Assignment 3

Group Members:
B.Veda – 201601010 – veda.b16@iiits.in
N.Srilekha – 201601055 – srilekha.n16@iiits.in

AIM:

Task:1:- Find k that is used in k-nearest neighbor classifier (k-NNC) using 3-fold cross validation technique and testing it on the given test data and classify the given test input.
Task:2:- Using the train data, probabilities are found and Employing Naive Bayes Classifier on the given test data and printing the accuracy.

LANGUAGE USED: ANSI C

DATASETS:

Pp_tra.dat: 6670*193
Pp_tes.dat: 3333*193

Task-1:method:-

k-NNC:

- K-NNC classifies data sets based on their similarity with neighbors.
- The k-nearest neighbors algorithm (k-NN) is a non-parametric method used for **classification** and regression. In both cases, the input consists of the k closest training examples in the feature space.
- It choses its class based on the neighbor's class who have the maximum same class which are of nearest proximity.
- Euclidean Distance is used to find distance.
- The value of k is found out using 3-fold cross validation technique.
- The error rates are calculated on the validation set for a range of k values.
- Then the mean of all the error rates (argmin{e_1, e_2, ….,e_j, …, e_m}) is considered as the k value.
- Using the obtained k value from 3-fold validation, the test data is tested, and accuracy is calculated based on the correct classifications.

Task-2:method:-

Naïve Bayes Classifier:

- Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem.
- It is assumed that the features are independent of each other.

- The probability of occurrence of a feature value is its frequency ratio.

$$P(y|x_1,...,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)...P(x_n)}$$

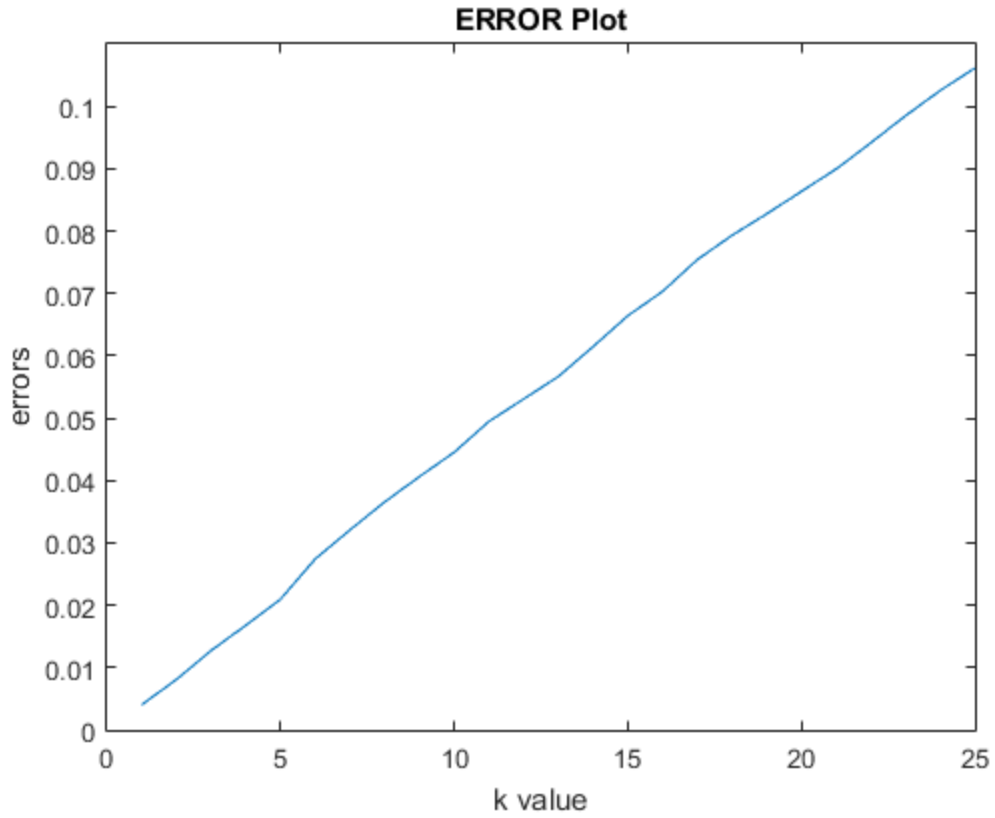
- The minimum error rate for each feature is given by $P(w_i/x)$
- The max. discrimination is the max posterior ($P(x/\omega_i)P(\omega_i)$) given by

$$P(X) = \sum_{j=1}^{c} P(X|\omega_j)P(\omega_j)$$

- Logarithm is applied based multiplication of small numbers becomes very small often result is zero.
- The class which maximum probability is picked out

$$y = argmax_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

RESULTS:
For k-NNC:
K value: 1
Accuracy: 90.009001 %
Error rates:



For Naïve Bayes Classifier:
Accuracy: 89.978998%

Conclusion: k-NNC is often not seen as a related classifier to Bayes classifier. This is because, distributions are not explicitly calculated. The time and space complexity of k-NNC are both equal to O(n) where as the time complexity of Naïve Bayes Classifier is O(1).