Cluster Analysis: Basic Concepts and Algorithms

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?

   Before compression, the data set requires $4 \times 32 \times 2^{20} = 134,217,728$ bytes. After compression, the data set requires $4 \times 32 \times 2^{16} = 8,388,608$ bytes for the prototype vectors and $2 \times 2^{20} = 2,097,152$ bytes for vectors, since identifying the prototype vector associated with each data vector requires only two bytes. Thus, 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 8.1.

   The solutions are also indicated in Figure 8.1.

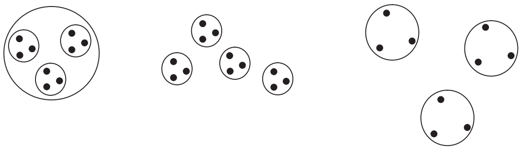


Figure 8.1. Points for Exercise 2.

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.

   (a) When there is hierarchical structure in the data. Most algorithms that automatically determine the number of clusters are partitional, and thus, ignore the possibility of subclusters.

   (b) When clustering for utility. If a certain reduction in data size is needed, then it is necessary to specify how many clusters (cluster centroids) are produced.

4. Given K equally sized clusters, the probability that a randomly chosen initial centroid will come from any given cluster is $1/K$, but the probability that each cluster will have exactly one initial centroid is much lower. (It should be clear that having one initial centroid in each cluster is a good starting situation for K-means.) In general, if there are K clusters and each cluster has n points, then the probability, p, of selecting in a sample of size K one initial centroid from each cluster is given by Equation 8.1. (This assumes sampling with replacement.) From this formula we can calculate, for example, that the chance of having one initial centroid from each of four clusters is $4!/4^4 = 0.0938$.

$$p = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K} \qquad (8.1)$$

   (a) Plot the probability of obtaining one point from each cluster in a sample of size K for values of K between 2 and 100.

   The solution is shown in Figure 4. Note that the probability is essentially 0 by the time $K = 10$.

   (b) For K clusters, $K = 10, 100$, and 1000, find the probability that a sample of size $2K$ contains at least one point from each cluster. You can use either mathematical methods or statistical simulation to determine the answer.

   We used simulation to compute the answer. Respectively, the probabilities are 0.21, $< 10^{-6}$, and $< 10^{-6}$.

   Proceeding analytically, the probability that a point doesn't come from a particular cluster is, $1 - \frac{1}{K}$, and thus, the probability that all $2K$ points don't come from a particular cluster is $(1 - \frac{1}{K})^{2K}$. Hence, the probability that at least one of the 200 points comes from a particular cluster is $1 - (1 - \frac{1}{K})^{2K}$. If we assume independence (which is too optimistic, but becomes approximately true for larger values of K), then an upper bound for the probability that all clusters are represented in the final sample is given by $(1 - (1 - \frac{1}{K})^{2K})^K$. The values given by this bound are 0.27, 5.7e-07, and 8.2e-64, respectively.

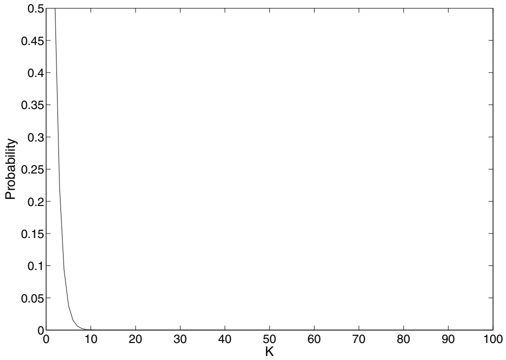


Figure 8.2. Probability of at least one point from each cluster. Exercise 4.

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.

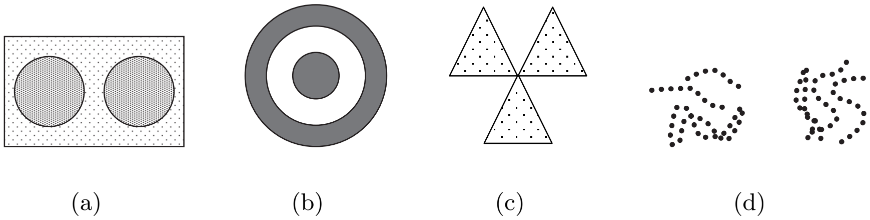


Figure 8.3. Clusters for Exercise 5.

(a) **center-based** 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 rings.
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 8.4 matches the corresponding part of this question, e.g., Figure 8.4(a) goes with part (a).

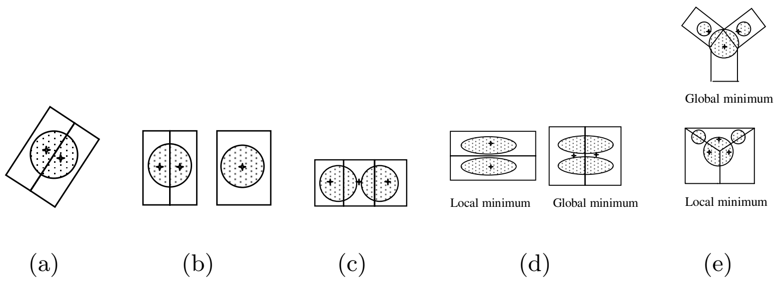


Figure 8.4. Diagrams for Exercise 6.

(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 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 $0° \leq \theta \leq 180°$ 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.

(b) $K = 3$. 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.

(c) $K = 3$. 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.

(d) $K = 2$.

In both case, 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.

(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.

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 actually 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.

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

8. Consider the mean of a cluster of objects from a binary transaction data set. What are the minimum and maximum values of the components of the mean? What is the interpretation of components of the cluster mean? Which components most accurately characterize the objects in the cluster?

   (a) The components of the mean range between 0 and 1.

   (b) For any specific component, its value is the fraction of the objects in the cluster that have a 1 for that component. If we have asymmetric binary data, such as market basket data, then this can be viewed as the probability that, for example, a customer in group represented by the the cluster buys that particular item.

   (c) This depends on the type of data. For binary asymmetric data, the components with higher values characterize the data, since, for most clusters, the vast majority of components will have values of zero. For regular binary data, such as the results of a true-false test, the significant components are those that are unusually high or low with respect to the entire set of data.

