

PECL3 - Fundamentos de la ciencia de datos

Mario Adán Herrero Alberto González Martínez
Branimir Stefanov Yanev Diego Gutiérrez Marco

29 de noviembre de 2020

Resumen

En el siguiente documento se presentan los resultados y solución de la PECL3 del laboratorio de Fundamentos de la Ciencia de Datos. Se realizará un análisis de clasificación de los datos utilizando R siguiendo los ejercicios realizados previamente en teoría. Para la resolución del ejercicio utilizaremos el algoritmo de construcción de árboles de decisión de Hunt y también realizaremos un análisis de regresión lineal. Los datos calculados se mostrarán en este pdf utilizando las herramientas Sweave y TinyTex/MikTex.

Índice

1. Introducción	3
2. Ejercicio 1- Análisis de clasificación de datos. Calificaciones - Planetas	3
2.1. Apartado 1 - Árboles de decisión. Algoritmo Hunt	3
2.1.1. Librería rpart	4
2.1.2. Librería tree	4
2.1.3. Apartado 2 - Análisis de regresión lineal	5
3. Ejercicio 2 - Análisis de clasificación de datos. Vehículos - Notas	5
3.1. Apartado 1 - Árboles de decisión. Algoritmo Hunt	5
3.2. Apartado 2 - Análisis de regresión lineal	6
3.2.1. Muestra 1	8
3.2.2. Muestra 2	9
3.2.3. Muestra 3	9
3.2.4. Muestra 4	10
4. Conclusiones	12

1. Introducción

***** CAMBIAR EXPLICACIÓN ***** La práctica consta de dos partes:

En la primera parte, se va a realizar un análisis de asociación, utilizando R, aplicando conceptos vistos en la teoría y haciendo uso del algoritmo Apriori, se resolverá el mismo problema visto en teoría.

En la segunda parte, el grupo desarrollará un enunciado, y su posterior solución, de un ejercicio que contenga modificaciones del ejercicio hecho en clase, en el que se realice un análisis de asociación con R.

2. Ejercicio 1- Análisis de clasificación de datos. Clasificaciones - Planetas

***** CAMBIAR EXPLICACIÓN ***** Para resolver este ejercicio haremos uso del paquete “arules”. Este paquete de R nos proporciona las funciones necesarias para realizar nuestro análisis de asociación.

En este caso en particular analizaremos las compras de varios usuarios para concluir qué productos son comprados juntos habitualmente.

El objetivo es obtener las asociaciones cuyo soporte sea igual o superior al 50 % y cuya confianza sea igual o superior al 80 %.

2.1. Apartado 1 - Árboles de decisión. Algoritmo Hunt

En el primer apartado, se pide que, a partir de la muestra de las clasificaciones de distintos alumnos “calificacionesMuestra.txt”, se obtenga la función de clasificación, utilizando para ellos, la medida de impureza Gini.

La clasificación se realizará a partir de los siguientes atributos, siendo clasificación el atributo clasificador:

1. Teoría
2. Laboratorio
3. Práctica
4. Calificación

A continuación, mostramos la tabla con los datos:

```
> # Primero leemos los datos.  
> calificaciones <- read.table("./data/calificacionesMuestra.txt")  
> muestra <- data.frame(calificaciones)  
> #Mostramos la tabla con los datos.  
> muestra
```

	Teoria	Lab	Pract	Calificacion
suceso1	A	A	B	Ap
suceso2	A	B	D	Ss
suceso3	D	D	C	Ss
suceso4	D	D	A	Ss

suceso5	B	C	B	Ss
suceso6	C	B	B	Ap
suceso7	B	B	A	Ap
suceso8	C	D	C	Ss
suceso9	B	A	C	Ss

Para poder realizar el análisis cargamos las librerías *tree* y *rpart*

```
> # Cargamos librerías necesarias.
> library('tree')
> library('rpart')
```

Para realizar la clasificación, podemos utilizar cualquiera de las librerías cargadas, rpart y tree. Estas dos librerías utilizan para realizar la clasificación la medida de impureza Gini, con lo cual, no hace falta realizar ningún cambio en ninguna de sus llamadas.

