**AI ASSIGNMENT-3**

Krishna Kumar Dey-201601046

Preethi Hena-201601024

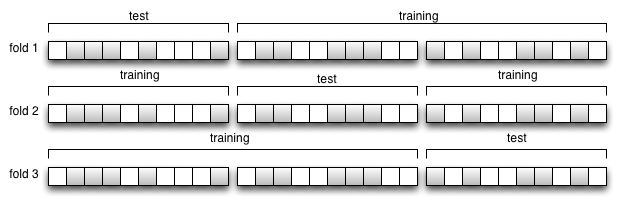
The assignment is divided into two tasks.One is finding k-value to be used in k-NNC algorithm and thus finding the error rate using this k.The other task involves using the Naive Bayes Classification.

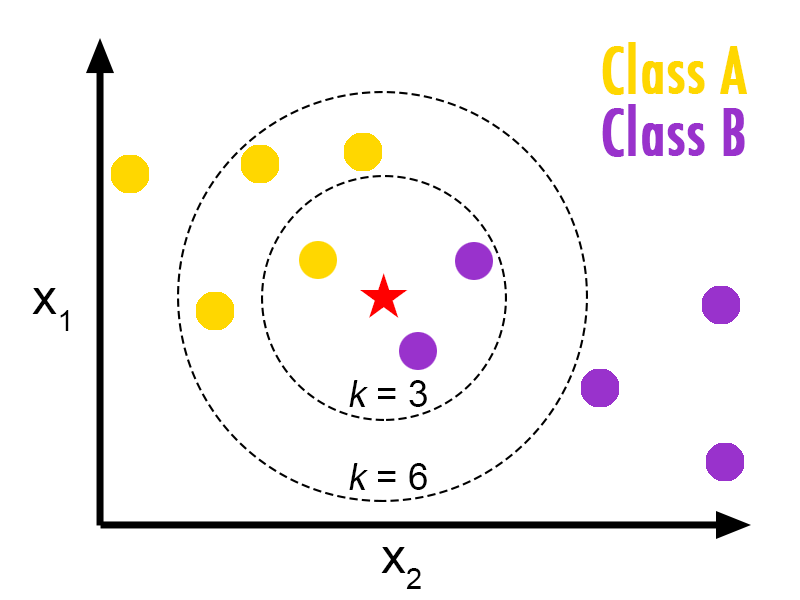
We are given with the training and test sets.There are totally 192 features and a class assigned to a single data point.Each feature is given a value from 0 to 4.

