Introducción

Supongamos que disponemos de un conjunto de individuos clasificados en distintos grupos de acuerdo con la observación de una determinada característica que los diferencia. Parece lógico pensar que esa característica observada que los diferencia pudiera estar relacionada con otras características más fácilmente observables, de forma que si dispusiéramos de ellas y conociéramos la relación existente entre éstas y aquélla (por ejemplo, mediante algún tipo de función "predictiva") podríamos tratar de anticipar su comportamiento más probable, con cierta fiabilidad, sin necesidad de esperar a observarla. Ello lógicamente es tanto más interesante cuanto mayor es el coste asociado a la observación final de la característica que expresa finalmente la clasificación real, y tanto más cuando la observación de esta característica conlleve la desaparición de la propia unidad observada (caso de que la característica sea, por ejemplo, la muerte).

Introduzcamos como ilustración el siguiente ejemplo simplificado. Imaginemos que a una inmobiliaria llegan una serie de compradores potenciales interesándose por la compra de vivienda. Pensemos que el acto final de acabar comprando o no la vivienda en cuestión puede ponerse en relación con (depende de) una serie de caracteríscas de los individuos que manifiestan el interés de comprar como pueden ser la propia posesión (o no) de una primera vivienda, su proximidad a un posible enlace matrimonial o convivencia, sus ahorros y capacidad de endeudamiento, etc.; y supongamos hipotéticamente que pensásemos que la decisión de compra, de una forma muy simplificada, estaría básicamente relacionada con sólo dos características fácilmente observables como son la cantidad de dinero que pensaban dejar para pagar a plazos por la compra de la vivienda (X_1) y el número de años que tardarían en pagarla (X_2).

Si en los archivos de la inmobiliaria existen 49 casos de situaciones anteriores similares en las que, además de conocer estas dos características, también se conoce la decisión final sobre la compra que adoptaron los correspondientes clientes, ¿podríamos establecer un procedimiento que nos permitiera saber, en base a esa experiencia acumulada por la inmobiliaria, si sería muy probable que un nuevo cliente, que dice que aplazaría 11 millones de pesetas a pagar en 6 años para adquirir la vivienda, terminase comprando la vivienda? ¿o, si por el contrario, sería más probable que no la comprara?

Los métodos de Análisis Discriminante, junto a los más recientes basados en Modelos de Respuestas Cualitativas, son las técnicas estadísticas empleadas por excelencia para resolver este tipo de problemas y sus generalizaciones.

Dado un conjunto de individuos, de los que se conocen sus características, clasificados en k grupos diferentes, el Análisis Discriminante trata de establecer las relaciones óptimas existentes entre aquéllas características de los individuos y sus grupos de pertenencia; lo que permitiría clasificar (*identificar*) nuevos individuos, a partir de sus características observadas, en uno de aquéllos grupos y mediante una regla de decisión óptima que permitirá predecir la clasificación de los nuevos individuos de la forma más fiable posible

con respecto a la realidad.

La pertenencia de un individuo a un grupo se modeliza mediante una variable categórica que toma tantos valores como grupos haya y que también se conoce como variable grupo o variable dependiente.

Las características observadas a partir de las que se va a proceder a la identificación de los individuos se conocen como variables clasificadoras, variables criterios, variables predictioras o variables explicativas, exigiéndoseles generalmente en el Análisis Discriminante estar medidas en escalas de intervalo.

Y para obtener la relación óptima existente entre las características de los individuos y sus grupos de pertenencia pueden plantearse varias opciones. La opción que parte del establecimiento de un modelo similar al de regresión que nos permita explicar la variable categórica en función de las demás variables clasificadoras y la resolución de los problemas teóricos que plantea, conduce a los mencionados Modelos de Respuesta Cualitativa, merecedores de atención propia en otro capítulo

Aquí, vamos a referirnos estrictamente a las técnicas tradicionalmente empleadas para el Análisis Discriminante iniciadas por Fisher en 1936, y que desarrollamos a continuación.

CLASIFICACIÓN CON 2 GRUPOS

En esta situación, partimos de que la población se divide en 2 grupos o subpoblaciones, G_1 y G_2, sobre cuyos individuos se observan, en general, p variables $x=(X_1,X_2,...,X_p)'$. Y supongamos que, en cada grupo G_j, (j=1,2), la variable absolutamente continua $x=(X_1,X_2,...,X_p)'$ se distribuye según una cierta función de densidad de probabilidad $f_j(x)$. Además, representaremos por μ y Σ el vector de medias y la matriz de varianzas y covarianzas poblacionales y, análogamente, por μ_1, μ_2, Σ_1 y Σ_2, los correspondientes vectores de medias y matrices de varianzas y covarianzas de los respectivos grupos G_1 y G_2.

En estas circunstancias, el Análisis Discriminante trata de establecer alguna regla que relacione características y grupos, de forma que permita la identificación (clasificación) óptima de individuos en función de sus características. Y para ello, veamos a continuación los principales criterios empleados más generalmente para ello.

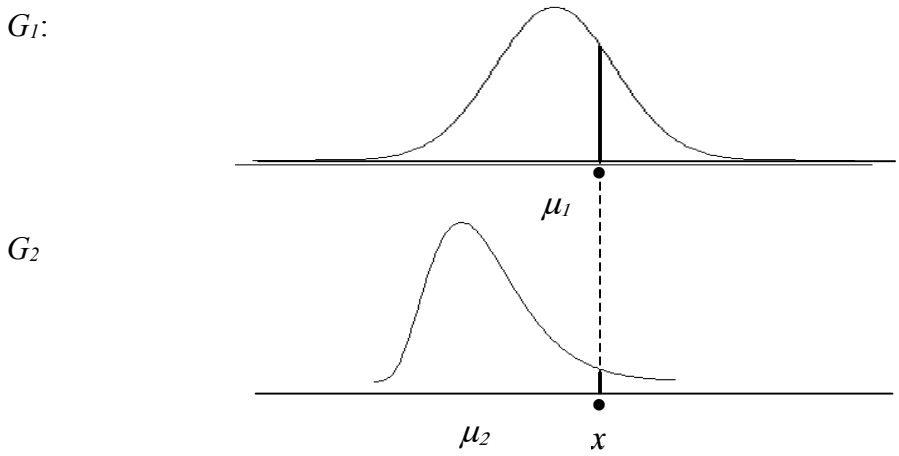
Regla Discriminante de Máxima Verosimilitud

El criterio de máxima verosimilitud siempre induce a considerar como solución del problema planteado aquélla que explique con máxima probabilidad lo que se observa en la realidad. Por tanto, la Regla de Máxima Verosimilitud aplicada al análisis discriminante para identificar (clasificar) un individuo de caracteristicas x en alguno de los 2 grupos existentes será:

$$\text{Asignar } x \text{ al grupo } G_1 \Leftrightarrow f_1(x) > f_2(x)$$

Es decir; la regla de máxima verosimilitud asigna el nuevo individuo, que presenta características x, al grupo G_j en el que dichas características presentan la máxima probabilidad o densidad de probabilidad.

Para ilustrar intuitivamente el proceder de esta regla, supongamos que tenemos una única característica unidimensional clasificadora continua de forma que, en los grupos G_1 y G_2 se distribuya y localice distintamente como aparece en los siguiente gráficos:



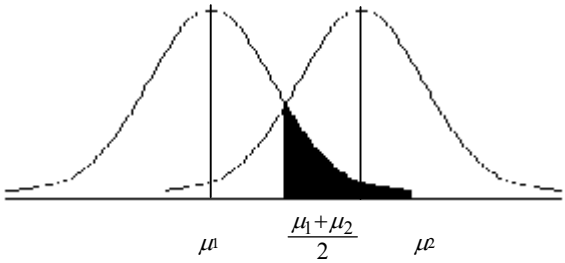
Como observamos en el gráfico, un individuo de característica x presenta una densidad de probabilidad en cada distribución de cada grupo. Así, la característica x del individuo en el grupo G_2 se encuentra en una zona muy improbable, por ser mayor de lo común en este grupo. Sin embargo, la característica x del individuo se encuentra en una zona más probable en el grupo G_1, ya que se encuentra más cercana a la moda. Así pues, la regla de máxima verosimilitud nos induciría a asignar los individuos que presentase característica x al grupo G_1, para el que la densidad de probabilidad en dicho valor de la característica, x, es más alta.

Caso Particular: 2 grupos normales univariantes con igual varianza (k=2, p=1)

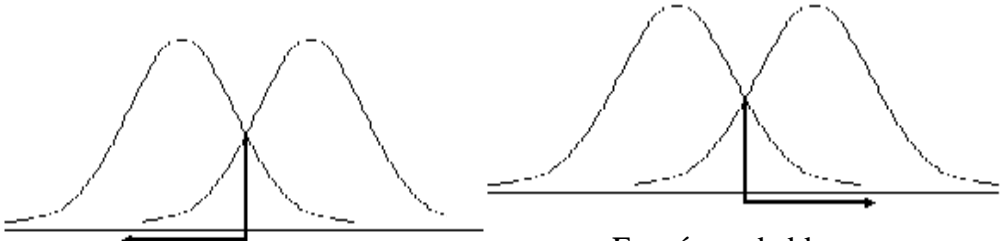
En este caso sólo habría dos grupos (k=2) para clasificar a los individuos en función de una única variable x (p=1). Además, supondremos que esta característica clasificadora x se distribuye normalmente en ambos grupos, con igual varianza σ pero con distintas medias μ_1 y μ_2.

