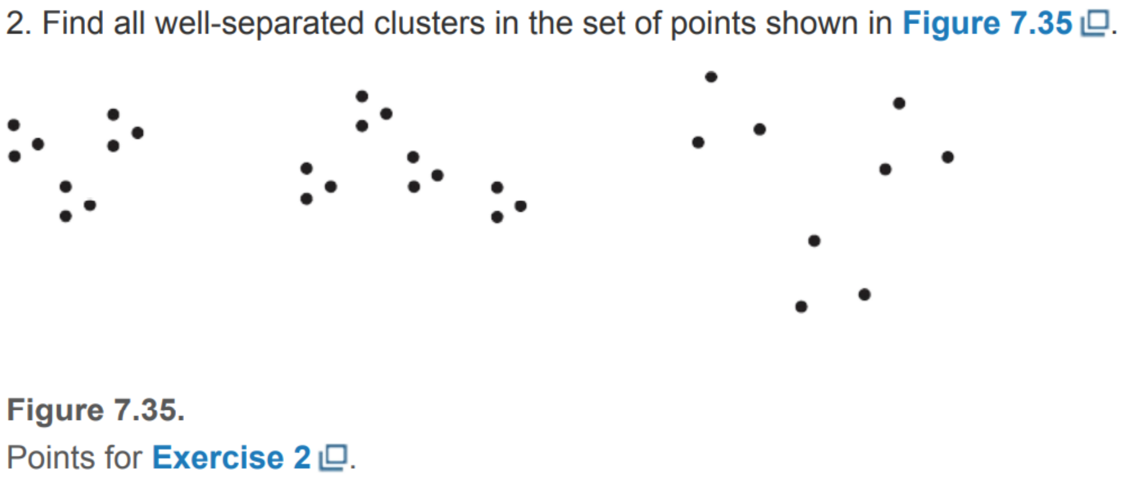
1.1 Tan, Chapter 7 (Cluster Analysis: Basic Concepts and Algorithms)
Exercises 2, 6, 7, 11, 12, 16.

2.)



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

Figure 7.35.
Points for **Exercise 2**.

- ➢ In the first picture, three different clusters can be seen. Each of the clusters is made up of three points.
- ➢ In the second picture, four different clusters can be seen. Each of the clusters is made of three points.
- ➢ In the third picture, three different clusters can be seen. Each of the clusters is made of three points.

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).

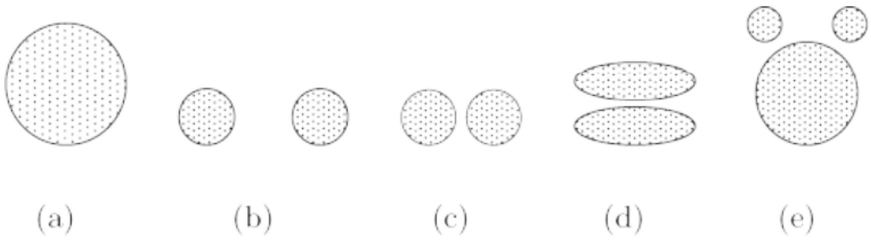


Figure 7.37. Diagrams for Exercise 6 .

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.)

In fact, there are infinite solutions to this problem given the constraints defined by the statement. It just must be done by a simple split of the points within the circle, for example with any line or two-dimensional space constraint dividing the circle and crossing it through the center. For instance, taking a valid solution such as splitting the circle by a line crossing the center of it, the data would be separated into two different sized clusters which each of the clusters' centroids would be in a line which is orthogonal to the splitting line. The solutions defined, would have a minimum global error.

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

Taking into account that both clusters are separated by a distance greater than a radius, and by starting with two clusters located in real coordinates of the two-dimensional plane, the algorithms would reach the correct solution. That is to say, it would end up by taking one of the circles as a cluster, and dividing the other one by a line crossing it through the center, thus leading to the other two clusters. The solutions previously mentioned have reached a global minimum error.

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

The resulting centroids would be some of the data points and the clusters will be located one close to each circle. And about the third one it could be said, that if the distance which separates both circles really is so small compared to the radius, a cluster could be formed between both circles clustering some of the data points in the center perimeter boundaries of the circles.

d. K = 2

For this case, each ellipse will be grouped as a cluster with centroids located at each of the elipse's centers. This way, a global minimal error could be reached.

A splitting can also be done by grouping the left-side points in a cluster, and the right-side points in another cluster, thus, leading to a symmetrical cluster division. But this clustering would not reach a global minimal error value, it would be a solution comprising a local minimal error.

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.

By taking into account the noticeable symmetry of the image, two different solutions came up to my mind. First of all, three clusters each of them mainly encompassing points of each circle. That is to say, the majority of the biggest circle's points would be encompassed by a centroid located somehow south than the center of the big circle. Then, each of the other two centroids, located close to the southeast and southwest perimeter's borders of the small circles, would encompass the rest of the points of the biggest circle, the ones in the north region, and the ones of the small circles. This could seem like a globally minimal error, but some further calculations should be done in order to demonstrate comparing the results with the one of the next proposal.

