Homework 3

CS 6720

Due: 1st March by the time class starts.

1. Implement OPTICS after generating the dataset through this script. Type the following command

java -jar data.jar <last 2 digits of roll number>

to get the dataset *output.dat*.

Do the following

a. Run OPTICS and generate the reachability plot for $\epsilon = 10, minPts = 10$. (20 points)
b. Extract the clusters from the reachability plot. You should properly explain the parameters you used and why? (20 points)
c. Make OPTICS faster by indexing the dataset using Range Tree. You can download any implementation from the internet. Plot the running time of OPTICS with Range Tree and without Range tree at 10,000 points, 25,000 points, 100,000 points, and 200,000. You can randomly sample points from the entire dataset to create the smaller datasets. Explain the results. (20 points)
d. Plot the running time of OPTICS with Range Tree and without Range tree at $\epsilon = 5,10,25,50,100$. Explain the results. (20 points)

2. Generate 1000 random points at dimensions 10, 100, 500, 1000, 10000. Each dimension value ranges between -1 and 1. For each number of dimensions, compute the angle between all pairs and plot the distribution of angles. Explain the results. (20 points)
3. This is a more research-like question and hence, open-ended. You will be graded based on how well you bring out the different aspects of the datasets and MCL.
   a. Use MCL to detect communities in these datasets. They have ground-truth communities. Use the parameters to optimize the results. Comment on the parameter values that produce best results and the possible reasons. Whether you want to present any plots, table, etc., it is completely up to you. (40 points)
      i. https://snap.stanford.edu/data/com-DBLP.html
      ii. https://snap.stanford.edu/data/com-Amazon.html