Assignment 6

Supervised learning Regression

- Aim: Generate a proper 2D dataset of N points, split the dataset into training Datasets and test Data set

i) Perform Linear Regression analysis with least square method.

ii) Plot graph for training MSE and test MSE and comment on curve fitting generalization error.

iii) Verify the effect of Data set Size and Bias Variance Trade off

iv) Apply cross validation and plot the graphs for error.

v) Apply subset selection method and plot the graph for errors.

vi) Describe your finding in each case

- Objective : To study

i) To study concept of supervised learning - Regression.

ii) How to download different input dataset.

iii) Linear regression analysis with least square methods

iv) adding and Removing Packages

v) Using MSE to evaluate model performance.

vi) concept of curve fitting and generalization errors.

Page :
Date : / /

- Theory:

Least square Method for linear Regression

This method is used to find coefficient of model parameter in ~~Non~~ linear Regression.

$(x_1, y_1)\ (x_2, y_2)\ \cdots\ (x_n, y_n)$

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

$$Y = B_0 + B_1 x$$

B_0 and B_1 are called the parameters of linear Regression.

These parameter can be found by using least square method.

* let the regression equation be

$$\hat{y} = B_0 + B_1 X$$

where $\hat{y}$ is predicted by regression line.

- Residuals or Errors:
The difference between the actual value of y ques in training data and predicted value of y i.e y predicted by linear regression.

Least square method finds values of B_0 & B_1 to result in Regression line for which sum of square of all residuals or (SSE) Errors is minimum

Page :
Date : / /

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

$$= \sum_{i=1}^{n} (y_i - \hat{B}_0 - \hat{B}_1 X_i)^2$$

To get the minimum value of SSE for B_0 and B_1 partial derivative of SSE unit B_0 and B_1 must be equal to 0.

$$\frac{\partial SSE}{\partial B_0} = 0$$

$$\frac{\partial SSE}{\partial B_0} \therefore$$

$$= \frac{\partial}{\partial B_0} = \sum_{i=1}^{n} (y_i - \hat{B}_0 - \hat{B}_1 X_i)^2 = 0$$

$$\therefore B_0 = \bar{y} - \hat{B}_1 \bar{x}$$

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

Effect of Data Size on Linear Regression There two complex cases that could be observed

- The
- Training dataset is relatively unrepresentative 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 datasets.

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

- K-fold cross validation.

It is one ways to improve over the holdout method. ⇒ The dataset is divided into K-subsets and the holdout method is repeated K times.

Each time one of the K-subsets is used as the test set out the other K-1 subset are put together to form a training set. Then the average error across all the K trial is computed.

Conclusion

In this assingment, we implemented Regression on the real estate dataset using R and applied concept like cross validation subset selection method.