Abstract:-

In this assignment,i have implemented k-nearest neighbor clustering algorithm using 3 different datasets. Also,i have utilized 3 different measures for classifying the examples. These distance measures included the simple euclidean distance and kernels such as Polynomial kernel and radial kernel. The accuracy obtained in classification of examples,by using the above mentioned 3 approaches is analyzed and compared with others.

Introduction:-

We are provided with 3 different datasets: ecoli.data,yeast.data and glass.data.We have to implement K nearest neighbors algorithm and classify examples using different measures. And then analyze and state which distant measures were more successful in classifying examples present in data files.
First the data is fetched from the data file,and then its attributes are normalized and 1o fold cross validation is performed. Then,classes of tuples in testing sets are determined using knn algorithm on training set. And then finally the accuracy of the prediction is calculated

K-nearest neighbour:

The k-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. The class of a data sample is determined by classes of k nearest data samples. The weights given to different samples may vary and different techniques can be used to calculate the distance.
In this assignment,through input the user provides k,no. of neighbors to be considered and which distant measure to use out of the following three:
1)Euclidean distance
2)Polynomial kernel
3)Radial basis kernel

(1) Polynomial kernel:

$$K(\boldsymbol{x}, \boldsymbol{y}) = (1 + \langle \boldsymbol{x}, \boldsymbol{y} \rangle)^p$$

(2) Radial basis kernel:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{\sigma^2}\right\}$$

Distance between samples has been calculated using:

$$d^2(\psi(\boldsymbol{x}), \psi(\boldsymbol{y})) = K(\boldsymbol{x}, \boldsymbol{x}) - 2K(\boldsymbol{x}, \boldsymbol{y}) + K(\boldsymbol{y}, \boldsymbol{y})$$

Datasets used:-

The following datasets were used:-

1)ecoli.data

Source: http://archive.ics.uci.edu/ml/datasets/Ecoli

Kenta Nakai
Institute of Molecular and Cellular Biology
Osaka, University
1-3 Yamada-oka, Suita 565 Japan
nakai@imcb.osaka-u.ac.jp
http://www.imcb.osaka-u.ac.jp/nakai/psort.html
Donor: Paul Horton (paulh@cs.berkeley.edu)
Date: September, 1996

Purpose:To find Localization site of protein

Number of Instances: 336 for the Ecoli dataset and

Number of Attributes for Ecoli dataset: 8 (7 predictive, 1 name)

Attribute Information.

1. Sequence Name: Accession number for the SWISS-PROT database
2. mcg: McGeoch's method for signal sequence recognition.
3. gvh: von Heijne's method for signal sequence recognition.
4. lip: von Heijne's Signal Peptidase II consensus sequence score. Binary attribute.
5. chg: Presence of charge on N-terminus of predicted lipoproteins. Binary attribute.
6. aac: score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins.
7. alm1: score of the ALOM membrane spanning region prediction program.
8. alm2: score of ALOM program after excluding putative cleavable signal regions from the sequence.

2)glass.data

Sources: http://archive.ics.uci.edu/ml/datasets/Glass+Identification
(a) Creator: B. German
-- Central Research Establishment
Home Office Forensic Science Service
Aldermaston, Reading, Berkshire RG7 4PN
(b) Donor: Vina Spiehler, Ph.D., DABFT
Diagnostic Products Corporation
(213) 776-0180 (ext 3014)
(c) Date: September, 1987

Purpose:To identify glass type.

Number of Instances: 214

Number of Attributes: 10 (including an Id#) plus the class attribute -- all attributes are continuously valued.

Attribute Information:

1. Id number: 1 to 214
2. RI: refractive index
3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
4. Mg: Magnesium
5. Al: Aluminum
6. Si: Silicon
7. K: Potassium
8. Ca: Calcium
9. Ba: Barium
10. Fe: Iron
11. Type of glass: (class attribute)
   -- 1 building_windows_float_processed
   -- 2 building_windows_non_float_processed
   -- 3 vehicle_windows_float_processed
   -- 4 vehicle_windows_non_float_processed (none in this database)
   -- 5 containers
   -- 6 tableware
   -- 7 headlamps

3)yeast.data

Source:http://archive.ics.uci.edu/ml/datasets/Yeast
Kenta Nakai
Institute of Molecular and Cellular Biology
Osaka, University
1-3 Yamada-oka, Suita 565 Japan
nakai@imcb.osaka-u.ac.jp
http://www.imcb.osaka-u.ac.jp/nakai/psort.html

Purpose:To find Protein Localization Site.

Number of Instances: 1484 for the Yeast dataset.

Number of Attributes for Yeast dataset: 9 (8 predictive, 1 name)

Attribute Information.

1. Sequence Name: Accession number for the SWISS-PROT database
2. mcg: McGeoch's method for signal sequence recognition.

3. gvh: von Heijne's method for signal sequence recognition.
4. alm: Score of the ALOM membrane spanning region prediction program.
5. mit: Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins.
6. erl: Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute.
7. pox: Peroxisomal targeting signal in the C-terminus.
8. vac: Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins.
9. nuc: Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

Result and analysis:

	Euclidean Distance	Polynomial Kernel	Radial Basis Kernel
Ecoli	85.724	86.68	88.286
Glass	61.744	63.087	65.77
Yeast	68.18	72.727	71.841

- The performance of distance measures is in order:euclidean distance<Polynomial Kernel<Radial basis Kernel for ecoli and yeast dataset.
- And performance is in order:Euclidean distance<Radial basis Kernel<Polynomial Kernel for glass dataset.
- Performance in general increases with k,peaking at around 8-10 and then dropping slightly(because of unnecessary weightage given to samples far from our example which have very less influence) if k is increased for almost all datasets and all distant measures.
- Optimum performance for polynomial kernel is obtained around p=14 and optimum performance for sigma is obtained around sigma=4 for ecoli and yeast dataset and sigma=1 for glass dataset.

Conclusion:

The Usage of Kernels indeed improve performance of knn in different dataset as compared to simple euclidean distance knn,but the selection of kernel depends on which dataset is being used and thus for performance improvement,kernel should be chosen wisely.