Matea Marinkovic
February 28th, 2022
Introduction to Data Science
Assignment 2

Assignment 2 – Classification with Nearest Neighbors

Exercise 1: Nearest Neighbor Classification

Test Accuracy: 0.9775
Training Accuracy: 1.0

In exercise 1, we are first creating our model and training them with the original training points. Therefore, the training accuracy of the 1-NN test is a perfect 1.0, as the model is reflecting the accuracy of the data that we trained it with. Test data point 1 displays the model seeking to label the data point based on the training data points (the y data set), and it is shown that in 97.75% of instances, the label will be accurate, and in 2.25% of times, the nearest neighbor will skew the accuracy to result in an incorrect label.

Exercise 2: Cross-Validation

Found Parameter Kbest: 3

For this exercise, I start by creating indices for cross validation by splitting the data over 5 folds. Then created an array of the K values to be tested, and a loop to test the different integers in the k array [1,3,5,7,9,11]. I then looped the accuracy test through the k array over all 5 cross folds to determine the accuracies of all of the integers, appending the results to a list and averaging the 5 cross fold accuracies for each k value. Then I used a mask function to determine the max accuracy with the lowest k value, which in this case was 3.

Exercise 3: Evaluation of Classification Performance

Test Accuracy: 0.9875
Train Accuracy: 0.9933333333333333

Exercise 4: Data Normalization

For this exercise, I chose to use the first normalization variant because it is the only variant out of the three that only normalizes the xTrain data. The reason this is important because if we also normalized the xTest data, then we would introduce bias into the model. If the model becomes biased, then we would have a biased estimate of the generalization performance of the model.

New Kbest after normalization: 11
Test Accuracy of new Kbest: 0.98
Train Accuracy of new Kbest: 0.9916666666666667

It is seen in this exercise that the normalization does influence the kbest value. After the normalization of the data, we have a new kbest value as the normalization makes so all the features are centered, which would remove any data features that are larger and could dominate the function, blocking the model from being able to learn accurately from all the

points. The value of the train data here cannot be one, as there are multiple neighbors (k value) being assessed instead of just one. The distance from these points causes “noise” which lowers the accuracy of the labeling of any one test point.