Si suponemos, sin perdida de generalidad, que $\mu_1<\mu_2$, la función de densidad de la característica para el grupo G_1 se encontraría a la izquierda de la correspondiente función de densidad para el grupo G_2, ya que, en el grupo G_1 los valores más probables están alrededor de μ_1 y en el grupo G_2, alrededor de μ_2.

Existen valores de la característica para los que podemos encontrar individuos ubicados en cualquiera de los dos grupos, aunque con distintas verosimilitudes. Así, en la zona sombreada del siguiente gráfico es más probable que el individuo pertenezca al grupo G_2 que al G_1; mientras que en la zona simétrica es más probable que el individuo pertenezca al grupo G_1 que al grupo G_2.



Intuitivamente, de lo dicho hasta aquí, podemos deducir cual será la forma de decidir sobre la pertenencia a los grupos de un elemento que presenta una característica *x*. El eje de simetría del gráfico pasa por el valor promedio de μ_1 y μ_2 y es éste punto el valor crítico que separa las dos zonas de máxima verosimilitud de cada grupo. En la zona de la izquierda es más probable la pertenencia al grupo G_1 pues la función de densidad de en este grupo es siempre superior a la del G_2, y en la zona opuesta ocurre lo contrario. Luego cualquier individuo con característica a la izquierda de la línea vertical debe ser asignado al grupo G_1, y todo individuo con característica a la derecha de esta línea debe ser asignado al grupo G_2.



Analíticamente, sustituyendo la densidad de probabilidad en cada grupo G_j por la expresión correspondiente a la de una distribución normal univariante de media μ_j y desviación típica σ, en la expresión de la regla discriminante de máxima verosimilitud antes definida, ésta quedaría como:

$$\text{Asignar } x \text{ al grupo } G_1 \Leftrightarrow f_1(x) > f_2(x) \Leftrightarrow$$
$$\Leftrightarrow (\mu_1 - \mu_2)\cdot x > (\mu_1 - \mu_2)\cdot\left(\frac{\mu_1 + \mu_2}{2}\right)$$

Para demostrarlo, basta tener en cuenta que los dos grupos siguen distribuciones normales con la misma varianza (misma dispersión): el grupo G_1 se distribuye según una normal con media μ_1 y desviación típica σ y el grupo G_2 se distribuye según otra normal con media μ_2 y desviación típica también σ. Por tanto, las correspondientes funciones de densidad serán de la forma:

$$G_1 \to N(\mu_1,\sigma) \Rightarrow f_1(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2}$$
$$G_2 \to N(\mu_2,\sigma) \Rightarrow f_2(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2}$$

de donde

$$f_1(x) > f_2(x) \Leftrightarrow \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} > \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2} \Leftrightarrow$$

$$\Leftrightarrow -\tfrac{1}{2}\left(\tfrac{x-\mu_1}{\sigma}\right)^2 > -\tfrac{1}{2}\left(\tfrac{x-\mu_2}{\sigma}\right)^2 \Leftrightarrow (x-\mu_1)^2 < (x-\mu_2)^2 \Leftrightarrow -2\mu_1 x + {\mu_1}^2 < -2\mu_2 x + {\mu_2}^2 \Leftrightarrow$$

$$\Leftrightarrow -2x(\mu_1 - \mu_2) < {\mu_2}^2 - {\mu_1}^2$$

y, en consecuencia,

$$f_1(x) > f_2(x) \Leftrightarrow (\mu_1 - \mu_2)\cdot x > (\mu_1 - \mu_2)\cdot \frac{\mu_1 + \mu_2}{2} \Leftrightarrow \begin{cases} x < \dfrac{\mu_1 + \mu_2}{2} & \text{, si} \quad \mu_1 < \mu_2 \\ x > \dfrac{\mu_1 + \mu_2}{2} & \text{, si} \quad \mu_1 > \mu_2 \end{cases}$$

Como puede verse, si $\mu_1 < \mu_2$ obtenemos la misma regla deducida intuitivamente de la anterior representación gráfica, mientras que si $\mu_1 > \mu_2$, la regla obtenida sería la simétrica.

Caso Particular: 2 grupos normales p-variantes con igual matriz Σ de varianzas y covarianzas (k=2, p=p)

La distribución normal p-variante, que denotaremos como $NM(\mu, \Sigma)$, depende de su vector (columna) de medias μ, indicativo del punto de máxima densidad de probabilidad y centro de la distribución,

$$\mu = \begin{pmatrix} \mu_1 & \cdots & \mu_p \end{pmatrix}'$$

donde cada μ_i representa la media de la marginal X_i, así como de la matriz, semidefinida positiva, de varianzas y covarianzas representada por

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \ddots & & \vdots \\ & & \ddots & \vdots \\ \sigma_{1p} & \cdots & \cdots & \sigma_p^2 \end{pmatrix}$$

compuesta como sabemos por las varianzas poblacionales de las marginales en la diagonal y, simétricamente en los triángulos superiores e inferiores, en el resto las covarianzas poblacionales.

Consideremos su función de densidad $f_p(x)$, donde x en un vector columna de dimensión p cuyas componentes representan los valores de las p variables marginales, normales univariantes, $x=(X_1, X_2, ...,X_p)'$ que constituyen la normal p-variante. Su expresión es:

$$f_p(x) = (2\pi)^{-p/2}\left|\Sigma\right|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}$$

Obsérvese que cuando p=1 esta expresión coincide con la correspondiente función de densidad de la normal unidimensional que conocemos.

Análogamente a como hemos procedido para una dimensión, la regla de máxima verosimilitud para asignar un individuo a uno de dos grupos (*k*=2), en base a *p* variables de que conjuntamente siguen una distribución normal multivariante en cada uno de los dos grupos y con idéntica matriz de varianzas y covarianzas, diferenciándose únicamente en los vectores de medias (que para el grupo G_1 sería μ_1 y para G_2 sería μ_2), nos dirá que se "el individuo sea asignado al grupo G_1 si y sólo si $f_1(x) > f_2(x)$".

Así, sustituyendo la expresión de la función de densidad y operando obtendremos que:

Un individuo de características x debe asignarse al grupo 1 si y solo si:

$$f_1(x) > f_2(x) \Leftrightarrow (x-\mu_1)'\Sigma^{-1}(x-\mu_1) < (x-\mu_2)'\Sigma^{-1}(x-\mu_2) \Leftrightarrow$$

$$\Leftrightarrow (\mu_1-\mu_2)'\Sigma^{-1}x > (\mu_1-\mu_2)'\Sigma^{-1}\left(\frac{\mu_1+\mu_2}{2}\right) \Leftrightarrow$$

$$\Leftrightarrow \alpha' x > \alpha'\left(\frac{\mu_1+\mu_2}{2}\right) \quad , \textit{siendo } \alpha' = (\mu_1-\mu_2)'\Sigma^{-1}$$

Como vemos, la regla lineal discriminante presenta una frontera (hiperplano) de división $\alpha' x = 0{,}5\cdot\alpha'(\mu_1+\mu_2)$ entre los puntos que se asignarían al grupo G_1 y los puntos que se asignarían al grupo G_2, , que es una función lineal. Esta recta, divide el espacio en dos zonas (semiespacios): el subespacio $\alpha' x > 0{,}5\cdot\alpha'(\mu_1+\mu_2)$ de los asignables al grupo G_1) y el subespacio $\alpha' x < 0{,}5\cdot\alpha'(\mu_1+\mu_2)$ de los asignables al grupo G_2.

Criterio geométrico de la Distancia D^2 de Mahalanobis

Con la notación empleada, Mahalanobis define su distancia de un individuo de características x a un grupo cuyos centroide y matriz de varianzas y covarianzas son respectivamente μ y Σ, como:

$$D^2(x,G) = (x-\mu)'\Sigma^{-1}(x-\mu)$$

Este criterio intuitivo consiste en asignar el individuo de características x al grupo más cercano según esta medida. Es decir,

$$\text{Asignar } x \text{ al grupo } G_1 \Leftrightarrow D^2(x,G_1) < D^2(x,G_2) \Leftrightarrow$$

$$\Leftrightarrow (x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) < (x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)$$

lo que conduce, en el caso de que los grupos sean homocedásticos, a la misma regla discriminante deducida para el caso de 2 grupos normales

$$\text{Asignar } x \text{ al grupo } G_1 \Leftrightarrow D^2(x,G_1) < D^2(x,G_2) \Leftrightarrow$$
$$\Leftrightarrow (x-\mu_1)'\Sigma^{-1}(x-\mu_1) < (x-\mu_2)'\Sigma^{-1}(x-\mu_2) \Leftrightarrow$$
$$\Leftrightarrow (\mu_1-\mu_2)'\Sigma^{-1}x > (\mu_1-\mu_2)'\Sigma^{-1}\left(\frac{\mu_1+\mu_2}{2}\right) \Leftrightarrow$$
$$\Leftrightarrow \alpha' x > \alpha'\left(\frac{\mu_1+\mu_2}{2}\right) \quad , \textit{siendo } \alpha' = (\mu_1-\mu_2)'\Sigma^{-1}$$

Como en la práctica los vectores y matrices μ_1, μ_2, Σ_1 y Σ_2, no serán conocidos, se estiman muestralmente quedando la *Regla Lineal Discriminante* como la introdujo *Fisher* en 1936 y cuyas propiedades maestrales estudiaron posteriormente Wald y Anderson en 1944:

$$\text{Asignar } x \text{ al grupo } G_1 \Leftrightarrow \alpha' x > \alpha'\left(\frac{\bar{x}_1+\bar{x}_2}{2}\right) \quad , \textit{siendo } \alpha' = (\bar{x}_1-\bar{x}_2)'S^{-1}$$

Regla Discriminante de Bayes

Supongamos, como hasta ahora, que la población se divide en 2 grupos o subpoblaciones, G_1 y G_2, sobre cuyos individuos se observan, en general, p variables absolutamente continuas $x=(X_1,X_2,...,X_p)'$. Y supongamos que, en cada grupo G_j, (j=1,2), la variable $x=(X_1,X_2,...,X_p)'$ se distribuye según una cierta función de densidad de probabilidad $f_j(x)$.

