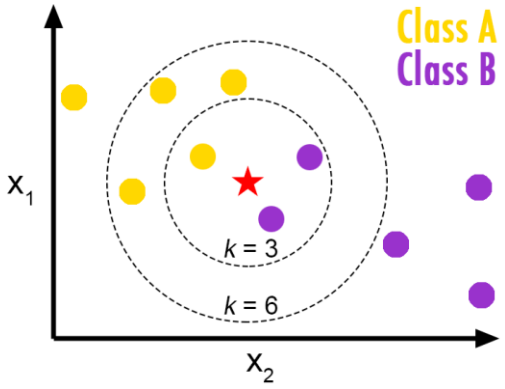
K-Nearest Neighbor Classifier

It is a supervised machine learning algorithm which classifies patterns according to the labels of its ‘k’ nearest neighbors.

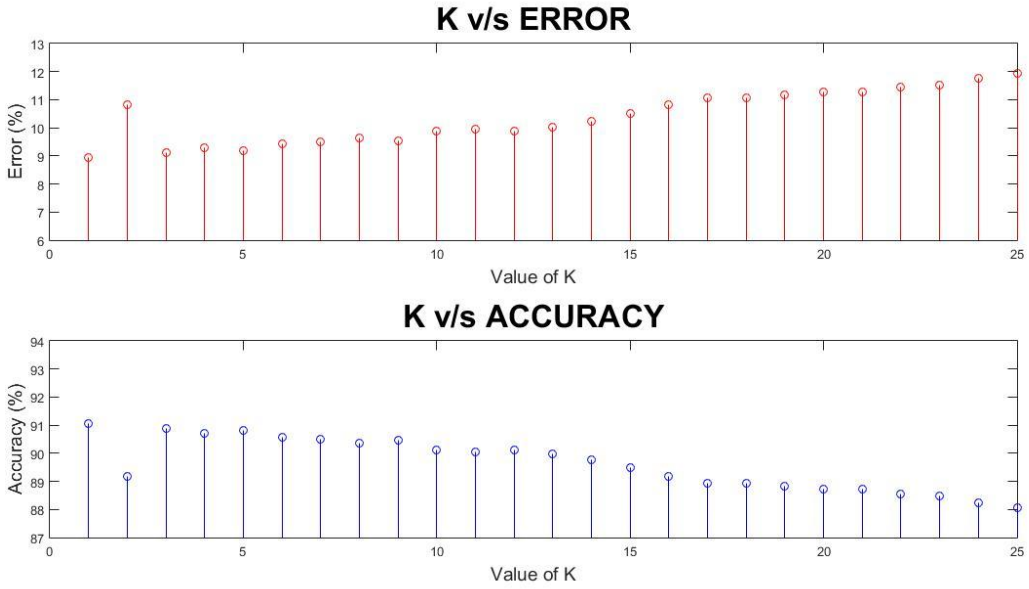


Here, we have taken the OCR hand written dataset. It is a 192 dimensional, 10 class problem with classes being the digits from 0 to 9. We have employed a 3 fold cross validation technique to determine the optimal value of ‘k’ to be used for classification of test data patterns.

Average error values for ‘k’ ranging from 1 to 25 have been calculated and ‘k’ with the least error has been chosen. For calculating nearness, Euclidean distance has been taken which is given as follows:

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_{192} - b_{192})^2}$$

Since, it's time consuming to compute this distance, it has first been calculated and stored in the file "dist_tra.txt" for the training set (during cross validation) and "dist_tes.txt" for the test set (during final test set prediction). The distances are loaded from these files into the C program and then the KNN program execution is continued.



As is evident from the graph, k=1 has been chosen since it has the least error. Accuracy obtained on the test set (using this k) is **92.17%**.