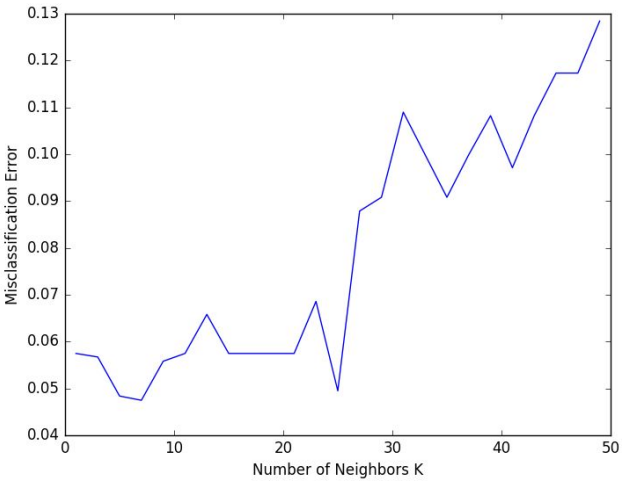
APPLIED MACHINE - LEARNING
ASSIGNMENT - 1

Saurav
ES16BTECH11007

Question - 1a)
The training error in kNN is achieved when the training data itself is used a testing data;

When K = 1, only one nearest neighbour is seen and the group of that neighbour is allotted to each point.

Now as the value of 'K' increases the neighbourhood changes. The majority class in the neighbourhood depends on the data. But it will not be zero, it (training error) will increase but the extent depends on the training data.



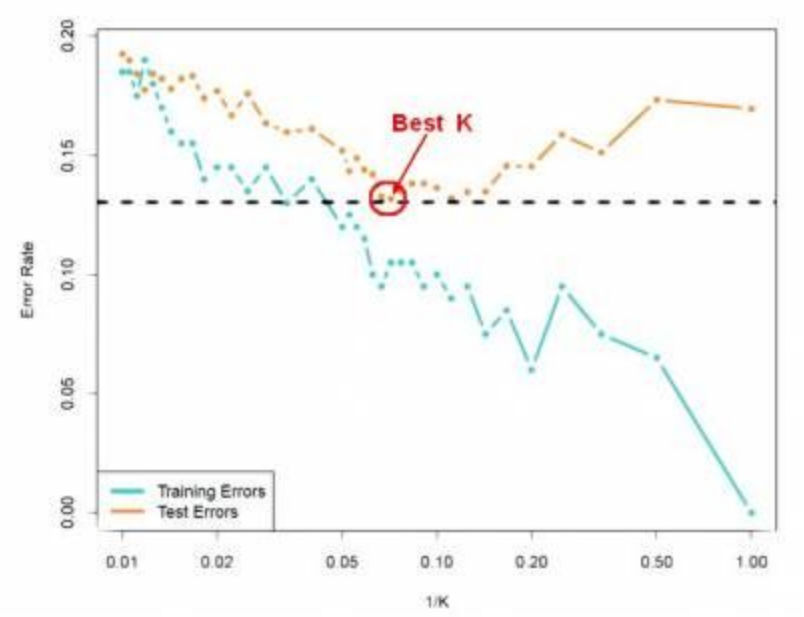
Now as the value approaches 'N' (total number of training points), the prediction becomes skewed and it misclassifies any new data point that actually belongs to a class that is not majority in number, and the prediction accuracy decreases and the training error will be high depending on how the training data is distributed.

The training error gradually increases from **zero** but actually depends on the training data on how high the error can reach.

In the Graph
The above graph is plotted with a data of three classes ('0', '1', '2') with dataset being N = 1000 data-points and the value of 'K' is varied from K=1 to K=10

Question 1b)

Assuming the data is not skewed to any of the classes, the generalization error in the real world will be lot higher than the training data as we are going to see a large number of data combinations which we may not have seen -. It may initially decrease for some range of 'K' where it approaches it minimum error rate. And then it will gradually increases as value of 'K' increases.



a) When 'K' is small

Initially the generalization error will be high (when 'K' is near to 1) as the new data will be easily influenced by noise and considering all possible cases and combinations of datasets, there is high possibility of misclassification.

b) When 'K' is near its optimal value

Then the generalization error will decrease as 'K' approaches its optimal value as it takes into account considerable number of neighbours into account to correctly predict the class of the new data point and also it doesn't consider too many points, so it prediction error(Generalization error) will be at its minimum

c) as 'K' increases from its optimal value

The generalization error will keep on increasing as the value of 'K' increases from its optimal value. This is because we are taking more number of points than we need to into consideration while predicting new data points and the extra data points may be misleading the prediction. Considering all possible cases and combination of real world data, the generalization error increases as 'K' increases, the increase may not be linear but there will be little ups and downs in the graph as 'K' increases but the generalization seen as a whole, it increases.

In the Graph

The above graph is for classification of data into two classes ('0','1')
as ((1/K) increases -> K decreases)) the data-set is as large as 50 data points.

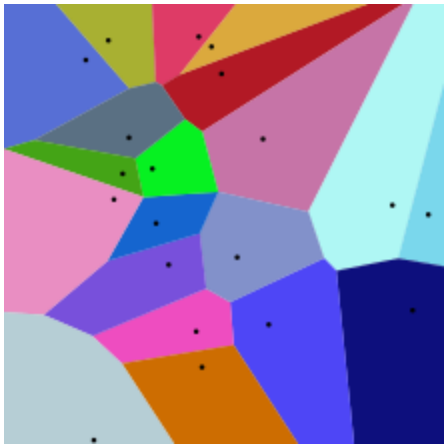
Question - 1c)

kNN suffers from the curse of High Dimensionality and is undesirable for data with high-dimensions because -

a) The distance metric becomes meaningless as the dimensions increases and the traditional euclidean distance won't work and we have to rely on other distance metrics.
b) All the points lie on one unit hypersphere and its difficult to classify
c) Increases computational complexity.

Question - 1d)

I think KNN with one nearest neighbour is not possible to make using decision trees.



We can understand the KNN with one nearest neighbours from the **voronoi diagram** where boundaries are drawn equidistant of two neighbouring points.

Thing is we cannot create a closed boundary between neighbouring points by just using constant threshold points no matter how multivariate the node will be.

So, i think it's difficult to make a decision tree that mimics one nearest neighbour