Date / /
Page No.

ASSIGNMENT 2

Aim: Generate a proper 2D dataset of N points. Split the data set into Training Data set and Test Data set.

i) Perform linear regression analysis with least Square Method

ii) Plot the graph for training MSE and Test MSE and comment on Curve Fitting and Generalization Error

iii) Verify the Effect of Data Set Size and Bias - Variance Trade off

iv) Apply Cross validation and plot the graph for errors

v) Apply subset selection and plot graph for errors

vi) Describe your finding

Theory:

When we have single input attribute (x) and we want to use linear regression, this is called simple linear regression.

If we had multiple input attributes, this would be called multiple linear regression, The procedure for linear regression is different and simpler than that for multiple linear regression, so it is a good place to start.

Date / /
Page No

With simple linear regression we want to model our data as follows:

$$y = B0 + B1 * x$$

This is a line where y is the output variable we want to predict, x is the input variable we know and B0 and B1 are coefficients that we need to estimate and move the line around.

Technically, B0 is called the intercept because it determines where the line intercepts the y-axis. In machine learning we can call this the bias, because it is added to offset all predictions that we make. The B1 term is called the slope because it affects the slope of the line or how x translates into a y value before we add our bias.

The goal is to find the best estimates for the coefficients to minimise the errors in predicting y from x.

Simple regression is great because rather than having to search for values by trial and error or calculate them analytically using more advanced linear algebra, we can estimate the directly from our data.

We can start off by estimating the value for B1 as:

$$B1 = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Where mean() is the average value for the variable in our dataset. The x_i and y_i refer to the fact that we need to repeat these calculations across all values in our dataset and i refers to i^{th} value of x or y.

$$B0 = \bar{y} - (B1 * \bar{y})$$

We can calculate an error for our predictions called the Root Mean Squared Error or RMSE

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Where sqrt() is the square root function, p is the predicted value and y is the actual value, i is the index for a specific instance, n is the number of predictions, because we must calculate the error across all predicted values.

Date / /
Page No:

Conclusion: The following objectives with respect to linear regression were implemented in this assignment sucessfully.