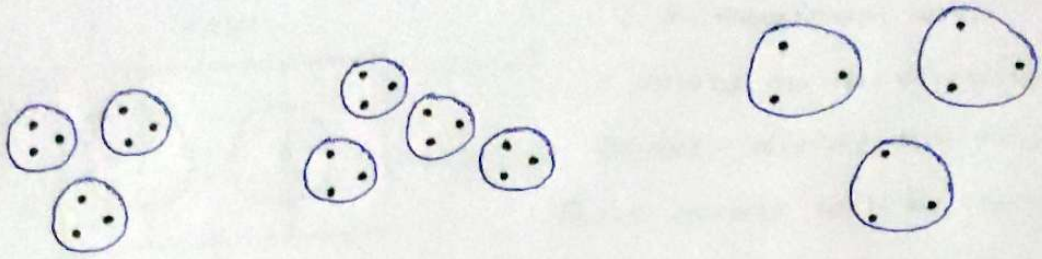
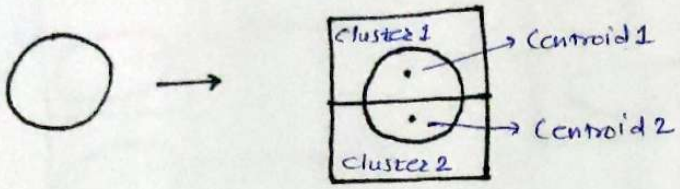
Q2 Find all well-seperated clusters in the set of points shown below

Ans:



Q6. For following set of two dimentional points .....

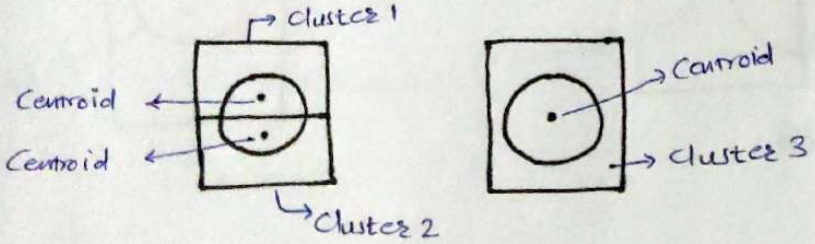
a) K=2 How many possible ways are there



We can partition given circle in infinite way into two clusters. Simple way is to draw diameter and partition circle into two parts will gives us two clusters. As we can draw infinite diameters so there are infinite ways to partition the circle And all the partitions will have almost same global minimum error.

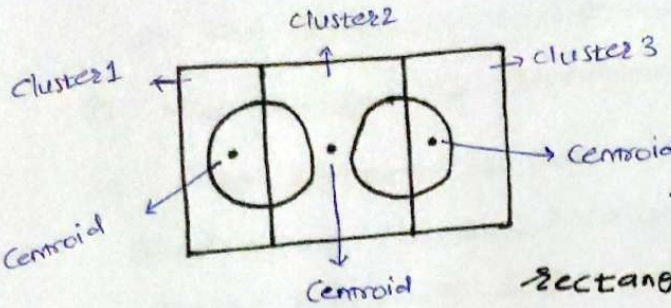
If we draw a perpendicular to a diameter and then midpoint between the centre and the circumference on both the halfs of the circle will be the position of centroids of the cluster.

b) K=3 distance between edges is greater than radii of circles



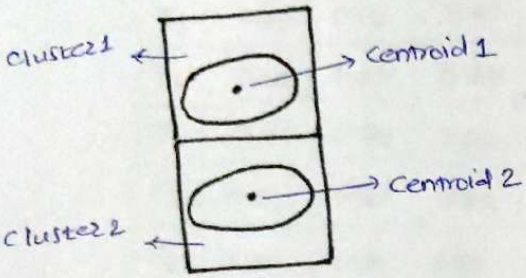
In this case we can make one circle as a cluster and divide another one into two parts. In first case center will become a centroid and in second case midpoint between the center and circumference on both halves of circle will be the position of 2nd & 3rd Centroids so that all the partitions will have same global minimum error.

c) K=3. The distance between the edges of circle is much less than radius of circle.

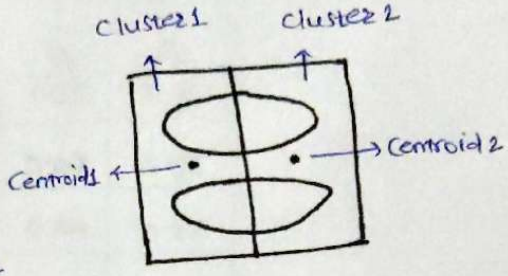


Take maximum distance between the 2 points on the circumference of the circles Equally divide the line into 3 parts. These points will be centroids and 3 rectangles that have these points as the center will be our three clusters.

d) K=2



i)



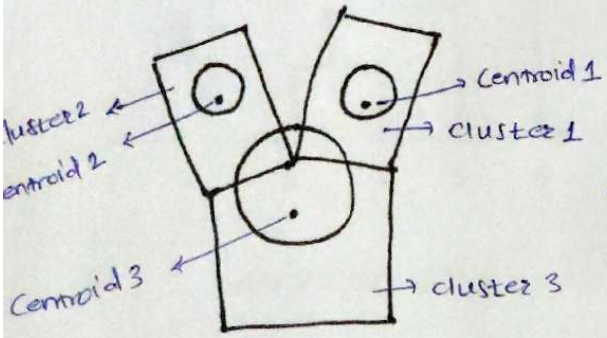
ii)

Above are two solutions for the problem where K=2.

