Data Mining
Problem Set 5

Chapter 8

2. Find all well-separated clusters in the set of points shown in Figure 8.1.

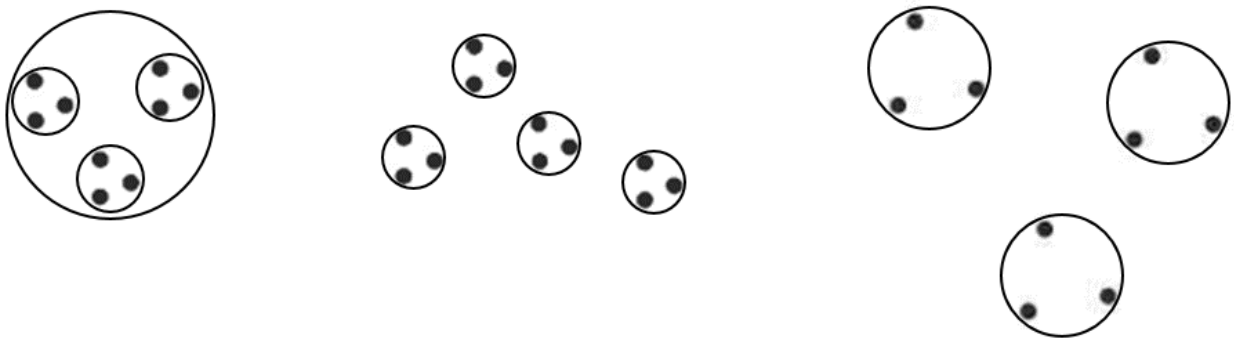


Figure 8.35. Points for Exercise 2.

5. Identify the clusters in Figure 8.3 using the center-, contiguity-, and density-based definitions. Also indicate the number of clusters for each case and give a brief indication of your reasoning. Note that darkness or the number of dots indicates density. If it helps, assume center-based means K-means, contiguity-based means single link, and density-based means DBSCAN.

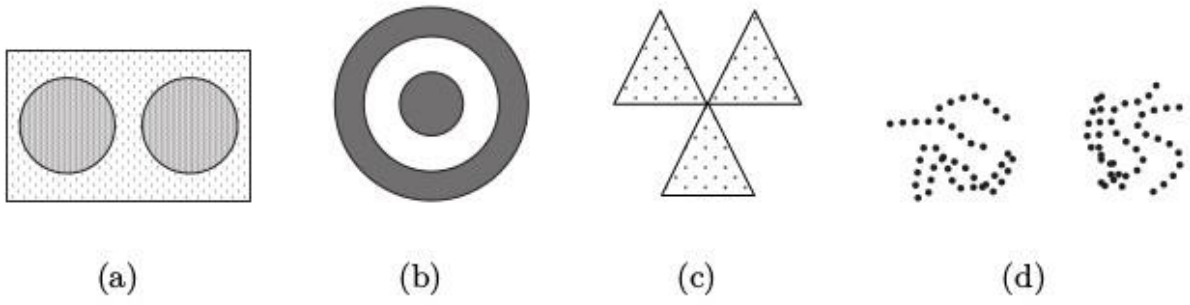


Figure 8.3. Clusters for Exercise 5.

a. **Center-based:** Splitting the rectangular region in half, we will have 2 clusters.
Contiguity-based: The noise joins the circular regions into 1 cluster.
Density-based: As the circles are the densest, each circular region will be a cluster, resulting in 2 clusters total.

b. **Center-based:** 1 cluster with both rings.
Contiguity-based: 1 cluster for each ring, so there are 2 clusters total.
Density-based: The two rings define two regions of high density with a region of low density between them, so there are 2 clusters total, one for each ring.

c. **Center-based:** 1 cluster containing all 3 triangles.
Contiguity-based: Because the triangles are touching, there will be 1 cluster.
Density-based: There will be 3 clusters despite them touching because of the density inside the triangles.

d. **Center-based:** There will be a total of 2 clusters, one for each main group of lines.
Contiguity-based: Each regular line will be its own cluster for 3 clusters in addition to the 2 different sets of lines that intertwine for 5 clusters total.
Density-based: The two main groups of lines have high density with a region of low density between them, so there will be 2 clusters.

