NPTEL Online Certification Courses

Indian Institute of Technology Kharagpur

Course Name: Introduction to Machine Learning

Assignment – Week 8 (Clustering)

TYPE OF QUESTION: MCQ/MSQ

Number of Question: 7 **Total Marks: 7x2 = 14**

1. For two runs of K-Mean clustering is it expected to get same clustering results?

   A) Yes
   B) No

Answer: (B)

K-Means clustering algorithm instead converses on local minima which might also correspond to the global minima in some cases but not always. Therefore, it's advised to run the K-Means algorithm multiple times before drawing inferences about the clusters.
However, note that it's possible to receive same clustering results from K-means by setting the same seed value for each run. But that is done by simply making the algorithm choose the set of same random no. for each run.

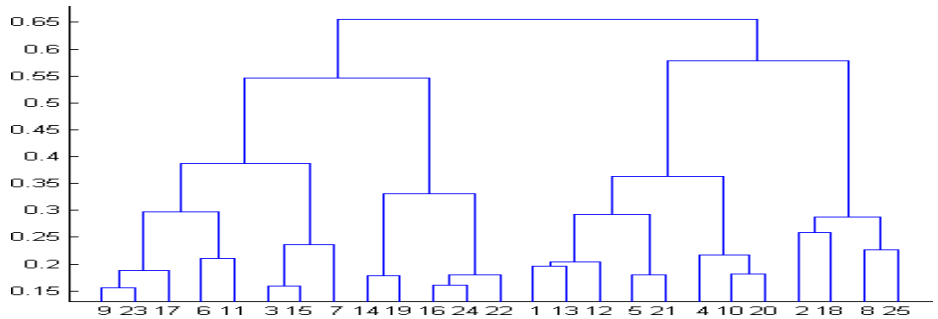
2. Which of the following can act as possible termination conditions in K-Means?

   I. For a fixed number of iterations.
   II. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
   III. Centroids do not change between successive iterations.
   IV. Terminate when RSS falls below a threshold

   A) I, III and IV
   B) I, II and III
   C) I, II and IV
   D) **All of the above**

Answer: D

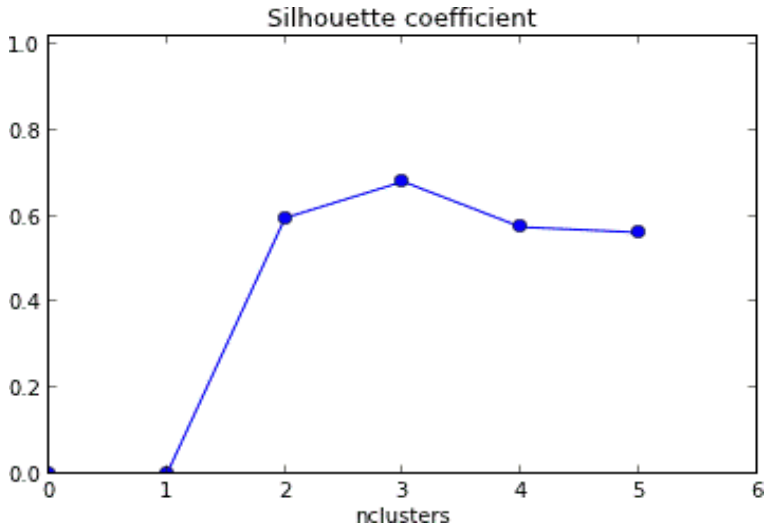
3. After performing K-Means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusion can be drawn from the dendrogram?



A) There were 28 data points in clustering analysis.
B) The best no. of clusters for the analysed data points is 4.
C) The proximity function used is Average-link clustering.
D) The above dendrogram interpretation is not possible for K-Means clustering analysis.

Answer: A dendrogram is not possible for K-Means clustering analysis. However, one can create a cluster gram based on K-Means clustering analysis.

4. What should be the best choice of no. of clusters based on the following results:



A) 1
B) 2
C) **3**
D) 4

Answer: C
The silhouette coefficient is a measure of how similar an object is to its own cluster compared to other clusters. Number of clusters for which silhouette coefficient is highest represents the best choice of the number of clusters.

5. Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

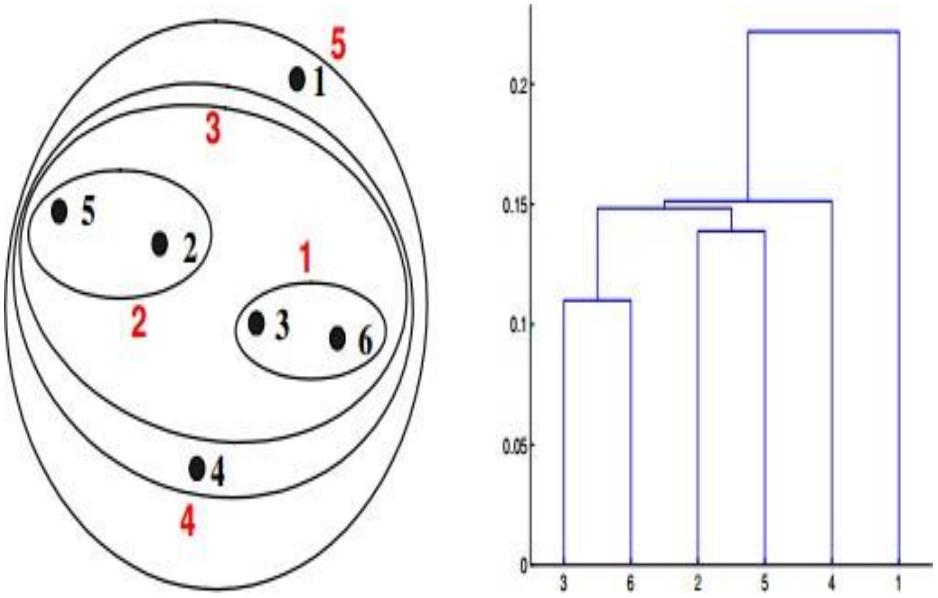
Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

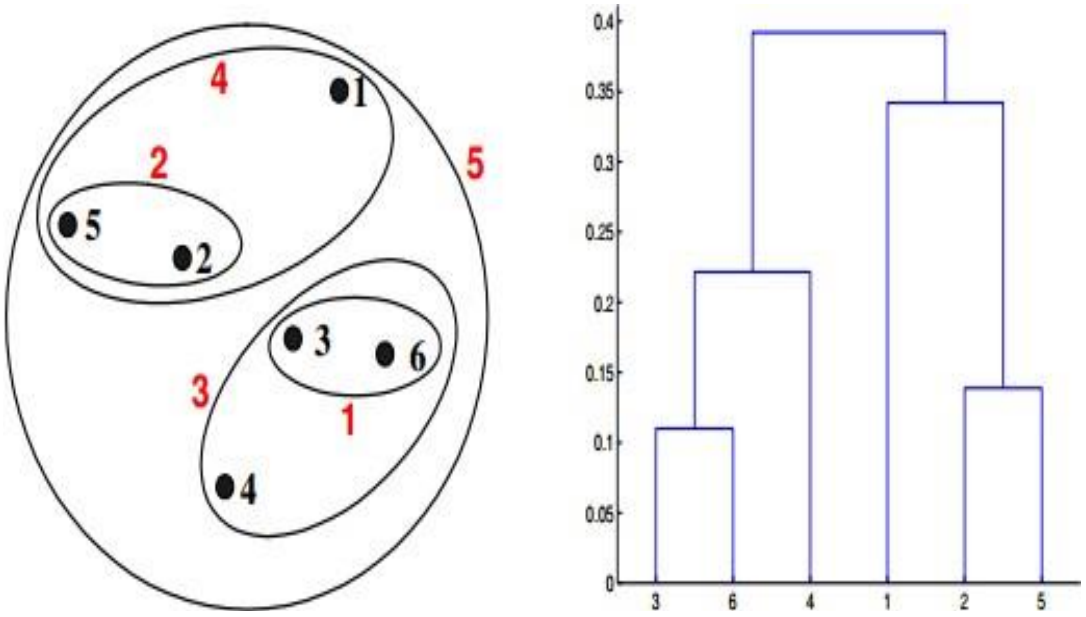
Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