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.

   Consider a data set that consists of three circular clusters, that are identical in terms of the number and distribution of points, and whose centers lie on a line and are located such that the center of the middle cluster is equally distant from the other two. Bisecting K-means would always split the middle cluster during its first iteration, and thus, could never produce the correct set of clusters. (Postprocessing could be applied to address this.)

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?

Time series data is dense high-dimensional data, and thus, the cosine measure would not be appropriate since the cosine measure is appropriate for sparse data. If the magnitude of a time series is important, then Euclidean distance would be appropriate. If only the shapes of the time series are important, then correlation would be appropriate. Note that if the comparison of the time series needs to take in account that one time series might lead or lag another or only be related to another during specific time periods, then more sophisticated approaches to modeling time series similarity must be used.

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.

12. The leader algorithm (Hartigan [4]) 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.

    Note that the algorithm described here is not quite the leader algorithm described in Hartigan, which assigns a point to the first leader that is within the threshold distance. The answers apply to the algorithm as stated in the problem.

    (a) What are the advantages and disadvantages of the leader algorithm as compared to K-means?

The leader algorithm requires only a single scan of the data and is thus more computationally efficient since each object is compared to the final set of centroids at most once. Although the leader algorithm is order dependent, for a fixed ordering of the objects, it always produces the same set of clusters. However, unlike K-means, it is not possible to set the number of resulting clusters for the leader algorithm, except indirectly. Also, the K-means algorithm almost always produces better quality clusters as measured by SSE.

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

Use a sample to determine the distribution of distances between the points. The knowledge gained from this process can be used to more intelligently set the value of the threshold.

The leader algorithm could be modified to cluster for several thresholds during a single pass.

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) If we have K K-means clusters, then the plane is divided into K Voronoi regions that represent the points closest to each centroid.

(b) The boundaries between clusters are piecewise linear. It is possible to see this by drawing a line connecting two centroids and then drawing a perpendicular to the line halfway between the centroids. This perpendicular line splits the plane into two regions, each containing points that are closest to the centroid the region contains.

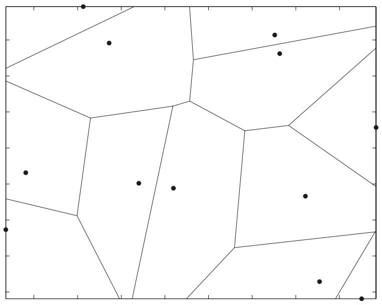


Figure 8.5. Voronoi diagram for Exercise 13.

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.

15. Traditional agglomerative hierarchical clustering routines merge two clusters at each step. Does it seem likely that such an approach accurately captures the (nested) cluster structure of a set of data points? If not, explain how you might postprocess the data to obtain a more accurate view of the cluster structure.

    (a) Such an approach does not accurately capture the nested cluster structure of the data. For example, consider a set of three clusters, each of which has two, three, and four subclusters, respectively. An ideal hierarchical clustering would have three branches from the root—one to each of the three main clusters—and then two, three, and four branches from each of these clusters, respectively. A traditional agglomerative approach cannot produce such a structure.

    (b) The simplest type of postprocessing would attempt to flatten the hierarchical clustering by moving clusters up the tree.

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.

    The solutions are shown in Figures 8.6(a) and 8.6(b).

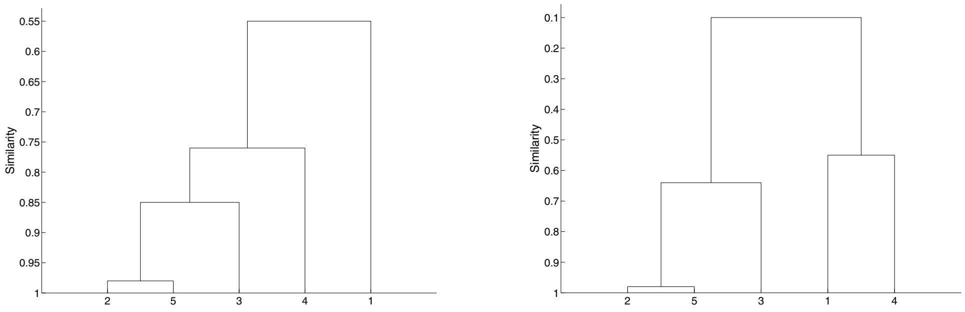
17. Hierarchical clustering is sometimes used to generate K clusters, $K > 1$ by taking the clusters at the K^{th} level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, 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

Table 8.1. 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



(a) Single link.

(b) Complete link.

Figure 8.6. Dendrograms for Exercise 16.

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 is 6, 12, 18, 24, 30.
Error = 360.
Second cluster is 42, 48.
Error = 18.
Total Error = 378

ii. $\{15, 40\}$ First cluster is 6, 12, 18, 24 .
Error = 180.
Second cluster is 30, 42, 48.
Error = 168.
Total 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?

Yes, both centroids are stable solutions.

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

The two clusters are {6, 12, 18, 24, 30} and {42, 48}.

(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.)

MIN produces the most natural clustering.

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

MIN produces contiguous clusters. However, density is also an acceptable answer. Even center-based is acceptable, since one set of centers gives the desired clusters.

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

K-means is not good at finding clusters of different sizes, at least when they are not well separated. The reason for this is that the objective of minimizing squared error causes it to "break" the larger cluster. Thus, in this problem, the low error clustering solution is the "unnatural" one.

18. Suppose we find K clusters using Ward's method, bisecting K-means, and ordinary K-means. Which of these solutions represents a local or global minimum? Explain.

Although Ward's method picks a pair of clusters to merge based on minimizing SSE, there is no refinement step as in regular K-means. Likewise, bisecting K-means has no overall refinement step. Thus, unless such a refinement step is added, neither Ward's method nor bisecting K-means produces a local minimum. Ordinary K-means produces a local minimum, but like the other two algorithms, it is not guaranteed to produce a global minimum.

