ANALISIS DISCRIMINANTE

Jorge Galbiati R.

Objetivo del Análisis Discriminante

El objetivo del análisis Discriminante es proporcionar una *regla discriminante* que permita asignar un nuevo individuo u objeto a una de varias poblaciones, clases o grupos previamente identificados. La regla se obtiene a partir de una muestra, consistente en un conjunto de observaciones multivariantes, en que una de las variables es la población a la que pertenece cada observación. Existen varios métodos para obtener la regla discriminante.

Se tienen g poblaciones $\Pi_1, \Pi_2, ..., \Pi_g$, $g \geq 2$. Se tiene una observación definida por el vector $\underline{x}$ en el espacio de números reales p-dimensional $\Re$. Se debe decidir de cuál de estas poblaciones proviene esta nueva observación.

Una *regla discriminante* d es una partición del espacio $\Re$ en g regiones disjuntas $R_1, R_2, ..., R_g$ tales que $R_1 \cup R_2 \cup ... \cup R_g = \Re$, de modo que si $\underline{x} \in R_j$ entonces $\underline{x}$ proviene de la población Π_j.

Para obtener una regla discriminante se dispone de una matriz de datos X, llamada *muestra de aprendizaje*. Una de sus variables (columnas) es un factor, que indica la poblacióina que pertenece cada una de las observaciones. Esta se considera como la *variable dependiente* . Las demás son consideradas como *variables independientes*.

Veremos dos formas de obtener reglas discriminantes, que son *máxima verosimilitud clásico* y el procedimiento *discriminante canónico*.

MODELO DE ANALISIS DISCRIMINANTE DE MAXIMA VEROSIMILITUD CLASICO

Se asume que a cada población Π_j se asocia una distribución de probabilidad normal $N_p(\underline{\mu}_j, \Sigma_j)$, con media respectiva $\underline{\mu}_j$ y matriz de varianzas-covarianzas respectiva Σ_j para $j = 1, 2, ..., g$. La regla discriminante asigna la nueva observación $\underline{x}$ a aquella población en que la verosimilitud es máxima.

La log-verosimilitud (sin los términos constantes) de una observación cualquiera $\underline{x}$ proveniente de la j-ésima población normal $N_p(\underline{\mu}_j, \Sigma_j)$ es

$$l(\underline{x} \mid \underline{\mu}_j, \Sigma_j) = -\tfrac{1}{2} \log |\Sigma_j| - \tfrac{1}{2}(\underline{x} - \underline{\mu}_j)' \Sigma_j^{-1} (\underline{x} - \underline{\mu}_j)$$

que es equivalente a

$$l(\underline{x} \mid \underline{\mu}_j, \Sigma_j) = -\tfrac{1}{2} \log |\Sigma_j| - \tfrac{1}{2} \underline{x}' \Sigma_j^{-1} \underline{x} + \underline{\mu}_j' \Sigma_j^{-1} \underline{x} - \tfrac{1}{2} \underline{\mu}_j' \Sigma_j^{-1} \underline{\mu}_j \qquad (1)$$

La situación cambia de acuerdo a los supuestos que se hacen sobre la estructura de las matrices de varianzas-covarianzas Σ_j, lo que da origen a diversos casos, de acuerdo a la estructura de covarianzas.

1. Caso Heterosedástico

Asume que las matrices de varianzas-covarianzas son distintas. La función discriminante es la función de log-verosimilitud tal como se plantea en (1)

$$d_j(\underline{x}) = -\tfrac{1}{2}(\log|\Sigma_j| + \underline{\mu}_j'\Sigma_j^{-1}\underline{\mu}_j) + \underline{\mu}_j'\Sigma_j^{-1}\underline{x}_j - \tfrac{1}{2}\underline{x}_j'\Sigma_j^{-1}\underline{x}_j$$

Esta expresión, como función de $\underline{x}$, tiene la forma

$$d_j(\underline{x}) = \beta_{j0} + \underline{\beta}_{j1}'\underline{x} + \underline{x}'\beta_{j2}\underline{x}$$

En que el escalar β_{j0}, el vector $\underline{\beta}_{j1}$ y la matriz β_{j2} no dependen de $\underline{x}$.