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

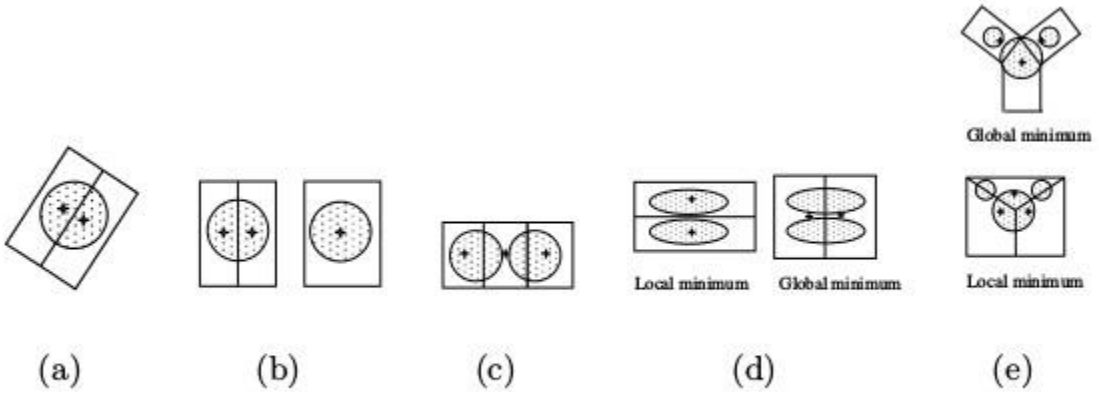


Figure 8.4. Diagrams for Exercise 6.

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 8.4 matches the corresponding part of this question, e.g., Figure 8.4(a) goes with part (a).

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 the exact centroid locations, just a qualitative description.)

i. There is an infinite number of ways to split a circle into two clusters in theory with any line that bisects the circle. This line can make any angle up to 180° with the x-axis. Then, the centroids will lie on the perpendicular bisector of

the line that splits the circle into two clusters and will be symmetrical in position. Each of these solutions is a global minimum.

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

i. The restriction forces this form of solution. As they are still circles, the bisector could have any angle, or it could be the other circle that is split. Again, any of these solutions have the same global minimum.

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

i. The three clusters will each be in the middle of their box.

d. K = 2.

i. The two possible solutions are shown. If bisected horizontally, the solution is the left figure, and it has a local minimum. If bisected vertically, the solution is the right figure, and it has is a global minimum.

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.

i. I believe only the top figure to be possible in practice. The top two clusters are enclosed in rectangles while the bottom cluster's shape is made up of a triangle and a rectangle. This is a global minimum.

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:

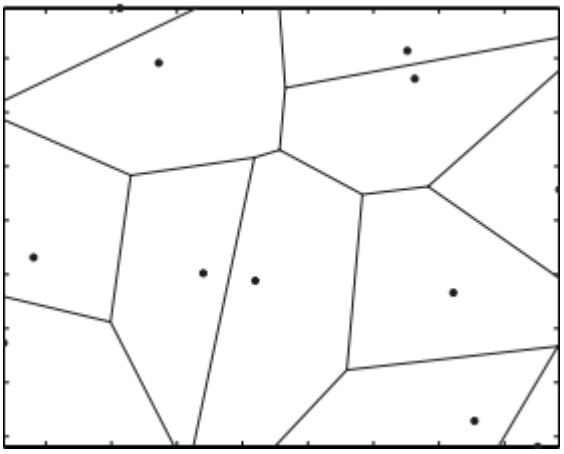
   (a) **Centroids should be equally distributed between more dense and less dense regions.**
   (b) **More centroids should be allocated to the less dense region.**
   (c) **More centroids should be allocated to the denser region.**

   - Option (b) should be chosen as less dense regions require more centroids to minimize the squared error.

9. Give an example of a data set consisting of three natural clusters, for which (almost always) K-means would likely find the correct clusters but bisecting K-means would not.

   a. To ensure K-means finds the correct clusters more easily, we will start with 3 identical circular clusters with centers on the same line. The middle cluster's center will be equidistant from the other two cluster centers. With the data set this way, bisecting K-means will always split the middle cluster to start which will always lead to the wrong set of clusters.

