El Analisis Discriminante

Gilbert Saporta

El Análisis Discriminante

Gilbert Saporta*

El objetivo de los métodos de discriminación consiste en predecir una variable cualitativa con k categorías con la ayuda de p predictores, generalmente numéricos.

Se puede considerar el análisis discriminante como una extensión del problema de regresión en el caso donde la variable a explicar es cualitativa; veremos por otra parte que en el caso de dos categorías, podemos referirnos exactamente a una regresión lineal múltiple.

Los datos están constituidos por n observaciones repartidas en k clases y descritas por p variables explivativas.

Se distinguen clásicamente dos aspectos en análisis discriminante:

a) **descriptivo:** buscar cuales son las combinaciones lineales de variables que permiten separar lo mejor posible las k categorías y dar una representación gráfica (así como en análisis factorial), que da cuenta lo mejor posible de esta separación;

b) **decisional:** un nuevo individuo se presenta para el que se conocen los valores de los predictores. Se trata entonces de decidir en cual categoría hay que asignarlo. Es un problema de clasificación (pero no en el sentido de clasificación automática).

Estos dos aspectos corresponden, *grosso modo,* a la distinción entre métodos geométricos y métodos probabilísticos.

Entre las innumerables aplicaciones del análisis discriminante citamos algunos campos:

- ayuda a la decisión en medicina: a partir de medidas de laboratorio, se busca una función que permita predecir lo mejor posible el tipo de afección de un enfermo, o su evolución probable con el fin de orientar el tratamiento;

- meteorología: prevención de avalanchas a partir de variables ligadas a la atmósfera y a la nieve;

- finanzas: prevención del comportamiento de los demandantes de crédito.

*Département de Mathématiques et Informatique, Centre National d'Arts et Métiers, París

1 Métodos geométricos

Estos métodos, esencialmente descriptivos, se basan solamente en los conceptos de distancia y no hacen intervenir hipótesis probabilísticas.

1.1 Datos y notaciones

Los n individuos e_i de la muestra constituyen una nube E de $\mathbb{R}^p$ dividida en k subnubes $E_1, E_2, \ldots, E_k$ con centros de gravedad $g_1, g_2, \ldots, g_k$ y matrices de varianzas $V_1, V_2, \ldots, V_k$ (figura 1).

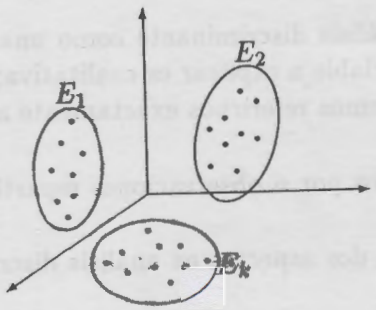

Figura 1

Sea g el centro de gravedad y V la matriz de varianza de E. Si a los n individuos $e_1, e_2, \ldots, e_n$ les son afectados los pesos $p_1, p_2, \cdots, p_n$, entonces los pesos $q_1, q_2, \cdots, q_k$ de cada subnube son entonces

$$q_j = \sum_{e_i \in E_j} p_i$$

Se tiene:

$$g_j = \frac{1}{q_j} \sum_{e_i \in E_j} p_i e_i$$

$$g = \sum_{j=1}^{k} q_j g_j \quad \text{y} \quad V_j = \frac{1}{q_j} \sum_{e_i \in E_j} p_i (e_i - g_j)(e_i - g_j)^t$$

Llamemos *matriz de varianza interclase*, la matriz de varianza B de los k centros de gravedad provistos de los pesos q_j:

$$B = \sum_{j=1}^{k} q_j (g_j - g)(g_j - g)^t$$

y matriz de varianza intraclase W, el promedio de las matrices V_j:

$$W = \sum_{j=1}^{k} q_j V_j$$

En general, W es invertible mientras que B no lo es, porque los k centros de gravedad están en un subespacio de dimensión $k-1$ de $\mathbb{R}^p$ (si $p > k-1$ lo que es generalmente el caso), entonces la matriz B es de tamaño p.

Tenenemos entonces la relación siguiente:

$$V = W + B$$

que se demuestra fácilmente y constituye una generalización de la relación clásica:

varianza total = promedio de las varianzas + varianza de las medias

Supondremos en adelante que $g = 0$, es decir, que las variables explicativas están centradas.

Si consideramos que la tabla de datos a estudiar se pone bajo la forma:

$$[A \mid X]$$

donde X es la matriz de p variables explicativas y A la tabla lógica asociada a la variable cualitativa, los k centros de gravedad $g_1, g_2, \cdots g_k$ son los filas de la matriz $(A^tDA)^{-1}(A^tDX)$.

(A^tDA) es la matriz diagonal de pesos q_j de las subnubes:

$$A^tDA = D_q = \begin{bmatrix} q_1 & & 0 \\ & q_2 & \\ & & \ddots \\ 0 & & q_k \end{bmatrix}$$

La matriz de varianza interclases se escribe entonces, si $g = 0$:

$$\begin{aligned} D &= ((A^tDA)^{-1}A^tDX)^tA^tDA((A^tDA)^{-1}A^tDX) \\ &= X^tDA(A^tDA)^{-1}A^tDX = (X^tDA)D_q^{-1}(A^tDX) \end{aligned}$$

En el caso donde $p_i = \frac{1}{n}$ las expresiones anteriores se simplifican e introduciendo los efectivos $n_1, n_2, \cdots, n_k$ de las k subnubes, tenemos:

$$B = \frac{1}{n}\sum_j n_j g_j g_j^t; \qquad g_j = \frac{1}{n_j}\sum_{e_i \in E_j} e_i; \qquad W = \frac{1}{n}\sum_j n_j V_j$$

Supondremos en adelante estar en este caso.

1.2 El análisis factorial discriminante (AFD)

A. Los ejes y variables discriminantes

El AFD consiste en la búsqueda de nuevas variables (las variables discriminantes) correspondientes a las direcciones de $\mathbb{R}^p$ que separan lo mejor posible en proyección a los k grupos de observaciones.

El eje 1 de la figura 2 posee un buen poder discriminante mientras que el eje 2 (que está en el eje principal usual) no permite separar en proyección los dos grupos.

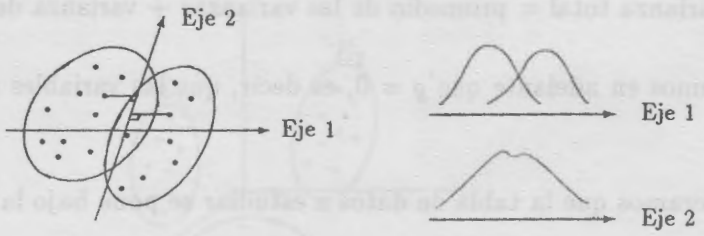

Figura 2

Supongamos que $\mathbb{R}^p$ está dotado de una métrica M. Notaremos a al eje discriminante, u al factor asociado $u = Ma$, la variable discriminante será Xu.

En proyección sobre el eje a, los k centros de gravedad deben ser tan separados como sea posible, mientras que cada subnube debe ser proyectada de manera agrupada alrededor de la proyección de su centro de gravedad.

En otras palabras, la inercia de la nube de los g_j proyectados sobre a debe ser máxima. La matriz de inercia de la nube de los g es MBM y la inercia de la nube proyectada sobre a es a^tMBMa si a es de M-norma igual a 1.

Es necesario también que en proyección sobre a, cada subnube debe ser bien agrupada, y por tanto que a^tMV_jMa sea pequeño para $j = 1, 2, \cdots, k$.

Se buscará por lo tanto minimizar el promedio $\sum_{j=1}^{k} q_j a^tMV_jMa$, sea a^tMWMa.

Ahora bien, la relación $V = B + W$ implica que $MVM = MBM + MWM$ y, por consiguiente:

$$a^tMVMa = a^tMBMa + a^tMWMa$$

Tomaremos entonces como criterio a maximizar, la razón de la inercia interclases entre la inercia total. Es decir,

$$\max_a \frac{a^t MBMa}{a^t MVMa}$$

Sabemos que el máximo se alcanza si a es vector propio de $(MVM)^{-1}MBM$ asociado a su mayor valor propio λ_1:

$$M^{-1}V^{-1}BMa = \lambda a$$

Al eje discriminante a es entonces asociado el factor discriminante u tal que $u = Ma$, de modo que $V^{-1}Bu = \lambda_1 u$.

