355
14.39	1290	12.37	678
14.06	1295	13.11	502
14.83	1045	12.37	510
13.86	1045	13.34	750
14.10	1510	12.21	718
14.12	1280	12.29	870
13.75	1320	13.86	410
14.75	1150	13.49	472
14.38	1547	12.99	985
13.63	1310		
14.30	1280		
13.83	1130		
14.19	1680		
13.64	845		



Clasificar mediante el discriminador lineal de Fisher la nueva observación $\mathbf{x}=(13.05,515)'$.

Calculamos los vectores de medias y las matrices de covarianzas muestrales. Para el primer tipo de vino:

$$\overline{\mathbf{x}}_1 = (14.0115, 1234.6000)', \quad \mathbf{S}_1 = \begin{pmatrix} 0.2115 & 42.6801 \\ 42.6801 & 52947.0400 \end{pmatrix},$$

y para el segundo tipo de vino:

$$\overline{\mathbf{x}}_2 = (12.7720, 596.6667)', \quad \mathbf{S}_2 = \begin{pmatrix} 0.3400 & -6.4900 \\ -6.4900 & 33019.9524 \end{pmatrix}.$$

La matriz de covarianzas ponderada es:

$$\mathbf{S}_p = (20\,\mathbf{S}_1 + 15\,\mathbf{S}_2)/33 = \begin{pmatrix} 0.2827 & 22.9167 \\ 22.9167 & 47098.1844 \end{pmatrix}$$

El valor del discriminador lineal de Fisher para la nueva observación $\mathbf{x} = (13.05, 515)'$ es:

$$\hat{L}(\mathbf{x}) = \left(\mathbf{x} - \tfrac{1}{2}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)\right)'\, \mathbf{S}_p^{-1}\, (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) =$$

$$\left(\begin{pmatrix} 13.05 \\ 515 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 26.7835 \\ 1831.2667 \end{pmatrix} \right)' \begin{pmatrix} 0.2827 & 22.9167 \\ 22.9167 & 47098.1844 \end{pmatrix}^{-1} \begin{pmatrix} 1.2395 \\ 637.9333 \end{pmatrix}$$

$$= -5.9288.$$

Puesto que $\hat{L}(\mathbf{x}) < 0$, asignaremos la nueva observación al segundo tipo de vino.

Observad que nos hemos creído una hipótesis que no hemos comprobado ¿cuál es? ¿qué contraste deberíamos realizar?

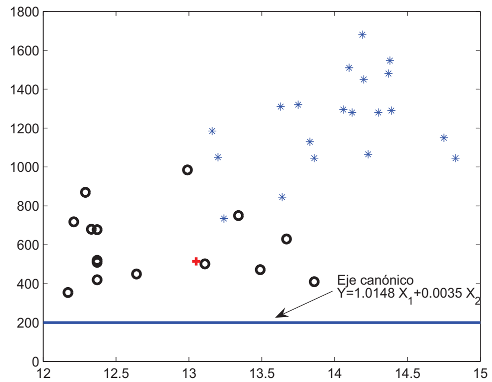
Criterio geométrico: la regla del discriminador lineal de Fisher equivale a asignar el nuevo individuo a la población más próxima según la distancia de Mahalanobis:

$$\delta_M^2(\mathbf{x}, \overline{\mathbf{x}}_i) = (\mathbf{x} - \overline{\mathbf{x}}_i)'\mathbf{S}_p^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i), \quad i = 1, 2.$$

En nuestro caso, tenemos que:

$$\delta_M^2(\mathbf{x}, \overline{\mathbf{x}}_1) = 12.3709, \quad \delta_M^2(\mathbf{x}, \overline{\mathbf{x}}_2) = 0.5134.$$

Puesto que $\delta_M^2(\mathbf{x}, \overline{\mathbf{x}}_2) < \delta_M^2(\mathbf{x}, \overline{\mathbf{x}}_1)$, la nueva observación se asigna al segundo tipo de vino.



2. Regla de máxima verosimilitud

Sean Ω_1 y Ω_2 dos poblaciones y $\mathbf{X} = (X_1, \ldots, X_p)'$ un vector con distribución de probabilidad conocida, dependiente de un parámetro $\boldsymbol{\theta}$ que toma el valor $\boldsymbol{\theta}_1$ si $\mathbf{X} \in \Omega_1$ y $\boldsymbol{\theta}_2$ si $\mathbf{X} \in \Omega_2$.

