31/03/18

Data Mining
Assignment - 3

Avinash Vellineni
A20406657

Problem 1.1 :-

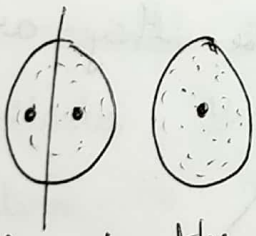
2)

6)

a)

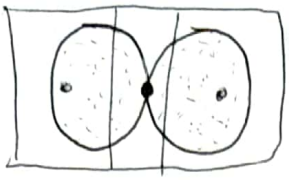
There are infinite number of ways to split a circle. A line is required to bisect the circle. The line makes an angle [0 - 180] with X axis. The centroids lie on the perpendicular bisector of the two semi circles and are symmetrically placed. They have the same global minimum error.

b)



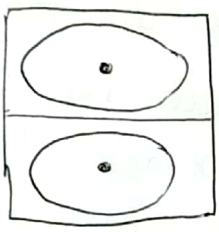
There are infinite number of ways to split a circle. A line is required to bisect the circle. The line makes an angle [0 - 180] with x-axis. The centroids lie on the perpendicular bisector of the two semi circles of the first circle and for the second circle centroid is placed in the centre of the circle They have the same global minimum error. Here the two circles lie slightly greater than the radii of the circles.

c)

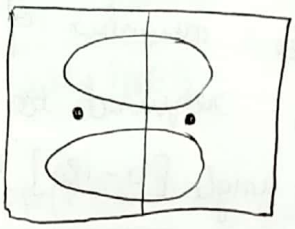


Here the split is made as shown in the figure. Here the distance between the edges of the circles is much less than the radii of the circles. Here the initial centroids are actual data points. Sideway centroids have low error & center has higher error.

d) K=2



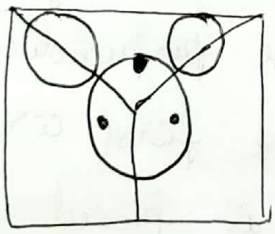
Local minimum



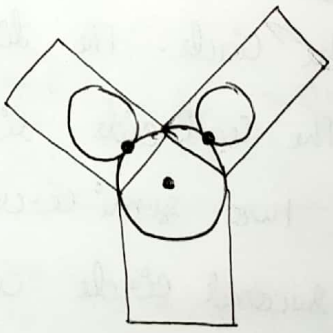
global minimum

There are two possible splits as shown above. In the first diagram, the two clusters are only local minimum, In the second case they are global minimum

e)



Local minimum



global minimum

There are two possible solutions as shown in the figure with local & global minimum respectively.

From the global minimum figure - there are two rectangular symmetrical cluster. the third cluster is in the form of a triangle

11)

If SSE of an attribute is low for all clusters, then it is a constant and gives less information when the observations are divided into groups.

If SSE of an attribute is low for just one cluster then it helps in defining the cluster.

If SSE of an attribute is high for all the clusters then the corresponding attribute could be a noise

If SSE of an attribute is high for only one cluster then the attributes with low SSE defines the cluster. This high SSE attribute is not part of defining the cluster.

Per variable SSE can be used to eliminate attributes that have poor classification between the clusters. As mentioned above that if SSE of an attribute is high for all the clusters then it could be a noise and doesn't help in defining the cluster.

12)

a) Leader algorithm is computationally efficient than K-means. Output of a leader algorithm is fixed, (ie) it produces same clusters everytime when there is fixed ordering of objects. K-means produces different clusters based on the centroid position.

In terms of SSE K-means has better accuracy over leader algorithm.

b) We can use a sample of observations to find the distribution of distance between the points. Using this helps to set the value of the threshold. By using different values of thresholds we can get different clusters.

ii)

Similarity matrix

	P_1	P_2	P_3	P_4	P_5
P_1	1	0.1	0.41	0.55	0.35
P_2	0.1	1	0.64	0.47	0.98
P_3	0.41	0.64	1	0.44	0.85
P_4	0.55	0.47	0.44	1	0.76
P_5	0.35	0.98	0.85	0.76	1