Los factores discriminantes y, por consiguiente, las variables discriminantes Xu, son independientes de la métrica M. Se escogerá por comodidad $M = V^{-1}$ que da $BV^{-1}a = \lambda a$ y $V^{-1}Bu = \lambda u$.

Se tiene siempre $0 \leq \lambda_1 \leq 1$ pues $0 \leq \dfrac{a^t MBMa}{a^t MVMa} \leq 1$.

$\lambda_1 = 1$ corresponde al caso siguiente: en proyección sobre a las dispersiones intraclases son nulas, las k nubes están por lo tanto en un hiperplano ortogonal a a (ver figura 3).

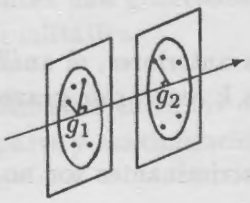

Figura 3

Hay evidentemente discriminación perfecta si los centros de gravedad se proyectaran en puntos diferentes.

$\lambda_1 = 0$ corresponde al caso en que el mejor eje no permite separar los centros de gravedad g_i, es el caso donde están confundidos y las nubes son, por lo tanto, concéntricas y ninguna separación lineal es posible (er figura 4).

Figura 4

Puede ocurrir, sin embargo, que existe la posibilidad de discriminación no lineal: la distancia al centro permite separar los grupos, pero se trata de una función cuadrática de las variables.

El valor propio λ es una medida pesimista del poder discrimiante de un eje. La figura 5 muestra que se puede dicriminar perfectamente pues los grupos están bien separados a pesar que $\lambda < 1$.

El número de los valores propios no nulos, y por tanto el de ejes discriminantes, es igual a $k-1$ en el caso habitual $n > p > k$ y donde las variables no están ligadas por relaciones lineales.

Figura 5

B. Un análisis en componentes principales (ACP) particular

De acuerdo con las ecuaciones anteriores, el análisis factorial discriminante no es otra cosa que el ACP de la nube de los k centros de gravedad con la métrica V^{-1}.

Se deduce que las variables discriminantes son no correlacionadas 2 a 2.

Así como en ACP, se podrá interpretar las variables discriminantes por medio del círculo de las correlaciones.

Representación gráfica

Si existe un segundo eje discriminante, es posible representar la nube de las n observaciones en proyección sobre el plano definido por estos dos ejes, este plano es entonces el que permite visualizar mejor la separación de las observaciones en clases.

Nosotros veremos más adelante que el análisis factorial discriminante equivale también al ACP de los g_i con la métrica W^{-1}.

C. Un análisis canónico particular

El análisis discriminante es el análisis canónico de las tablas A y X.

En efecto, la ecuación del análisis canónico de A y X que da las variables canónicas asociadas a X, se escribe:

$$(X^tDX)^{-1}X^tDA(A^tDA)^{-1}A^tDXu = \lambda u$$

esto es idéntico a $V^{-1}Bu = \lambda u$ de acuerdo con el párrafo 1. Esto es una nueva prueba de que las variables discriminantes son no correlacionadas dos a dos.

Si se designa por Aa la primera variable canónica asociada a A, solución de la otra ecuación del análisis canónico:

$$(A^tDA)^{-1}A^tDX(X^tDX)^{-1}X^tDAa = \lambda a$$

normada de tal forma que su proyección sobre el subespacio de $\mathbb{R}^n$ generado por las p variables explicativas sea indéntica a Xu, se puede presentar al análisis discriminante como la búsqueda de la codificación de la variable cualitativa que la vuelve más próxima del espacio generado por las columnas de X. Si las p variables explicativas son centradas, entonces la variable codificada lo es también y u es el vector de los coeficientes de regresión de Aa sobre X.

El primer valor propio λ_1 es entonces el cuadrado del coeficiente de correlación múltiple.

El análisis discriminante es entonces una generalización de la regresión múltiple en el caso donde la variable a explicar es cualitativa.

La figura 6 en $\mathbb{R}^n$ muestra la identidad entre las dos concepciones del análisis discriminante: análisis canónico por una parte y maximización de la varianza interclase dividida por la varianza total, por otra parte.

W_X es el espacio generado por las columnas de X; W_A es el espacio generado por las indicatrices de la variable a explicar.

Si se proyecta D-ortogonalmente la variable discriminante ξ sobre W_A en Aa, el teorema de Pitágoras se escribe:

$$\|\xi\|^2 = \|Aa\|^2 + \|Aa - \xi\|^2$$

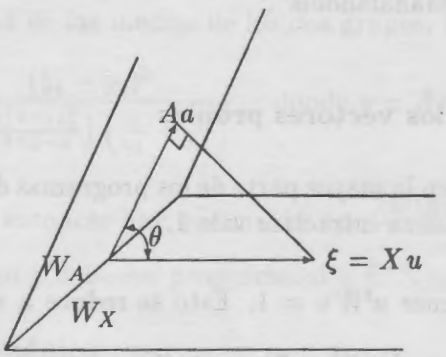

Figura 6

Varianza total de ξ = varianza interclase + varianza intraclase.

La maximización del cociente $\dfrac{\text{varianza interclase}}{\text{varianza total}}$ no es otra cosa que la maximización de $\cos^2\theta$, donde θ es el ángulo formado por Aa y ξ, lo que es el criterio del análisis canónico.

Los autores de habla inglesa llaman a este método **análisis discriminante canónico.**

D. Análisis de varianza y métrica W^{-1}

Si solamente hubiera una variable explicativa, se mediría la eficacia de su poder separador sobre la variable de grupo por medio de un análisis de varianza ordinario con un factor. La estadística F valdría entonces $\dfrac{\text{varianza inter}/k-1}{\text{varianza intra}\ /n-k}$.

Como hay p variables se puede buscar la combinanción lineal definida por unos coeficientes u dando el valor maximal para la estadística de test, lo cual se reduce a maximizar:

$$\frac{u^t B u}{u^t W u}$$

La solución es dada por la ecuación:

$$W^{-1}Bu = \mu u \quad \text{con } \mu \text{ maximal}$$

Los vectores propios de $W^{-1}B$ son los mismos que los de $V^{-1}B$ con $\mu = \dfrac{\lambda}{1-\lambda}$.

En efecto, $Bu = \lambda V u$ es equivalente a:

$$Bu = \lambda(W+B)u, \quad \text{es decir} \quad (1-\lambda)Bu = \lambda W u$$

de donde $W^{-1}Bu = \dfrac{\lambda u}{1-\lambda}$.

Si $0 \leq \lambda \leq 1$ se tiene en cambio $0 \leq \mu \leq \infty$ y $\lambda = \dfrac{\mu}{1+\mu}$.

La utilización de V^{-1} o de W^{-1} como métrica es por tanto indiferente. La métrica W^{-1} es llamada "métrica de Mahalanobis".

E. Normalización de los vectores propios

La convención usual en la mayor parte de los programas de computador es tener variables discriminantes cuya varianza intraclase vale 1.

Se debe por tanto tener $u^t W u = 1$. Esto se reduce a $u^t B u = \dfrac{\lambda}{1-\lambda} = \mu$ y a $u^t V u = \dfrac{1}{1-\lambda}$ puesto que $u^t B u = u^t \lambda (W+B)u = \lambda u^t V u$.

1.3 El caso de dos grupos

A. La función de Fisher

Solamente hay una variable discriminante puesto que $k - 1 = 1$.

El eje discriminante es entonces necesariamente la recta que une los dos centros de gravedad g_1 y g_2:

$$a = (g_1 - g_2)$$

La variable discriminante d se obtiene entonces de la proyección sobre a según la métrica V^{-1} o W^{-1} que tiene en cuenta la orientación de las nubes respecto a la recta de los centros.

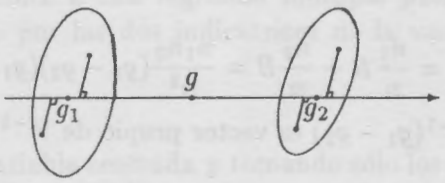

Figura 7

El factor discriminante u vale por tanto:

$$u = V^{-1}(g_1 - g_2) \quad \text{o} \quad u = W^{-1}(g_1 - g_2)$$

que son proporcionales.