Si son π_1 y π_2, con $\pi_1+\pi_2=1$, las probabilidades a priori de pertenencia a cada uno de los respectivos grupos, entonces, aplicando el Teorema de Bayes, las probabilidades a posteriori de que un individuo I que presenta características x pertenezca a cada uno de los grupos será:

$$\Pr ob(I \in G_j \mid x) = \frac{f_j(x)\pi_j}{f_1(x)\pi_1 + f_2(x)\pi_2} \quad , j = 1 \text{ ó } 2$$

lo que nos induce a actuar de la siguiente forma:

Un individuo de características x debe asignarse al grupo 1 si y solo si:

$$\Pr ob(I \in G_1 \mid x) > \Pr ob(I \in G_2 \mid x) \Leftrightarrow \frac{f_1(x)\pi_1}{f_1(x)\pi_1 + f_2(x)\pi_2} > \frac{f_2(x)\pi_2}{f_1(x)\pi_1 + f_2(x)\pi_2} \Leftrightarrow$$
$$\Leftrightarrow f_1(x)\pi_1 > f_2(x)\pi_2$$

Y si la casificación errónea del individuo llevase algún coste asociado, $c_{2|1}$ si el individuo es realmente del grupo 1 y se clasifica en el 2 y $c_{1|2}$ si el individuo es realmente del grupo 2 y se clasifica en el 1, (no produciendo costes la clasificación correcta) podemos plantear la regla de asignar un individuo al grupo que proporciona un menor coste esperado de

clasificación errónea, siendo el coste esperado de clasificar un individuo en un grupo el siguiente,

$$E\left[Coste(I \to G_1 \mid x)\right] = \frac{c_{1|2} f_2(x)\pi_2}{f_1(x)\pi_1 + f_2(x)\pi_2}$$
$$E\left[Coste(I \to G_2 \mid x)\right] = \frac{c_{2|1} f_1(x)\pi_1}{f_1(x)\pi_1 + f_2(x)\pi_2}$$

En este caso, la Regla Discriminante de Bayes diría:

Un individuo de características x debe asignarse al grupo 1 si y solo si:

$$E\left[Coste(I \to G_1 \mid x)\right] < E\left[Coste(I \to G_2 \mid x)\right] \Leftrightarrow$$
$$\Leftrightarrow \frac{c_{1|2} f_2(x)\pi_2}{f_1(x)\pi_1 + f_2(x)\pi_2} < \frac{c_{2|1} f_1(x)\pi_1}{f_1(x)\pi_1 + f_2(x)\pi_2} \Leftrightarrow \frac{f_1(x)\pi_1}{c_{1|2}} > \frac{f_2(x)\pi_2}{c_{2|1}}$$

Obsérvese que, a igualdad de costes y de probabilidades a priori, la Regla Discriminante de Bayes también coincide con la de Regla Discriminante de Máxima Verosimilitud.

Función Lineal Discriminante Canónica

Obsérvese que las reglas discriminantes proporcionadas por los métodos anteriores producen la división del espacio de características de los individuos en dos subespacios favorables a sendos grupos. Cuando los grupos son normales y homocedásticos, la partición del espacio la realiza una función lineal de la forma:

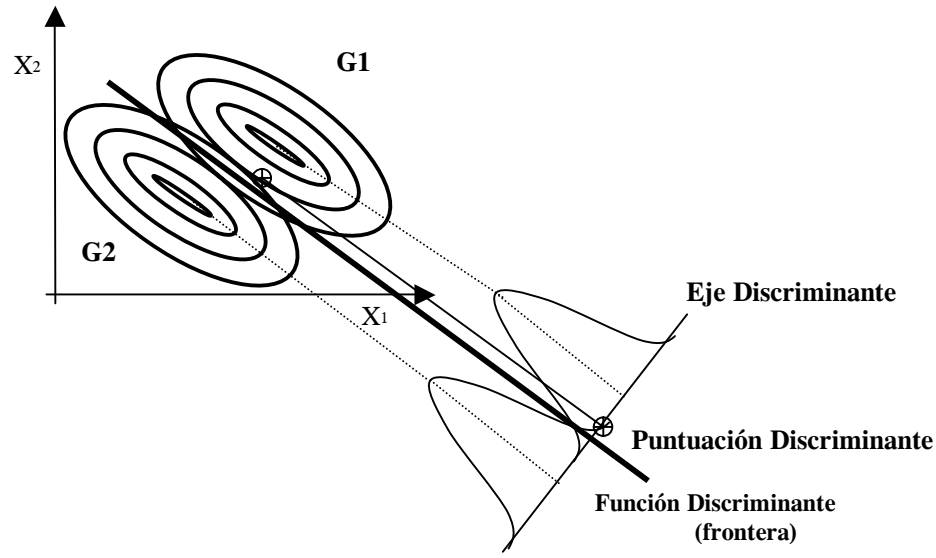
$$\alpha' x > \alpha' \left(\frac{\mu_1 + \mu_2}{2} \right) \quad , \textit{siendo } \alpha' = (\mu_1 - \mu_2)' \Sigma^{-1}$$

Llamaremos *función lineal discriminante* a la función $D(x) = \alpha' x$ que nos proporciona la regla de división y que permite proyectar todos los puntos del espacio sobre su dirección normal y característica α.

Consecuentemente, llamaremos *puntuación discriminante* de un individuo al valor que toma la función discriminante para dicho individuo en función de sus características $d(x) = \alpha' x$.

Se observa que todos los puntos situados sobre un hiperplano paralelo al hiperplano frontera presentan la misma puntuación discriminante. En consecuencia, llamaremos *eje discriminante* al eje (dirección) sobre el que estamos representando cada puntuación discriminante d como proyección de los todos los individuos del mismo hiperplano $\alpha' x = d = cte$ construido a partir de la función discriminante.

Se presenta a continuación una representación intuitiva, en dos dimensiones, de la situación de los datos, del hiperplano (línea recta) frontera, de los semiespacios (semiplanos) y del proceso de identificación que inducen éstas reglas discriminantes:



Zonas más densas del plano proporcionarán puntuaciones factoriales más frecuentes, y viceversa; por lo que sobre el eje discriminante podemos construir la distribución de puntuaciones discriminantes que presentarán mayor o menor densidades de probabilidad según las zonas proyectadas y que se encuentra igualmente representada.

Si en esta dirección los grupos se aprecian separados, las distribuciones de las puntuaciones factoriales (proyecciones de aquéllos) también lo estarán; y podemos intuir que cuanto más pronunciado (más bajo) sea el "valle" que se aprecie en una determinada dirección, mejor discriminación de los grupos se obtendrá, ya que se estará consiguiendo un menor solape de las campanas de Gauss que expresan los comportamientos normales en los dos grupos y, consecuentemente, de las distribuciones de puntuaciones discriminantes.

Esta idea será la que subyace en la aparición de la llamada *Función Lineal Discriminante Canónica*: buscar ser justamente la función lineal cuya dirección normal característica (vector de coeficientes) sea aquélla sobre la que se proyectan las puntuaciones discriminantes de forma óptima para dividir, separar o se distinguir más claramente los grupos.

La situación relativa de las distribuciones de puntuaciones discriminantes de los grupos puede ser muy diferente, según el eje (o dirección) sobre el que proyectemos dichas puntuaciones. Así, si imaginásemos las proyecciones que obtendríamos en las direcciones de los ejes coordenados, podríamos ver que producirían un mayor solape entre las distribuciones de puntuaciones discriminantes de los dos grupos G_1 y G_2, que la proyección representada en el gráfico.

