Section 7.7 - 1, 2, 3, 5, 6, 7, 11, 14, 16, 30

1. **Consider a data set consisting of 2^{20} data vectors, where each vector has 32 components and each component is a 4-byte value. Suppose that vector quantization is used for compression, and that 2^{16} prototype vectors are used. How many bytes of storage does that data set take before and after compression and what is the compression ratio?**

   Answer: Before compression, dataset takes $4*32*\mathbf{2^{20}}$ = 134.217,728 bytes.

   After compression, dataset takes $4*32*2^{16}$ = 8,388,608 bytes for the prototype vectors and $2*\mathbf{2^{20}}$ = 2,097,152 bytes for vectors. As identifying the prototype vector associated with each data vector requires only 2 bytes. Therefore, after compression 10,485,760 bytes are needed to represent the data.

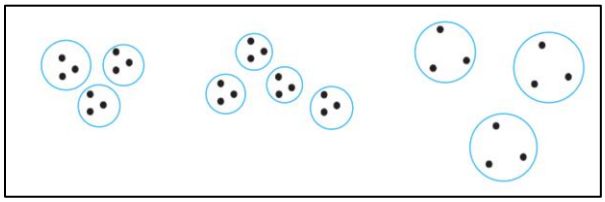
   The compression ratio is 12:8.

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

Figure 7.35.

Points for Exercise 2.

Answer:



3. **Many partitional clustering algorithms that automatically determine the number of clusters claim that this is an advantage. List two situations in which this is not the case.**

Answer: A partitional clustering is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. Except in following situations:

1) when the number of clusters calculated is greater than the system can handle.

2) When the dataset is known and running the algorithm doesn't return any additional information.

5. **Identify the clusters in Figure 7.36 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.**

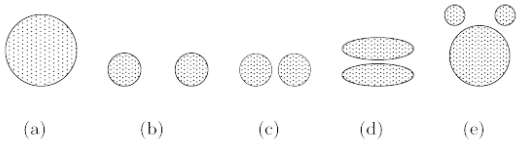
(a) center-based 2 clusters. The rectangular region will be split in half. Note that the noise is included in the two clusters.
contiguity-based 1 cluster because the two circular regions will be joined by noise.
density-based 2 clusters, one for each circular region. Noise will be eliminated.

(b) center-based 1 cluster that includes both rings.
contiguity-based 2 clusters, one for each ring.
density-based 2 clusters, one for each ring.

(c) center-based 3 clusters, one for each triangular region. One cluster is also an acceptable answer.
contiguity-based 1 cluster. The three triangular regions will be joined together because they touch.
density-based 3 clusters, one for each triangular region. Even though the three triangles touch, the density in the region where they touch is lower than throughout the interior of the triangles.

(d) center-based 2 clusters. The two groups of lines will be split in two.
contiguity-based 5 clusters. Each set of lines that intertwines becomes a cluster.
density-based 2 clusters. The two groups of lines define two regions of high density separated by a region of low density.

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).**



1. **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 1. 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 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 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.

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

3. **K=3. The distance between the edges of the circles is much less than the radii of the circles.**
   The three boxes show the three clusters that will result in the realistic case that the initial centroids are actual data points.

4. **K=2.**
   In both cases, the rectangles show the clusters. In the first case, the two clusters are only a local minimum while in the second case the clusters represent a globally minimal solution.

5. **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.**
   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. (The two smaller clusters in the drawing are supposed to be symmetrical.) I believe that the second solution—suggested by a student—is also possible, although it is a local minimum and might rarely be seen in practice for this configuration of points. Note that while the two pie shaped cuts out of the larger circle are shown as meeting at a point, this is not necessarily the case—it depends on the exact positions and sizes of the circles. There could be a gap between the two pie shaped cuts which is filled by the third (larger) cluster. (Imagine the small circles on opposite sides.) Or the boundary between the two pie shaped cuts could be a line segment.

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.

Answer: The correct answer is (C). Less dense regions require more centroids if the squared error is to be minimized.

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?

(a) 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.

(b) if the SSE of one attribute is relatively low for just one cluster, then this attribute helps define the cluster.

(c) If the SSE of an attribute is relatively high for all clusters, then it could well mean that the attribute is noise.

(d) 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. It could merely be the case that the clusters defined by this attribute are different from those defined by the other attributes, but in any case, it means that this attribute does not help define the cluster.

(e) 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. Note that attributes with high SSE for all clusters are particularly troublesome if they have a relatively high SSE with respect to other attributes (perhaps because of their scale) since they introduce a lot of noise into the computation of the overall SSE.

14. You are given a data set with 100 records and are asked to cluster the data. You use K-means to cluster the data, but for all values of K, $1 \leq K \leq 100$, the K-means algorithm returns only one non-empty cluster. You then apply an incremental version of K-means but obtain exactly the same result. How is this possible? How would single link or DBSCAN handle such data?