$W^{-1}(g_1 - g_2)$ es la función de Fisher (1936)

Para razones de estimación habitualmente no se toma W^{-1}, sino:

$$\frac{n_1 + n_2 - 2}{n_1 + n_2} W^{-1}$$

Se puede en efecto reencontrar el procedimiento de Fisher por el razonamiento siguiente: se busca la combinanción lineal de las variables explicativas tales que el cuadrado de la estadística del test T de igualdad de las medias de los dos grupos, toma un valor maximal:

$$\max \frac{(\bar{y}_1 - \bar{y}_2)^2}{\left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{donde } y = Xu$$

si tomamos $\hat{\Sigma} = \dfrac{n_1 + n_2}{n_1 + n_2 - 2} W$ entonces hay que maximizar $\dfrac{(u^t(g_1 - g_2))^2}{u^t \hat{\Sigma} u}$. u es definido salvo por un factor multiplicativo y debe ser proporcional a $\hat{\Sigma}^{-1}(g_1 - g_2)$.

B. Aplicación del análisis canónico

Se puede encontrar el único valor propio de $V^{-1}B$ observando que para dos grupos:

$$B = \frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)^t$$

En efecto: $B = \frac{n_1}{n} g_1 g_1^t + \frac{n_2}{n} g_2 g_2^t$; ahora bien:

$$g = \frac{n_1}{n} g_1 + \frac{n_2}{n} g_2 = 0$$

De donde $B = \frac{n_1}{n} g_1 g_1^t - \frac{n_1}{n} g_1 g_2^t = \frac{n_1}{n} g_1 (g_1^t - g_2^t)$ y simétricamente:

$$B = \frac{n_2}{n} g_2 (g_1^t - g_2^t)$$

luego:

$$B = \frac{n_1}{n} B + \frac{n_2}{n} B = \frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)^t$$

Se verifica que $u = V^{-1}(g_1 - g_2)$ es vector propio de $V^{-1}B$:

$$V^{-1}\frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)^t V^{-1}(g_1 - g_2) = \lambda V^{-1}(g_1 - g_2)$$

con:

$$\lambda = \frac{n_1 n_2}{n^2}(g_1 - g_2)^t V^{-1}(g_1 - g_2)$$

y:

$$\mu = \frac{\lambda}{1 - \lambda} = \frac{n_1 n_2}{n^2}(g_1 - g_2)^t W^{-1}(g_1 - g_2)$$

μ es, por lo tanto, proporcional a la D_p^2 de Mahalanobis estimada entre los dos grupos ([1], capítulo 15).

Se tiene exactamente:

$$\mu = \frac{n_1 n_2}{n(n-2)} D_p^2 \text{ pues } D_p^2 = \frac{n-2}{n}(g_1 - g_2)^t W^{-1}(g_1 - g_2)$$

Se puede hallar entonces:

$$W^{-1}(g_1 - g_2) = \left(1 + \frac{n_1 n_2}{n(n-2)} D_p^2\right) V^{-1}(g_1 - g_2)$$

El uso del convenio de normalización $u^t W u = 1$ presenta la ventaja siguiente:

Las coordenadas de los dos centros de gravedad sobre el eje discriminante tienen una diferencia igual a la distancia de Mahalanobis D_p.

En efecto, $g_1^t u$ y $g_2^t u$ son estas coordenadas donde u es el factor canónico normalizado. éste es proporcional a $W^{-1}(g_1 - g_2)$, la constante de proporcionalidad α es tal que $u^t W u = 1$, es decir:

$$[\alpha W^{-1}(g_1 - g_2)]^t W [\alpha W^{-1}(g_1 - g_2)] = \alpha^2 (g_1 - g_2)^t W^{-1}(g_1 - g_2)$$

Despreciando la corrección por $\frac{n}{n-2}$ (o utilizando $\hat{\Sigma}$ en lugar de W) sigue que $|\alpha| = \frac{1}{D_p}$.

De donde:

$$|g_1^t u - g_2^t u| = |(g_1 - g_2)^t u| = |\alpha|(g_1 - g_2)^t W^{-1}(g_1 - g_2) = \frac{D_p^2}{D_p} = D_p$$

C. Equivalencia con una regresión múltiple

El análisis canónico se reduce a una regresión múltiple puesto que después de haber centrado, el espacio generado por las dos indicatrices de la variable de los grupos es de dimensión 1.

Es suficiente definir una variable centrada y tomando sólo los dos valores a y b sobre los grupos 1 y 2 respectivamente ($n_1 a + n_2 b = 0$).

Se obtendrá entonces un vector de coeficientes de regresión proporcional a la función de Fisher para cualquier escogencia de a.

La escogencia $a = \frac{n}{n_1}$, $b = -\frac{n}{n_2}$ conduce entonces a $b = (X^t X)^{-1} X^t y = V^{-1}(g_1 - g_2)$.

Se tiene:

$$R^2 = \frac{D_p^2}{\frac{n(n-2)}{n_1 n_2} + D_p^2}$$

Se tendrá cuidado en el hecho que las hipótesis habituales de la regresión no son verificables. Por el contrario; aquí y no es aleatorio y X sí lo es. Las estadísticas usuales que provee un programa de regresión, se deben utilizar únicamente a título indicativo.

D. Un ejemplo

Los datos de la tabla 1 (suministrados por J. P. Nakache) concernientes a 101 víctimas de infartos del miocardo (51 fallecieron, 50 sobrevivieron) sobre los cuales han sido medidos en el momento de la admisión, 7 variables (frecuencia cardiaca, índice cardiaco, índice sistólico, presión diastólica, presión arterial pulmonar, presión ventricular y resistencia pulmonaria). La tabla 2 da las estadísticas elementales por grupo.

Se puede encontrar el único valor propio de $V^{-1}B$ observando que para dos grupos:

$$B = \frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)^t$$

En efecto: $B = \frac{n_1}{n} g_1 g_1^t + \frac{n_2}{n} g_2 g_2^t$; ahora bien:

$$g = \frac{n_1}{n} g_1 + \frac{n_2}{n} g_2 = 0$$

De donde $B = \frac{n_1}{n} g_1 g_1^t - \frac{n_1}{n} g_1 g_2^t = \frac{n_1}{n} g_1 (g_1^t - g_2^t)$ y simétricamente:

$$B = \frac{n_2}{n} g_2 (g_1^t - g_2^t)$$

luego:

$$B = \frac{n_1}{n} B + \frac{n_2}{n} B = \frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)^t$$

Se verifica que $u = V^{-1}(g_1 - g_2)$ es vector propio de $V^{-1}B$:

$$V^{-1}\frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)^t V^{-1}(g_1 - g_2) = \lambda V^{-1}(g_1 - g_2)$$

con:

$$\lambda = \frac{n_1 n_2}{n^2}(g_1 - g_2)^t V^{-1}(g_1 - g_2)$$

y:

$$\mu = \frac{\lambda}{1 - \lambda} = \frac{n_1 n_2}{n^2}(g_1 - g_2)^t W^{-1}(g_1 - g_2)$$

μ es, por lo tanto, proporcional a la D_p^2 de Mahalanobis estimada entre los dos grupos ([1], capítulo 15).

Se tiene exactamente:

$$\mu = \frac{n_1 n_2}{n(n-2)} D_p^2 \text{ pues } D_p^2 = \frac{n-2}{n}(g_1 - g_2)^t W^{-1}(g_1 - g_2)$$

Se puede hallar entonces:

$$W^{-1}(g_1 - g_2) = \left(1 + \frac{n_1 n_2}{n(n-2)} D_p^2\right) V^{-1}(g_1 - g_2)$$

El uso del convenio de normalización $u^t W u = 1$ presenta la ventaja siguiente:

Las coordenadas de los dos centros de gravedad sobre el eje discriminante tienen una diferencia igual a la distancia de Mahalanobis D_p.

En efecto, $g_1^t u$ y $g_2^t u$ son estas coordenadas donde u es el factor canónico normalizado. éste es proporcional a $W^{-1}(g_1 - g_2)$, la constante de proporcionalidad α es tal que $u^t W u = 1$, es decir:

$$[\alpha W^{-1}(g_1 - g_2)]^t W [\alpha W^{-1}(g_1 - g_2)] = \alpha^2 (g_1 - g_2)^t W^{-1}(g_1 - g_2)$$

Despreciando la corrección por $\frac{n}{n-2}$ (o utilizando $\hat{\Sigma}$ en lugar de W) sigue que $|\alpha| = \frac{1}{D_p}$.

