G4.P-1

David Emanuel Craciunescu Laura P<e9>rez Medeiro

November 5, 2019

1 Data analysis using Hunt decision algorithm and linear regression for qualifications and planets data

1.1

The data use in this exercise will be nine students marks and califications, as it has been done in the teorical classes. For this analysis, it wil be used the gain information meause, using Gini as the impurity measure.

First step in the analysis is read the data, which is contains in a txt file called qualifications.txt:

```
> library(utils)
> qualifications <- read.table("qualification.txt")
> sample = data.frame(qualifications)
```

In order to make the analysis, the package rpart will be used. this means, that it should be install before working with the dataset. In order to manage R packages, it will be used Packrat

```
> library(rpart)
> clasification = rpart(C.G ~ .,
+                       data = sample,
+                       method = "class",
+                       minsplit = 1)
> clasification

n= 9

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

1) root 9 3 Ss (0.3333333 0.6666667)
  2) Labo=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) Labo=C,D 4 0 Ss (0.0000000 1.0000000) *
```

Another package that can be used to do this analysis is tree:

```
> library(tree)
> (clasificationTree = tree(C.G ~ .,
+                           data = sample,
+                           mincut = 1,
+                           minsize = 2)
+ )

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

1) root 9 11.46 Ss ( 0.3333 0.6667 )
  2) Labo: A,B 5  6.73 Ap ( 0.6000 0.4000 )
    4) Pract: A,B 3  0.00 Ap ( 1.0000 0.0000 ) *
    5) Pract: C,D 2  0.00 Ss ( 0.0000 1.0000 ) *
  3) Labo: C,D 4  0.00 Ss ( 0.0000 1.0000 ) *

> clasificationTree

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

1) root 9 11.46 Ss ( 0.3333 0.6667 )
  2) Labo: A,B 5  6.73 Ap ( 0.6000 0.4000 )
    4) Pract: A,B 3  0.00 Ap ( 1.0000 0.0000 ) *
    5) Pract: C,D 2  0.00 Ss ( 0.0000 1.0000 ) *
  3) Labo: C,D 4  0.00 Ss ( 0.0000 1.0000 ) *
```

1.2

In this second part, tha dataset use is planets.txt. To this dataset, linear regression will be applied.

As it has been done before, the first step consist on reading data from a *txt* file:

```
> library(utils)
> data <- read.table("planets.txt")
> data = data.frame(data)
> names(data)

[1] "Radius"  "Density"
```

In order to quantify the correlation between the variables, it will be calculated the coeficient's matrix correlation:

```
> cor(data)

          Radius  Density
Radius  1.000000 0.371063
Density 0.371063 1.000000
```

Then, it will be calculated and representated the minimun square error line:

```
> regression <- lm( Density~Radius, data)
> summary(regression)

Call:
lm(formula = Density ~ Radius, data = data)

Residuals:
Mercurio    Venus   Tierra    Marte
 0.70312 -0.01253  0.24566 -0.93624

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.3624     1.2050   3.620   0.0685 .
Radius        0.1394     0.2466   0.565   0.6289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.846 on 2 degrees of freedom
Multiple R-squared:  0.1377,        Adjusted R-squared:  -0.2935
F-statistic: 0.3193 on 1 and 2 DF,  p-value: 0.6289
```

The equation's line is y = 4.3624 + 0.1394x

```
> library(gplots)
> par(mar = rep(2,4))
> plot(data$Density, data$Radius)
> abline(regression)
```

Finally, it is necessary to calculate ANOVA in order to analysize correctly the relation between variables.

```
> anova <- aov(Density~Radius, data)
> summary(anova)

            Df Sum Sq Mean Sq F value Pr(>F)
Radius       1 0.2286  0.2286   0.319  0.629
Residuals    2 1.4314  0.7157
```

2 Data analysis using Hunt decision algorithm and linear regression for vehicules and pairs of data

The following part consist on doing the same analysis as it has been done before, but now with new datasets.

```
> library(utils)
> vehicules <- read.table("vehiculos.txt")
> sampleV = data.frame(vehicules)
```

