CHAPTER

Clustering: Basic Concepts 11

11.1 INTRODUCTION

All the previous chapters were concerned with supervised classification. In the current and following chapters, we turn to the unsupervised case, where class labeling of the training patterns is not available. Thus, our major concern now is to "reveal" the organization of patterns into "*sensible*" clusters (groups), which will allow us to discover similarities and differences among patterns and to derive useful conclusions about them. This idea is met in many fields, such as the life sciences (biology, zoology), medical sciences (psychiatry, pathology), social sciences (sociology, archaeology), earth sciences (geography, geology), and engineering [Ande 73]. Clustering may be found under different names in different contexts, such as unsupervised learning and learning without a teacher (in pattern recognition), numerical taxonomy (in biology, ecology), typology (in social sciences), and partition (in graph theory). The following example is inspired by biology and gives us a flavor of the problem.

Consider the following animals: sheep, dog, cat (mammals), sparrow, seagull (birds), viper, lizard (reptiles), goldfish, red mullet, blue shark (fish), and frog (amphibians). In order to organize these animals into clusters, we need to define a *clustering criterion*. Thus, if we employ the way these animals bear their progeny as a clustering criterion, the sheep, the dog, the cat, and the blue shark will be assigned to the same cluster, while all the rest will form a second cluster (Figure 11.1a). If the clustering criterion is the existence of lungs, the goldfish, the red mullet, and the blue shark are assigned to the same cluster, while all the other animals are assigned to a second cluster (Figure 11.1b). On the other hand, if the clustering criterion is the environment where the animals live, the sheep, the dog, the cat, the sparrow, the seagull, the viper, and the lizard will form one cluster (animals living outside water); the goldfish, the red mullet, and the blue shark will form a second cluster (animals living only in water); and the frog will form a third cluster by itself, since it may live in the water or out of it (Figure 11.1c). It is worth pointing out that if the existence of a vertebral column is the clustering criterion, all the animals will lie in the same cluster. Finally, we may use composite clustering criteria as

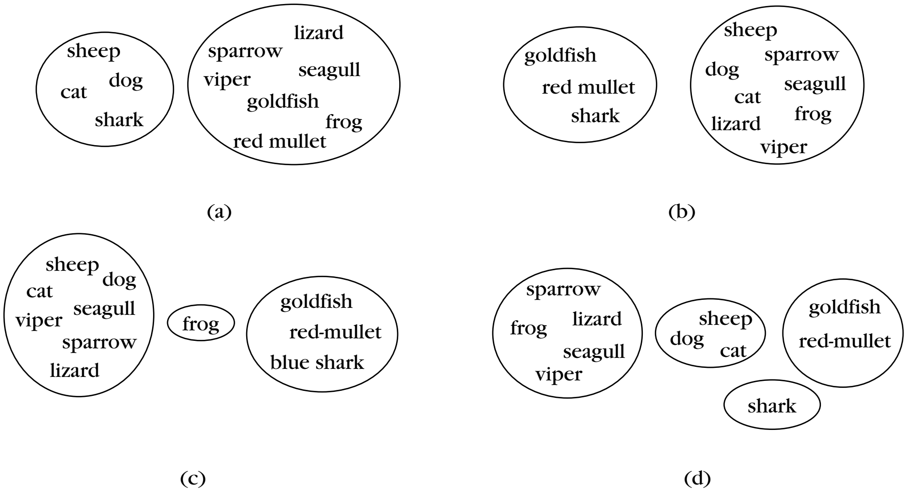

FIGURE 11.1

Resulting clusters if the clustering criterion is (a) the way the animals bear their progeny, (b) the existence of lungs, (c) the environment where the animals live, and (d) the way these animals bear their progeny and the existence of lungs.

well. For example, if the clustering criterion is the way these animals bear their progeny *and* the existence of lungs, we end up with four clusters as shown in Figure 11.1d.

This example shows that the process of assigning objects to clusters may lead to very different results, depending on the specific criterion used for clustering.

Clustering is one of the most primitive mental activities of humans, used to handle the huge amount of information they receive every day. Processing every piece of information as a single entity would be impossible. Thus, humans tend to categorize entities (i.e., objects, persons, events) into clusters. Each cluster is then characterized by the common attributes of the entities it contains. For example, most humans "possess" a cluster "dog." If someone sees a dog sleeping on the grass, he or she will identify it as an entity of the cluster "dog." Thus, the individual will infer that this entity barks even though he or she has never heard this specific entity bark before.

As was the case with supervised learning, we will assume that all patterns are represented in terms of *features*, which form l-dimensional feature vectors.

The basic steps that an expert must follow in order to develop a clustering task are the following:

- *Feature selection*. Features must be properly selected so as to encode as much information as possible concerning the task of interest. Once more, parsimony and, thus, minimum information redundancy among the features is a major goal. As in supervised classification, *preprocessing* of features may be necessary prior to their utilization in subsequent stages. The techniques discussed there are applicable here, too.

- *Proximity measure*. This measure quantifies how "similar" or "dissimilar" two feature vectors are. It is natural to ensure that all selected features contribute equally to the computation of the proximity measure and there are no features that dominate others. This must be taken care of during preprocessing.
- *Clustering criterion*. This criterion depends on the interpretation the expert gives to the term *sensible*, based on the type of clusters that are expected to underlie the data set. For example, a compact cluster of feature vectors in the l-dimensional space, may be sensible according to one criterion, whereas an elongated cluster may be sensible according to another. The clustering criterion may be expressed via a cost function or some other types of rules.
- *Clustering algorithms*. Having adopted a proximity measure and a clustering criterion, this step refers to the choice of a specific algorithmic scheme that unravels the clustering structure of the data set.
- *Validation of the results*. Once the results of the clustering algorithm have been obtained, we have to verify their correctness. This is usually carried out using appropriate tests.
- *Interpretation of the results*. In many cases, the expert in the application field must integrate the results of clustering with other experimental evidence and analysis in order to draw the right conclusions.

In a number of cases, a step known as *clustering tendency* should be involved. This includes various tests that indicate whether or not the available data possess a clustering structure. For example, the data set may be of a completely random nature, thus trying to unravel clusters would be meaningless.

As one may have already suspected, different choices of features, proximity measures, clustering criteria, and clustering algorithms may lead to totally different clustering results. *Subjectivity is a reality we have to live with from now on.* To demonstrate this, let us consider the following example. Consider Figure 11.2. How many "sensible" ways of clustering can we obtain for these points? The most "logical" answer seems to be two. The first clustering contains four clusters (surrounded by solid circles). The second clustering contains two clusters (surrounded by dashed lines). Which clustering is "correct"? It seems that there is no definite answer. Both clusterings are valid. The best thing to do is give the results to an expert and let the expert decide about the most sensible one. Thus, the final answer to these questions will be influenced by the expert's knowledge.

The rest of the chapter presents some basic concepts and definitions related to clustering, and it discusses proximity measures that are commonly encountered in various applications.

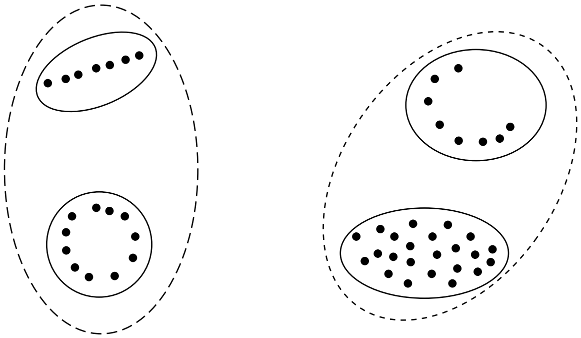

FIGURE 11.2

A coarse clustering of the data results in two clusters, whereas a finer one results in four clusters.

11.1.1 Applications of Cluster Analysis

Clustering is a major tool used in a number of applications. To enrich the list of examples already presented in the introductory chapter of the book, we summarize here four basic directions in which clustering is of use [Ball 71, Ever 01]:

- *Data reduction*. In several cases, the amount of the available data, N, is often very large and as a consequence, its processing becomes very demanding. Cluster analysis can be used in order to group the data into a number of "sensible" clusters, m ($\ll N$), and to process each cluster as a single entity. For example, in data transmission, a representative for each cluster is defined. Then, instead of transmitting the data samples, we transmit a code number corresponding to the representative of the cluster in which each specific sample lies. Thus, data compression is achieved.
- *Hypothesis generation*. In this case we apply cluster analysis to a data set in order to infer some hypotheses concerning the nature of the data. Thus, clustering is used here as a vehicle to suggest hypotheses. These hypotheses must then be verified using other data sets.
- *Hypothesis testing*. In this context, cluster analysis is used for the verification of the validity of a specific hypothesis. Consider, for example, the following hypothesis: "Big companies invest abroad." One way to verify whether this is true is to apply cluster analysis to a large and representative set of companies. Suppose that each company is represented by its size, its activities abroad, and its ability to complete successfully projects on applied research. If, after applying cluster analysis, a cluster is formed that corresponds to companies that are large and have investments abroad (regardless of their ability to complete successfully projects on applied research), then the hypothesis is supported by the cluster analysis.

