Tema 33

1. Análisis Discriminante.
2. Clasificación con 2 grupos.
3. Función discriminante de Fisher.
4. Clasificación con más de 2 grupos.
5. Funciones Clasificadoras.

1. ANÁLISIS DISCRIMINANTE

El Análisis Discriminante es una técnica multivariante de dependencia que permite clasificar a distintos individuos en grupos a partir de un conjunto de medidas sobre ellos, representadas a través de una serie de variables.

Cada individuo puede pertenecer a un sólo grupo, estos grupos son internamente homogéneos.

La pertenencia a uno u otro grupo se introduce en el análisis mediante una variable categórica que toma tantos valores como grupos existentes, esta variable es la variable dependiente (var. explicada).

Las variables que se utilizan para clasificar a los individuos se las conoce como variables clasificadoras, estas variables son las var. explicativas o predictoras.

(=~~var. indep~~

La información de las var. clasificadoras se sintetiza en las funciones discriminantes a partir de combinaciones lineales entre las variables, estas funciones permiten discriminar o identificar los grupos definidos por la variable dependiente (categórica).

El análisis discriminante se aplica para fines explicativos y/o predictivos.

- f. explicativos: determina la contribución de cada variable clasificadora a la clasificación correcta de cada uno de los individuos

- F. Predictivas: determina el grupo al que pertenece un individuo para el que se conocen los valores que toman las variables clasificadoras

Según el número de categorías de la variable dependiente, el análisis discriminante puede ser:

- Análisis simple o de dos grupos cuando la variable dependiente tiene dos categorías
- Análisis múltiple si tiene más de dos categorías.

Supuestos fundamentales en A.D.

Para que las conclusiones del A.D sean fiables, deben cumplirse una serie de condiciones (hipótesis)

1. La matriz de covarianzas intragrupo debe ser la misma (Σ) en todos los grupos. Hipótesis de homocedasticidad.
Para comprobar si las matrices de covarianzas intragrupo son iguales se utiliza el contraste de Box: (Barlett-Box)

$$H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_6$$

Para minimizar los efectos cuando no se cumple este supuesto, hay que utilizar tamaños muestrales elevados y matrices intragrupo específicas para clasificar

2.- Cada uno de los grupos ha de ser una muestra procedente de una población con distribución normal multivariante.
Para comprobar esta hipótesis se pueden utilizar test de carácter gráfico, o examinar las distribuc. de cada una de las variables clasificadoras, si cada var. se distribuye como una normal entonces la variable conjunta se dist como una normal multiv.
Cuando no se cumple esta hipótesis, los test de significación no son fiables por lo que habría que utilizar otra técnica de análisis multivariable como la regresión logística.

$$1+2 \Rightarrow X_g \overset{d}{\sim} N(\mu_g, \Sigma) \quad , \quad X_g = (X_1, X_2, \ldots, X_g)$$

3.- No debe existir multicolinealidad entre las variables clasificadoras.
El incumplimiento de esta hipótesis no supone un problema si su presencia es similar en todas las posibles muestras.

4.- Se supone que se ha extraído una muestra aleatoria multivariante en cada uno de los grupos.
Debe haber un mínimo de 20 observaciones en cada grupo por cada variable clasificadora.

5.- Las medias poblacionales de los grupos deben de ser significativamente distintas.
El contraste de esta hipótesis se puede realizar mediante:

- En el caso de dos grupos, el estadístico T^2 de Hotelling (generalización del estadístico t de Student).

$$H_0: \mu_I = \mu_{II}$$
$$H_1: \mu_I \neq \mu_{II}$$

$$T^2 = (\bar{y}_1 - \bar{y}_2)' \, S^{-1} (\bar{y}_1 - \bar{y}_2) \left(\frac{n_1 n_2}{n_1 + n_2} \right)$$

donde:

$$\frac{n_1 + n_2 - K - 1}{K} \cdot \frac{T^2}{n_1 + n_2 - 2} \overset{d}{\sim} F_{K, n_1 + n_2 - K - 1}$$

- En el caso de g ($g > 2$) grupos, el estadístico de Rao o el estadístico V de Barlett, construidos a partir de la Λ de Wilks

$$\Lambda = \frac{SCD}{STC} = \frac{SCD}{SCE + SCD}$$

Consideraciones generales del A.D:

- las variables clasificadoras deben ser variables cuantitativas. Cuando se utilizan variables categóricas, la función discriminante lineal de Fisher no tiene el carácter de óptima.

