Tema 1: Análisis Discriminante Lineal

Introducción al Análisis Discriminante

Supongamos que un conjunto de objetos se clasifica en una serie de grupos; el *Análisis Discriminante* equivale a un análisis de regresión donde la variable dependiente es categórica y tiene como categorías la etiqueta de cada uno de los grupos, y.las variables independientes son continuas y determinan a qué grupos pertenecen los objetos. Se trata de encontrar relaciones lineales entre las variables continuas que mejor discriminen en los grupos dados a los objetos. Además, se trata de definir una regla de decisión que asigne un objeto nuevo, que no sabemos clasificar previamente, a uno de los grupos prefijados.

Se presentan una serie de restricciones o supuestos:

(*i*) Se tiene una variable categórica y el resto de variables son de intervalo o de razón y son independientes respecto de ella.

(*ii*) Es necesario que existan al menos dos grupos y para cada grupo se necesitan dos o más casos.

(*iii*) El número de variables discriminantes debe ser menor que el número de objetos menos dos: $x_1, \ldots, x_p$, donde $p < (n-2)$ y n es el número de objetos.

(*iv*) Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes.

(*v*) El número máximo de funciones discriminantes es igual al mínimo entre el número de variables y el número de grupos menos 1 (con q grupos, $(q-1)$ funciones discriminantes).

(*vi*) Las matrices de covarianzas dentro de cada grupo deben ser aproximadamente iguales.

(*vii*) Las variables continuas deben seguir una distribución normal multivariante.

Algunos ejemplos de problemas de Análisis Discriminante son: asignar un cráneo a una de dos posibles especies animales, asignar un texto escrito a uno de entre dos posibles autores, decidir que una declaración de impuestos es potencialmente defraudadora o no, determinar que una empresa está en riesgo de quiebra o no, decidir que un nuevo método de fabricación es eficaz o no.

Existen varios enfoques posibles para este problema. El primero es el análisis discriminante clásico debido a Fisher. Este procedimiento está basado en la normalidad multivariante de las variables y es óptimo bajo dicho supuesto, aunque en la realidad no siempre se cumple éste. Un segundo enfoque más flexible está basado en los modelos con variables dependientes cualitativas, de los cuales el más utilizado es el modelo de regresión logística. Un tercer enfoque es el de los árboles de regresión. Este método conduce a buenos resultados cuando las variables son muy heterogéneas.

Clasificación entre dos poblaciones

Planteamiento del Problema

Sean P_1 y P_2 dos poblaciones donde se define una variable aleatoria vectorial $\mathbf{x}$ k-dimensional. Supondremos que $\mathbf{x}$ es absolutamente continua y que las funciones de densidad de ambas poblaciones, f_1 y f_2, son conocidas. Se estudia el problema de clasificar un nuevo individuo en una de estas poblaciones cuando se observa un vector $\mathbf{x}_0$. El problema puede enfocarse desde el punto de vista de la Inferencia o de la Teoría de Decisión,

incluyendo además probabilidades *a priori* (enfoque bayesiano) o no.

A continuación, se presenta la formulación del problema más general como un problema bayesiano de decisión.

Se consideran las hipótesis siguientes:

(i) Las probabilidades *a priori* de que un individuo tomado al azar provenga de cada población son conocidas: π_1, π_2, tales que $\pi_1 + \pi_2 = 1$.

(ii) Las consecuencias asociadas a los errores de clasificación son $c(2|1)$ y $c(1|2)$, donde $c(i|j)$ es el coste de clasificar en P_i de un objeto que pertenece realmente a P_j. Estos costes se suponen conocidos.

(iii) Las preferencias del decisor por las consecuencias de sus acciones son lineales, es decir, maximizar la función de utilidad equivale a minimizar el coste esperado o coste de oportunidad de la decisión. Por lo tanto, podemos minimizar los costes de oportunidad de la decisión mediante el criterio del valor esperado.

Las posibles decisiones en el problema son únicamente dos: asignar en P_1 ó en P_2.

Una regla de decisión equivale a hacer una partición del espacio muestral E_x (que en general será $\mathbb{R}^K$) en dos regiones: A_1 y $A_2 = E_x - A_1$, tales que:

1. Si $\mathbf{x}_0 \in A_1 \Longrightarrow d_1$ (asignar en P_1).

2. Si $\mathbf{x}_0 \in A_2 \Longrightarrow d_2$ (asignar en P_2).

Un vez observado el valor $\mathbf{x}_0$ podemos calcular la probabilidad *a posteriori* de que el elemento pertenezca a cada población.

Se denomina $P\left(1|\mathbf{x}_0\right)$ la probabilidad a posteriori de que un elemento que ha tomado un valor igual a $\mathbf{x}_0$ pertenezca a P_1. Por el teorema de Bayes esta probabilidad es:

$$P(1|\mathbf{x}_0) = \frac{P(\mathbf{x}_0|1)P(1)}{P(\mathbf{x}_0|1)P(1) + P(\mathbf{x}_0|2)P(2)}.$$

Las probabilidades $P(\mathbf{x}_0|1)$ y $P(\mathbf{x}_0|2)$ son proporcionales a $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$, por lo que la ecuación anterior puede escribirse como

$$P(1|\mathbf{x}_0) = \frac{f_1(\mathbf{x}_0)\pi_1}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2},$$

y para la segunda población

$$P(2|\mathbf{x}_0) = \frac{f_2(\mathbf{x}_0)\pi_2}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2}.$$

Así, si clasificamos al elemento en el grupo 2 las posibles consecuencias son:

(i) Acertar con probabilidad $P(2|\mathbf{x}_0)$, en cuyo caso no hay ningún coste de penalización.

