DATA MINING HOMEWORK 8

Name: Pavani Rangineni

CWID: A20516359

1.1.2 Find all well-separated clusters in the set of points shown in Figure 7.35.

ANS:

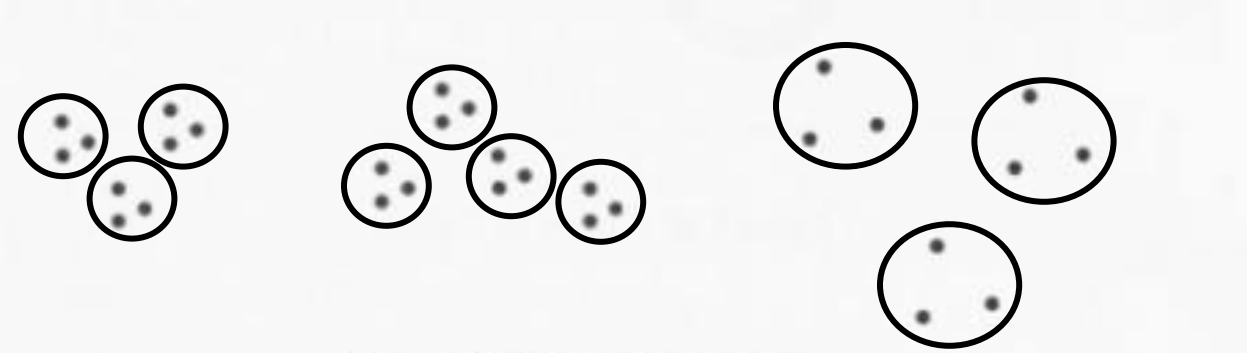


Figure 7.35. Points for Exercise 2.

1.1.6 For the following sets of two-dimensional points, (1) provide a sketch of how they would be split into clusters by K-means for the given number of clusters and (2) indicate approximately where the resulting centroids would be. Assume that we are using the squared error objective function. If you think that there is more than one possible solution, then please indicate whether each solution is a global or local minimum. Note that the label of each diagram in Figure 7.37 matches the corresponding part of this question, e.g., Figure 7.37(a) goes with part (a).