Pues bien, la idea de la Función Lineal Discriminante Canónica es encontrar cuál es el eje discriminante (dirección de proyección) que nos permite reconocer más claramente los grupos que estamos considerando. Y para ello utiliza uno de los resultados fundamentales del Análisis de la Varianza, empleado también por técnicas de clasificación no jerárquicas,

que dice que, cuando un conjunto de datos se particiona en un cierto número de grupos , *k*, la matriz de varianzas y covarianzas observada del conjunto original de datos, *S*, puede descomponerse como la suma de la llamada *matriz de varianzas y covarianzas intra-grupos, W,* más la llamada *matriz de varianzas y covarianzas inter-grupos, B*, según el siguiente resultado:

$$S = \frac{1}{n} X_c^{'} X_c = B + W$$

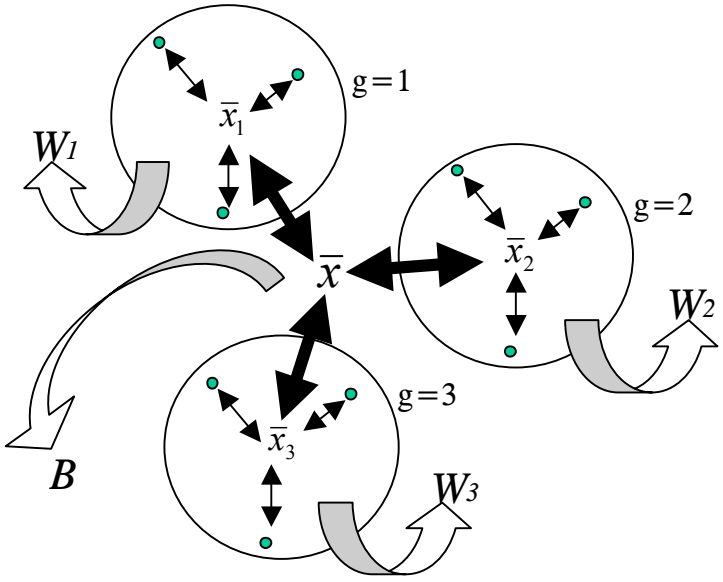
siendo el elemento genérico de fila *i* y columna *j* de estas matrices, *S, B* y *W* respectivamente, los siguientes:

$$S_{ij} = \frac{1}{n} \sum_{g=1}^{k} \sum_{l=1}^{n_g} (x_{gl,i} - \bar{x}_i)(x_{gl,j} - \bar{x}_j)$$

$$B_{ij} = \frac{1}{n} \sum_{g=1}^{k} n_g (\bar{x}_{g,i} - \bar{x}_i)(\bar{x}_{g,j} - \bar{x}_j)$$

$$W_{ij} = \frac{1}{n} \sum_{g=1}^{k} \sum_{l=1}^{n_g} (x_{gl,i} - \bar{x}_{g,i})(x_{gl,j} - \bar{x}_{g,j})$$

donde *n* representa el número de datos; X_c, la matriz de los datos centrados; $\bar{x}_i$, la media de la variable *i* en el conjunto de datos; n_g, el número de elementos del grupo *g*-ésimo; $\bar{x}_{g,i}$, la media de la variable *i* en el grupo *g*-ésimo; y $x_{gl,i}$, el valor que presenta el individuo *l*-ésimo del grupo *g*-ésimo para la variable *i*. Lo que gráficamente puede esquematizarse en el siguiente gráfico:



La dispersión dentro de cada grupo produce una matriz de varianzas y covarianzas W_i, indicativa de la homogeneidad interna en el grupos. Y sumando éstas para todos los grupos se obtendría la matriz de varianzas y covarianzas intra-grupos *W*, indicativa de la homogeneridad interna de los grupos en conjunto.

Por su parte, la matriz *B* o matriz de varianzas inter-grupos, se obtendría como varianza de

los centroides de los grupos con respecto al centroide de todos los datos en conjunto, ponderando aquéllos por el número de elementos del grupo al que corresponde y representa. Por tanto, la varianza inter-grupos es indicativa del la separación o heterogeneidad entre los grupos.

Por tanto, a medida que una determinada distribución en grupos de los objetos "aumenta" en algún sentido la matriz *W*, también hará "disminuir" en algún sentido la matriz *B*, ya que el conjunto de objetos, sus datos y la correspondiente matriz de varianzas y covarianzas *S* permanecen inalterados; y viceversa.

Dicho esto y volviendo a la idea intuitiva de la Función Lineal Discriminante Canónica, nuestro objetivo será encontrar este eje discriminante que nos permita la máxima diferenciación, la máxima claridad a la hora de decidir como asignar un individuo a un grupo o a otro; lo que se conseguirá "aumentando" en algún sentido la matriz *B*, con lo que consecuentemente "disminución" de la matriz *W*.

Obtención de la (primera) Función Lineal Discriminante Canónica

Para formalizar este problema supongamos, sin pérdida de generalidad, que las variables clasificadoras $X_{i\ ,\ i=1,2,\ldots,p}$ están centradas. Con ello se persigue simplificar las cosas y que el centro de la nube de puntos (valor de referencia) sea el cero.

Se propone como Función Lineal Discriminante la forma lineal:

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

donde los coeficientes $u_1,\ldots,u_p$ constituirán justamente el vector $u = (u_1,\ldots,u_p)'$ que identifica el eje discriminante y que pretendemos encontrar de forma que proporcione la proyección que mejor nos permita identificar los casos.

Con la notación anterior y si, como de costumbre, llamamos *X* a la matriz de datos, el cálculo de las puntuaciones discriminantes para cada caso, que se realiza sustituyendo las coordenadas de cada individuo en la función D anterior,

$$\begin{aligned} d_1 &= u_1\, x_{11} + \cdots + u_p\, x_{1p} \\ d_2 &= u_1\, x_{21} + \cdots + u_p\, x_{2p} \\ &\cdots \\ d_n &= u_1\, x_{n1} + \cdots + u_p\, x_{np} \end{aligned}$$

puede expresarse matricialmente como

$$d = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix} = X \cdot u$$

La distribución de D expresada por el vector de puntuaciones discriminantes de todos los puntos, *d*, será, consecuentemente, centrada en el origen, por ser una combinación lineal de

variables centradas.

Y su varianza, al ser D una variable centrada, se puede expresar directamente como su momento de segundo orden, lo que conduce a que:

$$Var(D) = \frac{1}{n} d'd = \frac{1}{n}(Xu)'Xu = \frac{1}{n}u'X'Xu = u'\left(\frac{1}{n}X'X\right)u = u'Su = u'(B+W)u = u'Bu + u'Wu$$

siendo $S = \frac{1}{n}X'X$ la matriz de varianzas y covarianzas de los datos que se ha podido como suma de las matrices de varianzas y covarianzas inter-grupos (B) e intra-grupos (W) y *u*, el vector dirección del eje discriminante.

De manera que la dispersión, la varianza, de las puntuaciones discriminantes se puede descomponer en dos sumandos: $u'Bu$ y $u'Wu$ que dependerán de la dirección del eje discriminante y de la agrupación de individuos que ésta determine, ya que de ella dependen las matrices *B* y *W*.

Así pues, la mejor dirección de proyección que vamos buscando será aquella que propicie que la dispersión dentro de cada grupo sea lo menor posible y que la separación entre grupos sea lo mayor posible; lo que en términos de la varianza de las puntuaciones discriminantes y de las matrices *B* y *W*, significa que la parte de varianza que se debe a la dispersión entre grupos (sumando $u'Bu$) sea lo mayor posible y la parte de varianza que se debe a la dispersión dentro de los grupos (sumando $u'Wu$) sea lo menor posible.

Teniendo en cuenta todo este razonamiento, la dirección (u_1,⋯,u_p) del eje discriminante que propone la Función Lineal Discriminante canónica será aquélla que haga máxima la expresión:

$$\lambda = \underset{u\in\Re^p}{Max}\left\{\frac{u'Bu}{u'Wu}\right\} = \underset{u\in\Re^p}{Max}\left\{\frac{\text{Varianza inter - grupos}}{\text{Varianza intra – grupos}}\right\}$$

La solución de este problema basta derivar e igualar a cero, para así obtener los puntos estacionarios, y comprobar cuál es el máximo; para lo que derivando el cociente, vectorialmente, e igualando a cero obtenemos:

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

donde la matriz W es normalmente definida positiva y va a poder invertirse (si no hay variables completamente redundantes, en cuyo caso sería sólo semidefinida positiva y deberemos eliminar la redundancia para poder proceder de acuerdo a este enfoque) y λ es el valor que toma el cociente a maximizar en la solución.

Teniendo esto en cuenta, las posibles soluciones al problema verificarán la siguiente expresión:

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

de donde, posibles direcciones óptimas u serán todos los autovectores (vectores propios o característicos) de la matriz $W^{-1}B$, siendo λ, valores que toma el cociente a maximizar en cada dirección *u*, sus respectivos autovalores (valores propios) asociados. Así, la dirección del eje discriminante canónico que buscamos sería la de aquel autovector que haga máximo este valor de λ, y por tanto, que esté asociado al máximo autovalor de la matriz $W^{-1}B$. Es decir, encontrando el mayor de los autovalores de esta matriz, su autovector (o vector propio) asociado marcará la dirección del eje discriminante que pretendíamos encontrar.

Procedimiento general de identificación en el caso de 2 grupos

Una vez calculada la puntuación discriminante (valor de D) para el individuo que se pretende clasificar (identificar), debemos compararlo con algún valor crítico que delimite las puntuaciones bajas indicativas de pertenencia a un grupo, de las puntuaciones altas indicativas de la pertenencia al otro grupo. A este punto se le conoce como *Punto de Corte Discriminante*.

Bajo la condición de que las distribuciones grupales sean unimodales, simétricas respecto de sus centroides y homocedásticas (igual matriz de varianzas y covarianzas en los grupos), la distribución de proyecciones discriminantes se manifiesta simétrica y el *Punto de Corte Discriminante*, *C*, puede situarse intuitivamente a medio camino entre las puntuaciones discriminantes D_1 y D_2 de los centroides de los grupos G_1 y G_2, respectivamente:

$$C = \frac{D_1 + D_2}{2}$$

Esta regla suele aplicarse de forma aproximada en situaciones de simple unimodalidad y homocedasticidad de los grupos. En otros casos, el *Punto de Corte Discriminante* debe determinarse mediante el examen de la variación de los errores de clasificación muestral cuando se varia la elección de dicho punto de corte.