(ii) Equivocarnos con probabilidad $P(1|\mathbf{x}_0)$, en cuyo caso incurrimos en el coste asociado $c(2|1)$ y el coste promedio o valor esperado de la decisión "clasificar $\mathbf{x}_0$ en P_2" será:

$$E(d_2) = c(2|1)P(1|\mathbf{x}_0) + 0P(2|\mathbf{x}_0) = c(2|1)P(1|\mathbf{x}_0).$$

Análogamente, el coste esperado de la decisión: clasificar en el grupo 1:

$$E(d_1) = 0P(1|\mathbf{x}_0) + c(1|2)P(2|\mathbf{x}_0) = c(1|2)P(2|\mathbf{x}_0).$$

Asignaremos el elemento al grupo 2 si su coste esperado es menor, es decir, sustituyendo en las expresiones anteriores, si:

$$\frac{f_2(\mathbf{x}_0)\pi_2}{c\,(2|1)} > \frac{f_1(\mathbf{x}_0)\pi_1}{c\,(1|2)},$$

que indica que, siendo iguales el resto de términos, clasificaremos en la población P_2 si

(i) la probabilidad a priori es más alta;

(ii) la verosimilitud de que provenga de P_2 es más alta;

(iii) el coste de equivocarnos al clasificarlo en P_2 es más bajo.

Simplificaciones

(a) Si suponemos que los costes son iguales, $c(1|2) = c(2|1) = c$, la decisión dependerá únicamente de las probabilidades y asignaremos el elemento al grupo más probable a posteriori.

Así, clasificaremos al elemento $\mathbf{x}_0$ en P_2 si:

$$\pi_2 f_2(\mathbf{x}_0) > \pi_1 f_1(\mathbf{x}_0)$$

(b) Si las probabilidades a priori y los costes son iguales la condición de clasificar en P_2 se reduce a:

$$f_2(\mathbf{x}_0) > f_1(\mathbf{x}_0)$$

y la regla de decisión expresa simplemente que clasificamos al individuo en el grupo que pueda generarlo con mayor probabilidad.

Poblaciones Normales: Función lineal discriminante

Vamos a aplicar el análisis anterior al caso en que f_1 y f_2 sean distribuciones normales con distintos vectores de medias pero idéntica matriz de varianzas. Entonces, su función de densidad es

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}.$$

La manera óptima es clasificar en la población P_2 si:

$$\frac{\pi_2 f_2(\mathbf{x})}{c(2|1)} > \frac{\pi_1 f_1(\mathbf{x})}{c(1|2)}.$$

Como ambos términos son siempre positivos, tomando logaritmos y sustituyendo $f(\mathbf{x})$ por su expresión, la ecuación anterior se convierte en:

$$\log \pi_2 - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) > \log \pi_1 - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - \log\left(\frac{c\,(1|2)}{c\,(2|1)}\right),$$

es decir, operando,

$$(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) > (\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - 2\log\left(\frac{c(1|2)}{c(2|1)}\frac{\pi_2}{\pi_1}\right) \qquad (1)$$

Llamando D_i a la distancia de Mahalanobis entre el punto observado, $\mathbf{x}$, y la población i:

$$D_i = (\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$$

y suponiendo iguales los costes y las probabilidades a priori, $c(1|2) = c(2|1)$; $\pi_1 = \pi_2$, la regla resultante es:

$$\text{Clasificar en 2 si } D_1^2 > D_2^2$$

es decir, clasificar la observación en la población de cuya media esté más próxima, usando la distancia de Mahalanobis. Observemos que si las variables $\mathbf{x}$ tuvieran matriz de covarianzas $\mathbf{V} = \mathbf{I}\sigma^2$ la regla equivaldría a utilizar la distancia euclídea.

Interpretación de la regla de clasificación

La regla general anterior puede escribirse de una forma equivalente que permite interpretar mejor el método de clasificación utilizado.

Si simplificamos 1,

$$\begin{aligned}(\mathbf{x} - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) &= \mathbf{x}'\mathbf{V}^{-1}\mathbf{x} - 2\boldsymbol{\mu}_1'\mathbf{V}^{-1}\mathbf{x} + \boldsymbol{\mu}_1'\mathbf{V}^{-1}\boldsymbol{\mu}_1 \\ (\mathbf{x} - \boldsymbol{\mu}_2)'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) &= \mathbf{x}'\mathbf{V}^{-1}\mathbf{x} - 2\boldsymbol{\mu}_2'\mathbf{V}^{-1}\mathbf{x} + \boldsymbol{\mu}_2'\mathbf{V}^{-1}\boldsymbol{\mu}_2\end{aligned}$$

entonces, la regla divide al conjunto posible de valores de $\mathbf{X}$ en dos regiones cuya frontera es (simplificando términos comunes en ambos miembros), la ecuación:

$$-2\boldsymbol{\mu}_1'\mathbf{V}^{-1}\mathbf{x} + \boldsymbol{\mu}_1'\mathbf{V}^{-1}\boldsymbol{\mu}_1 = -2\boldsymbol{\mu}_2'\mathbf{V}^{-1}\mathbf{x} + \boldsymbol{\mu}_2'\mathbf{V}^{-1}\boldsymbol{\mu}_2 - 2\log\frac{c(1|2)\pi_2}{c(2|1)\pi_1},$$

que, como función de $\mathbf{x}$, equivale a:

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}\mathbf{x} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}\left(\frac{\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1}{2}\right) - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1}.$$

Llamando:

$$\mathbf{w} = \mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \tag{2}$$

entonces, la frontera entre las regiones de clasificación para P_1 y P_2 puede escribirse como:

$$\mathbf{w}'\mathbf{x} = \mathbf{w}'\left(\frac{\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1}{2}\right) - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1} \tag{3}$$

que es la ecuación de un hiperplano. Esta ecuación indica que el procedimiento de clasificación puede resumirse así:

(1) Calcular el vector $\mathbf{w}$ con (2) y a continuación el segundo miembro de (3) que depende sólo de términos conocidos;

(2) Eescribir la función discriminante:

$$g(\mathbf{x}) = \mathbf{w}'\mathbf{x} = w_1x_1 + \ldots + w_kx_k$$

Esta función es una combinación lineal de los valores de la variable con los pesos dados por el vector $\mathbf{w}$.

