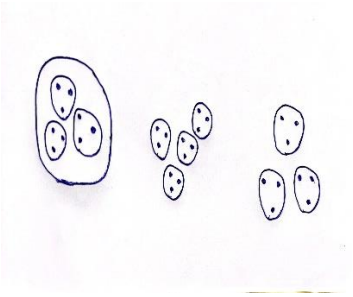
HW-8

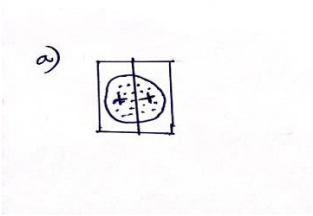
1-1.1-2.



1-1.1-6-a.

Given K=2

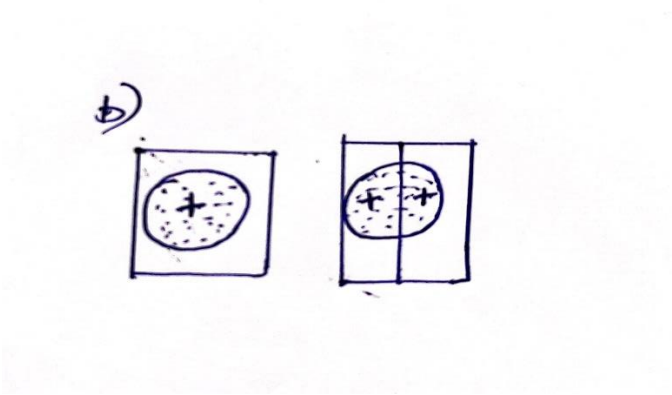
There are many no.of ways to cut the circle, to get two clusters just bisect the circle with a line. The centroids will be located on the perpendicular bisector of the line dividing the circle into two clusters. We can obtain many solutions, all of them has same error and globally minimal.



1-1.1-6-b

Given K=3

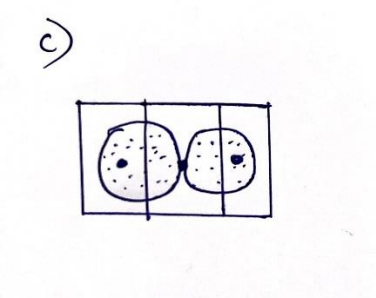
The distance between these circle is more than the radius of circle. We can bisect any circle, in any way we wanted, all possible splits have the same globally minimal error. This how it looks after splitting



1-1.1-6-c

Given K=3

Here there two clusters are close to each other, so we can split in the following way seen below, these three boxes below represents the 3 clusters after the splitting. This how clusters look with 3 centroids.

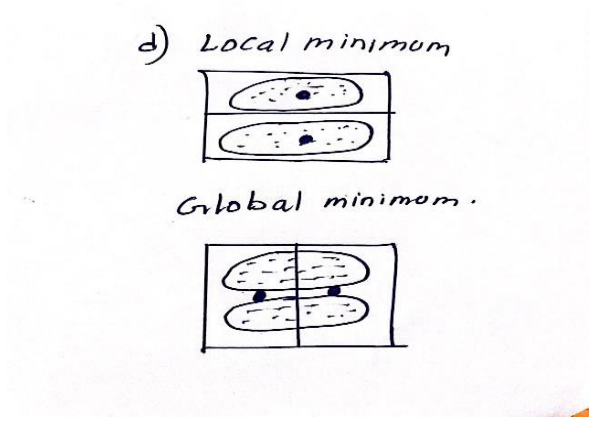


1-1.1-6-d

Given K=2, there are two possible solutions.

Case1: In below both the two circles are splited in a way they are two individual clusters, with the centroid in the center of those circles, which represents the local minimum.

Case2: The bisecting line went through these two circles, by splitting them into half's. And the centroids are placed at the intersection of those circles to the line.