- Es importante tener en cuenta los tamaños de los grupos, ya que los grupos mayores tienen una probabilidad más alta de clasificar casos por azar.

- El A.D. es sensible a la presencia de casos aislados, datos atípicos, por eso conviene detectarlos.
- El A.D. es sensible a la relación entre el número de casos y el número de variables, siendo recomendable

$$\frac{\text{Nº casos}}{\text{Nº variables}} > 10 \quad (\text{se puede admitir} > 5)$$

Selección de variables (Ampliado E. Uriel y J. Aldás, Thomson Análisis Multivariante Aplicado pág 298-299)

Cuando se dispone de un número elevado de variables potencialmente discriminantes es necesario aplicar un sistema que permita seleccionar las variables con mayor capacidad clasificadora.

Métodos:

1. Selección hacia adelante (forward): la variable que entra en el modelo es la que más contribuye a discriminar entre grupos. Finaliza cuando no hay variables que contribuyan significativamente a discriminar
2. Selección hacia atrás (backward): inicialmente todas las variables pertenecen al modelo y va saliendo la variable que menos contribuye a la discriminación
3. Selección paso a paso (stepwise): es una combinación de los dos métodos anteriores, las var. pueden salir o entrar en cualquiera de las etapas.

Se fija: F-min-para-entrar > F-max-para-salir

Nivel de tolerancia = $1 - r_i^2 > 0'0001$

↳ coef. determ. entre x_i y resto de var.

Estadístico a minimizar ≡ Regla de decisión.

→ Regla de decisión: minimizar un estadístico que puede ser:
- Λ de Wilks
- Distancia de Mahalanobis
- V de Rao

↳ Mide la bondad del ajuste en cada paso.

2. CLASIFICACIÓN EN DOS GRUPOS

Planteamiento general del problema

Sean P_1 y P_2 dos poblaciones donde tenemos definida una v.a. vectorial x, p-variante. Suponemos que x es absolutamente continua y que f_1 y f_2 (fc's densidad) son conocidas. Vamos a estudiar el problema de clasificar un nuevo elemento, x_0, con valores conocidos de las p variables en una de las dos poblaciones.

Sean Π_1, Π_2 las prob. a priori de que x_0 venga de cada una de las poblaciones, con $\Pi_1 + \Pi_2 = 1$. Su distrib. de prob. será:

$$f(x) = \Pi_1 f_1(x) + \Pi_2 f_2(x)$$

Una vez observado x_0, podemos calcular las prob. a posteriori de que x_0 haya sido generado por cada una de las poblaciones, $P(i/x_0)$ con $i=1,2$.

Por el Tª de Bayes:

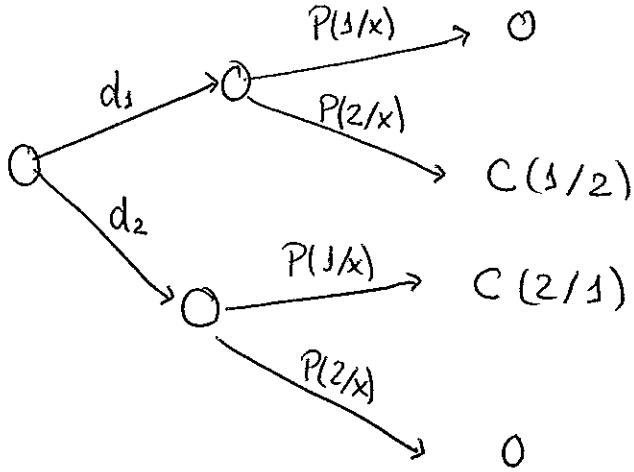
$$P(i/x_0) = \frac{P(x_0/i)\,\Pi_i}{\sum_{i=1}^{2} P(x_0/i)\,\Pi_i} \underset{P(x_0/i) = f_i(x_0)\,\Delta x_0}{=} \frac{f_i(x_0)\,\Pi_i}{\sum_{i=1}^{2} f_i(x_0)\,\Pi_i}$$

$\Rightarrow$ Clasificamos x_0 en P_2 si: $\Pi_2 f_2(x_0) > \Pi_1 f_1(x_0)$

Si $\Pi_1 = \Pi_2 \Rightarrow f_2(x_0) > f_1(x_0)$

Es decir, clasificamos x_0 en la población más probable

Si las consecuencias de un error de clasificación pueden cuantificarse, podemos incluirlas en la solución del problema formulándolo como un prob. bayesiano de decisión.



