MACHINE LEARNING ASSIGNMENT 1

QUESTION 1:

When the value of k is 1 or 3, the ***training score is much higher than the test score*** which shows our model might have been ***overfitted***. This is because when the k value is low, the model pays more attention to noise, thus failing to generalize picking up the closest data points. This implies that our model has ***Low bias and high variance***. Further when we increase the K values, the model performs well, (i.e) able to generalize well. This means our model has ***high bias and low variance***.

When the k value is 30, the test score is higher showing a better performance. But continuous increase in the k values, drops the performance (eg. 50 and 100). Furthermore, the predictions also seem to look more simple as K value increases. The first plot when the K value is 1 has a complicated looking prediction.

QUESTION 2:

Our best model found in the previous exercise is when k = 30. In this model, the test and the train error almost converge. It is in this model we have the lowest error rate. When we train the same distance(k = 30) with Manhattan, the model that we get is close to Normal classification. The classification boundary is also clearly defined and the two classes are well separated. The training error has decreased to a minor extent.

Our Model with Euclidean distance seems to perform the same as Manhattan as we have a low dimensional dataset. Manhattan based distance models might have significant improvement in case our dataset has high dimensionality.

The Bayes Error is the lower limit of the error that you can get with any classifier. We cannot plot this line since we only use KNN, not any other so we can't compare.

QUESTION 3:

With low values of k the training error rate is low, which increases as the k value increases. The test error starts decreasing as the k value increases.

In the region where ***test error was decreasing, the training error seems to be significantly lower***, possibly due to the effects of ***overfitting*** in that region. As the K increases, the test and training error rate start to converge, yo reflect the reduction in overfitting. As K gets large, both the training and test error rates eventually start increasing again, due to the ***underfitting***.

Initially training error is very low with higher test error which shows ***overfitting***. This is the region with ***low bias, high variance and high capacity***. After k = 50 the training error decreases and test error increases, which also could be the effects of overfitting. As k increases both the test and the training error starts to increase due to the effect of underfitting.

QUESTION 4:

We scaled the data using MinMaxScalar and then trained the model again. This improved our test and training accuracy a little bit.