SEIS 763 Machine Learning
Assignment 2
Due: midnight 10/4/21 on Canvas

Individual effort

You will be building regression models on the dataset provided.

Dataset: The Housing dataset contains information about houses in the suburbs of a US city in 1970s. The features of the 506 samples in the dataset are summarized here:

- `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 oxide 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 [people of African American descent] by town
- `LSTAT`: Percentage of lower status of the population
- `MEDV`: Median value of owner-occupied homes in $1000s

We will regard the house prices (`MEDV`) as our target variable—the variable that we want to predict using one or more of the 13 explanatory variables.

What you need to do: Create a jupyter notebook called **Assign2.ipynb**. Write code for each of the following questions by having a separate cell for every question. Copy the actual question in a markdown cell and right below that you should have a code cell

Q1) Load the dataset into a pandas dataframe and display the first 5 lines of the dataset along with the column headings. Note that the text file does not have the headers, which means you will have to add them to the dataframe.

Q2) Split the dataset into training (70%) and testing set (30%). Normalize the data using standardization.

Q3) **Model 1**: Using scikit learn, build a Linear Regression model with all the variables.

Q4) What are the weight parameters (including the intercept) you get for Model 1.

Q5) Use Model 1 to make a prediction on the test set. Calculate mean squared error.

Q6) **Model 2**: Build a linear regression model with all the variables using Normal Equations method.

Q7) What are the weight parameters you get for Model 2.

Q8) Use Model 2 to make a prediction on the test set. Calculate mean squared error.

Submission:

- Make sure each of the cells have been run along with the output shown right below. Now, export the notebook as .html file.
- Submit the **.html** file and **.ipynb** notebook on Canvas.

Note: Do not submit the data. Your code should be referencing data.txt when you load the data in your code.