Tan, Chapter 7 (Cluster Analysis: Basic Concepts and Algorithms)

Exercise 2: Find all well-separated clusters in the set of points shown in Figure 7.1. The solutions are also indicated in Figure 7.1.

Figure 7.1. Points for Exercise 2.

Exercise 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 7.4 matches the corresponding part of this question, e.g., Figure 7.4(a) goes with part (a).

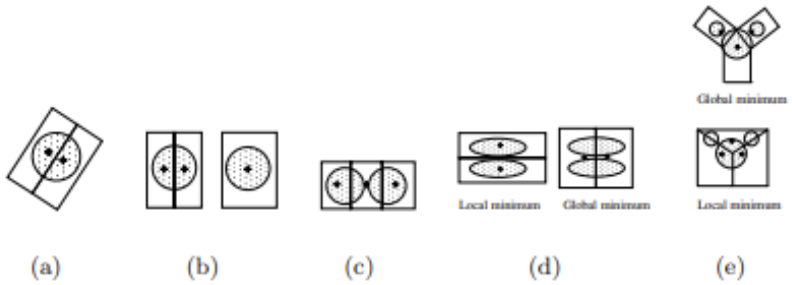


Figure 7.4. Diagrams for Exercise 6.

A. K = 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?
   a. **Solution**: In theory, there are an infinite number of ways to split the circle into two clusters - just take any line that bisects the circle. This line can 125 make any angle $0^\circ \leq \theta \leq 180^\circ$ with the x axis. The centroids will lie on the perpendicular bisector of the line that splits the circle into two clusters and will be symmetrically positioned. All these solutions will have the same, globally minimal, error.

B. K = 3. The distance between the edges of the circles is slightly greater than the radii of the circles.
    a. **Solution**: If you start with initial centroids that are real points, you will necessarily get this solution because of the restriction that the circles are more than one radius apart. Of course, the bisector could have any angle, as above, and it could be the other circle that is split. All these solutions have the same globally minimal error.
C. K = 3. The distance between the edges of the circles is much less than the radii of the circles.
    a. **Solution**:The three boxes show the three clusters that will result in the realistic case that the initial centroids are actual data points.
D. K = 2. In both cases, the rectangles show the clusters.
    a. **Solution**:In the first case, the two clusters are only a local minimum while in the second case the clusters represent a globally minimal solution.
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.
    a. **Solution**:For the solution shown in the top figure, the two top clusters are enclosed in two boxes, while the third cluster is enclosed by the regions defined by a triangle and a rectangle.

Exercise 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 are 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.

Solution: C, because less dense regions require more centroids if the squared error is to be minimized.

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?

Solution: If the SSE of one attribute is low for all clusters, then the variable is essentially a constant and of little use in dividing the data into groups. If the SSE of one attribute is relatively low for just one cluster, then this attribute helps define the cluster. If the SSE of an attribute is relatively high for all clusters, then it could well mean that the attribute is noise. If the SSE of an attribute is relatively high for one cluster, then it is at odds with the information provided by the attributes with low SSE that define the cluster. And if the idea is to eliminate attributes that have poor distinguishing power between clusters, i.e., low or high SSE for all clusters, since they are useless for clustering.

Exercise 12: The leader algorithm (Hartigan [4]) 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. Note that the algorithm described here is not quite the leader algorithm described in Hartigan, which assigns a point to the first leader that is within the threshold distance.

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

Solution: The leader algorithm requires only a single scan of the data and is thus more computationally efficient since each object is compared to the final set of centroids at most once. Although the leader algorithm is order dependent, for a fixed ordering of the objects, it always produces the same set of clusters.

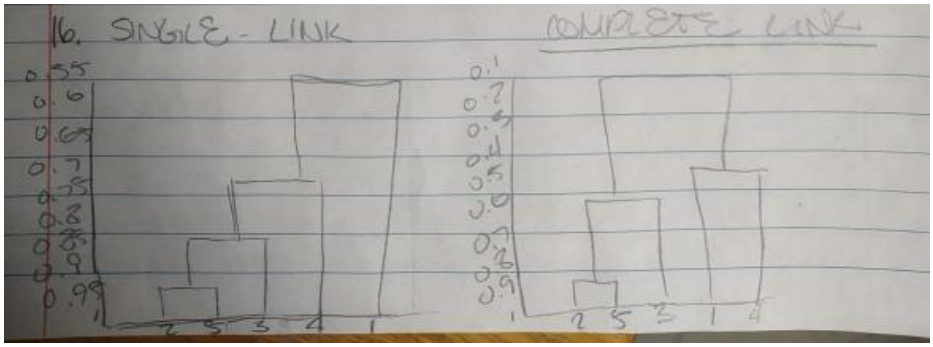
(b) Suggest ways in which the leader algorithm might be improved.

Solution: Use a sample to determine the distribution of distances between the points. The knowledge gained from this process can be used to more intelligently set the value of the threshold.

Exercise 16: Use the similarity matrix in Table 7.1 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

Table 7.1. Similarity matrix for Exercise 16.

	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	0.76
p5	0.35	0.98	0.85	0.76	1.00


16. SINGLE - LINK
COMPLETE LINK
0.55
0.6
0.7
0.98
1
2 5 3 4 1
0.1
0.2
0.4
0.5
0.9
1
2 5 3 1 4