Fundamentos de la Ciencia de Datos

Práctica 1

Grado en Ingeniería Informática
Universidad de Alcalá

Pablo García García
Abel López Martínez
Álvaro Jesús Martínez Parra
Raúl Moratilla Núñez

14 de noviembre de 2023

Índice general

Introducción

Parte 1

Ejercicios guiados

En esta primera parte de esta práctica, repetirán los ejercicios explicados y realizados por el profesor en las clases de laboratorio, utilizando los mismos procedimientos vistos plasmándolos en un documento LaTeX.

1.1. Descripción de los datos

"*El primer conjunto de datos, que se empleará para realizar el análisis de descripción de datos, estará formado por datos de una característica cualitativa, nombre, y otra cuantitativa, radio, de los satélites menores de Urano, es decir, aquellos que tienen un radio menor de 50 Km, dichos datos, los primeros cualitativos nominales, y los segundos cuantitativos continuos, son: (Nombre, radio en Km): Cordelia, 13; Ofelia, 16; Bianca, 22; Crésida, 33; Desdémona, 29; Julieta, 42; Rosalinda, 27; Belinda, 34; Luna-1986U10, 20; Calíbano, 30; Luna-999U1, 20; Luna 1999U2, 15.*"

Para comenzar con la resolución de este ejercicio, deberemos escribir los datos en un fichero `.txt`, cumpliendo las siguientes normas:

- Existirá una tabulación entre dato y dato.
- La primera columna numera las filas, y en la primera fila se introduce un espacio y el nombre de las variables.
- Se introducirá un salto de línea en la última fila
- Para los números decimales se utilizarán puntos.
- Al escribir nombres, no se deberán introducir espacios.

Obedeciendo a estas normas, copiamos los datos en un fichero llamado `satelites.txt`, y lo cargamos en R de la siguiente manera:

```
s <- read.table("data/satelites.txt")
print(s)

##          nombre radio
## 1      Cordelia    13
## 2        Ofelia    16
## 3        Bianca    22
## 4       Crésida    33
## 5     Desdémona    29
## 6       Julieta    42
## 7     Rosalinda    27
## 8       Belinda    34
## 9  Luna-1986U10    20
## 10     Calíbano    30
## 11   Luna-999U1    20
## 12  Luna-1999U2    15
```

Ahora en la variable `s` tenemos un dataframe con los datos de nuestros satélites. En los dataframes se accede por `[fila, columna]`, y también podemos consultar las dimensiones con la función `dim`. Sería de esperar que nos dijera que tiene 12 filas (los 12 datos), y 2 columnas (`nombre` y `radio`).

```
dim(s)

## [1] 12  2
```

También podemos ordenar el dataframe, en función de una de las magnitudes (columnas), usando la función `order` aplicando recursivamente el concepto de acceder por filas y columnas. Veamos un ejemplo, si en `s` teníamos guardado nuestro dataframe, y queremos ordenar por `radio`, la manera de hacerlo sería la siguiente:

```
s_ordered <- s[order(s$radio), ]
print(s_ordered)

##          nombre radio
## 1      Cordelia    13
## 12  Luna-1999U2    15
## 2        Ofelia    16
## 9  Luna-1986U10    20
## 11   Luna-999U1    20
## 3        Bianca    22
## 7     Rosalinda    27
## 5     Desdémona    29
## 10     Calíbano    30
## 4       Crésida    33
```

```
## 8       Belinda    34
## 6       Julieta    42
```

Podemos introducir nuevos criterios a la ordenación, como por ejemplo, hacerlo en orden descendente. Para esto usaremos la función `rev`.

```
s_ordered_rev <- s[rev(order(s$radio)), ]
print(s_ordered_rev)

##          nombre radio
## 6       Julieta    42
## 8       Belinda    34
## 4       Crésida    33
## 10     Calíbano    30
## 5     Desdémona    29
## 7     Rosalinda    27
## 3        Bianca    22
## 11   Luna-999U1    20
## 9  Luna-1986U10    20
## 2        Ofelia    16
## 12  Luna-1999U2    15
## 1      Cordelia    13
```

1.2. Asociación

"*El segundo conjunto de datos, que se empleará para realizar el análisis de asociación, estará formado por las siguientes 6 cestas de la compra:* {*Pan, Agua, Leche, Naranjas*}, {*Pan, Agua, Café, Leche*}, {*Pan, Agua, Leche*}, {*Pan, Café, Leche*}, {*Pan, Agua*}, {*Leche*}."