(3) Introducir en esta función los valores observados para el nuevo individuo a clasificar, $\mathbf{x}_0 = (x_{10}, \ldots, x_{k0})$. Según la ecuación (1) clasificaremos en la población 2 cuando el primer miembro sea mayor que el segundo.

En el caso particular de que $c(1|2)\pi_2 = c(2|1)\pi_1$ la regla de decisión se reduce entonces a clasificar en P_2 si:

$$\mathbf{w}'\mathbf{x} > \mathbf{w}'\left(\frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right). \tag{4}$$

Se puede comprobar que esta regla equivale a proyectar el punto $\mathbf{x}$ que queremos clasificar y las medias de ambas poblaciones sobre una recta, y después asignar el punto a aquella población de cuya media se encuentre más próxima en la proyección.

En resumen, el problema de clasificación cuando los costes y las probabilidades a priori se suponen idénticos y las variables normales, se reduce a definir una variable escalar, $z = \mathbf{w}'\mathbf{x}$, trasladar las medias y el punto observado a dicha escala, y asignarlo a la media más próxima. La distancia entre las medias proyectadas es igual a su distancia de Mahalanobis en el espacio. La varianza de la nueva variable escalar es igual a la distancia de Mahalanobis entre las medias

$$Var(z) = Var(\mathbf{w}'\mathbf{x}) = \mathbf{w}'Var(\mathbf{x})\mathbf{w} = \mathbf{w}'\mathbf{V}\mathbf{w} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = D^2$$

ya que $\mathbf{w} = \mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$.

Generalización para varias poblaciones normales

Planteamiento General

La generalización de estas ideas para varias poblaciones es simple: el objetivo es ahora dividir el espacio E_x en G regiones $A_1, \ldots A_g, \ldots, A_G$ tales que si $\mathbf{x}$ pertenece a A_i el punto se clasifica en la población P_i. Supondremos que los costes de clasificación son constantes y que no dependen de la población en que se haya clasificado. Entonces, la región A_g vendrá definida por aquellos puntos con máxima probabilidad de ser generados por P_g, es decir donde el producto de la *a priori* y la verosimilitud sean máximas:

$$A_g = \{\mathbf{x} \in E_x \text{ tal que } \pi_g f_g(\mathbf{x}) > \pi_i f_i(\mathbf{x}),\ \forall i \neq g\}$$

Si las probabilidades a priori son iguales $\forall i$, $\pi_i = 1/G$ y las distribuciones $f_i(\mathbf{x})$ son normales con la misma matriz de varianzas, la condición anterior equivale a calcular la distancia de Mahalanobis del punto observado al centro de cada población y clasificarlo en la población que haga esta distancia mínima.

Al minimizar las distancias de Mahalanobis $\left(\mathbf{x} - \boldsymbol{\mu}_g\right)' \mathbf{V}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_g\right)$ se llega a la ecuación

$$\begin{aligned} \mathbf{w}'_{ij}\mathbf{x} &= \mathbf{w}'_{ij}\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) \Longrightarrow \\ \mathbf{w}'_{ij}\left(\mathbf{x} - \boldsymbol{\mu}_i\right) &= \mathbf{w}'_{ij}(\boldsymbol{\mu}_j - \mathbf{x}). \end{aligned}$$

Esta ecuación admite la misma interpretación en el sentido de las proyecciones que en el caso de dos poblaciones. Se construye una dirección $\mathbf{w}_{ij}$ y se proyectan las medias y el punto $\mathbf{x}$ que tratamos de clasificar sobre esta dirección y se asigna el punto a la población de cuya media proyectada esté más próxima.

Para ilustrar el procedimiento operativo, supongamos cinco poblaciones con $k > 4$ variables medidas sobre ellas Tenemos dos formas de realizar el análisis. La primera es calcular para las G poblaciones las distancias de Mahalanobis y clasificar la observación en la más próxima.

La segunda es hacer el análisis comparando las poblaciones dos a dos. Supongamos que hemos obtenido de las comparaciones dos a dos los siguientes resultados ($i \succ j$ indica

que la población i es preferida a la j, es decir, el punto se encuentra más próximo a la media de la población i que a la de j):

$$1 \succ 2$$

$$2 \succ 3$$

$$4 \succ 3$$

$$5 \succ 4$$

Las poblaciones 2, 3 y 4 quedan descartadas (ya que $1 \succ 2 \succ 3$ y $5 \succ 4$). La duda se refiere a las poblaciones 1 y 5. Construyendo (a partir de las reglas anteriores) la regla para discriminar entre estas dos últimas poblaciones, supongamos que

$$5 \succ 1$$

entonces clasificaremos en la población 5.

Cuando $k < G - 1$ el máximo número de proyecciones linealmente independientes que podemos construir es k, y éste será el máximo número de direcciones a definir. Por ejemplo, supongamos que $k = 2$ y $G = 5$. Podemos definir una dirección de proyección cualquiera, por ejemplo,

$$\mathbf{w}_{12} = \mathbf{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

y proyectar todas las medias $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_5)$ y el punto $\mathbf{x}$ sobre dicha dirección. Entonces clasificaremos el punto en la población cuya media proyectada está más próxima. Ahora bien, es posible que sobre esta dirección coincidan las medias proyectadas de varias poblaciones. Si esto ocurre con, por ejemplo, las $\boldsymbol{\mu}_4$ y $\boldsymbol{\mu}_5$, resolveremos el problema proyectando sobre otra segunda dirección

$$\mathbf{w}_{45} = \mathbf{V}^{-1}(\boldsymbol{\mu}_4 - \boldsymbol{\mu}_5).$$

Poblaciones desconocidas. caso general

Regla estimada de clasificación

Se trata ahora el caso de aplicar la teoría anterior cuando en lugar de trabajar con poblaciones se consideran muestras. Abordaremos directamente el caso de G poblaciones posibles y, como caso particular, la discriminación clásica para $G = 2$.

