UTA ID: 1001557968 Name: Darshil Parikh

Project 1

Linear Regression: Linear Regression is a linear approach for modelling the relationship between a scalar dependent variable Y and one or more explanatory variables for independent variables denoted by X.

Training set and Test set: Given data set is divided into training and test sets, where training set is used to build a model and test set is used to validate it. Moreover, the test set is used to determine the accuracy of the model.

N - fold cross validation: Cross Validation is a technique for assessing how the result of statistical analysis will generalize with an independent set.

Data: Iris data set.

Project:

- Train a model via linear regression.
- Using the trained model to do the classification.
- Needs N-fold cross validation.

Method:

- Data was bifurcated into training and testing data using the n-fold cross validation procedure.
- Beta-cap which i.e. weights value was calculated by applying matrix operations on the trained data using the following formula: $\widehat{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}$
- Predicted value was calculated using:

$$f(\mathbf{x}) = w_0 x_0 + w_1 x_1 + w_2 x_2 + \ldots w_d x_d = \mathbf{w}^T \mathbf{x}$$

- Error was calculated using:

$$Error\ (\mathbf{w}) = \frac{1}{n} \sum_{i=1,..\,n} (y_i - f(\mathbf{x}_i, \mathbf{w}))^2$$

- Average Cross-Validation error was found using:

$$(\text{CV Error})^{(\lambda)} = K^{-1} \sum_{k=1}^{K} (\text{CV Error})_k^{(\lambda)}$$

Result and Reasoning:

- I selected various values of 'N' for N – fold cross validation ranging from 5 – 20 with step-size of 5.
- The average error rate was least when the N was chosen to be 15.
- The average error rate increased again when the value of 'N' was chosen as 20 (accuracy: 94.43%).
- As the average error rate for 15 – fold cross validation was minimum i.e. the accuracy was maximum, I chose the value 15.
- The average error rate for 15 folds is: 0.05502156981457746
- The accuracy is: 94.50%