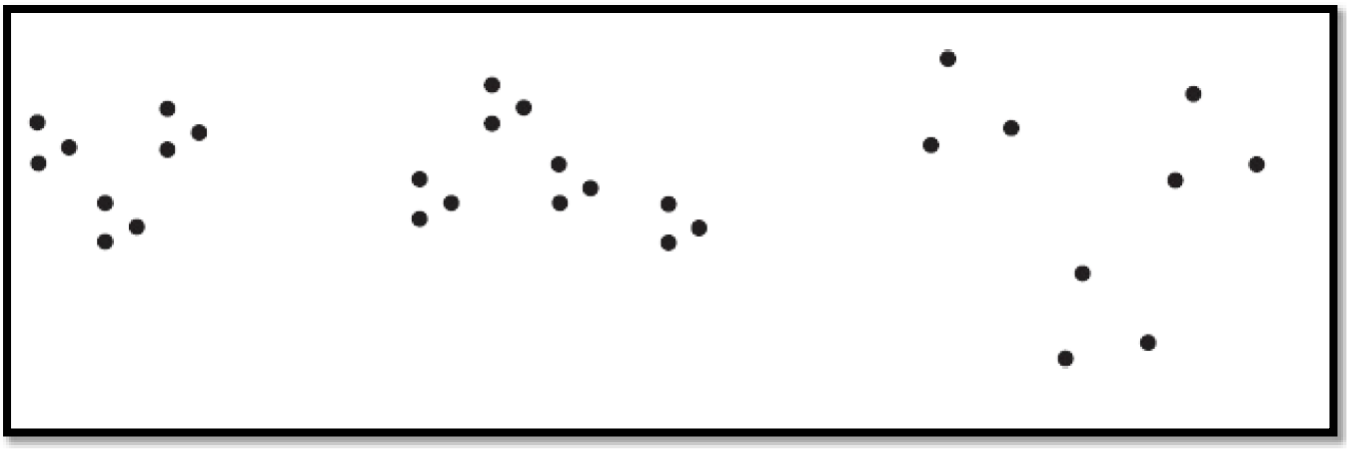
Javier-EchavarrenSuarez-HW8

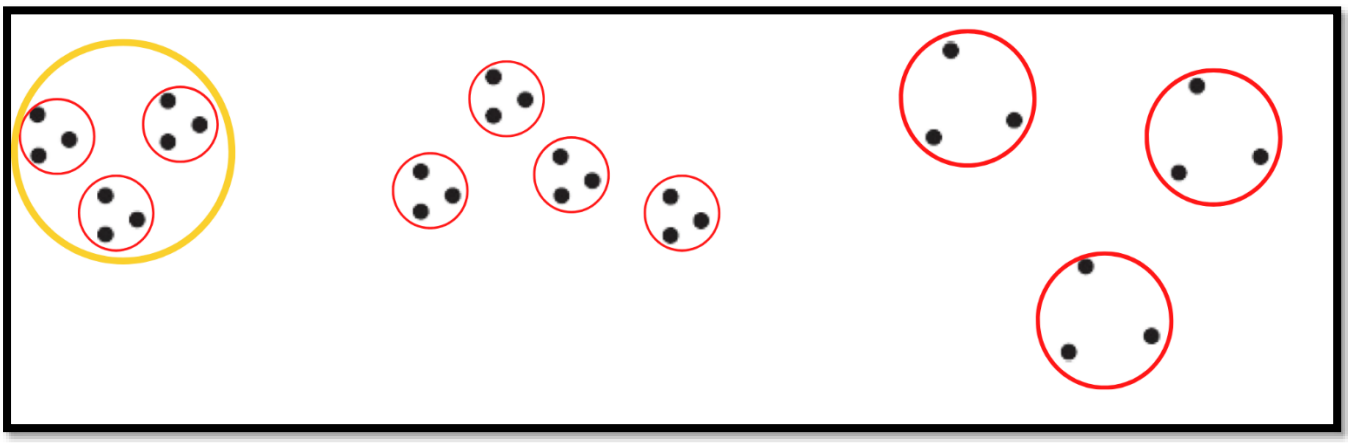
1. Exercises (3 points divided evenly among the questions)

1.1 Tan, Chapter 7 (Cluster Analysis: Basic Concepts and Algorithms) Exercise 2, 6, 7, 11, 12, 16. (For 16, note that Table 7.13 for Exercise 16 has a similarity matrix, not a distance matrix. Similarity and distance are related to each other by the formula distance = 1.0 – similarity.)