1.3. Detección de datos anómalos

1.3.1. Primer ejercicio

"*El tercer conjunto de datos, que se empleará para realizar el análisis de detección de datos anómalos utilizando técnicas con base estadística, estará formado por los siguientes 7 valores de resistencia y densidad para diferentes tipos de hormigón* {*Resistencia, Densidad*}*:* {*3, 2; 3.5, 12; 4.7, 4.1; 5.2, 4.9; 7.1, 6.1; 6.2, 5.2; 14, 5.3*}*. Aplicar las medidas de ordenación a la resistencia y las de dispersión a la densidad.*"

1.3.2. Segundo ejercicio

"*El cuarto conjunto de datos, que se empleará para realizar el análisis de detección de datos anómalos utilizando técnicas basadas en la proximidad y en la densidad, estará formado*

por las siguientes 5 calificaciones de estudiantes: 1. {4, 4}; 2. {4, 3}; 3. {5, 5}; 4. {1, 1}; 5. {5, 4} donde las características de las calificaciones son: (Teoría, Laboratorio)."

Parte 2

Ejercicios autónomos

2.1. Descripción de los datos

"*El primer conjunto de datos, que se empleará para realizar el análisis de descripción de datos, estará formado por datos de una característica cuantitativa, distancia, desde el domicilio de cada estudiantes hasta la Universidad, dichos datos, cuantitativos continuos, son: 16.5, 34.8, 20.7, 6.2, 4.4, 3.4, 24, 24, 32, 30, 33, 27, 15, 9.4, 2.1, 34, 24, 12, 4.4, 28, 31.4, 21.6, 3.1, 4.5, 5.1, 4, 3.2, 25, 4.5, 20, 34, 12, 12, 12, 12, 5, 19, 30, 5.5, 38, 25, 3.7, 9, 30, 13, 30, 30, 26, 30, 30, 1, 26, 22, 10, 9.7, 11, 24.1, 33, 17.2, 27, 24, 27, 21, 28, 30, 4, 46, 29, 3.7, 2.7, 8.1, 19, 16.*"

2.2. Asociación

"*El segundo conjunto de datos, que se empleará para realizar el análisis de asociación, estará formado por las siguientes conjuntos de extras incluidos en 8 ventas de coches: {X, C, N, B}, {X, T, B, C}, {N, C, X}, {N, T, X, B}, {X, C, B}, {N}, {X, B, C}, {T, A}. Donde: {X: Faros de Xenon, A: Alarma, T: Techo Solar, N: Navegador, B: Bluetooth, C: Control de Velocidad}, son los extras que se pueden incluir en cada coche.*"

2.3. Detección de datos anómalos

2.3.1. Primer ejercicio

"*El tercer conjunto de datos, que se empleará para realizar el análisis de detección de datos anómalos utilizando técnicas con base estadística, estará formado por los siguientes 10 valores de velocidades de respuesta y temperaturas normalizadas de un microprocesador {Velocidad, Temperatura}: {10, 7.46; 8, 6.77; 13, 12.74; 9, 7.11; 11, 7.81; 14, 8.84; 6, 6.08; 4, 5.39; 12, 8.15; 7, 6.42; 5, 5.73}. Aplicar las medidas de ordenación a la velocidad y las de dispersión a la temperatura.*"

2.3.2. Segundo ejercicio

"*El cuarto conjunto de datos, que se empleará para realizar el análisis de detección de datos anómalos utilizando técnicas basadas en la proximidad y en la densidad, estará formado por el número de Mujeres y Hombres inscritos en una serie de cinco seminarios que se han impartido sobre biología. Los datos son: {Mujeres, Hombres}: 1. {9, 9}; 2. {9, 7}; 3. {11, 11}; 4. {2, 1}; 5. {11, 9}.*"