Assignment 10 (Sol.)

Introduction to Machine Learning
Prof. B. Ravindran

1. 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

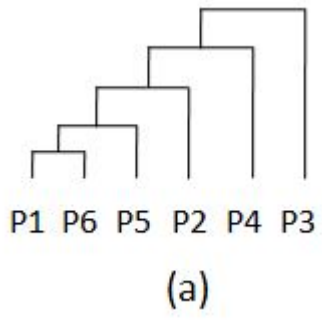
   (d) There is no wide variation in density among the data points

   Sol. (a) & (b)
   The Euclidean distance measure ensures that areas around a cluster centroid comprising points closest to that centroid (which is a cluster) is spherical in shape. Also, this particular distance measure prevents arbitrarily sized clusters since this typically violates the clustering criterion.

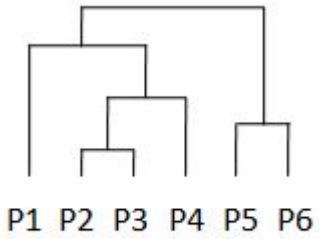
2. Consider the similarity matrix given below.

	P1	P2	P3	P4	P5	P6
P1	1.00	0.70	0.65	0.40	0.20	0.05
P2	0.70	1.00	0.95	0.70	0.50	0.35
P3	0.65	0.95	1.00	0.75	0.55	0.40
P4	0.40	0.70	0.75	1.00	0.80	0.65
P5	0.20	0.50	0.55	0.80	1.00	0.85
P6	0.05	0.35	0.40	0.65	0.85	1.00

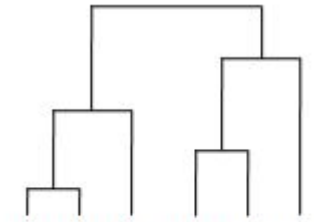
   Show the hierarchy of clustering created by the single-link clustering algorithm.



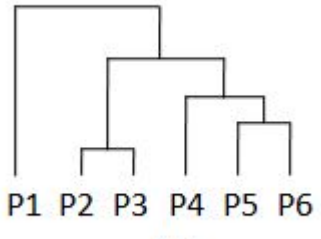
(a)



(b)



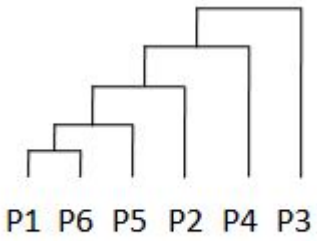
(c)



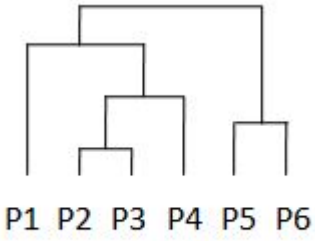
(d)

Sol. (d)
Note that the matrix shows similarity between points and not distance. Thus, high values between two points indicate that they are more similar/closer together.

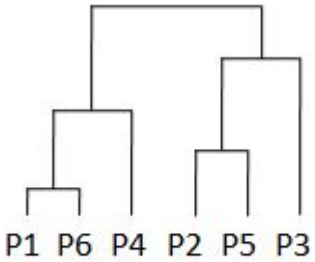
3. For the above similarity matrix, show the hierarchy of clustering obtained on performing complete-link clustering.



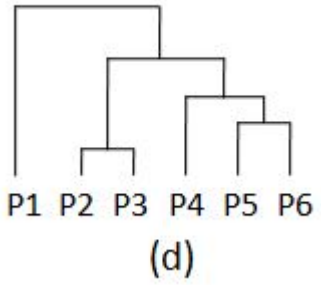
(a)



(b)



(c)



(d)

Sol. (b)

4. Considering single-link and 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? If so, in which approach will this tend to be observed?

   (a) No

   (b) Yes, single-link clustering

   (c) Yes, complete-link clustering

   (d) Yes, both single-link and complete-link clustering

   Sol. (d)
   This is possible in both single-link and complete-link clustering. In the single-link case, an example would be two parallel chains where many points are closer to points in the other chain/cluster than to points in their own cluster. In the complete-link case, this notion is more intuitive due to the clustering constraint (measuring distance between two clusters by the distance between their farthest points).

