Report Homework 5

Akshay Adlakha
akshayad

Problem 1: In this problem, we are given a spam dataset. After loading data, did a sanity check that the data has NA values or not. The last column in the dataset is the spam which is to be used to train neural network.

The commonly used words in the mail has a numerical value which is used to assign weights to the neural network.

To control training models, used trainingContorl method. The cross validation was executed for 5 folds and the weights for 100 iterations. Then, the confusion matrix to see the results.

```
Confusion Matrix and Statistics

          Reference
Prediction email spam
     email   539   27
     spam     41  314

               Accuracy : 0.9262
                 95% CI : (0.9073, 0.9422)
    No Information Rate : 0.6298
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.843

 Mcnemar's Test P-Value : 0.1149

            Sensitivity : 0.9293
            Specificity : 0.9208
         Pos Pred Value : 0.9523
         Neg Pred Value : 0.8845
             Prevalence : 0.6298
         Detection Rate : 0.5852
   Detection Prevalence : 0.6145
      Balanced Accuracy : 0.9251

       'Positive' Class : email
```

This result says that the cross validation for neural network gives us a 92% accuracy for recognizing the spam emails from the probability of keywords given in the dataset.

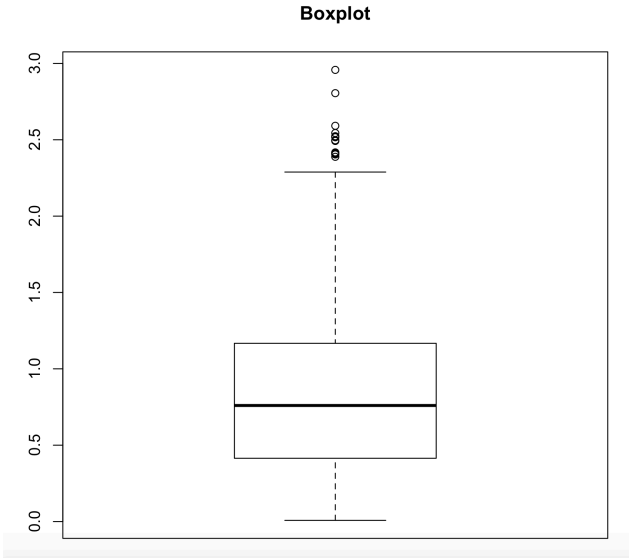
The p value is low which nullifies the null value hypothesis.

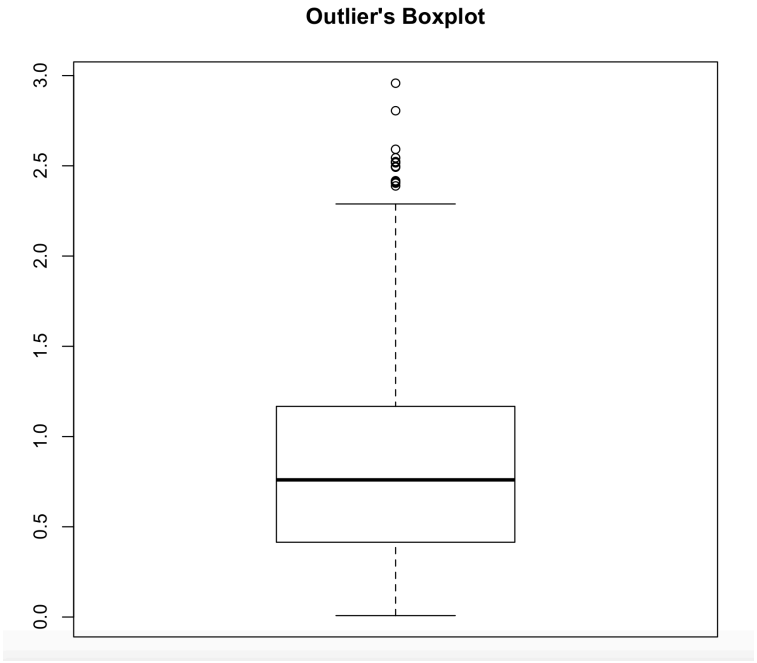
```
          Reference
Prediction email spam
     email   539   27
     spam     41  314
```

The neural network identified the 539 emails and 314 spams correctly out of all the observations.

Problem 2: In this problem, chosen the dataset of Gender Classification. After loading a dataset, did a sanity check that the data has NA values or not. Then, divided data into training and testing sets.

Then had an outlier in the training dataset. After that, plotted the box-plot with and without outlier. And, here are the results:





From the above two plots, we can say that the value we changed in the data is an outlier.

Then, I run the neural model on the original dataset and the dataset with outliers.

The accuracy when a neural model is used on the original dataset is mentioned below:

```
0.9621451 0.9731861 0.9652997 0.9652997 0.9652997 0.9747634 0.9684543 0.9668770
0.9668770 0.9652997 0.9684543 0.9700315 0.9716088 0.9747634 0.9794953 0.9747634
0.9763407 0.9763407 0.9763407 0.9716088
```

There is no much difference in accuracies.

