BT3041 Analysis and Interpretation of Biological Data

Assignment 1

Name: Madhav Maheshwari

Roll No: BS16B022

The code for the dbscan algorithm consists of three function. The basic algorithm of the code is loop based.

First function is the main function(**dbscan\_2**) which is called when the file is run with the input arguments as (dataset,eps,min\_pts,n) dataset in csv format and n is the figure number of the plot for use in the next part of the question. This function reads the csv file and puts it in a matrix, then sorts the matrix on x coordinate basis. Then it calls for each point the second function to check if the point is a core point or not, if it is not then it is labelled as noise(-1).

Here the second function is used to check weather the point is core or not, it takes the matrix, the point, epsilon, min points and index of the point in the matrix as input arguments and checks for the points within the circle of epsilon radius using the Euclidian distance formula for the points ranging from x-epsilon to x+epsilon since the matrix is sorted on x basis we need not to look beyond this range of x. If these no of points exceed or are equal to the min points then this point will be considered as core point else a noise.

Once we find the first core point it is given as input to the third function with several other input arguments. This function grows the cluster by taking two counters and adding all the points inside the circle in a matrix, one counter keeps track of the length of the matrix to append the next point and the other counter keeps track of the points inside the matrix which are checked if they are core or noncore. Once both the counters reaches the same position in the matrix the loop is beaked and the matrix is plotted on the x-y plane as a single cluster and the cluster labels are returned to the main function. Once all the clusters are plotted the main function plots all the noise points.

These labels are compared by the original labels given and hence calculate the accuracy of the clustering program. This accuracy is returned as the output.

For the second part of the question I created another function(**dbs**) which calls the dbscan\_2 function for different combinations of epsilon and min points and record their accuracy along with the epsilon and min points in a matrix. This matrix is plotted as a 3D scatter plot for epsilon vs min points vs accuracy.