Problem statement

Use the given dataset to predict the price of the houses using the given features.

Dataset

The Boston Housing Dataset has been taken from scikit-learn library.

The dataset contains information about different houses in Boston.

There are 13 features and 506 samples in this dataset. The objective is to predict the price (MEDV) of the housing using the given features. Of the 13 features, RM has a strong positive correlation with MEDV (0.7) and LSAT has a strong negative correlation with MEDV (-0.74). So, based on these observations, we will use the 2 features, RM and LSAT.

Next, we split the data into a training set and testing set. To split the data, we use train_test_split function provided by scikit-learn library. We train the model with 80% of the samples and test with the other 20% of the sample.

Files

1. training.csv – Contains training data, with the first row as headers, first column as serial numbers, last column as the target, and rest of the columns as features.
2. test.csv – Contains test data, with the first row as headers and first column as serial numbers, with all features other than the target in the same order as training.csv
3. sample_submission.csv – This file contains sample values for the feature to predict. The first row contains header, the first column contains serial numbers, and second column contains the target value.
4. correct_lables.csv – This file contains real values for the variable to predict and classify. The first row contains header, the first column contains serial numbers, and second column contains the actual values corresponding to test.csv.
5. source.m – Consists of the code that is used to solve the problem statement
6. notebook.ipynb – Consists of the pre-processing code to extract the database