Esta es una función cuadratica en el vector $\underline{x}$.

Se sustituyen Σ_j y μ_j por estimadores S_j y $\overline{x}_j$ respectivamente, a partir de las n_j observaciones que pertenecen a la población j-ésima de X.

De acuerdo al principio de máxima verosimilitud, se clasifica la nueva observación $\underline{x}$ en aquella población Π_j en que $d_j(\underline{x})$ es máximo.

2. Caso Homosedástico

Se supone que las matrices de varianzas-covarianzas son iguales: $\Sigma_1 = \Sigma_2 = ... = \Sigma_g$. Se usan todas las observaciones para estimar la varianza común, que denominaremos Σ, por medio del estimador S.

Aquí el término cuadrático $\frac{1}{2}\underline{x}'\Sigma_j^{-1}\underline{x}$ es constante para todas las poblaciones, luego para efectos de maximizar la función de verosimilitud, se descarta.

Las funciones discriminantes quedan

$$d_j(\underline{x}) = -\tfrac{1}{2}\underline{\mu}_j'\Sigma^{-1}\underline{\mu}_j + \underline{\mu}_j'\Sigma^{-1}\underline{x}$$

Que tiene la siguiente forma, con el escalar β_{j0} y el vector $\underline{\beta}_{j1}$ independientes de la observación $\underline{x}$,

$$d_j(\underline{x}) = \beta_{i0} + \underline{\beta}_{i1}'\underline{x}$$

Esta es una es una función lineal en el vector $\underline{x}$.

El caso heterosedástico es un caso general, mientras que el caso homosedástico es muy especial, en que se asume una gran cantidad de condiciones, que son $\Sigma_1 = \Sigma_2 = ... = \Sigma_g$. Hay casos intermedios, en que se imponen condiciones menos restrictivas sobre las matrices de varianzas-covarianzas.

3 Caso Covarianzas Proporcionales

En este caso se asumen las matrices de varianzas-covarianzas proporcionales, es decir,

$\Sigma_j = k_j^2 \Sigma_1$ para $j = 2, 3, ..., p$ en que Σ_1 es la matriz de varianzas-covarianzas de la primera poblacón.

La matriz Σ y las constantes k_j se estiman en forma iterativa.

Se reemplazan en las funciones discriminantes y se procede como en los casos anteriores.

4. Caso Matrices de Correlaciones Iguales

En este caso se asume que

$$\Sigma_j = D_j \Sigma_1 D_j \text{ con } D_j = diag\{k_{j1}, k_{j2}, ..., k_{jp}\} \qquad j = 2, 3, ..., p$$

en que Σ_1 es la matriz de varianzas-covarianzas de la primera poblacón. Σ_1 y los k_{jl} son parámetros que se estiman en forma iterativa.

5. Caso Esférico por Grupos

Asume que las variables son independientes, por lo tanto sus matrices de varianzas-covarianzas son diagonales, del tipo

$$\Sigma_j = diag\left\{\sigma_{j1}^2, \sigma_{j2}^2, ...\sigma_{jp}^2\right\} \qquad \text{para} \qquad j = 1, 2, ..., p$$

6. Caso Esférico

Es el caso más restrictivo. Es un caso especial del esférico por grupos, pero homosedástico, es decir, las matrices de varianzas-covarianzas de los grupos son todas iguales

$$\Sigma_j = diag\left\{\sigma_1^2, \sigma_2^2, ...\sigma_p^2\right\} \qquad \text{para} \qquad j = 1, 2, ..., g$$

7. Caso Componentes Principales Comunes

Asume que las matrices de varianzas-covarianzas Σ_j tienen los mismos vectores propios, pero se diferencian en sus valores propios. Si Γ es la matriz ortogonal de vectores propios de la matriz de varianzas-covarianzas de la primera poblacón Σ_1, y Λ_j es la matriz diagonal de valores propios de la matriz de varianzas-covarianzas de la j-ésima población Σ_j, $j = 1, 2,\ g,$, y si

$$\Lambda_j = diag\{\lambda_{j1}, \lambda_{j2}, ..., \lambda_{jp}\}$$

entonces se supone que

$$\Sigma_j = \Gamma \Lambda_j \Gamma'$$

8. Caso Componentes Principales con Covarianzas Proporcionales

Este es como el caso anterior, con la condición adicional de que los valores propios son proporcionales a los de la matriz de varianzas-covarianzas de la primera poblacón Σ_1.