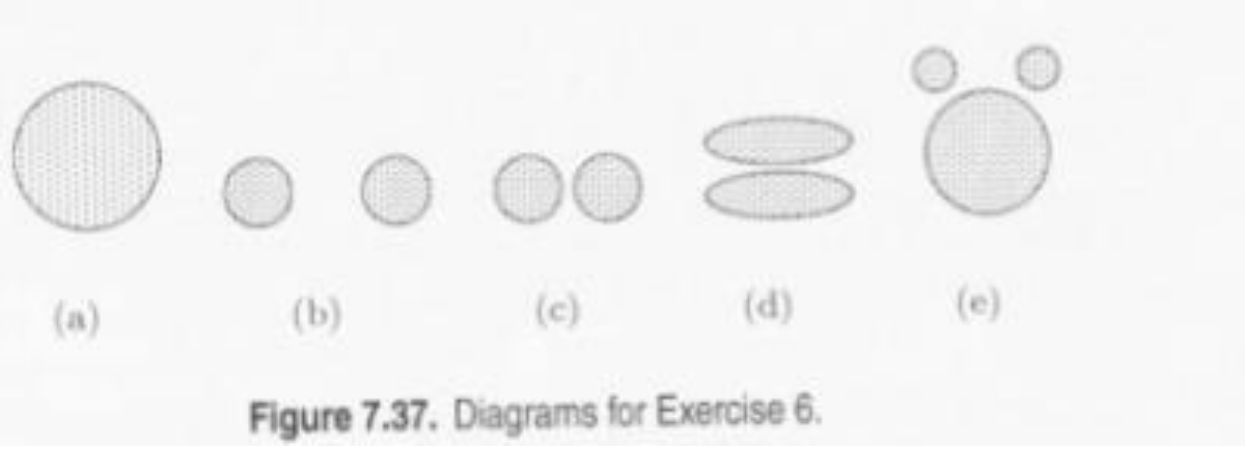
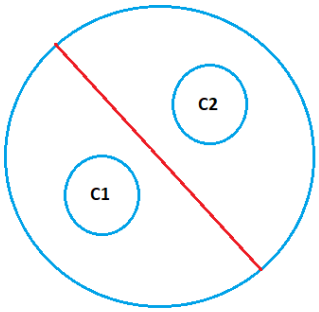
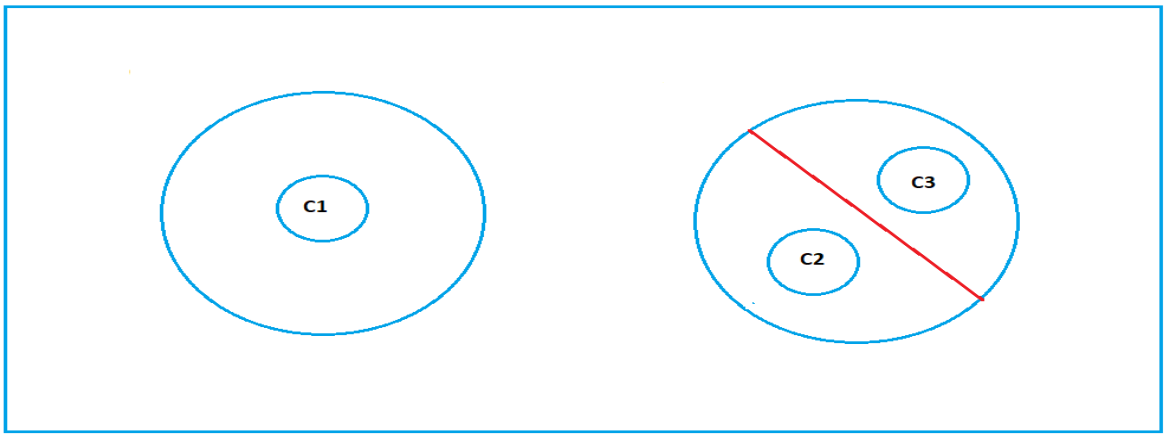


Figure 7.37. Diagrams for Exercise 6.

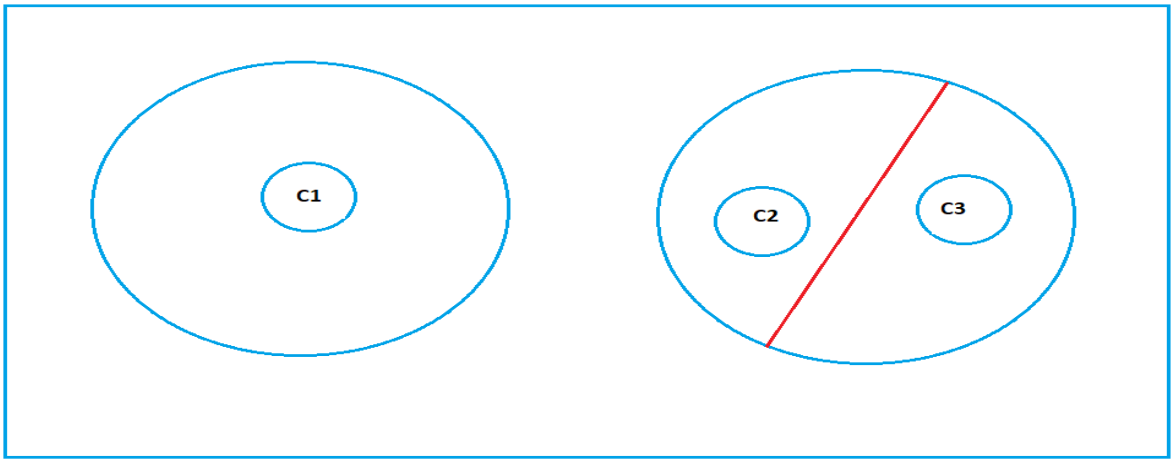
a) I< = 2. Assuming that the points are uniformly distributed in the circle, how many possible ways are there (in theory) to partition the points into two clusters? What can you say about the positions of the two centroids? (Again, you don't need to provide exact centroid locations, just a qualitative description.)



