ML Assignment – 1

Members:
Adwait Thattey (S20170010004)
Aman Gupta (S20170010009)

Approach

The KNN classifier is used to classify the labels of the given test sample.
The classifier uses 2 values K and P to perform the KNN.
First the classifier finds the MinKowski's distance of the new point to all the points. The Minkowski's distance is given by:

$$d = \sqrt[p]{\sum_{i=0}^{n} |x_i - y_i|^p}$$

Where '**n**' is the number of features. Using this distance it finds the nearest '**k**' points. The class to which maximum of the k points belong to , is given to the new point.

The **R-Fold Cross Validation** aims to find the optimum value for **k and p**.
The given dataset is divided into '**r**' parts. One of the parts is taken as testing set and others are taken as training set. The KNN is applied on all points on the testing set and the accuracy is found for the given k,p. Then some other rth part is taken as testing set and others as training set. This continues till we have 'r' accuracies. The average is taken and that is said to be the accuracy for the given 'k' and 'p'
The parameters of p and k are varied in a range so as to determine that for what value of p and k the classifier is giving the highest accuracy or least error. This is being run on a cross validation set.
accuracy of that particular element. So we will end up doing this (P*K) times and then the
greatest accuracy among all of them will give us the correct value of P and K.

Thus we repeat the experiment p*k times and find the best (optimum) value of k and p.

Code:
The explanation of the code is written in a file named README.txt included in the zip file

Results:

Dataset 1:

This is for wheat seeds dataset

Here is the attached results file for each k and p
Result Of Dataset 1

The optimum value of K and P is:
k:8 p:2 accuracy: 91.9216%

Dataset 2:

Here is the result for each k and p
Result of Dataset 2

The optimum value of k and p is:
k:1 p:2 accuracy: 98.4310%