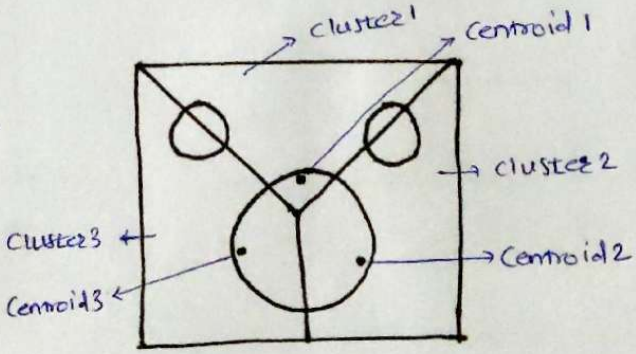
i) Produces local minimum error (left figure)

ii) Right figure will produce global minimum error.

e) K=3 Hint: Use the symmentry of the situation



i)



ii)

There are two solutions to the problem i) produces global minimum error and ii) produces local minimum error.

11) Total SSE is the sum of the SSE .....

i) If a given attribute is constant, If it's SSE is constant for all the clusters. And it has minimal impact on clustering result.

ii) If a given attribute variable has low SSE for just one cluster then that variable dominates in defining the cluster.

iii) If a given attribute variable has high SSE for all the clusters then there is a high probability that this attribute is noise

iv) If a given attribute variable has high SSE for only one cluster then this variable doesn't help in defining the cluster, and the attribute that has low SSE for that cluster dominated this cluster.

v) Per variable SSE information helps us eliminate attribute that has little impact is defining the clusters. Attributes that have low SSE for all the clusters are effectively constant and has a low impact on defining the cluster. Attributes that have high SSE for all the clusters are essentially noise, and they impact overall SSE.

12) Advantages of leader Algorithm:

i) In this algorithm as each object is compared to the final set of centroids so its computationally efficient and its complexity is $O(n)$ where n is number of elements.

ii) It will always return the same result if the order of the input element are same.

Disadvantages:

i) We cannot have cluster number in advance as we have it in k-means. so even if we know the value of k we can't control the algo to make exact number of cluster

2) This simplistic algo does not take SSE into consideration so sum of error is high. k means will almost always have a better result.

ways to improve leadership algo:

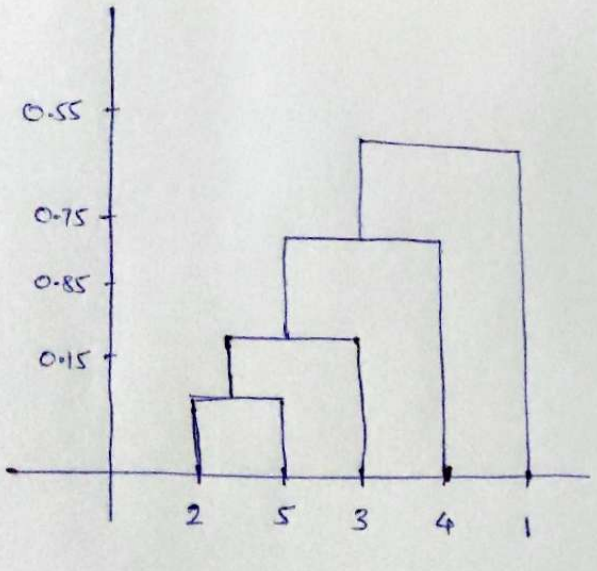
1) Run the algorithm in 1st pass with given threshhold as mean of all the consecutive distances

2) Now change the threshhold to get the predefined k clustures

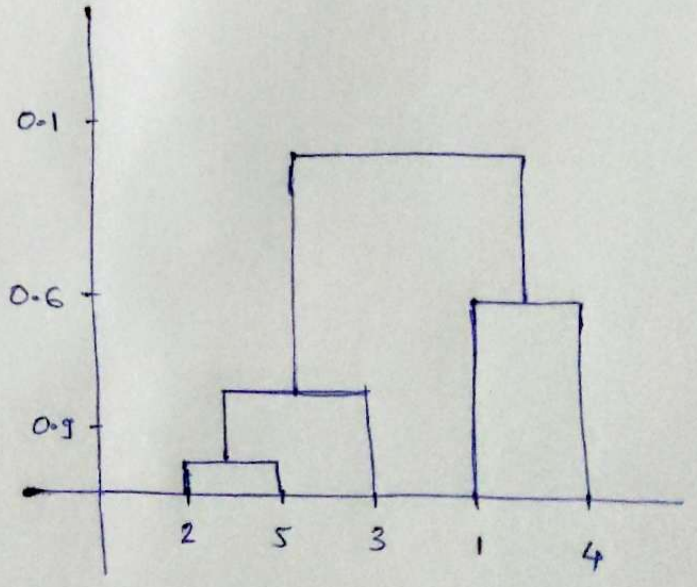
3) Also after each pass we can calculate total SSE and by modifying threshhold values we can check which configuration giving minimum SSE.

16) Single and complete link hierarchical clustering.

	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



a) Single Link



b) Complete Link