Universidad de La Laguna

Escuela de Doctorado y Estudios de Posgrado

Departamento de Ingeniería Informática y de Sistemas

Extracción de conocimiento en bases de datos

PRÁCTICA: AGRUPAMIENTO

J. Marcos Moreno Vega

UNIVERSIDAD DE LA LAGUNA

TUFTE-LATEX.GOOGLECODE.COM

Objetivo:
Obtener una buena segmentación de los clientes de un centro comercial.

Base de datos:
mailCustomers.csv

Lenguaje de programación:
R o Python.

Segmentación de clientes

La segmentación de clientes es una tarea relevante que pretende identificar los diferentes grupos de clientes que tiene una empresa. La segmentación puede hacerse atendiendo a criterios demográficos, geográficos, de comportamiento de compra o usando una combinación de ellos.

Base de datos

En la base de datos *mallCustomers.csv* se recogen los datos de 200 clientes de un centro comercial. El fichero consta de 5 variables, incluyendo el identificador de cada cliente.

En la tabla 1.1 se enumeran y describen todas las variables.

Variable	Descripción	Valores
CustomerID	identificador del cliente	Numérico
Gender	género del cliente	Male/Female
Age	edad del cliente	Numérico
Annual.income	ingresos anuales del cliente	Numérico
Spending.score	puntuación dada al cliente por el supermercado	Numérico

Tabla 1.1: Base de datos mallCustomers.csv

Tareas

a) Aplicando el algoritmo *k*-means, identificar un buen agrupamiento de los clientes fijando adecuadamente el valor de *k*.

b) Construir los dendrogramas de los agrupamientos jerárquicos que se obtienen usando los métodos *single*, *centroid* y *ward*. Comparar los dendrogramas obtenidos para identificar el que preserva mejor las distancias originales. Usar este dendrograma para agrupar a los clientes.

c) Repetir las tareas anteriores considerando por separado a los clientes con género *male* y *female*. ¿Hay diferencias entre los grupos de clientes encontrados al separar por el género? En caso afirmativo, enumerar las diferencias.

Qué debe presentar el alumno

a) Cuaderno Jupyter, debidamente comentado, usado para preparar y analizar los datos.

b) Una memoria en formato pdf en la que se describan brevemente el análisis realizado, los grupos de clientes identificados y una descripción de los mismos.