la mejor decisión es la que minimiza los costes esperados:

$$E[d_1] = 0\ P(1/x_0) + C(1/2)\ P(2/x_0)$$

$$E[d_2] = C(2/1) \cdot P(1/x_0) + 0 \cdot P(2/x_0)$$

$$x_0 \in P_2 \iff \boxed{\frac{f_2(x_0)\,\pi_2}{C(2/1)} > \frac{f_1(x_0)\,\pi_1}{C(1/2)}}\ *$$

Esta condición indica que, a igualdad de los otros términos, clasificaremos en la población P_2 si:

a) Su probabilidad a priori es más alta.

b) La verosimilitud de que x_0 provenga de P_2 es más alta

c) El costo de clasificarlo en P_2 es más bajo

* Este criterio es equivalente a minimizar la probab. total de error en la clasificación:

$$P_T(\text{error}) = P(1/x \in 2) + P(2/x \in 1) \quad , \quad P(i/x \in j) = \int_{A_i} f_j(x)\,dx$$

Poblaciones normales

Sup. f_1. f_2 distrib. normales con distintos vectores de medias e igual matriz de var. cov.

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\,|V|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)' V^{-1} (x-\mu_i)\right\}$$

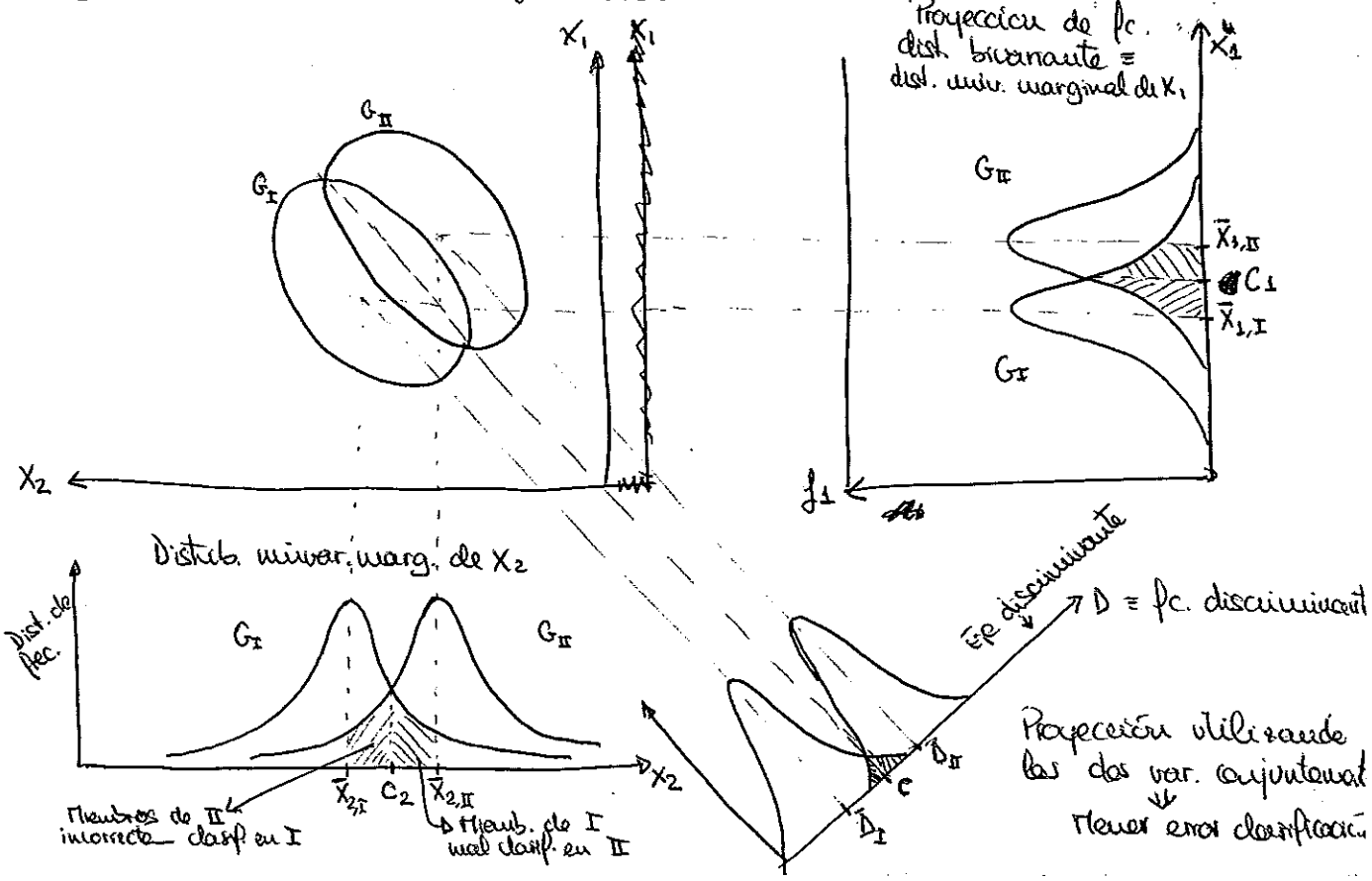
$$x \in P_2 \iff \frac{f_2(x)\,\pi_2}{C(2/1)} > \frac{f_1(x)\,\pi_1}{C(1/2)} \iff \text{(tomando logaritmos)}$$