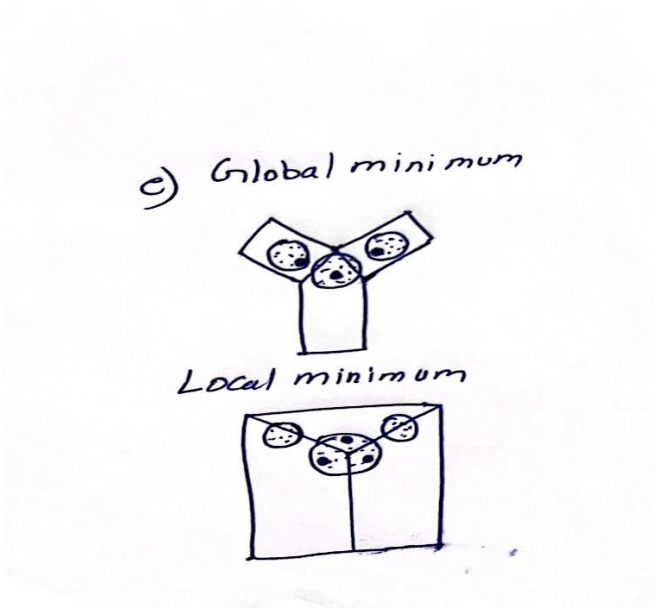


1-1.1-6-e

Given K=3, there are two possible solutions

Case1: The two circles in the top are taken into separate boxes, these small boxes have small portion of large circle, And the big circle is taken into separate box, and we can see the centroids of these clusters in the image below, which represents the Global minimum

Case2: In here the three circles are bisected by the lines, and these are slitted into half's. The centroids are in the big circle and they are close to each other we can see in the image below which represents the Local minimum.



1-1.1-7

Answer: (B)

The more centroids should be allocated to less dense region. The less dense refer to the less similarity in data points, so then more centroids are used to minimize the squared error.

1-1.1-11

What does it mean if the SSE for one variable is low for all clusters?

Answer: The SSE for one variable is low for all clusters, that means the variable is constant, and which may not be helpful while we divide the data into different groups.

Low for just one cluster?

Answer: The SSE is low for one cluster , that means that will be helpful to explain about the cluster.

High for all clusters?

Answer: The SSE is high for all clusters, that means the attribute is noise or outlier, which does not belong to any of the classes.

High for just one cluster?

Answer: The SSE high for one cluster, that means it does not be useful to explain about the cluster, it does not support the info given by the attributes with lower SSE that define the cluster.

How could you use the per variable SSE information to improve your clustering?

The variables with low and high SSE for all clusters are not useful for clustering, The attributes with high SSE with all clusters, these will put lot of noise into overall SSE. We should eliminate the attributes that have poor discriminative power between clusters.

1-1.1-12-a

In leader algorithm we cannot set the number of clusters to be produced, it will produce the same clusters for the given set of data. Its not like K-means we cannot fix with no.of clusters. By comparing both K-means and leader algorithm, the leader algo will produce better clusters as measured by SSE. The leader algorithm is a type of incremental clustering algorithm that is commonly used to cluster large data sets.

1-1.1-12-b

The Leader algorithm is order dependent, different clusters may produce based on the data set order given to the algorithm. In here we use threshold value, if the distance between the two points is less than the threshold then it will be added to cluster, if the distance is grater than the threshold then new cluster will be created. The knowledge gained by observing will help to set better threshold.

1-1.1-16

Similarity Matrix:

	P1	P2	P3	P4	P5
P1	1				
P2	0.10	1			
P3	0.41	0.64	1		
P4	0.55	0.47	0.44	1	
P5	0.35	0.98	0.85	0.76	1

Formula Distance = 1 – similarity

Single Link:

Here, we will get the new table

Distance Matrix

	P1	P2	P3	P4	P5
P1	0				
P2	0.9	0			
P3	0.59	0.36	0		
P4	0.45	0.53	0.56	0	
P5	0.65	0.02	0.15	0.24	0

Here the minimum value is ‘0.02’ (p5,p2)

D(P5,P2 -> P1) = MIN(D(P1 ->P2), D(P1->P5))

D(P5,P2 -> P1) = MIN(0.9,0.65)

D(P5,P2 -> P1)= 0.65

D(P5,P2 -> P3) = MIN(D(P3 ->P2), D(P3->P5))

D(P5,P2 -> P3) = MIN(0.0.36,0.15)

D(P5,P2 -> P3)= 0.15

D(P5,P2 -> P4) = MIN(D(P4 ->P2), D(P4->P5))

D(P5,P2 -> P4) = MIN(0.53,0.24)

D(P5,P2 -> P4)= 0.24

	P1	P2,P5	P3	P4
P1	0			
P2,P5	0.65	0		
P3	0.59	0.15	0	
P4	0.45	0.24	0.56	0

Here the minimum value is 0.15

D(P2,P3,P5 -> P1) = MIN(D(P1 ->P2,P5), D(P1->P3))

D(P2,P4,P5 -> P1) = MIN(0.65,0.59)

D(P2,P4,P5 -> P1)= 0.59

D(P2,P3,P5 -> P4) = MIN(D(P4 ->P2,P5), D(P3->P4))

D(P2,P3,P5 -> P4) = MIN(0.24,0.56)