13. The Voronoi diagram for a set of K points in the plane is a partition of all the points of the plane into K regions, such that every point (of the plane) is assigned to the closest point among the K specified points. (See Figure 8.5). What is the relationship between Voronoi diagrams and K-means clusters? What do Voronoi diagrams tell us about the possible shapes of K-means clusters?

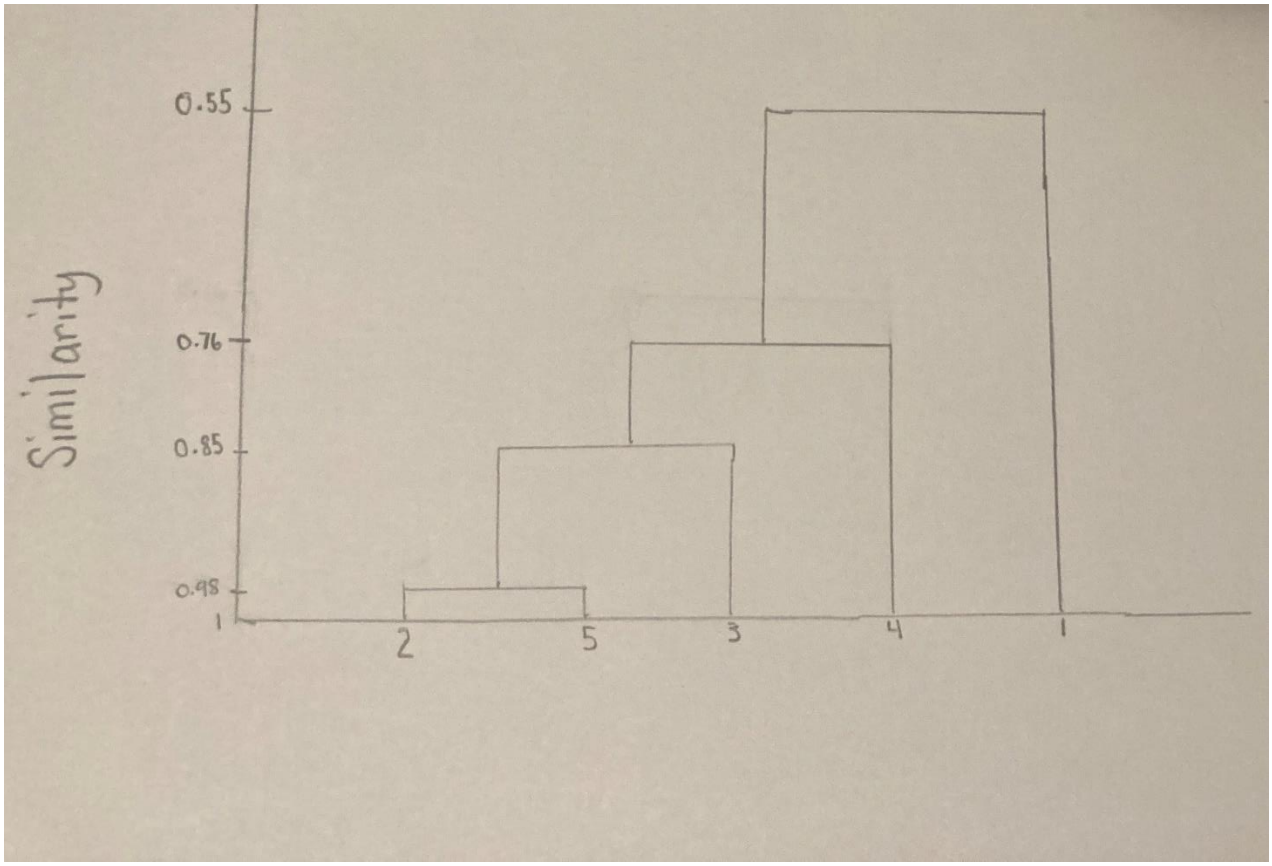


a. For *K* K-means clusters, the plane will be divided into *K* Voronoi regions since they represent the closest points among the centroid.
b. Voronoi diagrams tell us that the K-means clusters have boundaries between them that are piecewise linear.