Resumidamente, si es D la función lineal discriminante canónica y suponemos, sin perdida de generalidad que las puntuaciones discriminantes en los centroides de los grupos G_1 y G_2 se sitúan de forma que $D_1 < D_2$, el procedimiento general de identificación será:

1º Calcular el Punto de Corte Discriminante
Bajo condiciones de unimodalidad y homocedasticidad grupal:
$C=(D_1+D_2)/2$

2º Se calcula la puntuación discriminante del individuo "i" a clasificar, D_i

3º Si $D_i-C<0$, el individuo se asigna a G_1
En caso contrario, es decir, si $D_i-C>0$, el individuo se asigna a G_2

Observamos que la regla es similar a la que obteníamos con la regla de máxima verosimilitud, y con la regla de la distancia de Mahalanobis. De hecho se puede demostrar que, en el caso de 2 grupos normales homocedásticos con hasta 2 variables clasificadoras (los casos básicos que hemos visto), la Función Lineal Discriminante Canónica coincide con las Funciones Lineales Discriminantes obtenidas por aquellos métodos. En otros casos no tienen porqué coincidir.

CLASIFICACIÓN CON k GRUPOS (k>2)

Regla Discriminante de Máxima Verosimilitud

Partamos ahora de que la población se divide en *k* grupos o subpoblaciones, $G_1, G_2, \ldots, G_k$ y sobre cuyos individuos se observan *p* variables $x=(x_1,x_2,\ldots,x_p)$.

Supongamos que, en cada grupo $G_j, (j=1,2,\ldots,k)$, la variable $x=(x_1,x_2,\ldots,x_p)$ se distribuye con función de densidad de probabilidad $f_j(x)$.

En estas circunstancias, el criterio de máxima verosimilitud siempre induce a considerar como solución del problema planteado aquélla que explique con máxima probabilidad lo que se observa en la realidad. Por tanto, la Regla de Máxima Verosimilitud para identificar (clasificar) un individuo de caracteristicas *x* en alguno de los *k* grupos existentes será:

$$\text{Asignar } x \text{ al grupo } G_j \Leftrightarrow f_j(x) = \underset{g=1,2,\ldots,k}{Max}\{f_g(x)\}$$

Es decir; la regla de máxima verosimilitud asigna el nuevo individuo, que presenta características *x*, al grupo G_j en el que dichas características presentan la máxima probabilidad o densidad de probabilidad.

Criterio geométrico de la Distancia D^2 de Mahalanobis

Este criterio consistirá en asignar el individuo de características *x* al grupo más cercano según esta medida. Es decir,

$$\text{Asignar } x \text{ al grupo } G_h \Leftrightarrow D^2(x,G_h) = \underset{j=1,2,\ldots,k}{Min}\{D^2(x,G_j)\} \Leftrightarrow$$
$$\Leftrightarrow (x-\mu_h)'\Sigma_h^{-1}(x-\mu_h) < (x-\mu_j)'\Sigma_j^{-1}(x-\mu_j) \quad \forall j=1,2,\ldots,k \quad j \neq h$$

o, en el caso de que los grupos sean homocedásticos con matriz común de varianzas y covarianzas Σ:

$$\text{Asignar } x \text{ al grupo } G_h \Leftrightarrow D^2(x,G_h) = \underset{j=1,2,\ldots,k}{Min}\{D^2(x,G_j)\} \Leftrightarrow$$
$$\Leftrightarrow (x-\mu_h)'\Sigma^{-1}(x-\mu_h) < (x-\mu_j)'\Sigma^{-1}(x-\mu_j) \quad \forall j=1,2,\ldots,k \quad j \neq h$$

Regla Discriminante de Bayes

Si son π_i , i=1,2,….,k, con $\pi_1+\pi_2+\ldots+\pi_k=1$, las probabilidades a priori de pertenencia a cada uno de los respectivos grupos, entonces, aplicando el Teorema de Bayes, las probabilidades a posteriori de que un individuo I que presenta características *x* pertenezca a cada uno de los grupos será:

$$\Pr ob(I \in G_h \mid x) = \frac{f_h(x)\pi_h}{\sum_{j=1}^{k} f_j(x)\pi_j} \quad ,h = 1,2,...,\mathrm{k}$$

lo que nos induce a actuar de la siguiente forma:

Un individuo de características x debe asignarse al grupo *h* si y solo si:

$$\text{Asignar } x \text{ al grupo } G_h \Leftrightarrow \Pr ob(I \in G_h \mid x) = \underset{j=1,2,...,k}{Max} \left\{\Pr ob(I \in G_j \mid x)\right\} \Leftrightarrow$$

$$\Leftrightarrow \frac{f_h(x)\pi_h}{\sum_{j=1}^{k} f_j(x)\pi_j} > \frac{f_i(x)\pi_h}{\sum_{j=1}^{k} f_j(x)\pi_j} \quad \forall i = 1,2,...,k \quad i \neq h \Leftrightarrow$$

$$\Leftrightarrow f_h(x)\pi_h > f_i(x)\pi_i \quad \forall i = 1,2,...,k \quad i \neq h$$

Y si la casificación errónea del individuo llevase algún coste asociado, $c_{h|i}$ si el individuo es realmente del grupo *i* y se clasifica en el *h*, (no produce costes la clasificación correcta) podemos plantear la regla de asignar un individuo al grupo que proporciona un menor coste esperado de clasificación errónea, siendo el coste esperado de clasificar un individuo en un grupo el siguiente,

$$E\left[Coste(I \to G_h \mid x)\right] = \frac{\sum_{\substack{i=1 \\ i \neq h}}^{k} c_{h|i} f_i(x)\pi_i}{\sum_{j=1}^{k} f_j(x)\pi_j}$$

En este caso, la Regla Discriminante de Bayes diría:

Un individuo de características x debe asignarse al grupo *h* si y solo si:

$$E\left[Coste(I \to G_h \mid x)\right] = \underset{l=1,...,k}{Min}\, E\left[Coste(I \to G_l \mid x)\right] \Leftrightarrow$$

$$\Leftrightarrow \sum_{\substack{i=1 \\ i \neq h}}^{k} c_{h|i} f_i(x)\pi_i = \underset{l=1,...,k}{Min} \sum_{\substack{i=1 \\ i \neq l}}^{k} c_{l|i} f_i(x)\pi_i$$

Funciones Lineales Discriminantes Canónicas

Generalicemos el problema para cuando tenemos más de dos grupos, suponiendo que tendremos a los individuos clasificados en *k* grupos y observadas sobre ellos *p* variables clasificadoras.

Hemos visto que el problema de maximizar la la expresión:

$$\lambda = \underset{u \in \Re^p}{Max}\left\{\frac{u'Bu}{u'Wu}\right\} = \underset{u \in \Re^p}{Max}\left\{\frac{\text{Varianza inter - grupos}}{\text{Varianza intra – grupos}}\right\}$$

presentaba una solución (máximo absoluto) que venía dada por el autovector asociado al máximo autovalor.

Sin embargo, vimos que también eran puntos estacionarios de esta función el resto de los autovectores de la matriz $W^{-1}B$, algunos de los cuales podrían ser también máximos relativos, si no absolutos si el máximo autovalor fuese múltiple.

Si las magnitudes de los autovalores coinciden con los valores que toma la función objetivo en los puntos estacionarios (máximos, mínimos y puntos de silla), es obvio que los máximos deben ser los autovectores asociados a los mayores autovalores.

Basándonos en ello, podemos obtener varias Funciones Lineales Discriminantes canónicas de la forma

$$D_i = u_{i1}X_1 + ... + u_{ip}X_p$$

y cuyos vectores de coeficientes (dirección del eje discriminante) se obtienen de las soluciones del problema:

$$\underset{u_i}{Max}\frac{u_i' Bu_i}{u_i' Wu_i} = \lambda_i$$

que son, como puede demostrarse otra vez, los autovector u_i de $W^{-1}B$ y sus autovalores asociados λ_i.

Y como los autovalores λ_i son indicadores del poder discriminante del correspondiente eje discriminante definido por su autovector asociado u_i (cuanto mayor es un autovalor, mayor poder discriminante presentará su eje discriminante asociado y viceversa), entonces, ordenando los autovalores de menor a mayor, tendríamos sus correspondientes autovectores, como ejes discriminantes, ordenados de más a menos discriminantes.

Podemos observar además que si la matriz $W^{-1}B$ fuera simétrica, entonces siempre sería posible obtener un sistema de autovectores ortogonales (y ortonormales también). Pero, en esta aplicación al Análisis Discriminante, no tiene porqué ocurrir esto; es decir, al no tener que ser simétrica la matriz $W^{-1}B$, las direcciones de los ejes discriminantes no tienen porqué ser ortogonales pudiéndose cruzar con cualesquiera inclinaciones .

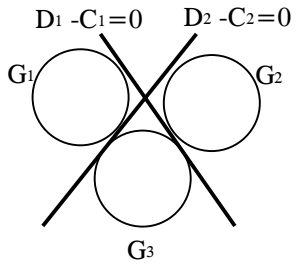
En cualquier caso, cada uno de los hiperplanos $D_i = C_i$, divide el espacio óptimamente (al menos en sentido relativo) en dos subespacios, siendo C_i el punto de corte discriminante determinado para la correspondiente función lineal discriminante (y que bajo la hipótesis de homocedasticidad se aproxima por la puntuación discriminante media de los centroides de los subespacios que separa).