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



Solution:



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).

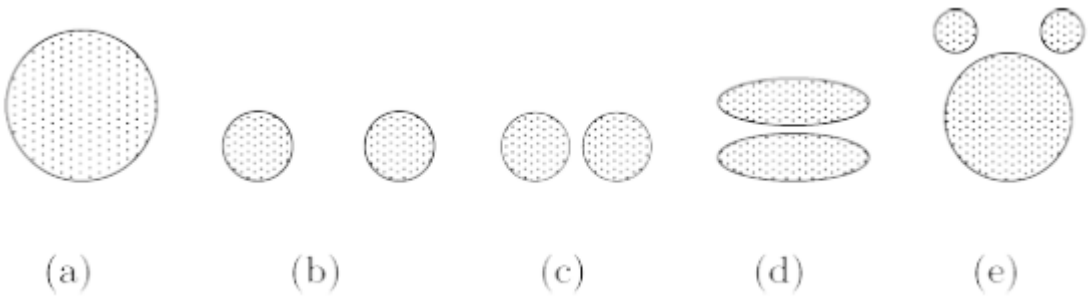
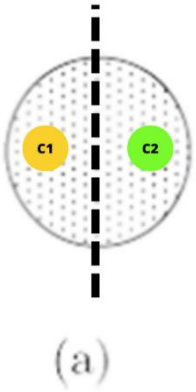


Figure 7.37.

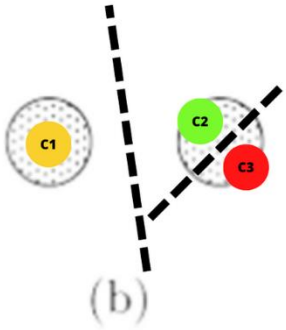
Diagrams for Exercise 6 .

a. **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? (Again, you don't need to provide exact centroid locations, just a qualitative description.)**
There are too many different ways to partition these points into clusters, as this distribution has radial symmetry. Any distribution of centroids would be like the one shown in the image but shifted around the center of the circle or one of the centroids being at the center of the circle and the other outside of it, with a distance to the center of the circle greater than two times the radius of the circle.



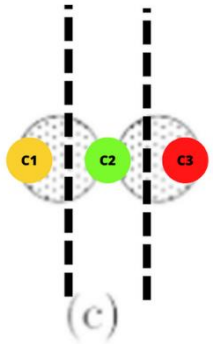
(a)

b. **K=3 The distance between the edges of the circles is slightly greater than the radii of the circles.**
If the distance between the edges of the circles is greater than the radii of the circles, It will be impossible that a centroid on one of the circles takes any of the points from the other circle as long as there is any centroid in the second circle on the nearest half to the first circle.
So one of the centroids will be at the center of one of the circles, and the other two will be in the second circle, with the same distance to the center of the circle and aligned with the center of the second circle.



(b)

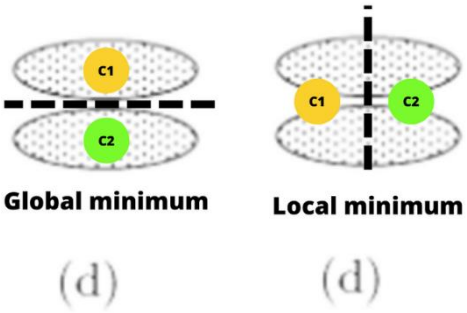
c. **K= 3 The distance between the edges of the circles is much less than the radii of the circles.**
In this case the three centroids will be aligned with the two centers of the circles. One of them being between the start of the first circle and the center of it, the second one will be between both circles, and the third one will be between the center of the second circle and its end.