Then, used a model on the data with outliers gives the following result when the number of neurons in the hidden layer is varied from 1 to 20.

```
0.9668770 0.9652997 0.9716088 0.9747634 0.9684543 0.9700315 0.9684543 0.9652997
0.9731861 0.9652997 0.9747634 0.9668770 0.9605678 0.9731861 0.9605678 0.9621451
0.9637224 0.9621451 0.9794953 0.9637224
```

From this result also, we can say that there is no much difference in accuracy.

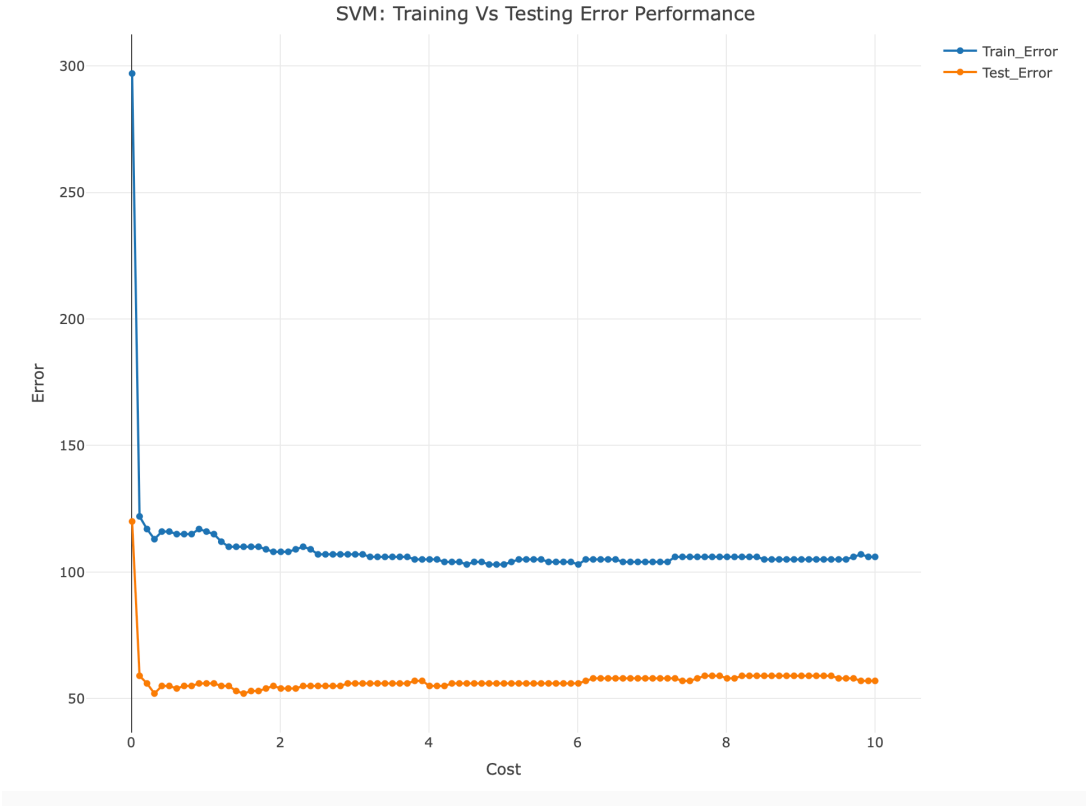
So, we can probably infer from the results that why there is no change in the accuracy in the second case when compared with the first because of the number of data points we have in the dataset.

One outlier we introduced in the dataset might not be causing too much variation to the weights that a neural network model learns.

Problem 3: In this problem, we are given a dataset of OJ from ISLR package and interested in the prediction of Purchase.