19. Hierarchical clustering algorithms require $O(m^2 \log(m))$ time, and consequently, are impractical to use directly on larger data sets. One possible technique for reducing the time required is to sample the data set. For example, if K clusters are desired and $\sqrt{m}$ points are sampled from the m points, then a hierarchical clustering algorithm will produce a hierarchical clustering in roughly $O(m)$ time. K clusters can be extracted from this hierarchical clustering by taking the clusters on the K^{th} level of the dendrogram. The remaining points can then be assigned to a cluster in linear time, by using various strategies. To give a specific example, the centroids of the K clusters can be computed, and then each of the $m - \sqrt{m}$ remaining points can be assigned to the cluster associated with the closest centroid.

For each of the following types of data or clusters, discuss briefly if (1) sampling will cause problems for this approach and (2) what those problems are. Assume that the sampling technique randomly chooses points from the total set of m points and that any unmentioned characteristics of the data or clusters are as optimal as possible. In other words, focus only on problems caused by the particular characteristic mentioned. Finally, assume that K is very much less than m.

(a) Data with very different sized clusters.

This can be a problem, particularly if the number of points in a cluster is small. For example, if we have a thousand points, with two clusters, one of size 900 and one of size 100, and take a 5% sample, then we will, on average, end up with 45 points from the first cluster and 5 points from the second cluster. Five points is much easier to miss or cluster improperly than 50. Also, the second cluster will sometimes be represented by fewer than 5 points, just by the nature of random samples.

(b) High-dimensional data.

This can be a problem because data in high-dimensional space is typically sparse and more points may be needed to define the structure of a cluster in high-dimensional space.

(c) Data with outliers, i.e., atypical points.

By definition, outliers are not very frequent and most of them will be omitted when sampling. Thus, if finding the correct clustering depends on having the outliers present, the clustering produced by sampling will likely be misleading. Otherwise, it is beneficial.

(d) Data with highly irregular regions.

This can be a problem because the structure of the border can be lost when sampling unless a large number of points are sampled.

(e) Data with globular clusters.

This is typically not a problem since not as many points need to be sampled to retain the structure of a globular cluster as an irregular one.

(f) Data with widely different densities.

In this case the data will tend to come from the denser region. Note that the effect of sampling is to reduce the density of all clusters by the sampling factor, e.g., if we take a 10% sample, then the density of the clusters is decreased by a factor of 10. For clusters that aren't very dense to begin with, this may means that they are now treated as noise or outliers.

(g) Data with a small percentage of noise points.

Sampling will not cause a problem. Actually, since we would like to exclude noise, and since the amount of noise is small, this may be beneficial.

(h) Non-Euclidean data.

This has no particular impact.

(i) Euclidean data.

This has no particular impact.

(j) Data with many and mixed attribute types.

Many attributes was discussed under high-dimensionality. Mixed attributes have no particular impact.

20. Consider the following four faces shown in Figure 8.7. Again, darkness or number of dots represents density. Lines are used only to distinguish regions and do not represent points.

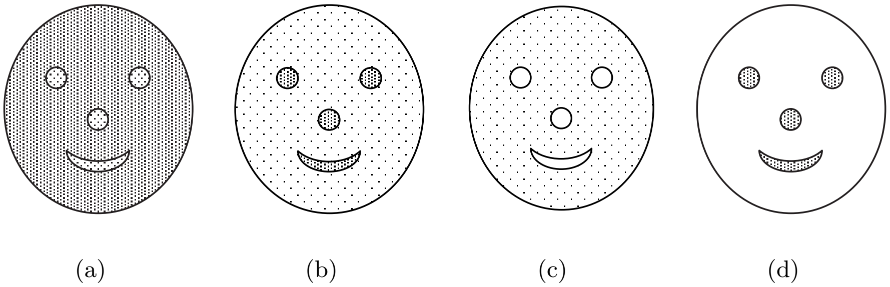


Figure 8.7. Figure for Exercise 20.

(a) For each figure, could you use single link to find the patterns represented by the nose, eyes, and mouth? Explain.

Only for (b) and (d). For (b), the points in the nose, eyes, and mouth are much closer together than the points between these areas. For (d) there is only space between these regions.

(b) For each figure, could you use K-means to find the patterns represented by the nose, eyes, and mouth? Explain.

Only for (b) and (d). For (b), K-means would find the nose, eyes, and mouth, but the lower density points would also be included. For (d), K-means would find the nose, eyes, and mouth straightforwardly as long as the number of clusters was set to 4.

(c) What limitation does clustering have in detecting all the patterns formed by the points in Figure 8.7(c)?

Clustering techniques can only find patterns of points, not of empty spaces.

21. Compute the entropy and purity for the confusion matrix in Table 8.2.

Table 8.2. Confusion matrix for Exercise 21.

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Total	Entropy	Purity
#1	1	1	0	11	4	676	693	0.20	0.98
#2	27	89	333	827	253	33	1562	1.84	0.53
#3	326	465	8	105	16	29	949	1.70	0.49
Total	354	555	341	943	273	738	3204	1.44	0.61

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?

Yes. The random points will have regions of lesser or greater density, while the uniformly distributed points will, of course, have uniform density throughout the unit square.

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

The random set of points will have a lower SSE.

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

DBSCAN will merge all points in the uniform data set into one cluster or classify them all as noise, depending on the threshold. There might be some boundary issues for points at the edge of the region. However, DBSCAN can often find clusters in the random data, since it does have some variation in density.

23. Using the data in Exercise 24, compute the silhouette coefficient for each point, each of the two clusters, and the overall clustering.

Cluster 1 contains {P1, P2}, Cluster 2 contains {P3, P4}. The dissimilarity matrix that we obtain from the similarity matrix is the following:

Table 8.3. Table of distances for Exercise 23

.

	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Let a indicate the average distance of a point to other points in its cluster. Let b indicate the minimum of the average distance of a point to points in another cluster.

Point P1: SC = 1- a/b = 1 - 0.1/((0.65+0.55)/2)= 5/6 = 0.833
Point P2: SC = 1- a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846
Point P2: SC = 1- a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556
Point P2: SC = 1- a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478

Cluster 1 Average SC = (0.833+0.846)/2 = 0.84
Cluster 2 Average SC = (0.556+0.478)/2 = 0.52
Overall Average SC = (0.840+0.517)/2 = 0.68

24. Given the set of cluster labels and similarity matrix shown in Tables 8.4 and 8.5, respectively, compute the correlation between the similarity matrix and the ideal similarity matrix, i.e., the matrix whose ij^{th} entry is 1 if two objects belong to the same cluster, and 0 otherwise.