De donde:

$$|g_1^t u - g_2^t u| = |(g_1 - g_2)^t u| = |\alpha|(g_1 - g_2)^t W^{-1}(g_1 - g_2) = \frac{D_p^2}{D_p} = D_p$$

C. Equivalencia con una regresión múltiple

El análisis canónico se reduce a una regresión múltiple puesto que después de haber centrado, el espacio generado por las dos indicatrices de la variable de los grupos es de dimensión 1.

Es suficiente definir una variable centrada y tomando sólo los dos valores a y b sobre los grupos 1 y 2 respectivamente ($n_1 a + n_2 b = 0$).

Se obtendrá entonces un vector de coeficientes de regresión proporcional a la función de Fisher para cualquier escogencia de a.

La escogencia $a = \frac{n}{n_1}$, $b = -\frac{n}{n_2}$ conduce entonces a $b = (X^t X)^{-1} X^t y = V^{-1}(g_1 - g_2)$.

Se tiene:

$$R^2 = \frac{D_p^2}{\frac{n(n-2)}{n_1 n_2} + D_p^2}$$

Se tendrá cuidado en el hecho que las hipótesis habituales de la regresión no son verificables. Por el contrario; aquí y no es aleatorio y X sí lo es. Las estadísticas usuales que provee un programa de regresión, se deben utilizar únicamente a título indicativo.

D. Un ejemplo

Los datos de la tabla 1 (suministrados por J. P. Nakache) concernientes a 101 víctimas de infartos del miocardo (51 fallecieron, 50 sobrevivieron) sobre los cuales han sido medidos en el momento de la admisión, 7 variables (frecuencia cardiaca, índice cardiaco, índice sistólico, presión diastólica, presión arterial pulmonar, presión ventricular y resistencia pulmonaria). La tabla 2 da las estadísticas elementales por grupo.

Tabla 1

FRCAR	INCAR	INSYS	PRDIA	PAPUL	PVENT	REPUL	PRONOSTICO
90	1.71	19.0	16	19.5	16.0	912	supervivencia
90	1.68	18.7	24	31.0	14.0	1476	fallecimiento
120	1.40	11.7	23	29.0	8.0	1657	fallecimiento
82	1.79	21.8	14	17.5	10.0	782	supervivencia
80	1.58	19.7	21	28.0	18.5	1418	fallecimiento
80	1.13	14.1	18	23.5	9.0	1664	fallecimiento
94	2.04	21.7	23	27.0	10.0	1059	supervivencia
80	1.19	14.9	16	21.0	16.5	1412	supervivencia
78	2.16	27.7	15	20.5	11.5	759	supervivencia
100	2.28	22.8	16	23.0	4.0	807	supervivencia
90	2.79	31.0	16	25.0	8.0	717	supervivencia
86	2.70	31.4	15	23.0	9.5	681	supervivencia
80	2.61	32.6	8	15.0	1.0	460	supervivencia
61	2.84	47.3	11	17.0	12.0	479	supervivencia
99	3.12	31.8	15	20.0	11.0	513	supervivencia
92	2.47	26.8	12	19.0	11.0	615	supervivencia
96	1.88	19.6	12	19.0	3.0	809	supervivencia
86	1.70	19.8	10	14.0	10.5	659	supervivencia
125	3.37	26.9	18	28.0	6.0	665	supervivencia
80	2.01	25.0	15	20.0	6.0	796	supervivencia
82	3.15	38.4	13	20.0	6.0	508	supervivencia
110	1.66	15.1	23	31.0	6.5	1494	fallecimiento
80	1.50	18.7	13	17.0	12.0	907	fallecimiento
118	1.03	8.7	19	27.0	10.0	2097	fallecimiento
95	1.89	19.9	25	27.0	20.0	1143	fallecimiento
80	1.45	17.1	19	23.0	15.0	1269	fallecimiento
85	1.30	15.1	13	18.0	10.0	1108	fallecimiento
105	1.84	17.5	18	22.0	10.0	957	fallecimiento
122	2.79	22.9	25	36.0	10.0	1032	supervivencia
81	1.77	21.9	18	27.0	11.0	1220	supervivencia
118	2.31	19.6	22	27.0	10.0	935	supervivencia
87	1.20	13.8	34	41.0	20.0	2733	fallecimiento
65	1.19	18.3	15	18.0	13.0	1210	fallecimiento
84	2.15	25.6	27	37.0	10.0	1377	supervivencia
103	0.91	8.8	30	33.5	10.0	2945	fallecimiento
75	2.54	33.9	24	31.0	16.0	976	supervivencia
90	2.08	23.1	20	28.0	6.0	1077	supervivencia
90	1.93	21.4	11	18.0	10.0	746	supervivencia
90	0.95	10.6	20	24.0	6.0	2021	fallecimiento
65	2.38	36.6	16	22.0	12.0	739	supervivencia
95	0.99	10.4	20	27.5	8.0	2222	fallecimiento
95	0.85	8.9	19	22.0	15.5	2071	fallecimiento
86	2.05	23.8	21	28.0	10.0	1093	supervivencia
82	2.02	24.6	16	22.0	14.0	871	supervivencia
70	1.44	20.6	19	26.5	11.0	1472	fallecimiento
92	3.06	33.3	10	15.0	6.0	392	supervivencia

FRCAR	INCAR	INSYS	PRDIA	PAPUL	PVENT	REPUL	PRONOSTICO
94	1.31	13.9	26	40.0	15.0	2443	fallecimiento
79	1.29	16.3	24	31.0	10.0	1922	fallecimiento
67	1.47	21.9	15	18.0	16.0	980	supervivencia
75	1.21	16.1	19	24.0	4.0	1587	fallecimiento
80	2.41	30.9	19	24.0	7.0	797	supervivencia
61	3.28	54.0	12	16.0	7.0	390	supervivencia
110	1.24	11.3	22	27.5	11.0	1774	fallecimiento
116	1.85	15.9	33	42.0	13.0	1816	fallecimiento
75	2.00	26.7	16	22.0	5.0	880	supervivencia
92	1.97	21.4	18.0	27.0	3.0	1096	fallecimiento
110	0.96	8.8	15.0	19.0	16.0	1583	supervivencia
95	2.56	26.9	8.0	13.0	3.0	406	supervivencia
75	2.32	30.9	8.0	10.0	6.00	345	supervivencia
80	2.65	33.1	13.0	19.0	9.0	574	supervivencia
102	1.60	15.7	24.0	31.0	16.0	1550	fallecimiento
86	1.67	19.4	18.0	23.0	8.5	1102	supervivencia
60	0.82	13.7	22.0	32.0	13.05	3122	fallecimiento
100	1.76	17.6	23.0	33.0	2.0	1500	supervivencia
80	3.28	41.0	12.0	17.0	2.0	415	supervivencia
108	2.96	27.4	24.0	35.0	6.5	946	supervivencia
92	1.37	14.8	25.0	46.0	11.0	2686	fallecimiento
100	1.38	13.8	20.0	31.0	11.0	1797	fallecimiento
80	2.85	35.6	25.0	32.0	7.0	898	supervivencia
87	2.51	28.8	16.0	24.0	20.0	765	fallecimiento
100	2.31	23.1	8.0	12.0	0.0	416	supervivencia
120	1.18	9.9	25.0	36.0	8.0	2441	fallecimiento
115	1.83	15.9	25.0	30.0	8.0	1311	fallecimiento
101	2.55	25.2	23.2	30.5	9.0	957	supervivencia
100	2.31	23.1	8.0	12.0	0.0	416	supervivencia
120	1.18	9.9	25.0	36.0	8.0	2441	fallecimiento
115	1.83	15.9	25.0	30.0	8.0	1311	fallecimiento
101	2.55	25.2	23.2	30.5	9.0	957	supervivencia
92	2.17	23.5	19.0	24.0	3.0	885	supervivencia
87	1.42	16.1	20.0	26.0	10.0	1465	fallecimiento
80	1.59	19.9	13.0	20.5	4.0	1031	supervivencia
88	1.47	16.7	23.0	32.5	10.0	1769	fallecimiento
104	1.23	11.8	27.0	33.0	11.0	2146	fallecimiento
90	1.45	16.1	17.0	24.0	8.5	1324	supervivencia
67	0.85	12.7	26.0	33.0	11.0	3106	fallecimiento
87	2.37	27.2	15.0	22.0	10.0	743	supervivencia
108	2.40	22.2	26.0	31.0	4.0	1033	supervivencia
120	1.91	15.9	18.0	27.0	15.0	1131	fallecimiento
108	1050	13.9	28.0	43.0	16.0	1813	fallecimiento
86	2.36	27.4	24.0	34.0	8.0	1153	supervivencia
112	1.56	13.9	24.0	29.0	4.0	1487	fallecimiento
80	1.34	17.0	16.0	25.0	16.0	1493	fallecimiento
95	1.65	17.4	20.0	33.0	7.0	1600	fallecimiento
90	2.04	22.7	28.0	41.0	10.0	1608	fallecimiento

