AI Assignment – 3 Report

Group Members:

Aakash Shukla – 201601001

Prashant Kumar Mahanta – 201601066

Training Dataset:

The training dataset has 6670 points where each point has 192 features and a class Label.

Test Dataset:

The test dataset has 3333 points where each point has same 192 features and a class Label.

1.Find k to be used in k-nearest neighbor classifier (k-NNC) using a 3-fold cross validation.

First, we need to divide the training set into 3 equal parts (each called a fold). Then we need to take one-fold at a time do as follows.

1. *Using the rest data-set (other two-fold) train the model.*
2. *Test the model using the reserve portion (chosen fold) of the data-set.*

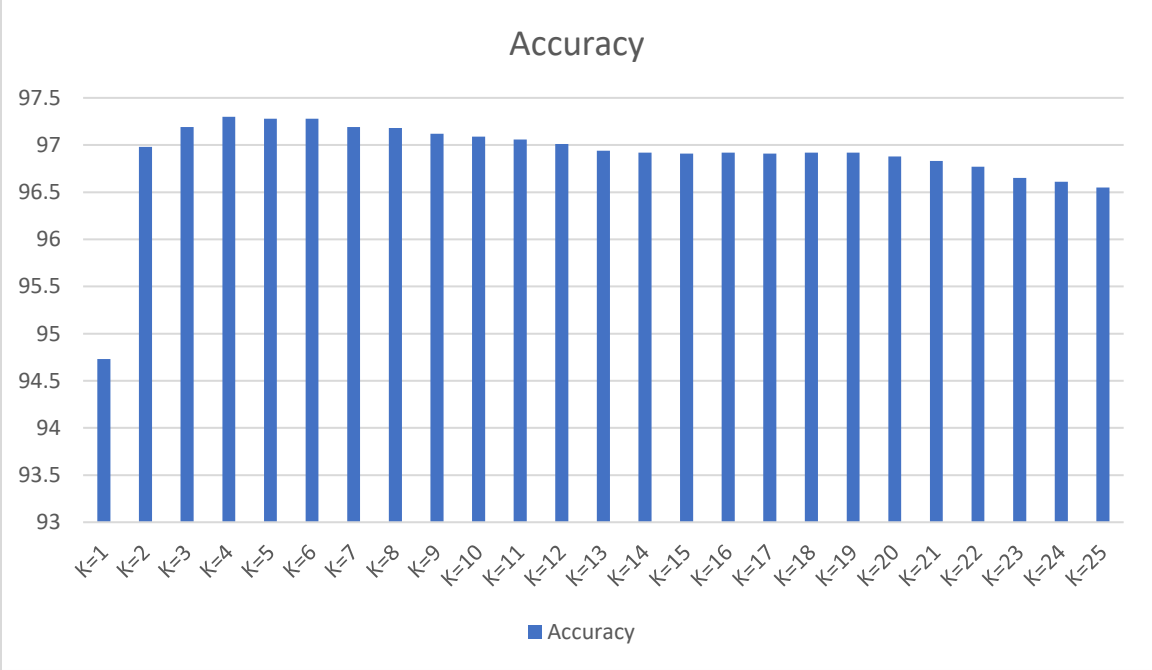
Algorithm

Let m be the number of training data samples. Let p be an unknown point.

1. *Store the training samples in an 2-array of data points arr[][]. This means each row of this 2-array represents 192 features(a point).*
2. *Make set S of K (varies from 1 to 25) smallest distances obtained. Each of these distances correspond to an already classified data point.*
3. *Return the majority label among S.*

- Check Whether the predicted label is same as the point's label. If the labels are different then we need to increase the error.
- We need to find the error for each fold (say e1 for 1st fold, e2 for 2nd fold, e3 for 3rd fold).
- Then error for a value of K (varies from 1 to 25) will be

  e=(e1+e2+e3)/3
- So, we need find error for each value of K.
- The K corresponds to minimum error will be the best K.



The best K we can see here is when K=4(i.e.,97.30%).

So here from the graph we can see that the best accuracy is for K=4.

NOTE: **We have shuffled the training dataset and used it.**

Then now as we have got the best K now we need to check this on test dataset.

So, after iterating through all the test instances, accuracy = **92.289%**

2. Employ the Naive Bayes classifier with the given data set. You can use log of the posterior (to overcome the small numbers problem). Report about your observations in your report.

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

The fundamental Naive Bayes assumption is that each feature makes an:

- independent
- equal

contribution to the outcome.

Bayes Theorem:

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:

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

Here, 192×5×10 3-D matrix is taken as a feature-matrix. 192 is the no. Of features, 5 represents different values of features and 10 is the no. Of class labels.

Hence, we reach to the result:

$$P(y|X) = \frac{P(x_1|y)P(x_2|y)P(x_3|y)\ldots P(x_{192}|y)}{P(x_1)P(x_2)P(x_3)\ldots\ldots\ldots P(x_{192})} \qquad y = \{0,1,2,3,4,5,6,7,8,9\}$$

$$P(y|X) = \text{argmax}_y\, P(y) \prod_{i=1}^{192} P(x_i|y)$$

So, after iterating through all the test instances, accuracy = **81.72%**