*ML Assignment -1*

*Group Members:*

Yagna Karthik Vaka (S20170010170)

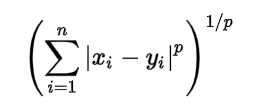
Sri Phani Sainath Kanala (S20170010063)

**Our understanding of the concept**

In **k – nearest neighbour classifier (k-NNC)**, we draw a sphere which contains **k** nearest neighbours of X.

Among *these* **k nearest neighbours**, according to the majority of classes around we decide the predicted class label.

But to obtain optimal **k** value for having most accuracy of the classifier we apply cross validation technique and to find the distances we use *Minkowski Distance formula*.



In **r-fold cross validation**, we partition the training set into **r blocks** and consider all except one as training set and the remaining as the test set.

For the range of **k** values (here 1 to 10) find the accuracy on the test set. We take the mean of all the accuracies of test set nodes for the whole range of 1 to 10. The maximum of mean of all the test accuracy rates,corresponding **k** value will be the best **k** value for the given classifier.

We can further have variety of variations in finding the distance between two nodes. This is done by changing the value of **p** (here 1 to 5) in *Minkowski distance formula.* It becomes Euclidean distance when **p** = 2.

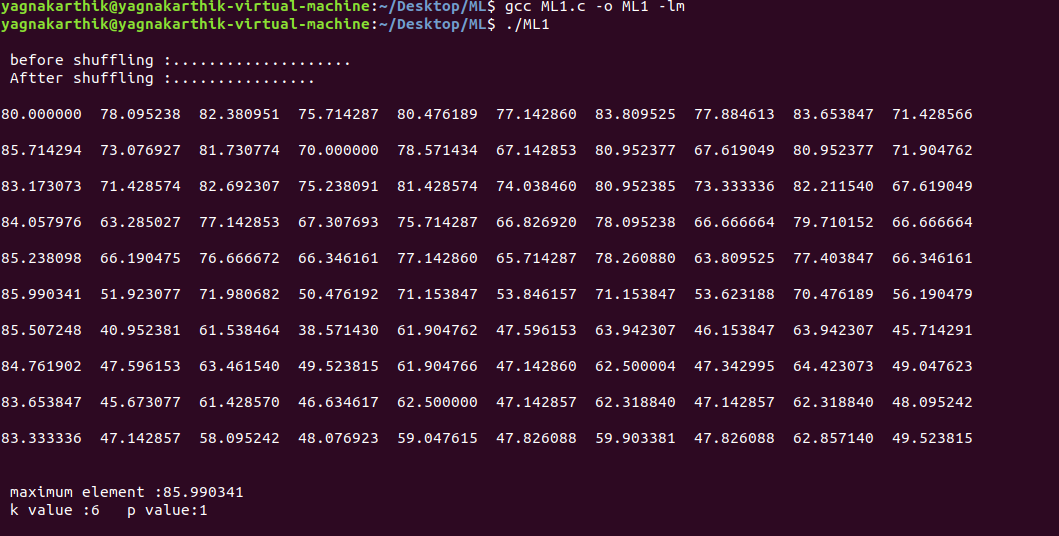
We repeat the above with all possible range of **k** values and **p** values. Our best bet for the classifier gives us the *maximum accuracy rate* on the test set with minimum error.

We explained our code in respective (ML1.c && ML2.c) comments.

***Results obtained….***

***1. Wheat seeds Data set.***

We obtained maximum accuracy when **p=1** and different values of **k (1 to 10)**.You can observe the values of accuracy percentages in below image.



As we are **shuffling the data** each time ,the pair of **k** and **p** values for which we are getting maximum accuracies is changing .Also we are getting maximum accuracies between *85 .0to 90.0%.*

**2.Optical recognition of Handwritten digits dataset:**

Here we compared our results with the results provided in that *dataset link(optdigits.names)*.

*Dataset results(expected):*

(K,p) values ----- accuracy

1,2 -------- 98.00%

2,2 -------- 97.30%

3,2 -------- 97.21%

4,2 -------- 97.43%

5,2 -------- 97.05%

6,2 -------- 97.03%

7,2 -------- 97.26%

8,2 -------- 97.17%

9,2 -------- 97.20%

10,2 -------- 97.74%

11,2 -------- 97.79%

*Dataset results(obtained):*

(K,p) values ----- accuracy

1,2 -------- 95.00%

2,2 -------- 94.30%

3,2 -------- 97.31%

4,2 -------- 94.63%

5,2 -------- 97.25%

6,2 -------- 97.43%

7,2 -------- 97.76%

8,2 -------- 97.47%

9,2 -------- 97.40%

10,2 -------- 97.84%

11,2 -------- 97.99%

Thank you……..