Lab 4- Regression Analysis

Create a python notebook "`Lab-4-<YOUR_NAME>.ipynb`". Write and execute your solution in the python notebook for the Exercises from 1 to 5. Use excel file for the Exercise 6. Submit both your python notebook and the excel file.

Total Marks: **20 + 2 (individual assessment) =22 marks**

Exercise-1 (4 Marks)

What linear regression equation best predicts statistics performance for the students assuming we have following data? If a student made an 80 on the test, what grade would we expect him to make in statistics? How well does the regression equation fit the data?

Student	test_score	statistics_grade
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

Note

- The above dataset is available in `student_score.csv` file

HINT

- Use `panda` library to load the csv file
- Use `stats` in `scipy` library for the regression analysis

Exercise-2 (1 Mark)

Plot linear regression line for the data given in Exercise-1 using `matplotlib` library

Exercise-3 (1 Mark)

Measure the R-squared value, goodness-of-fit for Exercise-1 linear regression model

Exercise-4 (4 Marks)

Consider the following data:

Y: [16,4,1,9,1,25,16,4,0,9,25]

X: [-4,-2,1,3,-1,-5,4,2,0,-3,5]

3.1) Visualize the scatter plot for the above data using `matplotlib` library

3.2) What type of regression model is it?

HINT

- Use `matplotlib` python library to draw a scatter plot

Exercise-5 (9 Mark)

1.In this part you will use Python to analyze the heart disease data set (the link and explanation is included here) by training and building a model with regression analysis. Test your model and discuss the result of your test with performance metrics. Make sure you separate training set and testing data properly. Then analyse the input data and explain which of them have more effects on output and modify your models by eliminating non significant variables. (5 marks)

Heart Disease Dataset: Here, is the link for heart disease dataset of patients.
http://archive.ics.uci.edu/ml/datasets/Heart+Disease
After going to this link you will find two folders: One: Data Folder and two: Dataset description. Data folder that has the dataset. It is better to use processed cleveland data. In the dataset description folder, you will find the description about the columns' names referring to the14 column of the dataset as the following: The last one attribute (number 14) is the result. Include your R source code of regression analysis, training and generating results. Here are the example of attributes and their Information (please see data set documents for more details)

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps) 5. #12 (chol)

6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
.....
13. #51 (thal)
--
14. #58 (num) ---------->result
For more information related to this assignment you can read Chapter 2 and Linear Regression section of Chapter 3 of "Doing Data Science" book.

2.Nonlinear Models

In this part you will use the heart disease data set to analyse the data with logistic regression analysis (or any other nonlinear classifier on your choice) and compare with linear regression analysis then answering which method is better. First use two models as the estimator (with

numerical result). Here you need to compare both methods by calculating Errors such as Mean Square Error (MSE) and other performance metrics (R-squared) to find which method can do prediction more accurately. Make sure you separate training set and testing data and there is no overfitting.

Exercise-6 (1 Mark)

Download the Excel file “Sample-probability-distributions-graph.xlsx” of the sample distributions from the lecture notes. By visual observation and running the regression analysis (for example by Excel regression analysis) find out which probability distribution is linear. You can examine fitting the distribution data by using linear regression model or by explaining the equation of each distribution.