Table 8.4. Table of cluster labels for Exercise 24.

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Table 8.5. Similarity matrix for Exercise 24.

Point	P1	P2	P3	P4
P1	1	0.8	0.65	0.55
P2	0.8	1	0.7	0.6
P3	0.65	0.7	1	0.9
P4	0.55	0.6	0.9	1

We need to compute the correlation between the vector $\mathbf{x} =< 1, 0, 0, 0, 0, 1 >$ and the vector $\mathbf{y} =< 0.8, 0.65, 0.55, 0.7, 0.6, 0.3 >$, which is the correlation between the off-diagonal elements of the distance matrix and the ideal similarity matrix.

We get:
Standard deviation of the vector $\mathbf{x}$: $\sigma_x = 0.5164$
Standard deviation of the vector $\mathbf{y}$: $\sigma_y = 0.1703$
Covariance of $\mathbf{x}$ and $\mathbf{y}$: $\mathrm{cov}(\mathbf{x}, \mathbf{y}) = -0.200$

Therefore, $\text{corr}(\mathbf{x}, \mathbf{y}) = \text{cov}(\mathbf{x}, \mathbf{y})/\sigma_x\sigma_y = -0.227$

25. Compute the hierarchical F-measure for the eight objects {p1, p2, p3, p4, p5, p6, p7, p8} and hierarchical clustering shown in Figure 8.8. Class A contains points p1, p2, and p3, while p4, p5, p6, p7, and p8 belong to class B.

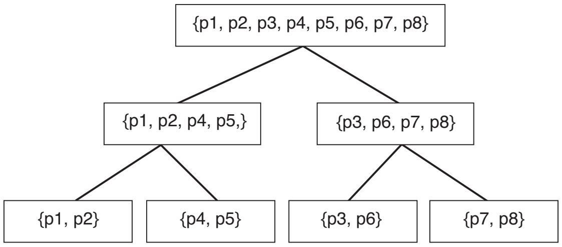


Figure 8.8. Hierarchical clustering for Exercise 25.

Let $R(i, j) = n_{ij}/n_i$ indicate the recall of class i with respect to cluster j.
Let $P(i, j) = n_{ij}/n_j$ indicate the precision of class i with respect to cluster j.
$F(i, j) = 2R(i, j) \times P(i, j)/(R(i, j) + P(i, j))$ is the F-measure for class i and cluster j.

For cluster #1 = $\{p1, p2, p3, p4, p5, p6, p7, p8\}$:
Class = A:
$R(A, 1) = 3/3 = 1,\ P(A, 1) = 3/8 = 0.375$
$F(A, 1) = 2 \times 1 \times 0.375/(1 + 0.375) = 0.55$
Class = B:
$R(B, 1) = 5/5 = 1,\ P(A, 1) = 5/8 = 0.625,\ F(A, 1) = 0.77$

For cluster #2= {p1,p2,p4,p5}
Class = A:
$R(A, 2) = 2/3, P(A, 2) = 2/4, F(A, 2) = 0.57$
Class = B:
$R(B, 2) = 2/5, P(B, 2) = 2/4, F(B, 2) = 0.44$

For cluster #3= {p3, p6, p7, p8}
Class = A:
$R(A, 3) = 1/3,\ P(A, 3) = 1/4,\ F(A, 3) = 0.29$
Class =B:
$R(B, 3) = 3/5, P(B, 3) = 3/4, F(B, 3) = 0.67$

For cluster #4={p1, p2}
Class = A:

$R(A,4) = 2/3,\ P(A,4) = 2/2,\ F(A,4) = 0.8$
Class =B:
$R(B,4) = 0/5,\ P(B,4) = 0/2,\ F(B,4) = 0$

For cluster #5 = {p4, p5}
Class = A:
$R(A,5) = 0,\ P(A,5) = 0,\ F(A,5) = 0$
Class =B:
$R(B,5) = 2/5,\ P(B,5) = 2/2,\ F(B,5) = 0.57$

For cluster #6 = {p3, p6}
Class = A:
$R(A,6) = 1/3,\ P(A,6) = 1/2,\ F(A,6) = 0.4$
Class =B:
$R(B,6) = 1/5,\ P(B,6) = 1/2,\ F(B,6) = 0.29$

For cluster #7 = {p7, p8}
Class = A:
$R(A,7) = 0,\ P(A,7) = 1,\ F(A,7) = 0$
Class = B:
$R(B,7) = 2/5,\ P(B,7) = 2/2,\ F(B,7) = 0.57$

Class A: $F(A) = \max\{F(A,j)\} = \max\{0.55, 0.57, 0.29, 0.8, 0, 0.4, 0\} = 0.8$
Class B: $F(B) = \max\{F(B,j)\} = \max\{0.77, 0.44, 0.67, 0, 0.57, 0.29, 0.57\} = 0.77$
Overall Clustering: $F = \sum_1^2 \frac{n_i}{n} \max_i F(i,j) = 3/8 * F(A) + 5/8 * F(B) = 0.78$

26. Compute the cophenetic correlation coefficient for the hierarchical clusterings in Exercise 16. (You will need to convert the similarities into dissimilarities.)

    This can be easily computed using a package, e.g., MATLAB. The answers are single link, 0.8116, and complete link, 0.7480.

27. Prove Equation 8.14.

$$
\begin{aligned}
\frac{1}{2|C_i|}\sum_{x\in C_i}\sum_{y\in C_i}(x-y)^2 &= \frac{1}{2|C_i|}\sum_{x\in C_i}\sum_{y\in C_i}((x-c_i)-(y-c_i))^2\\
&= \frac{1}{2|C_i|}\Big(\sum_{x\in C_i}\sum_{y\in C_i}(x-c_i)^2 - 2\sum_{x\in C_i}\sum_{y\in C_i}(x-c_i)(y-c_i)\\
&+ \sum_{x\in C_i}\sum_{y\in C_i}(y-c_i)^2\Big)\\
&= \frac{1}{2|C_i|}\Big(\sum_{x\in C_i}\sum_{y\in C_i}(x-c_i)^2 + \sum_{x\in C_i}\sum_{y\in C_i}(y-c_i)^2\Big)\\
&= \frac{1}{|C_i|}\sum_{x\in C_i}|C_i|(x-c_i)^2\\
&= \text{SSE}
\end{aligned}
$$

The cross term $\sum_{x\in C_i}\sum_{y\in C_i}(x-c_i)(y-c_i)$ is 0.