Single linkage:-

highest is 0.98 $P_2 - P_5$

	P_1	P_2-P_5	P_3	P_4
P_1	1			
P_2-P_5	0.35	1		
P_3	0.41	0.85	1	
P_4	0.55	0.76	0.44	1

highest is 0.85

Combine $P_2 - P_5 - P_3$

	P_1	P_2-P_5-P_3	P_4
P_1	1		
P_2-P_5-P_3	0.41	1	
P_4	0.55	0.76	1

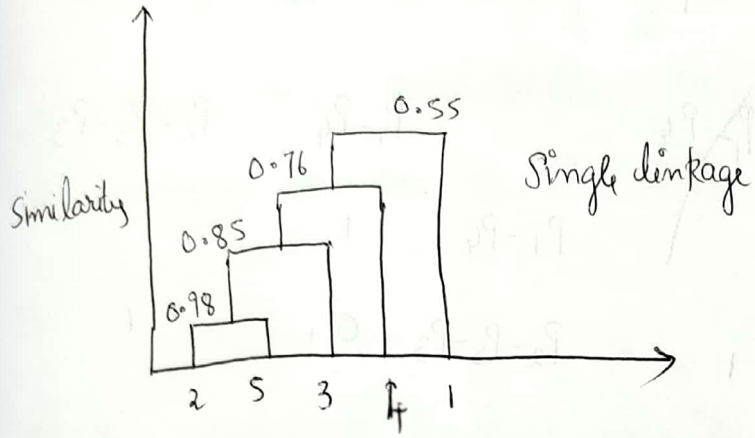
highest is 0.76

join $P_2 - P_5 - P_3 - P_4$

	P_1	P_2-P_5-P_3-P_4
P_1	1	
P_2-P_5-P_3-P_4	0.55	1

So we have

$P_2 - P_5 - P_3 - P_4 - P_1$



Single linkage

Complete linkage:-

	P_1	P_2	P_3	P_4	P_5
P_1	1				
P_2	0.1	1			
P_3	0.41	0.64	1		
P_4	0.55	0.47	0.44	1	
P_5	0.35	**0.98**	0.85	0.76	1

P_2-P_5 highest 0.98

Choosing minimum similarity.

	P_1	P_2-P_5	P_3	P_4
P_1	1			
P_2-P_5	0.1	1		
P_3	0.41	**0.64**	1	
P_4	0.55	0.47	0.44	1

highest is 0.64 so

$P_2 - P_5 - P_3$ ✓

P_1 P_2-P_5-P_3 P_4

0.55

highest is P_1 So P_2-P_5-P_3-P_1

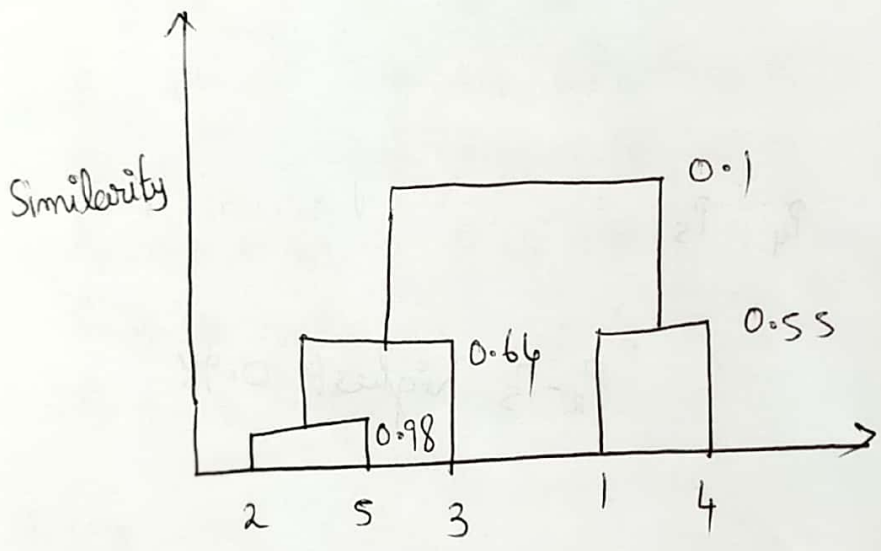
	P_1	P_2-P_5-P_3	P_4
P_1	1		
P_2-P_5-P_3	0.1	1	
P_4	0.55	0.44	1

P_2-P_5-P_3 P_1-P_4

P_1-P_4 P_2-P_5-P_3

	P_1-P_4	P_2-P_5-P_3
P_1-P_4	1	
P_2-P_5-P_3	0.1	1

P_2-P_5-P_3 P_1 P_4



Complete linkage