Si éstos son $\lambda_{11}, \lambda_{12}, ..., \lambda 1p$, entonces los valores propios de Σ_j son

$$\lambda_{jr} = k_j^2 \lambda_{1r} \qquad j = 2, 3, ..., p, \qquad r = 1, 2, ..., g$$

Viene a ser un caso especial del caso 3, Covarianzas Proporcionales, en que todas las matrices de varianzas-covarianzas son proporcionales.

Los ocho casos vistos, en los que se estiman los parametros por máxima verosimilitud, bajo el supuesto de normalidad, se muestran en la siguiente tabla resumen, junto con el respectivo número de parámetros a estimar:

Caso	Estructura de Covarianzas	Número de Parámetros a estimar
Heterosedástico	$\Sigma_j \qquad j = 1, 2, .., g$	$gp(p+1)/2$
Correlaciones iguales	$\Sigma_j = D_j \Sigma_1 D_j$	$gp + p(p-1)/2$
Proporcional	$\Sigma_j = k_j^2 \Sigma_1$	$g - 1 + p(p+1)/2$
Esférico por grupos	$\Sigma_j = D_j \qquad$ diagonal	gp
Homosedástico	$\Sigma_j = \Sigma$	$p(p+1)/2$
Esférico	$\Sigma_j = D \qquad$ diagonal	p
C. Principales Comunes	$\Lambda_j = diag\{\lambda_{j1}, \lambda_{j2}, ..., \lambda_{jp}\}$	$gp + p^2$
C. Pr. Cov. proporcionales	$\lambda_{jr} = k_j^2 \lambda_{1r}$	$g - 1 + p^2 + p$

FUNCION DISCRIMINANTE CANONICA

Es un método totalmente diferente al de Máxima Verosimilitud. Se aplica sólo a la estructura de varianzas-covarianzas homosedástica, y no asume una distribución conocida para las poblaciones.

Para discriminar entre las poblaciones, se utiliza un conjunto de funciones lineales $\underline{a}_1'\underline{x}$, $\underline{a}_1'\underline{x}$,... tal que la primera maximiza el cuociente entre la suma de cuadrados entre grupos y la suma de cuadrados dentro de los grupos, en la muestra de entrenamiento.

La segunda maximiza lo mismo, pero en el espacio ortogonal a $\underline{a}_1$, la tercera igual, en el espacio ortogonal a $\underline{a}_1$ y a $\underline{a}_2$, y así.

La primera función es la que discrimina más, la segunda menos que la primera, la tercera menos que la segunda, etc. Se dice que la primera tiene mayor *poder discriminante*.

Lo normal es que se baste la primera de estas funciones para discriminar. A veces ésta no discrimina lo suficiente y se requiere la segunda para ayudar a discriminar bien. En algunos casos muy especiales puede requerirse la tercera, adems de las dos primeras.

El desarrollo siguiente corresponde conduce a la obtención de la primera función discriminante, que denominaremos $\underline{a}_1'\underline{x}$ para simplificar la notación. Después se generalizará con objeto de obtener las restantes funciones discriminantes.

Sea $X_{n\times p}$ la matriz de datos de la muestra de entrenamiento, que incluye solo las variables independientes (excluye la columna con el factor que indica la población).