28. Prove Equation 8.15.

$$
\begin{aligned}
\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(c_j-c_i)^2 &= \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|((m-c_i)-(m-c_j))^2\\
&= \frac{1}{2K}\Big(\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(m-c_i)^2 - 2\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(m-c_i)(m-c_j)\\
&+ \sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(m-c_j)^2\Big)\\
&= \frac{1}{2K}\Big(\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(m-c_i)^2 + \sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(m-c_j)^2\Big)\\
&= \frac{1}{K}\sum_{i=1}^{K}K|C_i|(m-c_i)^2\\
&= \text{SSB}
\end{aligned}
$$

Again, the cross term cancels.

29. Prove that $\sum_{i=1}^{K}\sum_{x\in C_i}(x-m_i)(m-m_i) = 0$. This fact was used in the proof that TSS = SSE + SSB on page 557.

$$\begin{aligned}
\sum_{i=1}^{K}\sum_{x\in C_i}(x-c_i)(c-c_i) &= \sum_{i=1}^{K}\sum_{x\in C_i}(xc-c_ic-xc_i+c_i^2)\\
&= \sum_{i=1}^{K}\sum_{x\in C_i}xc-\sum_{i=1}^{K}\sum_{x\in C_i}c_ic-\sum_{i=1}^{K}\sum_{x\in C_i}xc_i+\sum_{i=1}^{K}\sum_{x\in C_i}c_i^2\\
&= m_ic_ic-m_ic_ic-m_ic_i^2+m_ic_i^2\\
&= 0
\end{aligned}$$

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.

    (a) 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.

    (b) 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.

31. We can represent a data set as a collection of object nodes and a collection of attribute nodes, where there is a link between each object and each attribute, and where the weight of that link is the value of the object for that attribute. For sparse data, if the value is 0, the link is omitted. Bipartite clustering attempts to partition this graph into disjoint clusters, where each cluster consists of a set of object nodes and a set of attribute nodes. The objective is to maximize the weight of links between the object and attribute nodes of a cluster, while minimizing the weight of links between object and attribute links in different clusters. This type of clustering is also known as **co-clustering** since the objects and attributes are clustered at the same time.

    (a) How is bipartite clustering (co-clustering) different from clustering the sets of objects and attributes separately?

    In regular clustering, only one set of constraints, related either to objects or attributes, is applied. In co-clustering both sets of constraints are applied simultaneously. Thus, partitioning the objects and attributes independently of one another typically does not produce the same results.

(b) Are there any cases in which these approaches yield the same clusters?

Yes. For example, if a set of attributes is associated only with the objects in one particular cluster, i.e., has 0 weight for objects in all other clusters, and conversely, the set of objects in a cluster has 0 weight for all other attributes, then the clusters found by co-clustering will match those found by clustering the objects and attributes separately. To use documents as an example, this would correspond to a document data set that consists of groups of documents that only contain certain words and groups of words that only appear in certain documents.

(c) What are the strengths and weaknesses of co-clustering as compared to ordinary clustering?

Co-clustering automatically provides a description of a cluster of objects in terms of attributes, which can be more useful than a description of clusters as a partitioning of objects. However, the attributes that distinguish different clusters of objects, may overlap significantly, and in such cases, co-clustering will not work well.

32. In Figure 8.9, match the similarity matrices, which are sorted according to cluster labels, with the sets of points. Differences in shading and marker shape distinguish between clusters, and each set of points contains 100 points and three clusters. In the set of points labeled 2, there are three very tight, equal-sized clusters.

