Home Exercise 3

Statistical Learning (AMI22T)

[This exercise consists of 20 points. In order to get full points, you have to answer all the questions, below. Your answer must be clearly written, well structured, and easy to follow. Your methods and models should be appropriately motivated. A cover page to your solution should make clear of your identifier, e.g. name, e-mail etc. The length of your solution should not exceed 12 pages, Times New Roman fonts with font size 12 and single line space. You may write any mathematical derivations by hand, then scan it and insert it in your report. Upload your solution on Learn by June 09, 2020; 09:00 CET. For any further questions, do not hesitate to contact Moudud Alam, e-mail:maa@du.se]

Problem 1:

This problem is about investigating some inferential properties of the linear regression model. Let us assume a linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad (1)$$

Further assume that we know the true values of the coefficients being $\beta_0 = -1$, and $\beta_1 = 2$, the predictor $x_i \sim N(0,1)$, and the irreducible error $\epsilon_i \sim N(0,1)$. You can simulate 100 observations from this model using the following R codes

```
set.seed(100)
x<-rnorm(100)
y<- -1+2*x+rnorm(100)
```

a) Estimate the model parameters using the simulated data. You can do this by running lm(y~x), in R. Now take the variables in deviation form. This can be done in R as flows:
   u<- x-mean(x)
   v<- y-mean(y)
   Now run a linear model of v and u as lm(v~u). Do you expect the slope parameters of these two regression models (y on x, and v on u) to be exactly the same? Explain why.
b) Because the intercept parameter in the regression of v on u is 0, better you run this regression by omitting the intercept term, e.g. run lm(v~u+0). Also run a reverse regression i.e. lm(u~v+0). Notice that, you get exactly the same t-statistic in the both models. Is this a coincidence or did you expect it? Explain with a mathematical proof that the equality of the t-statistic is expected, or disprove by showing a counter example.

Problem 2:

In the assignment folder on learn you will find the five_personality.txt dataset. I concerns data from more than 1.000.000 individuals that took an online personality test consisting of 50 statements. For every statement, the participants could give a numerical response based on their agreement with the statement (on a scale 1-5). The 50 statements can be found at the codebook.txt file on the same folder on learn. Based on the responses, the personality of an

individual can be described as detailed in: https://en.wikipedia.org/wiki/Big_Five_personality_traits

In this exercise, you are asked to perform clustering in the five_personality dataset and create clusters that classify people based on their responses to the quiz. Since the dataset is quite large, you should do sampling of an appropriate subset and built the clusters based on this subset. Try to validate the clusters by assigning appropriate labels and visualizing the results. Check if k-means clustering and hierarchical clustering provide the same clusters.

Finally, take **a different subset** and make **predictions** on where the observations of the second dataset are clustered based on the algorithms you have built in the first subset.