**Data Science and Machine Learning (MSc)**

DAMA51: Foundations in Computer Science

Academic Year: 2022–2023

|  |  |
| --- | --- |
| #3 Written Assignment | |
| Submission Deadline | Wed, 8 Mar 2023, 11:59 PM |
| Student Name: | TRANTEROU VASILEIA |

# Remarks

The deadline is definitive.

An indicative solution will be posted online along with the returning of the graded assignments.

The assignment is due via the STUDY submission system. **You are expected to turn in a document (.DOC, .ODT, .PDF) and a compressed (.ZIP, .RAR) file containing all your work:**

* **1 document file (this document) with the answers to all the questions, along with the R code and the results of the execution of the code**
* **1 compressed file with 4 R scripts with the code that answers to each one of the problems to the Topics 3 and 5.**

**You should not make any changes in the written assignment file other than providing your own answers.** You should also type all of your answers into Word and not attach any handwritten notes as pictures into your work otherwise a 5% reduction of your final grade will be applied. Make sure to name all the files (ZIP file, DOC file and R script files) with **your last name first followed by a dash symbol and the names of each component at the end**. For example for the student with last name Aggelou the files should be named as follows: Aggelou-HW4.zip, Aggelou-HW4.doc, Aggelou-Topic3.R, and Aggelou-Topic5.R. The R script files should automatically run with the **source** command and generate the correct results. Also, please include comments before each command to explain the functionality of the command that follows. In the computations, use three decimal places.

|  |  |  |
| --- | --- | --- |
| Topic | Points | Grades |
| 1. **Online Quiz** | 40 |  |
| 1. **Article review** | 5 |  |
| 1. **Model Fiiting & Gradient Descent** | 20 |  |
| 1. **Confusion Matrix and**   **ROC curve** | 20 |  |
| 1. **Outlier Detection** | 20 |  |
| **TOTAL** | **105 (max 100)** | **/100** |

# Topic 1: Online QUIZ

Complete the corresponding online quiz available at:

<https://study.eap.gr/mod/quiz/view.php?id=25406>

You have one effort and unlimited time to complete the quiz, up to the submission deadline. **(40 points)**

# Topic 2: Article Review

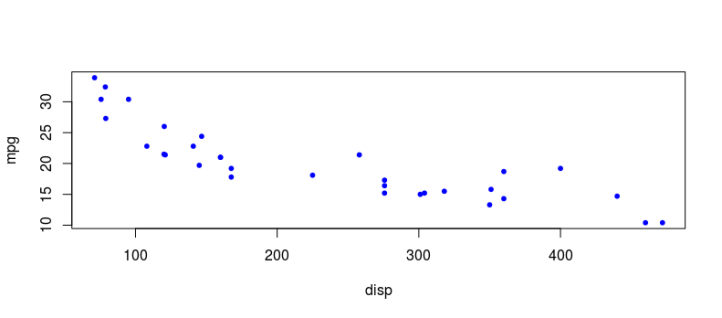
The article "Challenges in Deploying Machine Learning: A Survey of Case Studies" (<https://dl.acm.org/doi/10.1145/3533378>) describes *data poisoning* and *model stealing* as two particularly relevant threats facing organizations which rely on ML projects. Select a business domain that is most familiar or interesting to you and describe a scenario where one of the threats might be considerably more serious than the other.

Note: You should write up your answer to a maximum of 300 words. Any text in excess of 300 words will not be taken into consideration.

**(5 points)**

|  |
| --- |
|  |

# Topic 3: Model Fitting and Gradient Descent

We will use the *mtcars* data set. We want to investigate the relationship between miles per gallon (*mpg*) and engine displacement volume (*disp*) of the cars. A simple scatterplot is an excellent visual tool to assess the linearity between two variables. Below is a scatterplot of these two variables.

1. Does the plot imply that a linear relationship between disp and mpg might exist? **(2 Points)**

In my opinion, there is a linear relationship between disp and mpg and specifically, negative linear relationship because there is a tendency between disp and mpg. For example, the high values of mpg are decreasing while the low values of disp are increasing.