D(P2,P4,P5 -> P4)= 0.24

	P1	P2,P4,P5	P3
P1	0		
P2,P3,P5	0.59	0	
P4	0.45	0.24	0

Here the minimum is 0.24

D(P2,P3,P4,P5 -> P1) = MIN(D(P1 ->P2,P3,P5), D(P1->P4))

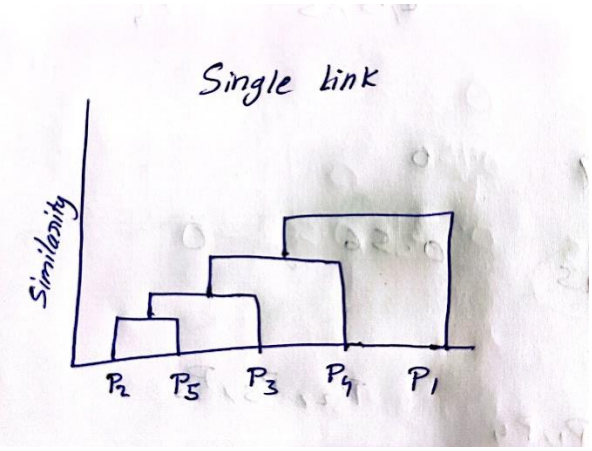
D(P2,P3,P4,P5 -> P1) =min(0.45,0.59)

D(P2,P3,P4,P5 -> P1)= 0.45

	P1	P2,P3,P4,P5
P1	0	
P2,P3,P4,P5	0.45	0

	P1,P2,P3,P4,P5
P1,P2,P3,P4,P5	0

Dendrogram:



Complete Link:

	P1	P2	P3	P4	P5
P1	0				
P2	0.9	0			
P3	0.59	0.36	0		
P4	0.45	0.53	0.56	0	
P5	0.65	0.02	0.15	0.24	0

Here the minimum value 0.02

D(P5,P2 -> P1) = MAX(D(P1 ->P2), D(P1->P5))

D(P5,P2 -> P1) = MAX(0.9,0.65)

D(P5,P2 -> P1)= 0.9

D(P5,P2 -> P3) = MAX(D(P3 ->P2), D(P3->P5))

D(P5,P2 -> P3) = MAX(0.36,0.15)

D(P5,P2 -> P3)= 0.36

D(P5,P2 -> P4) = MAX(D(P4->P2), D(P4->P5))

D(P5,P2 -> P4) = MAX(0.24,0.53)

D(P5,P2 -> P4)= 0.53

	P1	P2, P5	P3	P4
P1	0			
P2, P5	0.9	0		
P3	0.59	0.36	0	
P4	0.45	0.53	0.56	0

D(P2, P3, P5 -> P1) = MAX(D(P1 ->P2, P5), D(P1->P3))

D(P2, P3, P5 -> P1) = MAX(0.9,0.59)

D(P2, P3, P5 -> P1)= 0.9

D(P2, P3, P5 -> P4) = MAX(D(P4 ->P2, P5), D(P4->P3))

D(P2, P3, P5 -> P4) = MAX(0.53,0.56)

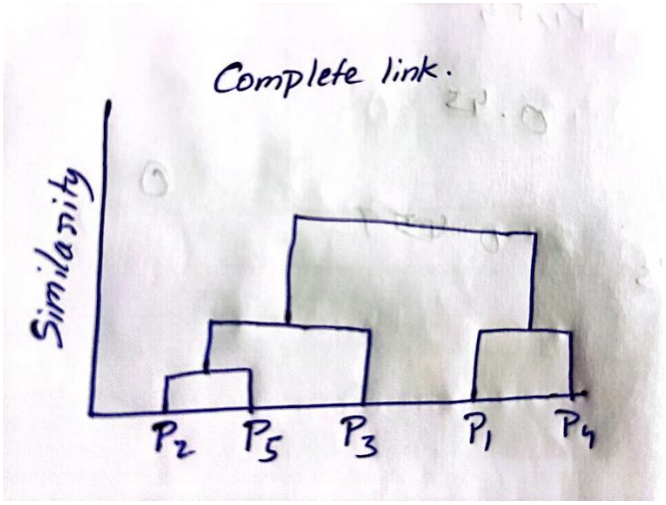
D(P2, P3, P5 -> P4)= 0.56

	P1	P2, P3, P5	P 4
P1	0		
P2, P3, P5	0.9	0	
P4	0.45	0.56	0

Here the minimum is 0.45

	P1, P4	P2, P3, P5
P1, P4	0.45	
P2, P3, P5	0.9	0

	P1,P2,P3,P4,P5
P1,P2,P3,P4,P5	0


Complete link.
Similarity
P2
P5
P3
P1
P4