$$\iff -\frac{1}{2}\underbrace{(x-\mu_2)' V^{-1}(x-\mu_2)}_{D_2^2} + \log\frac{\pi_2}{C(2/1)} > -\frac{1}{2}\underbrace{(x-\mu_1)' V^{-1}(x-\mu_1)}_{D_1^2} + \log\frac{\pi_1}{C(1/2)}$$

D_2^2 = distancia de Mahalanobis entre x y μ_2

$$\iff \boxed{D_1^2 - \log\frac{\pi_1}{C(1/2)} > D_2^2 - \log\frac{\pi_2}{C(2/1)}}$$

Representación gráfica de la clasificación con dos grupos y dos variables clasificadoras



3.- FUNCIÓN DISCRIMINANTE DE FISHER

El caso de dos grupos

La función lineal discriminante para dos grupos fue deducida por primera vez por Fisher en 1936 por un procedimiento intuitivo: encontrar una variable escalar tal que maximice la distancia entre las medias proyectadas en relación a la variabilidad resultante en la proyección.

La fc. discriminante de Fisher D se obtiene como función lineal de p variables explicativas X:

$$D = u_1 X_1 + u_2 X_2 + \dots + u_p X_p$$

La fc. discriminante para las n observaciones:

$$D_i = u_1 X_{1i} + u_2 X_{2i} + \dots + u_p X_{pi}$$

$$d_{n\times 1} = X_{n\times p} u_{p\times 1}$$

donde $D_i \equiv$ puntuación discriminante correspondiente a la observación i-ésima.

La variabilidad de la función discriminante se puede expresar

$$d'd = u' \underbrace{X'X} u$$

Matriz de suma de cuadrados y productos cruzados (SCPC) total de las variables X.

$$X'X = T = F + W$$

$$F = \sum_{g=1}^{G} n_g (\bar{X}_g - \bar{X}_T)(\bar{X}_g - \bar{X}_T)' \equiv \text{SCPC entre-grupos}$$

$$W = \sum_{g=1}^{G} \sum_{j=1}^{n_g} (X_{jg} - \bar{X}_g)(X_{jg} - \bar{X}_g)' \equiv \text{SCPC residual o intra-grupos}$$

$$T = \sum_{g=1}^{G} \sum_{j=1}^{n_g} (X_{jg} - \bar{X}_T)(X_{jg} - \bar{X}_T)' \equiv \text{SCPC total}$$

$d'd = u'Tu = u'Fu + u'Wu$

donde T, F, W se pueden calcular con los datos muestrales

Para estimar los coeficientes u_i, Fisher utilizó el siguiente criterio:

Maximización de $\frac{\text{Variabilidad entre-grupos}}{\text{Variabilidad intra-grupos}}$

En este criterio se obtiene el eje discriminante de forma que las distribuciones proyectadas sobre el mismo estén lo más separadas posible entre sí y, al mismo tiempo, que cada una de las distribuciones esté lo menos dispersa posible.