- *Prediction based on groups*. In this case, we apply cluster analysis to the available data set, and the resulting clusters are characterized based on the characteristics of the patterns by which they are formed. In the sequel, if we are given an unknown pattern, we can determine the cluster to which it is more likely to belong, and we characterize it based on the characterization of the respective cluster. Suppose, for example, that cluster analysis is applied to a data set concerning patients infected by the same disease. This results in a number of clusters of patients, according to their reaction to specific drugs. Then for a new patient, we identify the most appropriate cluster for the patient and, based on it, we decide on his or her medication (e.g., see [Payk 72]).

11.1.2 Types of Features

A feature may take values from a continuous range (subset of $\mathcal{R}$) or from a finite discrete set. If the finite discrete set has only two elements, then the feature is called *binary* or *dichotomous*.

A different categorization of the features is based on the relative significance of the values they take [Jain 88, Spat 80]. We have four categories of features: *nominal*, *ordinal*, *interval-scaled*, and *ratio-scaled*.

The first category, nominal, includes features whose possible values code states. Consider for example a feature that corresponds to the sex of an individual. Its possible values may be 1 for a male and 0 for a female. Clearly, any quantitative comparison between these values is meaningless. The next category, ordinal, includes features whose values can be *meaningfully ordered*. Consider, for example, a feature that characterizes the performance of a student in the pattern recognition course. Suppose that its possible values are 4, 3, 2, 1 and that these correspond to the ratings "excellent," "very good," "good," "not good." Obviously, these values are arranged in a meaningful order. However, the difference between two successive values is of no meaningful quantitative importance.

If, for a specific feature, the difference between two values is meaningful while their ratio is meaningless, then it is an interval-scaled feature. A typical example is the measure of temperature in degrees Celsius. If the temperatures in London and Paris are 5 and 10 degrees Celsius, respectively, then it is meaningful to say that the temperature in Paris is 5 degrees higher than that in London. However, it is meaningless to say that Paris is twice as hot as London.

Finally, if the ratio between two values of a specific feature is meaningful, then this is a ratio-scaled feature, the fourth category. An example of such a feature is weight, since it is meaningful to say that a person who weighs 100 kg is twice as fat as a person whose weight is 50 kg.

By ordering the types of features as nominal, ordinal, interval-scaled, and ratio scaled, one can easily notice that each type of feature possesses all the properties of the types that are before it. For example, an interval-scaled feature has all the properties of the ordinal and nominal types. This information will be of use in Section 11.2.2.

Example 11.1

Suppose that we want to group companies according to their prospects of progress. To this end, we may take into account whether a company is private or public, whether or not the company has activities abroad, its annual budgets for the last, say, three years, its investments, and its rates of change of the budgets and investments. Therefore, each company is represented by a 10×1 vector. The first component of the vector corresponds to a nominal feature, which codes the state "public" or "private." The second component indicates whether or not there are activities abroad. Its possible values are 0, 1, and 2 (discrete range of values), which correspond to "no investments," "poor investments," and "large investments." Clearly, this component corresponds to an ordinal feature. All the remaining features are ratio-scaled.

11.1.3 Definitions of Clustering

The definition of clustering leads directly to the definition of a single "cluster." Many definitions have been proposed over the years (e.g., [John 67, Wall 68, Ever 01]). However, most of these definitions are based on loosely defined terms, such as *similar*, and *alike*, etc., or they are oriented to a specific kind of cluster. As pointed out in [Ever 01], most of these definitions are of vague and of circular nature. This fact reveals the difficulty of having a universally acceptable definition for the term cluster.

In [Ever 01], the vectors are viewed as points in the l-dimensional space, and the clusters are described as "continuous regions of this space containing a relatively high density of points, separated from other high density regions by regions of relatively low density of points." Clusters described in this way are sometimes referred to as *natural clusters*. This definition is closer to our visual perception of clusters in the two- and three-dimensional spaces.

Let us now try to give some definitions for "clustering," which, although they may not be universal, give us an idea of what clustering is. Let X be our data set, that is,

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}. \tag{11.1}$$

We define as an *m-clustering* of X, $\Re$, the partition of X into m sets (*clusters*), $C_1, \ldots, C_m$, so that the following three conditions are met:

- $C_i \neq \emptyset, i = 1, \ldots, m$
- $\cup_{i=1}^{m} C_i = X$
- $C_i \cap C_j = \emptyset, i \neq j, i, j = 1, \ldots, m$

In addition, the vectors contained in a cluster C_i are "more similar" to each other and "less similar" to the feature vectors of the other clusters. Quantifying the terms *similar* and *dissimilar* depends very much on the types of clusters involved. For example, other *measures* (measuring similarity) are required for compact clusters (e.g., Figure 11.3a), others for elongated clusters (e.g., Figure 11.3b), and different ones for shell-shaped clusters (e.g., Figure 11.3c).

Note that, under the preceding definitions of clustering, each vector belongs to a single cluster. For reasons that will become clear later on, this type of clustering is sometimes called *hard* or *crisp*. An alternative definition is in terms of the *fuzzy sets*, introduced by Zadeh [Zade 65]. A fuzzy clustering of X into m clusters is characterized by m functions u_j where

$$u_j : X \rightarrow [0, 1], \quad j = 1, \ldots, m \tag{11.2}$$

and

$$\sum_{j=1}^{m} u_j(\boldsymbol{x}_i) = 1, \quad i = 1, 2, \ldots, N, \qquad 0 < \sum_{i=1}^{N} u_j(\boldsymbol{x}_i) < N, \quad j = 1, 2, \ldots, m \tag{11.3}$$

These are called *membership functions*. The value of a fuzzy membership function is a mathematical characterization of a set, that is, a cluster in our case, which may not be precisely defined. That is, each vector $\boldsymbol{x}$ belongs to more than one cluster simultaneously "up to some degree," which is quantified by the corresponding value of u_j in the interval [0,1]. Values close to unity show a high "grade of membership" in the corresponding cluster and values close to zero, a low grade of membership. The values of these membership functions are indicative of the structure of the data set, in the sense that if a membership function has close to unity values for two vectors of X, that is, $\boldsymbol{x}_k$, $\boldsymbol{x}_n$, they are considered similar to each other [Wind 82].

The right condition in (11.3) guarantees that there are not trivial cases where clusters exist that do not share any vectors. This is analogous to the condition $C_i \neq \emptyset$ of the aforementioned definition.

The definition of clustering into m distinct sets C_i, given before, can be recovered as a special case of the fuzzy clustering if we define the fuzzy membership functions u_j to take values in $\{0, 1\}$, that is, to be either 1 or 0. In this sense, each data vector belongs exclusively to one cluster and the membership functions are now called *characteristic functions* ([Klir 95]).

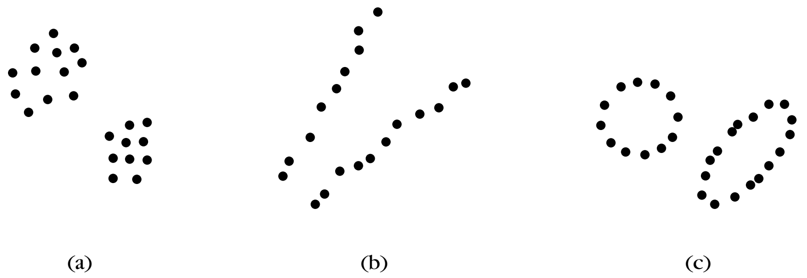

FIGURE 11.3

(a) Compact clusters. (b) Elongated clusters. (c) Spherical and ellipsoidal clusters.

11.2 PROXIMITY MEASURES

11.2.1 Definitions

We begin with definitions concerning measures between vectors, and we will extend them later on to include measures between subsets of the data set X.

A *dissimilarity measure* (DM) d on X is a function.

$$d : X \times X \rightarrow \mathcal{R}$$

where $\mathcal{R}$ is the set of real numbers, such that

$$\exists d_0 \in \mathcal{R} : -\infty < d_0 \leq d(\boldsymbol{x}, \boldsymbol{y}) < +\infty, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X \tag{11.4}$$

$$d(\boldsymbol{x}, \boldsymbol{x}) = d_0, \quad \forall \boldsymbol{x} \in X \tag{11.5}$$

and

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X \tag{11.6}$$

If in addition

$$d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \quad \text{if and only if} \quad \boldsymbol{x} = \boldsymbol{y} \tag{11.7}$$

and

$$d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}), \quad \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X \tag{11.8}$$

d is called a *metric DM*. Inequality (11.8) is also known as the *triangular inequality*. Finally, equivalence (11.7) indicates that the minimum possible dissimilarity level value d_0 between any two vectors in X is achieved when they are identical. Sometimes we will refer to the dissimilarity level as distance, where the term is not used in its strict mathematical sense.

