Chapter 5

Clustering

5.1 Introduction

Chapters 3 and 4 describe how samples may be classified if a training set is available to use in the design of a classifier. However, there are many situations where the classes themselves are initially undefined. Given a set of feature vectors sampled from some population, we would like to know if the data set consists of a number of relatively distinct subsets. If it does and we can determine these subsets, we can define them to be classes. This is sometimes called **class discovery**. The techniques from Chapters 3 and 4 can then be used to further analyze or model the data or to classify new data if desired. **Clustering** refers to the process of grouping samples so that the samples are similar within each group. The groups are called **clusters**.

In some applications, the main goal may be to discover the subgroups rather than to model them statistically. For example, the marketing director of a firm that supplies business services may want to know if the businesses in a particular community fall into any natural groupings of similar companies so that specific service packages and marketing plans can be designed for each of these subgroups. Reading the public data on these companies might give an idea of what some of these subgroups could be, but the process would be difficult and unreliable, particularly if the number of features or companies is large. Fortunately, clustering techniques allow the division into subgroups to be done automatically, without any preconceptions about what kinds of groupings should be found in the community being analyzed. Cluster analysis has been applied in many fields. For example, in 1971, Paykel used cluster analysis to group 165 depressed patients into four clusters which were then called "anxious," "hostile," "retarded psychotic," and "young depressive." In image analysis, clustering can be used to find groups of pixels with similar gray levels, colors, or local textures, in order to discover the various regions in the image.

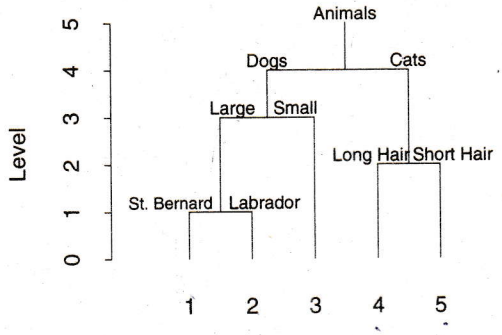

Figure 5.1: A hierarchical clustering.

In cases where there are only two features, clusters can be found through visual inspection by looking for dense regions in a scatterplot of the data if the subgroups or classes are well separated in the feature space. If, for example, there are two bivariate normally distributed classes and their means are separated by more than two standard deviations, two distinct peaks form if there is enough data. In Figure 4.20 at least one of the three classes forms a distinct cluster, which could be found even if the classes were unknown. However, distinct clusters may exist in a high-dimensional feature space and still not be apparent in any of the projections of the data onto a plane defined by a pair of the feature axes. One general way to find candidates for the centers of clusters is to form an n-dimensional histogram of the data and find the peaks in the histogram. However, if the number of features is large, the histogram may have to be very coarse to have a significant number of samples in any cell, and the locations of the boundaries between these cells are specified arbitrarily in advance, rather than depending on the data.

5.2 Hierarchical Clustering

A **hierarchy** can be represented by a tree structure such as the simple one shown in Figure 5.1. The patients in an animal hospital are composed of two main groups, dogs and cats, each of which is composed of subgroups. Each subgroup is, in turn, composed of subgroups, and so on. Each of the individual animals, 1 through 5, is represented at the lowest level of the tree. **Hierarchical clustering** refers to a clustering process that organizes the data into large groups, which contain smaller groups, and so on. A hierarchical clustering may be drawn as a **tree** or **dendrogram**. The finest grouping is at the bottom of the dendrogram; each sample by itself forms

a cluster. The coarsest grouping is at the top of the dendrogram, where all samples are grouped into one cluster. In between, there are various numbers of clusters. For example, in the hierarchical clustering of Figure 5.1, at level 0 the clusters are

$$\{1\},\{2\},\{3\},\{4\},\{5\},$$

each consisting of an individual sample. At level 1, the clusters are

$$\{1,2\},\{3\},\{4\},\{5\}.$$

At level 2, the clusters are

$$\{1,2\},\{3\},\{4,5\}.$$

At level 3, the clusters are

$$\{1,2,3\},\{4,5\}.$$

At level 4, the single cluster

$$\{1,2,3,4,5\}$$

consists of all the samples.