Si X se particiona en la forma:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ X_g \end{bmatrix}$$

en que X_j es la submatriz de datos correspondiente a las observaciones de la población Π_j. Se define el vector:

$$\underline{y} = X\underline{a} = \begin{bmatrix} X_1\underline{a} \\ . \\ . \\ X_g\underline{a} \end{bmatrix} = \begin{bmatrix} \underline{y}_1 \\ . \\ . \\ \underline{y}_1 \end{bmatrix}_{n\times 1} \qquad \text{en que } n = \sum_{i=1}^{g} n_i$$

Entonces se define la suma de cuadrados total como el numerador de la matriz de varianzas-covarianzas S, pre y post multiplicado por $\underline{a}$

$$SCT = n\underline{a}'S\underline{a} = \underline{a}'X'(I - \tfrac{1}{n}\underline{1}\underline{1}')X\underline{a} = \underline{y}'(I - \tfrac{1}{n}\underline{1}\underline{1}')\underline{y} = \underline{y}'H\underline{y}$$

Esta suma de cuadrados se puede descomponer en dos partes:
La *suma de cuadrados dentro de los grupos* y la *suma de cuadrados entre grupos.*

$$SCT = SCD + SCE$$

Suma de cuadrados dentro de los grupos

Es una medida de la variabilidad existente entre las observaciones que están en un mismo grupo. Se define como la suma de los numeradores de las varianzas dentro de cada grupo

$$\begin{aligned} SCD = \sum_{j=1}^{g} n_j S_{y_j} &= \sum_{j=1}^{g} \underline{y}_j'(I - \tfrac{1}{n_j}\underline{1}\underline{1}')\underline{y}_j \\ &= \sum_j \underline{a}X_j'(I - \tfrac{1}{n_j}\underline{1}\underline{1}')X_j\underline{a} \\ &= \sum_j \underline{a}'X_j'H_jX_j\underline{a} = \underline{a}'D\underline{a} \end{aligned}$$

en que $D_{p\times p} = \sum_{j=1}^{g} X_j'H_jX_j$ y tiene rango mximo, p

H_j es la matriz de centrado $n_j \times n_j$. S_{y_j} es la matriz de varianzas-covarianzas muestral en el grupo j-ésimo.

Suma de cuadrados entre grupos

Mide la variabilidad entre los promedios de los grupos, cada promedio ponderado por el número de observaciones que contiene el grupo.

$$\begin{aligned} SCE = \sum_{j=1}^{g} n_j \left(\overline{y}_j - \overline{y}\right)^2 &= \sum_j n_j \left[\underline{a}'(\overline{x}_j - \overline{x})\right]^2 \\ &= \underline{a}'E\underline{a} \end{aligned}$$

en que

$$E_{p\times p} = \sum_{j=1}^{g} n_j(\overline{x}_j - \overline{x})(\overline{x}_j - \overline{x})' \qquad \text{y tiene rango } g-1$$

$\overline{y}_j$ es la media de los elementos del vector $\underline{y}_j$

$\overline{y}$ es la media de todos los elementos de $\underline{y}$

Obtención de $\underline{a}$:

El principio que aplica la Función Discriminante Canónica es quelos grupos estén constituidos por elementos parecidos, mientras que haya una notoria diferencia entre los grupos.

Esto se traduce en que la variabilidad entre los grupos debe ser los grande posible, mientras la variabilidad dentro de los grupos debe pequeña.

Entonces nuestro problema se puede reducir a maximizar $\dfrac{SCE}{SCD} = \dfrac{\underline{a}'E\underline{a}}{\underline{a}'D\underline{a}}$ con respecto de $\underline{a}$

El cuociente es invariante para cambios de escala de $\underline{a}$. Entonces el problema es equivalente a maximizar la forma cuadrtica $\underline{a}'E\underline{a}$ sujeto a la condición $\underline{a}'D\underline{a} = 1$

Se resuelve usando multiplicadores de Lagrange. La función a derivar respecto de $\underline{a}$ y λ es

$$\Phi(\underline{a}, \lambda) = \underline{a}'E\underline{a} - \lambda\underline{a}'D\underline{a}$$

y las derivadas, igualadas a cero, son

$$\begin{aligned} E\underline{a} - \lambda D\underline{a} = 0 \qquad & m \text{ ecuaciones} \\ \underline{a}'D\underline{a} - 1 = 0 \qquad & 1 \text{ ecuación} \end{aligned}$$