A *similarity measure* (SM) s on X is defined as

$$s : X \times X \rightarrow \mathcal{R}$$

such that

$$\exists s_0 \in \mathcal{R} : -\infty < s(\boldsymbol{x}, \boldsymbol{y}) \leq s_0 < +\infty, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X \tag{11.9}$$

$$s(\boldsymbol{x}, \boldsymbol{x}) = s_0, \quad \forall \boldsymbol{x} \in X \tag{11.10}$$

and

$$s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X \tag{11.11}$$

If in addition

$$s(\boldsymbol{x}, \boldsymbol{y}) = s_0 \quad \text{if and only if} \quad \boldsymbol{x} = \boldsymbol{y} \tag{11.12}$$

and

$$s(\boldsymbol{x}, \boldsymbol{y}) s(\boldsymbol{y}, \boldsymbol{z}) \leq [s(\boldsymbol{x}, \boldsymbol{y}) + s(\boldsymbol{y}, \boldsymbol{z})] s(\boldsymbol{x}, \boldsymbol{z}), \quad \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X \tag{11.13}$$

s is called a *metric SM*.

Example 11.2

Let us consider the well-known Euclidean distance, d_2

$$d_2(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{l}(x_i - y_i)^2}$$

where $\boldsymbol{x},\boldsymbol{y} \in X$ and x_i, y_i are the ith coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. This is a dissimilarity measure on X, with $d_0 = 0$; that is, the minimum possible distance between two vectors of X is 0. Moreover, the distance of a vector from itself is equal to 0. Also, it is easy to observe that $d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{y},\boldsymbol{x})$.

The preceding arguments show that the *Euclidean distance is a dissimilarity measure*. In addition, the Euclidean distance between two vectors takes its minimum value $d_0 = 0$, when the vectors coincide. Finally, it is not difficult to show that the triangular inequality holds for the Euclidean distance (see Problem 11.2). Therefore, the Euclidean distance is a metric dissimilarity measure.

For other measures, the values d_0 (s_0) may be positive or negative.

Not all clustering algorithms, however, are based on proximity measures between vectors. For example, in the hierarchical clustering algorithms[1] one has to compute distances between pairs of sets of vectors of X. In the sequel, we extend the preceding definitions in order to measure "proximity" between subsets of X. Let U be a set containing subsets of X. That is, $D_i \subset X, i = 1, \ldots, k$, and $U = \{D_1, \ldots, D_k\}$. A ***proximity measure*** $\wp$ ***on*** U is a function

$$\wp : U \times U \rightarrow \mathcal{R}$$

Equations (11.4)–(11.8) for dissimilarity measures and Eqs. (11.9)–(11.13) for similarity measures can now be repeated with D_i, D_j in the place of $\boldsymbol{x}$ and $\boldsymbol{y}$ and U in the place of X.

Usually, the proximity measures between two sets D_i and D_j are defined in terms of proximity measures between elements of D_i and D_j.

Example 11.3

Let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6\}$ and $U = \{\{\boldsymbol{x}_1, \boldsymbol{x}_2\}, \{\boldsymbol{x}_1, \boldsymbol{x}_4\}, \{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}, \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}\}$. Let us define the following dissimilarity function:

$$d_{\min}^{ss}(D_i,\ D_j) = \min_{\boldsymbol{x}\in D_i,\ \boldsymbol{y}\in D_j} d_2(\boldsymbol{x},\ \boldsymbol{y})$$

where d_2 is the Euclidean distance between two vectors and $D_i,\ D_j \in U$.

The minimum possible value of $d_{\min}^{ss}$ is $d_{\min,0}^{ss} = 0$. Also, $d_{\min}^{ss}(D_i,\ D_i) = 0$, since the Euclidean distance between a vector in D_i and itself is 0. In addition, it is easy to see that the

[1] These algorithms are treated in detail in Chapter 13.

commutative property holds. Thus, this dissimilarity function is a measure. It is not difficult to see that $d_{\min}^{ss}$ is not a metric. Indeed, Eq. (11.7) for subsets of X does not hold in general, since the two sets D_i and D_j may have an element in common. Consider, for example the two sets $\{\boldsymbol{x}_1, \ \boldsymbol{x}_2\}$ and $\{\boldsymbol{x}_1, \ \boldsymbol{x}_4\}$ of U. Although they are different, their distance $d_{\min}^{ss}$ is 0, since they both contain $\boldsymbol{x}_1$.

Intuitively speaking, the preceding definitions show that the DMs are "opposite" to SMs. For example, it is easy to show that if d is a (metric) DM, with $d(\boldsymbol{x},\boldsymbol{y}) > 0$, $\forall \boldsymbol{x},\boldsymbol{y} \in X$, then $s = a/d$ with $a > 0$ is a (metric) SM (see Problem 11.1). Also, $d_{\max} - d$ is a (metric) SM, where $d_{\max}$ denotes the maximum value of d among all pairs of elements of X. It is also easy to show that if d is a (metric) DM on a finite set X, such that $d(\boldsymbol{x},\boldsymbol{y}) > 0, \forall \boldsymbol{x},\boldsymbol{y} \in X$, then so are $-\ln(d_{\max} + k - d)$ and $kd/(1 + d)$, where k is an arbitrary positive constant. On the other hand, if s is a (metric) SM with $s_0 = 1 - \varepsilon$, where ε is a small positive constant, then $1/(1 - s)$ is also a (metric) SM. Similar comments are valid for the similarity and dissimilarity measures between sets $D_i, D_j \in U$.

In the sequel, we will review the most commonly used proximity measures between two points. For each measure of similarity we give a corresponding measure of dissimilarity. We will denote by $b_{\min}$ and $b_{\max}$ the corresponding minimum and maximum values that they take for a finite data set X.

11.2.2 Proximity Measures between Two Points

Real-Valued Vectors

A. Dissimilarity Measures

The most common DMs between real-valued vectors used in practice are:

- The *weighted* l_p metric DMs, that is,

$$d_p(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{i=1}^{l} w_i |x_i - y_i|^p\right)^{1/p} \tag{11.14}$$

where x_i, y_i are the ith coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$, $i = 1, \ldots, l$, and $w_i \geq 0$ is the ith *weight coefficient*. They are used mainly on real-valued vectors. If $w_i = 1, i = 1, \ldots, l$, we obtain the *unweighted* l_p metric DMs. A well-known representative of the latter category of measures is the *Euclidean distance*, which was introduced in Example 11.2 and is obtained by setting $p = 2$.

The weighted l_2 metric DM can be further generalized as follows:

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T B(\boldsymbol{x} - \boldsymbol{y})} \tag{11.15}$$

where B is a symmetric, positive definite matrix (Appendix B).

This includes the Mahalanobis distance as a special case, and it is also a metric DM.

Special l_p metric DMs that are also encountered in practice are the (weighted) l_1 or *Manhattan norm*,

$$d_1(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{l} w_i|x_i - y_i| \tag{11.16}$$

and the (weighted) l_∞ *norm*,

$$d_\infty(\boldsymbol{x},\boldsymbol{y}) = \max_{1\leq i\leq l} w_i|x_i - y_i| \tag{11.17}$$

The l_1 and l_∞ norms may be viewed as overestimation and underestimation of the l_2 norm, respectively. Indeed, it can be shown that $d_\infty(\boldsymbol{x},\boldsymbol{y}) \leq d_2(\boldsymbol{x},\boldsymbol{y}) \leq d_1(\boldsymbol{x},\boldsymbol{y})$ (see Problem 11.6). When $l = 1$ all l_p norms coincide.

Based on these DMs, we can define corresponding SMs as $s_p(\boldsymbol{x},\boldsymbol{y}) = b_{\max} - d_p(\boldsymbol{x},\boldsymbol{y})$.

- Some additional DMs are the following [Spat 80]:

$$d_G(\boldsymbol{x},\boldsymbol{y}) = -\log_{10}\left(1 - \frac{1}{l}\sum_{j=1}^{l}\frac{|x_j - y_j|}{b_j - a_j}\right) \tag{11.18}$$

where b_j and a_j are the maximum and the minimum values among the jth features of the N vectors of X, respectively. It can easily be shown that $d_G(\boldsymbol{x},\boldsymbol{y})$ is a metric DM. Notice that the value of $d_G(\boldsymbol{x},\boldsymbol{y})$ depends not only on $\boldsymbol{x}$ and $\boldsymbol{y}$ but also on the whole of X. Thus, if $d_G(\boldsymbol{x},\boldsymbol{y})$ is the distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ that belong to a set X and $d'_G(\boldsymbol{x},\boldsymbol{y})$ is the distance between the same two vectors when they belong to a different set X', then, in general, $d_G(\boldsymbol{x},\boldsymbol{y}) \neq d'_G(\boldsymbol{x},\boldsymbol{y})$. Another DM is [Spat 80]

$$d_Q(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\frac{1}{l}\sum_{j=1}^{l}\left(\frac{x_j - y_j}{x_j + y_j}\right)^2} \tag{11.19}$$

Example 11.4

Consider the three-dimensional vectors $\boldsymbol{x} = [0,\ 1,\ 2]^T$, $\boldsymbol{y} = [4,\ 3,\ 2]^T$. Then, assuming that all w_i's are equal to 1, $d_1(\boldsymbol{x},\boldsymbol{y}) = 6, d_2(\boldsymbol{x},\boldsymbol{y}) = 2\sqrt{5}$, and $d_\infty(\boldsymbol{x},\boldsymbol{y}) = 4$. Notice that $d_\infty(\boldsymbol{x},\boldsymbol{y}) < d_2(\boldsymbol{x},\boldsymbol{y}) < d_1(\boldsymbol{x},\boldsymbol{y})$.