Se puede demostrar que el número autovectores de la matriz $W^{-1}B$ que producen un

máximo (absoluto o relativo) de la función objetivo $\underset{u_i}{Max}\frac{u_i' B u_i}{u_i' W u_i} = \lambda_i$ es el mínimo entre el número de grupos menos uno, *k-1* y el número de variables *p;* por lo que:

el número máximo de ejes discriminantes=h = min{k-1,p}

Por ejemplo, si tenemos tres grupos y al menos dos variables , podremos establecer dos funciones lineales discriminantes, que serán rectas, como las que intuitivamente representamos en el siguiente gráfico:



de forma que los grupos pueden localizarse o caracterizarse por su posición relativa con respecto a las funciones discriminantes y, análogamente, podremos identificar a un individuo en uno u otro grupo.

En cualquier caso, el proceso de identificación se complica. Así, en el gráfico anterior, los elementos del grupo G_1 se caracterizan por cumplir que $D_2>C_2$; los elementos del grupo G_2 se caracterizan por cumplir que $D_1>C_1$; y los elementos del grupo G_3 se caracterizan por cumplir que $D_1<C_1$ y que $D_2<C_2$.

Como puede verse, la complejidad de uso de estas funciones clasificadoras crece potencialmente con el número de grupos.

Funciones Lineales Discriminantes Clasificadoras de Fisher

Manejar las reglas lineales discriminantes vistas hasta ahora es fácil cuando se tienen sólo dos grupos; algo menos cuando se tienen tres grupos; y bastante menos a medida que el número de grupos aumenta (cuatro, cinco,...)

Las funciones lineales discriminantes clasificadoras de Fisher tienen por objetivo eliminar esta dificultad de manejo creciente que tienen aparejado la utilización de las funciones lineales discriminantes canónicas. Constituyen pues un procedimiento alternativo en el que se definirá una función lineal clasificadora asociada a cada grupo, indicadora de la afinidad con grupo de los individuos con unas determinadas características observadas. De esta forma, comparando los valores indicativos de la afinidad del individuo con todos los grupos, podremos identificarlo con el que presente un resultado mayor. La complejidad de uso de estas funciones clasificadoras crece sólo linealmente con el número de grupos.

Si los grupos son homocedásticos, el criterio geométrico de la D^2 de Mahalanobis nos dice

que:

$$\text{Asignar } x \text{ al grupo } G_h \Leftrightarrow D^2(x,G_h) = \underset{j=1,2,\ldots,k}{Min}\left\{D^2(x,G_j)\right\} =$$

$$= \underset{j=1,2,\ldots,k}{Min}\left\{(x-\mu_j)'\Sigma^{-1}(x-\mu_j)\right\} = x'\Sigma^{-1}x + \underset{j=1,2,\ldots,k}{Min}\left\{-2\mu_j'\Sigma^{-1}x + \mu_j'\Sigma^{-1}\mu_j\right\} =$$

$$= x'\Sigma^{-1}x - 2\underset{j=1,2,\ldots,k}{Max}\left\{\mu_j'\Sigma^{-1}x - \frac{1}{2}\mu_j'\Sigma^{-1}\mu_j\right\} \Leftrightarrow$$

$$\frac{1}{2}\left(x'\Sigma^{-1}x - D^2(x,G_h)\right) = \underset{j=1,2,\ldots,k}{Max}\left\{\mu_j'\Sigma^{-1}x - \frac{1}{2}\mu_j'\Sigma^{-1}\mu_j\right\}$$

Bajo esta hipótesis y muestralmente, las Funciones Lineales Discriminantes Clasificadoras de Fisher se definen, para cada grupo G_g, g=1,2,…,k, como:

$$F^{(g)}(x) = v_{g1}X_1 + v_{g2}X_2 + \ldots + v_{gp}X_p - C_g = \bar{x}_g'S^{-1}x - \frac{1}{2}\bar{x}_g'S^{-1}\bar{x}_g = \frac{1}{2}\left(x'S^{-1}x - D^2(x,\bar{x}_g)\right)$$

De forma que cuanto más diste un elemento del centroide de un grupo, menor valor proporcionará para la función clasificadora de ese grupo, y viceversa. Así, cada una de ellas informa de la afinidad o grado de pertenencia relativa a su grupo asociado.

La identificación de un individuo con un grupo mediante estas funciones se realiza por el siguiente procedimiento,

1º Se calculan estas funciones, $F(1)_i$, $F(2)_i$, … , $F(k)_i$, para el individuo “i” a clasificar

2º Si $F(g)_i > F(j)_i$, para todos los demás grupos j, el individuo se asigna a al grupo g.

Así, si tenemos tres grupos, y un individuo presenta características clasificadoras $\tilde{x}$, entonces se evalúan las funciones $F^I(\tilde{x})$, $F^{II}(\tilde{x})$, y $F^{III}(\tilde{x})$, y asignaremos el individuo al grupo G_g tal que $F^g(\tilde{x}) = max\{F^I(\tilde{x}),\ F^{II}(\tilde{x}),\ F^{III}(\tilde{x})\}$. Es decir, la función que muestre una evaluación máxima indicará el grupo al que debe ser asignado el individuo.

Caso de 2 grupos

En este caso, las funciones clasificadoras de Fisher de cada uno de los grupos se relacionan de forma que su diferencia coincide con la función lineal discriminante canónica.

Así, si se construye la función lineal clasificadora para cada grupo, $F^{(1)}$ para el primer grupo y $F^{(2)}$ para el segundo grupo,

$$F^{(1)}=v_1X_1+v_2X_2+\ldots+v_pX_p\text{-C'} \text{ y } F^{(2)}=w_1X_1+w_2X_2+\ldots+w_pX_p\text{-C''},$$

entonces, se cumple que

$$D\text{-}C = D\text{-}\frac{D_1 + D_2}{2} \propto F^{(1)}\text{-}F^{(2)}$$

Concretamente, para SPSS:

$$D\text{-}C = D\text{-}\frac{D_1 + D_2}{2} = \frac{F^{(1)}\text{-}F^{(2)}}{D_1 - D_2}$$

La identificación de un individuo con un grupo mediante estas funciones se realiza por el siguiente procedimiento,

1º Se calculan estas funciones, $F^{(2)}_i$ y $F^{(1)}_i$, para el individuo "i" a clasificar
2º Si $F^{(2)}_i < F^{(1)}_i$, el individuo se asigna a G_1
Si $F^{(2)}_i > F^{(1)}_i$, el individuo se asigna a G_2

Determinación de las Funciones Discriminantes Significativas

El significado de cada autovalor λ_i asociado a cada función lineal discriminante canónica no es otro que el del cociente entre las varianzas inter-grupo e intra-grupo de las puntuaciones discriminantes sobre el correspondiente eje discriminante asociado, y que hace máximo este cociente .

$$\lambda_i = \frac{u_i' B u_i}{u_i' W u_i} = \frac{\text{varianza inter - grupos del eje discriminante i - ésimo}}{\text{varianza intra - grupos del eje discriminante i - ésimo}}$$

Su cociente respecto a la suma de los autovalores nos indica la importancia relativa de cada eje discriminante con respecto de los otros; es decir, es un indicador de la potencia discrimiante del mismo

$$\text{Importancia relativa del eje discriminante i - ésimo} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

Se llama coeficiente de correlación canónica asociado al eje discriminante i-ésimo a la raíz cuadrada del cociente entre varianza explicada por la diferencia entre los grupos (varianza inter-grupos) y la varianza total de las puntuaciones discrimiantes en dicho eje discriminante.

$$\rho_i = \sqrt{\frac{\text{varianza inter - grupos del eje discriminante i - ésimo}}{\text{varianza total del eje discriminante i - ésimo}}} = \sqrt{\frac{u_i' B u_i}{u_i' S u_i}} = \sqrt{\frac{\lambda_i}{1 + \lambda_i}}$$

Su cuadrado, denominado en nomenclatura del análisis de la varianza coeficiente η_i^2, representa pues la proporción de la varianza total de las puntuaciones discriminantes en dicho eje discriminante que se debe a la diferencia existente entre los grupos.

$$\eta_i^2 = \rho_i^2 = \frac{\text{varianza inter - grupos del eje discriminante i - ésimo}}{\text{varianza total del eje discriminante i - ésimo}} = \frac{u_i' B u_i}{u_i' S u_i} = \frac{\lambda_i}{1+\lambda_i}$$

La importáncia relativa nos permitía ordenar los ejes discrimiantes de más a menos discriminantes. El coeficiente de correlación canónica y el coeficiente eta al cuadrado nos permiten conocer hasta qué punto los ejes discriminantes relativamente más importantes son realmente discriminantes para los grupos existentes, indicándonos el grado de importancia real o significación de éstos ejes.

Contraste de significación de los ejes discriminantes

El contraste para determinar el número de ejes discriminantes significativos se basa en la distribución del estadístico Λ de Wilks que se calcula como:

$$\Lambda = \frac{|W|}{|S|} = \prod_{j=1}^{p} \left(\frac{1}{1+\lambda_j} \right)$$

ya que

$$\frac{1}{\Lambda} = \frac{|S|}{|W|} = |W^{-1}S| = |W^{-1}(W+B)| = |I + W^{-1}B| = \prod_{j=1}^{p} (1+\lambda_j)$$

Valores de Λ cercanos a 1 se obtienen cuando todos los autovalores λ_i son cercanos a cero, indicando un poder discriminante casi nulo de los ejes. Valores de Λ cercanos a 0 se obtienen cuando algunos de los autovalores λ_i son relativamente grandes, indicando un posiblemente importante poder discriminante.