Si premultiplicamos la primera ecuación por D^{-1} (D es el numerador de una matriz de varianzas-covarianzas, luego es definida positiva, por lo tanto existe su inversa), y si la multiplicamos por $\underline{a}'$ y la combinamos por la segunda ecuación, obtenemos

$$D^{-1}E\underline{a} = \lambda\underline{a}$$
$$\lambda = \underline{a}'E\underline{a}$$

Queda claro que la solución que se obtiene es que $\underline{a}$ es vector propio asociado al mayor valor propio de la matriz simétrica $D^{-1}E$

El rango de $D^{-1}E$ es $g-1$, por lo tanto tiene $g-1$ valores propios no cero. Cada uno da origen a una función discriminante.

Dado que siempre se debe maximizar $\underline{a}'E\underline{a}$ para cada $\underline{a}$ (que da origen a su respectiva función discriminante), y se vió que los $\underline{a}$ son los valores propios de $D^{-1}E$, entonces éstos constituyen medidas de cun bien discrimina cada una de las funciones discriminantes.

Se define el *poder discriminantes* de una función discriminante $y_j = \underline{a}'X$ como la magnitud del valor propio respectivo.

Si la primera función discriminante no es suficiente para discriminar bien (su valor propio es de una magnitud baja en porcentaje respecto de todas), se puede incluir una segunda función discriminante, una tercera, etc.

Clasificación de una nueva observación

Para discriminar una nueva observación $\underline{x}_0$, es decir, decidir a qué población pertenece, se debe seguir el siguiente procedimiento:

Sea $\underline{x}_0$ la nueva observación que se va a clasificar.

Se le aplican las funciones de discriminación a esta observación. Tantas funciones como las que se vayan a usar para clasificar:

$$y_{10} = \underline{a}_1' \underline{x}_0$$
$$y_{20} = \underline{a}_2' \underline{x}_0$$
$$\cdots\cdots$$
etc.

Se calculan los promedios $\overline{\underline{x}}_1, \overline{\underline{x}}_2, ..., \overline{\underline{x}}_g$ de las observaciones pertenecientes a las respectivas poblaciones $\Pi_1, \Pi_2, ...\Pi_g$.

Luego se les aplican las funciones discriminantes que van a ser usadas para clasificar:

$$\underline{a}_1' \overline{\underline{x}}_1, \quad \underline{a}_1' \overline{\underline{x}}_2, \quad ..., \underline{a}_1' \overline{\underline{x}}_g$$
$$\underline{a}_2' \overline{\underline{x}}_1, \quad \underline{a}_2' \overline{\underline{x}}_2, \quad ..., \underline{a}_2' \overline{\underline{x}}_g$$
$$\cdots\cdots$$
etc.

Con estos valores se forman los centroides de las observaciones provenientes de las distintas poblaciones,

$$\overline{\underline{y}}_1 = \begin{bmatrix} \underline{a}_1' \overline{\underline{x}}_1 \\ \underline{a}_2' \overline{\underline{x}}_1 \\ . \\ . \end{bmatrix}, \qquad \overline{\underline{y}}_2 = \begin{bmatrix} \underline{a}_1' \overline{\underline{x}}_2 \\ \underline{a}_2' \overline{\underline{x}}_2 \\ . \\ . \end{bmatrix}, \qquad ... \qquad \overline{\underline{y}}_g = \begin{bmatrix} \underline{a}_1' \overline{\underline{x}}_g \\ \underline{a}_2' \overline{\underline{x}}_g \\ . \\ . \end{bmatrix}$$

Finalmente se deben calcular las distancias del vector correspondiente a la nueva observación. Se utiliza la distancia euclidea, que en el caso de una sóla función discriminante se reduce a la diferencia entre los dos valores, en valor absoluto.

$$\underline{y}_0 = \begin{bmatrix} y_{10} \\ y_{20} \\ . \\ . \end{bmatrix}$$

y cada uno de los centroides

$$\overline{\underline{y}}_1, \overline{\underline{y}}_2, ... \overline{\underline{y}}_g,$$

La distancia ms corta determina en qué población Π_k se clasifica $\underline{x}_0$.

