BT3041 Analysis and Interpretation of Biological Data

Assignment 1

Name: Priyanka Mowlali

Roll No: BE16B021

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

First function is the main function(**dbscan\_1**) 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).

The second function is to check weather a point is a core or not, since the array is sorted, we can just check for the points for x-eps to x+eps for weather they are inside the epsilon radius circle of the point by calculating Euclidian distance between the points. It returns true if the no of points are greater than or equal to min points and false otherwise.

If the check core function returns true then that point is fed to the third function for starting the clustering. This function creates a matrix and put all the points inside this circle in that matrix. Now from this matrix we will grow the cluster, we will check for these points in the matrix to be core or not by second function and if it is core then the same third function is called recursively and the same matrix is generated for that point also and the process is repeated until for all the point sit is checked in the matrices generated. Once this is done this function returns the label of the points in the matrices as same label.

The main function uses these labels to plot the points on the x-y plane. The points are plotted with different colour for different cluster labels and another colour for noise as well. This function also compares the labels generated with the given labels and calculate the accuracy percentage and returns this accuracy as output.

For the next parts of the question I have written another function(**dbscan**) which for different combinations of epsilon and min points calls the dbscan\_1 function and stores the eps, min points and the accuracy returned by the function into a matrix and this is plotted in a 3d scatter plot.