2.1.1. Librería rpart

```
> clasificacionRpart <- rpart(Calificacion~, data=muestra, method="class", minsplit=1)
> #Mostramos el resultado
> print(clasificacionRpart)

n= 9

node), split, n, loss, yval, (yprob)
 * denotes terminal node

1) root 9 3 Ss (0.3333333 0.6666667)
  2) Lab=A,B 5 2 Ap (0.6000000 0.4000000)
    4) Pract=A,B 3 0 Ap (1.0000000 0.0000000) *
    5) Pract=C,D 2 0 Ss (0.0000000 1.0000000) *
  3) Lab=C,D 4 0 Ss (0.0000000 1.0000000) *
```

Como podemos observar la clasificación utilizando la librería rpart se realiza de forma correcta.

2.1.2. Librería tree

```
> calificacionTree <- tree(Calificacion~, data=muestra, mincut=1, minsize=2)
> print(calificacionTree)

node), split, n, deviance, yval
 * denotes terminal node

1) root 9 NA NA *
```

Vemos que utilizando la librería tree el resultado de la clasificación es el mismo, esto se debe a que, como hemos dicho anteriormente, las dos librería utilizan como medida de impureza Gini por defecto.

2.1.3. Apartado 2 - Análisis de regresión lineal

Para este apartado, se pide que se realice un análisis de regresión lineal a partir de la muestra de los planetas “planetasMuestra.txt”.

```
> # Leemos y mostramos los datos  
> planetas <- read.table("./data/planetasMuestra.txt")  
> planetas
```

	R	D
Mercurio	2.4	5.4
Venus	6.1	5.2
Tierra	6.4	5.5
Marte	3.4	3.9

***** CAMBIAR EXPLICACIÓN ***** Una vez cargados los datos realizamos el análisis de regresión, entre la densidad (D) y el radio (R), obteniendo los coeficientes asociados:

```
> # Calculamos la regresión  
> regresion = lm(D~R, data=planetas)  
> print(regresion)
```

Call:
lm(formula = D ~ R, data = planetas)

Coefficients:
(Intercept) R
4.3624 0.1394

***** EXPLICAR REGRESION Y INTERPRETAR DATOS SI ES POSIBLE *****

3. Ejercicio 2 - Análisis de clasificación de datos. Vehículos - Notas

Para este segundo ejercicio se deben realizar los mismo cálculos que para el ejercicio 1, pero con otras muestras y cambiando la forma en la que se calculan los resultados. Las muestras utilizadas para este ejercicio serán “muestra1.txt” y “muestra2.txt”.

3.1. Apartado 1 - Árboles de decisión. Algoritmo Hunt

Para el primer apartado, a partir de una muestra con datos sobre las características de los vehículos, se pide realizar un análisis de clasificación de los datos sobre el atributo TipoVehiculo.

La clasificación se realizará a partir de los siguientes atributos, siendo clasificación el atributo clasificador tipo vehículos:

1. Tipo carnet
2. Número de ruedas

3. Número de pasajeros

4. Tipo vehículo

```
> # Leemos los datos de la primera muestra
> datos <- read.table("./data/muestra1.txt")
> muestra1 <- data.frame(datos)
> # Mostramos la tabla
> muestra1

  TipoCarnet NumeroRuedas NumeroPasajeros TipoVehiculo
1          B             4                  5      Coche
2          A             2                  2      Moto
3          N             2                  1  Bicicleta
4          B             6                  4     Camion
5          B             4                  6      Coche
6          B             4                  4      Coche
7          N             2                  2  Bicicleta
8          B             2                  1      Moto
9          B             6                  2     Camion
10         N             2                  1  Bicicleta

> # Realizamos y mostramos la clasificación de los datos en función del campo TipoVehiculo
> clasificacion <- rpart(TipoVehiculo~, data=muestra1, method="class", minsplit=1)
> print(clasificacion)

n= 10

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 10 7 Bicicleta (0.3000000 0.2000000 0.3000000 0.2000000)
  2) TipoCarnet=N 3 0 Bicicleta (1.0000000 0.0000000 0.0000000 0.0000000) *
  3) TipoCarnet=A,B 7 4 Coche (0.0000000 0.2857143 0.4285714 0.2857143)
    6) NumeroRuedas>=3 5 2 Coche (0.0000000 0.4000000 0.6000000 0.0000000)
      12) NumeroRuedas>=5 2 0 Camion (0.0000000 1.0000000 0.0000000 0.0000000) *
      13) NumeroRuedas< 5 3 0 Coche (0.0000000 0.0000000 1.0000000 0.0000000) *
    7) NumeroRuedas< 3 2 0 Moto (0.0000000 0.0000000 0.0000000 1.0000000) *
```

