Chiranjeev
P21CS007
chiranjeev.1@iitj.ac.in

Assignment - 2
Linear Regression

05-09-2021

I. Dataset Description:

a. **Name of Dataset :** Boston Housing Dataset (link - Dataset link)

b. **About Dataset:** The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA.

c. **Dataset features :**

- 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 centres
- 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 blacks by town
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in $1000's

d. **Data Pre-processing :**
Found that there are not any Not Available (NA) values in the dataset.

e. **Features Selected :**
All features mentioned above are chosen for training and testing.

f. **Target Value to be Predicted :**
MEDV (Median Value of owner-occupied homes in $1000’s)

II. Splitting the Dataset:

Used train_test_split of sklearn to split the dataset into train and test.

Split the Dataset into: 80% - train set, 20% test set

III. Techniques Used:

A. Pseudo-Inverse Method :

Steps Followed:

1. Calculated **'A'** Matrix of training set which is:

$$\mathbf{A} = \sum_{i=1}^{m} \boldsymbol{X\_train}_i \ (\boldsymbol{X\_train}_i)^T$$

2. Calculated **'b' :**

$$\mathbf{b} = \sum_{i=1}^{m} Y\_train_i \ . \ \boldsymbol{X\_train_i}$$

3. Now, to find Weight vector(**W**), we will use Pseudo-inverse method, that is :

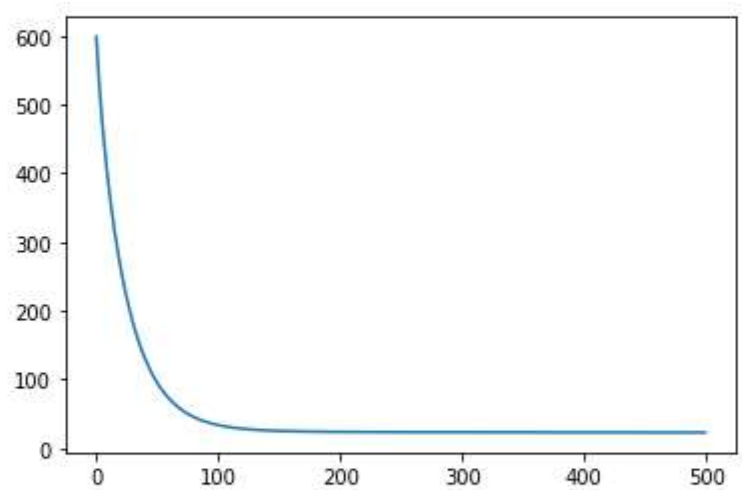
$$\mathbf{W} = \mathbf{A}^{\dagger}\,\mathbf{b}$$

Results :

Metric Used	**Train Set**	**Test Set**
Mean Squared Error (MSE)	24.9611205	22.75666
Root Mean Squared Error (RMSE)	4.99611054	4.7703947
R2 - score	0.70942	0.70934

B. <u>Gradient Descent Method</u> :

1. **Loss Minimized:** Mean Squared Error
2. **Number of Epochs:** 500
3. **Learning rate:** 0.01
4. **Graph Plot: No. of epochs (on X-axis) Vs Loss (on Y-axis):**



Results :

Metric Used	**Train Set**	**Test Set**
Mean Squared Error (MSE)	22.782801	20.657890
Root Mean Squared Error (RMSE)	4.7731333	4.5450906
R2 - score	0.73478	0.73614

IV. Appendix :

1. Mean Squared Error Loss (MSE):

$$L_s(h) = \frac{1}{m}\sum_{i=1}^{m}(h(\boldsymbol{x}_i) - y_i)^2$$

Where,

m is the number of samples

$\boldsymbol{x}_i$ is the i^{th} feature vector

y_i is the i^{th} actual target value

$h(\boldsymbol{x}_i)$ is the hypothesis function

2. Root Mean Squared Error Loss (RMSE):

$$\textbf{\textit{RMSE loss}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(\boldsymbol{x}_i) - y_i)^2}$$

Where,

m is the number of samples

$\boldsymbol{x}_i$ is the i^{th} feature vector

y_i is the i^{th} actual target value

$h(\boldsymbol{x}_i)$ is the hypothesis function

2. R2-score:

$$\mathbf{R^2} = \mathbf{1} - \frac{\boldsymbol{RSS}}{\boldsymbol{TSS}} = 1 - \frac{\sum_{i=1}^{m}(y\_true_i - y\_predict_i)^2}{\sum_{i=1}^{m}(y\_true_i - mean(y\_true))^2}$$

Where,

m is the number of samples

R^2 is coefficient of determination

RSS is sum of square residuals

TSS is total sum of squares

y_true_i is the i^{th} actual target value

y_pred_i is the i^{th} predicted target value