Data Mining: Assignment Week 7: Clustering

1. A good clustering is one with_____?

A. Low inter-cluster distance and low intra-cluster distance

B. Low inter-cluster distance and high intra-cluster distance

C. High inter-cluster distance and low intra-cluster distance

D. High inter-cluster distance and high intra-cluster distance

Ans: C

Explanation: A good clustering technique is one which produces high quality clusters in which intra-cluster similarity (i.e. intra cluster distance) is low and the inter-cluster similarity (i.e. inter cluster distance) is high.

2. The leaves of a dendrogram in hierarchical clustering represent?

A. Individual data points

B. Clusters of multiple data points

C. Distances between data points

D. Cluster membership value of the data points

Ans: A

Explanation: Refer to Dendrogram usage in HAG clustering.

3. Which of the following is a hierarchical clustering algorithm?

A. Single linkage clustering

B. K-means clustering

C. DBSCAN

D. None of the above

Ans: A

Explanation: single-linkage clustering is one of several methods of hierarchical clustering. It is based on grouping clusters in bottom-up fashion (agglomerative clustering), at each step combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other.

4. Which of the following is not true about the DBSCAN algorithm?

A. It is a density based clustering algorithm

B. It requires two parameters MinPts and epsilon

C. The number of clusters need to be specified in advance

D. It can produce non-convex shaped clusters

Ans: C

Explanation: Density-based spatial clustering of applications with noise (DBSCAN) is a density-based clustering non-parametric algorithm. DBSCAN requires two parameters: ε (epsilon) and the minimum number of points required to form a dense region (minPts).

5. Which of the following clustering algorithm uses a minimal spanning tree concept?

A. Complete linkage clustering

B. Single linkage clustering

B. Average linkage clustering

C. DBSCAN

Ans: B

Explanation: The naive algorithm for single-linkage clustering has time complexity $O(n^3)$. An alternative algorithm is based on the equivalence between the naive algorithm and Kruskal's algorithm for minimum spanning trees. Instead of using Kruskal's algorithm, Prim's algorithm can also be used.

6. Distance between two clusters in single linkage clustering is defined as:

A. Distance between the closest pair of points between the clusters

B. Distance between the furthest pair of points between the clusters

C. Distance between the most centrally located pair of points in the clusters

D. None of the above

Ans: A

Explanation: Mathematically, the linkage function – the distance $D(X,Y)$ between clusters $X$and Y is described by the expression:

D(X,Y) = min d(x,y) s.t. x ϵ X and y ϵ Y where X and Y are any two sets of elements considered as clusters, and $d(x,y)$ denotes the distance between the two elements x and y.

7. Distance between two clusters in complete linkage clustering is defined as:

A. Distance between the closest pair of points between the clusters

B. Distance between the furthest pair of points between the clusters

C. Distance between the most centrally located pair of points in the clusters

D. None of the above

Ans : B

Explanation: Mathematically, the linkage function – the distance $D(X,Y)$ between clusters X and Y is described by the expression:

D(X,Y) = max d(x,y) s.t. x ϵ X and y ϵ Y where X and Y are any two sets of elements considered as clusters, and $d(x,y)$ denotes the distance between the two elements x and y.

8. Consider a set of five 2-dimensional points $p_1=(0, 0)$, $p_2=(0, 1)$, $p_3=(5, 8)$, $p_4=(5, 7)$, and $p_5=(0, 0.5)$. Euclidean distance is the distance function used. Single linkage clustering is used to cluster the points into two clusters. The clusters are:

A. {p_1, p_2, p_3} {p_4, p_5}

B. {p_1, p_4, p_5} {p_2, p_3}

C. {p_1, p_2, p_5} {p_3, p_4}

D. {p_1, p_2, p_4} {p_3, p_5}

Ans : C

Explanation: find the Euclidean distance between the points and cluster together points having minimum Euclidean distance.

	P1	P2	P3	P4	P5
P1	0				
P2	1	0			
P3	9.4	8.602	0		
P4	8.602	7.81	1	0	
P5	**0.5**	**0.5**	9.01	8.2	0

{P1, P5} and {P2, P5} has minimum distance. We will choose {P1, P5} and cluster them together.

We will evaluate the distance of all the points from the cluster {P1, P5}. Taking minimum distance.

	P1, P5	P2	P3	P4
P1, P5	0			
P2	**0.5**	0		
P3	9.01	8.602	0	
P4	8.2	7.81	1	0

(P1, P5) and P2 has minimum distance. We will cluster them together.

	P1, P2, P5	P3	P4
P1, P2, P5	0		
P3	8.602	0	
P4	7.81	**1**	0

(P3, P4) has minimum distance. They will be clustered together.

We have got two clusters the process of clustering stops.

Two clusters obtained are **{P1, P2, P5} and {P3, P4}**.

9. Which of the following is not true about K-means clustering algorithm?

A. It is a partitional clustering algorithm

B. The final cluster obtained depends on the choice of initial cluster centres

C. Number of clusters need to be specified in advance

D. It can generate non-convex cluster shapes

Ans: D

Explanation: K-means clustering cannot generate non-convex cluster shapes.

10. Consider a set of five 2-dimensional points $p_1=(0, 0)$, $p_2=(0, 1)$, $p_3=(5, 8)$, $p_4=(5, 7)$, and $p_5=(0, 0.5)$. Euclidean distance is the distance function. The k-means algorithm is used to cluster the points into two clusters. The initial cluster centers are p_1 and p_4. The clusters after two iterations of k-means are:

A. $\{p_1, p_4, p_5\}$ $\{p_2, p_3\}$

B. $\{p_1, p_2, p_5\}$ $\{p_3, p_4\}$

C. $\{p_3, p_4, p_5\}$ $\{p_1, p_2\}$

D. $\{p_1, p_2, p_4\}$ $\{p_3, p_5\}$

Ans: B

Explanation: 1st iteration

Initial centres are P1 and P4

	c1 =P1= (0,0)	c2 =P4= (5,7)	Closest Centre
P1	0	8.602	c1
P2	1	7.81	c1
P3	9.4	1	c2
P4	8.602	0	c2
P5	0.5	8.2	c1

2nd iteration

Clusters after 1st iteration are:
C1 = {P1, P2, P5} cluster centre is c1= (0, 0.5)
C2 = {P3, P4} cluster centre is c2= (5, 7.5)

	c1= (0, 0.5)	c2= (5, 7.5)	Closest centre
P1	0.5	9.01	c1
P2	0.5	8.2	c1
P3	9.01	0.5	c2
P4	8.2	0.5	c2
P5	0	8.6	c1

Clusters formed after 2nd iteration are {P1, P2, P5} and {P3, P4}.