In this file, TC = license type, NR = number of roads that has the vehicule, NP = number of people that can be in the vehicule and TV = vehicule's type.

```
> library(rpart)
> clasificationV = rpart(TV ~.,
+                        data = sampleV,
+                        method="class",
+                        minsplit =1)
> clasificationV

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) TC=N 3 0 Bicicleta (1.0000000 0.0000000 0.0000000 0.0000000) *
   3) TC=A,B 7 4 Coche (0.0000000 0.2857143 0.4285714 0.2857143)
     6) NR>=3 5 2 Coche (0.0000000 0.4000000 0.6000000 0.0000000)
      12) NR>=5 2 0 CamiÃ³n (0.0000000 1.0000000 0.0000000 0.0000000) *
      13) NR< 5 3 0 Coche (0.0000000 0.0000000 1.0000000 0.0000000) *
     7) NR< 3 2 0 Moto (0.0000000 0.0000000 0.0000000 1.0000000) *

> library(tree)
> (clasificationTreeV = tree(TV ~.,
+                           data = sampleV,
+                           mincut = 1,
+                           minsize = 2)
+ )

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

1) root 10 27.32 Bicicleta ( 0.3 0.2 0.3 0.2 )
  2) NR < 3 5  6.73 Bicicleta ( 0.6 0.0 0.0 0.4 )
    4) TC: A,B 2  0.00 Moto ( 0.0 0.0 0.0 1.0 ) *
    5) TC: N 3  0.00 Bicicleta ( 1.0 0.0 0.0 0.0 ) *
  3) NR > 3 5  6.73 Coche ( 0.0 0.4 0.6 0.0 )
    6) NR < 5 3  0.00 Coche ( 0.0 0.0 1.0 0.0 ) *
    7) NR > 5 2  0.00 CamiÃ³n ( 0.0 1.0 0.0 0.0 ) *

> clasificationTreeV

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

1) root 10 27.32 Bicicleta ( 0.3 0.2 0.3 0.2 )
  2) NR < 3 5  6.73 Bicicleta ( 0.6 0.0 0.0 0.4 )
    4) TC: A,B 2  0.00 Moto ( 0.0 0.0 0.0 1.0 ) *
    5) TC: N 3  0.00 Bicicleta ( 1.0 0.0 0.0 0.0 ) *
  3) NR > 3 5  6.73 Coche ( 0.0 0.4 0.6 0.0 )
    6) NR < 5 3  0.00 Coche ( 0.0 0.0 1.0 0.0 ) *
    7) NR > 5 2  0.00 CamiÃ³n ( 0.0 1.0 0.0 0.0 ) *
```

2.1

Then, the lineal regression anaylisis will be done:

```
> library(utils)
> pairs <- read.table("pair1.txt")
> dataP = data.frame(pairs)
> names(dataP)
```

```
[1] "x"       "y"       "muestra"
```

The coeficient's matrix correlation:

```
> cor(dataP)
```

```
                x             y       muestra
x       1.0000000  8.163662e-01  0.000000e+00
y       0.8163662  1.000000e+00 -5.248517e-05
muestra 0.0000000 -5.248517e-05  1.000000e+00
```

Minimun square error line:

```
> regression <- lm( x~y, dataP)
> summary(regression)
```

```
Call:
lm(formula = x ~ y, data = dataP)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9845 -1.5525 -0.2928  1.3838  4.2011

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.9991     1.1273  -0.886    0.381
y             1.3331     0.1455   9.161 1.44e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.869 on 42 degrees of freedom
Multiple R-squared:  0.6665,       Adjusted R-squared:  0.6585
F-statistic: 83.92 on 1 and 42 DF,  p-value: 1.437e-11
```

The equation's line is y = -0.9991 + 1.3331x, and the anova is:

```
> anova <- aov(x~y, dataP)
> summary(anova)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
y            1  293.2  293.24   83.92 1.44e-11 ***
Residuals   42  146.8    3.49
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3