L3S Research Center
Large Scale Data Mining, SS 2016
Dr. Avishek Anand
Solution to Assignment 1, due: **28 April 2016**

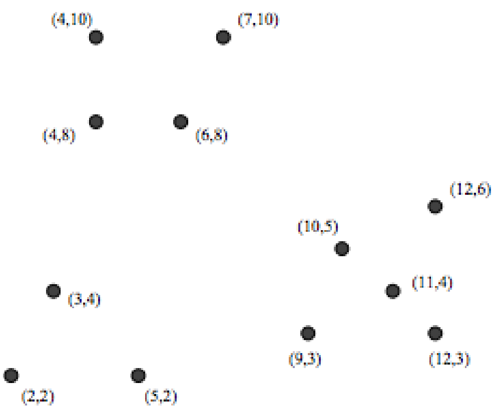
Problem 1.

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.

2. How would the clustering of the figure below (you can also see Example 7.2, page 246 of the MMDS book) change if we used for the distance between two clusters:

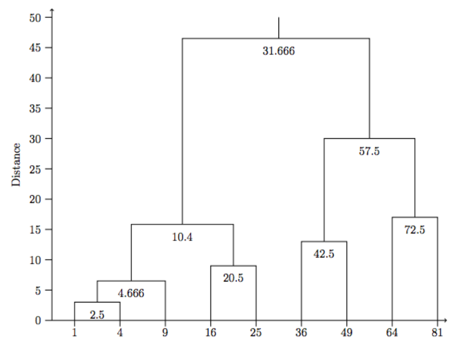
   (a) The minimum of the distances between any two points, one from each cluster.

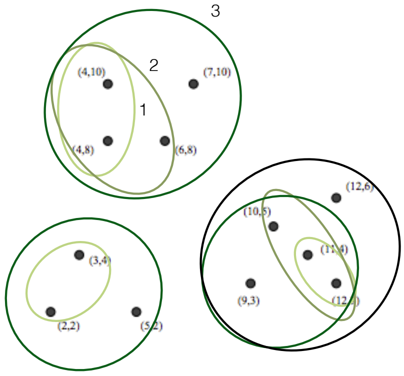
   (b) The average of the distances between pairs of points, one from each of the two clusters.



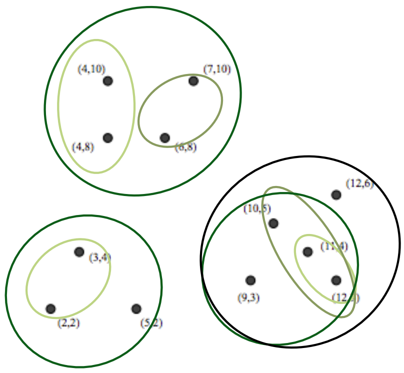
Solution:

(a)




3
2
1
(4,10)
(7,10)
(4,8)
(6,8)
(12,6)
(10,5)
(3,4)
(9,3)
(2,2)
(5,2)

(b) i.


(4,10)
(7,10)
(4,8)
(6,8)
(12,6)
(10,5)
(3,4)
(9,3)
(2,2)
(5,2)

ii.

Problem 2. In k-means algorithm we compute the cluster centroid (prototype) as the mean of the points in the cluster, as this minimizes the sum of squared errors. Consider a variation of k-means for one-dimensional data where we want to minimize the sum of *absolute* errors, that is, our goal is to find clusters $C_1, C_2, \ldots, C_k$ that minimize

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} |\nu_j - x_i|,$$

where ν_j is the prototype of cluster C_j. Prove that we should use the median of points in C_j as ν_j. *Hint:* compute the derivative of sum of absolute errors w.r.t. cluster prototypes.
Solution: For a cluster C_j, lets look at the derivative of the objective function w.r.t ν_j.

$$\frac{\partial}{\partial \nu_j} \sum_{x_i \in C_j} |\nu_j - x_i| = \sum_{x_i \in C_j} \frac{(\nu_j - x_i)}{\sqrt{(x_i - \nu_j)^2}}$$

The derivative can take the following values:

$$\frac{\partial}{\partial \nu_j} \sum_{x_i \in C_j} |\nu_j - x_i| = \begin{cases} +1 & \text{if } \nu_j > x_i \\ -1 & \text{if } \nu_j < x_i \\ \text{undefined} & \text{if } \nu_j = x_i \end{cases}$$

The objective function is not differentiable when $\nu_j = x_i$ since the derivative becomes undefined. For all other values the derivative is either +1 (when $\nu_j > x_i$) or -1 when ($\nu_j < x_i$). Hence, the derivative indicates how many x_i's are smaller than ν_j.
To minimize the sum of absolute errors, we need to find the value of ν_j for which the derivative takes the value zero. It can do so if there are equal number of x_i's that are smaller and larger than ν_j (for even number of x_i's). If there is an odd number of x_i's then the derivative is -1 left of the median value and +1 right of it, hence ν_j should be the median points in C_j.

Problem 3.

1. Consider the K-Means algorithm. For the points of the figure in Problem 1, if we select three starting points using the methods described in the lecture, and the first point we choose is (3,4), which other points are selected.
2. For the three clusters in the figure
   (a) Compute the representation of the cluster as in the BFR Algorithm. That is, compute N, SUM, and SUMSQ.
   (b) Compute the variance and standard deviation of each cluster in each of the two dimensions.

Solution:

1. The starting points are $(3, 4)$ $(7, 10)$ $(12, 6)$
2. (a) Cluster 1: $N = 4$, SUM=(21,36), SUMSQ=(117,328)
   Cluster 2: $N = 3$, SUM=(10,8), SUMSQ=(38,24)
   Cluster 3: $N = 5$, SUM=(54,21), SUMSQ=(590,95)

(b) Cluster 1: Variance in x = 1.68, SD = 1.29; Variance in y = 1, SD = 1
Cluster 2: Variance in x = 1.55, SD = 1.24; Variance in y = 0.39, SD = 0.94
Cluster 3: Variance in x = 1.36, SD = 1.16; Variance in y = 1.36, SD = 1.16

Problem 4. Give an example of a dataset and a selection of k initial centroids such that when the points are reassigned to their nearest centroid at the end, at least one of the initial k points is reassigned to a different cluster.

Solution:
Let the 2 initial cluster centroids be the green and red dot. Notice how the red dot is assigned to a different cluster at iteration 2.