**AI ASSIGNMENT – 3 REPORT**

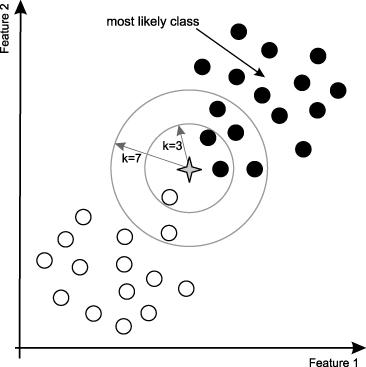
**GROUP MEMBERS:**

JUNAID N Z - 201601036

ROHAN S - 201601073

**K-Nearest Neighbor Classifier**

K-Nearest Neighbor Classifier (k-nnc) is a supervised machine learning algorithm which classifies patterns according to the labels of its ‘k’ nearest neighbors.

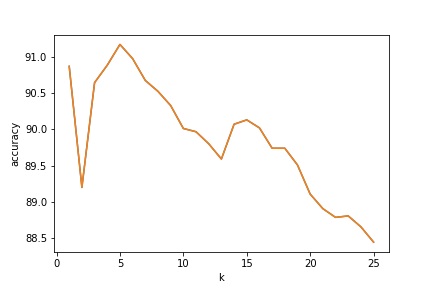


For this assignment, the OCR hand written dataset has been taken. It is a 192 dimensional, 10 class problem with classes whose values varies from 0 to 9. 3-fold cross validation technique has n been used to determine the optimal of ‘k’ for classification of test data patterns.

Mean accuracy for different values of ‘k’ ranging from 1 to 25 have been calculated and ‘k’ with the highest mean accuracy was chosen. Euclidean distance has been taken for calculating the closeness of the elements. It is given as follows:

The i index is varied throughout the range of features.

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=5 has been chosen since it has the least error. Accuracy obtained on the cross-validation set (using this k) is **91.17%** .

The test set accuracy is **92%** with **k=1** and **91%** with **k=5**.

**Naïve Bayes Classifier**

This is also a supervised learning method wherein the probability of occurrence of each test data in a particular class depends on the individual probabilities of each independent feature from the training data. Here each feature can be assumed to be independent and thus the name naïve.

Bayes’ Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes’ theorem is stated mathematically as the following equation:



where A and B are events and P(B) ? 0.

* Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.
* P(A) is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
* P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

Now, with regards to our dataset, we can apply Bayes’ theorem in following way:



Where y is each independent feature, and X is the class(0 to 9).

Here each feature has 5 discrete values(0-4) and 192 features with 10 classes.

An accuracy of **81.6%** is observed.