As a second solution, and based on the symmetrical properties of the layout, the three centroids could be located in the biggest circle. One close to the north perimeter of the biggest circle and encompassing the points located at the north part of the biggest circle, as well as the northern-center points of the smallest circles. The other two centroids would be located at the x-axis line crossing the center of the biggest circle, but at the center-left and right side respectively encompassing the southwest and southeast points of the biggest circle respectively, as well as the remaining points located at the southwest and shouteast regions of the smallest circles. This solution seems to be a local minimal error based solution, but some calculations should be made in order to compare it with the previous proposed solution. As the data was not given, the author cannot make any further quantitative analysis, as the scope of this response would be limited by a qualitative analysis of the figures.

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***
- ***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.

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.

Given the dataset described in the statement, in which we can encounter half of the data points clustered in half of the clusters in a highly dense area, and the other ones in a low density area, we are asked to determine the distribution of centroids. Therefore, among the three options given, the second one, the ***"b"***, is chosen.

Taking into account that the optimization for this algorithm is based on getting the lowest value of squared errors, the less dense regions must contain a higher number of centroids. This way, the more sparse regions will be able to be covered by more centroids and clusters, thus leading to a decrease in the sum of squared errors due to the fact that the dispersed data will be better fitted.

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?

SSE for one variable is low for all clusters?

It means that the attribute value is not really significant for the clustering, probably a constant not really useful for the clustering. That is to say, that taking only into account the values of the data points for that attribute, they could not be easily clustered into K clusters. In other words, probably the clustering has not been made mainly depending on the value of that attribute.

Low for just one cluster?

It means that the attribute really helps in order to perform the cluster. That is to say, the attribute is really significant for that clustering.

High for all clusters?

If the sum of squared errors is high for all the clusters by taking into account the data of an attribute, it could mean that the attribute is not really useful for the clustering. In fact, it could be noise, really harmful for accurate clusterings.

High for just one cluster?

It means that the attribute does not really help in order to perform the cluster. That is to say, the attribute is not really significant for that clustering, hence the clustering will be performed according to the attributes that show a low value of sum squared errors.

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

It must be clear that whenever it has to be chosen among attributes, the key is to conserve the useful ones and to eliminate the ones that do not help in distinguishing between clusters.
For example, whenever we encounter attributes with really low or really high SSE values for all the clusters, we can neglect them thereby leading to an enhancement of the clustering.

Summary:

The attributes with a low value of SSE will be the ones that define the clustering whereas the ones with high SSE values will not be helpful for the clustering.

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 leader algorithm is computationally less expensive due to the fact that each object is compared to the final centroids only one time in the worst case scenario. In addition to this, the leader algorithm depends on the order of the items, but for the same order, always gets the same clustering or outcome.

Nevertheless, in the leader algorithm the number of clusters cannot be chosen as it can be in the K-means algorithm. Finally, the K- means algorithm is usually more accurate and leads to better results in terms of Sum of Squared Errors.

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

Taking a subset of the dataset and using it as an experimental sample for figuring out the distances between the data points. With this information, a more sensible choice will be made, when it comes to setting up a 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.

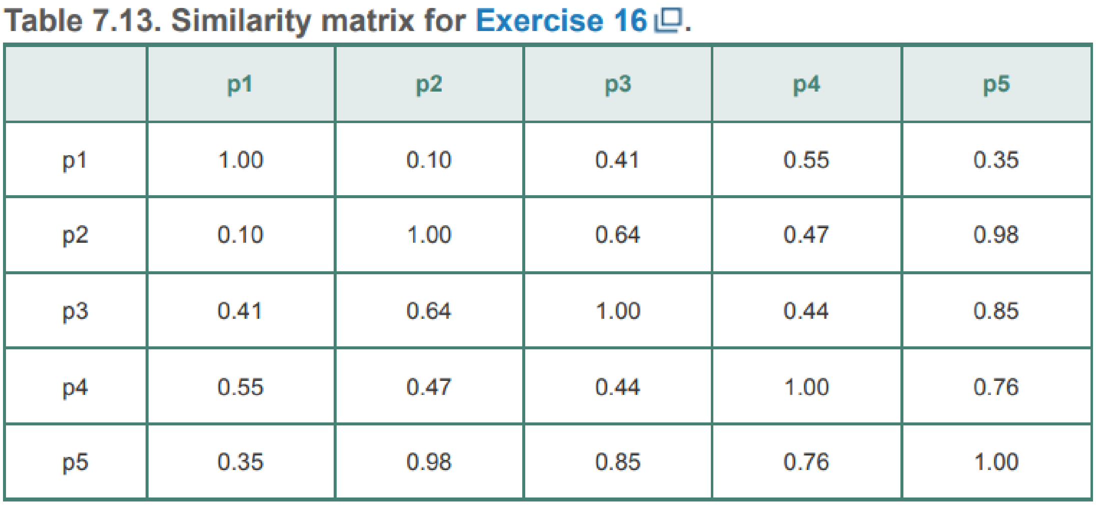


Table 7.13. Similarity matrix for Exercise 16.

	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

(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.)