Sea $\mathbf{x} = (x_1, \ldots, x_p)'$ el vector de observaciones de $\mathbf{X}$ sobre un individuo ω. La probabilidad o verosimilitud de la observación $\mathbf{x}$ en Ω_i es $\mathcal{L}_i(\mathbf{x}) = f(x_1, \ldots, x_p; \boldsymbol{\theta}_i)$.

La regla discriminante de máxima verosimilitud consiste en asignar ω a la población Ω_i para la cual la verosimilitud de la observación es mayor. Esta regla tiene asociada la siguiente función discriminante

$$V(\mathbf{x}) = \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}).$$

Si $V(\mathbf{x}) > 0$, ω se asigna a Ω_1. En caso contrario, ω se asigna a Ω_2.

3. Regla de Bayes

En ciertas situaciones, se conocen las probabilidaddes *a priori* de que ω pertenezca a cada una de las poblaciones:

$$q_1 = P(\omega \in \Omega_1), \quad q_2 = P(\omega \in \Omega_2), \quad q_1 + q_2 = 1.$$

Una vez se dispone de las observaciones $\mathbf{x} = (x_1, \ldots, x_p)'$, la probabilidad *a posteriori* de que ω pertenezca a las poblaciones (teorema de Bayes) es:

$$P(\omega \in \Omega_i | \mathbf{x}) = \frac{q_i\, \mathcal{L}_i(\mathbf{x})}{q_1\, \mathcal{L}_1(\mathbf{x}) + q_2\, \mathcal{L}_2(\mathbf{x})}, \quad i = 1, 2.$$

La regla discriminante de Bayes consiste en asignar ω a la población Ω_i para la que $P(\omega \in \Omega_i | \mathbf{x})$ es máxima.

La regla de Bayes tiene asociada la siguiente función discriminante, que se conoce como **discriminador de Bayes**:

$$B(\mathbf{x}) = \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}) + \log(q_1/q_2),$$

Si $B(\mathbf{x}) > 0$, ω se asigna a Ω_1 y, en caso contrario, se asigna ω a Ω_2.

Propiedades:

1. Cuando $q_1 = q_2 = 1/2$, entonces $B(\mathbf{x}) = V(\mathbf{x})$.
2. La regla de Bayes minimiza la probabilidad de clasificación errónea.

Clasificación en poblaciones normales

Supongamos ahora que:
$\mathbf{X} = (X_1, \ldots, X_p)' \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ en Ω_1
$\mathbf{X} = (X_1, \ldots, X_p)' \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ en Ω_2,
es decir:

$$\mathcal{L}_i(\mathbf{x}) = \frac{|\boldsymbol{\Sigma}_i|^{-1/2}}{(2\,\pi)^{p/2}} \exp\left\{ -\frac{1}{2}\,(\mathbf{x} - \boldsymbol{\mu}_i)'\,\boldsymbol{\Sigma}_i^{-1}\,(\mathbf{x} - \boldsymbol{\mu}_i) \right\}, \quad i = 1, 2.$$

Caso 1: Si $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ y $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, entonces:

a) Los clasificadores de máxima verosimilitud y lineal coinciden:

$$\begin{aligned} V(\mathbf{x}) &= \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}) \\ &= \frac{1}{2}\Big((\mathbf{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) - (\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) \Big) \\ &= L(\mathbf{x}). \end{aligned}$$

b) Si $\mathbf{x} \in \mathbb{R}^p$ proviene de alguna de las poblaciones Ω_i, para $i = 1, 2$, entonces el discriminador lineal de Fisher tiene distribución normal:

$$L(\mathbf{x}) = \left(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = (\mathbf{x} - \boldsymbol{\mu})' \mathbf{a} = \mathbf{a}' (\mathbf{x} - \boldsymbol{\mu}),$$

donde $\mathbf{a} = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ y $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$.