FRCAR	INCAR	INSYS	PRDIA	PAPUL	PVENT	REPUL	PRONO
90	2.04	22.7	28.0	41.0	10.0	1608	fallecimiento
90	3.03	33.6	17.0	23.5	7.0	620	supervivencia
94	1.21	12.9	17.0	22.0	3.0	1455	fallecimiento
51	1.34	26.3	11.0	17.0	6.0	1015	fallecimiento
110	1.17	10.6	29.0	35.0	10.5	2393	fallecimiento
96	1.74	18.1	24.0	29.0	6.0	1333	fallecimiento
132	1.31	9.9	23.0	28.0	12.0	1710	fallecimiento
135	0.95	7.0	15.0	20.0	7.0	1684	fallecimiento
105	1.92	18.3	18.0	24.0	3.0	1000	fallecimiento
99	0.83	8.4	23.0	27.0	8.0	2602	fallecimiento
116	0.60	5.2	33.0	38.0	10.0	5067	fallecimiento
112	1.54	13.8	25.0	31.0	8.0	1610	fallecimiento

Un análisis factorial discriminante entre los sobrevivientes y los fallecidos dan los resultados siguientes: la distancia de Mahalanobis al cuadrado vale:

$$D_7^2 = 4.942 \text{ de donde } \quad D_7 = 2.223$$

Tabla 2

Variable	N	Media	Desviación estándar
FRCAR	51	95.90196078	17.97693511
INCAR	51	1.39470588	0.37619332
INSYS	51	14.99607843	4.63900682
PRDIA	51	21.09803922	5.14183152
PAPUL	51	29.09803922	6.81910523
PVENT	51	10.64705882	4.34429985
REPUL	51	1797.27450980	739.87296419
FRCAR	50	88.34000000	13.84109527
INCAR	50	2.30580000	0.56055035
INSYS	50	26.75200000	8.08319597
PRDIA	50	16.50400000	5.15304388
PAPUL	50	22.84000000	6.46532352
PVENT	50	8.33000000	4.05398519
REPUL	50	841.38000000	303.68256050

Bajo las hipótesis de multinormalidad ([1], capítulo 15), es el valor correspondiente a $F = 16476$:

$$\frac{n_1 n_2}{n} \frac{n-p-1}{p(n-2)} D_p^2 = F$$

El valor crítico al 1% para un valor $F(7;93)$ es 2.84, el D^2 es significativo de una diferencia neta entre los dos grupos.

Se halla $R^2 = \lambda = 0.5576$ y $\mu = 1.2604$.

La variable discriminante se obtiene entonces por la combinación lineal de las 7 variables centradas sobre la media general de los dos grupos (var tabla 3).

Tabla 3

FRCAR	-0.026445290
INCAR	2.768181397
INSYS	-0.075037835
PRDIA	0.009115031
PAPUL	-0.074211897
PVENT	-0.021086258
REPUL	0.000084078

Si no se centra se suma la constante 1.22816 a la combinación lineal anterior de los datos brutos.

Los coeficientes de correlación lineales de la variable discriminante con las 7 variables (los dos grupos confundidos), son idénticos sobre la tabla 4.

Tabla 4

FRCAR	-0.3097
INCAR	0.9303
INSYS	0.8976
PRDIA	-0.6321
PAPUL	-0.5751
PVENT	-0.3591
REPUL	-0.8676

Las medias de los dos grupos sobre las variables discriminantes son:

Fallecimientos -1.1005
Sobrevivencia 1.1225

Se vuelve a encontrar $D_7 = +1.1005 + 1.1225 = 2.2230$

1.4 Reglas geométricas de asignación

Habiendo encontrado la mejor representación de la separación en k clases de n individuos, se puede entonces intentar asignar una observación e a uno de los grupos.

La regla natural consiste en calcular las distancias de la observación a clasificar, a cada uno de los k centros de gravedad y asignar según la distancia más pequeña. Falta definir la métrica que se va a utilzar.

A. Regla de Mahalanobis–Fisher

Consiste en utilizar la métrica W^{-1} (o V^{-1} que es equivalente):

$$d^2(e, g_i) = (e - g_i)^t W^{-1}(e - g_i)$$

Desarrollando esta cantidad se encuentra:

$$d^2(e, g_i) = e^t W^{-1} e + g_i^t W^{-1} g_i - 2e^t W^{-1} g_i$$

Puesto que $e^t W^{-1} e$ no depende del grupo i, la regla consiste en buscar el mínimo de $g_i^t W^{-1} g_i - 2e^t W^{-1} g_i$ o el máximo de $e^t W^{-1} g_i - \frac{g_i^t W^{-1} g_i}{2}$.

Se ve que esta regla es lineal respecto a las coordenadas de e.

Se debe calcular para cada individuo, k funciones lineales de sus coordenadas y buscar el valor maximal.

La tabla 5 da para el ejemplo de los infartos las dos funciones de clasificación.

Tabla 5

	deceso	supervivencia
CONSTANTE	-91.57481116	-89.97134555
FRCAR	1.53609883	1.47730875
INCAR	-52.09444392	-45.94054613
INSYS	5.44165359	5.27483824
PRDIA	-0.64815662	-0.62789315
PAPUL	0.70738671	0.54240748
PVENT	0.85037707	0.80350057
REPUL	0.00638975	0.00657667

En el caso de los dos grupos se decidirá asignar al grupo 1 si:

$$e^t W^{-1} g_1 - \frac{1}{2}(g_1^t W^{-1} g_1) > e^t W^{-1} g_2 - \frac{1}{2} g_2^t W^{-1} g_1$$

o sea

$$e^t W^{-1}(g_1 - g_2) > \frac{1}{2}(g_1 + g_2)^t W^{-1}(g_1 - g_2)$$

Puesto que $W^{-1}(g_1 - g_2)$ es la función de Fisher, la regla consiste en asignar al grupo 1 si el valor de la función discriminante es superior al umbral:

$$\frac{1}{2}(g_1 + g_2)^t W^{-1}(g_1 - g_2)$$

Cuando los dos grupos son de igual efectivo $g_1 + g_2 = 0$; se asigna al grupo 1 si la función $e^t W^{-1}(g_1 - g_2)$ es positiva.

En el ejemplo de los infartos es poco más o menos el caso puesto que $n_1 = 50$ y $n_2 = 51$. Se predecirá la supervivencia si la función discriminante es positiva.

Observemos que la aplicación de la regla geométrica se puede hacer indiferentemente en el espacio $\mathbb{R}^p$ o en el espacio factorial $\mathbb{R}^{k-1}$.

En particular si $k = 3$, las fronteras de asignación a los grupos son hiperplanos ortogonales al plano de los tres centros de gravedad. Se pueden leer directamente las distancias de Mahalanobis a g_1, g_2, g_3 utilizando el gráfico de las dos variables canónicas discriminantes normalizadas a 1 (en el sentido de la varianza interclase).

B. Insuficiencia de las reglas geométricas

La utilización de la regla anterior conduce a asignaciones incorrectas cuando las dispersiones de los grupos son muy diferentes entre ellos, nada justifica entonces el uso de la misma métrica para los diferentes grupos.

En efecto, si se considera la figura 8, aunque e sea más próximo de g_1 que de g_2 en el sentido habitual, es más natural asignar e a la segunda clase que a la primera cuyo "poder de atracción" es más pequeño.

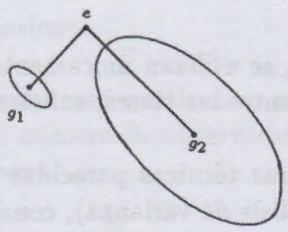

Figura 8

Diversas soluciones utilizan métricas locales M_i tales como:

$$d^2(e, g_i) = (e - g_i)^t M_i (e - g_i)$$

que han sido propuestas, tomando en general, M_i proporcional a V_i^{-1}.