16. Use the similarity matrix in Table 8.1 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.

a. Single Link

Points {2} and {5} are merged first = 0.98
{2,5} merged with {3} = 0.85
{2,5,3} merged with {4} = 0.76
{2,5,3,4} merged with {1} = 0.55



b. Complete link

17.	P1	P2	P3	P4	P5
P1	1.0	0.10	0.41	0.55	0.35
P2	0.10	1.0	0.64	0.47	0.98
P3	0.41	0.64	1.0	0.44	0.85
P4	0.55	0.47	0.44	1.0	0.76
P5	0.35	0.98	0.85	0.76	1.0

min{(2,5), 1} = min{(2,1), (5,1)} = min(0.1, 0.35) = 0.1
min{(2,5), 3} = min{(2,3), (5,3)} = min(0.64, 0.85) = 0.64

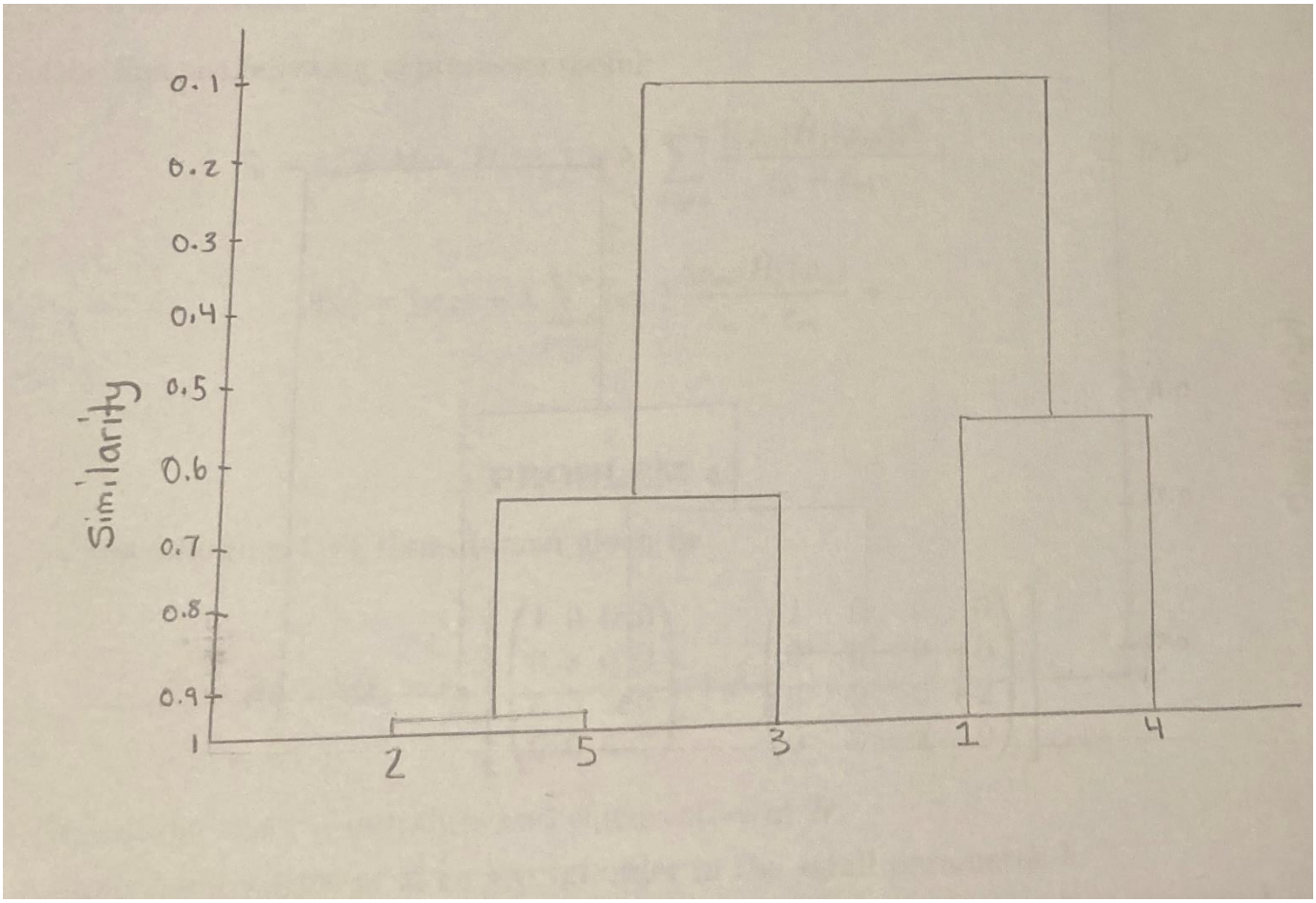
min{(2,5), 4} = min{(2,4), (5, 4)} = min(0.47, 0.76) = 0.47