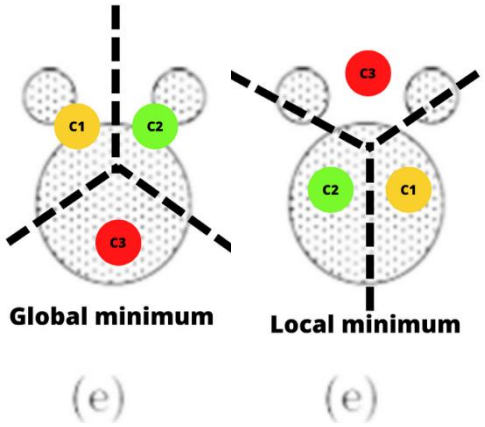
(c)

d. **K= 2**

In this case, there is a local minimum where both centroids are between the groups of points, and a global maximum where each centroid is in the center of each group of points.



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.**
In this case, there will be two cases, in the global minimum, the centroids of the first two clusters will be between the little groups of data and the big group, having symmetry on the y axis, and the other centroid will be in the big group of data, on the symmetry axis, and lower than the center of the circle.
In the local minimum case, a centroid will be between the three groups of data, on the axis of symmetry, and the other two will be in the big circle of data, having horizontal symmetry.



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.
False

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

True

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

False

Explanation for a,b,c: There should be more centroids in the less dense region, since the data in that region will be more disperse, and their distance to the centroids is higher, we would have less squared error if we put more centroids in the less dense zone.

Note: Do not get distracted by special cases or bring in factors other than density. However, if you feel the true answer is different from any given above, justify your response.

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?

If the SSE for one variable is low for all clusters, it means that this variable is a constant, and doesn't help us distinguish between clusters.

Low for just one cluster?

If the SSE of one variable is low for just one cluster it means that variable has high correlation with that cluster, as points in that cluster have similar values of that variable. This variable can serve to distinguish data in and out of that cluster.

High for all clusters?

If SSE for a variable is high for all clusters, it means that the variable is not giving us any information about the cluster that the data is in, it is randomness, and we should not give attention to that variable in the clusterization.

High for just one cluster?

If SSE is high for only one cluster, it means that variable doesn't help us define that cluster, and doesn't help us discriminate between that cluster and others.

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

We could see what variables are important for clustering and what other variables are not, so we can discard the variables that don't help and that way we can reduce noise and make a better clusterization with the relevant variables.

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?

The advantages are that this method is adaptable to n number of clusters, depending on the dataset, and is computationally cheap, as there is not iteration for all the points, once a point is evaluated, there is not further evaluation of that point. However, it is dependent on the initial point chosen, as the clusters start growing from there. It also has the problem of defining the threshold, which may not be the correct one. Finally, it has the size problem, as this method is not capable of selecting clusters with a size greater than two times the threshold.

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

We could keep adding to the cluster all the points that are with a distance to any of the points in the cluster lower than the threshold, and this way we would be able to select clusters that are bigger than the threshold, as long as they have at least distances between the points lower than the threshold.

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 dendrogram should clearly show the order in which the points are merged.

	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

Table 7.13. Similarity matrix for Exercise 16 .

Distance Matrix					
	p1	p2	p3	p4	p5
p1	0	0,9	0,59	0,45	0,65
p2	0,9	0	0,36	0,53	0,02
p3	0,59	0,36	0	0,56	0,15
p4	0,45	0,53	0,56	0	0,24
p5	0,65	0,02	0,15	0,24	0

We take the two points with less distance between them (2 and 5) and refresh the table.

Distance Matrix- iteration 1				
	p1	p2,p5	p3	p4
p1	0	0,65	0,59	0,45
p2,p5	0,65	0	0,15	0,24
p3	0,59	0,15	0	0,56
p4	0,45	0,24	0,56	0

We take again the less distance between points or clusters (2+5, p3) and keep iterating.

Distance Matrix- iteration 2			
	p1	p2,p5,p3	p4
p1	0	0,59	0,45
p2,p5,p3	0,59	0	0,24
p4	0,45	0,24	0

We take again the less distance between points or clusters (2+5+3,4) and keep iterating.

Distance Matrix- iteration 3		
	p1	p2,p5,p3,p4
p1	0	0,45
p2,p5,p3,p4	0,45	0

And we finally add the last value to the cluster.

The corresponding dendrogram is shown below.