Bajo las hipótesis habituales en el MANOVA de que las variables en cada uno de los k grupos se distribuyen según normales p-dimensionales con vectores de medias respectivos μ_i e igual matriz de varianzas y covarianzas Σ, es decir $X \sim N_p(\mu_i ; \Sigma)$ en cada grupo i-ésimo, y bajo la hipótesis nula de que todos los grupos poseen la misma media (o equivalentemente, que es imposible disccriminar los grupos), entonces el estadístico Λ de Wilks tiene una distribución, compleja de calcular, que recibe su nombre (distribución de Wilks) y que depende de 3 parámetros: la dimensión del espacio, p; los grados de libertad de W, n-k; y los grados de libertad de B, k-1.

$$\Lambda \xrightarrow{BajoH_0} \Lambda(p; n-k; k-1)$$

Son los términos $(1+\lambda_j)^{-1}$ los que informan del poder discriminante del eje j-ésimo, tanto más discriminantes cuanto menor es este valor, ya que

$$(1+\lambda_j)^{-1} = \frac{1}{1+\lambda_j} = \frac{1}{1+\frac{u_j' B u_j}{u_j' W u_j}} = \frac{u_j' W u_j}{u_j' W u_j + u_j' B u_j} = \frac{u_j' W u_j}{u_j' S u_j}$$

Así pues, se puede construir, por analogía, un estadístico similar que nos hable del poder discriminante de los últimos ejes discriminantes, que serán lógicamente los menos discriminantes, para tratar de determinar si son o no despreciables desde el punto de vista discriminante, reteniendo como significativos sólo los *r* primeros.

$$\Lambda_r = \prod_{j=r+1}^{p}\left(\frac{1}{1+\lambda_j}\right) \qquad \text{siendo } \Lambda_0 = \Lambda$$

Con todo esto, el contraste puede plantearse de la siguiente forma:

$$\begin{cases} H_0 : \lambda_{r+1} = 0 \quad \left(= \lambda_{r+2} = ... = \lambda_p\right) \\ H_1 : \lambda_{r+1} > 0 \end{cases}$$

Bajo las hipótesis habituales en el MANOVA de que las variables en cada uno de los k grupos se distribuyen según normales p-dimensionales con vectores de medias respectivos μ_i e igual matriz de varianzas y covarianzas Σ, es decir $X \sim N_p(\mu_i ; \Sigma)$ en cada grupo i-ésimo, y bajo esta hipótesis, puede demostrarse que:

$$\Lambda_r \xrightarrow{BajoH_0} \Lambda(p-r; n-k; k-r-1)$$

de donde aplicando la aproximación de Barlett[1] obtenemos que el estadístico experimental Λ_r se distribuye asintóticamente, bajo la hipótesis nula como:

$$-\left[n - \frac{p+k}{2} - 1\right] \ln \Lambda_r \xrightarrow[n\to\infty]{BajoH_0} \chi^2_{(p-r)(k-r-1)}$$

Evaluación de Errores

Hemos hablado del proceso de identificación o clasificación que inducen las distintas reglas discriminantes expuestas. Ahora trataremos de caracterizar nuestras decisiones con alguna medida que nos informa de la bondad o fiabilidad de las identificaciones realizadas.

Podemos evaluar los errores de clasificación a partir de la llamada *matriz de confusión*. Esta matriz de confusión enfrenta la información sobre la pertenencia a los distintos grupos que conocemos por los datos históricos del problema (grupos de pertenencia real observados) con los resultados que obtenemos si aplicamos la regla discriminante a evaluar sobre estos mismos datos históricos. Generalmente y en cada grupo real, encontraremos individuos correctamente clasificados e individuos mal clasificados que son asignados a grupos erróneos. La *matriz de confusión* puede describirse, entonces, como la tabla de contingencia de las características "grupo real" y "grupo asignado" para el conjunto de individuos

[1] La aproximación de Barlett para la distribución Λ de Wilks es: $-\left[n - \frac{s-t+1}{2}\right] \ln \Lambda(s; n; t) \xrightarrow{n\to\infty} \chi^2_{s \cdot t}$

histórico, como reflejamos a continuación.

Matriz de Confusión

		Grupos Asignados			
		1	2		K
Grupos reales	1	m_{11}	m_{12}		m_{1k}
	2	m_{21}	m_{22}		m_{2k}
	…	…	…		…
	k	m_{k1}	m_{k2}		m_{kk}

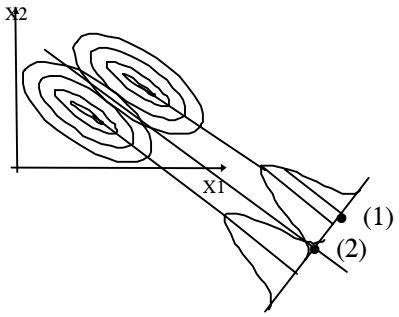
Si la regla discriminante presenta un buen comportamiento, los individuos históricos, quedarían bien clasificados; es decir, si realmente un individuo es del grupo G1, la regla discriminante debería asignarlo a ese grupo y análogamente si es del grupo G2. Y en consecuencia, si la regla discriminante funciona correctamente, la diagonal principal de la matriz de confusión presentaría frecuencias altas, mientras que los errores –que quedarán contabilizados en la diagonal secundaria– deberían ser escasos.

Así, a partir de esta matriz de confusión podemos establecer algunas medidas de error basadas en el cociente entre el número de casos bien (mal) identificados que se encuentra en la diagonal principal (fuera de la diagonal principal) y el número total de casos que estemos considerando como conjunto de referencia. Así podemos definir las siguientes tasas que nos informan de la proporción de casos mal clasificados, bien de una forma global, bien de cada grupo.

$$\text{Tasa de error Global}: \quad 1 - \frac{\sum_{i=1}^{k} m_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}}$$

$$\text{Tasa de error Parcial para el grupo i}: \quad 1 - \frac{m_{ii}}{\sum_{j=1}^{k} m_{ij}}$$

También nos interesa caracterizar la fiabilidad de cada identificación realizada. Es decir, evaluar qué garantías hay de que una asignación de un individuo de determinadas características a un grupo concreto sea correcta. Si miramos las distribuciones de las puntuaciones discriminantes de los grupos en el siguiente gráfico, los individuos 1 y 2 deben ser asignados al grupo G; pero estas las asignaciones no parecen igual de claras o fiables.



Lógicamente, parece que un individuo que está en (1) presenta mayor seguridad de pertenecer a G_1 que un individuo que esté en (2) pues la lejanía con respecto del grupo G_2 es bastante mayor para el primero que para el segundo.

Una forma de evaluar la fiabilidad de esa identificación la proporciona la aplicación del Teorema de Bayes que permite obtener la probabilidad a posteriori de que un cierto individuo pertenezca a cada grupo, conocida su puntuación discriminante D, a partir de la probabilidad o creencia a priori de pertenencia al grupo como se indica a continuación:

Probabilidad de pertenencia a un grupo

$$P(i \mid D) = \frac{\pi_i \cdot \Pr ob(D \mid i)}{\sum_{g=1}^{k} \pi_g \cdot \Pr ob(D \mid g)}$$

siendo π_i es la probabilidad a priori de pertenencia al grupo *i*, que puede evaluarse de distintas formas (igual para todos los grupos, proporcional a los tamaños de los grupos, de acuerdo con juicios expertos, subjetivamente,...), *Prob*(*D*|*i*) es la probabilidad de que un individuo del grupo *i* saque una puntuación discriminante D, probabilidad que se evalúa estudiando las puntuaciones discriminantes de los elementos conocidos del grupo *i*, y *Prob*(*i*|*D*) es la probabilidad a posteriori de pertenencia al grupo *i* calculada por la regla del Teorema de Bayes y que se tomará como indicadora de la fiabilidad de la identificación del individuo en el grupo i-ésimo.

Para evitar que el utilizar los mismos datos para construir la regla discriminante y evaluar los errores reclasificación mediante los procedimientos anteriores pueda sesgar la evaluación realizada de éstos, y siempre que el tamaño del conjunto de datos lo permita, se suele particionar los datos en dos subconjuntos: uno para la obtención de las reglas discriminantes y otro para probarlas y evaluar su eficacia.

Ejemplo:

A una inmobiliaria le interesa conocer la posibilidad de que un nuevo cliente que se interesa por la compra de una vivienda finalmente la compre, con el fin de orientar la actitud de su personal hacia el potencial cliente.

Supongamos hipotéticamente que la decisión de compra, de una forma deliberadamente muy simplificada en aras de la sencillez analítica e ilustrativa, estuviese básicamente relacionada con los millones de pesetas que el cliente pensaba dejar para pagar a plazos por la compra de la vivienda (X_1) y el número de años que tardaría en pagarlos (X_2), variables fácilmente observables por los agentes de la inmobiliaria ya que los clientes suelen contestar a ellas prácticamente en su primera visita.

En los archivos de la inmobiliaria existen 49 casos de situaciones anteriores similares en las que, además de conocer estas dos características, también se conoce la decisión final sobre la compra que adoptaron los correspondientes clientes (*variable Grupo*), con los resultados de la siguiente tabla:

Caso	X1	X2	Grupo
1	12	8	1
2	7	5	2
3	9	7	1
.	.	.	.
.	.	.	.
.	.	.	.
48	10	7	2
49	10	5	1

donde la variable Grupo presenta el valor 1 en los clientes que finalmente No Compraron, y el valor 2 si terminaron comprando la vivienda. Lógicamente, esta variable clasifica a los 49 individuos en dos grupos: el grupo de los que toman el valor 1 (los que no compraron la vivienda) y el grupo de los que toman el valor 2 (los que sí compraron la vivienda).