EJEMPLO

Se tienen seis observaciones de dos variables, v1 y v2, pertenecientes a dos poblaciones, que se indican en la primera columna:

$$X = \begin{bmatrix} 1 & 1 & 4 \\ 1 & 3 & 3 \\ 1 & 5 & 2 \\ 2 & 4 & 7 \\ 2 & 7 & 6 \\ 2 & 8 & 3 \end{bmatrix}$$

Se desea construir reglas discriminantes canónicas y luego aplicarlas para clasificar la observación $\underline{x}_0 = \begin{bmatrix} 9 \\ 1 \end{bmatrix}$

Promedio general o centroide:

$$\overline{\underline{x}} = \begin{bmatrix} 4.667 & 4.167 \end{bmatrix}$$

Promedios o centroides por grupo:

$$\overline{\underline{x}}_1' = \begin{bmatrix} 3.000 & 3.000 \end{bmatrix} \qquad \overline{\underline{x}}_2' = \begin{bmatrix} 6.333 & 5.333 \end{bmatrix}$$

Matrices de sumas de cuadrados y productos dentro de cada grupo:

$$D_{y_1} = \begin{bmatrix} 8.000 & -4.000 \\ -4.000 & 2.000 \end{bmatrix} \qquad D_{y_2} = \begin{bmatrix} 8.667 & -7.333 \\ -7.333 & 8.667 \end{bmatrix}$$

Matriz de suma de cuadrados dentro de los grupos:

$$\mathrm{D} = \begin{bmatrix} 16.667 & -11.333 \\ -11.333 & 10.667 \end{bmatrix}$$

La inversa de esta matriz es

$$D^{-1} = \begin{bmatrix} 0.216 & 0.230 \\ 0.230 & 0.338 \end{bmatrix}$$

Matriz de Sumas de cuadrados y productos entre los grupos:

$$E = \begin{bmatrix} 16.6676 & 11.667 \\ 11.667 & 8.167 \end{bmatrix}$$

El producto $D^{-1}E$ es igual a

$$D^{-1}E = \begin{bmatrix} 6.284 & 4.399 \\ 7.770 & 5.439 \end{bmatrix}$$

Se deben obtener los valores de esta matriz. Son

$$\lambda_1 = 11.723 \qquad \text{y} \qquad \lambda_2 = 0{,}000$$

Dado que el segundo valor propio es 0,000, significa que no tiene ningún poder discriminante, y luego la primera función discriminante es suficiente para discriminar totalmente los dos conjuntos de observaciones.

Los correspondientes vectores propios estandarizados de modo que $\underline{a}' D \underline{a}$ sea iguial a 1 son

$$\underline{a}_1 = \begin{bmatrix} 0.899 \\ 1.112 \end{bmatrix} \qquad \text{y} \qquad \underline{a}_2 = \begin{bmatrix} 0.410 \\ \text{-}0.586 \end{bmatrix}$$

La primera función discriminante (y la única que se usar, pues la segunda tiene poder disacriminante cero) es

$$y_1 = \underline{a}_1' \underline{x} = 0{,}0{,}899x_1 + 1{,}112x_2$$

en que $\underline{x}$ es la observación que se desea clasificar.

Supóngase que se desea clasificar la nueva observación $\underline{x}_0 = (9, 1)$

La función discriminante aplicada a la nueva observación $\underline{x}_0$ da lo siguiente

$$y_0 = 0{,}899 \times 9 + 1{,}112 \times 1 = 9{,}2047$$

La función discriminante aplicada a los centroides da

$$y_1 = 0{,}899 \times 3{,}000 + 1{,}112 \times 3{,}000 = 5{,}697 \text{ e } y_2 = 0{,}899 \times 6{,}333 + 1{,}112 \times 5{,}333 = 11{,}624$$

Por último, las distancias de la observación a clasificar a los centroides, son

$$|y_0 - y_1| = |5{,}697 - 9{,}204| = 3{,}506$$
$$|y_0 - y_2| = |11{,}624 - 9{,}204| = 2{,}420$$