1. Use R to build a linear model for *mpg* (response variable) and *disp* (predictor variable). Print the corresponding slope and intercept. **(4 points)**

*slope:* -0.041

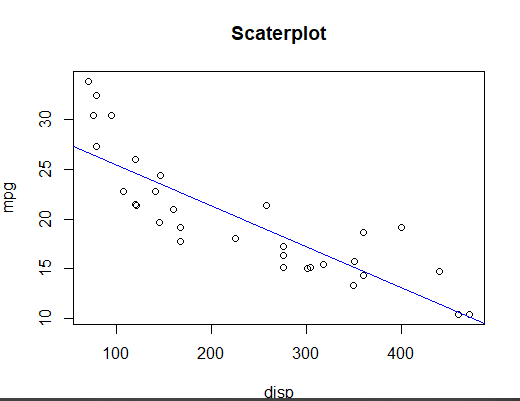
*intercept:* 29.599

*R code:* data(mtcars)

mtcars.model<-lm(mpg~disp,data=mtcars)

summary(mtcars.model)

1. Use R to draw the regression line on top of the scatter plot presented above. Then, calculate the sum of squared errors (SSE). Hint: You have to draw the scatter plot first, in order to draw the regression line. **(4 points)**

**

plot(x,y,main='Scaterplot',xlab='disp',ylab='mpg')

abline(lm(y ~ x, data = mtcars), col = "blue")

*sse<-sum((fitted(mtcars.model)-mtcars$mpg)^2)*

*SSE = 317.1587*

1. Use the first four rows of mtcars in order to calculate by hand the first and second iteration of the gradient descent algorithm. You will use *disp* and *mpg*, **min-max normalized**, as the response and predictor variable, respectively.

Hint: are the intercept, slope, and vector of predictions of *mpg*, respectively. Use learning rate equal to 0.001. **(10 points)**

1. Provide the min-max normalized values of *disp* and *mpg* in the following table. **(2 points)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *disp* | *disp normalized* | *mpg* | *mpg normalized* |
| 1. | 160 | 0.346 | 21.000 | 0 |
| 2. | 160 | 0.346 | 21.000 | 0 |
| 3. | 108 | 0 | 22.800 | 1 |
| 4. | 258 | 1 | 21.400 | 0.22 |

1. Provide the results and calculations of the first iteration in the following spaces. **(3 points)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  | *SSE* = \*\* |
| 1.  2.  3.  4. | 0.650 | 0.350 | 0.7711  0.7711  0.650  1 | 0.959 |
| \*\* | | | | |

Calculations:

1. Xscaled= (X-Xmin)/(Xmax-Xmin) Yscaled=(Y-Ymin)/(Ymax-Ymin)

X1,2= (160-108)/(258-108)=0.346 Y1,2=(21000-21000)/(22800-21000)=0

X3=(108-108)/(258-108)=0 Y3=(22800-21000)/(22800-21000)=1

X4=(258-108)/(258-108)=1 Y4=(21400-21000)/(22800-21000)=0.22

1. Yp=a+bX

Y1,2=0.650+0.350\*0.346=0.7711

Y3=0.650+0.350\*0=0.650

Y4=0.650+0.350\*1=1

SSE= ½ [+ + + ]=

=½ (0.594+0.594+0.1225+0.608)=1.918/2=0.959

1. Provide the results and calculations of the second iteration in the following spaces. **(5 points)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  | *SSE* = \*\* |
| 1.  2.  3.  4. | 0.651 | 0.351 | 0.767  0.767  0.647  0.994 | 0.949 |
| \*\* | | | | |

Calculations:

SSE/a= - [(-0.7711)+(-0.7711)+(0.350)+(-0.78)]=- (-2.672)= 2.672

SSE/b= - [(-0.7711\*0.346)+(-0.7711\*0.346)+(0.350\*0)+(-0.78\*1)]= -(-2.322)= 2.322

a\_new=a-r\*(SSE/a)=0.650-0.001\*2.672=0.651

b\_new=b-r\*(SSE/b)=0.350-0.001\*2.322=0.351

Yp=a+b\*X

Y1,2=0.647+0.347\*0.346=0.767

Y3= 0.647+0.347\*0=0.647

Y4=0.647+0.347\*1=0.994

SSE= ½ [+ + + ]=

=½ (0.588+0.588+0.124+0.599)=1.899/2=0.949

# Topic 4: Confusion Matrix & ROC curve

A diagnostic test is performed on a number of individuals that aims to identify whether these individuals have been infected or not by COVID-19. The following table quotes the results of the test and the true state of each individual. The result of each test is the probability of each tested individual to be infected. **(20 points)**

|  |  |  |
| --- | --- | --- |
|  | result | true state |
| 1. | 0.51 | N |
| 2. | 0.67 | N |
| 3. | 0.88 | I |
| 4. | 0.34 | N |
| 5. | 0.22 | N |
| 6. | 0.01 | N |
| 7. | 0.71 | I |
| 8. | 0.23 | N |

Infected (I), Non-infected (N)

1. Set the threshold to 0.50 and build the corresponding confusion matrix. An individual whose result is above or equal to the threshold is classified as Infected. **(8 points)**

|  |  |  |
| --- | --- | --- |
| Predicted \ Reference | Infected | Non-infected |
| Infected | 2 | 2 |
| Non-infected | 0 | 4 |

