DATA MINING

CS422- Section 01

Illinois Institute of Technology

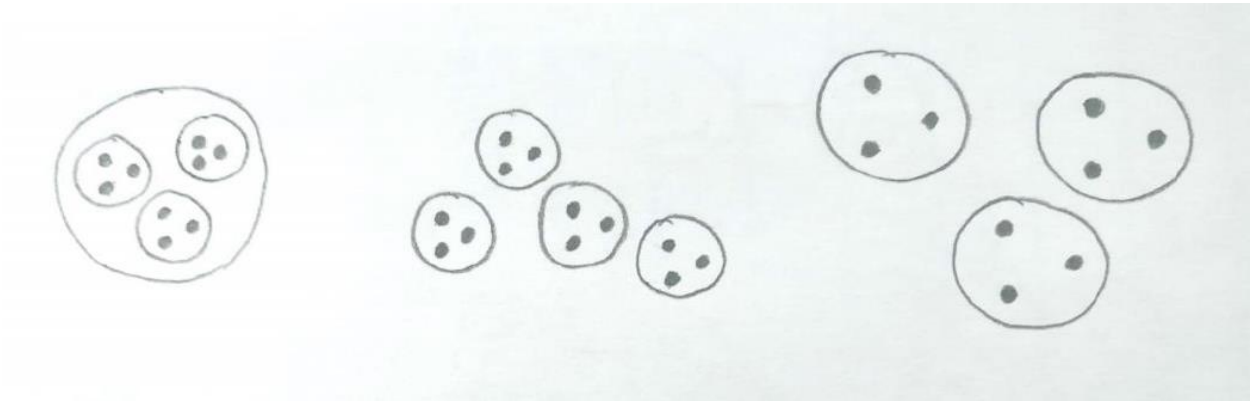
HOMEWORK – 3

Megha Tatti (CWID: A20427027)

Exercise 1:

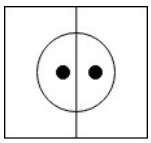
Section 1.1: Chapter 7

Question 2:
Answer:

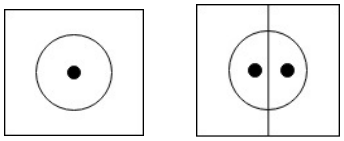


Question 6:
Answer:

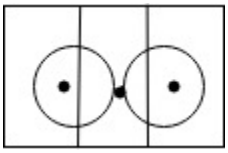
(a) There are many ways to split the circle into k=2(2 clusters). We must take any line that bisects the circle. The centroids will lie on the perpendicular bisector of the line that splits the circle into two clusters and will be symmetrically positioned. All solutions will have the same globally minimal error.



(b) The solution is as shown because of the restriction that the circles are more than one radius apart. The bisector could have any angle, as shown, and it could be the other circle that is split. All these have the same globally minimal error.

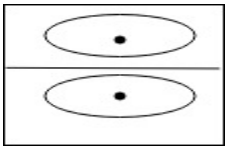


(c) The below shown are the 3 clusters in 3 boxes that will result in the realistic case that the initial centroids are actual data points.

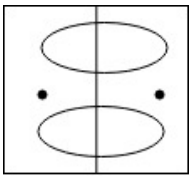


(d) This can be split in ways:

(1) Local minimum: These 2 clusters are in local minimum.



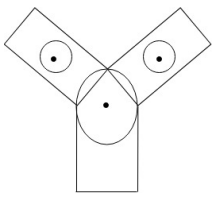
(2) Global minimum: The clusters are split with global minimum. We can see that the centroids are between the 2 clusters.



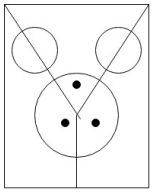
(e) The two top clusters are enclosed in two boxes, while the third cluster is enclosed by the regions defined by a triangle and a rectangle.

Note that while the two pie shaped cuts out of the larger circle are shown as meeting at a point, this is not necessarily the case—it depends on the exact positions and sizes of the circles. There could be a gap between the two pie shaped cuts which is filled by the third (larger) cluster. Or the boundary between the two pie shaped cuts could actually be a line segment.

Global minimum:



Local minimum:



Question 11:

Ans:

- If the SSE of one attribute is low for all clusters, then the variable is a constant and of little use in dividing the data into groups.
- If the SSE of one attribute is low for just one cluster, then this attribute helps to define the cluster.
- If the SSE of an attribute is high for all clusters, then it could mean that the attribute is noise.
- If the SSE of an attribute is high for one cluster, then it is at odds with the information provided by the attributes with low SSE that define the cluster.
- We can eliminate attributes that have poor distinguishing power between clusters, i.e., low or high SSE for all clusters, since they are useless for clustering. The attributes with high SSE for all clusters are particularly troublesome if they have a relatively high SSE with respect to other attributes since they introduce a lot of noise into the computation of the overall SSE.

Question 12:

Ans: **(a)** Advantages of leader algorithm as compared to K-means:

- The leader algorithm requires only a single scan of the data and is thus more computationally efficient since each object is compared to the final set of centroids at most once.
- Although the leader algorithm is order dependent, for a fixed ordering of the objects, it always produces the same set of clusters.