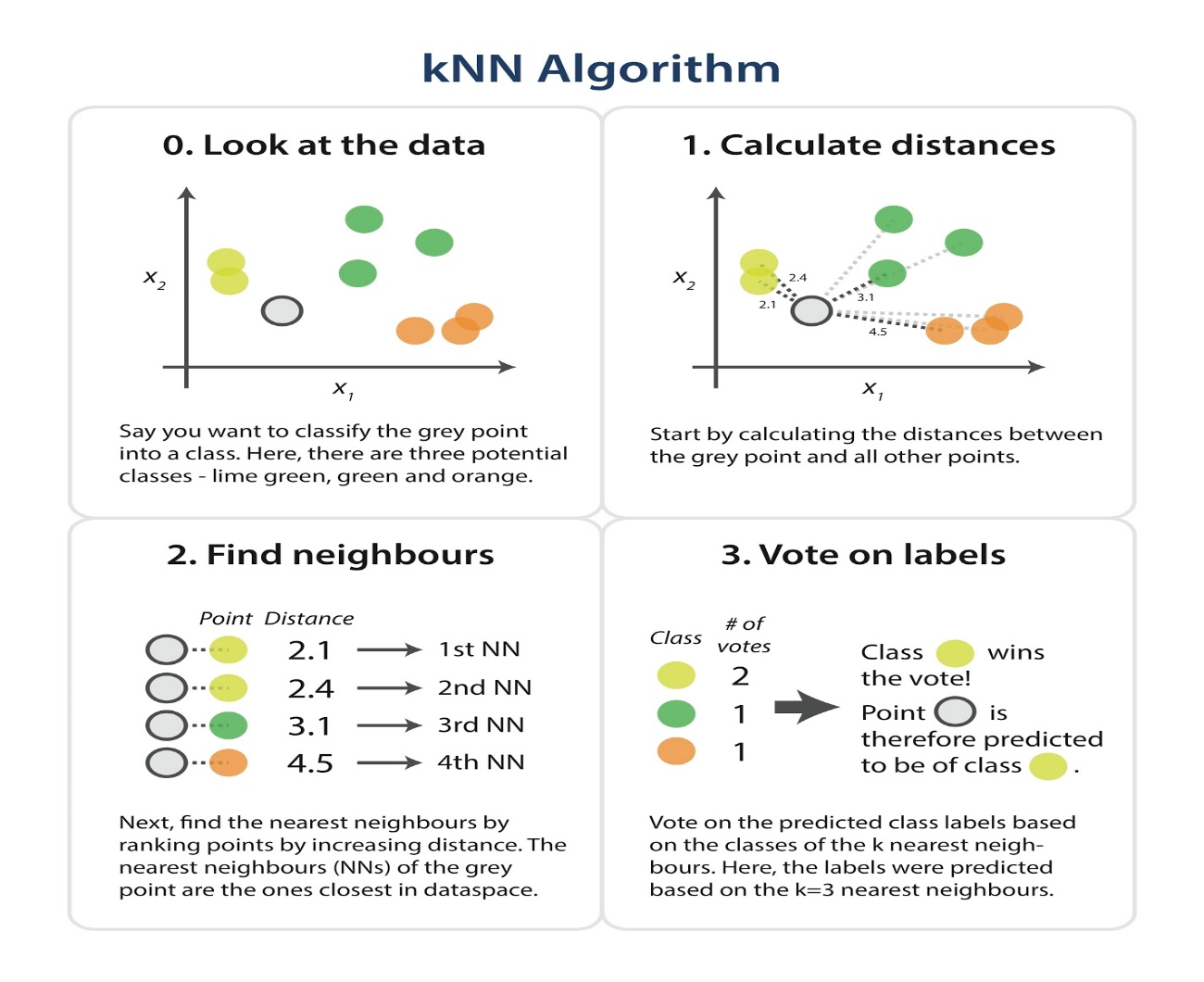
**k-NNC(kth-Nearest Neighbor Classification Algorithm):**Here we used 3-fold cross validation and applied k-NNC on that to find the value of k.

**3-fold cross validation**: Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

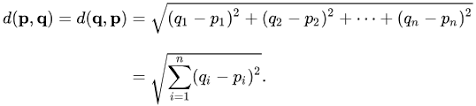
In 3-fold cross-validation, the original sample(here,the training set) is partitioned into 3 equal size subsamples. Of the 3 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 2 subsamples are used as training data. The cross-validation process is then repeated 3 times (the folds), with each of the 3 subsamples used exactly once as the validation data. The 3 results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.



**k-NNC**:In this algorithm ,the training set is used to make model and each point’s class in the test set is found by checking the k-nearest neighbour classes and out of them the majority class is found and the new data point is assigned that class.



In each of the folds, the two training sets make the model and each of the data point in test set is assigned a class using the k-NNC algorithm each time taking different k over the range of 1 to 25.The distance is found using Euclidean distance formula.



Errors are found for each k i.e over the range of 1 to 25 found for every fold and the average of errors for each k is found.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| K=1 | K=2 | K=3 | K=4…………….. | K=25 |
| Avg error when k=1 of all the three folds | Avg error when k=2 of all the three folds |  |  |  |

The k value which has the minimum average error is found and that k is finalised to be used to perform k-NNC and find the error rate on actual test set .

The graphs found by doing the above procedure are attached below:

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

Independent and equal contribution to the outcome. We assume that no pair of features are dependent. Secondly, each feature is given the same weight(or importance).

i.e P(A,B) = P(A)P(B)

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



where, y is class variable and X is a dependent feature vector (of size n) where:



hence

which can be expressed as:



Now, as the denominator remains constant for a given input, we can remove that term:



For this, we find the probability of given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as:

where x1, x2.. xn are features of vector X, y is the class.

So, finally, we are left with the task of calculating P(y) and P (xi | y).Please note that P(y) is also called class probability and P(xi | y) is called conditional probability.

Graphs which are obtained after using this classifier are attached below:

A close up of a map

Description generated with very high confidence