Si llegase un cliente que dice que aplazaría 11 millones de pesetas a pagar en 6 años, y según la experiencia acumulada por la inmobiliaria, ¿sería probable que terminase comprando la vivienda o, por el contrario, sería más probable que no la comprara?

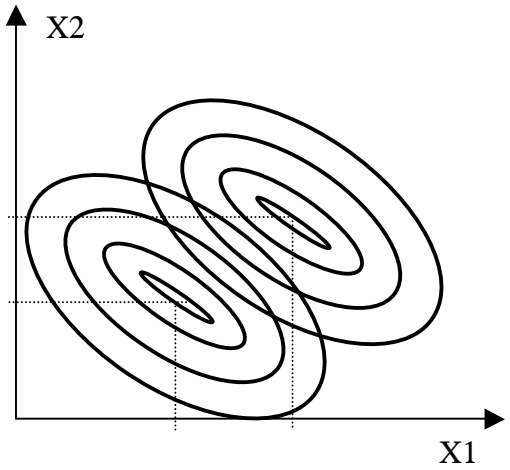
Disponemos de 49 individuos, dos variables explicativas (X1 y X2), por lo que p=2, y una variable categórica presentando 2 modalidades distintas, por lo que disponemos de los casos clasificados en k=2 grupos.

Antes de determinar la forma particular que adopta la regla de máxima verosimilitud en este caso, presentaremos a continuación los principales estadísticos resúmenes de los 49 casos observados y que básicamente nos informan de los vectores de medias (de los grupos y total) y de la matriz de varianzas y covarianzas de ambas variables clasificadoras, así como una aproximación intuitiva a la representación gráfica de los datos, de acuerdo con estos estadísticos y supuestos un comportamiento normal bivariante para (X1 y X2), con una misma matriz de varianzas y covarianzas, como exige la aplicación

de la regla de máxima verosimilitud:

(Medias)	No Compraron	Compraron
X1	11,49	8,50
X2	7,97	4,75
N	37	12

Matriz de Varianzas y Covarianzas	
6.5609	0.5377
0.5377	4.6930
Determinante:	30.5012



Examinados los 49 casos de la historia, se aprecia que hay 37 individuos en el grupo de los que no compraron y 12 en el de los que sí lo hicieron, a lo largo del período examinado (historia). Las medias de los que no compraron para las variables X1 y X2 son 11.49 y 7.97 respectivamente. Es decir, los que finalmente no compraron dijeron que iban a aplazar, por término medio 11,49 millones y que iban a aplazarlos por 7,97 años; mientras que los del grupo que compraron, por término medio aplazaron 8.50 millones y declararon que lo devolverían en 4,75 años. Y si a los mismos datos se les calculala matriz de varianzas y covarianzas común, tendríamos la anteriormente expuesta, en la que se observa cierta correlación positiva entre ambas variables calsificadoras: a más cantidad aoplazada, mayor tiempo para pagarla.

Supuesto el comportamiento normal bivariante para (X1 y X2), con una misma matriz de varianzas y covarianzas, la regla de máxima verosimilitud nos dice que:

Un individuo de características $\tilde{x}$ debe asignarse al grupo G_1 si y solo si:

$$f_1(x) > f_2(x) \Leftrightarrow \alpha' x > 0{,}5 \cdot \alpha'(\mu_1 + \mu_2) \quad , \textit{siendo } \alpha' = (\mu_1 - \mu_2)'\Sigma^{-1}$$

siendo

$$\alpha' = (\tilde{\mu}_1 - \tilde{\mu}_2)'\Sigma^{-1}$$

Inversa de la Matriz de Varianzas y Covarianzas	
0.1539	-0.0176
-0.0176	0.2151

Para calcular el vector α necesitamos la inversa de la matriz de varianzas y covarianzas, que existe ya que su determinante es no nulo y toma el valor 30,5012, y sería

$$\alpha' = (\tilde{\mu}_1 - \tilde{\mu}_2)'\Sigma^{-1} = \begin{pmatrix} 11{,}49 - 8.50 \\ 7.97 - 4.75 \end{pmatrix}' \begin{pmatrix} 6.5609 & 0.5377 \\ 0.5377 & 4.6930 \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} 2.99 & 3.22 \end{pmatrix} \begin{pmatrix} 0.1539 & -0.0176 \\ -0.0176 & 0.2151 \end{pmatrix}$$

$$= \begin{pmatrix} 0.4035 & 0.6400 \end{pmatrix}$$

mientras que el miembro derecho de la regla sería:

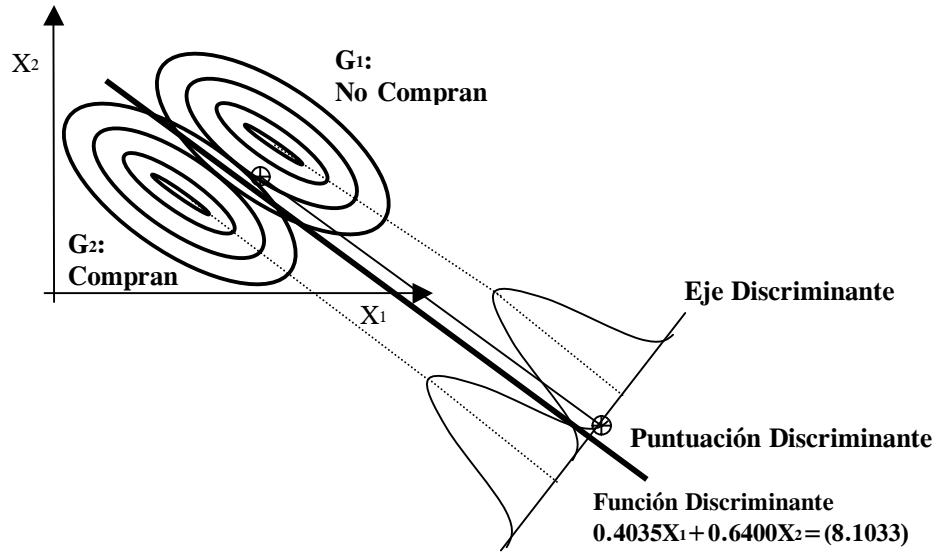
$$0.5\alpha'(\tilde{\mu}_1 + \tilde{\mu}_2) = 0.5\begin{pmatrix} 0.4035 & 0.6400 \end{pmatrix} \begin{pmatrix} 11.49 + 8.50 \\ 7.97 + 4.75 \end{pmatrix} = 0.5\begin{pmatrix} 0.4035 & 0.6400 \end{pmatrix} \begin{pmatrix} 19.99 \\ 12.72 \end{pmatrix} = 8.1033$$

de donde la **regla discriminante de máxima verosimilitud** para un individuo genérico $\tilde{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, inclinaría a asignar al individuo en el grupo G_1 si y solo si:

$$\alpha\tilde{x} = \begin{pmatrix} 0.4035 & 0.6400 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} > 8.1033 \Leftrightarrow 0.4035x_1 + 0.6400x_2 > 8.1033$$

En caso contrario (menor) lo asignaríamos al grupo G_2.

Como vemos, la regla discriminante presenta una frontera de división entre los puntos que se asignarían al grupo G_1 y los puntos que se asignarían al grupo G_2, $0.4035x_1 + 0.6400x_2 = 8.1033$, que es una función lineal y, en este caso concreto, una recta. Esta recta, divide el espacio en dos zonas (semiplanos): el de los asignables al grupo G_1 (los que cumplen que $0.4035x_1 + 0.6400x_2 > 8.1033$) y la de los asignables al grupo G_2 (los que cumplen que $0.4035x_1 + 0.6400x_2 < 8.1033$). Y representamos gráficamente la situación de los datos, junto con la línea frontera y ambos semiplanos, junto al proceso de identificación que induce esta regla, obtendríamos aproximadamente el gráfico que se presenta a continuación:



En él podemos observar la situación de los dos grupos: G_1, los que no compran – los que piden más cantidad a más tiempo y por tanto están más alejados del origen de coordenadas– , y G_2, los que terminan comprando. La función $0.4035x_1 + 0.6400x_2 = 8.1033$ muestra la división de los dos grupos en dos zonas.
Para identificar un individuo cualquiera, como el marcado en el gráfico, debemos evaluar la *Función Discriminante* $0.4035x_1 + 0.6400x_2$ en sus características, lo que nos proporcionaría un *Puntuación Discriminante,* que podemos representar sobre la dirección normal característica de la función discriminantes y que denominamos *Eje Discriminante*, que debemos comparar con valor de la función discriminante en la línea frontera (8.1033); de forma que si la puntuación discriminante es mayor –como es el caso del dibujo– asociaríamos el individuo al grupo G_1, y en caso contrario al grupo G_2.

¿Qué decir de un individuo que llega a la inmobiliaria interesándose por una vivienda para cuya adquisición aplazará 11 millones en 6 años?. Este individuo tendrá en nuestro epacio de características clasificadoras unas coordenadas x1=11 y x2=6.

Calculando la regla discriminante de máxima verosimilitud en su caso, tendremos que:

$$\alpha\widetilde{x} = \begin{pmatrix} 0.4035 & 0.6400 \end{pmatrix}\begin{pmatrix} 11 \\ 6 \end{pmatrix} = 8.2785 > 8.1033 = 0.5\alpha'(\widetilde{\mu}_1 + \widetilde{\mu}_2)$$

por lo que lo asignaríamos al grupo primero, G_1, de los que probablemente no comprarán.