El problema de la optimalidad de la regla de decisión geométrica no puede ser resuelto sin referencia a un modelo probabilístico. En efecto, el problema es de saber cómo se comportará esta regla frente a nuevas observaciones, lo que lleva a hacer hipótesis distribucionales sobre la distribución en el espacio de estas nuevas observaciones. Se alcanzan, por lo tanto, los límites de los métodos descriptivos. Veremos más adelante en qué condiciones ellas conducen a unas reglas optimales.

1.5 Un método de discriminación sobre variables cualitativas: el método Disqual

Cuando los predictores son p variables cualitativas $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$ con $m_1, m_2 \ldots, m_p$ modalidades respectivamente, se puede utilizar el procedimiento siguiente: se efectúa en una primera etapa el análisis de correspondencias múltiples de las variables $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$, es decir el análisis de correspondencias de la tabla disyuntiva $X = (X_1|X_2 \ldots |X_p)$.

Se reemplazan entonces las p variables cualitativas por las q coordenadas sobre los ejes factoriales y se efectúa luego un análisis factorial discriminante sobre estas q variables numéricas $z_1, z_2, \ldots, z_q$.

Un factor discriminante d es una combinación lineal de los z_j que son combinaciones lineales de las indicatrices de los $\mathcal{X}_i$. Se expresa entonces directamente d como combinación lineal de las indicatrices de los $\mathcal{X}_i$, o lo que se reduce a atribuir a cada categoría de cada variable un valor numérico *puntaje*. d es entonces simplemente igual a la suma de los puntajes obtenidos en las categorías de las p varibles. Esto se reduce a trasformar cada variable cualitativa en una variable numérica discreta con m valores.

Cuando $k = 2$ este método es óptimo en el sentido siguiente: tomando todos los factores posibles del Análisis de las Correspondencias Múltiples (ACM) $\left(\sum_i m_i - p\right)$ la cuantificación de las variables $\mathcal{X}_i$ es la que da la distancia de Mahalanobis más grande entre los dos grupos.

En la práctica, sin embargo, se utilizan únicamente los factores que presentan a la vez una inercia y en poder separar entre las clases, suficientes.

Igualmente, se podrían utilizar técnicas parecidas a las expuestas en el capítulo 17 de [1] (modelo lineal general y análisis de varianza), consistentes en anular los coeficientes de ciertas variables indicatrices y efectuar un análisis discriminante sobre $\sum_i m_i - p$ columnas de X.

La ventaja del análisis de correspondencias es proveer además de una descripción de las relaciones entre las variables explicativas, unas componentes ortogonales (el D_q^2 es entonces la suma de los D_1^2 sobre las diversas componentes con la métrica V^{-1}).

En estas mismas memorias, se detalla un poco más este método y se hace una aplicación al puntaje en crédito (*credit-scoring*).

2 Métodos probabilísticos

2.1 La regla Bayesiana

Se supone que los k grupos están en proporción $p_1, p_2, \ldots, p_k$ en la población total y que la distribución de probabilidad del vector observación $x = (x_1, \ldots, x_p)$ es dado para cada grupo j por una densidad (o una ley discreta) $f_j(x)$.

Observando un punto de coordenadas $(x_1, x_2, \ldots, x_p)$ la probabilidad de que provenga del grupo j es dada por la fórmula de Bayes:

$$P(G_j/x) = \frac{p_j f_j(x)}{\sum_{j=1}^{k} p_j f_j(x)}$$

La regla bayesiana consiste entonces en asignar la observación x al grupo que tiene la probabilidad *a posteriori* máxima.

Los denominadores siendo los mismos para los k grupos, se debe entonces buscar el máximo de:

$$p_j f_j(x)$$

Es entonces necesario conocer o estimar $f_j(x)$. Diversas posibilidades existen.

A. Métodos no paramétricos

No se hace hipótesis específica sobre la familia de leyes de probabilidad. Variantes multidimensionales del método del núcleo permiten estimar $f_j(x)$ por

$$\hat{f}_j(x) = \frac{1}{n_j h} \sum_{i=1}^{n_j} K\left(\frac{x - x_i}{h}\right)$$

donde K es una densidad multidimensional.

La discriminación "por bolas" es un caso particular: se traza alrededor de x una bola en $\mathbb{R}^p$ de radio ρ dado y se cuenta el número de observaciones k_j del grupo j en esta bola. Se estimará directamente $P(G_j/x)$ por:

$$\frac{k_j}{\sum_j k_j}$$

(*Observación:* la bola puede ser vacía si ρ es muy pequeño).

Uno de los métodos más utilizados es el método de los k vecinos más cercanos. Se buscan los k puntos más próximos de x en el sentido de una métrica que se precisa y se clasifica x en el grupo más representado: la probabilidad *a posteriori* se obtiene igual que para la discriminación por bolas, pero no tiene mucho sentido si k es pequeño.

B. Métodos paramétricos

Se da una familia parametrizada de leyes de probabilidad para $f_j(x)$ y se utiliza la muestra para estimar los parámetros. El caso normal p-dimensional es el más clásico y será desarrollado más adelante.

Se puede igualmente dar una expresión parametrizada de la probabilidad *a posteriori* y estimarla directamente: será el caso de la regresión logística tratada en la sección §2.4.

2.2 El modelo normal multidimensional

Se supondrá que x sigue una ley $N_p(\mu, \Sigma_j)$ para cada grupo:

$$f_j(x) = \frac{1}{(2\pi)^{p/2}(\det \Sigma_j)^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_j)^t \Sigma_j^{-1}(x-\mu_j)\right]$$

A. El caso general

La regla bayesiana $\max p_j f_j(x)$ se reduce, pasando a logaritmos, a minimizar:

$$(x-\mu_j)^t \Sigma_j^{-1}(x-\mu_j) - 2\ln p_j + \ln(det\Sigma_j)$$

Cuando los Σ_j son diferentes esta regla es cuadrática y se debe comparar k funciones cuadráticas de x. Σ_j es en general estimado por $\frac{n}{n-1}V_j$ y μ_j por g_j.

B. El caso de igualdad de matrices de varianza–covarianza

Si $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k = \Sigma$, la regla se vuelve lineal. En efecto $\ln(\det \Sigma_j)$ es una constante y $(x-\mu_j)^t\Sigma^{-1}(x-\mu_j)$ es entonces igual a la distancia de Mahalanobis teórica de x a μ_j: $\Delta^2(x, \mu_j)$.

Desarrollando y eliminando $x^t\Sigma^{-1}x$ que no depende del grupo, se tiene:

$$\text{máx}\left\{x^t\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^t\Sigma^{-1}\mu_j + \ln p_j\right\}$$

Si Σ es estimada por $\frac{n}{n-k}W$, la regla bayesiana corresponde a la regla geométrica cuando hay igualdad de probabilidades *a priori*. La regla geométrica es entonces óptima.

La probabilidad *a posteriori* de pertenencia al grupo j es proporcional a:

$$p_j \exp\left(-\frac{1}{2}\Delta^2(x, \mu_j)\right)$$

C. Dos grupos con igualdad de las matrices de varianza

Se asignará x al grupo 1 si:

$$x^t\Sigma^{-1}(\mu_1 - \mu_2) > \frac{1}{2}(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) + \ln\frac{p_2}{p_1}$$

Si $p_1 = p_2 = 0.5$ se encuentra la regla Fisher estimando Σ por $\frac{n}{n-2}W$.

Es decir:

$$S(x) = x^t\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) + \ln\frac{p_2}{p_1}$$

Se asignará a x al grupo 1 si $S(x) > 0$ y al grupo 2 si $S(x) < 0$.

La función $S(x)$ llamada *puntaje* o estadística de Anderson está ligada simplemente a la probabilidad *a posteriori* de pertenencia al grupo 1.