Assume now that these vectors belong to a data set X that contains N vectors with maximum values per feature 10, 12, 13 and minimum values per feature 0, 0.5, 1, respectively. Then $d_G(\boldsymbol{x},\boldsymbol{y}) = 0.0922$. If, on the other hand, $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to an X' with the maximum (minimum) values per feature being 20, 22, 23 $(-10,\ -9.5,\ -9)$, respectively, then $d_G(\boldsymbol{x},\boldsymbol{y}) = 0.0295$.

Finally, $d_Q(\boldsymbol{x},\boldsymbol{y}) = 0.6455$.

B. Similarity Measures

The most common similarity measures for real-valued vectors used in practice are:

- *The inner product.* It is defined as $s_{\text{inner}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{l} x_i y_i$. In most cases, the inner product is used when the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are normalized, so that they have the same length a. In these cases, the upper and the lower bounds of s_{inner} are $+a^2$ and $-a^2$, respectively, and $s_{\text{inner}}(\boldsymbol{x}, \boldsymbol{y})$ depends exclusively on the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$.

  A corresponding dissimilarity measure for the inner product is $d_{\text{inner}}(\boldsymbol{x}, \boldsymbol{y}) = b_{\max} - s_{\text{inner}}(\boldsymbol{x}, \boldsymbol{y})$.

  Closely related to the inner product is the *cosine similarity measure*, which is defined as

$$s_{\text{cosine}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \tag{11.20}$$

  where $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{l} x_i^2}$ and $\|\boldsymbol{y}\| = \sqrt{\sum_{i=1}^{l} y_i^2}$ are the lengths of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. This measure is invariant to rotations but not to linear transformations.

- *Pearson's correlation coefficient.* This measure can be expressed as

$$r_{\text{Pearson}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}_d{}^T\boldsymbol{y}_d}{\|\boldsymbol{x}_d\|\|\boldsymbol{y}_d\|} \tag{11.21}$$

  where $\boldsymbol{x}_d = [x_1 - \bar{x}, \ldots, x_l - \bar{x}]^T$ and $\boldsymbol{y}_d = [y_1 - \bar{y}, \ldots, y_l - \bar{y}]^T$, with x_i, y_i being the ith coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and $\bar{x} = \frac{1}{l}\sum_{i=1}^{l} x_i$, $\bar{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$. Usually, $\boldsymbol{x}_d$ and $\boldsymbol{y}_d$ are called difference vectors. Clearly, $r_{\text{Pearson}}(\boldsymbol{x}, \boldsymbol{y})$ takes values between -1 and $+1$. The difference from s_{inner} is that s_{Pearson} does not depend directly on $\boldsymbol{x}$ and $\boldsymbol{y}$ but on their corresponding difference vectors. A related dissimilarity measure can be defined as

$$D(\boldsymbol{x}, \boldsymbol{y}) = \frac{1 - r_{\text{Pearson}}(\boldsymbol{x}, \boldsymbol{y})}{2} \tag{11.22}$$

  This takes values in the range $[0, 1]$. This measure has been used in the analysis of gene-expression data ([Eise 98]).

- Another commonly used SM is the *Tanimoto measure*, which is also known as Tanimoto distance [Tani 58]. It may be used for real- as well as for discrete-valued vectors. It is defined as

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \boldsymbol{x}^T\boldsymbol{y}} \tag{11.23}$$

  By adding and subtracting the term $\boldsymbol{x}^T\boldsymbol{y}$ in the denominator of (11.23) and after some algebraic manipulations, we obtain

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{1 + \frac{(\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y})}{\boldsymbol{x}^T\boldsymbol{y}}}$$

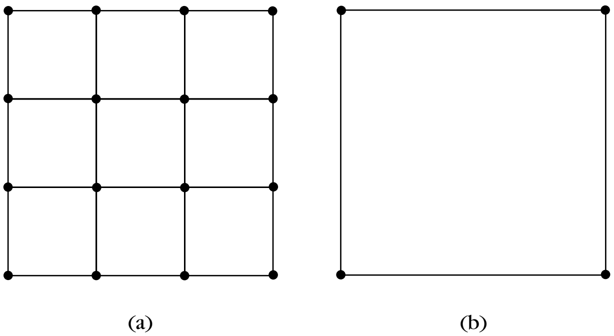

FIGURE 11.4

(a) The $l = 2$ dimensional grid for $k = 4$. (b) The H_2 hypercube (square).

That is, the Tanimoto measure between $\boldsymbol{x}$ and $\boldsymbol{y}$ is inversely proportional to the squared Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ divided by their inner product. Intuitively speaking, since the inner product may be considered as a measure of the correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$, $s_T(\boldsymbol{x}, \boldsymbol{y})$ is inversely proportional to the squared Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$, divided by their correlation.

In the case in which the vectors of X have been normalized to the same length a, the last equation leads to

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{-1 + 2\frac{a^2}{\boldsymbol{x}^T\boldsymbol{y}}}$$

In this case, s_T is inversely proportional to $a^2/\boldsymbol{x}^T\boldsymbol{y}$. Thus, the more correlated $\boldsymbol{x}$ and $\boldsymbol{y}$ are, the larger the value of s_T.

- Finally, another similarity measure that has been proved useful in certain applications [Fu 93] is the following:

$$s_c(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{d_2(\boldsymbol{x}, \boldsymbol{y})}{\|\boldsymbol{x}\| + \|\boldsymbol{y}\|} \tag{11.24}$$

$s_c(\boldsymbol{x}, \boldsymbol{y})$ takes its maximum value (1) when $\boldsymbol{x} = \boldsymbol{y}$ and its minimum (0) when $\boldsymbol{x} = -\boldsymbol{y}$.

Discrete-Valued Vectors

We will now consider vectors $\boldsymbol{x}$ whose coordinates belong to the finite set $F = \{0, 1, \ldots, k - 1\}$, where k is a positive integer. It is clear that there are exactly k^l vectors $\boldsymbol{x} \in F^l$. One can imagine these vectors as vertices in an l-dimensional grid as depicted in Figure 11.4. When $k = 2$, the grid collapses to the H_l (unit) hypercube.

Consider $\boldsymbol{x}, \boldsymbol{y} \in F^l$ and let

$$A(\boldsymbol{x}, \boldsymbol{y}) = [a_{ij}] \qquad i, j = 0, 1, \ldots, k-1 \tag{11.25}$$

be a $k \times k$ matrix, where the element a_{ij} is the number of places where the first vector has the i symbol and the corresponding element of the second vector has the j symbol, $i, j \in F$. This matrix is also known as a *contingency table*. For example, if $l = 6, k = 3$ and $\boldsymbol{x} = [0, 1, 2, 1, 2, 1]^T, \boldsymbol{y} = [1, 0, 2, 1, 0, 1]^T$, then matrix $A(\boldsymbol{x}, \boldsymbol{y})$ is equal to

$$A(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

It is easy to verify that

$$\sum_{i=0}^{k-1} \sum_{j=0}^{k-1} a_{ij} = l$$

Most of the proximity measures between two discrete-valued vectors may be expressed as combinations of elements of matrix $A(\boldsymbol{x}, \boldsymbol{y})$.

A. Dissimilarity Measures

- *The Hamming distance* (e.g., [Lipp 87, Gers 92]). It is defined as the number of places where two vectors differ. Using the matrix A, we can define the Hamming distance $d_H(\boldsymbol{x}, \boldsymbol{y})$ as

$$d_H(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=0}^{k-1} \sum_{j=0, j \neq i}^{k-1} a_{ij} \tag{11.26}$$

that is, the summation of all the off-diagonal elements of A, which indicate the positions where $\boldsymbol{x}$ and $\boldsymbol{y}$ differ.

In the special case in which $k = 2$, the vectors $\boldsymbol{x} \in F^l$ are binary valued and the Hamming distance becomes

$$d_H(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{l} (x_i + y_i - 2x_i y_i) = \sum_{i=1}^{l} (x_i - y_i)^2 \tag{11.27}$$

In the case where $\boldsymbol{x} \in F_1^l$, where $F_1 = \{-1, 1\}$, $\boldsymbol{x}$ is called bipolar vector and the Hamming distance is given as

$$d_H(\boldsymbol{x}, \boldsymbol{y}) = 0.5 \left(l - \sum_{i=1}^{l} x_i y_i \right) \tag{11.28}$$

Obviously, a corresponding similarity measure of d_H is $s_H(\boldsymbol{x}, \boldsymbol{y}) = b_{\max} - d_H(\boldsymbol{x}, \boldsymbol{y})$.

- *The l_1 distance.* It is defined as in the case of the continuous-valued vectors, that is,

$$d_1(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{l} |x_i - y_i| \tag{11.29}$$

The l_1 distance and the Hamming distance coincide when binary-valued vectors are considered.

B. Similarity Measures

A widely used similarity measure for discrete-valued vectors is *the Tanimoto measure*. *It is inspired by the comparison of sets*. If X and Y are two sets and n_X, n_Y, $n_{X\cap Y}$ are the cardinalities (number of elements) of X, Y, and $X \cap Y$, respectively, the Tanimoto measure between two sets X and Y is defined as

$$\frac{n_{X\cap Y}}{n_X + n_Y - n_{X\cap Y}} = \frac{n_{X\cap Y}}{n_{X\cup Y}}$$

In other words, the Tanimoto measure between two sets is the ratio of the number of elements they have in common to the number of all different elements.

