Universidad Austral
Maestría en Ciencia de Datos
Introducción al Data Mining

INTRODUCCIÓN A LA MINERÍA DE DATOS

ANÁLISIS CLUSTER

- Ejercitación -

Docentes
Mg. Leandro Kovalevski
Ing. Pablo Beltramone

Universidad Austral
Maestría en Ciencia de Datos
Introducción al Data Mining

Practica Análisis Cluster

1. Método de aglomeración jerárquico

Los cereales en copos constituyen uno de los tantos alimentos que componen la base de la pirámide de alimentación. Para estudiar de las características nutricionales de los cereales en copos de consumo habitual, se obtuvo una muestra de 43 productos disponibles en el mercado.
Se registraron las variables que se detallan a continuación:

- *Marca*: marca comercial del cereal.
- *Nombre*: nombre comercial del cereal.
- *Público*: indica si el cereal es para adultos o para chicos.
- *Paquete*: indica si está envasado en caja o en bolsa.
- *Preciox350*: precio, en pesos, llevado a una base de 350 gramos.
- *Calorías*: Kilocalorías por cada 100 gramos de cereal.
- *Proteínas* (en gramos por cada 100 gramos de cereal).
- *Carbohidratos* (en gramos por cada 100 gramos de cereal).
- *Fibras* (en gramos por cada 100 gramos de cereal).
- *Lípidos* (en gramos por cada 100 gramos de cereal).
- *Sodio* (en miligramos por cada 100 gramos de cereal).
- *Potasio* (en miligramos por cada 100 gramos de cereal).
- *Vitamina* C (en miligramos por cada 100 gramos de cereal).
- *Hierro* (en miligramos por cada 100 gramos de cereal).

El conjunto de datos se encuentra en el archivo 'datoscereales.xls'.

1.1 Describir el conjunto de datos.

1.2 Aplicar un método de agrupamiento jerárquico considerando sólo las características nutricionales.

1.3 De ser necesario, estandarizar las variables. Justifique su elección.

1.4 ¿Cuantos grupos se identifican en el dendograma? ¿Existe alguna manera de determinar el número óptimo de clusters a considerar?

1.5 ¿Existe alguna manera de determinar el número óptimo de clusters a considerar?

1.6 ¿Qué características tienen los grupos formados?

1.7 ¿Qué sucede si eligiéramos trabajar con un cluster más? ¿Cuáles de los grupos se divide? ¿en qué se diferencian los nuevos grupos formados?

1.8 ¿Cómo se distribuyen las variables categóricas 'marca', 'publico' y 'paquete' en los clusters encontrados? ¿Qué se puede interpretar?

1.9 A través del análisis de componentes principales realizado previamente, ¿se podía intuir la presencia de distintos grupos y sus características?

1.10 Repetir el análisis con otras medida de distancia de Canberra y de Minkowski con p = 5? . Cambian los grupos formados? Qué esperaría que ocurra con los 'outliers'?

1.11 ¿Cómo se pueden incorporar esas variables categóricas al análisis? ¿Cómo varía el agrupamiento?

2. Método de aglomeración no jerárquico

Utilizando los datos del archivo 'consumos_2018.csv' donde se presenta la información sobre más de 10.000 hogares referida a características de la región donde se encuentran los hogares y al consumo de distintos productos y servicios:

2.1 Aplique un método de agrupamiento no jerárquico considerando sólo las variables socio-demográficas.
2.2 ¿Qué características tienen los grupos formados?
2.3 Imagínese que luego de presentar los resultados, la persona que evalúa los resultados le cuestiona por qué no incluyó la variable 'zona'. ¿Qué opina usted? ¿Es correcto incluirla? Jusitifique.
2.4 Dado que 'zona' es una variable categórica en escala ordinal, ¿cómo podría incluirla en el análisis?
2.5 Realice una nueva segmentación incluyendo también las variables de consumo: 'var_0001', 'var_0002', …, 'var_0012'.
2.6 ¿Podría aplicarse un agrupamiento jerárquico a este conjunto de datos?

3. Aglomeración con variables binarias

A partir de la base de datos 'nacimientos.xls' que contiene información de 51 nacimientos en tres centros de salud diferentes sobre el recién nacido, la madre y el parto en las siguientes variables:

Variable	Descripción
id	indicadora del parto.
centro	centro donde se realizó el parto ('A', 'B' o 'C').
eg_menor_30	variable indicadora de si la edad gestacional del RN fue menor a 30 meses. (0: No, 1: Si).
ematerna_mayor_40	variable indicadora de si la edad de la madre era mayor a 40. (0: No, 1: Si).
hta	variable indicadora de si la madre presentaba hipertensión arterial. (0: No, 1: Si).
sexo_masculino	variable indicadora de si el sexo del bebé era masculino. (0: No, 1: Si).
corti	variable indicadora de si hubo aplicación de corticoides. (0: No, 1: Si).
Cesarea	variable indicadora de si el parto fue por cesárea. (0: No, 1: Si).
peso_menor_1000	variable indicadora de si el peso del RN fue menor a 1000 g. (0: No, 1: Si)

3.1 Describa el conjunto de datos.
3.2 Aplicar un método de agrupamiento jerárquico considerando sólo las variables binarias.
3.3 ¿Cuál es el número óptimo de clusters a considerar?
3.4 Describa las características de los grupos considerados
3.5 ¿Existe alguna relación en la distribución de los nacimientos de los centros ('A', 'B' y 'C') entre los distintos clusters? ¿Qué se puede interpretar?