Seoul National University

M1522.001400 Introduction to Data Mining

Spring 2016, Kang

Homework 6: Clustering (Chapter 7)

Due: May 16, 09:30 AM

Reminders

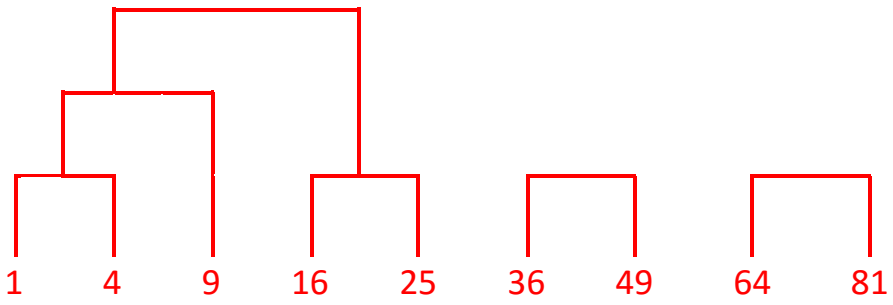
- The points of this homework add up to 100.
- Like all homeworks, this has to be done individually.
- Lead T.A.: Minsoo Jung (qtyp456987@gmail.com)
- Please type your answers *in English*. Illegible handwriting may get no points, at the discretion of the graders.
- If you have a question about assignments, please upload your question in eTL.
- If you want to use slipdays or consider late submission with penalties, please note that you are allowed one week to submit your assignment after the due date.

Remember that:

- Whenever you are making an assumption, please state it clearly.

Question 1

Perform a hierarchical clustering of the one-dimensional set of points 1, 4, 9, 16, 25, 36, 49, 64, 81, assuming clusters are represented by their centroid (average), and at each step the clusters with the closest centroids are merged. Stop combining clusters when the number of clusters is 3. [15 points]



Question 2

For the points of **Figure 1**, perform a hierarchical clustering. Assume that the distance between two clusters is:

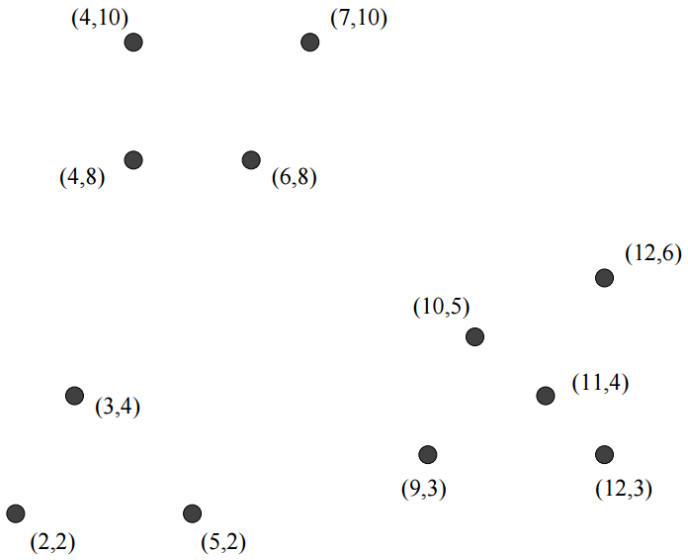
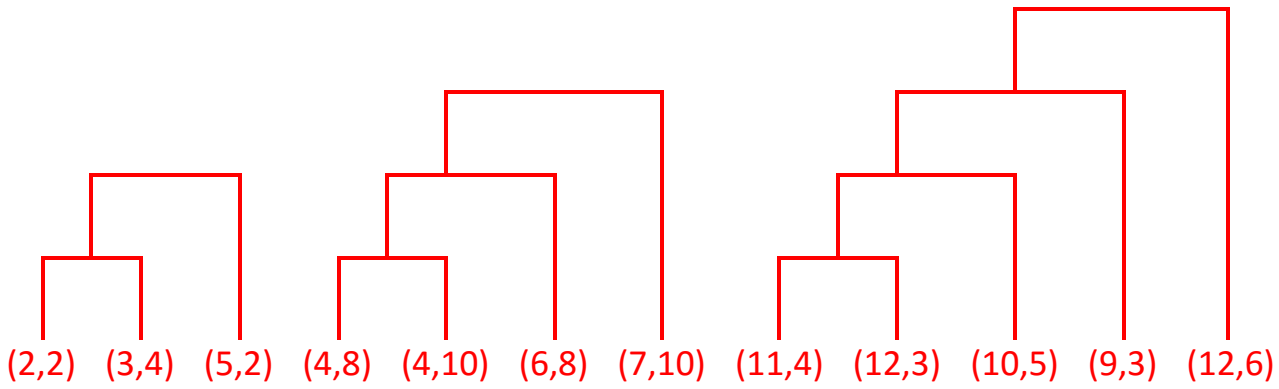
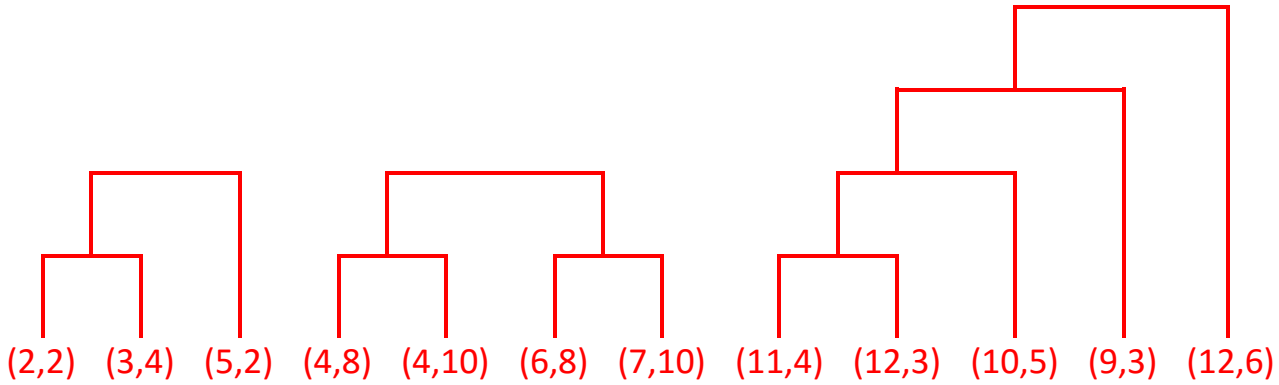


