Tema 0: Introducción al Análisis Multivariante

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE I

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones

Índice

Definición

Definición y ejemplos

- Análisis Multivariante:

En un sentido amplio, se refiere a todos los métodos estadísticos que analizan simultáneamente medidas múltiples de cada individuo u objeto sometido a investigación.

Cualquier análisis simultáneo de más de dos variables puede ser considerado aproximadamente como un análisis multivariante.

Muchas técnicas multivariantes son extensiones del análisis univariante.

Objetivos y técnicas multivariantes

Objetivos

1. Resumir los datos mediante un pequeño conjunto de nuevas variables, construidas como transformaciones de las originales, con la mínima pérdida de información.
2. Ordenar y agrupar casos (u observaciones) similares o variables similares, en base a una serie de características comunes
3. Examinar las relaciones de dependencia entre variables
4. Clasificar nuevas observaciones en grupos definidos, es decir, predecir la clasificación de futuras observaciones, en base a los datos observados anteriormente en mediciones comunes
5. Testear y validar hipótesis estadísticas, formuladas en base a parámetros poblacionales multivariantes

Objetivos y técnicas multivariantes

- Ejemplos indicativos de aplicaciones en distintas disciplinas:

1. **Administración de empresas**: construir tipologías de clientes
2. **Agricultura**: clasificar terrenos de cultivo por fotos aéreas
3. **Arqueología**: clasificar restos arqueológicos
4. **Biometría**: identificar los factores que determinan la forma de un organismo vivo.
5. **Ciencias de la computación**: diseñar algoritmos de clasificación automática
6. **Ciencias de la educación**: investigar la efectividad de las dimensiones de la contaminación ambiental
7. **Documentación**: clasificar revistas por sus artículos y construir indicadores bibliométricos
8. **Economía**: identificar las dimensiones del desarrollo económico.
9. **Geología**: clasificar sedimentos
10. **Historia**: determinar la importancia relativa de los factores que caracterizan los períodos prerrevolucionarios.
11. **Ingeniería**: transmitir óptimamente señales por canales digitales
12. **Lingüística**: encontrar patrones de asociación de palabras
13. **Medicina**: identificar tumores mediante imágenes digitales
14. **Psicología**: determinar los factores que componen la inteligencia humana
15. **Sociología y Ciencias políticas**: construir tipologías de los votantes de un partido

- Algunas técnicas estadísticas para el aprendizaje (técnicas multivariantes):

Métodos de dependencia

1. **Regresión múltiple**
2. **Análisis discriminante múltiple**
3. **Análisis logit**
4. **Análisis multivariante de la varianza (MANOVA)**
5. **Análisis de la correlación canónica**

Métodos de interdependencia:

1. **Análisis de Componentes Principales**
2. **Análisis Factorial**
3. **Análisis Clúster**
4. **Análisis de Correspondencias**

Componentes Principales:

- Variables cuantitativas
- Se analiza si es posible representar adecuadamente la información de p variables con un número menor de variables construidas como combinaciones lineales de las originales.
- Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos.

Análisis clúster:

- Variables cuantitativas

- El objetivo es agrupar las variables, o individuos, en una serie de grupos, de manera que los que pertenecen al mismo grupo presenten más similitudes que con el resto de variables que forman los demás grupos.

- Los grupos no están predefinidos.

- Normalmente se hace en dos etapas.
  - Primera etapa: definición de los grupos con análisis clúster
  - Segunda etapa: describir las personas o variables para determinar su composición.

Notación

Notación

- **p variables numéricas univariantes**. El conjunto de las p variables forman una variable vectorial o multivariante.

- Los valores de las p variables en cada uno de los n elementos pueden representarse en una matriz X de dimensiones (n x p), que llamaremos matriz de datos.

