Week 8 Assignments
Topic: Clustering

1. Suppose you run K-means clustering algorithm on a given dataset. What are the factors on which the final clusters depend on ?
   I. The value of K
   II. The initial cluster seeds chosen
   III. The distance function used.

   A) I only
   B) II only
   C) I and II only
   D) I, II and III

K-means clustering algorithm depends on all the three predefined factors.

2. Which of the following statements are true about the different types of linkages.
   A. single linkage suffers from chaining.
   B. Average linkage suffers from crowding.
   C. In single linkage clustering the similarity between two clusters depends on all the elements in the two clusters.
   D. Complete linkage avoids chaining but suffers from crowding.

Single linkage suffers from chaining and complete linkage avoids chaining but suffers from crowding.

3. Consider agglomerative hierarchical clustering and proceeding to a stage where every cluster has at least two points. You may be using single-link or complete-link hierarchical clustering. Is it possible for a point to be closer to points in other clusters than to points in its own cluster in some or all of these methods? Mark the methods where this can happen.

1. It is not possible in either method.
2. Only in single-link clustering
3. Only in complete-link clustering
4. **In both single-link and complete-link clustering**

Please refer lecture notes.

4. Choose **ALL** the statements that are true for hierarchical agglomerative clustering

A) The number of clusters need to be pre-specified.
B) **The output of the clustering algorithm depends on the choice of the similarity metric.**
C) The number of merge operations depends on the number of clusters desired.
D) The number of merge operations depends on the characteristics of the data set.

Please refer to the lecture notes.

5. We would like to cluster the natural numbers from 1 to 1024 into two clusters using hierarchical agglomerative clustering. We will use Euclidian distance as our distance measure. We break ties by merging the clusters in which the lowest natural number resides. For example, if the distance between clusters A and B is the same as the distance between cluster C and D, we would choose A and B as the next clusters to merge if $min|A,B| < min|C,D|$, where {A,B} are the set of natural numbers assigned to clusters A and B. For complete linkage clustering method specify the number of elements assigned to each of the clusters obtained by cutting the dendrogram at the root. In complete linkage clustering the distance between two clusters is the distance of the farthest members of the clusters.

A) 1,1023
B) 512, 512
C) 1022,2
D) None of these

Hints. The cluster assignments of the numbers will be like: ((1,2), (3,4)...

∴ The number of elements assigned to each of the final cluster will be: 512, 512.

6. Which of the following is not a clustering approach

A) Partitioning
B) Hierarchical
C) Density-based
D) **Bagging**

Refer lecture notes.

7. Which of the following options is a measure of internal evaluation of a clustering algorithm?

A) Rand index
B) Davies-Bouldin index
C) Jaccard index
D) F-measure

Refer video lectures.

8. Which among the following is/are some of the assumptions made by the k-means algorithm (assuming Euclidean distance measure)?

A) Clusters are spherical in shape
B) Clusters are of similar sizes
C) Data points in one cluster are well separated from data points of other clusters
D) There is no wide variation in density among the data points

Refer lecture notes.

9. We are given the following four data points in two dimension: x_1 = (2,2), x_2 =(8,6), x_3 =(6,8), $x_4 = (2,4)$. We want to cluster the data points into two clusters C_1 and C_2 using the K-Means algorithm. Euclidean distance is used for clustering. To initialize the algorithm we consider C_1 ={ x_1, x_3} and C_2 ={x_2, x_4}. After two iteration of the K-means algorithm, the cluster memberships are:

a. C_1 = {x_1 , x_2} and C_2 = {x_3, x_4}
b. C_1 = {x_1 , x_4} and C_2 = {x_2, x_3}
c. C_1 = {x_1 , x_3} and C_2 = {x_2, x_4}
d. None of these.

Hints. Before iteration 1 : cluster center for C1 = (4,5)
cluster center for C2 = (5,5).

Calculate distance (Euclidean) from all the data points to cluster center C1 and C2.

After iteration 1 : new cluster center for C1 = (2,3)
new cluster center for C2 = (7,7).

Same way after iteration 2, C1 = {x_1, x_4} and C2 = {x_2, x_3}.

10. With respect to k-means clustering, which of the following are the correct descriptions of the expectation (E) and maximization (M) steps respectively?

A) E-step: assign points to nearest cluster center, M-step: estimate model parameters that maximize the likelihood for the given assignment of points.

B) E-step: estimate model parameters that maximize the likelihood for the given assignment of points, M-step: assign points to nearest cluster center.

C) None of A or B.

D) Both A and B

Refer lecture notes.