Ciências ULisboa | Estatística e Investigação Operacional

FOUNDATIONS IN DATA SCIENCE

S3 – STATISTICS (20%)

Assignment - deliver by November 29^{th} 2024

Informations:

- The groups should have 3 or 4 elements per group.
- The resolution of the assignment will be carried out in software R.
- Create a R Markdown file in which the title of the document must contain the first and last name and student number of all elements in the group.
- All groups must send the .rmd and .pdf file with the following, First name_Last name_number, to the Professor email (eitrigueirao@fc.ul.pt).
- **The groups must choose between A, B or C.**
- Always justify your answers.

A

CSV File: airpollution.csv

In a study on pollution, carried out in 41 cities in the USA, the following variables were considered:

- so2: Sulphur dioxide content of air in micrograms per cubic meter
- temp: Average annual temperature (F)
- manuf: No. of manufacturing enterprises employing 20 or more workers
- pop: Population size (1970 census) in thousands
- wind: Average wind speed in miles per hour
- precip: Average annual precipitation in inches
- days: Average number of days with precipitation per year

Tasks:

1. Perform Principal Components Analysis of the data set.
2. Choose the principal components retained and the importance of each one.
3. Explain the importance of the variables for the explanation of each of the principal components retained.
4. Make a graphical representation of the principal components and present relevant results.
5. Perform a k-means clustering.
6. Make a graphical representation of the clusterings obtained.
7. Write a brief description of each cluster.

B

CSV File: cars.csv

NOTE: After reading the csv file in R, please consider the following code:

```
cars2 <- cars[,-1]

rownames(cars2) <- cars[,1]
```

The cars data set comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles.

The dataset includes fuel consumption and 10 aspects of automotive design and performance for 32 automobiles:

- mpg: Miles/(US) gallon
- cyl: Number of cylinders
- disp: Displacement (cu.in.)
- hp: Gross horsepower
- drat: Rear axle ratio
- wt: Weight (1000 lbs)
- qsec: 1/4 mile time
- vs: Engine (0 = V-shaped, 1 = straight)
- am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

Tasks:

1. Make a graphical representation between the distance traveled with a gallon of fuel (mpg) and the weight of a car (wt). What is the relationship between the two variables?
2. Perform a linear regression between the distance traveled with a gallon of fuel (mpg) and the weight of a car (wt).
3. Based on regression model summary obtained before, interpret the results.
4. Are the assumptions regarding the residuals violated?
5. Perform a multiple linear regression model by adding both horsepower(hp) and displacement(disp) in the linear regression model obtained in (2).
6. Based on regression model summary obtained before, interpret the results. What can you say about the quality of the adjustment of the model?
7. Make a graphical representation of the residuals. What do you conclude?

C

txt files: tcga63.txt and timestatus.txt

NOTE: To read the txt files, write the following code:

```
genes <- read.table("tcga63.txt",dec = ",", header = TRUE)

data.time.status<-read.table("./timestatus.txt",dec = ",", header = TRUE)
```

The ovarian cancer dataset is based on gene expression data of oncological patients and is constituted by 517 observations over 12042 covariates and the event time (time) and death (status). This data was obtained from The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/). Notice that the tcga63.txt only contains the gene expression for 63 genes.

Tasks:

1. Install package `glmnet` and perform a lasso regularization technique.
2. Based on the results, was any variable selected?
3. Perform the elastic net regularization. Consider $\alpha = \{0.2, 0.4, 0.5, 0.6\}$, and the number of folds equal to 5.
4. How many variables were selected in each model? Present the name of the variables selected, and conclude if there are any overlaps between the variables selected in each model.
5. Make a graphical representation (Kaplan Meier) of the models obtained.
6. Repeat the for the α values considered before, the elastic net regularization, with number of folds equal to 3.
7. Compare the results with the previously models (3).