La matriz general de datos $\mathbf{X}$ de dimensiones $k \times n$ (k variables y n individuos), se puede particionar ahora en G matrices correspondientes a las subpoblaciones. Se denomina x_{ijg} a los elementos de estas submatrices, donde i representa el individuo, j la variable y g el grupo o submatriz. Llamaremos n_g al número de elementos en el grupo g y el número total de observaciones se denomina:

$$n = \sum_{g=1}^{G} n_g$$

Vamos a llamar $\mathbf{x}'_{ig}$ al vector fila $(1 \times k)$ que contiene los k valores de las variables para el individuo i en el grupo g, es decir,

$$\mathbf{x}'_{ig} = (x_{i1g}, \ldots, x_{ikg})$$

El vector de medias dentro de cada clase o subpoblación es:

$$\overline{\mathbf{x}}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{x}_{ig}$$

que es un vector columna de dimensión k que contiene las k medias para las observaciones de la clase g. La matriz de varianzas y covarianzas para los elementos de la clase g será:

$$\widehat{\mathbf{S}}_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'$$

donde hemos dividido por $n_g - 1$ para tener estimaciones centradas de las varianzas y las covarianzas. Si suponemos que las G subpoblaciones tienen la misma matriz de varianzas y covarianzas, su mejor estimación centrada con todos los datos será una combinación lineal de las estimaciones centradas de cada población con un peso proporcional a su

precisión. Por tanto,

$$\widehat{\mathbf{S}}_w = \sum_{g=1}^{G} \frac{n_g - 1}{n - G} \widehat{\mathbf{S}}_g$$

Llamaremos $\mathbf{W}$ a la matriz de sumas de cuadrados *dentro* de las clases que viene dada por

$$\mathbf{W} = (n - G)\widehat{\mathbf{S}}_w$$

Para obtener las funciones discriminantes utilizaremos $\overline{\mathbf{x}}_g$ como estimación de μ_g, y $\widehat{\mathbf{S}}_w$ como estimación de $\mathbf{V}$. En concreto, suponiendo iguales las probabilidades a priori y los costes de clasificación, clasificaremos al elemento en el grupo que lleve a un valor mínimo de la distancia de Mahalanobis entre el punto $\mathbf{x}$ y la media del grupo. De todos modos, conviene antes de construir una regla de clasificación realizar un test para comprobar que los grupos son realmente distintos, es decir, que no todas las medias $\boldsymbol{\mu}_g$ son iguales.

El cálculo de probabilidades de error puede hacerse sin aplicar la hipótesis de normalidad, aplicando la función discriminante a las n observaciones para clasificarlas. En el caso de dos grupos, obtendríamos la tabla:

	Clasificado	
Realidad	P_1	P_2
P_1	n_{11}	n_{12}
P_2	n_{21}	n_{22}

donde n_{ij} es el número de datos que viniendo de la población i se clasifica en j.

El error aparente de la regla es:

$$\text{Error} = \frac{n_{12} + n_{21}}{n_{11} + n_{22}} = \frac{\text{Total de mal clasificados}}{\text{Total de bien clasificados}}.$$

Un procedimiento mejor es clasificar cada elemento con una regla en la que dicho elemento no ha intervenido: método de *validación cruzada*. Para ello, podemos construir n funciones discriminantes con las n muestras de tamaño $n-1$ que resultan al eliminar uno a uno cada elemento de la población y clasificar después cada dato con la regla construida sin él. Este método conduce a una mejor estimación del error de clasificación.

Variables canónicas discriminantes

El enfoque anterior puede generalizarse para encontrar variables canónicas que tengan el máximo poder discriminante para clasificar nuevos elementos respecto a las poblaciones. El objetivo es, en lugar de trabajar con las variables originales $\mathbf{x}$, definir r variables canónicas, z_i $(i = 1, \ldots, r)$ donde $r = min(G-1, k)$, que sean combinación lineal de las originales: $z_i = \mathbf{a}_i'\mathbf{x}$ de modo que

1. Las medias de las poblaciones $\boldsymbol{\mu}_g$ se expresan en términos de las variables canónicas donde $\overline{\mathbf{z}}_1, \ldots, \overline{\mathbf{z}}_g$ son vectores $r \times 1$ cuyas coordenadas son las proyecciones de las medias sobre las r variables canónicas;

2. Se hace lo mismo para el punto $\mathbf{x}_0$ a clasificar donde se denota a $\mathbf{z}_0$ como dicho vector;

3. Clasificamos el punto en aquella población de cuya media se encuentre más próxima, con la distancia euclídea, en el espacio de las variables canónicas z; es decir, lo clasificaremos en la población i si

$$(\mathbf{z}_0 - \mathbf{z}_i)'(\mathbf{z}_0 - \mathbf{z}_i) = \min_g (\mathbf{z}_0 - \mathbf{z}_g)'(\mathbf{z}_0 - \mathbf{z}_g)$$

A continuación se trata el problema de la obtención de las variables canónicas $\mathbf{z}$.