We turn now to the Tanimoto measure between two discrete-valued vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. The measure takes into account all pairs of corresponding coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$, except those whose corresponding coordinates (x_i, y_i) are both 0. This is justified if we have ordinal features and interpret the value of the ith coordinate of, say, $\boldsymbol{y}$ as the degree to which the vector $\boldsymbol{y}$ possesses the ith feature. According to this interpretation, the pairs $(x_i, y_i) = (0, 0)$ are less important than the others. We now define $n_{\boldsymbol{x}} = \sum_{i=1}^{k-1} \sum_{j=0}^{k-1} a_{ij}$ and $n_{\boldsymbol{y}} = \sum_{i=0}^{k-1} \sum_{j=1}^{k-1} a_{ij}$, where a_{ij} are elements of the $A(\boldsymbol{x},\boldsymbol{y})$ matrix (see Figure 11.5). In words, $n_{\boldsymbol{x}}$ $(n_{\boldsymbol{y}})$ denotes the number of the nonzero coordinates of $\boldsymbol{x}$ $(\boldsymbol{y})$. Then, the Tanimoto measure is defined as

$$s_T(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_{\boldsymbol{x}} + n_{\boldsymbol{y}} - \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} a_{ij}} \tag{11.30}$$

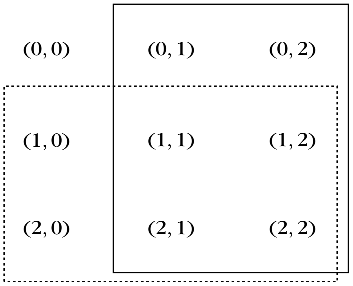

FIGURE 11.5

The elements of a contingency table taken into account for the computation of the Tanimoto measure.

In the special case $k = 2$, this equation results in [Tani 58, Spat 80]

$$s_T(\boldsymbol{x}, \boldsymbol{y}) = \frac{a_{11}}{a_{11} + a_{01} + a_{10}} \tag{11.31}$$

Other similarity functions between $\boldsymbol{x}, \boldsymbol{y} \in F^l$ can be defined using elements of $A(\boldsymbol{x}, \boldsymbol{y})$. Some of them consider only the number of places where the two vectors agree and the corresponding value is not 0, whereas others consider all the places where the two vectors agree. Similarity functions that belong to the first category are

$$\frac{\sum_{i=1}^{k-1} a_{ii}}{l} \quad \text{and} \quad \frac{\sum_{i=1}^{k-1} a_{ii}}{l - a_{00}} \tag{11.32}$$

A representative of the second category is

$$\frac{\sum_{i=0}^{k-1} a_{ii}}{l} \tag{11.33}$$

When dealing with binary-valued vectors (i.e., $k = 2$), probabilistic similarity measures have also been proposed [Good 66, Li 85, Broc 81]. For two binary-valued vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, a measure of this kind, s, is based on the number of positions where $\boldsymbol{x}$ and $\boldsymbol{y}$ agree. The value of $s(\boldsymbol{x}, \boldsymbol{y})$ is then compared with the distances of pairs of randomly chosen vectors, in order to conclude whether $\boldsymbol{x}$ and $\boldsymbol{y}$ are "close" to each other. This task is carried out using statistical tests (see also Chapter 16).

Dynamic Similarity Measures

The proximity measures discussed so far apply to vectors with the same dimension, l. However, in certain applications, such as the comparison of two strings st_1 and st_2 stemming from two different texts, this is not the case. For example, one of the two strings may be shifted with respect to the other. In these cases the preceding proximity measures fail. In such cases, dynamic similarity measures, such as the Edit distance, discussed in Chapter 8, can be used.

Mixed Valued Vectors

An interesting case, which often arises in practice, is when the features of the feature vectors are not all real or all discrete valued. In terms of Example 11.1, the third to the tenth features are real valued, and the second feature is discrete valued. A naive way to attack this problem is to adopt proximity measures (PMs) suitable for real-valued vectors. The reason is that discrete-valued vectors can be accurately compared in terms of PMs for real-valued vectors, whereas the opposite does not lead, in general, to reasonable results. A good PM candidate for such cases is the l_1 distance.

Example 11.5

Consider the vectors $\boldsymbol{x} = [4,\ 1,\ 0.8]^T$ and $\boldsymbol{y} = [1,\ 0,\ 0.4]^T$. Their (unweighted) l_1 and l_2 distances are

$$d_1(\boldsymbol{x},\ \boldsymbol{y}) = |4-1| + |1-0| + |0.8-0.4| = 3 + 1 + 0.4 = 4.4$$

and

$$d_2(\boldsymbol{x},\ \boldsymbol{y}) = \sqrt{|4-1|^2 + |1-0|^2 + |0.8-0.4|^2} = \sqrt{9 + 1 + 0.16} = 3.187$$

respectively. Notice that in the second case, the difference between the first coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$ specifies almost exclusively the difference between the two vectors. This is not the case with l_1 distance (see also related comments in Chapter 5, Section 5.2).

Another method that may be employed is to convert the real-valued features to discrete-valued ones, that is, to discretize the real-valued data. To this end, if a feature x_i takes values in the interval $[a, b]$, we may divide this interval into k subintervals. If the value of x_i lies in the rth subinterval, the value $r - 1$ will be assigned to it. This strategy leads to discrete-valued vectors, and as a consequence, we may use any of the measures discussed in the previous section.

In [Ande 73] the types nominal, ordinal, and interval-scaled types of features are considered and methods for converting features from one type to another are discussed. These are based on the fact (see Section 11.1.2) that as we move from nominal to interval scaled, we have to impose information on the specific feature, and when we move along the opposite direction, we have to give up information.

A similarity function that deals with mixed valued vectors, without making any conversions to the type of features, is proposed in [Gowe 71]. Let us consider two l-dimensional mixed valued vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Then, the similarity function between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as

$$s(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\sum_{q=1}^{l} s_q(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sum_{q=1}^{l} w_q} \tag{11.34}$$

where $s_q(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the similarity between the qth coordinates of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and w_q is a weight factor corresponding to the qth coordinate. Specifically, if at least one of the qth coordinates of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is undefined, then $w_q = 0$. Also, if the qth coordinate is a binary variable and it is 0 for both vectors, then $w_q = 0$. In all other cases, w_q is set equal to 1. Finally, if all w_q's are equal to 0 then $s(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is undefined. If the qth coordinates of the two vectors are binary then

$$s_q(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} 1, & \text{if } x_{iq} = x_{jq} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{11.35}$$

If the qth coordinates of the two vectors correspond to nominal or ordinal variables, then $s_q(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1$ if x_{iq} and x_{jq} have the same values. Otherwise, $s_q(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$.

Finally, if the qth coordinates correspond to interval or ratio scaled variables, then

$$s_q(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \frac{|x_{iq} - x_{jq}|}{r_q} \tag{11.36}$$

where r_q is the length of the interval where the values of the qth coordinates lie. One can easily observe that for the case of intervals or ratio-scaled variables, when x_{ik} and x_{jk} coincide, $s_q(\boldsymbol{x}_i, \boldsymbol{x}_j)$ takes its maximum value, which equals 1. On the other hand, if the absolute difference between x_{iq} and x_{jq} equals r_q, then $s_q(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$. For any other value of $|x_{iq} - x_{jq}|$, $s_q(\boldsymbol{x}_i, \boldsymbol{x}_j)$ lies between 0 and 1.

Example 11.6

Let us consider the following four 5-dimensional feature vectors, each representing a specific company. More specifically, the first three coordinates (features) correspond to their annual budget for the last three years (in millions of dollars), the fourth indicates whether or not there is any activity abroad, and the fifth coordinate corresponds to the number of employees of each company. The last feature is ordinal scaled and takes the values 0 (small number of employees), 1 (medium number of employees), and 2 (large number of employees). The four vectors are

Company	1st bud.	2nd bud.	3rd bud.	Act. abr.	Empl.
1 ($\boldsymbol{x}_1$)	1.2	1.5	1.9	0	1
2 ($\boldsymbol{x}_2$)	0.3	0.4	0.6	0	0
3 ($\boldsymbol{x}_3$)	10	13	15	1	2
4 ($\boldsymbol{x}_4$)	6	6	7	1	1

(11.37)

For the first three coordinates, which are ratio scaled, we have $r_1 = 9.7$, $r_2 = 12.6$, and $r_3 = 14.4$. Let us first compute the similarity between the first two vectors. It is

$$s_1(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 - |1.2 - 0.3|/9.7 = 0.9072$$

$$s_2(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 - |1.5 - 0.4|/12.6 = 0.9127$$

$$s_3(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 - |1.9 - 0.6|/14.4 = 0.9097$$

$$s_4(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0$$

and

$$s_5(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0$$

Also, $w_4 = 0$, while all the other weight factors are equal to 1. Using Eq. (11.34), we finally obtain $s(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0.6824$.