Tenemos en efecto

$$P(G_1/x) = P = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)}$$

de donde

$$\frac{1}{P} = 1 + \frac{p_2 f_2(x)}{p_1 f_1(x)} = 1 + \frac{p_2}{p_1} \exp\left[-\frac{1}{2}(x-\mu_2)\Sigma^{-1}(x-\mu_2) + \frac{1}{2}(x-\mu_1)\Sigma^{-1}(x-\mu_1)\right]$$

$$\frac{1}{P} - 1 = \frac{p_2}{p_1} \exp\left[\frac{1}{2}\Delta^2(x,\mu_1) - \frac{1}{2}\Delta^2(x,\mu_2)\right]$$

de donde $\ln\left(\frac{1}{P} - 1\right) = -S(x)$.

Es decir:

$$P = \frac{1}{1 + \exp(-S(x))} = \frac{\exp(S(x))}{1 + \exp(S(x))}$$

Se dice que P es función "logística" del *puntaje*.

Cuando $p_1 = p_2 = \frac{1}{2}$:

$$P = \frac{1}{1 + \exp\left(-\frac{1}{2}(\Delta^2(x,\mu_1) - \Delta^2(x,\mu_2))\right)}$$

He aquí a manera de ejemplo la tabla 6 que da las asignaciones de las 45 primeras observaciones de los datos de infartos según la regla anterior. El asterisco indica un error de clasificación.

Tabla 6

	Grupo real	Grupo atribuido		$P(G_1/x)$	$P(G_2/x)$
1	sobreviviente	sobreviviente		0.4515	0.5485
2	fallecido	fallecido		0.8140	0.1860
3	fallecido	fallecido		0.9597	0.0403
4	sobreviviente	sobreviviente		0.2250	0.7750
5	fallecido	fallecido		0.8112	0.1888
6	fallecido	fallecido		0.8928	0.1072
7	sobreviviente	sobreviviente		0.3202	0.6798
8	sobreviviente	fallecido	*	0.8711	0.1289
9	sobreviviente	sobreviviente		0.0984	0.9016
10	sobreviviente	sobreviviente		0.0797	0.9203
11	sobreviviente	sobreviviente		0.0138	0.9862
12	sobreviviente	sobreviviente		0.0160	0.9840
13	sobreviviente	sobreviviente		0.0052	0.9948
14	sobreviviente	sobreviviente		0.0105	0.9895
15	sobreviviente	sobreviviente		0.0019	0.9981
16	sobreviviente	sobreviviente		0.0258	0.9742
17	sobreviviente	sobreviviente		0.2011	0.7989
18	sobreviviente	sobreviviente		0.2260	0.7740
19	sobreviviente	sobreviviente		0.0022	0.9978
20	sobreviviente	sobreviviente		0.1222	0.8778
21	sobreviviente	sobreviviente		0.0014	0.9986
22	fallecido	fallecido		0.8629	0.1371
23	fallecido	sobreviviente	*	0.4804	0.5196
24	fallecido	fallecido		0.9900	0.0100
25	fallecido	fallecido		0.5845	0.4155
26	fallecido	fallecido		0.7447	0.2553
27	fallecido	fallecido		0.7067	0.2933
28	fallecido	sobreviviente	*	0.4303	0.5697
29	sobreviviente	sobreviviente		0.1118	0.8882
30	sobreviviente	fallecido	*	0.5734	0.4266
31	sobreviviente	sobreviviente		0.2124	0.7876
32	fallecido	fallecido		0.9928	0.0072
33	fallecido	fallecido		0.7301	0.2699
34	sobreviviente	fallecido	*	0.5354	0.4646
35	fallecido	fallecido		0.9943	0.0057
36	sobreviviente	sobreviviente		0.1218	0.8782
37	sobreviviente	sobreviviente		0.2757	0.7243
38	sobreviviente	sobreviviente		0.1759	0.8241
39	fallecido	fallecido		0.9555	0.0445
40	sobreviviente	sobreviviente		0.0695	0.9305
41	fallecido	fallecido		0.9762	0.0238
42	fallecido	fallecido		0.9785	0.0215
43	sobreviviente	sobreviviente		0.3240	0.6760
44	sobreviviente	sobreviviente		0.2121	0.7879
45	fallecido	fallecido		0.7880	0.2120

Bajo reserva del carácter realista de la hipótesis de multinormalidad, sus resultados son

entonces más precisos que una simple decisión según la distancia más corta. El cálculo de probabilidades *a posteriori* muestra aquí que 4 clasificaciones erróneas sobre 5 se produjeron en una zona de incertidumbre.

D. Acerca de ciertas pruebas

La hipótesis de igualdad de las matrices Σ_i puede ser probada por medio de una prueba de Box que generaliza la de Bartlett para el caso unidimensional.

Si la hipótesis $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ *es verdadera, la cantidad:*

$$\left(1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right)\left[\left(\sum_i \frac{1}{n_i - 1} - \frac{1}{n-k}\right)(n-k)\ln\left|\frac{n}{n-k}W\right| - \sum_i (n_i - 1)\ln\left|\frac{n_i}{n_i - 1}V_i\right|\right]$$

sigue aproximadamente una ley χ^2 *con* $\dfrac{p(p+1)(k-1)}{2}$ *grados de libertad.*

Si se rechaza la hipótesis de igualdad, ¿debemos utilizar las reglas cuadráticas? Esto no es seguro en todos los casos. Para empezar, el test de Box no es perfectamente confiable, además que el uso de las reglas cuadráticas implica la estimación de más parámetros que la regla lineal porque se debe estimar cada Σ_j. Cuando las muestras son pequeñas las funciones obtenidas son muy poco robustas y es mejor utilizar una regla lineal a pesar de todo.

Para dos grupos el resultado siguiente está en el origen de los métodos clásicos de selección de variables:

Sea un subgrupo de l *variables entre las* p *componentes de* x. *Supongamos que* $\Delta_p^2 = \Delta_l^2$, *en otros términos las* $p-1$ *variables restantes no dan ninguna información para separar las dos poblaciones; entonces:*

$$\frac{(n_1 + n_2 - p - 1)n_1 n_2 (D_p^2 - D_l^2)}{(p-l)(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_l^2} = F(p-l; n_1 + n_2 - p - 1)$$

Se puede así probar el crecimiento de la distancia de Mahalanobis aportada por una nueva variable a un grupo ya constituido tomando $l = p - 1$.

Cuando se hace la discriminación entre más de dos grupos, las pruebas son las que utilizan el Λ de Wilks.

El test de igualdad de las k esperanzas $\mu_1 = \mu_2 = \cdots = \mu_k$ es el siguiente

$$\Lambda = \frac{|W|}{|V|} = \frac{|W|}{|W + B|} = \frac{1}{|W^{-1}B + I|}$$

sigue la ley de Wilks de parámetros $p, n-k, k-1$ *bajo* $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.

Porque nV, nW, nB siguen respectivamente las leyes de Wishart con $n-1, n-k, k-1$ grados de libertad.

Si $k = 3$ se utilizará la ley exacta de Λ y no una aproximación

$$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} = \frac{p}{n-k-p+1} F(2p; 2(n-k-p+1))$$

Si $k = 2$, la prueba de Wilks y el test de la distancia de Mahalanobis ($H_0 : \Delta_p^2 = 0$) son idénticos porque B siendo de rango 1 se tiene:

$$\Lambda = \frac{1}{1 + D_p^2 \dfrac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)}} = \frac{1}{\mu+1} = 1-\lambda$$

La prueba $H_0 : \mu_i = \mu \quad \forall i$ puede efectuarse igualmente utilizando como estadística de prueba la traza de $W^{-1}B$ llamada estadística de Lawley-Hotelling que sigue la ley T_0^2 generalizada de Hotelling aproximable por un $\chi^2_{p(k-1)}$.

La traza de $V^{-1}B$ es llamada traza de Pillai. Para la introducción paso a paso de variables discriminando en k grupos, se utiliza usualmente la prueba de variación Λ medida por:

$$\frac{n-k-p}{k-1}\left(\frac{\Lambda_p}{\Lambda_{p+1}} - 1\right)$$

que se compara con un $F_{k-1;n-k-p}$

2.3 Medidas de eficacia de las reglas de clasificación

El criterio usual es la probabilidad de clasificar bien una observación cualquiera. Se comparará la eficacia de los diversos métodos de clasificación en términos de tasas de error.