Búsqueda del espacio de máximo poder discriminante

Comenzamos buscando un vector $\mathbf{u}'$ tal que, cuando proyectamos los puntos sobre él, se obtiene la máxima variabilidad entre los grupos en relación a la variabilidad dentro de los grupos. La media de las observaciones del grupo g en esta nueva variable será:

$$\overline{z}_g = \mathbf{u}'\overline{\mathbf{x}}_g$$

donde $\overline{\mathbf{x}}_g$ es el vector $k \times 1$ que contiene las medias de las k variables en dicho grupo. La media para todos los datos será:

$$z_T = \mathbf{u}'\overline{\mathbf{x}}_T$$

donde $\overline{\mathbf{x}}_T$ es el vector $k \times 1$ que contiene las medias de las k variables uniendo todos los grupos.

Se desea encontrar el vector $\mathbf{u}$ de manera que la separación entre las medias de los grupos sea máxima. Una medida de la distancia entre las medias $\overline{z}_1, \ldots, \overline{z}_g$ es la suma de cuadrados entre las medias dada por $\sum_{g=1}^{G} n_g(\overline{z}_g - \overline{z}_T)^2$. Para juzgar si éste término es grande o pequeño, debemos compararlo con la variabilidad intrínseca de los datos dada por $\sum_i \sum_g (z_{ig} - \overline{z}_g)^2$.

En definitiva, el criterio para encontrar la mejor dirección de proyección consiste en maximizar la separación relativa entre las medias, dada por:

$$\phi = \frac{\sum n_g(\overline{z}_g - \overline{z}_T)^2}{\sum\sum(z_{ig} - \overline{z}_g)^2}.$$

Este criterio se puede formular en función de los datos originales.

La suma de cuadrados *dentro* de grupos o variabilidad no explicada para los puntos proyectados, es:

$$V_{NE} = \sum_{j=1}^{n_g}\sum_{g=1}^{G}(z_{jg} - \overline{z}_g)^2 = \sum_{j=1}^{n_g}\sum_{g=1}^{G}\mathbf{u}'(\mathbf{x}_{jg} - \overline{\mathbf{x}}_g)(\mathbf{x}_{jg} - \overline{\mathbf{x}}_g)'\mathbf{u} = \mathbf{u}'\mathbf{W}\mathbf{u}$$

donde $W = (n - G)\widehat{\mathbf{S}}_w$. Esta matriz tiene dimensiones $k \times k$ y tendrá, en general, rango k, suponiendo que $n - G \geq k$ y estima la variabilidad de los datos respecto a las medias de grupo que es la misma, por hipótesis, en todos ellos.

La suma de cuadrados *entre* grupos, o variabilidad explicada, para los puntos proyectados es

$$\begin{aligned} V_E &= \sum_{g=1}^{G} n_g(\overline{z}_g - \overline{z}_T)^2 = \\ &= \sum n_g \mathbf{u}'(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)'\mathbf{u} = \\ &= \mathbf{u}'\mathbf{B}\mathbf{u} \end{aligned}$$

siendo $\mathbf{B}$ la matriz de suma de cuadrados entre grupos, que puede escribirse como

$$\mathbf{B} = \sum_{g=1}^{G} n_g \mathbf{a}_g \mathbf{a}_g'$$

siendo $\mathbf{a}_g = \overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T$.

La matriz $\mathbf{B}$ es cuadrada $k \times k$, simétrica y se obtiene como suma de G matrices de rango uno formadas por los vectores $\mathbf{a}_g$ que no son independientes, ya que están ligados por la relación:

$$\sum_{g=1}^{G} n_g \mathbf{a}_g = \mathbf{0},$$

que implica que el rango de $\mathbf{B}$ será $G - 1$.

En resumen, la matriz $\mathbf{W}$ mide las diferencias dentro de grupos y la $\mathbf{B}$ las diferencias entre grupos. La función a maximizar puede escribirse también como

$$\phi = \frac{\mathbf{u}_1' \mathbf{B} \mathbf{u}_1}{\mathbf{u}_1' \mathbf{W} \mathbf{u}_1}.$$

Derivando e igualando a cero de la forma habitual:

$$\frac{d\phi}{du_1} = 0 = \frac{2\mathbf{B}\mathbf{u}_1(\mathbf{u}_1'\mathbf{W}\mathbf{u}_1) - 2(\mathbf{u}_1'\mathbf{B}\mathbf{u}_1)\mathbf{W}\mathbf{u}_1}{(\mathbf{u}_1'\mathbf{W}\mathbf{u}_1)^2} = 0,$$

entonces,

$$\mathbf{B}\mathbf{u}_1 = \mathbf{W}\mathbf{u}_1 \left(\frac{\mathbf{u}_1'\mathbf{B}\mathbf{u}_1}{\mathbf{u}_1'\mathbf{W}\mathbf{u}_1} \right)$$

es decir, como

$$\mathbf{B}\mathbf{u}_1 = \phi \mathbf{W}\mathbf{u}_1$$

y suponiendo que $\mathbf{W}$ es una matriz no singular,

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{u}_1 = \phi \mathbf{u}_1$$

que implica que $\mathbf{u}_1$ debe de ser un vector propio de $\mathbf{W}^{-1}\mathbf{B}$ y ϕ es su valor propio asociado. Como queremos maximizar ϕ, $\mathbf{u}$ será el vector propio asociado al mayor valor propio.

Podemos plantearnos obtener un segundo eje tal que maximice ϕ con la condición de que la nueva variable canónica $z_2 = \mathbf{u}_2'\mathbf{x}$ esté incorrelada con la primera, $z_1 = \mathbf{u}_1'\mathbf{x}$. Se puede demostrar que esto ocurre si tomamos las raíces características y vectores de $\mathbf{W}^{-1}\mathbf{B}$ (puede también obtenerse derivando la función lagrangiana).

Los vectores propios de la matriz $\mathbf{W}^{-1}\mathbf{B}$ no serán, en general, ortogonales ya que aunque las matrices $\mathbf{W}^{-1}$ y $\mathbf{B}$ son simétricas, su producto no necesariamente lo es.

