We saw earlier that the use of $d_{\text{max}}(\cdot, \cdot)$ causes the smallest possible stepwise increase in the diameter of the partition. Another simple example is provided by the sum-of-squared-error criterion function $J_e$. By an analysis very similar to that used in Section 10.8, we find that the pair of clusters whose merger increases $J_e$ as little as possible is the pair for which the "distance"

$$d_e(D_i, D_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} \| \mathbf{m}_i - \mathbf{m}_j \|$$

is minimum (Problem 36). Thus, in selecting clusters to be merged, this criterion takes into account the number of samples in each cluster as well as the distance between clusters. In general, the use of $d_e(\cdot, \cdot)$ tends to favor growth by merging singletons or small clusters with large clusters over merging medium-sized clusters. While the final partition may not minimize $J_e$, it usually provides a very good starting point for further iterative optimization.
10.7 CRITERION FUNCTIONS FOR CLUSTERING

We have just considered the first major issue in clustering: how to measure "similarity." Now we turn to the second major issue: the criterion function to be optimized. Suppose that we have a set \( D = \{ x_1, \ldots, x_n \} \) of \( n \) samples that we want to partition into exactly \( c \) disjoint subsets \( D_1, \ldots, D_c \). Each subset is to represent a cluster, with samples in the same cluster being somehow more similar to each other than they are to samples in other clusters. One way to make this into a well-defined problem is to define a criterion function that measures the clustering quality of any partition of the data. Then the problem is one of finding the partition that extremizes the criterion function. In this section we examine the characteristics of several basically similar criterion functions, postponing until later the question of how to find an optimal partition.

10.7.1 The Sum-of-Squared-Error Criterion

The simplest and most widely used criterion function for clustering is the sum-of-squared-error criterion. Let \( n_i \) be the number of samples in \( D_i \) and let \( m_i \) be the mean of those samples,

\[
m_i = \frac{1}{n_i} \sum_{x \in D_i} x.
\]  

Then the sum-of-squared errors is defined by

\[
J_e = \sum_{i=1}^{c} \sum_{x \in D_i} ||x - m_i||^2.
\]  

This criterion function has a simple interpretation: For a given cluster \( D_i \), the mean vector \( m_i \) is the best representative of the samples in \( D_i \) in the sense that it minimizes the sum of the squared lengths of the "error" vectors \( x - m_i \) in \( D_i \). Thus, \( J_e \) measures the total squared error incurred in representing the \( n \) samples \( x_1, \ldots, x_n \) by the \( c \) cluster centers \( m_1, \ldots, m_c \). The value of \( J_e \) depends on how the samples are grouped into clusters and the number of clusters; the optimal partitioning is defined as one that minimizes \( J_e \). Clusterings of this type are often called minimum variance partitions.

What kind of clustering problems are well-suited to a sum-of-squared-error criterion? Basically, \( J_e \) is an appropriate criterion when the clusters form compact clouds that are rather well-separated from one another. A less obvious problem arises there are great differences in the number of samples in different clusters. In that case
FIGURE 10.10. When two natural groupings have very different numbers of points, the clusters minimizing a sum-squared-error criterion $J_e$ of Eq. 54 may not reveal the true underlying structure. Here the criterion is smaller for the two clusters at the bottom than for the more natural clustering at the top.

It can happen that a partition that splits a large cluster is favored over one that maintains the integrity of the natural clusters, as illustrated in Fig. 10.10. This situation frequently arises because of the presence of "outliers" or "wild shots" and brings up the problem of interpreting and evaluating the results of clustering. Because little can be said about that problem, we shall merely observe that if additional considerations render the results of minimizing $J_e$ unsatisfactory, then these considerations should be used, if possible, in formulating a better criterion function.
\[ d(x, x') = \left( \sum_{k=1}^{d} |x_k - x'_k|^q \right)^{1/q}, \]  

(49)

where \( q \geq 1 \) is a selectable parameter—the general Minkowski metric we considered in Chapter 4. Setting \( q = 2 \) gives the familiar Euclidean metric while setting \( q = 1 \) gives the Manhattan or city block metric—the sum of the absolute distances along each of the \( d \) coordinate axes. Note that only \( q = 2 \) is invariant to an arbitrary rotation or translation in feature space. Another alternative is to use some kind of metric based on the data itself, such as the Mahalanobis distance.

More generally, one can abandon the use of distance altogether and introduce a nonmetric similarity function \( s(x, x') \) to compare two vectors \( x \) and \( x' \). Conventionally, this is a symmetric functions whose value is large when \( x \) and \( x' \) are somehow "similar." For example, when the angle between two vectors is a meaningful measure of their similarity, then the normalized inner product

\[ s(x, x') = \frac{x'x'}{\|x\| \|x'\|} \]  

(50)

may be an appropriate similarity function. This measure, which is the cosine of the angle between \( x \) and \( x' \), is invariant to rotation and dilation, though it is not invariant to translation and general linear transformations.

When the features are binary-valued (0 or 1), the similarity function of Eq. 50 has a simple nongeometrical interpretation in terms of shared features or shared attributes. Let us say that a sample \( x \) possesses the \( i \)th attribute if \( x_i = 1 \). Then \( x'x' \) is merely the number of attributes possessed by both \( x \) and \( x' \), and \( \|x\| \|x'\| = (x'xx'x')^{1/2} \) is the geometric mean of the number of attributes possessed by \( x \) and the number possessed by \( x' \). Thus, \( s(x, x') \) is a measure of the relative possession of common attributes. Some simple variations are

\[ s(x, x') = \frac{x'x'}{d}, \]  

(51)

the fraction of attributes shared, and

\[ s(x, x') = \frac{x'x'}{x'x + x''x' - x'x'}, \]  

(52)

the ratio of the number of shared attributes to the number possessed by \( x \) or \( x' \). This latter measure (sometimes known as the Tanimoto coefficient or Tanimoto distance) is frequently encountered in the fields of information retrieval and biological taxonomy. Related measures of similarity arise in other applications, with the variety of measures testifying to the diversity of problem domains.

Fundamental issues in measurement theory are involved in the use of any distance or similarity function. The calculation of the similarity between two vectors always involves combining the values of their components. Yet in many pattern recognition applications the components of the feature vector measure seemingly noncomparable quantities, such as meters and kilogrammes. Recall our example of classifying fish: How can we compare the lightness of the skin to the length or weight of the fish? Should the comparison depend on whether the length is measured in meters or inches? How do we treat vectors whose components have a mixture of nominal, ordinal, interval and ratio scales? Ultimately, there are rarely clear methodological answers to these
is minimum.

36. Assume we are clustering using the sum-of-squared error criterion $J_e$ (Eq. 54). Show that a “distance” measure between clusters can be derived, Eq. 83, such that merging the “closest” such clusters increases $J_e$ as little as possible.