Working in the same way, we find that $s(\boldsymbol{x}_1, \boldsymbol{x}_3) = 0.0541$, $s(\boldsymbol{x}_1, \boldsymbol{x}_4) = 0.5588$, $s(\boldsymbol{x}_2, \boldsymbol{x}_3) = 0$, $s(\boldsymbol{x}_2, \boldsymbol{x}_4) = 0.3047$, $s(\boldsymbol{x}_3, \boldsymbol{x}_4) = 0.4953$.

Fuzzy Measures

In this section, we consider real-valued vectors $\boldsymbol{x}, \boldsymbol{y}$ whose components x_i and y_i belong to the interval $[0, 1]$, $i = 1, \ldots, l$. In contrast to what we have said so far, *the values of x_i are not the outcome of a measuring device*. The closer the x_i to 1 (0), the more likely $\boldsymbol{x}$ possesses (does not possess) the ith feature (characteristic).[2] As x_i approaches 1/2, we become less certain about the possession or not of the ith feature from $\boldsymbol{x}$. When $x_i = 1/2$ we have absolutely no clue whether or not $\boldsymbol{x}$ possesses the ith feature. It is easy to observe that this situation is a generalization of binary logic, where x_i can take only the value 0 or 1 ($\boldsymbol{x}$ possesses a feature or not). In binary logic, there is a certainty about the occurrence of a fact (for example, it will rain or it will not rain). The idea of fuzzy logic is that nothing is happening or not happening with absolute certainty. This is reflected in the values that x_i takes. The binary logic can be viewed as a special case of fuzzy logic where x_i takes only the value 0 or 1.

Next, we will define the similarity between two real-valued variables in $[0, 1]$. We will approach it as a generalization of the equivalence between two binary variables. The equivalence of two binary variables a and b is given by the following relation:

$$(a \equiv b) = ((NOT\ a)\ AND\ (NOT\ b))\ OR\ (a\ AND\ b) \tag{11.38}$$

Indeed, if $a = b = 0$ (1), the first (second) argument of the *OR* operator is 1. On the other hand if $a = 0$ (1) and $b = 1$ (0), then none of the arguments of the *OR* operator becomes 1.

An interesting observation is that the *AND* (*OR*) operator between two binary variables may be seen as the min (max) operator on them. Also, the *NOT* operation of a binary variable a may be written as $1-a$. In the fuzzy logic context and based on this observation, the logical *AND* is replaced by the operator min, while the logical *OR* is replaced by the operator max. Also, the logical *NOT* on x_i is replaced by $1 - x_i$ [Klir 95]. This suggests that the degree of similarity between two real-valued variables x_i and y_i in $[0, 1]$ may be defined as

$$s(x_i, y_i) = \max(\min(1 - x_i, 1 - y_i), \min(x_i, y_i)) \tag{11.39}$$

Note that this definition includes the special case where x_i and y_i take binary values and results in (11.38).

When we now deal with vectors in the l-dimensional space ($l > 1$), the vector space is the H_l hypercube. In this context, the closer a vector $\boldsymbol{x}$ lies to the center of H_l $(1/2, \ldots, 1/2)$, the greater the amount of uncertainty. That is, in this case we have almost no clue whether $\boldsymbol{x}$ possesses any of the l features. On the other hand, the closer $\boldsymbol{x}$ lies to a vertex of H_l, the less the uncertainty.

Based on similarity s between two variables in $[0, 1]$ given in (11.39), we are now able to define a similarity measure between two vectors. A common similarity

[2] The ideas of this section follow [Zade 73].

measure between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$s_F^q(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{l} s(x_i, y_i)^q \right)^{1/q} \tag{11.40}$$

It is easy to verify that the maximum and minimum values of s_F are $l^{1/q}$ and $0.5l^{1/q}$, respectively. As $q \to +\infty$, we get $s_F(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \le i \le l} s(x_i, y_i)$. Also, when $q = 1, s_F(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{l} s(x_i, y_i)$ (Problem 11.7).

Example 11.7

In this example we consider the case where $l = 3$ and $q = 1$. Under these circumstances, the maximum possible value of s_F is 3. Let us consider the vectors $\boldsymbol{x}_1 = [1, 1, 1]^T$, $\boldsymbol{x}_2 = [0, 0, 1]^T$, $\boldsymbol{x}_3 = [1/2, 1/3, 1/4]^T$, and $\boldsymbol{x}_4 = [1/2, 1/2, 1/2]^T$. If we compute the similarities of these vectors with themselves, we obtain

$$s_F^1(\boldsymbol{x}_1, \boldsymbol{x}_1) = 3\max(\min(1 - 1, 1 - 1), \min(1, 1)) = 3$$

and similarly, $s_F^1(\boldsymbol{x}_2, \boldsymbol{x}_2) = 3$, $s_F^1(\boldsymbol{x}_3, \boldsymbol{x}_3) = 1.92$, and $s_F^1(\boldsymbol{x}_4, \boldsymbol{x}_4) = 1.5$. *This is very interesting. The similarity measure of a vector with itself depends not only on the vector but also on its position in the H_l hypercube.* Furthermore, we observe that the greatest similarity value is obtained at the vertices of H_l. As we move toward the center of H_l, the similarity measure between a vector and itself decreases, attaining its minimum value at the center of H_l.

Let us now consider the vectors $\boldsymbol{y}_1 = [3/4, 3/4, 3/4]^T$, $\boldsymbol{y}_2 = [1, 1, 1]^T$, $\boldsymbol{y}_3 = [1/4, 1/4, 1/4]^T$, $\boldsymbol{y}_4 = [1/2, 1/2, 1/2]^T$. Notice that in terms of the Euclidean distance $d_2(\boldsymbol{y}_1, \boldsymbol{y}_2) = d_2(\boldsymbol{y}_3, \boldsymbol{y}_4)$. However, $s_F^1(\boldsymbol{y}_1, \boldsymbol{y}_2) = 2.25$ and $s_F^1(\boldsymbol{y}_3, \boldsymbol{y}_4) = 1.5$. These results suggest that the closer the two vectors to the center of H_l, the less their similarity. On the other hand, the closer the two vectors to a vertex of H_l, the greater their similarity. *That is, the value of $s_F^q(\boldsymbol{x}, \boldsymbol{y})$ depends not only on the relative position of $\boldsymbol{x}$ and $\boldsymbol{y}$ in H_l but also on their closeness to the center of H_l.*

Missing Data

A problem that is commonly met in real-life applications is that of missing data. This means that for some feature vectors we do not know all of their components. This may be a consequence of a failure of the measuring device. Also, in cases such as the one mentioned in Example 11.1, missing data may be the result of a recording error. The following are some commonly used techniques that handle this situation [Snea 73, Dixo 79, Jain 88].

1. Discard all feature vectors that have missing features. This approach may be used when the number of vectors with missing features is small compared to the total number of available feature vectors. If this is not the case, the nature of the problem may be affected.

2. For the ith feature, find its mean value based on the corresponding available values of all feature vectors of X. Then, substitute this value for the vectors where their ith coordinate is not available.

3. For all the pairs of components x_i and y_i of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ define b_i as

$$b_i = \begin{cases} 0, & \text{if both } x_i \text{ and } y_i \text{ are available} \\ 1, & \text{otherwise} \end{cases} \tag{11.41}$$

   Then, the proximity between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$\wp(\boldsymbol{x}, \boldsymbol{y}) = \frac{l}{l - \sum_{i=1}^{l} b_i} \sum_{\text{all } i: b_i = 0} \phi(x_i, y_i) \tag{11.42}$$

   where $\phi(x_i, y_i)$ denotes the proximity between the two scalars x_i and y_i. A common choice of ϕ when a dissimilarity measure is involved, is $\phi(x_i, y_i) = |x_i - y_i|$. The rationale behind this approach is simple. Let $[a, b]$ be the interval of the allowable values of $\wp(\boldsymbol{x}, \boldsymbol{y})$. The preceding definition ensures that the proximity measure between $\boldsymbol{x}$ and $\boldsymbol{y}$ spans all $[a, b]$, regardless of the number of unavailable features in both vectors.

4. Find the average proximities $\phi_{\text{avg}}(i)$ between all feature vectors in X along all components $i = 1, \ldots, l$. It is clear that for some vectors $\boldsymbol{x}$ the ith component is not available. In that case, the proximities that include x_i are excluded from the computation of $\phi_{\text{avg}}(i)$. We define the proximity $\psi(x_i, y_i)$ between the ith components of $\boldsymbol{x}$ and $\boldsymbol{y}$ as $\phi_{\text{avg}}(i)$ if at least one of the x_i and y_i is not available, and as $\phi(x_i, y_i)$ if both x_i and y_i are available ($\phi(x_i, y_i)$ may be defined as in the previous case). Then,

$$\wp(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{l} \psi(x_i, y_i) \tag{11.43}$$

Example 11.8

Consider the set $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$, where $\boldsymbol{x}_1 = [0, 0]^T$, $\boldsymbol{x}_2 = [1, *]^T$, $\boldsymbol{x}_3 = [0, *]^T$, $\boldsymbol{x}_4 = [2, 2]^T$, $\boldsymbol{x}_5 = [3, 1]^T$. The "$*$" means that the corresponding value is not available.

