Performance of Linear Regression on Boston Housing Dataset

Shashank Pathak(15BCE1287)*,
*School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: shashank.pathak2015@vit.ac.in*

***Abstract*—In this Experiment we use Linear Regression to get a regression line that helps us to predict the cost of houses in Boston. For this we have used the Boston Housing Dataset.Using the concept of correlation we will establish relation between different attributes that helps us to derive a regression line that will help to predict the prices of house.**

1. Introduction

Linear regression involves finding the best-fitting curve of a functional form that relates the value of an explanatory variable, X, and the mean value of a response variable, Y, given X. The goals of regression model are to determine whether Y and X are associated in some systematic way, and to estimate or predict the value of Y, or its mean, corresponding to a known value of X. Simple linear regression is linear regression in which we have one explanatory variable X and in multiple linear regression we have more than one explanatory variable. Linear Regression finds its application in various practical purposes like prediction, forecasting, error reduction.

2. Methodology

Linear regression analysis is the most widely used of all statistical techniques: it is the study of linear, additive relationships between variables. Let Y denote the dependent variable whose values you wish to predict, and let X1, ,Xk denote the independent variables from which you wish to predict it, with the value of variable Xi in period t (or in row t of the data set) denoted by Xit. Then the equation for computing the predicted value of Yt is a linear equation in X's.

This formula has the property that the prediction for Y is a straight-line function of each of the X variables, holding the others fixed, and the contributions of different X variables to the predictions are additive. The slopes of their individual straight-line relationships with Y are the constants b1, b2, , bk, the so-called coefficients of the variables. That is, bi is the change in the predicted value of Y per unit of change in Xi, other things being equal. The additional constant b0, the so-called intercept, is the prediction that the model would make if all the Xs were zero (if that is possible). The coefficients and intercept are estimated by least squares, i.e., setting them equal to the unique values that minimize the sum of squared errors within the sample of data to which the model is fitted. And the model's prediction errors are typically assumed to be independently and identically normally distributed.

3. Database - Boston Housing Dataset

Before we implement our linear regression model, we will introduce the dataset, the Housing Dataset, which contains information about houses in the suburbs of Boston collected by D. Harrison and D.L. Rubinfeld in 1978. The Housing Dataset has been made freely available and can be downloaded from the UCI machine learning repository at https://archive.ics.uci.edu/ml/datasets/Housing.The features of the 506 samples may be summarized as shown in the figure

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

Figure 1. Boston Housing Dataset Attributes

4. Algorithm

When using linear models for prediction, it turns out very conveniently that the only statistics of interest (at least for purposes of estimating coefficients to minimize squared error) are the mean and variance of each variable and the correlation coefficient between each pair of variables. The coefficient of correlation between X and Y is commonly denoted by rXY, and it measures the strength of the linear relationship between them on a relative (i.e., unitless) scale of -1 to +1. That is, it measures the extent to which a linear model can be used to predict the deviation of one variable

from its mean given knowledge of the other's deviation from its mean at the same point in time.

The correlation coefficient is most easily computed if we first standardize the variables, which means to convert them to units of standard-deviations-from-the-mean, using the population standard deviation rather than the sample standard deviation, i.e., using the statistic whose formula has n rather than n-1 in the denominator, where n is the sample size. The standardized version of X will be denoted here by X*.

Now, the correlation coefficient is equal to the average product of the standardized values of the two variables within the given sample of n observations:

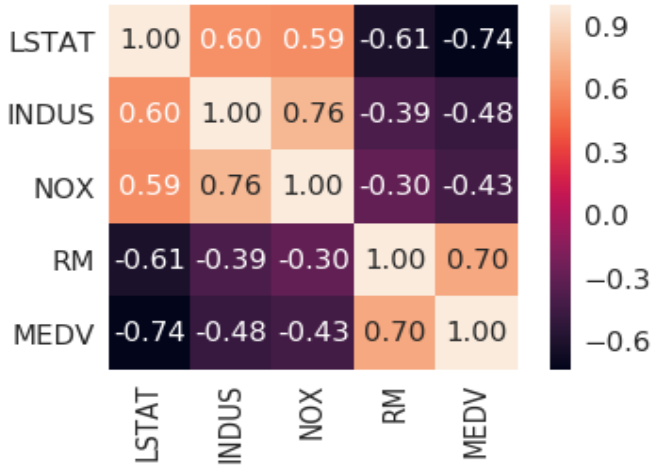

Figure 3.

