Introduction to Data Science - Assignment 2

Shahriyar Mahdi Robbani

February 23, 2020

Exercise 1

```
Accuracy on Training Data: 1.0
Accuracy on Testing Data: 0.945993031358885
```

As expected, the classifier has perfect accuracy (100%) when used on the training data since it uses the training data to make a prediction. The classifier has an accuracy of 94.5% when used on the testing data. This is quite accurate classification and shows the classifier can be used on unknown data to obtain reasonable prediction.

Exercise 2

```
Loss for 1 neighbors: 0.046
Loss for 3 neighbors: 0.037000000000000005
Loss for 5 neighbors: 0.044
Loss for 7 neighbors: 0.05
Loss for 9 neighbors: 0.05499999999999999
Loss for 11 neighbors: 0.055999999999999994
Best k: 3
```

The provided cross-validation code was used to split the training data into five subsets. Four of these subsets were used to train the data while the fifth subset was used to test the data. The loss for each subset was recorded and the average for all five subsets is displayed. The loss is smallest when k is 3.

Exercise 3

```
Accuracy on Training Data: 0.971
Accuracy on Testing Data: 0.9494773519163763
```

Exercise 4

In order to normalize the data, we have to calculate the scalar transformation from the training data and apply it to both the training and the test data. Version 1 does this therefore it is the correct version. Version 2 is incorrect because it uses two different scalar transformations for the training and the test data. The test data will not be normalized correctly if different scalars are used. Version 3 is incorrect because it combines the training and test data into one large dataset and then calculates the scalar. This is incorrect because using the test data to calculate the scalar transform of the training data introduces bias so the wrong transformation is obtained and applied to both datasets.

```
Loss for 1 neighbors: 0.041
Loss for 3 neighbors: 0.036000000000000004
Loss for 5 neighbors: 0.044
Loss for 7 neighbors: 0.047
Loss for 9 neighbors: 0.048
Loss for 11 neighbors: 0.051000000000000004
Best k: 3
Accuracy on Training Data: 0.972
Accuracy on Testing Data: 0.9599303135888502
```

The best value of k is once again 3, but after normalization is done, the loss is reduced for all values of k and the accuracy of the best k value is increased on the test data and increased very slightly on the training data. This is because normalization reduces the effect of outliers and extreme values when we calculate the squared distances, and this improves our predictions.