Figure 1: An example twelve points for **Question 2 and 3**

(a) The minimum of the distances between any two points, one from each cluster. Stop combining clusters when the number of clusters is 3. [15 points]



(b) The average of the distances between pairs of points, one from each of the two clusters. Stop combining clusters when the number of clusters is 3. [15 points]



Question 3

For the points of **Figure 1**:

(a) Compute the representation of the cluster as in the BFR Algorithm. That is, compute N, SUM, and SUMSQ. Assume the number of clusters is 3. [15 points]

Cluster	Points	N	SUM	SUMSQ
1	(2,2),(3,4),(5,2)	3	(10,8)	(38,24)
2	(4,8),(4,10),(6,8),(7,10)	4	(21,36)	(117,328)
3	(9,3),(10,5),(11,4),(12,3),(12,6)	5	(54,21)	(590,95)

(b) Compute the variance and standard deviation of each cluster in each of the two dimensions. [10 points]

Cluster	Points	Variance	Std
1	(2,2),(3,4),(5,2)	(1.56,0.89)	(1.25,0.94)
2	(4,8),(4,10),(6,8),(7,10)	(1.69,1)	(1.30,1)
3	(9,3),(10,5),(11,4),(12,3),(12,6)	(1.36,1.36)	(1.17,1.17)

Question 4

Suppose a cluster of three-dimensional points has standard deviations of 2, 3, and 5, in the three dimensions, in that order. Compute the Mahalanobis distance between the origin (0, 0, 0) and the point (1, −3, 4). [15 points]

$$\frac{3}{10}\sqrt{21} \approx 1.37$$

Question 5

Consider two clusters that are a circle and a surrounding ring, as in **Figure 2**. Suppose:

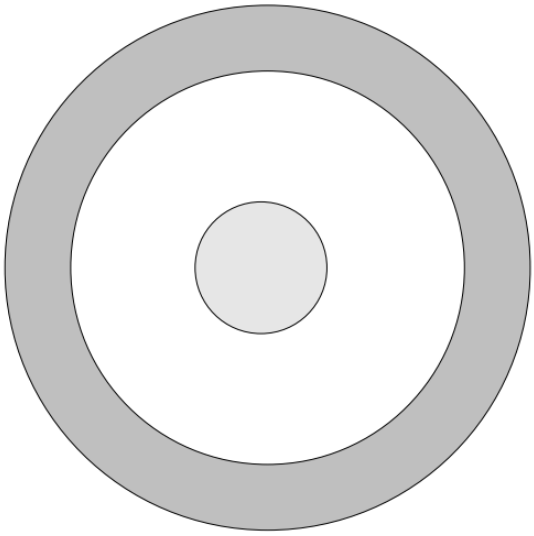


Figure 2: Two clusters, one surrounding the other, for **Question 5**.

i. The radius of the circle is c.

ii. The inner and outer circles forming the ring have radii i and o, respectively.

iii. All representative points for the two clusters are on the boundaries of the clusters.

iv. Representative points are moved 20% of the distance from their initial position toward the centroid of their cluster.

v. Clusters are merged if, after repositioning, there are representative points from the two clusters at distance d or less.

In terms of d, c, i, and o, under what circumstances will the ring and circle be merged into a single cluster? [15 points]

$$0.8 * (i - c) \leq d$$