CS488 Homework 2 -100 pts

Due: October, 2 (Upload soft copy on Canvas as a .docx or .pdf format)

Note: Label the homework numbers and attach all python code under appendix in your word/pdf format homework.

1. Use Iris data from sklearn datasets or download it from the UCI ML repository below: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/
   Write a python program for Iris data visualization that implements i) Correlation coefficient – display both matrix and heatmaps ii) Feature distribution analysis – display all the color coded features and iii) Plot a histogram of the individual features. (15 points)

1b. Discuss the a) implications of data distribution on data analysis b) inferences that can be drawn from i), ii) iii) - what were the data patterns/trends observed and how do they influence data analysis. (5 points)

2. For the data given below compute the linear regression parameters by hand:
   – *Problem statement:* Company M&M has a current product XYZ which is targeted towards the luxury goods market. It has observed a profit as below: (10 pts)

Months	**Profit (in K$, 1K=1000$)**
3	100
5	250
7	330
9	590
12	660
15	780
18	890

2b. What is the expected rate of profit over the next year (12 months)? What are the inferences conveyed through this predictive linear regression model? (5 pts)

2c. Company M&M wants to invest in a new product ABC if the current product XYZ has not produced a 1.5 times increase in profit over the next year. As a Data Scientist, would you advise company M&M to invest in a new product ABC or make changes to the current product XYZ? Provide your reasoning based on facts and figures to substantiate your decision-making process. (10 pts)

2d. Write a python program to implement questions 1-3. Provide code documentation and compare results obtained using linear regression function from sklearn with your own linear

regression model as discussed in class). Provide data visualization and plot the regression line for all cases. (20 pts)

3. Use Iris dataset for a Linear Regression (LR) analysis using sklearn function in python. Drop the 'petal length' feature and train the LR model on:
   i) 30% samples (i.e. train size = 0.3)
   ii) 70% samples (i.e. train size = 0.7).
   Compare the LR parameters and perform quantitative performance analysis using the root mean squared error (RMSE) measure obtained in each case. Which case was better and why? Draw your analysis based on the evaluated parameters. (20 pts)

   3b. Predict the 'petal length' for sample 50 for both case i) and ii). Compare the LR predictions with the actual value and evaluate the RMSE of each prediction. (10 pts)

   3c. Provide your analysis on which case did better and why? Draw your analysis based on the evaluated parameters. (5 pts)

APPENDIX – (Insert Code below with corresponding HW question numbers**)**