By using the equation distance = 1.0 - similarity, the distance table is obtained from the similarity data.

	p1	p2	p3	p4	p5
p1	0.00	0.90	0.59	0.45	0.65
p2	0.90	0.00	0.36	0.53	0.02
p3	0.59	0.36	0.00	0.56	0.15
p4	0.45	0.53	0.56	0.00	0.24
p5	0.65	0.02	0.15	0.24	0.00

The exercise could be solved based on both, the similarity data or the distance data. It will be solved by using the similarity data.

SINGLE LINK HIERARCHICAL CLUSTERING

	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

	p1	***p2, p5***	***p3***	***p4***
p1	1.00	0.35	0.41	0.55
p2, p5	0.35	1.00	***0.85***	0.76
p3	0.41	***0.85***	1.00	0.44
p4	0.55	0.76	0.44	1.00

	p1	***p2, p5, p3***	***p4***
p1	1.00	0.41	0.55
p2, p3, p5	0.41	1.00	***0.76***
p4	0.55	***0.76***	1.00

	p1	***p2, p5, p3, p4***
p1	1.00	***0.55***
p2, p3, p5, p4	***0.55***	1.00

P5 → P2
P3 → P5, P2
P4 → P5, P2, P3
P1 → P5, P2, P3, P4

COMPLETE LINK HIERARCHICAL CLUSTERING

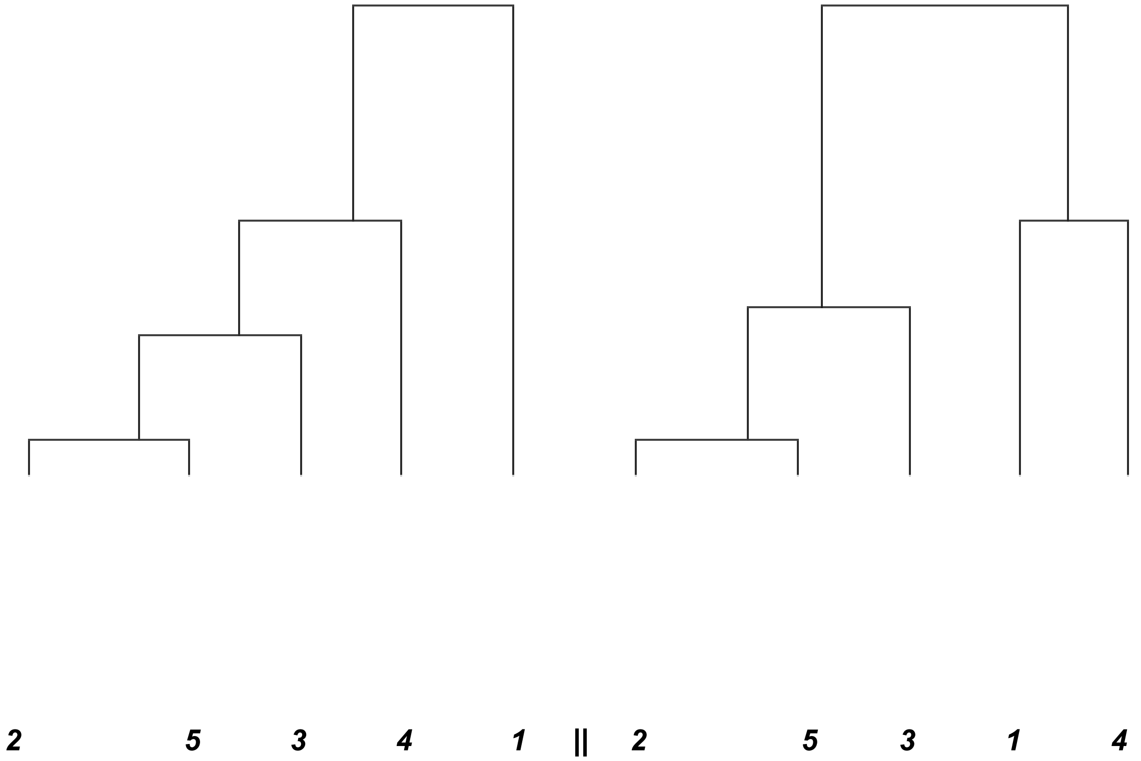
	p1	***p2***	***p3***	***p4***	***p5***
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	***0.98***
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	***0.98***	0.85	0.76	1.00

	p1	***p2, p5***	***p3***	***p4***
p1	1.00	0.10	0.41	0.55
p2, p5	0.10	1.00	***0.64***	0.47
p3	0.41	***0.64***	1.00	0.44
p4	0.55	0.47	0.44	1.00

	p1	***p2, p5, p3***	***p4***
p1	1.00	0.10	***0.55***
p2, p3, p5	0.10	1.00	0.44
p4	***0.55***	0.44	1.00

	p1, p4	***p2, p5, p3***
p1, p4	1.00	***0.10***
p2, p3, p5	***0.10***	1.00

P5 → P2
P3 → P5, P2
P4 → P1
P1, P4 → P5, P2, P3



SINGLE LINK

COMPLETE LINK