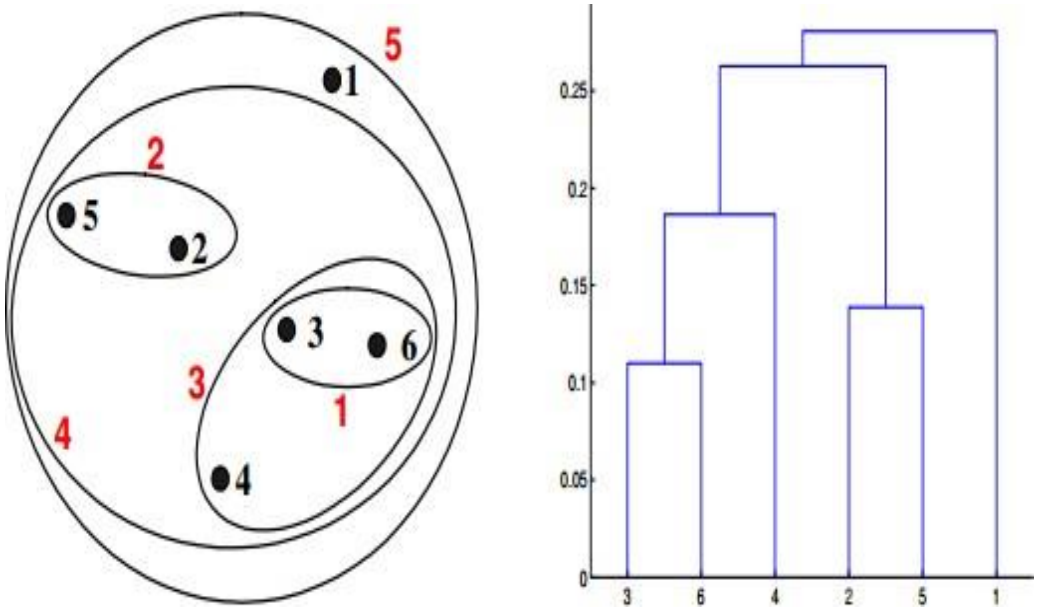
A)



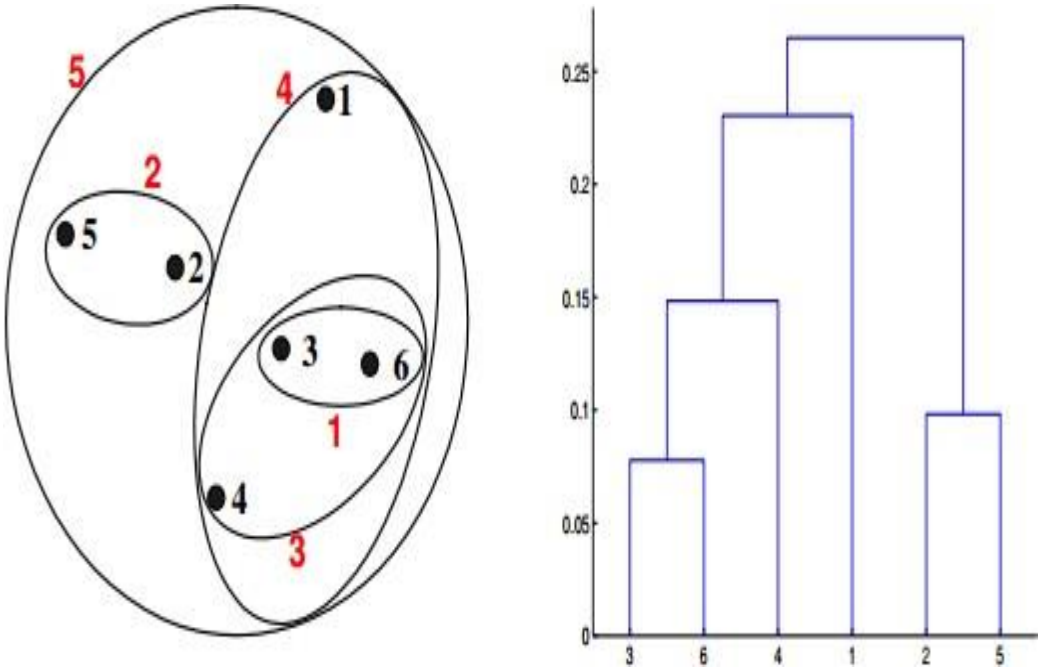
B)


4
1
2
5
5
2
3
6
3
1
4
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
3
6
4
1
2
5

C)


5
1
2
5
2
3
6
3
1
4
4
0.25
0.2
0.15
0.1
0.05
0
3
6
4
2
5
1

D)



Solution: A)

Answer: For the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined to be the minimum of the distance between any two points in the different clusters. For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram. As another example, the distance between clusters {3, 6} and {2, 5} is given by dist ({3, 6}, {2, 5}) = min (dis (3, 2), dist (6, 2), dist (3, 5), dist (6, 5)) = min (0.1483, 0.2540, 0.2843, 0.3921) = 0.1483.

6. Which of the following algorithms are most sensitive to outliersWhatWh?

A) K-means clustering
B) K-medians clustering
C) K-modes clustering
D) K -medoids clustering

Answer: A
K-means is the most sensitive because it uses the mean of the cluster data points to find the cluster center.

7. What is the possible reason(s) for producing two different dendograms using agglomerative clustering for the same data set?

A) Proximity function
B) No. of data points
C) Variables used
D) All of these

Answer: E
Change in either of the proximity function, no of variables used and data points will change the dendograms.