5. A graph is said to be k-connected if there does not exist a set of k-1 vertices whose removal disconnects the graph. If we define clusters as comprising of k-connected components of the thresholded graphs, does this result in a well-defined clustering algorithm?

   (a) Yes

   (b) No

   Sol. (a)
   In normal hierarchical clustering, as the threshold value is relaxed (increased or decreased depending upon whether we are using distance or similarity as a metric) more edges are added to the threshold graph. However, it is crucial that on adding edges, the clusters identified before do not break up. In the case of defining clusters as comprising k-connected components of the threshold graph, for a particular value of k, as more edges are added to the threshold graph, existing clusters may merge, but do not break up. Thus, this results in a well-defined clustering algorithm.

6. A set of nodes forms a p-cluster, if at least p percentage of the edges from the nodes in the set go to another node in the set. If we define clusters as comprising of p-clusters of the thresholded graphs, does this result in a well-defined clustering algorithm?

   (a) Yes

   (b) No

   Sol. (b)
   Suppose we have a p value of 0.5. Consider a set of points Q which form a cluster at a particular threshold level. This means that at least 50% of all edges in the set Q go to nodes in the same set. Now, if we relax the threshold, adding more edges, it may turn out that for the set of points Q, 50% of the edges no longer go to nodes in the same set (this can happen if a lot of the edges added to the nodes of set Q link to nodes outside the set). Thus, as the threshold level is relaxed, it is possible that points in an existing cluster are assigned to separate clusters. This goes against the notion of hierarchical clustering, and hence this definition of clustering is not valid.

7. In the CURE clustering algorithm, representative points of a cluster are moved a fraction of the distance between their original location and the centroid of the cluster. Would it make more sense to move them all a fixed distance towards the centroid instead? Why or why not?

   (a) Yes, because this approach will ensure that the original cluster shape is preserved.

   (b) No, because this approach will not be as effective against outliers as the original approach.

   Sol. (b)
   Moving representative points towards the cluster centroid helps in overcoming the effects of outliers. In the proposed approach, moving only a fixed distance towards the centroid would be less effective against outliers, since the distance between the outliers and the centroid may be much larger than the distance between the other points and the centroid.

8. Suppose while performing DBSCAN we randomly choose a point which has less than MinPts number of points in its neighbourhood. Which among the following is true for such a point?

   (a) It is treated as noise, and not considered further in the algorithm

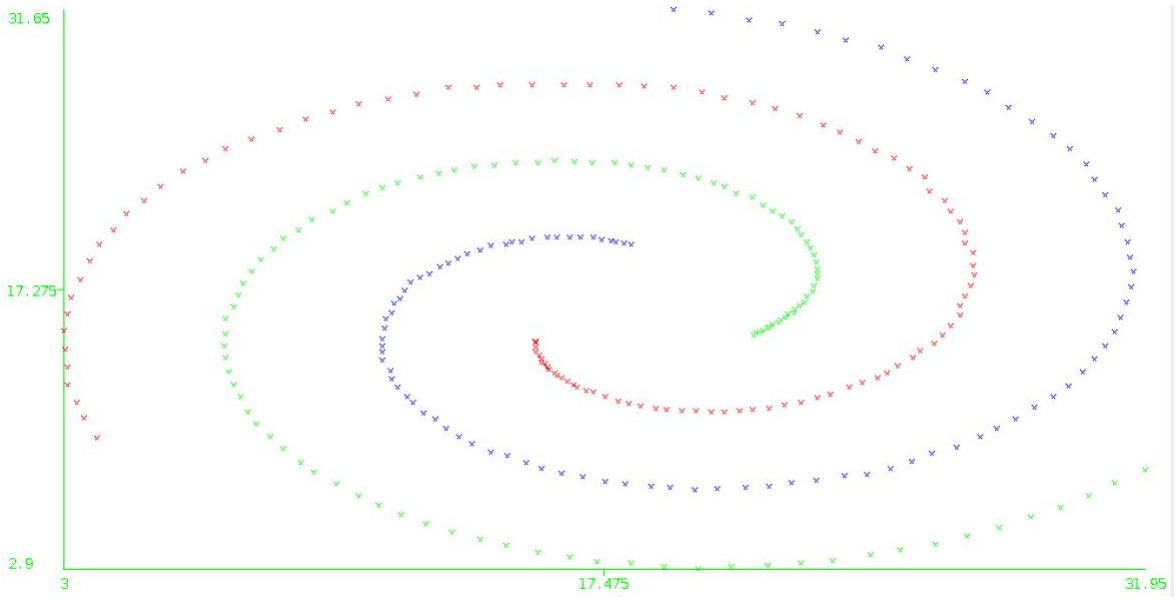
   (b) It becomes part of its own cluster

   (c) Depending upon other points, it may later turn out to be a core point

   (d) Depending upon other points, it may be density connected to other points

   Sol. (d)
   Since there are less than MinPts number of points in its neighbourhood, it is not a core point. However, it is not necessarily a noise point, since if there exists a core point in whose neighbourhood this point lies in, it can be a boundary point.

9. Consider the following image showing data points belonging to three different clusters (indicated by the colours of the points). Which among the following clustering algorithms will perform well in accurately clustering the given data?



(a) K-means

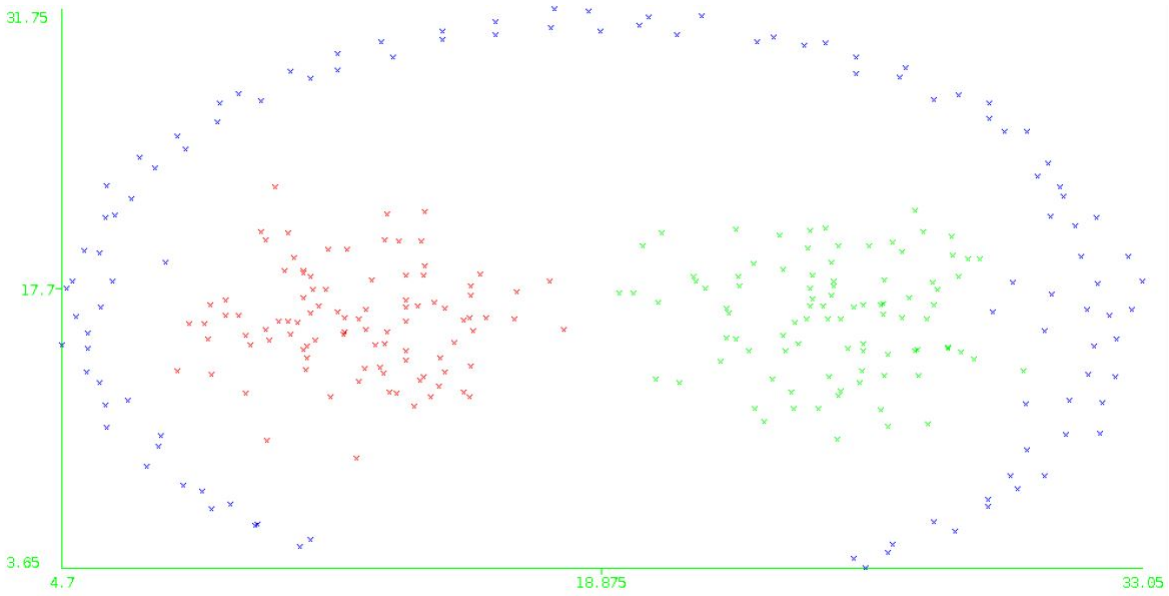
(b) Single-link hierarchical

(c) Complete-link hierarchical

(d) DBSCAN

Sol. (b) & (d)
K-means will clearly not work for this data set. Complete-link clustering will also not be able to recover the desired clustering since in the spiral structure in which the data points lie, points in the same cluster are actually quite far from other points in their own clusters. Single-link clustering is ideally suited for this data set as well as DBSCAN, since there is enough distance between points belonging to the different clusters.

10. Consider the following image showing data points belonging to three different clusters (indicated by the colours of the points). Which among the following clustering algorithms will perform well in accurately clustering the given data?



(a) K-means

(b) Single-link hierarchical

(c) Complete-link hierarchical

(d) DBSCAN

Sol. (d)
K-means and complete-link hierarchical clustering do not do well as in the previous question. The problem with single-link clustering is that there are a few points which belong to different clusters but are close enough to each other so that it may result in combining different clusters. DBSCAN with appropriate values of MinPts and Eps will be able to overcome the problem posed by such points.