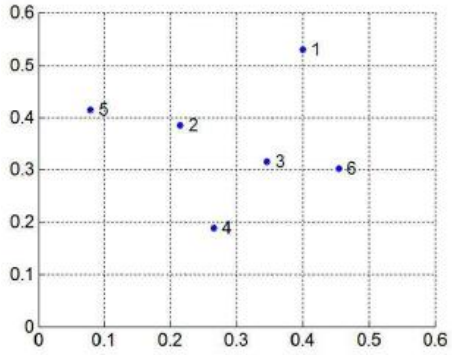
Homework 3

CS 870

Due: 10th Sept 11:59PM. No late submissions will be allowed. Please start early.

1

a. Draw the dendrogram for single linkage clustering on the data below (5 points):



Point	x	y
1	0.40	0.53
2	0.22	0.38
3	0.35	0.32
4	0.26	0.19
5	0.08	0.41
6	0.45	0.30

b. What is the computation complexity of clustering n points without using the support of any other data structure? Explain your reasoning. (5 points)

c. Propose an $O(n^2 \log n)$ algorithm. Explain your reasoning clearly. (10 points)

2. Implement OPTICS to produce the reachability plot. You are free to use any programming language but it must compile in an unix based OS. You need to provide three scripts

   a. rollno_compile.sh: Compile your code

   b. rollno_optics.sh <minPts> <epsilon>: Compute reachability data

   c. rollno_plot.sh: Plot the reachability data. You MUST use matplotlib for this part.

   The dataset for your optics code would be in the following format. You SHOULD NOT assume the numbers to be integers.

3 4 5 …

1 7 8 …

…

Here each line corresponds to an n-dimensional point. Each dimension is separated by a space. The number of dimensions and lines can be of any value till 5 and 500,000 respectively.

i. Gets correct plot (25 points)

ii. Competition based on running time as in last homework. We will only consider the running time of rollno_optics.sh. You must produce the correct reachability plot to be eligible for this competition. Do not use multi-threaded programming. However, you are free to explore the support of index structures. (15 points)

Deliverables: A PDF report for part 1 and readme for part 2. All codes and scripts.