El criterio de Fisher expresado de forma analítica:

$$\text{Max } \lambda = \frac{u'Fu}{u'Wu}$$

Solución:

$$\frac{\partial \lambda}{\partial u} = 0 \Leftrightarrow \frac{2Fu(u'Wu) - 2Wu(u'Fu)}{(u'Wu)^2} = 0 \Leftrightarrow$$

$$\Leftrightarrow \frac{2Fu}{2Wu} = \frac{u'Fu}{u'Wu} = \lambda \Leftrightarrow Fu = Wu\lambda$$

⇒ Ecuación para la obtención del eje discriminante:

$$W^{-1}Fu = \lambda u$$

Los centros de gravedad o centroides $(\bar{X}_I, \bar{X}_{II})$ se proyectan en el eje discriminante $(\bar{D}_I, \bar{D}_{II})$:

$$\bar{D}_I = u_1 \bar{X}_{1,I} + u_2 \bar{X}_{2,I} + \ldots + u_p \bar{X}_{p,I}$$
$$\bar{D}_{II} = u_1 \bar{X}_{1,II} + u_2 \bar{X}_{2,II} + \ldots + u_p \bar{X}_{p,II}$$

El punto de corte discriminante C se calcula promediando $\bar{D}_I$ y $\bar{D}_{II}$:

$$C = \frac{\bar{D}_I + \bar{D}_{II}}{2}$$

Si $D_i < C \Rightarrow$ se clasifica el individuo i en grupo I
Si $D_i > C \Rightarrow$ " " " " " " grupo II

En general, cuando se aplica el análisis discriminante, la función discriminante viene dada por:

$$D - C = u_1 X_1 + \dots + u_p X_p - C$$

En ocasiones se dispone de información de la probabilidad a priori sobre pertenencia de un individuo a cada uno de los grupos, así como el coste que una clasificación errónea puede tener. El punto de corte discriminante $C_{p.c}$ que se obtiene en este caso es:

$$C_{p.c} = \frac{\bar{D}_I + \bar{D}_{II}}{2} - \ln \frac{\Pi_{II} \cdot \text{Coste}(I/II)}{\Pi_I \cdot \text{Coste}(II/I)}$$

→ Como medida de evaluación de la bondad del ajuste se utiliza el coeficiente de determinación obtenido al realizar la regresión entre la variable categórica (dicotómica) y las puntuaciones discriminantes. A la raíz cuadrada de este coeficiente se le denomina correlación canónica. Una expresión alternativa de la correl. canónica es:

$$\eta = \sqrt{\frac{\lambda}{1+\lambda}}$$

$\lambda \equiv$ Lambda de Wilks

4.- CLASIFICACIÓN CON MÁS DE DOS GRUPOS

El enfoque de Fisher puede generalizarse para encontrar variables canónicas que tengan un máximo poder discriminante para clasificar nuevos elementos entre G poblaciones:

1.- Se definen $z = (z_1 \dots z_r)'$ de r variables canónicas donde $r = \min(G-1, p)$ t.q. $z_i = u_i' X$.

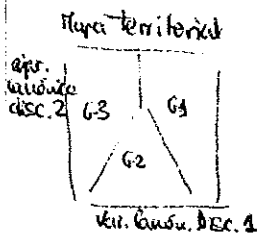
2.- Se proyectan las medias de las variables de los grupos, $\bar{x}_g$, sobre el espacio determinado por las r var. canónicas. ($\bar{z}_1, \dots \bar{z}_g$ son las ver. $r \times 1$ cuyas coord. son estas proyecciones).

3.- Se proyecta x_0 (pto a clasificar) t.q z_0 es su proyección.

4.- Se clasifica el pto en aquella población cuya media se encuentre más próxima.

$$x_0 \in P_i \Leftrightarrow (z_0 - \bar{z}_i)'(z_0 - \bar{z}_i) = \min_g (z_0 - \bar{z}_g)'(z_0 - \bar{z}_g)$$

(ej)



Obtención de las funciones discriminantes

$$D_j = u_{j1} X_1 + u_{j2} X_2 + \dots + u_{jp} X_p \quad \text{con } j = 1, 2, \dots, G-1 \ (\min(G-1, p))$$

los G-1 ejes discriminantes vienen def. por $u_1, u_2, \dots u_{G-1}$:

$$u_1 = \begin{pmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{pmatrix}; \quad u_2 = \begin{pmatrix} u_{21} \\ u_{22} \\ \vdots \\ u_{2p} \end{pmatrix}; \dots; \quad u_{G-1} = \begin{pmatrix} u_{G-1,1} \\ u_{G-1,2} \\ \vdots \\ u_{G-1,p} \end{pmatrix}$$

El criterio para obtener los ejes discriminantes es:

$$\text{Maximización de } \lambda_j = \frac{u_j' F u_j}{u_j' W u_j} \qquad j = 1, 2, \ldots, G-1$$

$$\Downarrow$$

$$W^{-1} F u_j = \lambda_j u_j \quad \text{ecuación para la obtención de eje disc. } j.$$