Disadvantages of leader algorithm as compared to K-means:

- It is not possible to set the number of resulting clusters for the leader algorithm(unlike K-means), except indirectly.
- the K-means algorithm almost always produces better quality clusters as measured by SSE.

(b) The knowledge gained from the process of using a sample to determine the distribution of distances between the points, can be used to set the value of the threshold. The leader algorithm could be modified to cluster for several thresholds during a single pass.

Question 16:
Ans:

	P1	**P2**	**P3**	**P4**	**P5**
P1	1.00	0.10	0.41	0.55	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.85
P4	0.55	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

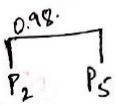
Single-Linkage Cluster

	P1	P2	P3	P4	P5
P1	1.00	0.10	0.41	0.55	0.35
P2	0.10	1.00	0.64	0.47	(0.98)
P3	0.41	0.64	1.00	0.44	0.85
P4	0.55	0.47	0.44	1.00	0.76.
P5	0.35	(0.98)	0.85	0.76	1.00.

Taking the maximum value,

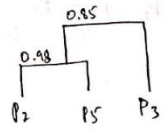
Highest = 0.98 [P_2 and P_5).

Merging both,



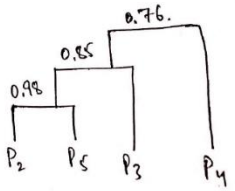
	P1	P2&P5	P3	P4
P1	1.0	0.35	0.41	0.55
P2&P5	0.35	1.0	(0.85)	0.76
P3	0.41	(0.85)	1.0	0.44
P4	0.55	0.76	0.44	1.0.

Highest = 0.85 (P_2, P_5 & P_3)



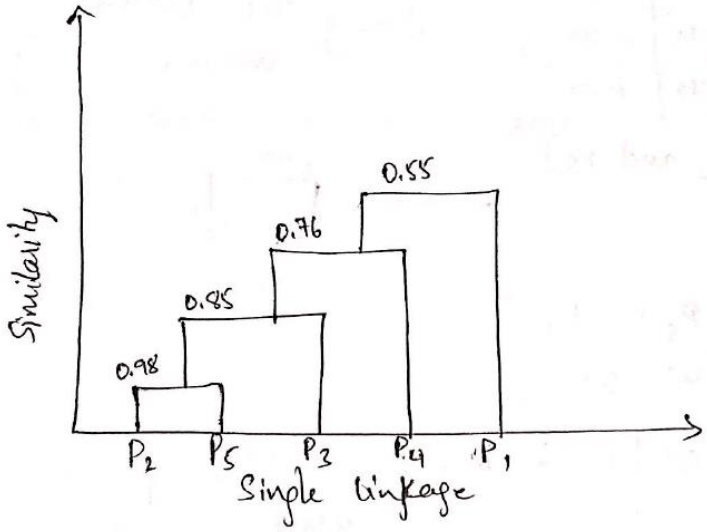
	P1	P2,P5,P3	P4
P1	1.00	0.41	0.55
P2,P5,P3	0.41	1.00	(0.76).
P4	0.55	(0.76)	1.00

Highest = 0.76 (P_2, P_5, P_3 & P_4)



	P_1	P_2, P_5, P_3, P_4
P_1	1.00	0.55
P_2, P_5, P_3, P_4	0.55	1.00

So, the final Single Linkage Cluster is :-

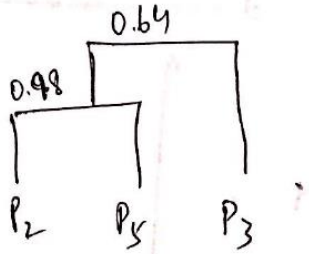


Complete Linkage :-

	P_1	P_2	P_3	P_4	P_5
P_1	1.00	0.10	0.41	0.55	0.35
P_2	0.10	1.00	0.64	0.47	0.98
P_3	0.41	0.64	1.00	0.44	0.85
P_4	0.55	0.47	0.44	1.00	0.76
P_5	0.35	0.98	0.85	0.76	1.00

Highest = 0.98 (P_2 and P_5).

	P_1	P_2 & P_5	P_3	P_4
P_1	1.00	0.10	0.41	0.55
P_2 & P_5	0.10	1.00	0.64	0.47
P_3	0.41	0.64	1.00	0.44
P_4	0.55	0.47	0.44	1.00



Choosing minimum similarity

Highest = 0.64 (P_2, P_5 & P_3)

	P_1	P_2, P_5, P_3	P_4
P_1	1.00	0.10	0.55
P_2, P_5, P_3	0.10	1.00	0.44
P_4	0.55	0.44	1.00

Highest = 0.55 (P_1, P_4)

	P_1, P_4	P_2, P_5, P_3
P_1, P_4	1.00	0.10
P_2, P_5, P_3	0.10	1.00

Complete linkage Clusters