Answers: 1 - D, 2 - C, 3 - A, 4 - B

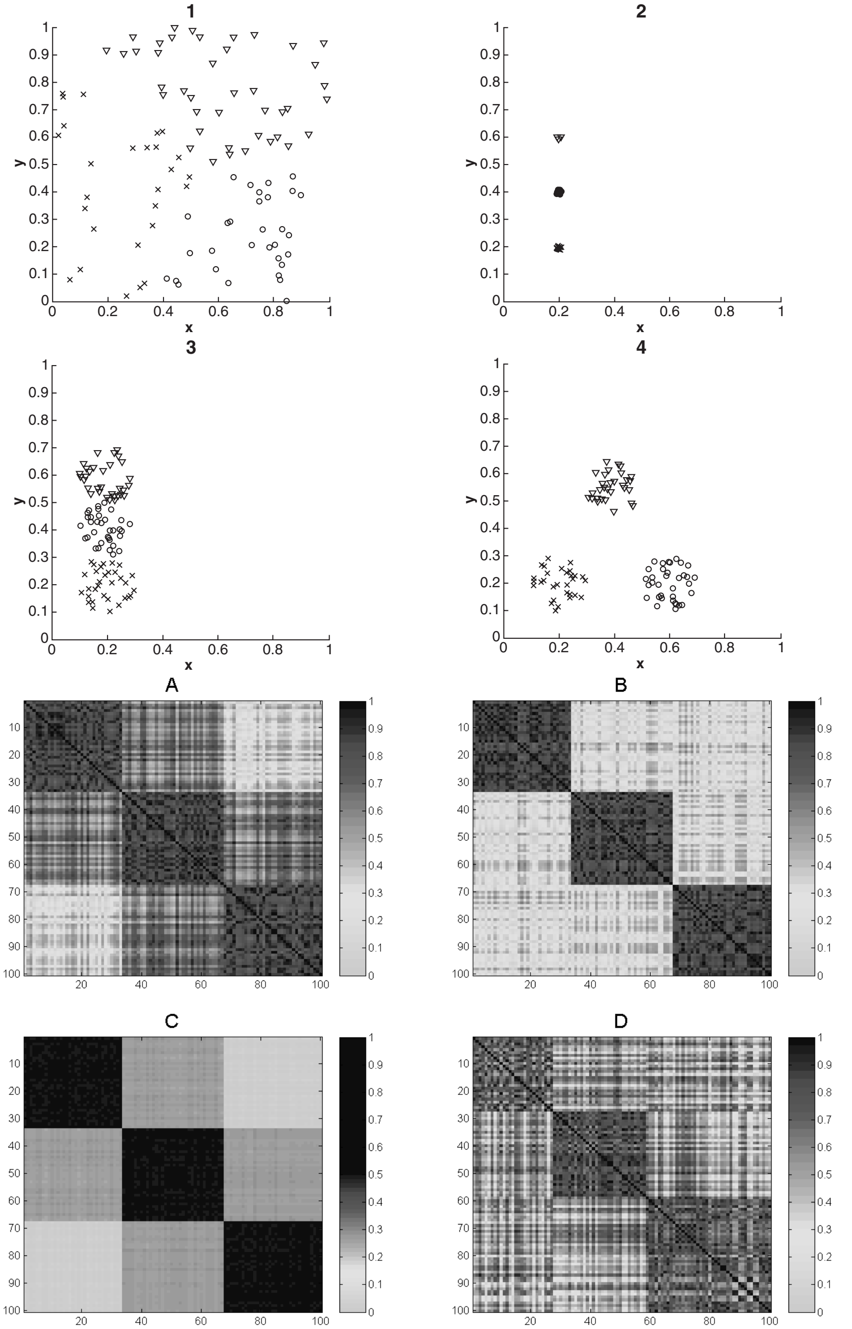


Figure 8.9. Points and similarity matrices for Exercise 32.

Cluster Analysis: Additional Issues and Algorithms

1. For sparse data, discuss why considering only the presence of non-zero values might give a more accurate view of the objects than considering the actual magnitudes of values. When would such an approach not be desirable?

   Consider document data. Intuitively, two documents are similar if they contain many of the same words. Although we can also include the frequency with which those words occur in the similarity computation, this can sometimes give a less reliable assessment of similarity. In particular, if one of the words in a document occurs rather frequently compared to other words, then this word can dominate the similarity comparison when magnitudes are taken into account. In that case, the document will only be highly similar to other documents that also contain the same word with a high frequency. While this may be appropriate in many or even most cases, it may lead to the wrong conclusion if the word can appear in different contexts, that can only be distinguished by other words. For instance, the word, 'game,' appears frequently in discussions of sports and video games.

2. Describe the change in the time complexity of K-means as the number of clusters to be found increases.

   As the number of clusters increases, the complexity of K-means approaches $O(m^2)$.

3. Consider a set of documents. Assume that all documents have been normalized to have unit length of 1. What is the "shape" of a cluster that consists of all documents whose cosine similarity to a centroid is greater than some specified constant? In other words, $\cos(d, c) \geq \delta$, where $0 < \delta \leq 1$.

Once document vectors have been normalized, they lie on am n-dimensional hypershpere. The constraint that all documents have a minimum cosine similarity with respect to a centroids is a constraint that the document vectors lie within a cone, whose intersection with the sphere is a circle on the surface of the sphere.

4. Discuss the advantages and disadvantages of treating clustering as an optimization problem. Among other factors, consider efficiency, non-determinism, and whether an optimization-based approach captures all types of clusterings that are of interest.

   Two key advantage to treating clustering as an optimization problem are that (1) it provides a clear definition of what the clustering process is doing, and (2) it allows the use of powerful optimization techniques that have been developed in a wide variety of fields. Unfortunately, most of these optimization techniques have a high time complexity. Furthermore, it can be shown that many optimization problems are NP hard, and therefore, it is necessary to use heuristic optimization approaches that can only guarantee a locally optimal solution. Often such techniques work best when used with random initialization, and thus, the solution found can vary from one run to another. Another problem with optimization approaches is that the objective functions they use tend to favor large clusters at the expense of smaller ones.

5. What is the time and space complexity of fuzzy c-means? Of SOM? How do these complexities compare to those of K-means?

   The time complexity of K-means $O(I * K * m * n)$, where I is the number of iterations required for convergence, K is the number of clusters, m is the number of points, and n is the number of attributes. The time required by fuzzy c-means is basically the same as K-means, although the constant is much higher. The time complexity of SOM is also basically the same as K-means because it consists of multiple passes in which objects are assigned to centroids and the centroids are updated. However, because the surrounding centroids are also updated and the number of passes can be large, SOM will typically be slower than K-means.

6. Traditional K-means has a number of limitations, such as sensitivity to outliers and difficulty in handling clusters of different sizes and densities, or with non-globular shapes. Comment on the ability of fuzzy c-means to handle these situations.

   Fuzzy c-means has all the limitations of traditional K-means, except that it does not make a hard assignment of an object to a cluster.

7. For the fuzzy c-means algorithm described in this book, the sum of the membership degree of any point over all clusters is 1. Instead, we could only require that the membership degree of a point in a cluster be between 0 and 1. What are the advantages and disadvantages of such an approach?

The main advantage of this approach occurs when a point is an outlier and does not really belong very strongly to any cluster, since in that situation, the point can have low membership in all clusters. However, this approach is often harder to initialize properly and can perform poorly when the clusters are not are not distinct. In that case, several cluster centers may merge together, or a cluster center may vary significantly from one iteration to another, instead of changing only slightly, as in ordinary K-means or fuzzy c-means.

8. Explain the difference between likelihood and probability.

   Probability is, according to one common statistical definition, the frequency with which an event occurs as the number of experiments tends to infinity. Probability is defined by a probability density function which is a function of the attribute values of an object. Typically, a probability density function depends on some parameters. Considering probability density function to be a function of the parameters yields the likelihood function.

9. Equation 9.12 gives the likelihood for a set of points from a Gaussian distribution as a function of the mean μ and the standard deviation σ. Show mathematically that the maximum likelihood estimate of μ and σ are the sample mean and the sample standard deviation, respectively.

   First, we solve for μ.

$$\begin{aligned}\frac{\partial \ell((\mu,\sigma)|\mathcal{X})}{\partial \mu} &= \frac{\partial}{\partial \mu}\left(-\sum_{i=1}^{m}\frac{(x_i-\mu)^2}{2\sigma^2} - 0.5m\log 2\pi - m\log\sigma\right)\\ &= -\sum_{i=1}^{m}\frac{2(x_i-\mu)}{2\sigma^2}\end{aligned}$$

   Setting this equal to 0 and solving, we get $\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$.

   Likewise, we can solve for σ.

$$\begin{aligned}\frac{\partial \ell((\mu,\sigma)|\mathcal{X})}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left(-\sum_{i=1}^{m}\frac{(x_i-\mu)^2}{2\sigma^2} - 0.5m\log 2\pi - m\log\sigma\right)\\ &= \sum_{i=1}^{m}\frac{2(x_i-\mu)^2}{2\sigma^3} - \frac{m}{\sigma}\end{aligned}$$

   Setting this equal to 0 and solving, we get $\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu)^2$.

10. We take a sample of adults and measure their heights. If we record the gender of each person, we can calculate the average height and the variance of the height, separately, for men and women. Suppose, however, that this information was not recorded. Would it be possible to still obtain this information? Explain.

The height of men and women will have separate Gaussian distributions with different means and perhaps different variances. By using a mixture model approach, we can obtain estimates of the mean and variance of the two height distributions. Given a large enough sample size, the estimated parameters should be close to those that could be computed if we knew the gender of each person.

11. Compare the membership weights and probabilities of Figures 9.1 and 9.4, which come, respectively, from applying fuzzy and EM clustering to the same set of data points. What differences do you detect, and how might you explain these differences?

    The fuzzy clustering approach only assigns very high weights to those points in the center of the clusters. Those points that are close to two or three clusters have relatively low weights. The points that are on the far edges of the clusters, away from other clusters also have lower weights than the center points, but not as low as points that are near two or three clusters.

    The EM clustering approach assigns high weights both to points in the center of the clusters and those on the far edges. The points that are near two or three clusters have lower weights, but not as much so as with the fuzzy clustering procedure.

    The main difference between the approaches is that as a point on the far edge of a cluster gets further away from the center of the cluster, the weight with which is belongs to a cluster becomes more equal among clusters for the fuzzy clustering approach, but for the EM approach the point tends to belong more strongly to the cluster to which it is closest.

12. Figure 9.1 shows a clustering of a two-dimensional point data set with two clusters: The leftmost cluster, whose points are marked by asterisks, is somewhat diffuse, while the rightmost cluster, whose points are marked by circles, is compact. To the right of the compact cluster, there is a single point (marked by an arrow) that belongs to the diffuse cluster, whose center is farther away than that of the compact cluster. Explain why this is possible with EM clustering, but not K-means clustering.

    In EM clustering, we compute the probability that a point belongs to a cluster. In turn, this probability depends on both the distance from the cluster center and the spread (variance) of the cluster. Hence, a point that is closer to the centroid of one cluster than another can still have a higher probability with respect to the more distant cluster if that cluster has a higher spread than the closer cluster. K-means only takes into account the distance to the closest cluster when assigning points to clusters. This is equivalent to an EM approach where all clusters are assumed to have the same variance.

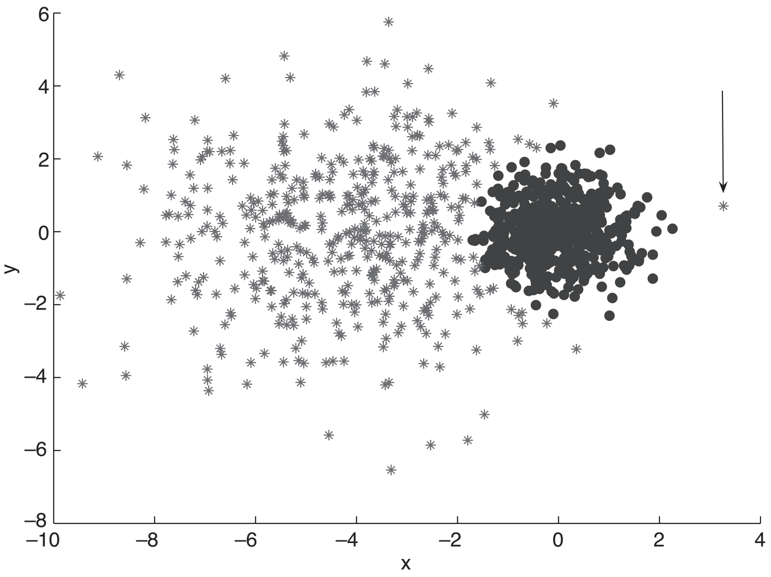


Figure 9.1. Data set for Exercise 12. EM clustering of a two-dimensional point set with two clusters of differing density.

13. Show that the MST clustering technique of Section 9.4.2 produces the same clusters as single link. To avoid complications and special cases, assume that all the pairwise similarities are distinct.

    In single link, we start with with clusters of individual points and then successively join two clusters that have the pair of points that are closest together. Conceptually, we can view the merging of the clusters as putting an edge between the two closest points of the two clusters. Note that if both clusters are currently connected, then the resulting cluster will also be connected. However, since the clusters are formed from disjoint sets of points, and edges are only placed between points in different clusters, no cycle can be formed. From these observations and by noting that we start with clusters (graphs) of size one that are vacuously connected, we can deduce by induction that at any stage in single link clustering process, each cluster consists of a connected set of points without any cycles. Thus, when the last two clusters are merged to form a cluster containing all the points, we also have a connected graph of all the points that is a spanning tree of the graph. Furthermore, since each point in the graph is connected to its nearest point, the spanning tree must be a minimum spanning tree. All that remains to establish the equivalence of MST and single link is to note that MST essentially reverses the process by which single link built the minimum spanning tree; i.e., by breaking edges beginning with the longest and proceeding until the smallest. Thus, it generates the same clusters as single link, but in reverse order.

14. One way to sparsify a proximity matrix is the following: For each object (row in the matrix), set all entries to 0 except for those corresponding to the objects k-nearest neighbors. However, the sparsified proximity matrix is typically not symmetric.

    (a) If object a is among the k-nearest neighbors of object b, why is b not guaranteed to be among the k-nearest neighbors of a?