Para poder realizar la visualización de los datos utilizamos la herramienta `plot`

```
> plot(clasificacion, uniform=TRUE, main="Arbol de clasificacion vehiculos")
> text(clasificacion, use.n=TRUE, all=TRUE, cex=.7, fancy=TRUE, fwidth=0.5, fheight=0.7)

***** COMENTAR Y EXPLICAR EL GRÁFICO *****
```

3.2. Apartado 2 - Análisis de regresión lineal

Para este apartado, se pide que se realice un análisis de regresión lineal a partir “muestra2.txt”. El archivo de datos muestra2 ha sido generado mediante

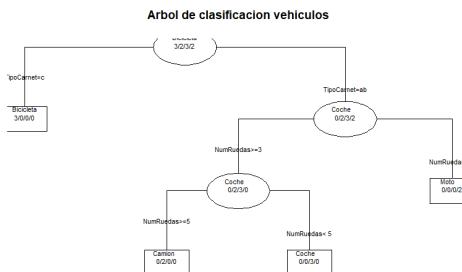


Figura 1: Árbol de decisión - Muestra 1

un script de python el cual genera 40 pares de valores (X, Y) entre 1 y 20, siendo X un número entero e Y un número decimal.

Para la resolución de este ejercicio dividiremos nuestra muestra en 4 muestras de 10 pares de valores cada una para estudiar la regresión.

```

> # Carga de los datos
> datos <- read.table("./data/muestra2.txt", header=TRUE)
> # Mostramos los datos
> datos

```

	X	Y
1	9	8.92
2	9	9.83
3	14	15.70
4	7	7.04
5	13	10.38
6	16	19.91
7	18	15.62
8	18	16.32
9	11	13.63
10	12	15.97
11	16	18.60
12	9	11.29
13	8	11.15
14	5	4.32
15	5	7.76
16	9	8.01
17	6	7.38
18	13	16.05
19	14	13.67
20	7	8.35
21	15	18.90
22	7	10.28
23	6	8.14
24	13	14.52
25	18	13.69

```

26 9 9.16
27 7 10.33
28 13 13.93
29 14 18.53
30 15 13.37
31 15 11.23
32 10 14.18
33 8 5.92
34 4 7.25
35 9 10.32
36 16 16.84
37 11 7.64
38 9 7.47
39 11 12.73
40 9 12.14

> # Dividimos nuestra muestra en 4 muestras de menor tamaño
> muestras <- split(datos, factor(sort(rank(row.names(datos))%%4)))

```

3.2.1. Muestra 1

Calcularemos la regresión para los datos de la primera muestra.

```

> # Datos de la primera muestra
> muestras[1]

$`0`
      X     Y
1   9 8.92
2   9 9.83
3  14 15.70
4   7 7.04
5  13 10.38
6  16 19.91
7  18 15.62
8  18 16.32
9  11 13.63
10 12 15.97

```

Calculamos la regresión de la siguiente forma:

```

> # Calculamos la regresión de la muestra 1
> regresion1 = lm(Y~X, data=muestras[[1]])
> regresion1

Call:
lm(formula = Y ~ X, data = muestras[[1]])

Coefficients:
(Intercept)          X
2.3091        0.8679

***** PODEIS COMENTAR UN POCO VALOR DE REGRESION *****
*****
```

3.2.2. Muestra 2

Calcularemos la regresión para los datos de la segunda muestra.

```
> # Datos de la segunda muestra  
> muestras[2]
```

```
$`1`  
X Y  
11 16 18.60  
12 9 11.29  
13 8 11.15  
14 5 4.32  
15 5 7.76  
16 9 8.01  
17 6 7.38  
18 13 16.05  
19 14 13.67  
20 7 8.35
```

Calculamos la regresión de la siguiente forma:

```
> # Calculamos la regresión de la muestra 2  
> regresion2 = lm(Y~X, data=muestras[[2]])  
> regresion2  
  
Call:  
lm(formula = Y ~ X, data = muestras[[2]])  
  
Coefficients:  
(Intercept) X  
0.9475 1.0555  
  
***** PODEIS COMENTAR UN POCO VALOR DE REGRESION  
*****
```

