Assignment-3: Linear Regression

Problem Statement:

A real estate agency aims to predict the median value of homes in Boston based on various features, such as crime rate, average number of rooms, and pupil-teacher ratio, using the Boston Housing Dataset for more informed property valuation. The data is shared in the following link: 'BostonHousingDataset.csv'. Develop a code in Python to do exploratory data analysis, correlation analysis, and linear regression from the data.

Implementation: [3+2=5 Marks]

(a) Implementation of Closed form Linear Regression (**LR_ClosedForm**) from scratch (without using builtin functions or scikit-learn).

(b) Implementation of Linear Regression using Gradient Descent (**LR_Gradient**) from scratch (without using builtin functions or scikit-learn).

** Implement **[LR_ClosedForm]** and **[LR_Gradient]** from scratch. You may make use of the numpy library to perform basic operations (e.g., matrix arithmetics).

** In general, you may use libraries to process and handle data.

** Use Pandas and Numpy for the implementation and Matplotlib and/or Seaborn for visualization.

Experiments: [3+3+2+3+4=15 Marks]

Experiment 1: Load the given dataset into a pandas dataframe. Delete the columns "B" and "LSTAT". Delete all the rows with "NaN" values in any column. Convert all integer values to floating point values. This altered dataset will serve as the base dataset for the subsequent experiments, let us call this dataset as "dataset_altered". Display the first 10 rows of dataset_altered.

Experiment 2: Plot histograms of "NOX", "RM" and "AGE" for dataset_altered. Tabulate the correlation coefficients for all the columns (including the target "MEDV"). Plot the corresponding correlation matrix heatmap.

Experiment 3: Divide dataset_altered into two subsets: dataset_altered_features (having all columns except "MEDV") and dataset_altered_target (having only "MEDV" column). This

constitutes data and corresponding labels. Divide each of dataset_altered_features and dataset_altered_label into training and testing subsets (use train_test_split from sklearn.model_selection, 90% data for training, 10% data for testing, no shuffling, random state of 100). Print the shapes of both training and testing subsets for dataset_altered_features and dataset_altered_label.

Experiment 4: Design a Closed Form Linear Regression Model to predict the "MEDV" value, given the input features — **[LR_ClosedForm]**. Print the values for coefficients and intercept. Predict the values for testing data. Calculate the RMSE (root mean squared error) of predicted data with the testing data (write a method to calculate the RMSE).

Experiment 5: Design a Linear Regression Model using Gradient Descent — **[LR_Gradient]**. Study the effect of different learning rates (a) 0.001, (b) 0.01, (c) 0.1. Calculate the RMSE of predicted data with the different learning rates. Report the optimal learning rate out of the above choices, and print the corresponding optimal values for coefficients and intercept.

Dataset:

The dataset contains the following columns. Column names and descriptions are:

CRIM: per capita crime rate by town
ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS: proportion of non-retail business acres per town
CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX: nitric oxides concentration (parts per 10 million)
RM: average number of rooms per dwelling
AGE: proportion of owner-occupied units built prior to 1940
DIS: weighted distances to five Boston employment centers
RAD: index of accessibility to radial highways
TAX: full-value property-tax rate per $10,000
PTRATIO: pupil-teacher ratio by town
B: 1000(Bk - 0.63)^2 where Bk is the proportion of Black people by town
LSTAT: % lower status of the population
MEDV: Median value of owner-occupied homes in $1000's

Link:
https://drive.google.com/file/d/1Ma7ZLw1LvlNRJkvmA_JYsbrjCvgwCiHO/view?usp=sharing

Submission:

A .zip file containing the python source code and a PDF report file. The final name should follow the template: <Assign-No>_<Your Roll No>.zip. For example, if your roll no is 15CE30021, the filename for Assignment 3 will be: `Assign-3_15ce30021.zip`

1. A single python code (.py) containing the implementations of the models and experiments with comments at function level. The first two lines should contain your name and roll no.

2. A report [PDF] containing **[1+1+1+1+1 points]**
    a. Experiment 1: A table containing the first 10 rows of dataset_altered.
    b. Experiment 2: Histograms of “NOX”, “RM” and “AGE” for dataset_altered; table containing correlation coefficients; correlation matrix heatmap. State what all you can infer from the correlation matrix.
    c. Experiment 3: Print the shape of individual data matrices.
    d. Experiment 4: Values for coefficients and intercept; RMSE value of predicted data with the testing data.
    e. Experiment 5: A bar plot of RMSEs vs learning rate. State the optimal learning rate, and corresponding values for coefficients and intercept.

Responsible TAs:

Please write to the following TAs for any doubt or clarification regarding Assignment 1.

Raj Krishan Ghosh: rajkrishanghosh@gmail.com

Deadline:

The deadline for submission is **30th January (Tuesday), 11:55 PM, IST**. Irrespective of the time in your device, once submission in moodle is closed, no request for submission post-deadline will be entertained. No email submission will be considered. So, it is suggested that you start submitting the solution at least one hour before the deadline**.**

Plagiarism policy: Binary marking (two parties)