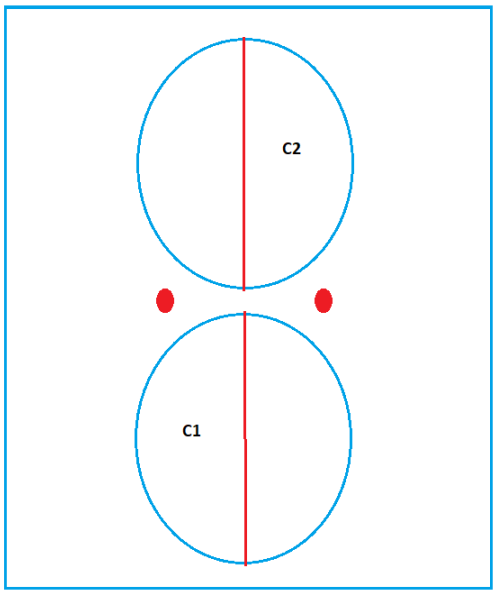
(b) K = 3. The distance between the edges of the circles is slightly greater than the radii of the circles.



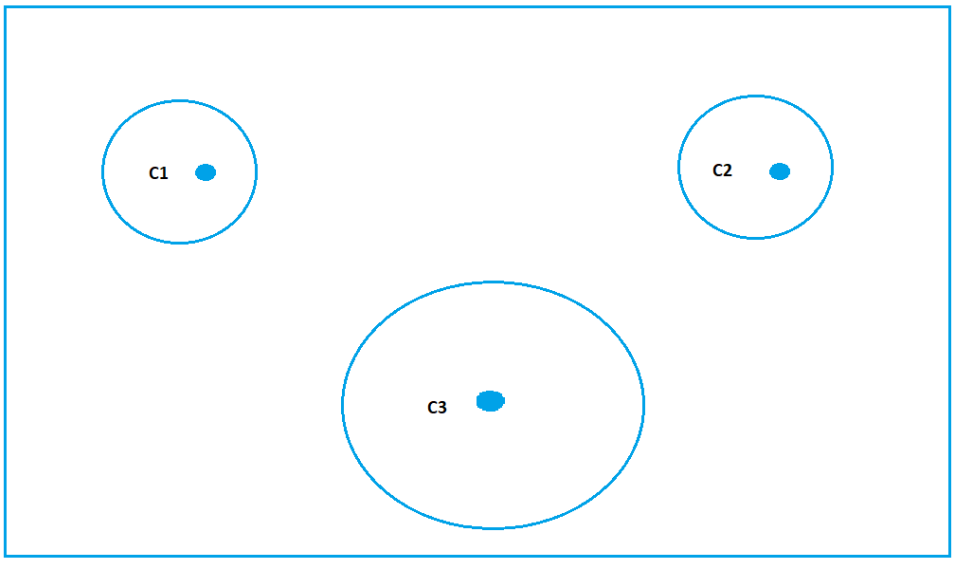
c) K = 3. The distance between the edges of the circles is much less than the radii of the circles.



(d) K = 2.



(e) K = 3. Hint: Use the symmetry of the situation and remember that we are looking for a rough sketch of what the result would be.



1.1.7 Suppose that for a data set

- there are m points and K clusters,
- half the points and clusters are in "more dense" regions.
- half the points and clusters arc in "less dense" regions, and

- the two regions are well-separated from each other.

For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters:

(a) Centroids should be equally distributed between more dense and less dense regions.

(b) More centroids should be allocated to the less dense region.

(c) More centroids should be allocated to the denser region.