According to the second technique, we find the average value of the second feature, which is 1, and we substitute it for the "$*$"s. Then, we may use any of the proximity measures defined in the previous sections.

Assume now that we wish to find the distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ using the third technique. We use the absolute difference as the distance between two scalars. Then $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{2}{2-1} 1 = 2$. Similarly, $d(\boldsymbol{x}_2, \boldsymbol{x}_3) = \frac{2}{2-1} 1 = 2$.

Finally, if we choose the fourth of the techniques, we must first find the average of the distances between any two values of the second feature. We again use the absolute difference as the distance between two scalars. The distances between any two available values of the second feature are $|0-2|=2$, $|0-1|=1$, and $|2-1|=1$, and the average is $4/3$. Thus, the distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 + 4/3 = 7/3$.

11.2.3 Proximity Functions between a Point and a Set

In many clustering schemes, a vector $\boldsymbol{x}$ is assigned to a cluster C taking into account the proximity between $\boldsymbol{x}$ and C, $\wp(\boldsymbol{x}, C)$. There are two general directions for the definition of $\wp(\boldsymbol{x}, C)$. According to the first one, all points of C contribute to $\wp(\boldsymbol{x}, C)$. Typical examples of this case include:

- The *max proximity function:*

$$\wp_{\max}^{ps}(\boldsymbol{x}, C) = \max_{\boldsymbol{y}\in C} \wp(\boldsymbol{x}, \boldsymbol{y}) \tag{11.44}$$

- The *min proximity function:*

$$\wp_{\min}^{ps}(\boldsymbol{x}, C) = \min_{\boldsymbol{y}\in C} \wp(\boldsymbol{x}, \boldsymbol{y}) \tag{11.45}$$

- The *average proximity function:*

$$\wp_{\text{avg}}^{ps}(\boldsymbol{x}, C) = \frac{1}{n_C} \sum_{\boldsymbol{y}\in C} \wp(\boldsymbol{x}, \boldsymbol{y}) \tag{11.46}$$

  where n_C is the cardinality of C.

In these definitions, $\wp(\boldsymbol{x}, \boldsymbol{y})$ may be any proximity measure between two points.

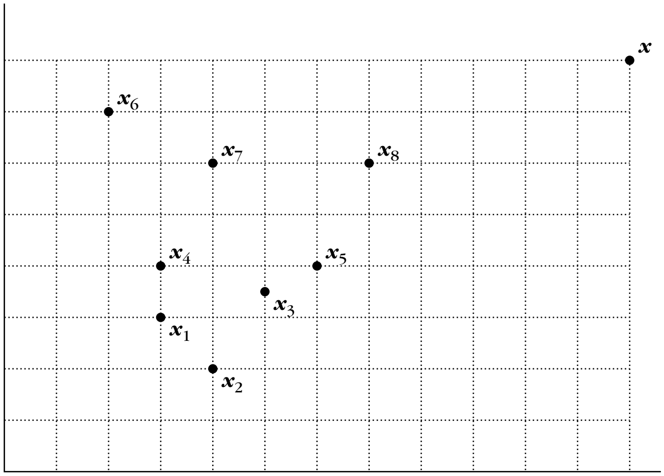

FIGURE 11.6

The setup of Example 11.9.

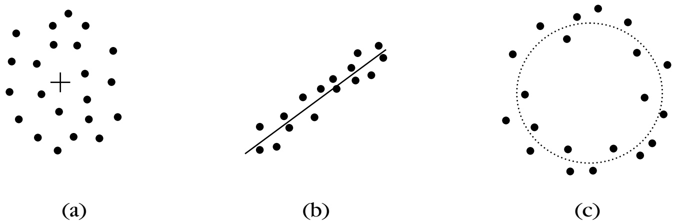

FIGURE 11.7

(a) Compact cluster. (b) Hyperplanar (linear) cluster. (c) Hyperspherical cluster.

Example 11.9

Let $C = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_8\}$, where $\boldsymbol{x}_1 = [1.5, 1.5]^T$, $\boldsymbol{x}_2 = [2, 1]^T$, $\boldsymbol{x}_3 = [2.5, 1.75]^T$, $\boldsymbol{x}_4 = [1.5, 2]^T$, $\boldsymbol{x}_5 = [3, 2]^T$, $\boldsymbol{x}_6 = [1, 3.5]^T$, $\boldsymbol{x}_7 = [2, 3]^T$, $\boldsymbol{x}_8 = [3.5, 3]^T$, and let $\boldsymbol{x} = [6, 4]^T$ (see Figure 11.6). Assume that the Euclidean distance is used to measure the dissimilarity between two points. Then $d_{\max}^{ps}(\boldsymbol{x}, C) = \max_{\boldsymbol{y} \in C} d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{x}_1) = 5.15$. For the other two distances we have $d_{\min}^{ps}(\boldsymbol{x}, C) = \min_{\boldsymbol{y} \in C} d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{x}_8) = 2.69$ and $d_{\text{avg}}^{ps}(\boldsymbol{x}, C) = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} d(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{8} \sum_{i=1}^{8} d(\boldsymbol{x}, \boldsymbol{x}_i) = 4.33$.

According to the second direction, C is equipped with a representative and the proximity between $\boldsymbol{x}$ and C is measured as the proximity between $\boldsymbol{x}$ and the representative of C. Many types of representatives have been used in the literature. Among them, the point, the hyperplane, and the hypersphere are most commonly used.[3] Point representatives are suitable for compact clusters (Figure 11.7a) and hyperplane (hyperspherical) representatives for clusters of linear shape (Figure 11.7b) (hyperspherical shape, Figure 11.7c).

Point Representatives

Typical choices for a point representative of a cluster are:

- The *mean vector* (or *mean point*)

$$\boldsymbol{m}_p = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} \boldsymbol{y} \tag{11.47}$$

  where n_C is the cardinality of C. This is the most common choice when point representatives are employed, and we deal with data of a continuous space. However, it may not work well when we deal with points of a discrete space F^l. This is because it is possible for $\boldsymbol{m}_p$ to lie outside F^l. To cope with this problem, we may use the mean center $\boldsymbol{m}_c$ of C, which is defined next.

[3] In Chapter 14 we discuss the more general family of hyperquadric representatives, which include hyperellipsoids, hyperparabolas, and pairs of hyperplanes.

- The *mean center* $\boldsymbol{m}_c \in C$ is defined as the point for which

$$\sum_{y \in C} d(\boldsymbol{m}_c, \boldsymbol{y}) \leq \sum_{y \in C} d(\boldsymbol{z}, \boldsymbol{y}), \quad \forall \boldsymbol{z} \in C \tag{11.48}$$

where d is a dissimilarity measure between two points. When similarity measures are involved, the inequality is reversed.

Another commonly used point representative is the median center. It is usually employed when the proximity measure between two points is not a metric.

- The *median center* $\boldsymbol{m}_{\text{med}} \in C$ is defined as the point for which

$$\text{med}(d(\boldsymbol{m}_{\text{med}}, \boldsymbol{y})|\boldsymbol{y} \in C) \leq \text{med}(d(\boldsymbol{z}, \boldsymbol{y})|\boldsymbol{y} \in C), \quad \forall \boldsymbol{z} \in C \tag{11.49}$$

where d is a dissimilarity measure between two points. Here med(T), with T being a set of q scalars, is the minimum number in T that is greater than or equal to exactly $[(q+1)/2]$ numbers of T. An algorithmic way to determine med(T) is to list the elements of T in increasing order and to pick the $[(q+1)/2]$ element of that list.

Example 11.10

Let $C = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$, where $\boldsymbol{x}_1 = [1,1]^T$, $\boldsymbol{x}_2 = [3,1]^T$, $\boldsymbol{x}_3 = [1,2]^T$, $\boldsymbol{x}_4 = [1,3]^T$, and $\boldsymbol{x}_5 = [3,3]^T$ (see Figure 11.8). All points lie in the discrete space $\{0,1,2,\ldots,6\}^2$. We use the Euclidean distance to measure the dissimilarity between two vectors in C. The mean point of C is $\boldsymbol{m}_p = [1.8, 2]^T$. It is clear that $\boldsymbol{m}_p$ lies outside the space where the elements of C belong.

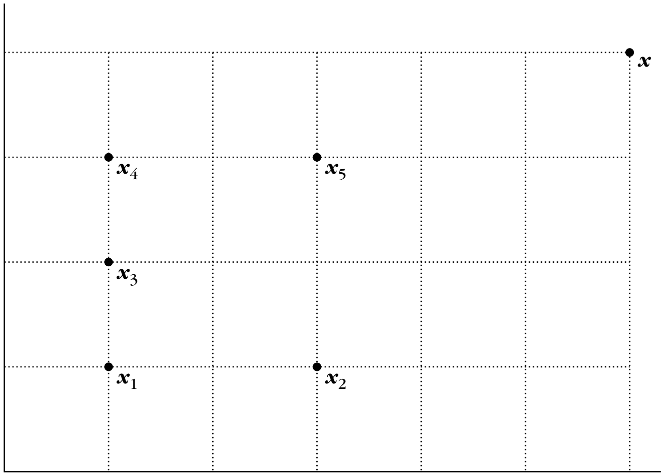

FIGURE 11.8

The setup of Example 11.10.