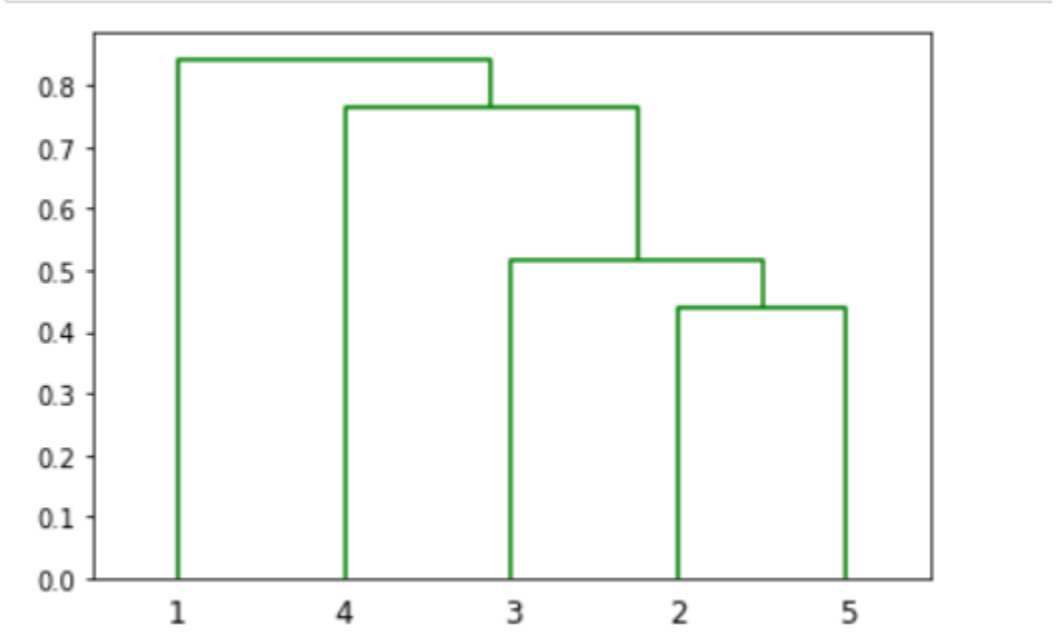
(a) The data consists completely of duplicates of one object.

(b) Single link (and many of the other agglomerative hierarchical schemes) would produce a hierarchical clustering, but which points appear in which cluster would depend on the ordering of the points and the exact algorithm. However, if the dendrogram were plotted showing the proximity at which each object is merged, then it would be obvious that the data consisted of duplicates. DBSCAN would find that all points were core points connected to one another and produce a single cluster.

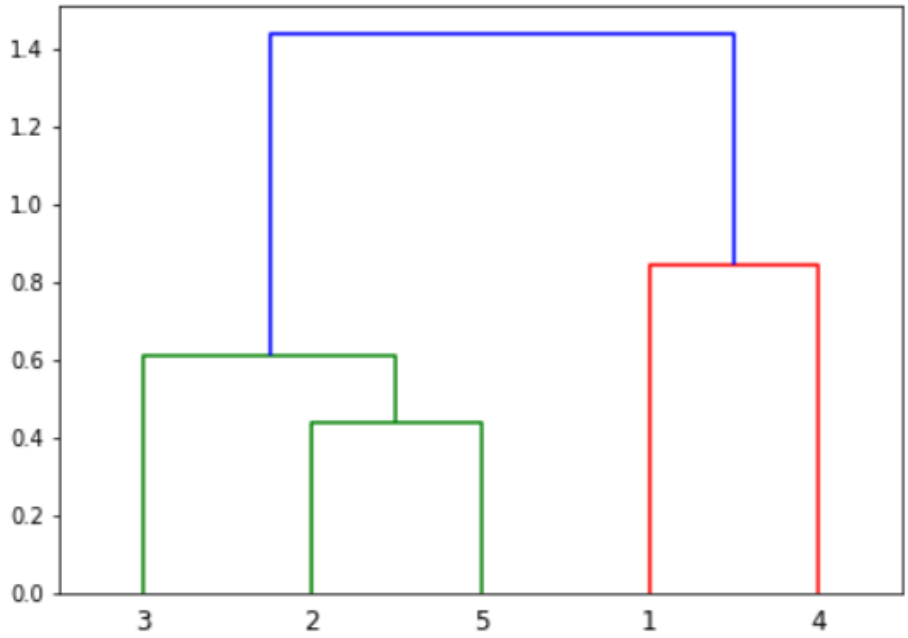
16. Use the similarity matrix in Table 7.13 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. Table 7.13. Similarity matrix for Exercise 16.

	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

Single link



Complete Link:



30. **Clusters of documents can be summarized by finding the top terms (words) for the documents in the cluster, e.g., by taking the most frequent k terms, where k is a constant, say 10, or by taking all terms that occur more frequently than a specified threshold. Suppose that K-means is used to find clusters of both documents and words for a document data set.**
    1. **How might a set of term clusters defined by the top terms in a document cluster differ from the word clusters found by clustering the terms with K-means?**

       First, the top words clusters could, and likely would, overlap somewhat. Second, it is likely that many terms would not appear in any of the clusters formed by the top terms. In contrast, a K-means clustering of the terms would cover all the terms and would not be overlapping.

    2. **How could term clustering be used to define clusters of documents?**

       An obvious approach would be to take the top documents for a term cluster; i.e., those documents that most frequently contain the terms in the cluster.