5. Experiments

On this dataset we have applied linear regression.In this we first used pairplot to find the corelation between different pairs of attribute and then plotted the corelation values in the form of correlation coefficent heatmap and then based on the values of these corelation we have found out the pairs on which linear regression has to be applied in our case we have taken RM and MEDV as our chosen attribute and found out the linear regression line between these two.

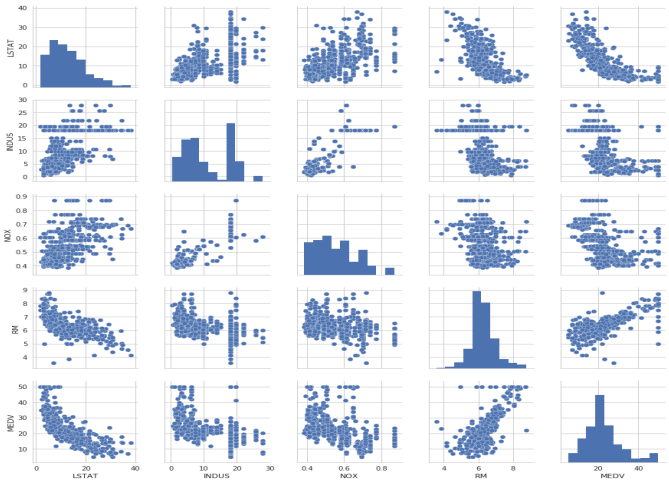

Figure 2.

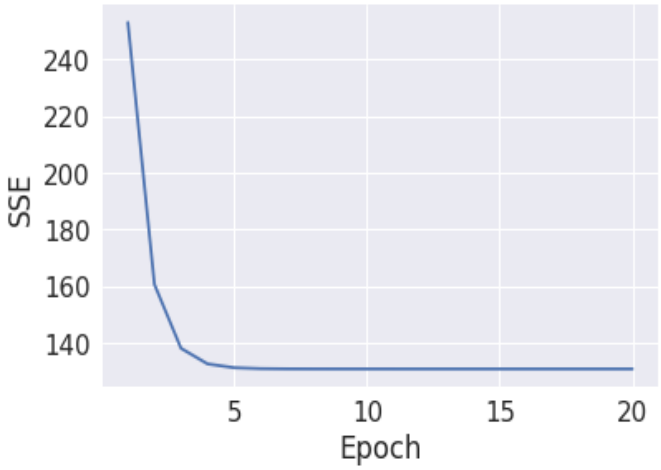

Figure 4. Train error vs Number of Epochs.

Using the linear regression model we get the slope and intercept of the regression line.There are many other algorithm like RANSAC for performing the linear regression

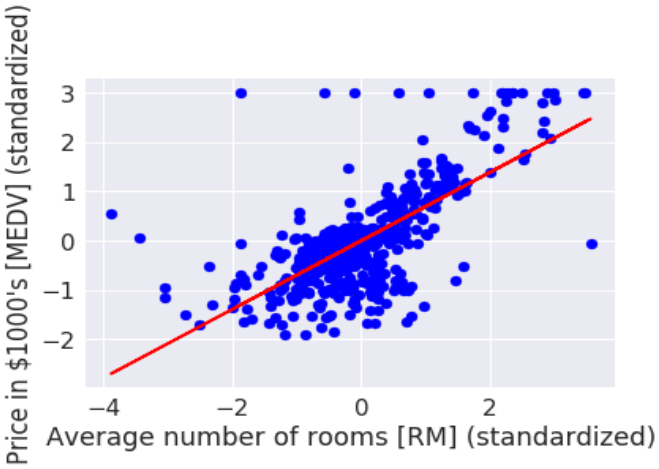

Figure 5. Two class classification performance

6. Conclusion

After the classification we can see there are many datapoint where the regression line was apt although there were datapoints which belong are very distnt from the line.This indeed suggest that the Linear regression model is a simple model which helps to get a hypothesis which is closely maps the real function but not exactly the real function.As we get a real valued line we will be able to predicts the housing prices afterwards which is the main aim of linear regression.