Además, el rango de esta matriz, $\mathbf{W}^{-1}\mathbf{B}$, será $r = \text{mín}(k, G-1)$ (ya que el rango del producto de dos matrices es menor o igual que el de las originales) y éste es el máximo número de factores discriminantes que podemos obtener.

Variables canónicas

Este procedimiento proporciona $r = \text{mín}(k, G-1)$ variables canónicas que vienen dadas por

$$\mathbf{z} = \mathbf{U}'\mathbf{x} \tag{5}$$

donde $\mathbf{U}'$ es una matriz $r \times k$ y $\mathbf{x}$ un vector $k \times 1$. El vector $\mathbf{z}$ de dimensión $r \times 1$, recoge los valores de las variables canónicas para el elemento $\mathbf{x}$.

Las variables canónicas así obtenidas resuelven el problema de clasificación. En efecto, para clasificar un nuevo individuo $\mathbf{x}_0$ basta calcular sus coordenadas $\mathbf{z}_0$ con la expresión (5) y asignarle al grupo de cuya media transformada esté más próxima con la distancia euclídea.

Se demuestra que, si estandarizamos adecuadamente estas variables, el cálculo de las distancias euclídeas en el espacio de dimensión r de las variables canónicas es equivalente a calcular las distancias de Mahalanobis en el espacio de dimensión k de las variables originales.

Discriminación cuadrática. Discriminación de poblaciones no normales

Cuando se supone normalidad y la hipótesis de igualdad entre varianzas no es admisible, el procedimiento de clasificación se reduce a calcular el grupo donde se alcanza la máxima probabilidad a posteriori

$$\min_{j\in\{1,\ldots G\}} \left[\frac{1}{2}\log|\mathbf{V}_j| + \frac{1}{2}(\mathbf{x}_0 - \boldsymbol{\mu}_j)'\mathbf{V}_j^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_j) - \ln(c_j\pi_j)\right].$$

Como habitualmente desconocemos $\mathbf{V}_j$ y $\boldsymbol{\mu}_j$, éstas se estiman mediante $\mathbf{S}_j$ y $\overline{\mathbf{x}_j}$.

Se puede observar que al no anularse el término $\mathbf{x}_0'\mathbf{V}_j^{-1}\mathbf{x}_0$ ahora las funciones discriminantes no son lineales y tendrán un término de segundo grado. En todo caso, el procedimiento operativo es análogo.

Aparece también un problema de discriminación cuadrática en el análisis de determinadas poblaciones no normales. En el caso general de poblaciones arbitrarias tenemos dos alternativas:

(*i*) Aplicar la teoría general y obtener la función discriminante que puede ser complicada.

(*ii*) Aplicar la teoría de poblaciones normales, tomar como medida de distancia la distancia de Mahalanobis y clasificar $\mathbf{x}$ en la población P_j para la cual

$$\mathbf{D}^2 = (\mathbf{x} - \overline{\mathbf{x}}_j)'\widehat{\mathbf{V}}_j^{-1}(\mathbf{x} - \overline{\mathbf{x}}_j)$$

sea mínima.

Para poblaciones discretas estas aproximaciones no son buenas. Se han propuesto métodos alternativos basados en la distribución multinomial o en la distancia χ^2 con relativa eficacia.

El enfoque bayesiano permite una solución general del problema sean o no iguales las matrices de covarianzas. Utilizando el teorema de Bayes para poblaciones normales con distinta varianza, y poniendo una distribución a priori no informativa, puede obtenerse que la razón de las probabilidades a posteriori entre la población i y la j es

$$\frac{P(i|\mathbf{x})}{P(j|\mathbf{x})} = c_{ij}\frac{\pi_i}{\pi_j}\cdot\frac{|\mathbf{S}_i|^{1/2}}{|\mathbf{S}_j|^{1/2}}\cdot\frac{\left(1 + a_j(\mathbf{x}_0 - \overline{\mathbf{x}}_j)'\mathbf{S}_j^{-1}(\mathbf{x}_0 - \overline{\mathbf{x}}_j)\right)^{n_j/2}}{\left(1 + a_i(\mathbf{x}_0 - \overline{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}_0 - \overline{\mathbf{x}}_i)\right)^{n_i/2}},$$

donde π_i son las probabilidades a priori, $\mathbf{S}_j$ las matrices de covarianzas estimadas,

$$a_i = n_i/(n_i^2 - 1),$$

y

$$c_{ij} = \left[\frac{(n_j^2 - 1)}{(n_i^2 - 1)}\,\frac{n_i}{n_j}\right]^{k/2}\frac{\Gamma\left(\frac{n_i}{2}\right)\Gamma\left(\frac{n_j - k}{2}\right)}{\Gamma\left(\frac{n_j}{2}\right)\Gamma\left(\frac{n_i - k}{2}\right)}.$$

Si los tamaños muestrales son aproximadamente iguales, $n_i \simeq n_j$, entonces $c_{ij} \simeq 1$.

Procedimientos Stepwise

Una vez obtenida una función discriminante es interesante evaluar el papel de cada una de las variables en dicha función. La manera habitual de hacerlo es comparando los posibles modelos que se pueden plantear con los distintos subconjuntos de variables que se pueden tomar.

Si se elimina un grupo de variables se puede calcular las correspondientes sumas de cuadrados (varianza residual) que mide el impacto de las variables eliminadas. Denominamos a la varianza residual asociada con la puntuación discriminante correspondiente a k variables como SS_k, y a la asociada con la puntuación discriminante con q variables eliminadas como SS_{k-q}, el estadístico

$$F = \frac{\left(SS_{k-q} - SS_k\right)/q}{SS_k/(n-k-1)}$$

tiene una distribución F de Snedecor con q y $n-(k+1)$ grados de libertad cuando las q variables eliminadas no tienen influencia en la discriminación. Esto es equivalente al método usado en regresión múltiple.

En `R` se puede usar considerando la librería `klaR`.

