43304

## Assignment - 06

* **Aim →** Supervised learning - Regression

Generate a 2D data sets of N points. Split the data into training and testing

1. Perform linear regression with least square method
2. Plot the graph for training MSE and test MSE
3. Verify the effect of data - set size and bias - variance
4. Describe your findings in each case.

* **Theory →**

- **Least Square Method for Linear Regression :**

1. This method is used to find coefficient of model parameters in LR.

Given -

$(x_1, y_1)$, $(x_2, y_2)$, .... $(x_n, y_n)$

Target is to find simple linear regression model between independent variables x and dependent variable Y; as

$$Y = \beta_0 + \beta_1 X + e$$

$\beta_0$, $\beta_1$ are called parameters of linear regression. These parameters can be found using 2 methods → i) Least Square Method
ii) Maximum likelihood estimation

let regression eq^n be, $\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X$

where, $\hat{Y}$ is predicted by the regression line

- **Residuals or Errors :**

The difference btw^n the actual value of Y, given in training data and the predicted value of $\hat{Y}$ predicted by linear regression.

least square method finds values of $\beta_0$ and $\beta_1$, to result in regression line for which sum of squares of all residuals or SSE is minimum.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

$$\therefore SSE = \sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta_1} x_i)^2$$

To get minimum value of SSE for $\beta_0$ and $\beta_1$, partial derivative of SSE w.r.t $\beta_0$ and $\beta_1$ must be equal to 0

$$\frac{\partial SSE}{\partial \beta_0} = 0 \qquad \frac{\partial SSE}{\partial \beta_1} = 0$$

$$\rightarrow \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \left(y_i - \hat{\beta_0} - \hat{\beta_1} x_i\right)^2 = 0$$

$$\therefore \hat{\beta_1} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- **Effect of Data size on Linear Regression :**

There are 2 common cases that could be observed. Training dataset is relatively unrepresentable. It means that the training dataset does not provide sufficient information to learn the problem, relative to the validation set used to evaluate it. This may occur if the training dataset has too few examples as compared to validation dataset.

Validation dataset is relatively unrepresentative. It means that the validation dataset does not provide enough information to evaluate the ability of the model, to generalize.

This may occur if the validation dataset has too few examples as compared to the training dataset.

- **K-Fold Cross Validation :**

It is one way to improve over the holdout ~~data~~ method. The dataset is divided into K subsets, and the holdout method is repeated K-times. Each time, one of the (K-1) subsets is used as training data and the $K^{th}$ subset is used as the test data. The average error across all the K-trials is computed

The advantage of this method is that no matter how the data gets divided, each data point gets to be in the test dataset exactly once, and gets to be in the training set K-1 times. The variance of the resulting estimate is reduced as K is increased.

Disadvantages of this method is that the training algorithm has to rerun from scratch K times, which means that it takes K times as much computation to make an evaluation.

* Conclusion → Thus in this assignment, we implemented regression on the real estate dataset, using R and applied concepts.