In a hierarchical clustering, if at some level two samples belong to a cluster, they belong to the same cluster at all higher levels. For example, in Figure 5.1, at level 2 samples 4 and 5 belong to the same cluster; samples 4 and 5 also belong to the same cluster at levels 3 and 4.

Hierarchical clustering algorithms are called **agglomerative** if they build the dendrogram from the bottom up and they are called **divisive** if they build the dendrogram from the top down.

The general agglomerative clustering algorithm is straightforward to describe. The total number of samples will be denoted by n.

Agglomerative Clustering Algorithm

1. Begin with n clusters, each consisting of one sample.
2. Repeat step 3 a total of $n-1$ times.
3. Find the most similar clusters C_i and C_j and merge C_i and C_j into one cluster. If there is a tie, merge the first pair found.

Different hierarchical clustering algorithms are obtained by using different methods to determine the similarity of clusters. One way to measure the similarity between clusters is to define a function that measures distance between clusters. This distance function typically is induced by an underlying function that measures the distance between pairs of samples. In cluster analysis as in nearest neighbor techniques (Section 4.4), the most popular distance measures are Euclidean distance and city block distance.

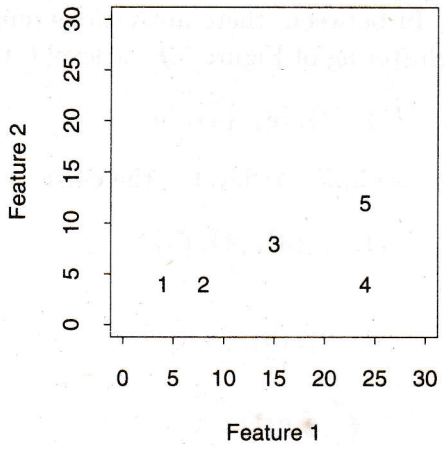

Figure 5.2: Samples for clustering.

The Single-Linkage Algorithm

The **single-linkage algorithm** is also known as the **minimum method** and the **nearest neighbor method**. The latter title underscores its close relation to the nearest neighbor classification method. The single-linkage algorithm is obtained by defining the distance between two clusters to be the smallest distance between two points such that one point is in each cluster. Formally, if C_i and C_j are clusters, the distance between them is defined as

$$D_{SL}(C_i, C_j) = \min_{a \in C_i, b \in C_j} d(a, b),$$

where $d(a, b)$ denotes the distance between the samples a and b.

Example 5.1 *Hierarchical clustering using the single-linkage algorithm.*

Perform a hierarchical clustering of five samples using the single-linkage algorithm and two features, x and y. A scatterplot of the data is shown in Figure 5.2. Use Euclidean distance for the distance between samples. The following tables give the feature values for each sample and the distance d between each pair of samples:

	x	y
1	4	4
2	8	4
3	15	8
4	24	4
5	24	12

	1	2	3	4	5
1	—	4.0	11.7	20.0	21.5
2	4.0	—	8.1	16.0	17.9
3	11.7	8.1	—	9.8	9.8
4	20.0	16.0	9.8	—	8.0
5	21.5	17.9	9.8	8.0	—

For the single-sample clusters $\{a\}$ and $\{b\}$, $D_{SL}(\{a\},\{b\}) = d(a,b)$.

The algorithm begins with five clusters, each consisting of one sample. The two nearest clusters are then merged. The smallest number in (5.1) is 4, which is the distance between samples 1 and 2, so the clusters $\{1\}$ and $\{2\}$ are merged. At this point there are four clusters

$$\{1,2\},\{3\},\{4\},\{5\}.$$

Next obtain the matrix that gives the distances between these clusters:

	{1,2}	3	4	5
{1,2}	—	8.1	16.0	17.9
3	8.1	—	9.8	9.8
4	16.0	9.8	—	8.0
5	17.9	9.8	8.0	—

The value 8.1 in row $\{1,2\}$ and column 3 gives the distance between the clusters $\{1,2\}$ and $\{3\}$ and is computed in the following way. Matrix (5.1) shows that $d(1,3) = 11.7$ and $d(2,3) = 8.1$. In the single-linkage algorithm, the distance between clusters is the minimum of these values, 8.1. The other values in the first row are computed in a similar way. The values in other than the first row or first column are simply copied from the previous table (5.1). Since the minimum value in this matrix is 8, the clusters $\{4\}$ and $\{5\}$ are merged. At this point there are three clusters:

$$\{1,2\},\{3\},\{4,5\}.$$

Next obtain the matrix that gives the distance between these clusters:

	{1,2}	3	{4,5}
{1,2}	—	8.1	16.0
3	8.1	—	9.8
{4,5}	16.0	9.8	—

Since the minimum value in this matrix is 8.1, the clusters $\{1,2\}$ and $\{3\}$ are merged. At this point there are two clusters:

$$\{1,2,3\},\{4,5\}.$$

The next step will merge the two remaining clusters at a distance of 9.8. The hierarchical clustering is complete. The dendrogram is shown in Figure 5.3.

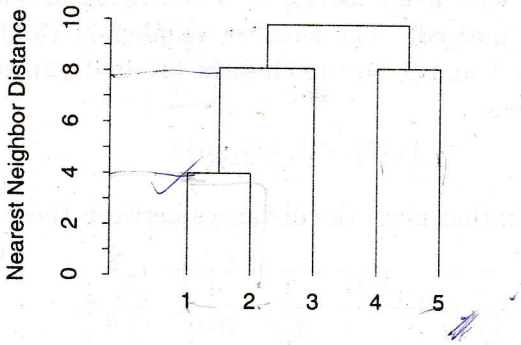

Figure 5.3: Hierarchical clustering using the single-linkage algorithm. The distance D_{SL} between clusters that merge is shown on the vertical axis.

The Complete-Linkage Algorithm

The **complete-linkage algorithm** is also called the **maximum method** or the **farthest neighbor method**. It is obtained by defining the distance between two clusters to be the largest distance between a sample in one cluster and a sample in the other cluster. If C_i and C_j are clusters, we define

$$D_{CL}(C_i, C_j) = \max_{a \in C_i, b \in C_j} d(a, b).$$

Example 5.2 *Hierarchical clustering using the complete-linkage algorithm.*

Perform a hierarchical clustering using the complete-linkage algorithm on the data shown in Figure 5.2. Use Euclidean distance (4.1) for the distance between samples.

As before, the algorithm begins with five clusters, each consisting of one sample. The nearest clusters {1} and {2} are then merged to produce the clusters

$$\{1,2\}, \{3\}, \{4\}, \{5\}.$$

Next obtain the matrix that gives the distances between these clusters:

	{1,2}	3	4	5
{1,2}	—	11.7	20.0	21.5
3	11.7	—	9.8	9.8
4	20.0	9.8	—	8.0
5	21.5	9.8	8.0	—

The value 11.7 in row $\{1,2\}$ and column 3 gives the distance between the clusters $\{1,2\}$ and $\{3\}$ and is computed in the following way. Matrix (5.1) shows that $d(1,3) = 11.7$ and $d(2,3) = 8.1$. In the complete-linkage algorithm, the distance between clusters is the maximum of these values, 11.7. The other values in the first row are computed in a similar way. The values in other than the first row or first column are simply copied from (5.1). Since the minimum value in this matrix is 8, the clusters $\{4\}$ and $\{5\}$ are merged. At this point the clusters are

$$\{1,2\}, \{3\}, \{4,5\}.$$

Next obtain the matrix that gives the distance between these clusters:

	{1,2}	3	{4,5}
{1,2}	—	11.7	21.5
3	11.7	—	9.8
{4,5}	21.5	9.8	—

Since the minimum value in this matrix is 9.8, the clusters $\{3\}$ and $\{4,5\}$ are merged. At this point the clusters are

$$\{1,2\}, \{3,4,5\}.$$

Notice that these clusters are different from those obtained at the corresponding point of the single-linkage algorithm.

At the next step, the two remaining clusters will be merged. The hierarchical clustering is complete. The dendrogram is shown in Figure 5.4.

A cluster, by definition, contains similar samples. The single-linkage algorithm and the complete-linkage algorithm differ in how they determine when samples in two clusters are similar so they can be merged. The single-linkage algorithm says that two clusters C_i and C_j are similar if there are any elements a in C_i and b in C_j that are similar, in the sense that the distance between a and b is small. In other words, in the single-linkage algorithm, it takes a *single* similar pair a, b with a in C_i and b in C_j in order to merge C_i and C_j. (Readers familiar with graph theory will recognize this procedure as that used by Kruskal's algorithm to find a minimum spanning tree.) On the other hand, the complete-linkage algorithm says that two clusters C_i and C_j are similar if the maximum of $D_{CL}(a,b)$ over *all* a in C_i and b in C_j is small. In other words, in the complete-linkage algorithm *all* pairs in C_i and C_j must be similar in order to merge C_i and C_j.