1. Calculate the sensitivity and specificity of the test based on the confusion matrix of the previous topic. **(4 points)**

Sensitivity : 1

Specificity : 0.67

1. Draw the ROC curve of the following table that displays the True Positive Rates (TPR) and False Positive Rates (FPR) for a COVID-19 diagnostic test. **(2 points)**

|  |  |  |
| --- | --- | --- |
| **Threshold** | **TPR** | **FPR** |
| 0.80 | 0.33 | 0.12 |
| 0.60 | 0.47 | 0.19 |
| 0.40 | 0.68 | 0.34 |
| 0.20 | 0.91 | 0.64 |
| 0.00 | 1.00 | 1.00 |

*Curve: tpr <-c(0.33, 0.47, 0.68, 0.91, 1)*

*fpr<-c(0.12, 0.19, 0.34, 0.64, 1)*

*plot(fpr,tpr, type="b", col="red")*

**

1. What is the trade-off between TPR and FPR by relaxing the threshold in the previous topic? **(2 point)**

If we lower the classification threshold, we will classify more observations as positive, increasing True Positives. But this will cause even the false positives to increase.

1. Given an AUC = 0.89, what is the probability p that the diagnostic test used, ranks a random positive case higher than a random negative case? **(2 points)**

p = 0.89

1. Assume that the total number of individuals tested is 200. Assume, also, that the number of infected and non-infected individuals is 40 and 160, respectively. What will be the number of these individuals that are expected to be classified as Infected, given a random classifier biased to the Non-infected class with probability equal to 0.60? **(2 points)**

Truly Infected individuals classified as Infected = 16

Truly Non-infected individuals classified as Infected = 64

# Topic 5: Outlier Detection

You will work on this topic using the fastfood dataset in library(openintro) which contains 515 observations on nutritional facts regarding the products of 8 different fastfood chains. You are requested to provide your R code and the result of its execution in every answer box. **(20 points)**

1. Print a summary of the dataset and indicate the minimum and maximum values of attribute “calcium” (**2 points**)

Answer:

R code:

library(openintro)

head(fastfood)

summary(fastfood$calcium)

summary(fastfood)

summary(fastfood)

restaurant item calories

Length:515 Length:515 Min. : 20.0

Class :character Class :character 1st Qu.: 330.0

Mode :character Mode :character Median : 490.0

Mean : 530.9

3rd Qu.: 690.0

Max. :2430.0

cal\_fat total\_fat sat\_fat trans\_fat

Min. : 0.0 Min. : 0.00 Min. : 0.000 Min. :0.000

1st Qu.: 120.0 1st Qu.: 14.00 1st Qu.: 4.000 1st Qu.:0.000

Median : 210.0 Median : 23.00 Median : 7.000 Median :0.000

Mean : 238.8 Mean : 26.59 Mean : 8.153 Mean :0.465

3rd Qu.: 310.0 3rd Qu.: 35.00 3rd Qu.:11.000 3rd Qu.:1.000

Max. :1270.0 Max. :141.00 Max. :47.000 Max. :8.000

cholesterol sodium total\_carb fiber

Min. : 0.00 Min. : 15 Min. : 0.00 Min. : 0.000

1st Qu.: 35.00 1st Qu.: 800 1st Qu.: 28.50 1st Qu.: 2.000

Median : 60.00 Median :1110 Median : 44.00 Median : 3.000

Mean : 72.46 Mean :1247 Mean : 45.66 Mean : 4.137

3rd Qu.: 95.00 3rd Qu.:1550 3rd Qu.: 57.00 3rd Qu.: 5.000

Max. :805.00 Max. :6080 Max. :156.00 Max. :17.000

NA's :12

sugar protein vit\_a vit\_c

Min. : 0.000 Min. : 1.00 Min. : 0.00 Min. : 0.00

1st Qu.: 3.000 1st Qu.: 16.00 1st Qu.: 4.00 1st Qu.: 4.00

Median : 6.000 Median : 24.50 Median : 10.00 Median : 10.00

Mean : 7.262 Mean : 27.89 Mean : 18.86 Mean : 20.17

3rd Qu.: 9.000 3rd Qu.: 36.00 3rd Qu.: 20.00 3rd Qu.: 30.00

Max. :87.000 Max. :186.00 Max. :180.00 Max. :400.00

NA's :1 NA's :214 NA's :210

calcium salad

Min. : 0.00 Length:515

1st Qu.: 8.00 Class :character

Median : 20.00 Mode :character

Mean : 24.85

3rd Qu.: 30.00

Max. :290.00

NA's :210

> summary(fastfood$calcium)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.00 8.00 20.00 24.85 30.00 290.00 210

