Departamento de Estadística e Investigación Operativa
3ºDDiplomatura de Estadística

Programa de la asignatura Ampliación de Análisis de Datos Multivariantes.
3º Diplomatura de Estadística. Curso 2011-12
Programa de Teoría.

Tema 1: Introducción al Análisis Cluster. Consideraciones Generales.

1. El problema de la clasificación.
2. El Análisis Cluster.
3. Cluster por individuos y por variables.
4. Clasificación de las técnicas cluster.
   (a) Métodos Jerárquicos.
   (b) Métodos no Jerárquicos.
5. Etapas en el Análisis Cluster

Tema 2: Medidas de Asociación.

1. Introducción.
2. Distancias y similaridades.
   (a) Distancias. Propiedades.
   (b) Similaridades. Propiedades.
3. Medidas de asociación entre variables.
   (a) Coseno del ángulo de vectores.
   (b) Coeficiente de correlación.
   (c) Medidas para datos binarios.
   (d) Medidas basadas en probabilidades condicionadas.
4. Medidas de asociación entre individuos.
   (a) Distancia euclídea, de Minkowski y de Mahalanobis.
   (b) Correlación entre individuos.
   (c) Distancias derivadas de la distancia χ^2.
   (d) Medidas no métricas. Coeficiente de Bray-Curtis.
   (e) Medidas para datos binarios.

Tema 3: Métodos Jerárquicos de Análisis Cluster.

1. Introducción.
2. Métodos jerárquicos aglomerativos.
   (a) Estrategia de la distancia mínima o similitud máxima.
   (b) Estrategia de la distancia máxima o similitud mínima.
   (c) Estrategia de la distancia o similitud promedio no ponderado.
   (d) Estrategia de la distancia o similitud promedio ponderado.

(e) Métodos basados en el centroide.
    i. Método del promedio ponderado.
    ii. Método de la Mediana.
(f) Método de Ward.

3. Ejemplo numérico de Análisis Cluster.
4. Fórmula de recurrencia de Lance-Williams.
5. Métodos jerárquicos disociativos.
6. La matriz cofenética. Coeficiente de correlación cofenético.
7. El problema del número de clusters a determinar. Técnicas de validación.

Tema 4: Métodos no Jerárquicos de Análisis Cluster.

1. Introducción.
2. Puntos semilla. Métodos de elección.
3. Particiones iniciales. Métodos de elección.
4. Métodos que fijan el número de clusters.
   (a) Método de Forgy y variante de Jancey. Ejemplo.
   (b) Método de las K-Medias de MacQueen.
   (c) Algunas cuestiones sobre estos métodos.
   (d) Propiedades de convergencia.
5. Métodos con el número final de clusters variable.
   (a) Nueva versión del método de las K-Medias.
   (b) Variante de Wishart del método de las K-Medias.
   (c) El método Isodata.

Bibliografía.

- **Aldenderfer, M.S. y Blashfield, R.K.** (1989). *Cluster Analysis*. Series: Quantitative Applications in the Social Sciences. Sage University Paper.
- **Anderberg, M.R.** (1973). *Cluster Analysis for applications*. Academic Press.
- **Duran Benjamin, S. y Odell, P.L.** (1974). *Cluster Analysis*. Lecture Notes in Economics and Mathematical Systems. Springer-Verlag.
- **Escudero, L.F.** (1977). *Reconocimiento de patrones.* Paraninfo.
- **Everitt, B.S.** (2001). *Cluster Analysis*. Edward Arnold.
- **Gutiérrez, R.; González, A.; Torres, F. y Gallardo, J.A.** (1994). *Técnicas de Análisis de datos multivariable. Tratamiento computacional.*
- **Hair, J.** (2000). *Análisis Multivariante.* Prentice Hall.
- **Peña, D.** (2002). *Análisis de datos multivariantes.* McGraw-Hill.
- **Romesburg, H.C.** (1984). *Cluster Analysis for researchers*. Lifetime Learning Publications.
- **Späth, H.** (1982). *Cluster Analysis algorithms*. John Wiley & Sons.

Programa de Prácticas.

El objetivo perseguido en la realización de las prácticas de ordenador, es mostrar al alumno la resolución de ejercicios prácticos directamente relacionados con las técnicas teóricas estudiadas, de forma que estos desarrollos adquiridos sepa traducirlos en la resolución de casos prácticos. Para las prácticas de ordenador se utilizarán los paquetes estadísticos BMDP, (capítulos 1M, 2M y KM), SPSS y R. Para la realización de Técnicas de Validación se emplearán programas de propia elaboración.

Una parte de las prácticas se encuentra completamente resuelta, y se proponen otras para su resolución. El programa comprende las siguientes prácticas en ordenador:

1. Dos prácticas de Análisis Cluster por variables. Programas utilizados: BMDP (capítulo 1M), SPSS y R.
2. Dos prácticas de Análisis Cluster por individuos. Programas utilizados: BMDP (capítulo 2M), SPSS y R.
3. Dos prácticas de Análisis Cluster no jerárquico mediante el procedimiento de las K-Medias. Programas utilizados: BMDP (capítulo KM), SPSS y R.

Pautas a seguir en la realización de las prácticas:

1. Para la realización de las prácticas se hará una introducción de las diferentes órdenes que deben utilizarse en los capítulos del BMDP, y las secuencias y menús usados en SPSS y R.
2. La primera práctica de cada bloque tiene como objetivo mostrar al alumno la aplicación directa y pormenorizada de cada uno de los pasos seguidos en el desarrollo teórico y será expuesta de manera muy exhaustiva.
3. Las siguientes prácticas resueltas de cada bloque, ofrecen una visión de las diferentes variantes que se pueden plantear en cada técnica estudiada, y llevan una graduación de menor a mayor dificultad.
4. Por último se propone al alumno la realización completa de una de las prácticas, siguiendo el mismo modelo que las prácticas resueltas.

Evaluación de la asignatura.

El alumno deberá realizar un ejercicio escrito sobre las diferentes estrategias utilizadas en el análisis cluster.

Finalmente podrá presentar por escrito los resultados de un ejemplo práctico realizado. La nota de este trabajo puede permitir al alumno aumentar en un escalón la obtenida en el ejercicio escrito.