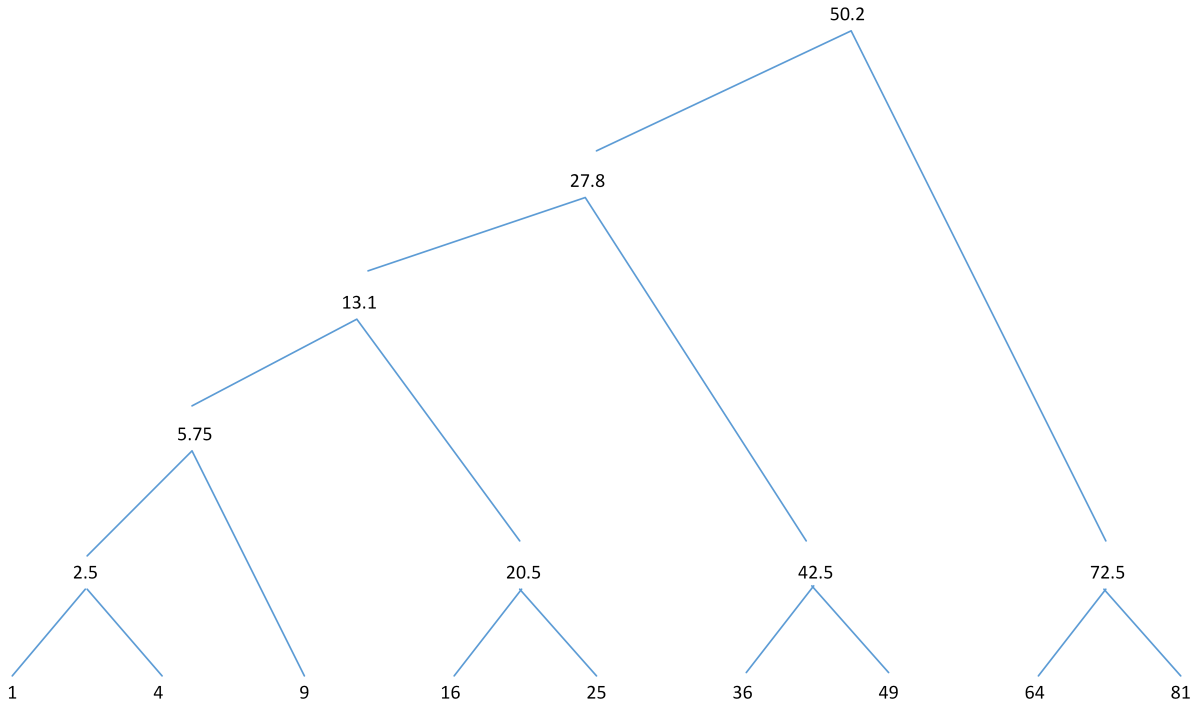
Intelligent Data Management - Exercise 6

Name: Triet Ho Anh Doan
Date: December 8, 2017

Assignment 1: Hierarchical Clustering

a. 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.



b. We can select clustroids for clusters, even if the space is Euclidean. Consider the three natural clusters in Fig. 7.2, and compute the clustroids of each, assuming the criterion for selecting the clustroid is the point with the minimum sum of distances to the other point in the cluster.

- To choose clustroids, first we compute the distance between every points in each cluster. Then, choose the point having the minimum distance to the rest.
- Cluster 1:

Point	(4, 10)	(7, 10)	(4, 8)	(6, 8)	Sum
(4, 10)	0	3	2	2.83	7.83
(7, 10)	3	0	3.61	2.24	8.85
(4, 8)	2	3.61	0	2	7.61
(6, 8)	2.83	2.24	2	0	7.07
Clustroid	(6, 8)				

- Cluster 2:

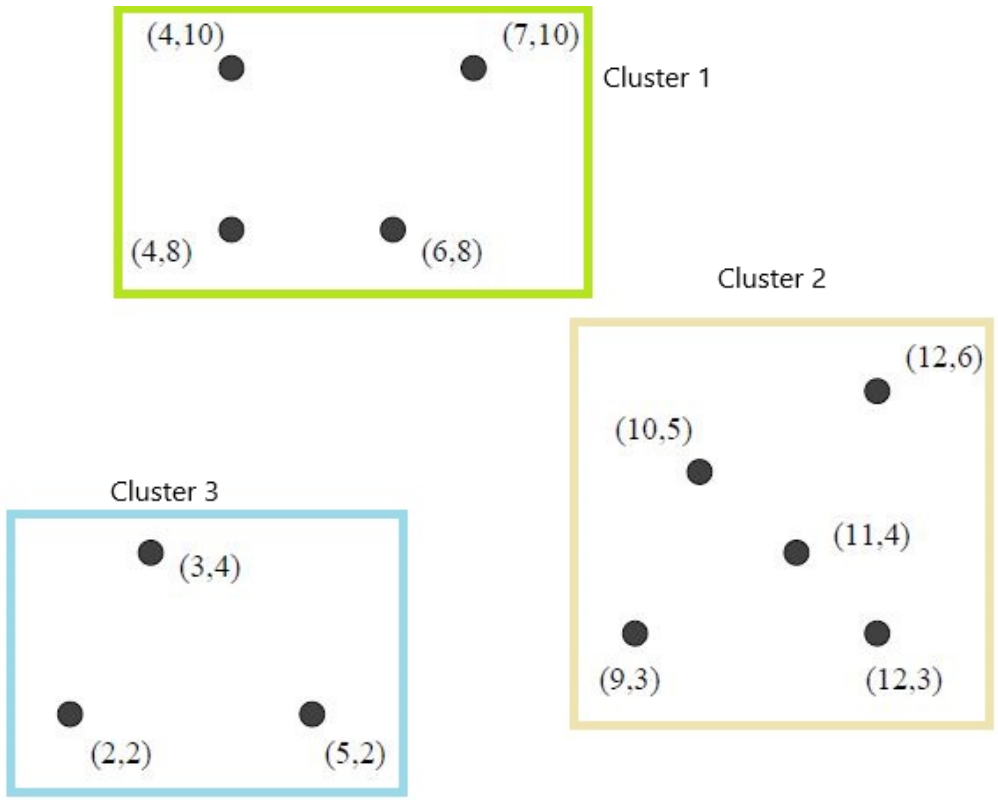


Figure 1: Example data with cluster notation

Point	(3, 4)	(2, 2)	(5, 2)	Sum
(3, 4)	0	2.24	2.83	5.07
(2, 2)	2.24	0	3	5.24
(5, 2)	2.83	3	0	5.83
Clustroid	(3, 4)			

- Cluster 3:

Point	(10, 5)	(12, 6)	(11, 4)	(9, 3)	(12, 3)	Sum
(10, 5)	0	2.24	1.41	2.24	2.83	8.72
(12, 6)	2.24	0	2.24	4.24	3	11.72
(11, 4)	1.41	2.24	0	2.24	1.41	7.3
(9, 3)	2.24	4.24	2.24	0	3	11.72
(12, 3)	2.83	3	1.41	3	0	10.24
Clustroid	(11, 4)					

Assignment 2: K-means Clustering

a. For the three clusters of Fig. 7.8, compute the representation of the cluster as in the BFR Algorithm. That is, compute N, SUM , and SUMSQ. (*This is done together with question b*)

b. For the three clusters of Fig. 7.8, compute the variance and standard deviation of

each cluster in each of the two dimensions.

$$\sigma^2 = \frac{SUMSQ_i}{N} - \left(\frac{SUM_i}{N}\right)^2$$

Cluster	N	SUM	SUMSQ	Variance	Standard Deviation
1	4	[21, 36]	[117, 328]	[1.69, 1]	[7.54, 1]
2	3	[10, 8]	[38, 24]	[1.56, 0.89]	[1.25, 0.94]
3	5	[54, 21]	[590, 95]	[1.36, 1.36]	[1.17, 1.17]

c. 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).

$$\begin{aligned}\text{Distance} &= \sqrt{\sum_{i=1}^{d}\left(\frac{p_i - c_i}{\sigma_i}\right)^2} \\ &= \sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{-3}{3}\right)^2 + \left(\frac{4}{5}\right)^2} \\ &= \sqrt{0.25 + 1 + 0.64} \approx 1.37\end{aligned}$$