AI ASSIGNMENT – 3 REPORT

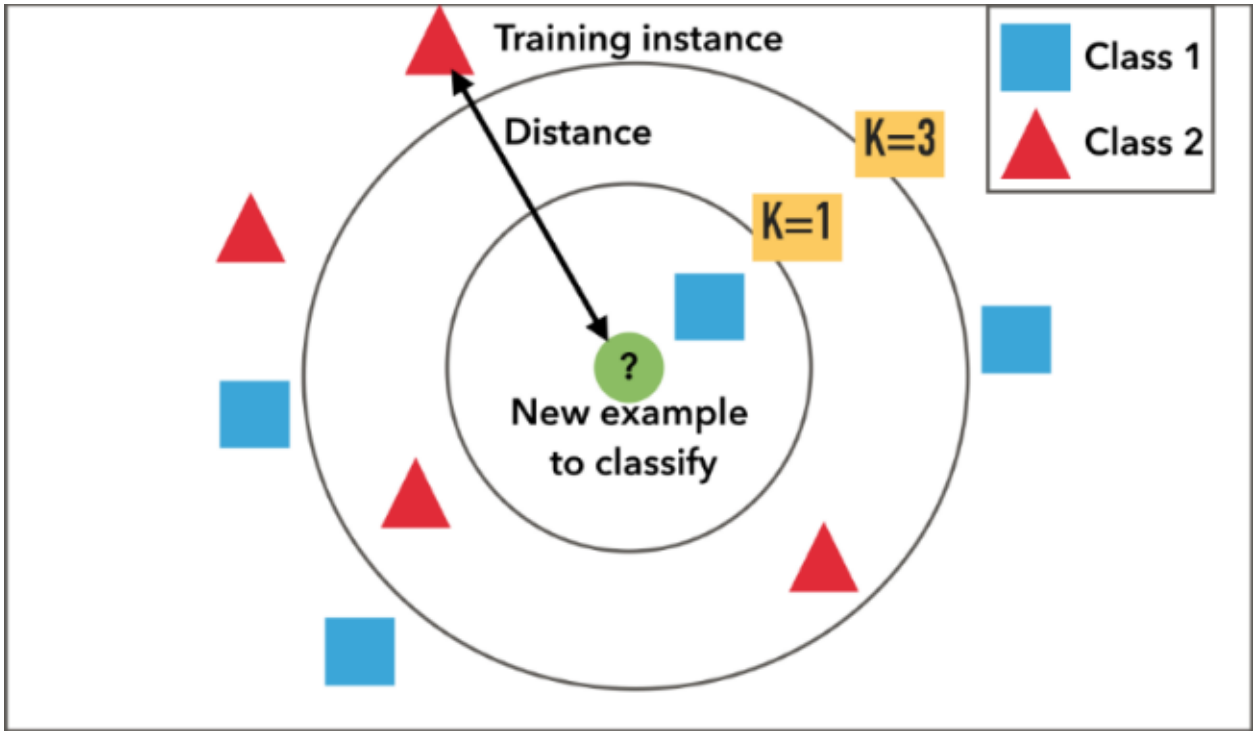
GROUP MEMBERS:

ADITYA HARIDAS MENON - 201601002

SHUBHAM GUPTA - 201601088

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 has been used to determine the optimal value of ‘k’ for classification of test data patterns. The test set compromises

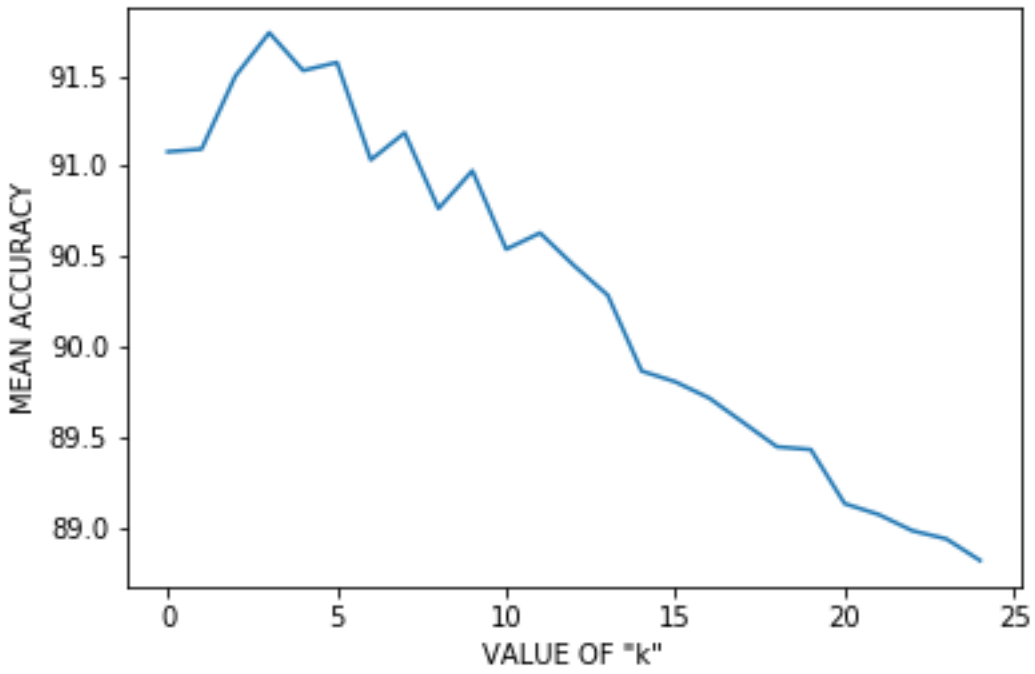
of 3333 observations and each observation has 192 features and 1 class label ranging from 0-9 .

Mean accuracies for different values of ‘k’ from 1 to 25 have been calculated and ‘k’ with the highest mean accuracy has been selected for classifying the test dataset. Euclidean distance has been taken for measuring the closeness of the points. It is given as follows:

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

For finding the value of ‘k’ by cross validation technique, the training set has been shuffled for getting a diverse set of data. The shuffled dataset has been stored in “shuffled.dat”.

The distances between each and every points are calculated and stored in a 2D matrix (dt[][]) for faster computation.



The above graph shows how the accuracy changes for different values of 'k'. From the graph and also by computation it has been found out that, the highest mean accuracy is **91.73789 and for k = 4.**

The test set is classified with k = 4 and the accuracy on the test data is found to be **93.189316**

Naïve Bayes Classifier

Naïve Bayes Classifier is a supervised learning algorithm where 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.

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:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are events and P(B) is not equal to 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:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Where y is each independent feature, and X is the class,

i.e. 0-9.

The accuracy after classifying the test set is found to be **81.548155.**