	(2,5)	1	**3**	4
(2,5)	1	0.1	0.64	0.47
1	0.1	1	0.41	0.55
3	0.64	0.41	1	0.44
4	0.47	0.55	0.44	1

min{((2,5),3), 1} = min{((2,5), 1), (3,1)} = min(0.1, 0.41) = 0.1
min{((2,5),3), 4} = min{((2,5),4), (3,4)} = min(0.47, 0.44) = 0.44

	((2,5),3)	1	4
((2,5),3)	1	0.1	0.44
1	0.1	1	0.55
4	0.44	0.55	1

min{(1,4), ((2,5),3)} = min{(1, ((2,5),3)), (4, ((2,5),3))} = min(0.1, 0.44) = 0.1



17. Hierarchical clustering is sometimes used to generate *K* clusters, $K > 1$ by taking the clusters at the K^{th} level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.
The following is a set of one-dimensional points: {6, 12, 18, 24, 30, 42, 48}.

(a) For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids.

i. **{18,45}**
First cluster for 18: 6, 12, 18, 24, 30
Squared Error = $(18-6)^2 + (18-12)^2 + (18-18)^2 + (18-24)^2 + (18-30)^2 = 360$
Second cluster for 45: 42, 48
Squared Error = $(45-42)^2 + (45-48)^2 = 18$
Total squared error = 378

ii. **{15,40}**
First cluster for 15: 6, 12, 18, 24
Squared Error = $(15-6)^2 + (15-12)^2 + (15-18)^2 + (15-24)^2 = 180$
Second cluster for 40: 30, 42, 48
Squared Error = $(40-30)^2 + (40-42)^2 + (40-48)^2 = 168$
Total squared error = 348

(b) Do both sets of centroids represent stable solutions; i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated?

i. Yes, both sets of centroids represent stable solutions.

(c) What are the two clusters produced by single link?

i. {6, 12, 18, 24, 30} as they're the same distance and then {42, 48} since the distance from 42 to 48 is smaller than the distance from 30 and 42.

(d) Which technique, K-means or single link, seems to produce the "most natural" clustering in this situation? (For K-means, take the clustering with the lowest squared error.)

i. Single link produces the most natural clustering.

(e) What definition(s) of clustering does this natural clustering correspond to? (Well-separated, center-based, contiguous, or density.)

i. Contiguous clustering.

(f) What well-known characteristic of the K-means algorithm explains the previous behavior?

i. K-means cannot handle clusters of different sizes well.

22. You are given two sets of 100 points that fall within the unit square. One set of points is arranged so that the points are uniformly spaced. The other set of points is generated from a uniform distribution over the unit square.
    a. **Is there a difference between the two sets of points?**
        i. There is a difference in density between the sets of points. In the uniformly spaced set, the density is uniform throughout the unit square, but there will be spots of lesser or greater density in the set of random points.
    b. **If so, which set of points will typically have a smaller SSE for K=10 clusters?**
        i. The random set of points will typically have smaller SSE.
    c. **What will be the behavior of DBSCAN on the uniform data set? The random data set?**
        i. Depending on the threshold, DBSCAN might classify the entire uniform data set as noise or merge all points into one cluster. For the random data set, DBSCAN should be able to find clusters because of the varying densities.