COMP6245 Foundations of Machine Learning – Lab 2

Supritha Konaje
Email: ssk1n21@soton.ac.uk
Student ID: 32864477

1. Normalizing each Feature

Mean = 0: Mean is calculated for each feature. In every iteration it is subtracted by each of the element. Standard Deviation = 1: Standard Deviation is calculated. Similarly, like mean, SD is calculated by subtracted by each of the element in dataset.

2. Width Parameter of Basis Function

The average of the pairwise distance is comparatively better than choosing any two random points. The reason might be because the points might not be linear in this given space.

3. K-Means Clustering

Using the *KMeans* package from *sklearn.cluster*, we train our model using *kmeans.fit* which takes training dataset as it's parameter. *kmeans.cluster_centers_* gives the center of the cluster. The centers of the basic function were set to the center of the centers of the K means cluster center.

4. Training and Testing Sets

In Figure 1, we have the scatter plot for training and test set. To achieve this plot, all the previous steps were followed.

Figure 1: Training and Test Set when M = 200

5. Overtraining

If a model is overtrained, the accuracy of the model is increased but sometimes it might happen that it might generalize the data that are not important (noise) for making any prediction. A model is said to overfit if the test set fits closely to the training set. It basically learns from the noise of the training dataset rather than the actual relationship of the features and the targets.

In Figure 2, we can see that when we change the basis function M, it changes the reality of the data.

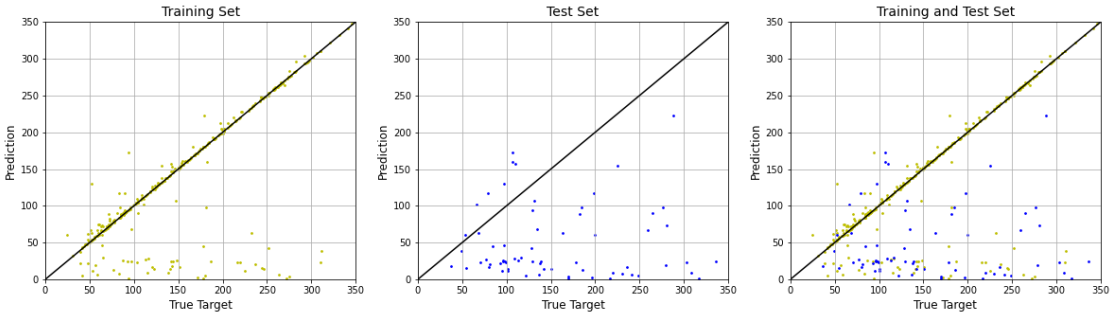

Figure 2: Training and Test set when M=300

6. 10-Fold Cross Validation

K-Fold cross validation improves the performance of a model. Unlike, training the model only once, in K-Fold, we can train the model K-1 number of time and the K^{th} set is used to test. We must perform 10-fold cross validation. So, the 9-sub data set will be our training model and the 10^{th} data set will be our test set in which we will test our model's prediction.

To achieve 10 fold, we are using the *KFold* from the package *sklearn.model_selection.* The mean result of the 10 folds is less noisy and has better accuracy. Also, this is method to avoid overtraining a model.

In Figure 3, we can see the accuracy of the RBF Model to be 54.909 and after 10-Fold Cross Validation, the accuracy score is almost 68.5685 which is far better than the previous accuracy.

```
print('Accuracy Score : %.3f' % (score*100))

Model Accuracy Score : 54.909

In [23]: scores = cross_val_score(model, X, y, cv=10)
         print ('Cross-validated scores:', scores*100)

Cross-validated scores: [55.61441118 23.05609191 35.35777697 62.19049778 26.58760166 61.81933759
 41.81591608 43.51523197 43.43698265 68.56851417]
```

Figure 3: Accuracy for the RBF model without and with KFold cross validation

7. Boxplots of test results for RBF and Linear Regression Models

In Figure 4 and Figure 5, the Boxplot for Linear Regression Model and RBF are plotted.

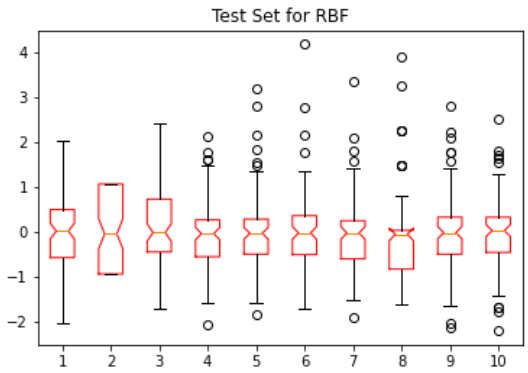

Figure 3: Boxplot for RBF

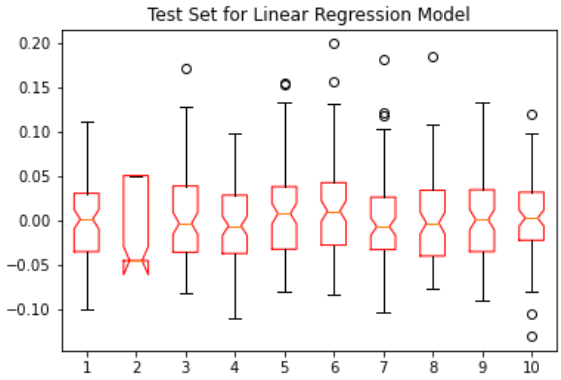

Figure 4: Boxplot for Linear Regression Model