A. Tasa de error para dos grupos con $\Sigma_1 = \Sigma_2$ y distribución normal

Cuando $p_1 = p_2$, la regla de clasificación teórica es asignar al grupo 1 si:

$$S(x) = x^t\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1+\mu_2)^t\Sigma^{-1}(\mu_1-\mu_2) > 0$$

La probabilidad de error de clasificación es entonces:

$$P(S(x) > 0 / x \in N_p(\mu_2; \Sigma))$$

La ley de $S(x)$ es una ley de Gauss de una dimensión como combinación lineal de componentes de x.

$$E(S(x)) = \mu_2^t \Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^t \Sigma^{-1}(\mu_1 - \mu_2)$$

$$= \frac{1}{2}(\mu_1 - \mu_2)^t \Sigma^{-1}(\mu_1 - \mu_2) = -\frac{1}{2}\Delta_p^2$$

$$V(S(x)) = (\mu_1 - \mu_2)^t \Sigma^{-1} \Sigma \Sigma^{-1}(\mu_1 - \mu_2) = \Delta_p^2$$

de donde

$$S(x) \quad \text{sigue una} \quad LG\left(-\frac{1}{2}\Delta_p^2; \Delta_p\right) \quad \text{si} \quad x \in G_2.$$

La probabilidad de clasificar en el grupo 1 una observación del grupo 2 es:

$$P(1/2) = P\left(U > \frac{\Delta_p}{2}\right)$$

que es igual a $P(2/1)$. Esta relación da una interpretación concreta de la distancia de Mahalanobis.

Si $p_1 \neq p_2$ se tiene:

$$P(1/2) = P\left(U > \frac{\Delta_p}{2} + \frac{1}{\Delta_p} \ln \frac{p_2}{p_1}\right)$$

$$P(2/1) = P\left(U > \frac{\Delta_p}{2} - \frac{1}{\Delta_p} \ln \frac{p_2}{p_1}\right)$$

Cuando μ_1, μ_2 y Σ son estimados $S(x)$ no sigue una ley normal y utilizar D_p como estimación del Δ_p conduce a una estimación sesgada de las probabilidades de error de clasificación: hay en promedio, subestimación de la probabilidad global de error, $p_1 P(2/1) + p_2 P(1/2)$ debido entre otras razones al hecho que D_p^2 sobreestima Δ_p^2 ([1] capítulo 15, sección 15.5.6c).

Para el ejemplo de los infartos, como $D_p = 2.223$ se llega a una estimación de tasas de error igual a $P(U > 1.11) = 0.13$.

La utilización de la estimación sin sesgo de Δ^2, $\frac{n-p-1}{n-2} D^2 - p\frac{n}{n_1 n_2} = 4.37$ conduce a una estimación de la tasa de error cercana al 15%.

B. El método de resustitución

éste consiste en reasignar las n observaciones según las funciones discriminantes encontradas. En el ejemplo de los infartos se obtienen los resultados dados en la tabla 7, es decir, con una tasa de error de 13%.

Tabla 7

grupo atribuido / grupo real	deceso	sobrevivencia
deceso	46	5
sobrevivencia	8	42

Este método tiene un gran defecto: subestima sistemáticamente la tasa de error porque se utilizan las mismas observaciones que sirvieron para encontrar las funciones discriminantes. La regla óptima para la muestra da buenos resultados si se aplica sobre ella.

C. Los métodos de validación cruzada

Para evitar el defecto del método de resustitución se aconseja partir la muestra en dos submuestras: una servirá para la elaboración de reglas de clasificación (muestra de base o de aprendizaje), la otra para la aplicación de las reglas de clasificación (muestra–prueba).

La tasa de error medida sobre la muestra–prueba será entonces una estimación sin sesgo de la tasa verdadera. Esto supone que se tienen numerosos datos para poder sustraer sin riesgo una parte considerable de los datos (25% es aconsejado).

En el caso de muestras pequeñas la técnica siguiente de Lachenbruch y Mickey (comparable al *press* en regresión) permite obtener una estimación realista de la tasa de error.

Se efectúan n análisis discriminantes sobre cada una de las n muestras de las $n-1$ observaciones obtenidas poniendo de lado cada vez una de las observaciones. Se clasifica entonces la observación que fue dejada aparte y se cuenta el porcentaje de error de clasificación.

Observación: Los procedimientos usuales de selección de variables optimisan criterios probabilísticos: ley de Λ Wilks o distancia de Mahalanobis, que no necesariamente optimizan el porcentaje correctamente.

2.4 La regresión logística

Cuando sólo hay dos grupos, bajo la hipótesis de normalidad e igualdad de las matrices de varianza, se ha visto que la probabilidad *a posteriori* era una función logística del *puntaje*, el cual era una función lineal de las variables explicativas.

Se tiene pues:

$$\ln\left(\frac{f_1(x)}{f_2(x)}\right) = \beta_0 + \beta^t x \quad \text{donde} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

El modelo de regresión logística consiste en partir de la relación anterior y estimar los $p+1$ parámetros según el máximo de verosimilitud.

Respecto a la discriminación lineal usual, el modelo logístico implica menos parámetros y cubre una amplia gama de leyes de probabilidades (las variables explicativas pueden ser binarias).

Se tiene, por tanto:

$$P(G_1/x) = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)} = \frac{\frac{p_1}{p_2}\frac{f_1(x)}{f_2(x)}}{1 + \frac{p_1}{p_2}\frac{f_1(x)}{f_2(x)}}$$

$$= \frac{\exp\left(\beta_0 + \ln\frac{p_1}{p_2} + \beta^t x\right)}{1 + \exp\left(\beta_0 + \ln\frac{p_1}{p_2}\beta^t x\right)}$$

$$P(G_2/x) = \frac{1}{1 + \exp\left(\beta_0 + \ln\frac{p_1}{p_2} + \beta^t x\right)}$$

La verosimilitud de las β (suponiendo n_1 y n_2 fijos y no aleatorios) es:

$$L = \prod_{i \in G_1} f_1(x_i) \prod_{i \in G_2} f_2(x_i)$$

Como:

$$f_1(x) = \frac{P(G_1/x) f(x)}{p_1} \quad \text{y} \quad f_2(x) = \frac{P(G_2/x) f(x)}{p_2}$$

Con $f(x) = p_1 f_1(x) + p_2 f_2(x)$, se tiene:

$$L = \frac{1}{p_1^{n_1} p_2^{n_2}} \prod_{i \in G_1} P(G_1/x_i) \prod_{i \in G_2} P(G_2/x_i) \prod_{i=1} f(x_i)$$

$$L = \frac{L_1 L_2}{p_1^{n_1} p_2^{n_2}}$$

donde L_1 es la verosimilitud condicional de los parámetros, conociendo las x_i y la densidad incondicional de las x_i, L_2.

No siendo conocida f se estimarán $\beta_0, \beta_1 \cdots \beta_p$ por un método de máxima verosimilitud condicional:

$$\max_{\beta} \prod_{i \in G_1} \frac{\exp\left(\beta_0 + \ln \frac{p_1}{p_2} + \beta^t x_i\right)}{1 + \exp\left(\beta_0 + \ln \frac{p_1}{p_2} + \beta^t x_i\right)} \prod_{i \in G_2} \frac{1}{1 + \exp\left(\beta_0 + \ln \frac{p_1}{p_2} + \beta^t x_i\right)}$$

Para esto hay que utilizar un método numérico, puesto que no hay solución análitica a la ecuación de la verosimilitud.

Siendo estimadas las β, la regla Bayesiana puede ser aplicada para las clasificaciones. Como:

$$\ln \frac{P(G_1/x)}{P(G_2/x)} = \beta_0 + \ln \frac{p_1}{p_2} + \beta^t x$$

se asignará al grupo 1 si $\beta_0 + \ln \frac{p_1}{p_2} + \beta^t x > 0$.

Cuando los datos provienen de dos pobalciones normales con $\Sigma_1 = \Sigma_2$ la regresión logística es menos adecuada que el análisis discriminante clásico, pues la solución dada por $S(x)$ corresponde a un máximo de verosimilitud verdadera y no a un máximo de verosimilitud condicional (se utiliza menos información en la regresión logísta puesto que sólo f_1/f_2 se supone conocido y no f_1 y f_2).

Parece que la regresión logística da resultados verdaderamente mejores que la regla geométrica, sólo para poblaciones claramente no normales o con Σ_1 muy diferente de Σ_2, pero al precio de un procedimiento de cálculo mucho más complejo que la simple inversión de la matriz W.

Bibliografía

[1] Saporta, G. (1988) *Théorie et Méthodes de la Statistique*, 2a. edición. Technip, París.