    Consider a dense set of $k+1$ objects and another object, an outlier, that is farther from any of the objects than they are from each other. None of the objects in the dense set will have the outlier on their k-nearest neighbor list, but the outlier will have k of the objects from the dense set on its k-nearest neighbor list.

    (b) Suggest at least two approaches that could be used to make the sparsified proximity matrix symmetric.

    One approach is to set the ij^{th} entry to 0 if the ji^{th} entry is 0, or vice versa. Another approach is to set the ij^{th} entry to 1 if the ji^{th} entry is 1, or vice versa.

15. Give an example of a set of clusters in which merging based on the closeness of clusters leads to a more natural set of clusters than merging based on the strength of connection (interconnectedness) of clusters.

    An example of this is given in the Chameleon paper that can be found at http://www.cs.umn.edu/ karypis/publications/Papers/PDF/chameleon.pdf. The example consists of two narrow rectangles of points that are side by side. The top rectangle is split into two clusters, one much smaller than the other. Even though the two rectangles on the top are close, they are not strongly connected since the strong links between them are across a small area. On the other hand, the largest rectangle on the top and the rectangle on the bottom are strongly connected. Each individual connection is not as strong, because these two rectangles are not as close, but there are more of them because the area of connection is large. Thus, an approach based on connectivity will merge the largest rectangle on top with the bottom rectangle.

16. Table 9.1 lists the two nearest neighbors of four points.

    Calculate the SNN similarity between each pair of points using the definition of SNN similarity defined in Algorithm 9.10.

    The following is the SNN similarity matrix.

17. For the definition of SNN similarity provided by Algorithm 9.10, the calculation of SNN distance does not take into account the position of shared

Table 9.1. Two nearest neighbors of four points.

Point	First Neighbor	Second Neighbor
1	4	3
2	3	4
3	4	2
4	3	1

Table 9.2. Two nearest neighbors of four points.

Point	1	2	3	4
1	2	0	0	1
2	0	2	1	0
3	0	1	2	0
4	1	0	0	2

neighbors in the two nearest neighbor lists. In other words, it might be desirable to give higher similarity to two points that share the same nearest neighbors in the same or roughly the same order.

(a) Describe how you might modify the definition of SNN similarity to give higher similarity to points whose shared neighbors are in roughly the same order.

This can be done by assigning weights to the points based on their position in the nearest neighbor list. For example, we can weight the i^{th} point in the nearest neighbor list by $n - i + 1$. For each point, we then take the sum or product of its rank on both lists. These values are then summed to compute the similarity between the two objects. This approach was suggested by Jarvis and Patrick [5].

(b) Discuss the advantages and disadvantages of such a modification.

Such an approach is more complex. However, it is advantageous if it is the case that two objects are more similar if the shared neighbors are roughly of the same rank. Furthermore, it may also help to compensate for arbitrariness in the choice of k.

18. Name at least one situation in which you would *not* want to use clustering based on SNN similarity or density.

When you wish to cluster based on absolute density or distance.

19. Grid-clustering techniques are different from other clustering techniques in that they partition space instead of sets of points.

(a) How does this affect such techniques in terms of the description of the resulting clusters and the types of clusters that can be found?

In grid-based clustering, the clusters are described in terms of collections of adjacent cells. In some cases, as in CLIQUE, a more compact description is generated. In any case, the description of the clusters is given in terms of a region of space, not a set of objects. (However, such a description can easily be generated.) Because it is necessary to work in terms of rectangular regions, the shapes of non-rectangular clusters can only be approximated. However, the groups of adjacent cells can have holes.

(b) What kind of cluster can be found with grid-based clusters that cannot be found by other types of clustering approaches? (Hint: See Exercise 20 in Chapter 8, page 564.)

Typically, grid-based clustering techniques only pay attention to dense regions. However, such techniques could also be used to identify sparse or empty regions and thus find patterns of the absence of points. Note, however, that this would not be appropriate for a sparse data space.

20. In CLIQUE, the threshold used to find cluster density remains constant, even as the number of dimensions increases. This is a potential problem since density drops as dimensionality increases; i.e., to find clusters in higher dimensions the threshold has to be set at a level that may well result in the merging of low-dimensional clusters. Comment on whether you feel this is truly a problem and, if so, how you might modify CLIQUE to address this problem.

This is a real problem. A similar problem exists in association analysis. In particular, the support of association patterns with a large number of items is often low. To find such patterns using an algorithm such as Apriori is difficult because the low support threshold required results in a large number of association patterns with few items that are of little interest. In other words, association patterns with many items (patterns in higher-dimensional space) are interesting at support levels (densities) that do not make for interesting patterns when the size of the association pattern (number of dimensions) is low. One approach is to decrease the support threshold (density threshold) as the number of items (number of dimensions) is increased.

21. Given a set of points in Euclidean space, which are being clustered using the K-means algorithm with Euclidean distance, the triangle inequality can be used in the assignment step to avoid calculating all the distances of each point to each cluster centroid. Provide a general discussion of how this might work.

Charles Elkan presented the following theorem in his keynote speech at the Workshop on Clustering High-Dimensional Data at SIAM 2004.

Lemma 1:Let x be a point, and let b and c be centers.
If $d(b,c) \geq 2d(x,b)$ then $d(x,c) \geq d(x,b)$.

Proof:
We know $d(b,c) \leq d(b,x) + d(x,c)$.
So $d(b,c) - d(x,b) \leq d(x,c)$.
Now $d(b,c) - d(x,b) \geq 2d(x,b) - d(x,b) = d(x,b)$.
So $d(x,b) \leq d(x,c)$.

This theorem can be used to eliminate a large number of unnecessary distance calculations.

22. Instead of using the formula derived in CURE—see Equation 9.19—we could run a Monte Carlo simulation to directly estimate the probability that a sample of size s would contain at least a certain fraction of the points from a cluster. Using a Monte Carlo simulation compute the probability that a sample of size s contains 50% of the elements of a cluster of size 100, where the total number of points is 1000, and where s can take the values 100, 200, or 500.

This question should have said "contains *at least* 50%."

The results of our simulation consisting of 100,000 trials was 0, 0, and 0.54 for the sample sizes 100, 200, and 500 respectively.