The Average-Linkage Algorithm

The single-linkage algorithm allows clusters to grow long and thin whereas the complete-linkage algorithm produces more compact clusters. Both clusterings are susceptible to

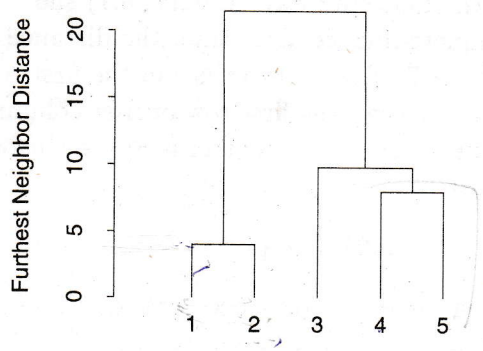

Figure 5.4: Hierarchical clustering using the complete-linkage algorithm.

distortion by outliers or deviant observations. The **average-linkage algorithm** is an attempt to compromise between the extremes of the single- and complete-linkage algorithms.

The average-linkage clustering algorithm, also known as the **unweighted pair-group method using arithmetic averages (UPGMA)**, is one of the most widely used hierarchical clustering algorithms. The average-linkage algorithm is obtained by defining the distance between two clusters to be the average distance between a point in one cluster and a point in the other cluster. Formally, if C_i is a cluster with n_i members and C_j is a cluster with n_j members, the distance between the clusters is

$$D_{AL}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{a \in C_i, b \in C_j} d(a, b).$$

Example 5.3 *Hierarchical clustering using the average-linkage algorithm.*

Perform a hierarchical clustering using the average-linkage algorithm on the data shown in Figure 5.2. Use Euclidean distance (4.1) for the distance between samples.

The algorithm begins with five clusters, each consisting of one sample. The nearest clusters $\{1\}$ and $\{2\}$ are then merged to form the clusters

$$\{1,2\}, \{3\}, \{4\}, \{5\}.$$

Next obtain the matrix that gives the distance between these clusters:

	{1,2}	3	4	5
{1,2}	—	9.9	18.0	19.7
3	9.9	—	9.8	9.8
4	18	9.8	—	8.0
5	19.7	9.8	8.0	—

The value 9.9 in row $\{1,2\}$ and column 3 gives the distance between the clusters $\{1,2\}$ and $\{3\}$ and is computed in the following way. Matrix (5.1) shows that $d(1,3) = 11.7$ and $d(2,3) = 8.1$. In the average-linkage algorithm, the distance between clusters is the average of these values, 9.9. The other values in the first row are computed in a similar way. The values in other than the first row or first column are simply copied from (5.1). Since the minimum value in this matrix is 8, the clusters $\{4\}$ and $\{5\}$ are merged. At this point the clusters are

$$\{1,2\}, \{3\}, \{4,5\}.$$

Next obtain the matrix that gives the distance between these clusters:

	{1,2}	3	{4,5}
{1,2}	—	9.9	18.9
3	9.9	—	9.8
{4,5}	18.9	9.8	—

Since the minimum value in this matrix is 9.8, the clusters $\{3\}$ and $\{4,5\}$ are merged. At this point the clusters are

$$\{1,2\}, \{3,4,5\}.$$

At the next step, the two remaining clusters are merged and the hierarchical clustering is complete.

An example of the application of the average-linkage algorithm to a larger data set using the SAS statistical analysis software package is presented in Appendix B.4.

Ward's Method

Ward's method is also called the **minimum-variance method**. Like the other algorithms, Ward's method begins with one cluster for each individual sample. At each iteration, among all pairs of clusters, it merges the pair that produces the smallest **squared error** for the resulting set of clusters. The squared error for each cluster is defined as follows. If a cluster contains m samples $\mathbf{x}_1, \ldots, \mathbf{x}_m$ where $\mathbf{x}_i$ is the feature