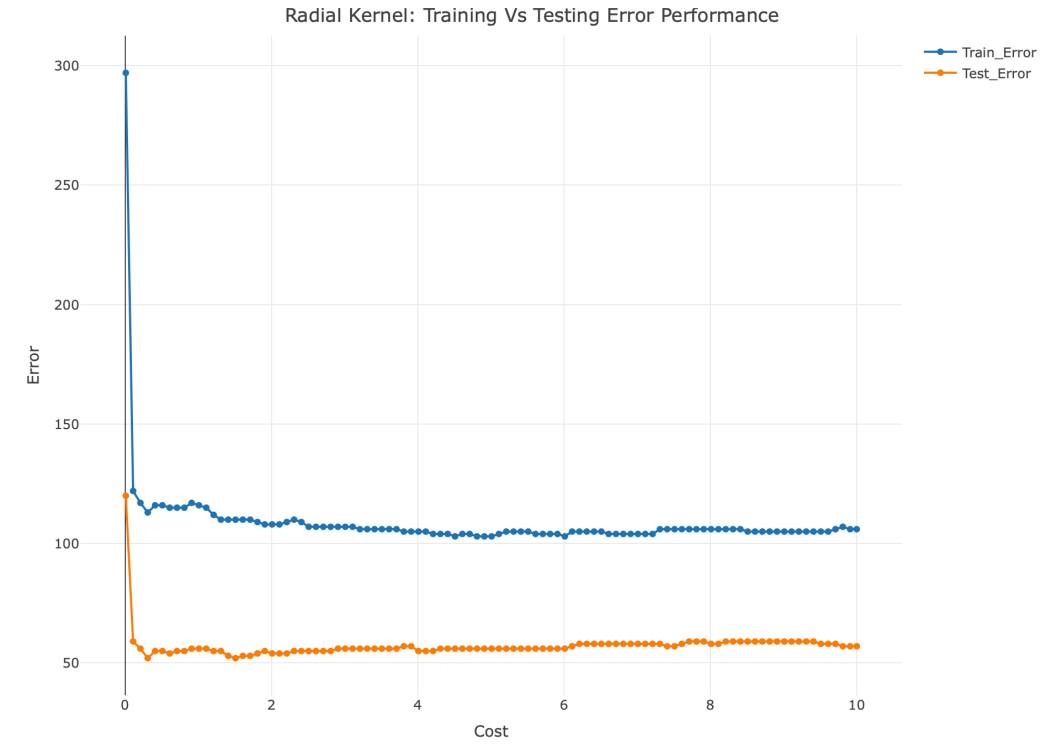
After loading data, did a sanity check over data that it has NA values or not. There are some features in the dataset which are to be categorial but actually are numeric. So, changed them to categorial values. Then, divided data into training and testing set.

Then, used the hyper parameter and iterated over cost from 0.01 to 10 and recorded the error. Here is the graph for the error.



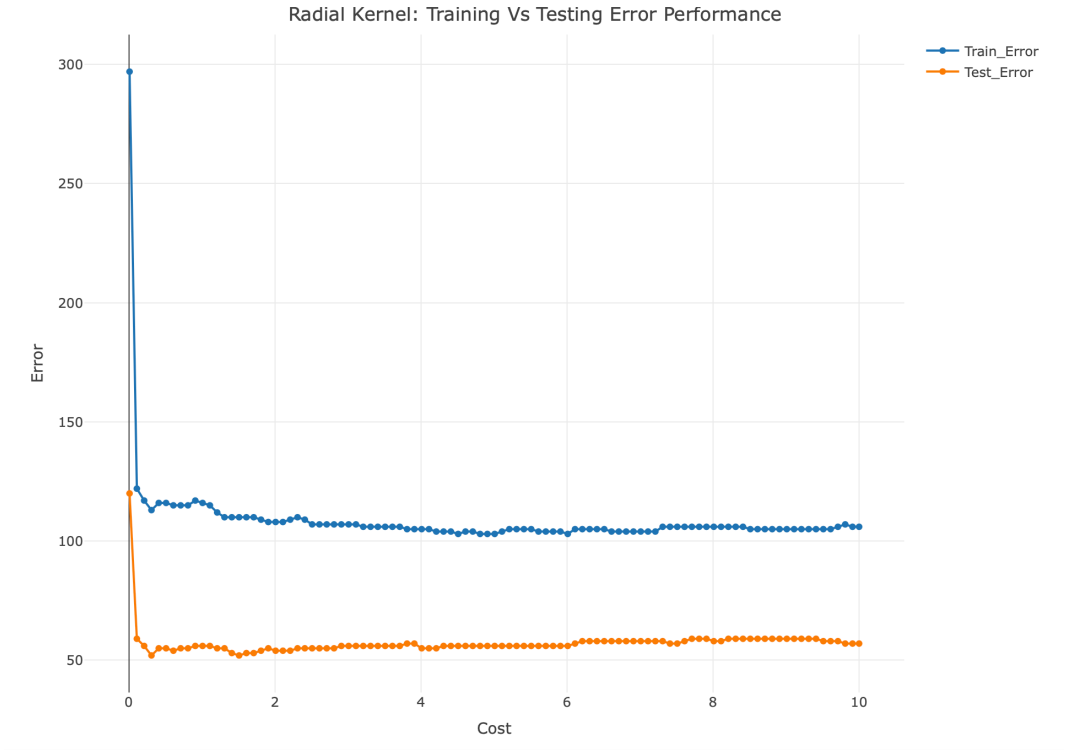
From this graph, we can say that the error for cost greater than around 2 are worst. So the cost around 0.1 to 1 are doing good.

Then, Used the radial kernel and iterated the hyper parameter for cost 0.01 to 10 and calculated the errors. Here is the graph.



From this graph, we can see that it is not performing good for the cost value greater than 2.

Then used the polynomial kernel with degree 2 for cost 0.01 to 10, and calculated the errors and plotting it gave the following result.



We can see that it is the same kind of scenario. Here the error seems that for cost greater than 3 are worst.

Unlike these graphs we can say that model seem to be performing good with the increase in cost hyper parameter. Then, it reaches a state after a while and no change can be seen.