Data Mining

Problem Set 5

Chapter 8

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

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.

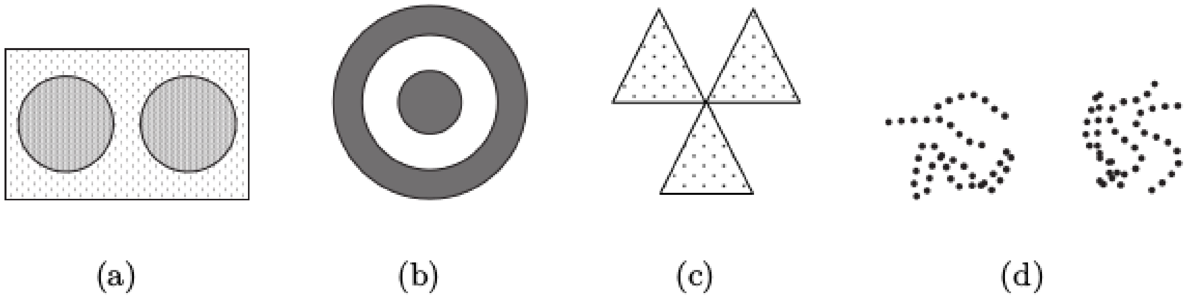

Figure 8.3. Clusters for Exercise 5.

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

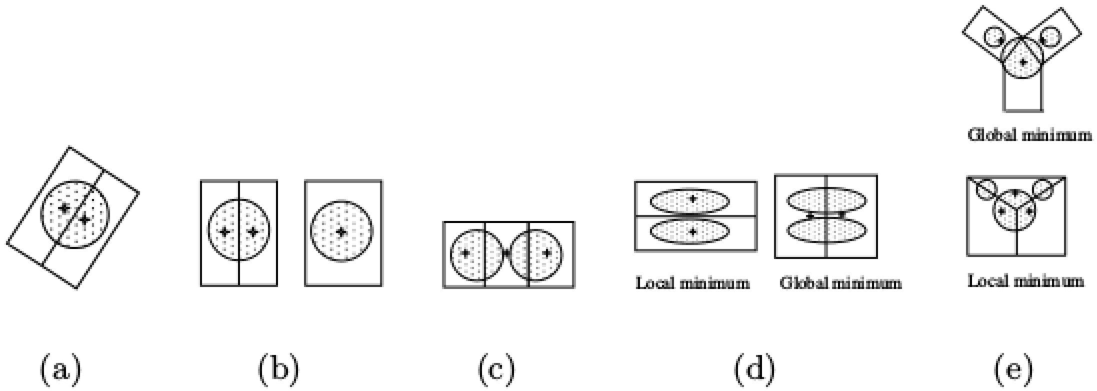

Figure 8.4. Diagrams for Exercise 6.

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.

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.

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.

10. Would the cosine measure be the appropriate similarity measure to use with K-means clustering for time series data? Why or why not? If not, what similarity measure would be more appropriate?

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?

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.

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?

(b) If so, which set of points will typically have a smaller SSE for K=10 clusters?

(c) What will be the behavior of DBSCAN on the uniform data set? The random data set?