CLL788

Assignment 1

To be submitted before 3rd February 2023 23:50

1. **Data Visualization :**
    a. Plot and report scatter, histogram, heatmap, box plots for data_1.xlsx.
    b. Now perform the same for data_3.xlsx.
    c. Calculate and report the statistics for both data sets.
    d. Detect the outliers in data_3.xlsx using standard deviation approach and MAD approach.
    e. Feel free to use in-build packages for the same.

2. **LMS:** You are CEO of a clothing company with outlets in many cities. You have decided to open an outlet in a new city. To help with the decision of selecting a city, you decide to look at population vs profit data and apply linear regression to see if any relation exists between population & profit with population being the independent variable.
    a. Apply Batch LMS, Mini batch, Stochastic LMS and Least Square closed form solution and compare the results. Plot the graphs of the obtained results and training data. Use the learning rate of 0.001. Analyze the results for convergence time, accuracy etc.) (Don't use in-built packages.)
    b. Manually perform the locally weighted least linear regression using the first four data points given in excel sheet. Query point is 6.2532 and bandwidth parameter is 0.5. Perform four iterations by using stochastic LMS.
    c. Compare and report the results of Elastic net, Lasso and Ridge regression. (Use in-built packages)

Note: Data for the question one is provided in the excel file "PL1.xlsx". Negative values in the profit column mean a loss.

3. **Logistic Regression:** A university conducts 2 exams – Aptitude & Verbal as its entrance test to a 2-year program. Based on the scores of these 2 papers, admission is given to students. University has not mentioned the exact criteria of selection. Based on historical data, you need to predict whether a student will get admission based on his/her scores in the 2 exams. Data is provided in q2train.csv & q2test.csv. Train.csv contains training data. First column contains the score of Aptitude exam, 2nd column contains the score of verbal exam and 3rd column indicates whether that student got admission or not. 0 indicates not selected whereas 1 means selected. q2test.csv contains test data.
    a. Visualize the dataset to find any visible patterns. You can use any package here.

b. Apply logistic regression on training data with the first 2 columns as input data and the third column as output. You can play around with learning to find the suitable learning rate.
c. Now using the trained model, predict admission results on test data (q2test.csv) and print the result in output1.txt with every line of the text file containing either 0 or 1. Plot the results. (Don't use in built packages.)

4. **Decision Tree:** A medical consultancy service wants to predict how much their patients are likely to spend on healthcare in the coming future. You are assigned the task of building a "Regression Tree" using past patient data to predict the medical costs for new patients. Use medical_data.csv file as a dataset. To perform the modeling use the following conditions.
    a. Build the model from scratch (Don't use in-built packages)
    b. Use "charges" column as target variable
    c. Use sum-of-the-squares as a criterion to select tree root and internal nodes
    d. The minimum number of samples required to split an internal node is 5
    e. Split the dataset into training and testing datasets in the ratio of 70:30
    f. Visualize the final tree in a readable and understandable view (use built-in packages is needed)
    g. Print the final training and testing errors ((ground_truth - predicted) ** 2)
    h. Compare your model results with the sklearn's DecisionTreeRegressor

5. Visualize a decision tree that predicts whether tennis will be played on a particular day. Find out the root node of that tree. Use (play_tennis.xslx) dataset for creating the classification tree. Solve this problem manually using the ID3 algorithm.

Submission Instructions

1. Submit a zip file on Moodle named "EntryNumber.zip" with all the code files and a pdf with all the graphs and analysis. Only python (preferred) and MATLAB are allowed.
2. For any doubts in the assignment, contact Aayush Kumar Tyagi: aiz218615@scai.iitd.ac.in and Deepak Kumar: chz228174@iitd.ac.in