Obsrv: $W^{-1}F$ no es simétrica $\Rightarrow$ en general, los ejes discriminantes no serán ortogonales

5.- FUNCIONES CLASIFICADORAS

Existe una forma alternativa a la utilización de la función discriminante, que consiste en construir fc's discriminantes para cada grupo $F_I, F_{II}, \ldots F_G$. Se clasifica a un individuo en el grupo para el que la función F_j sea mayor.

$$F_j = a_{j1} X_1 + a_{j2} X_2 + \ldots + a_{jp} X_p - C_j$$

En el caso de dos grupos ($j = I, II$), se pueden obtener los coeficientes de la fc. discriminante:

$$F_{II} - F_I = (a_{II1} - a_{I1}) X_1 + \ldots + (a_{IIp} - a_{Ip}) X_p - (C_{II} - C_I) =$$

$$= u_1 X_1 + \ldots + u_p X_p - C = D - C$$

→ Si se dispone de información a priori:

$$i \in I \iff F_I \ln \Pi_I > F_{II} \ln \Pi_{II}$$

CRITERIOS ALTERNATIVOS DE CLASIFICACIÓN

Análisis de regresión

Si se realiza un ajuste por mínimos cuadrados, tomando como var dependiente la var. categórica (dicotómica) y como var. explicativas las var. clasificadoras, se obtienen unos coeficientes que guardan una estricta proporcionalidad con los coeficientes de la fc. discrim. de Fisher.
A partir del coef de determinación de la regresión, se puede pasar fácilmente a la distancia de Mahalanobis entre la centroide de los grupos
(Ver "Análisis de datos multivariante" de David Peña)

Distancia de Mahalanobis

$$D_{ij}^2 = (x_i - x_j)' V^{-1} (x_i - x_j) \quad ; \; j = I, II$$

Asigna cada individuo al grupo para el que la distancia de Mahalanobis es menor.
La dist. de Mahalanobis clasifica a los individuos exactamente igual que lo hace la fc. discrim. de Fisher.
La diferencia es que, mientras la dist. de Mahalan. se calcula en el espacio de las var. originales, en el criterio de Fisher se sustituyen todas las var en la fc. discriminante, que es la utilizada para realizar la clasificación.

Discriminación en poblaciones no normales

Las funciones discriminantes no son lineales, son cuadráticas y el número de parámetros a estimar es mucho mayor, lo que hace que la discriminación cuadrática sea bastante inestable (excepto en muestras muy grandes).

Para poblaciones arbitrarias existen dos alternativas:

- Aplicar la teoría general y obtener la fc. discriminante que puede ser complicada
- Utilizar como medida de distancia la distancia de Mahalanobis.

Para poblaciones discretas estas aproximaciones no son buenas y existen métodos alternativos basados en la dist. multinomial o en la distancia X^2.

Discriminación en poblaciones con matrices de var-cov distintas

En este caso la discriminación es cuadrática y es bastante inestable por lo que se obtienen generalmente, mejores resultados con la función lineal que con la cuadrática.

OBSERVACIONES

Cálculo de probabilidades de pertenencia a una población

Además de la clasificación de un individuo a un grupo, es interesante tener información sobre la probabilidad de su pertenencia a cada grupo, ya que ello permite realizar análisis más matizados.

- Prob. a posteriori sin información a priori:

$$P(g/D) = \frac{e^{F_g}}{e^{F_I} + e^{F_{II}}} \qquad g = I, II$$

↓ g F_I, F_{II} fc's de clasificación

- Prob a posteriori con información a priori:

$$P(g/D) = \frac{\Pi_g e^{F_g}}{\Pi_I e^{F_I} + \Pi_{II} e^{F_{II}}} \qquad g = I, II$$

Criterio de clasif. con fc's de clasif.: $F_I \ln \Pi_I > F_{II} \ln \Pi_{II}$

$\Rightarrow i \in G_I$

- Coste total de clasificación errónea:

$$\Pi_I \cdot P(II/I) \cdot \text{Coste}(II/I) + \Pi_{II} \cdot P(I/II) \cdot \text{Coste}(I/II)$$

Cálculo de probabilidades de error

$$\text{Error} = \frac{\text{Total mal clasificados}}{\text{Total bien clasificados}}$$

Este método subestima la prob de error ya que los mismos datos se utilizan para estimar los parámetros y para evaluar la regla resultante.

Mejor utilizar la validación cruzada (dejando 1 fuera).