Ejemplos

En los siguientes ejemplos se muestra, además de los problemas, el código en `R` correspondiente.

Ejemplo 1

Se consideran los datos recogidos sobre 32 cráneos en el Tibet.

	Longitud	Anchura	Altura	Altura.Cara	.Anchura.Cara	Tipo
1	190.50	152.50	145.00	73.50	136.50	1
2	172.50	132.00	125.50	63.00	121.00	1
3	167.00	130.00	125.50	69.50	119.50	1
4	169.50	150.50	133.50	64.50	128.00	1
5	175.00	138.50	126.00	77.50	135.50	1
6	177.50	142.50	142.50	71.50	131.00	1
7	179.50	142.50	127.50	70.50	134.50	1
8	179.50	138.00	133.50	73.50	132.50	1
9	173.50	135.50	130.50	70.00	133.50	1
10	162.50	139.00	131.00	62.00	126.00	1
11	178.50	135.00	136.00	71.00	124.00	1
12	171.50	148.50	132.50	65.00	146.50	1
13	180.50	139.00	132.00	74.50	134.50	1
14	183.00	149.00	121.50	76.50	142.00	1
15	169.50	130.00	131.00	68.00	119.00	1
16	172.00	140.00	136.00	70.50	133.50	1
17	170.00	126.50	134.50	66.00	118.50	1
18	182.50	136.00	138.50	76.00	134.00	2
19	179.50	135.00	128.50	74.00	132.00	2
20	191.00	140.50	140.50	72.50	131.50	2
21	184.50	141.50	134.50	76.50	141.50	2
22	181.00	142.00	132.50	79.00	136.50	2
23	173.50	136.50	126.00	71.50	136.50	2
24	188.50	130.00	143.00	79.50	136.00	2
25	175.00	153.00	130.00	76.50	142.00	2
26	196.00	142.50	123.50	76.00	134.00	2
27	200.00	139.50	143.50	82.50	146.00	2
28	185.00	134.50	140.00	81.50	137.00	2
29	174.50	143.50	132.50	74.00	136.50	2
30	195.50	144.00	138.50	78.50	144.00	2
31	197.00	131.50	135.00	80.50	139.00	2
32	182.50	131.00	135.00	68.50	136.00	2

Los datos corresponden a dos tipos raciales diferentes en los que se practicaron diferentes medidas antropométricas de longitudes, anchuras de cráneo y de cara. El programa R tiene una función llamada `lda` (en la librería `MASS`) para calcular un análisis lineal discriminante.

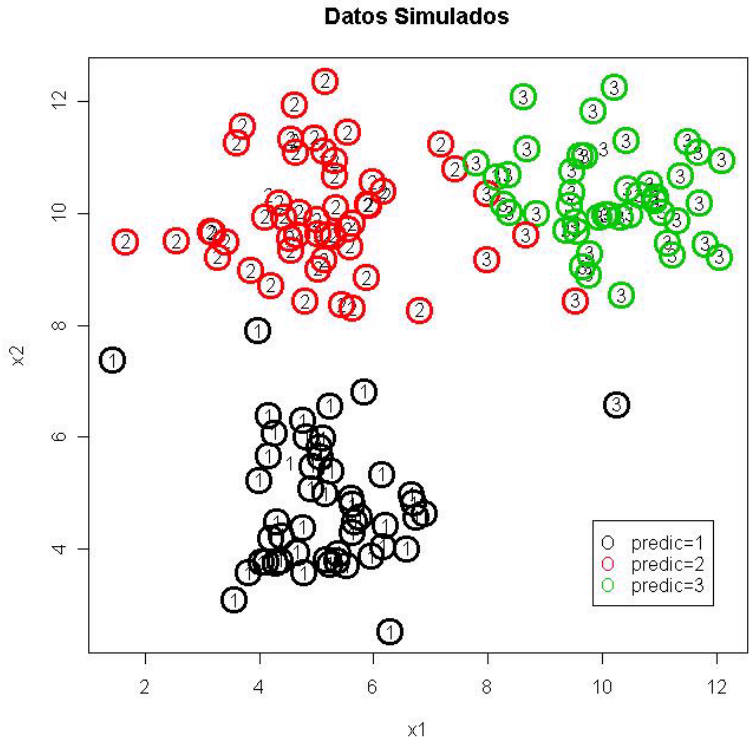
```
# Leo los datos
Tibet <- read.table("c:/directorio/Craneos.txt")
dimnames(Tibet)[[2]] <- c("Longitud", "Anchura", "Altura",
"Altura.Cara", "Anchura.Cara", "Tipo")
attach(Tibet)

m1 <- apply(Tibet[Tipo==1,-6],2,mean)
m2 <- apply(Tibet[Tipo==2,-6],2,mean)
l1 <- length(Tipo[Tipo==1])
l2 <- length(Tipo[Tipo==2])
x1 <- Tibet[Tipo==1,-6]
x2 <- Tibet[Tipo==2,-6]
S123 <- ((l1-1)*var(x1)+(l2-1)*var(x2))/(l1+l2-2)
T2 <- t(m1-m2)%*%solve(S123)%*%(m1-m2)
Fstat <- (l1+l2-5-1)*T2/(l1+l2-2)*5
pvalue <- 1-pf(Fstat,5,26)
Fstat
pvalue

m1 <- apply(Tibet[Tipo==1,-6],2,mean)
m2 <- apply(Tibet[Tipo==2,-6],2,mean)
l1 <- length(Tipo[Tipo==1])
l2 <- length(Tipo[Tipo==2])
x1 <- Tibet[Tipo==1,-6]
x2 <- Tibet[Tipo==2,-6]
S123 <- ((l1-1)*var(x1)+(l2-1)*var(x2))/(l1+l2-2)
```

```
a <- solve(S123)%*%(m1-m2)
z12 <- (m1%*%a+m2%*%a)/2
z1 <- m1%*%a
z2 <- m2%*%a
z1
z2
a
z12

# Se carga la libreria MASS
library(MASS)

# Se hace un analisis discriminante lineal
dis <- lda(Tipo ~ Longitud + Anchura + Altura + Altura.Cara
+ Anchura.Cara, data=Tibet, prior=c(0.5,0.5))

# Se consideran las medidas de dos nuevos craneos
nuevosdatos <- rbind(c(171,140.5,127.0,69.5,137.0),
c(179.0,132.0,140.0,72.0,138.5))

# Asigno a los dos nuevos datos los nombres de las variables
colnames(nuevosdatos) <- colnames(Tibet[,-6])

nuevosdatos <- data.frame(nuevosdatos)

# Se predice el grupo de pertenencia de los nuevos datos
predict(dis,newdata=nuevosdatos)$class
```

```
# Se predicen los datos originales en los grupos segun
# la function discriminante
grupo <- predict(dis,method="plug-in")$class

# Se observa el numero de datos originales bien y mal clasificados
table(grupo,Tipo)

# Se usa un procedimiento Stepwise
library(klaR)
sc <- greedy.wilks(Tipo ~ .,data=Tibet, prior=c(0.5,0.5), "lda",
niveau = 0.05)
sc

# O bien
sc <- stepclass(Tipo ~ .,data=Tibet, prior=c(0.5,0.5), "lda")
sc
```

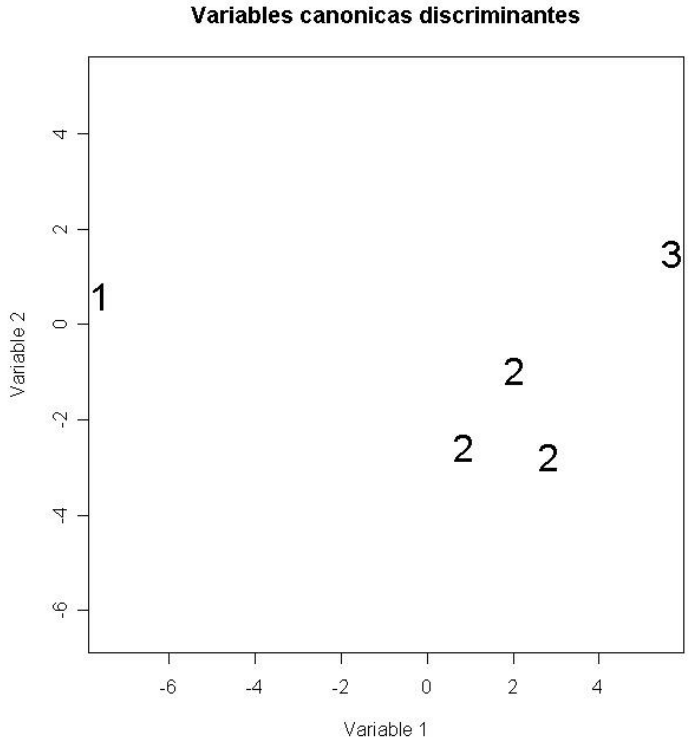
Ejemplo 2

En este ejemplo se ha generado una base de datos con 3 grupos donde las medias de dos variables x_1, x_2 son distintas. La siguiente figura ilustra la posición de las observaciones en el espacio original.



Se puede ver que los indicadores de grupos en las observaciones se disponen en tres grupos distintos y que ninguna de las dos variables observadas x_1, x_2 tiene mayor poder discriminante. Se entrena la función discriminante con 5 observaciones cogidas al azar entre las 150.

La proyección de la muestra de entrenamiento en las variables canónicas discriminantes se puede ver en la siguiente figura.



Se ve que la primera variable discrimina más entre los grupos. En particular entre el grupo 1 y 3 que son los más alejados en el espacio original. La segunda variable tiene menor poder discriminante (7 % frente a 93 % de la variable 1) y el grupo 2 es el más cercano al 3 en el nuevo espacio. El resultado de la clasificación de las restantes 145 observaciones (probabilidad *a priori* igual a 1/3 por cada grupo) se puede ver en la primera gráfica. Algunas observaciones entre el grupo 2 y 3 han sido confundidas por el clasificador.

```
rm(list=ls(all=TRUE))

# Inicializo el generador de numeros aleatorios
# para obtener los mismos datos y la misma muestra de entrenamiento
set.seed(17)

# Matriz de datos
```

```
x1 <- c(rnorm(100,mean=5),rnorm(50,mean=10))
x2 <- c(rnorm(50,mean=5),rnorm(50,mean=10),rnorm(50,mean=10))

# Genero un factor con 3 niveles y 50 replicas de cada uno
Ig <- gl(3,50)

# Dibujo los datos
plot(x1,x2,pch=as.character(Ig),main="Datos Simulados")

misdatos <- data.frame(x1,x2,Ig)

# Elijo la muestra de entrenamiento
train <- sample(1:150,5)
table(misdatos$Ig[train])

library(MASS)
z <- lda(Ig ~ ., misdatos, prior = c(1,1,1)/3, subset = train)

predigo <- predict(z, misdatos[-train, ])$class

# Dibujo las predicciones sobre los datos
plot(x1,x2,pch=as.character(Ig),main="Datos Simulados")
points(x=misdatos[-train,1],y=misdatos[-
train,2],pch="O",cex=2,col=as.numeric(predigo))

# Pongo la etiqueta en el grafico donde se quiera
a <- locator(1)
```

```
legend(a,c("predic=1","predic=2","predic=3"),col=1:3,pch="O")

# Dibujo de las variables canonicas discriminantes
plot(z, xlab="Variable 1",ylab="Variable 2",cex=2,
main="Variables canonicas discriminantes")
```