- Una matriz de datos está formada por filas que son los individuos y por columnas que son las variables.

$$X = \{x_{ij}\} \; donde$$

$$i = 1,....,n \; representa \; el \; individuo$$

$$j = 1,...,p \; representa \; la \; variable$$

Notación

Para cada uno de los individuos tenemos el valor de las variables

$$
\begin{array}{lccccccc}
 & \text{var } 1 & \text{var } 2 & . & . & \text{var } j & . & . & \text{var } p \\
Ind\ 1 & x_{11} & x_{12} & . & . & x_{1j} & . & . & x_{1p} \\
Ind\ 2 & x_{21} & x_{22} & . & . & x_{2j} & . & . & x_{2p} \\
. & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
Ind\ i & x_{i1} & x_{i2} & . & . & x_{ij} & . & . & x_{ip} \\
. & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . \\
Ind\ n & x_{n1} & x_{n2} & . & . & x_{nj} & . & . & x_{np}
\end{array}
= X_{n\,x\,p}
$$

Notación

- Representación del individuo $i \in \Re^p$

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ x_{ip} \end{pmatrix}$$

- Representación de una variable $j \in \Re^n$

$$x^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ . \\ . \\ x_{nj} \end{pmatrix}$$

Estadística descriptiva multivariante

Definición.-

Dado un vector aleatorio X definiremos su vector media como $\mu' = (\mu_1, \mu_2, \ldots, \mu_p)$.

Donde $\mu_i = E(X_i)$

Definición.-

La varianza de la i-ésima componente de X es

$$Var(X_i) = E\left[\left(X_i - \mu_i\right)^2\right] = E(X_i^2) - \mu_i^2 = \sigma_i^2$$

Definición.-

La covarianza de dos variables X_i y X_j se define como

$$Cov(X_i, X_j) = E\left[\left(X_i - \mu_i\right)\left(X_j - \mu_j\right)\right] = E(X_i X_j) - \mu_i \mu_j = \sigma_{ij}$$

La extensión de la noción de varianza a un vector aleatorio p-dimensional X es la matriz de covarianzas (de orden pxp y simétrica)

$$E\left\{\left[X-E(X)\right]\left[X-E(X)\right]'\right\}=\Sigma=\begin{pmatrix}\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2\end{pmatrix}$$

Definición.-

El coeficiente de correlación entre dos variables X_i y X_j se define como:

$$\rho_{ij} = \sigma_{ij} \Big/ \sqrt{\sigma_i^2}\sqrt{\sigma_j^2}$$

Si las variables son independientes, su covarianza, y por tanto su correlación, son cero.

Definición.-

Todo vector aleatorio p-dimensional tiene asociada una matriz de correlaciones:

$$P = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

Si denotamos por D la matriz diagonal de las desviaciones típicas de las componentes de X, las matrices de covarianzas y correlaciones se pueden relacionar mediante

$$P = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix} \qquad D = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_p \end{pmatrix}$$

$$\sum = DPD$$

$$P = D^{-1} \sum D^{-1}$$

Propiedades:

- La varianza siempre es no negativa. $\sum$ y P son dos matrices semidefinidas positivas. Tiene traza, valores propios y determinante no negativos.

- rang ($\sum$)=rang (P)

- Si rang ($\sum$) < p entonces existen una o más componentes del vector aleatorio X que son combinación lineal del resto.

Cuando trabajamos con una muestra, la matriz de covarianzas muestral la denotamos por $S=[S_{ij}]$, y se definirá como:

$$S = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix}$$

donde S_{ij} representa la covarianza muestral entre las variables X_i y X_j:

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

La matriz de correlaciones muestral se denota por $R=[r_{ij}]$, y se define como:

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

donde r_{ij} es la correlación muestral entre las variables X_i y X_j:

$$r_{ij} = s_{ij} \Big/ \sqrt{s_i^2}\sqrt{s_j^2}$$

Combinaciones lineales de variables

Si $a' = (a_1, a_2, \ldots, a_p)$ es un vector de constantes, entonces podemos definir

$$Y = a'X = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p$$

En tal caso la media poblacional de Y vendrá dada por

$$E(Y) = E(a'X) = a'E(X) = a'\mu$$

y la varianza poblacional por:

$$Var(Y) = a' \sum a$$

Distribución normal multivariante

Bibliografía

- Cuadras, C. M. (2007). Nuevos métodos de análisis multivariante. Barcelona: CMC editions.

- Peña, D. (2002). Análisis de datos multivariantes. McGraw-Hill.