README FILE

Shirsha Chowdhury

18EE10068

Imports Used:

Pandas to read the dataset

Numpy for mathematical operations on vectors

Math for mathematical tasks used for formulas

Functions used:

Zscores(arr): To find the Z scores of quantitative data.

vdm(train,a,x,y): To find the Value Distance Metric of Nominal data

distance(sample,train, test): This function finds the distance between 2 vectors.

getDistance(sample,test): Gets the distance vector for all the distances between the training examples and the single test example

getNeighbors(dist, num_neighbors): Function which returns the k nearest neighbors according to the distance array.

majorityVoting(data,arr): This function takes in the neighbor array and returns the majority class in present in the neighbor array

Output(sample,test,k): function to print the optimal k and user defined k. The optimal K came out to be at k=19, which we obtained by dividing the training set into 2 parts in the (training set and validation set) in the ration 0.85:0.15 and got the highest accuracuy of 86.67% at k=19. Also in general the optimal value of K lies in between the value of root(N), where N is the number of training examples.

Idea: The Main Idea is to use KNN algorithm to classify the test examples as “0” or “1”. In the training set [‘chol’,’age’,’trestbps’,’thalack’,’oldpeak’] were quantitative data and was replaced by Zscore and then Euclidean distance was used to find the distance between two data points. The rest were nominal attributes for which vdm was used to get the distance between them. In general the K should be chosen around the root of number of training examples.

The output was printed as 0 1 0 1.