3.2.3. Muestra 3

Calcularemos la regresión para los datos de la tercera muestra.

```
> # Datos de la tercera muestra  
> muestras[3]
```

```
$`2`  
X Y  
21 15 18.90  
22 7 10.28  
23 6 8.14  
24 13 14.52  
25 18 13.69  
26 9 9.16  
27 7 10.33
```

```

28 13 13.93
29 14 18.53
30 15 13.37

```

Calculamos la regresión de la siguiente forma:

```

> # Calculamos la regresión de la muestra 3
> regresion3 = lm(Y~X, data=muestras[[3]])
> regresion3

Call:
lm(formula = Y ~ X, data = muestras[[3]])

Coefficients:
(Intercept)          X
5.1626           0.6771

***** PODEIS COMENTAR UN POCO VALOR DE REGRESION
*****
```

3.2.4. Muestra 4

Calcularemos la regresión para los datos de la cuarta muestra.

```

> # Datos de la segunda muestra
> muestras[4]

$`3`
      X      Y
31 15 11.23
32 10 14.18
33  8  5.92
34  4  7.25
35  9 10.32
36 16 16.84
37 11  7.64
38  9  7.47
39 11 12.73
40  9 12.14

```

Calculamos la regresión de la siguiente forma:

```

> # Calculamos la regresión de la muestra 4
> regresion4 = lm(Y~X, data=muestras[[4]])
> regresion4

Call:
lm(formula = Y ~ X, data = muestras[[4]])

Coefficients:
(Intercept)          X
3.6999           0.6737

```

***** PODEIS COMENTAR UN POCO VALOR DE REGRESION

Una vez calculada la regresión de las muestras, necesitamos una forma de poder visualizar el resultado, para ello, hemos optado por hacer una figura con las gráficas de regresión de cada una de las muestras.

```
> # Dibujamos las gráficas resultantes de las regresiones de cada muestra
> par(mfrow=c(2,2))
> # Muestra1
> plot(muestras[[1]]$X, muestras[[1]]$Y, main="Muestra 1", xlab="x", ylab="y")
> abline(regresion1, col="blue")
> # Muestra2
> plot(muestras[[2]]$X, muestras[[2]]$Y, main="Muestra 2", xlab="x", ylab="y")
> abline(regresion2, col="red")
> # Muestra3
> plot(muestras[[3]]$X, muestras[[3]]$Y, main="Muestra 3", xlab="x", ylab="y")
> abline(regresion3, col="orange")
> # Muestra4
> plot(muestras[[4]]$X, muestras[[4]]$Y, main="Muestra 4", xlab="x", ylab="y")
> abline(regresion4, col="black")
> mtext(expression(paste(bold("Datos de regresión de las muestras"))),
+           side = 3, line = -2, outer = TRUE)
```

Generamos los siguientes gráficos:

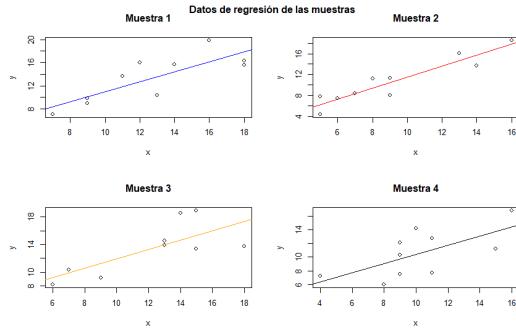


Figura 2: Cálculo de regresión de las muestras

***** COMENTAR Y EXPLICAR EL GRÁFICO *****
DECIR TAMBIEN QUE ESTÁN DISPERSOS DEBIDO A QUE SON VALORES ALEATORIOS *****

4. Conclusiones

***** CAMBIAR EXPLICACIÓN *****

Mediante esta práctica hemos empleado herramientas que permiten clasificar elementos partiendo de una muestra. Observamos que existe gran cantidad de soporte y paquetes por parte de la comunidad para realizar este tipo de análisis de datos. Contemplamos también la facilidad de uso a la hora de realizar estos análisis, y la extensión y potencia de estos análisis.