CS488/588 Homework 2 -40 pts

Due: February 25 (Upload soft copy on Canvas as a single file with .docx or .pdf format) Note: Label the homework numbers and attach all python code under appendix in your word/pdf format homework.

1a. Use Iris data from sklearn/scikit datasets or download it from the UCI ML repository

Link to iris from UCI repository: Iris - UCI Machine Learning Repository

Note you can also import it from the python library:

sklearn.datasets.load_iris — scikit-learn 1.4.0 documentation

Write a python program for Iris data visualization that implements the following:

i) Correlation coefficient – display correlation matrix as heat map (as in demo)

ii) Feature analysis/visualization (plot of individual features like in the demo) – display all the color coded features. (10 pts: 5 pts/category data visualization plots)

*Note only provide data visualizations here

1b. Present your data analysis based on the following:

a) Implications of data distribution on data analysis

b) Inferences that can be drawn from i), ii) in 1a - what were the data patterns/trends observed and how do they influence data analysis. (5 pts)

* Use data visualization to draw analysis

2.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) 20% samples (i.e. train size = 0.2)

ii) 80% samples (i.e. train size = 0.8).

Predict the ‘petal length’ for an unknown sample X (e.g. random sample index X[121] or whichever sample is not part of the training set) for both cases i) and ii).

a) Output the LR parameters (slope, intercept), LR predictions of the actual versus predicted petal length value and perform quantitative performance analysis using the root mean squared error (RMSE) measure obtained in each prediction case i) and ii). (20 pts; 10pts/case i) and ii))
b) Which case was better and why? Draw your analysis based on the evaluated parameters. (5 pts)

*Note only provide the actual petal length and predicted petal values, LR and RMSE parameters here. For choosing a random index for X, check the indices selected in the train set, and pick any index that was excluded from the train set to predict the petal length and compare with its actual value in the dataset.

APPENDIX

All code goes here