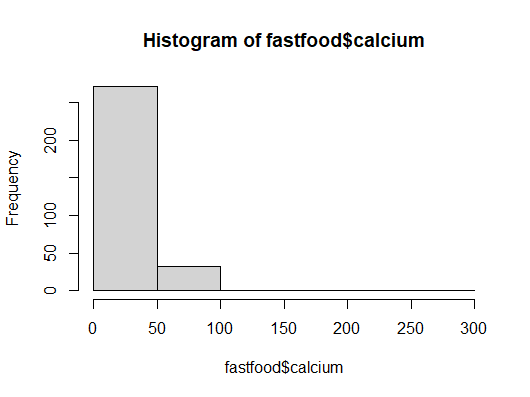
Calcium’s min and max respectively are: 0 and 290.

1. Create a histogram for the attribute “calcium”. Do you detect any potential outliers?

(**2 points**)

Answer:

R code: hist(fastfood$calcium)

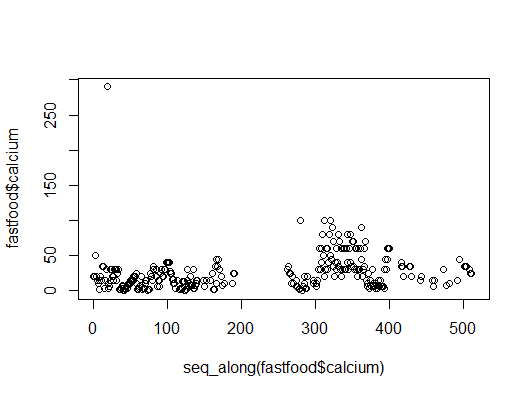


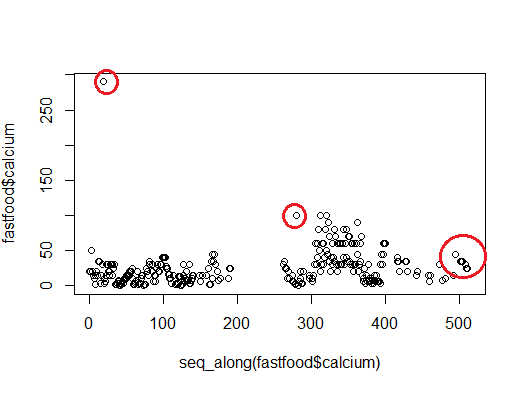
No, we can not detect any outliers.

1. Now, create a scatterplot for the attribute “calcium”. On the scatterplot, indicate any points you consider to be outliers (for example, by drawing arrowed lines pointing to the outliers, or by drawing small circles around the outliers). (**3 points**)

Answer:

R code: plot(fastfood$calcium, type="p")

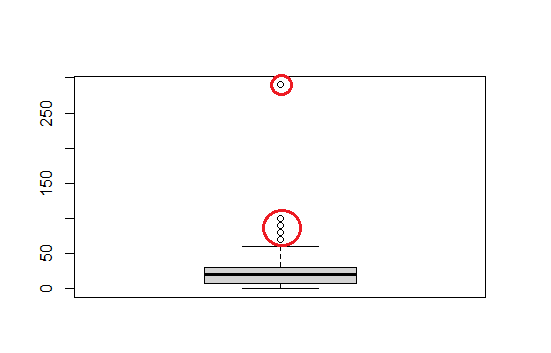




1. Finally, create a boxplot for the attribute “calcium”. On the boxplot, indicate any points you consider to be outliers (for example, by drawing arrowed lines pointing to the outliers, or by drawing small circles around the outliers). (**3 points**)

Answer:

R code: boxplot(fastfood$calcium)



1. Identify all the outliers of the attribute “calcium”, by performing a Grubbs test (use only the grubbs.test function). Hint: Start by identifying the first outlier, replace it in the data with NA and then proceed with the identification of the next one and so on. Insert your answers in the table provided. Do not forget to provide your R code as a separate file. (**10 points**)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Answer  R code:  fastfood$calcium  install.packages('outlier')  library(outliers)  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 290]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 100]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 90]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 80]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 70]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 60]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 50]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 45]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 40]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 35]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 30]<- NA  grubbs.test(fastfood$calcium)  fastfood$calcium[fastfood$calcium == 27]<- NA  grubbs.test(fastfood$calcium)   |  |  | | --- | --- | | Outlier | Index of outlier | | 290 | 19 | | 100 | 280 | | 100 | 312 | | 100 | 321 | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |