Capítulo 8

Análisis Discriminante

- Técnica de clasificación donde el objetivo es obtener una función capaz de clasificar a un nuevo individuo a partir del conocimiento de los valores de ciertas variables discriminadoras.
- A diferencia del A. Cluster, se deben conocer los grupos previamente y a qué grupo pertenecen ciertos individuos, de los que también se conoce sus valores en las variables discriminantes.

EJEMPLO

Se dispone de una muestra de pacientes a los que se les mide previamente un conjunto de variables. El investigador puede dividir la muestra en dos (o más) grupos de diagnóstico. Más tarde se mide a un nuevo enfermo el mismo grupo de variables y, por los valores obtenidos, el Análisis Discriminante permite asignar dicho paciente al grupo de máxima probabilidad, cuantificando a la vez el valor de ella.

INTERÉS

Extensión a los campos de las Ciencias de la Vida en la que la clasificación de individuos a través de un perfil observado constituye un frecuente problema de investigación.

DEFINICIÓN

Técnica de Análisis Multivariante que permite asignar o clasificar nuevos individuos dentro de grupos previamente reconocidos o definidos.

PLANTEAMIENTO DEL PROBLEMA

- **Punto de partida:** Tabla de datos de N individuos en que se han medido p variables (caso habitual), una variable cualitativa adicional (dependiente o clasificativa) con dos o más categorías que define, por otros medios, el grupo a que cada individuo pertenece.

- **Dimensión de la matriz:** $N \times (p+1)$.

- Cada caso figura con un perfil y una asignación de grupo.

- **Ojetivo:** Obtener un modelo matemático discriminante contra el cual sea contrastado el perfil de un nuevo individuo cuyo grupo se desconoce para, en función de un resultado numérico, ser asignado al grupo más probable.

- **Nota:** Cuanto mejor sea el grupo de partida más fiable será el resultado de asignaciones posteriores.

- **Doble Finalidad**:

  - ♠ Por una parte explicar la pertenencia de cada caso del fichero de datos original a uno u otro grupo en función de las variables de su perfil para comprobar su pertenencia o no al grupo preestablecido y cuantificar el peso de cada una de ellas en la discriminación.

  - ♠ Por otra parte predecir a qué grupo más probable habrá de pertenecer un nuevo individuo del que únicamente se conoce su perfil de variables.

En el primer caso se explica y
en el segundo se predice la variable categórica o grupo

EJEMPLO

- Dos grupos definidos.
- 3 variables explicativas: $V1, V2$ y $V3$
- 10 individuos de los que se conoce su perfil y su asignación

$$\begin{pmatrix} V1 & V2 & V3 & \text{Grupo} \\ 15 & 41 & 32 & 1 \\ 17 & 40 & 56 & 1 \\ 32 & 35 & 46 & 2 \\ 16 & 42 & 50 & 1 \\ 30 & 33 & 45 & 2 \\ 32 & 32 & 33 & 2 \\ 33 & 30 & 37 & 2 \\ 21 & 39 & 35 & 1 \\ 20 & 38 & 44 & 1 \\ 30 & 31 & 45 & 2 \end{pmatrix}$$

- El valor que toma el primer individuo en la primera variable es 15, el valor que toma este individuo en la segunda variable es 41 y el valor que toma en la tercera es 32. El grupo al que pertenece este individuo es el 1.
- Los individuos 1, 2, 4, 8 y 9 son del grupo 1 y los individuos 3, 5, 6, 7 y 10 son del grupo 2.
- La pertenencia a cada grupo no viene determinada por el valor de las variables. (Nuestro objetivo es relacionar estas dos informaciones).
- Se conoce la pertenencia a cada grupo por otro medios.
- La variable que asigna los individuos a los grupos es cualitativa (se transforma en numérica para su tratamiento estadístico).
- El objetivo será, dado un nuevo individuo, obtener los valores de $V1, V2$ y $V3$ y utilizar esta información para clasificarlo en el grupo 1 ó 2.
- También es de nuestro interés determinar qué variables tienen más peso a la hora de asignar el individuo al grupo más probable al que pertenece.

ENFOQUES DE ANÁLISIS

1) Basado en la obtención de funciones discriminantes de cálculo similar a las ecuaciones regresión lineal múltiple. Consiste en conseguir, a partir de las variables explicativas, unas funciones lineales de éstas con capacidad para clasificar otros individuos. A cada nuevo caso se aplican dichas ecuaciones y la función de mayor valor define el grupo al que pertenece.

2) Basado en técnicas de correlación canónica y de componentes principales (Análisis Factorial) denominado Análisis Discriminante Canónico.

CLASIFICACIÓN EN DOS GRUPOS

- Estudiamos la aplicación del Análisis Discriminante a la clasificación de individuos en el caso de que se puedan asignar solamente a dos grupos a partir de k variables discriminadoras.

- Problema resuelto por Fisher mediante su función discriminante:

$$D = u_1X_1 + u_2X_2 + u_3X_3 + \ldots + u_kX_k$$

- Las puntuaciones discriminantes son los valores que se obtienen al dar valores a $X_1, X_2, \ldots, X_k$ en la ecuación anterior.

- Se trata de obtener los coeficientes de ponderación u_j.

■ Si se considera N observaciones $\Longrightarrow$ La función discriminante

$$\boxed{D_i = u_1X_{1i} + u_2X_{2i} + u_3X_{3i} + \ldots + u_kX_{ki} \qquad \forall i = 1, \ldots, N}$$

D_i es la puntuación discriminante correspondiente a la observación $i-$ésima.

■ Función discriminate en forma matricial:

$$\boxed{\begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_N \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \ldots & X_{k1} \\ X_{12} & X_{22} & \ldots & X_{k2} \\ \vdots & \vdots & & \vdots \\ X_{1N} & X_{2N} & \ldots & X_{kN} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{pmatrix}}$$

■ Expresando el modelo en función de las desviaciones a la media

$$\boxed{\begin{pmatrix} D_1 - \overline{d}_1 \\ D_2 - \overline{d}_2 \\ \vdots \\ D_N - \overline{d}_N \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \ldots & X_{k1} \\ X_{12} & X_{22} & \ldots & X_{k2} \\ \vdots & \vdots & & \vdots \\ X_{1N} & X_{2N} & \ldots & X_{kN} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{pmatrix}}$$

Es decir:

$$\boxed{d = Xu}$$

■ La variabilidad de la función discriminante (suma de cuadrados de las desviaciones de las variables discriminantes con respecto a su media) se expresa:

$$\boxed{d'd = u'X'Xu}$$

♣ $X'X$ es una matriz simétrica que expresa las desviaciones cuadráticas con respecto a la media de las variables (suma de cuadrados total). Se puede descomponer en suma de cuadrados entre grupos F y suma de cuadrados intragrupos V .

$$\boxed{X'X = F + V}$$

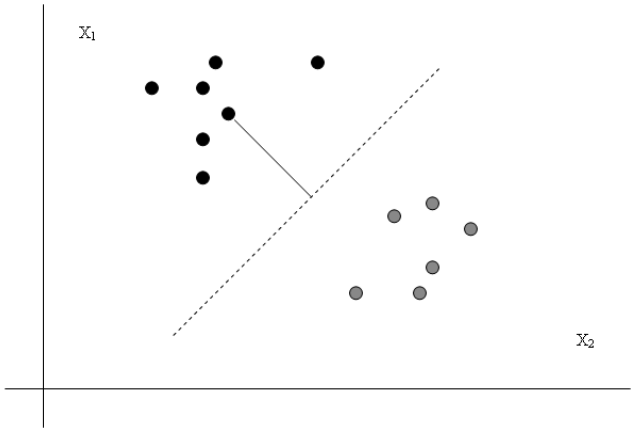
$$\Downarrow$$

$$\boxed{d'd = u'X'Xu = u'(F+V)u = u'Fu + u'Vu}$$

- Los ejes discriminantes vienen dados por los vectores propios asociados a los valores propios de la matriz $V^{-1}F$ ordenados de mayor a menor.
- Las puntuaciones discriminantes se corresponden con los valores obtenidos al proyectar cada punto del espacio $k-$dimensional de las variables originales sobre el eje discriminante.

EJEMPLO

Consideremos las variables X_1 y X_2 que se miden en un conjunto de 13 individuos. La nube de puntos resultante es



Los centros de gravedad o centroides (vector de medias) resumen la información sobre los grupos.

CENTROIDES PARA CADA GRUPO (GRUPO I Y GRUPO II)

$$\overline{x}_I = \begin{pmatrix} \overline{X}_{1I} \\ \overline{X}_{2I} \\ \vdots \\ \overline{X}_{kI} \end{pmatrix} \qquad \overline{x}_{II} = \begin{pmatrix} \overline{X}_{1II} \\ \overline{X}_{2II} \\ \vdots \\ \overline{X}_{kII} \end{pmatrix}$$

Los subíndices I y II indican a qué grupo pertenece la variable.

PARA CADA GRUPO SE OBTIENE

$$\overline{D}_I = u_1\overline{X}_{1I} + u_2\overline{X}_{2I} + \ldots + u_k\overline{X}_{kI}$$

$$\overline{D}_{II} = u_1\overline{X}_{1II} + u_2\overline{X}_{2II} + \ldots + u_k\overline{X}_{kII}$$

CRITERIO PARA CLASIFICAR A UN INDIVIDUO

★ Si $D_i < C$ se clasifica al individuo i en el grupo I .

★ Si $D_i > C$ se clasifica al individuo i en el grupo II .

♦ **C: punto de corte discriminante**

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

EN GENERAL

$$D - C = u_1X_1 + u_2X_2 + \ldots + u_kX_k - C$$

Se clasifica dependiendo de si $D - C$ es positivo o negativo

OBSERVACIONES

■ **Relación entre el análisis de regresión y el análisis discriminante con dos grupos:**

Si se realiza una recta por mínimos cuadrados, tomando como variable dependiente la variable que define la pertenencia a uno u otro grupo y como variables explicativas a las variables clasificadoras; los coeficientes que se obtienen guardan una estricta proporcionalidad con la función discriminante de Fisher.

■ **Contrastes de significación y evaluación de la bondad del ajuste:**

Responden a las cuestiones

★ ¿Se cumple la hipótesis de homocedasticidad del modelo?

★ ¿Se cumple la hipótesis de normalidad?

★ ¿Difieren significativamente las medias poblacionales de los dos grupos?

CLASIFICACIÓN EN MÁS DE DOS GRUPOS

ANÁLISIS DISCRIMINANTE MÚLTIPLE

■ Número máximo de ejes discriminantes mín $(G-1, k)$ (G=número de categorías). Se obtienen $G-1$ ejes discriminantes si el número de variables explicativas es mayor o igual que $G-1$ (hecho que suele ser generalmente cierto).

■ Cada una de las funciones discriminantes D_i se obtiene como función lineal de las k variables explicativas.

$$D_i = u_{i1}X_1 + u_{i2}X_2 + \ldots + u_{ik}X_k \qquad i = 1, \ldots, G-1$$

■ Los $G-1$ ejes vienen definidos respectivamente por los vectores u_1, u_2, ... , u_{G-1}.

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

CONCLUSIÓN

■ Los ejes discriminantes son las componentes de los vectores propios normalizados asociados a los valores propios de la matriz $V^{-1}F$ ordenados.

CONTRASTES DE SIGNIFICACIÓN

■ Se plantean contrastes específicos para determinar si cada uno de los valores propios obtenidos contribuye a la discriminación entre los diferentes grupos.

Bibliografía utilizada:

★ **R. Gutiérrez, A. González, F. Torres, J.A. Gallardo (1994).** *"Técnicas de Análisis de datos Multivariable. Tratamiento computacional". Universidad de Granada.*

★ **B. Visauta Vinacua (1998).** *"Análisis estadístico con SPSS para Windows, volumen II: Estadísitca multivariante". McGraw Hill .*

♦ **Temporalización:** Dos horas