

Foundation of Data Science (Practical) - Lab 5 (Linear Regression) Assignments

You must submit assignment in report format; including related theories or background knowledge required for the lab, codes and outputs for each question and a concluding remark/ discussion highlighting what you learned. For all plots and code there must be your name and/or roll number mentioned in some form (For instance in title of each plot or at beginning of your code in comment or variable naming ...)

In the given Boston housing dataset, using multiple linear regression formulation derived in your lab session, fit a regression line to the data in order to predict the housing price (i.e. *median value of owner occupied homes* (in 1000s)). And perform following:

- I. Verify that your result is correct by using LinearRegression of sklearn library.
- II. Evaluate your model on the test set using appropriate metrics and comment on it.
- III. Choose any appropriate data point values of your liking, except these: $CRIM = 0.002 * (\text{your_roll_number})$, $NOX = 0.005 * \text{your_roll_number} + 0.35$, $DIS = 1 + 0.1 * \text{your_roll_number}$, $TAX = 200 + 3 * \text{your_roll_number}$, $RAD = (\text{your_roll_number} \bmod 2)$, and predict the median housing price for it.
- IV. Note each coefficient values and comment on what it could signify.
- V. Plot residual plot for any two independent variable of your choice and comment on the plot.

Hint:- The dataset and its associated information is in the text file. You need to extract out data part as a separate file before importing using pandas. Separate the data out into test and train set using the indices provided in 'test_indices.txt' and 'train_indices.txt' files respectively. Convert pandas object to numpy and apply our formula for getting the coefficient values.

Multiple Linear Regression Coefficients Formula:

$$\beta = (X^T X)^{-1} X^T y$$