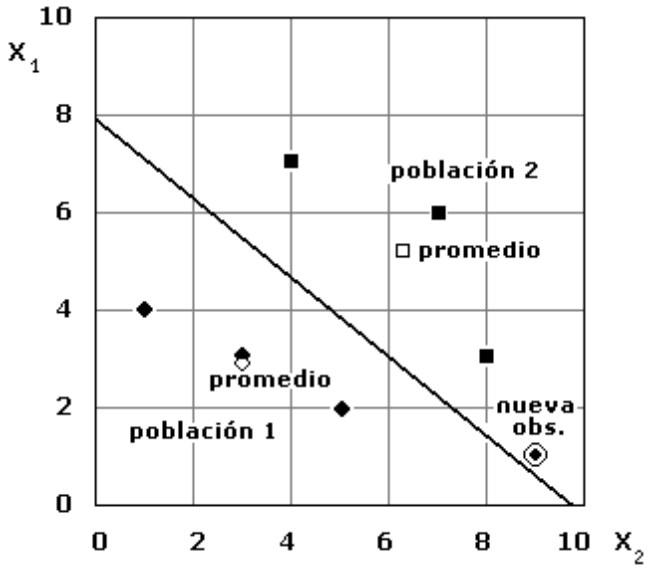

Figura 1: Regiones de discriminacón.

Por lo tanto estamos cerca del centroide del segundo grupo de datos, luego se clasifica la nueva observación $\underline{x}_0 = \begin{bmatrix} 9 \\ 1 \end{bmatrix}$ en la población Π_2.

La Figura 1 ilustra los seis puntos en el plano cartesiano, los dos centroides y la nueva observación a clasificar.

La recta oblicua divide el plano en dos: Las nuevas observaciones que caen en a parte inferior izquierda se clasifican en la población Π_1; las que caen en la parte superior derecha se clasifican en la población Π_2.

CRITERIO BAYESIANO DE DISCRIMINACION

A veces tiene sentido asumir que las poblaciones tienen destribuciones a priori $\pi_1, \pi_2, ..., \pi_g$

En tal caso se define la regla discriminante de Bayes, como aquella que asigna una observación $\underline{x}$ a la población Π_j si

$$\pi_j f_j(\underline{x}) \text{ es máximo} \qquad (j = 1, 2, ..., g).$$

Esta expresión es la verosimilitud a posteriori de Π_j.

Hay algunas formas estndar de asignar probabilidades a priori:

Una es de manera uniforme, que equivale a darles el mismo valor $\pi_j = \frac{1}{g}$ a cada una.

Otra forma es darles valores proporcionales a los tamaños de las submuestras de cada población presentes en la matriz de datos, $\pi_j = \frac{n_j}{n}$. Esto es válido si se asume que la muestra de entrenamiento se formó seleccionando n individuos al azar. De éstas, n_j resultaron de la población $\Pi_j, (j = 1, 2, ..., g)$.

No es válido si inicialmente se decidieron los tamaños n_j de las diferentes poblaciones.

ANALISIS DE ERRORES:

Una estimación de la tasa de estimación errónea provee de una medida cuantitativa del poder de discriminación de una regla discriminante. Algunas técnicas:

Validación Cruzada:

Consiste en estimar la función discriminante dejando fuera una observación, y luego usar la regla para clasificar la observación.

Se repite el procedimieno para todas las observaciones y se calcula la tasa de clasificación errónea.

Para los modelos homosedástico, heterosedástico y esférico, la mayor parte de los cálculos se hace una vez.

Tasas de error estimadas en base a Probabilidades a Posteriori.

Las probabilidades a posteriori, de pertenencia a una población,

$$\tau_j = \frac{\Pi_j f_j(\underline{x})}{\sum_{k=1}^{g} \Pi_k f_k(\underline{x})}$$

Se estiman mediante los promedios a través de las observaciones de cada una de las poblaciones, en la matriz de datos. Con ellas se obtienen probabilidades de clasificación errónea de las observaciones de la matriz de datos.