ANS: The correct answer is (c). If the squared error is to be minimized, less dense regions require more centroids.

1.1.11 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?

ANS: If the SSE of one variable is low across all clusters, the variable is fundamentally constant and has little use in grouping data.

Low for just one cluster?

ANS: If the SSE of a single variable is relatively low for only one cluster, then the variable contributes to the cluster's definition.

High for all clusters?

ANS: If the SSE of one variable is relatively high across all clusters, it is possible that the variable is noise.

High for just one cluster?

ANS: If the SSE of a variable is relatively high for one cluster, it is inconsistent with the information provided by the variables with low SSE that define the cluster. This could simply be because the clusters defined by variables differ from those defined by the other variables, but in any case, it indicates that the variable does not help define the cluster.

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

ANS: The main idea here is to eliminate variables that have low discriminating power between clusters. That is, low/high SSE for all clusters, as they are ineffective for clustering. Remind the variables with high-SSE for all clusters are especially troublesome if they have a relatively high-SSE compared to the other variables because they introduce a lot of noise into the overall-SSE calculation.

1.1.12. The leader algorithm (Hartigan [533]) 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?

ANS: K-Means Advantages:
1) If the variables are large, K-Means is usually computationally faster than hierarchical clustering if k is kept small.

2) K-Means clusters are tighter than hierarchical clustering, especially when the clusters are globular.

K-Means Disadvantages:
1) Difficult to predict K-Value.
2) Different initial partitions can result in different final clusters.
3) It does not work well with clusters of Different size and Different density.

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

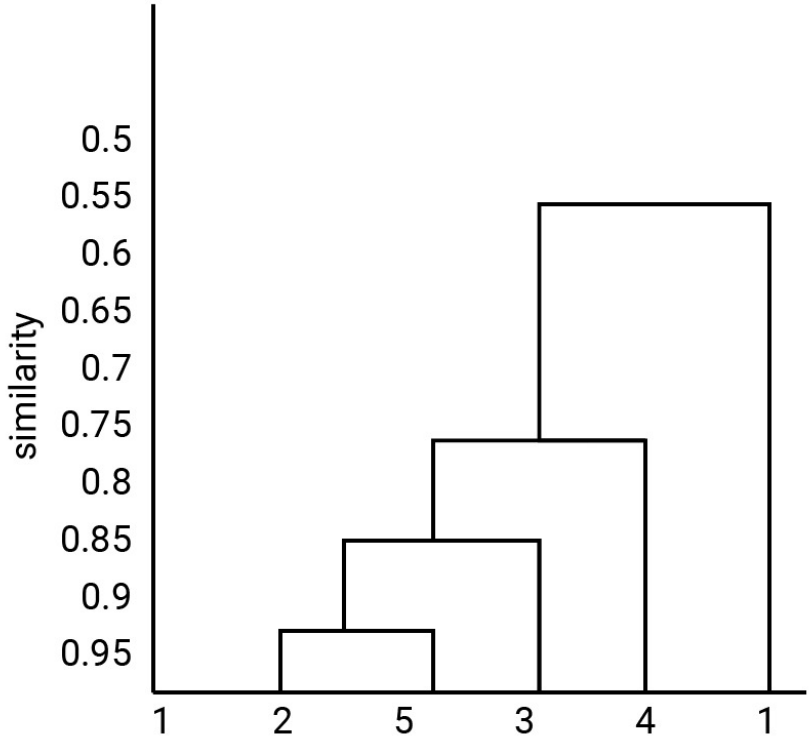
ANS: Determining the distribution of distances between points using a sample. The knowledge gained from this process can be used to intelligently set the value of the threshold. The leader algorithm could be modified to cluster for multiple thresholds in a single pass.

1.1.16 Use the similarity matrix in Table 7.13 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The den-drogram should clearly show the order in which the points are merged.

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	0.75
P5	0.35	0.98	0.84	0.76	1.00

Single link chart:



Complete link chart: