CS422: Data Mining
HW3 **A20402686**

Tan Chapter 8:

Exercise 11 - Total SSE is the sum of the SSE for each separate attribute. What does it mean if the SSE for one variable is low for all clusters? Low for just one cluster? High for all clusters? High for just one cluster? How could you use the per variable SSE information to improve your clustering?

Answer -

(a) If the SSE of one attribute is low for all clusters, then the variable will be constant and of little use in braking the data into groups.

(b) If the SSE of one attribute is relatively low for just one cluster, then that attribute helps to define the cluster.

(c) If the SSE of an attribute is high for all clusters, then that attribute is a noise.

(d) If the SSE of an attribute is high for one cluster, then it will be odd with the information provided by the low SSE attributes that define the cluster. It could merely be the case that the clusters defined by this attribute are different from those defined by the other attributes. But for all case, it means that this attribute does not help to define the cluster.

(e) The main goal is to eliminate attributes that have poor distinguishing power between clusters, hence low or high SSE for all clusters, because they will be useless for clustering.
Generally, attributes that has a relatively high SSE with respect to other attributes are disadvantageous if they because they introduce a lot of noise into the computation of the overall SSE.

Exercise 12 - The leader algorithm (Hartigan 1394]) represents each cluster using a point, known as a leader, and assigns each point to the cluster corresponding to the closest leader, unless this distance is above a user-specified threshold. In that case, the point becomes the leader of a new cluster.

(a) What are the advantages and disadvantages of the leader algorithm as compared to K-means?

Answer - The K-means algorithm almost always produces better quality clusters measured by SSE. The leader algorithm requires a single scan of the data and hence it is more computationally efficient because each object is compared to the final set of centroids at most once. Leader algorithm always produces the same set of clusters for a fixed ordering of the objects. Except indirectly, it is not possible to set the number of resulting clusters for the leader algorithm.

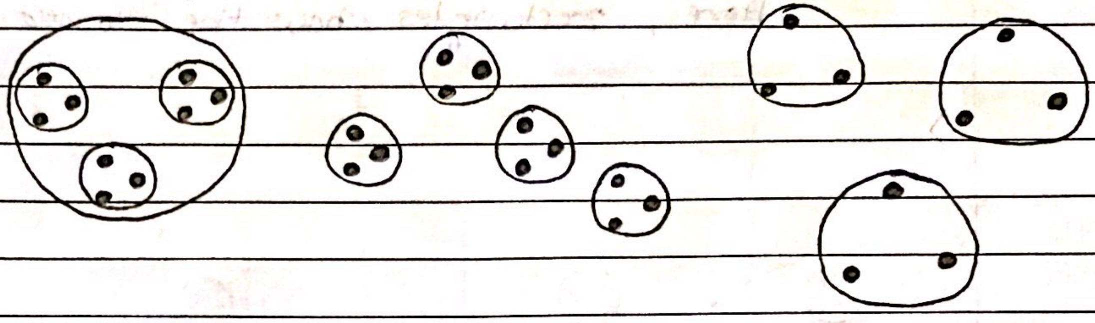
(b) Suggest ways in which the leader algorithm might be improved.

Answer – We should use a sample to determine the distribution of distances between the points. The information obtained from above process can be used to set the value of the threshold and the leader algorithm can be modified to cluster for several thresholds while performing a single pass.

1.1 → Tan chapter 8

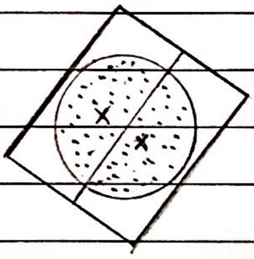
Exercise 2 :-

Find all well-separated clusters in the set of points shown in figure.



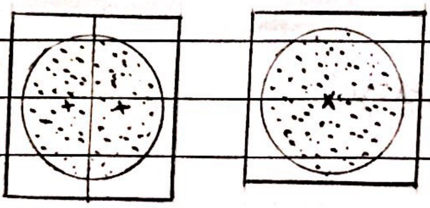
Exercise - 6 :-

a)



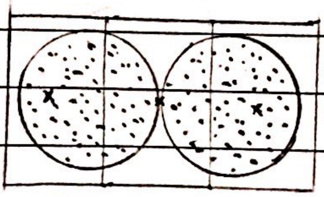
The centroids will be lie on the perpendicular bisector of the line which splits the circle into two clusters & they will be symmetrical. Every solution will have same global minimal erro

b)



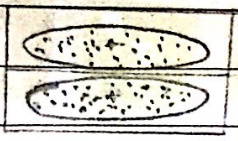
The bisector could have any angle. All these solutions have same globally minimal error.

c)

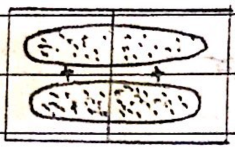


Here, three boxes represent three clusters.

d)



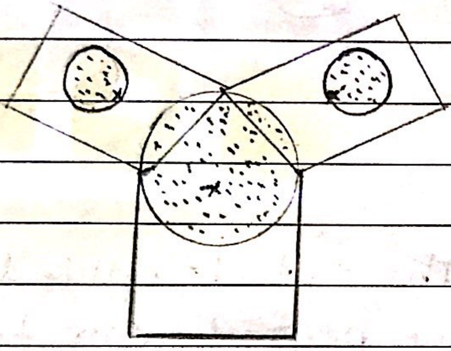
local minimum



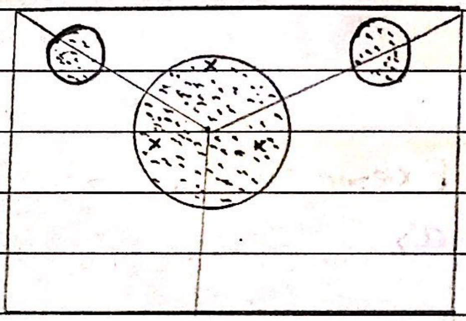
global minimum

Here, rectangles show the cluster.

e)



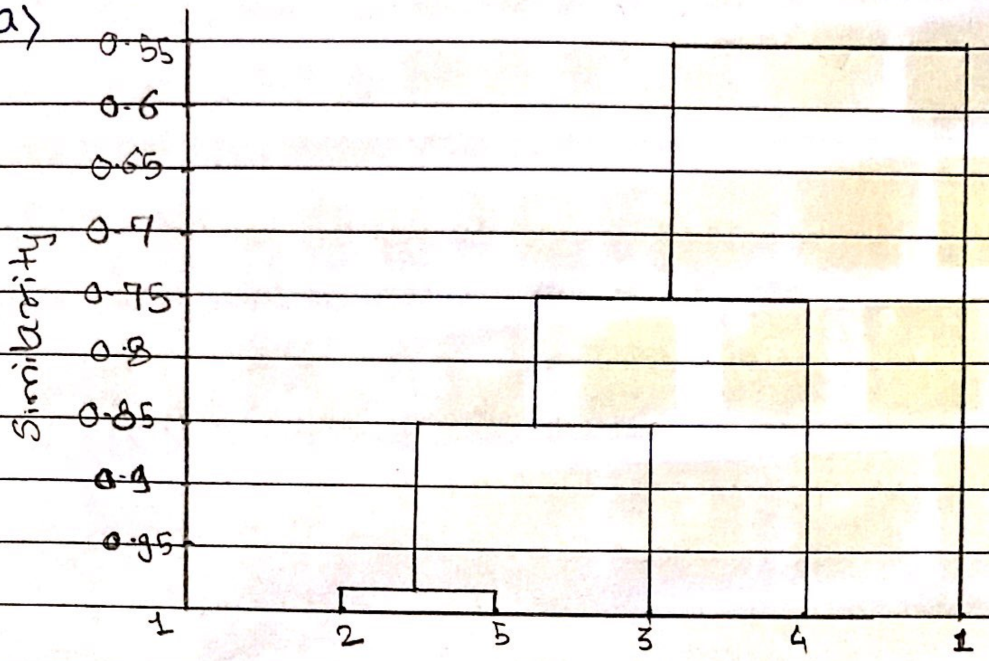
global minimum



local minimum

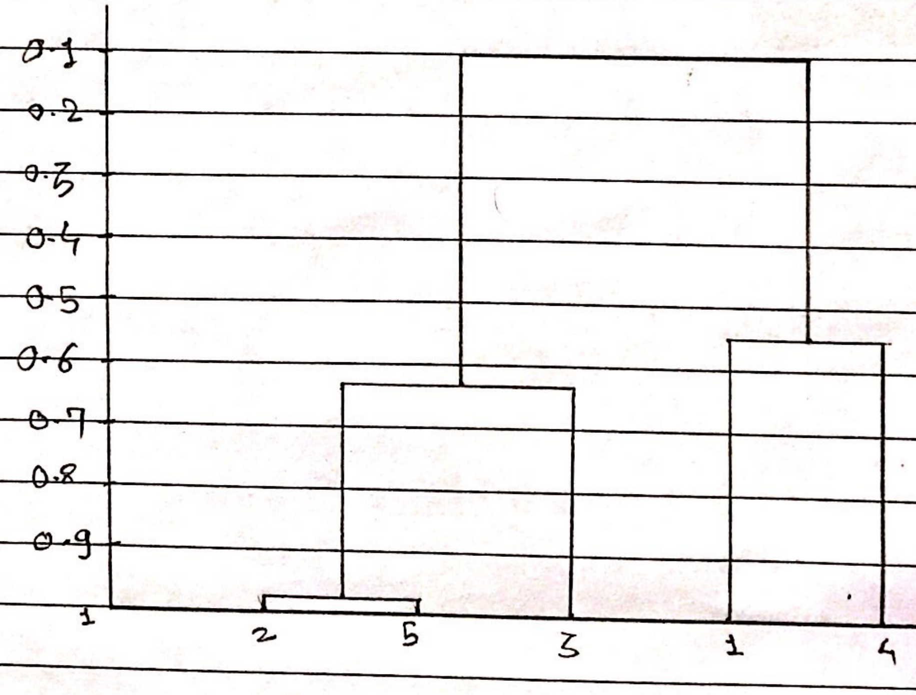
Exercise 16 :-

a)



a) Single Link

b)



b) Complete Link