Su varianza y esperanza son:
$\mathsf{var}(L(\mathbf{x})) = \mathsf{var}(\mathbf{a}' (\mathbf{x} - \boldsymbol{\mu})) = \mathbf{a}' \boldsymbol{\Sigma} \mathbf{a} = M^2$,

$$\mathsf{E}(L(\mathbf{x})) = \mathbf{a}' \, \mathsf{E}(\mathbf{x} - \boldsymbol{\mu}) = \begin{cases} \frac{1}{2} \mathbf{a}' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2} M^2, & \text{si } \mathbf{x} \in \Omega_1, \\ -\frac{1}{2} \mathbf{a}' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = -\frac{1}{2} M^2, & \text{si } \mathbf{x} \in \Omega_2, \end{cases}$$

Por tanto,

$$L(\mathbf{x}) \sim N\Big(\frac{1}{2}M^2, M^2\Big) \text{ si } \mathbf{x} \in \Omega_1, L(\mathbf{x}) \sim N\Big(-\frac{1}{2}M^2, M^2\Big) \text{ si } \mathbf{x} \in \Omega_2.$$

Puesto que $L(\mathbf{x})$ tiene distribución de probabilidad conocida, puede calcularse la **probabilidad de clasificación errónea**.

Se dice que el individuo $\mathbf{x}$ se clasifica erróneamente cuando se asigna a la población Ω_1 y en realidad proviene de Ω_2, o bien, cuando se asigna a la población Ω_2 y en realidad proviene de Ω_1.

Luego la probabilidad de clasificación errónea es:

$$\frac{1}{2} P\left(L(\mathbf{x}) > 0 | \mathbf{x} \in \Omega_2\right) + \frac{1}{2} P\left(L(\mathbf{x}) < 0 | \mathbf{x} \in \Omega_1\right) = \Phi\left(-\frac{M}{2}\right),$$

donde Φ es la función de distribución de la ley $N(0, 1)$.

c) Si conocemos las probabilidades *a priori* $q_1 = P(\omega \in \Omega_1)$, $q_2 = P(\omega \in \Omega_2)$, con $q_1 + q_2 = 1$, entonces el discriminador de Bayes es: $B(\mathbf{x}) = L(\mathbf{x}) + \log(q_1/q_2)$.

Caso 2: Si $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ y $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, entonces:

a) La regla de máxima verosimilitud proporciona el **discriminador cuadrático**:

$$\begin{aligned} V(\mathbf{x}) &= \log \mathcal{L}_1(\mathbf{x}) - \log \mathcal{L}_2(\mathbf{x}) \\ &= \frac{1}{2} \mathbf{x}'(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})\mathbf{x} + \mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2) \\ &+ \frac{1}{2} \boldsymbol{\mu}_2' \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2 - \frac{1}{2} \boldsymbol{\mu}_1' \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \log|\boldsymbol{\Sigma}_2| - \frac{1}{2} \log|\boldsymbol{\Sigma}_1| \\ &= Q(\mathbf{x}), \end{aligned}$$

b) Si conocemos las probabilidades *a priori* $q_1 = P(\omega \in \Omega_1)$, $q_2 = P(\omega \in \Omega_2)$, con $q_1 + q_2 = 1$, entonces el discriminador de Bayes es: $B(\mathbf{x}) = Q(\mathbf{x}) + \log(q_1/q_2)$.

Clasificación en $g \geq 2$ poblaciones

- **Criterio geométrico**. Se asigna el nuevo individuo a la población más próxima según la distancia de Mahalanobis,

$$\min_{1 \leq i \leq g} \delta_M^2(\mathbf{x}, \boldsymbol{\mu}_i).$$

- **Máxima verosilimitud**. Se asigna el nuevo individuo a la población con mayor verosimilitud,

$$\max_{1 \leq i \leq g} \mathcal{L}_i(\mathbf{x}) = \max_{1 \leq i \leq g} f(x_1, \ldots, x_p; \boldsymbol{\theta}_i).$$

- **Regla de Bayes**. Se asigna el nuevo individuo a la población con mayor probabilidad *a posteriori*,

$$\max_{1 \leq i \leq g} P(\omega \in \Omega_i | \mathbf{x}) = \max_{1 \leq i \leq g} \frac{q_i \mathcal{L}_i(\mathbf{x})}{\sum_{i=1}^g q_i \mathcal{L}_i(\mathbf{x})},$$

donde $q_i = P(\omega \in \Omega_i)$ son las *a priori* con $\sum_{i=1}^g q_i = 1$.