To find the mean center $\boldsymbol{m}_c$, we compute, for each point $\boldsymbol{x}_i \in C$, $i = 1, \ldots, 5$, the sum A_i of its distances from all other points of C. The resulting values are $A_1 = 7.83$, $A_2 = 9.06$, $A_3 = 6.47$, $A_4 = 7.83$, $A_5 = 9.06$. The minimum of these values is A_3. Thus, $\boldsymbol{x}_3$ is the mean center of C.

Finally, for the computation of the median center $\boldsymbol{m}_{\text{med}}$ we work as follows. For each vector $\boldsymbol{x}_i \in C$ we form the $n_C \times 1$ dimensional vector T_i of the distances between $\boldsymbol{x}_i$ and each of the vectors of C. Working as indicated, we identify $\text{med}(T_i)$, $i = 1, \ldots, 5$. Thus, $\text{med}(T_1) = \text{med}(T_2) = 2$, $\text{med}(T_3)=1$, $\text{med}(T_4) = \text{med}(T_5) = 2$. Then we choose $\text{med}(T_j) = \min_{i=1,\ldots,n_C}\{\text{med}(T_i)\} = \text{med}(T_3)$, and we identify $\boldsymbol{x}_3$ as the median vector of C. In our example, the mean center and the median center coincide. In general, however, this is not the case.

The distances between $\boldsymbol{x} = [6, 4]^T$ and C when the mean point, the mean center, and the median center are used as representatives of C are 4.65, 5.39, and 5.39, respectively.

Hyperplane Representatives

Linear shaped clusters (or hyperplanar in the general case) are often encountered in computer vision applications. This type of cluster cannot be accurately represented by a single point. In such cases we use lines (hyperplanes) as representatives of the clusters (e.g., [Duda 01]).

The general equation of a hyperplane H is

$$\sum_{j=1}^{l} a_j x_j + a_0 = \boldsymbol{a}^T \boldsymbol{x} + a_0 = 0 \tag{11.50}$$

where $\boldsymbol{x} = [x_1, \ldots, x_l]^T$ and $\boldsymbol{a} = [a_1, \ldots, a_l]^T$ is the weight vector of H. The distance of a point $\boldsymbol{x}$ from H is defined as

$$d(\boldsymbol{x}, H) = \min_{\boldsymbol{z} \in H} d(\boldsymbol{x}, \boldsymbol{z}) \tag{11.51}$$

In the case of Euclidean distance between two points and using simple geometric arguments (see Figure 11.9a), we obtain

$$d(\boldsymbol{x}, H) = \frac{|\boldsymbol{a}^T \boldsymbol{x} + a_0|}{\|\boldsymbol{a}\|} \tag{11.52}$$

where $\|\boldsymbol{a}\| = \sqrt{\sum_{j=1}^{l} a_j^2}$.

Hyperspherical Representatives

Clusters of another type are those that are circular (hyperspherical in higher dimensions). These are also frequently encountered in computer vision applications. For such clusters, the ideal representative is a circle (hypersphere).

The general equation of a hypersphere Q is

$$(\boldsymbol{x} - \boldsymbol{c})^T (\boldsymbol{x} - \boldsymbol{c}) = r^2 \tag{11.53}$$

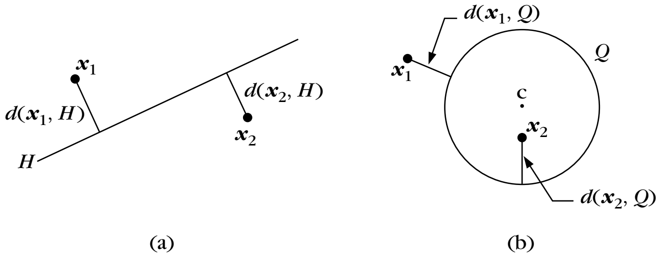

FIGURE 11.9

(a) Distance between a point and a hyperplane. (b) Distance between a point and hypersphere.

where $\boldsymbol{c}$ is the center of the hypersphere and r its radius. The distance from a point $\boldsymbol{x}$ to Q is defined as

$$d(\boldsymbol{x}, Q) = \min_{\boldsymbol{z} \in Q} d(\boldsymbol{x}, \boldsymbol{z}) \tag{11.54}$$

In most of the cases of interest, the Euclidean distance between two points is used in this definition. Figure 11.9b provides geometric insight into this definition. However, other nongeometric distances $d(\boldsymbol{x}, Q)$ have been used in the literature (e.g., [Dave 92, Kris 95, Frig 96]).

11.2.4 Proximity Functions between Two Sets

So far, we have been concerned with proximity measures between points in l-dimensional spaces and proximity functions between points and sets. Our major focus now is on defining proximity functions between sets of points. As we will soon see, some of the clustering algorithms are built upon such information. Most of the proximity functions $\wp^{ss}$ used for the comparison of sets are based on proximity measures, $\wp$, between vectors (see [Duda 01]). If D_i, D_j are two sets of vectors, the most common proximity functions are:

- The *max proximity function*:

$$\wp^{ss}_{\max}(D_i, D_j) = \max_{\boldsymbol{x} \in D_i, \boldsymbol{y} \in D_j} \wp(\boldsymbol{x}, \boldsymbol{y}) \tag{11.55}$$

  It is easy to see that if $\wp$ is a dissimilarity measure, $\wp^{ss}_{\max}$ is not a measure, since it does not satisfy the conditions in Section 11.2.1. $\wp^{ss}_{\max}$ is fully determined by the pair $(\boldsymbol{x}, \boldsymbol{y})$ of the most dissimilar (distant) vectors, with $\boldsymbol{x} \in D_i$ and $\boldsymbol{y} \in D_j$. On the other hand, if $\wp$ is a similarity measure, $\wp^{ss}_{\max}$ is a measure but it is not a metric (see Problem 11.12). In that case $\wp^{ss}_{\max}$ is fully determined by the pair $(\boldsymbol{x}, \boldsymbol{y})$ of the most similar (closest) vectors, with $\boldsymbol{x} \in D_i$ and $\boldsymbol{y} \in D_j$.

- The *min proximity function*:

$$\wp^{ss}_{\min}(D_i, D_j) = \min_{\boldsymbol{x} \in D_i, \boldsymbol{y} \in D_j} \wp(\boldsymbol{x}, \boldsymbol{y}) \tag{11.56}$$

When $\wp$ is a similarity measure, $\wp_{\min}^{ss}$ is not a measure. In this case $\wp_{\min}^{ss}$ is fully determined by the pair $(\boldsymbol{x}, \boldsymbol{y})$ of the most dissimilar (distant) vectors, with $\boldsymbol{x} \in D_i$ and $\boldsymbol{y} \in D_j$. On the other hand, if $\wp$ is a dissimilarity measure, $\wp_{\min}^{ss}$ is a measure, but it is not a metric (see Problem 11.12). In this case $\wp_{\min}^{ss}$ is fully determined by the pair $(\boldsymbol{x}, \boldsymbol{y})$ of the most similar (closest) vectors, with $\boldsymbol{x} \in D_i$ and $\boldsymbol{y} \in D_j$.

- The *average proximity function*:

$$\wp_{\text{avg}}^{ss}(D_i, D_j) = \frac{1}{n_{D_i} n_{D_j}} \sum_{\boldsymbol{x} \in D_i} \sum_{\boldsymbol{y} \in D_j} \wp(\boldsymbol{x}, \boldsymbol{y}) \tag{11.57}$$

where n_{D_i} and n_{D_j} are the cardinalities of D_i and D_j, respectively. It is easily shown that $\wp_{\text{avg}}^{ss}$ is not a measure even though $\wp$ is a measure. In this case, all vectors of both D_i and D_j contribute to the computation of $\wp_{\text{avg}}^{ss}$.

- The *mean proximity function*:

$$\wp_{\text{mean}}^{ss}(D_i, D_j) = \wp(\boldsymbol{m}_{\boldsymbol{D_i}}, \boldsymbol{m}_{\boldsymbol{D_j}}) \tag{11.58}$$

where $\boldsymbol{m}_{\boldsymbol{D_i}}$ is the representative of D_i, $i = 1, 2$. For example, $\boldsymbol{m}_{\boldsymbol{D_i}}$ may be the mean point, the mean center, or the median of D_i. Obviously, this is the proximity function between the representatives of D_i and D_j. It is clear that the mean proximity function is a measure provided that $\wp$ is a measure.

- Another proximity function that will be used later on is based on the mean proximity function and is defined as[4]

$$\wp_{e}^{ss}(D_i, D_j) = \sqrt{\frac{n_{D_i} n_{D_j}}{n_{D_i} + n_{D_j}}} \wp(\boldsymbol{m}_{\boldsymbol{D_i}}, \boldsymbol{m}_{\boldsymbol{D_j}}) \tag{11.59}$$

where $\boldsymbol{m}_{\boldsymbol{D_i}}$ is defined as in the previous case.

In the last two alternatives we consider only the cases in which D_i's are represented by points. The need for a definition of a proximity function between two sets via their representatives, when the latter are not points, is of limited practical interest.

Example 11.11

(a) Consider the set $D_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4\}$ and $D_2 = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4\}$, with $\boldsymbol{x}_1 = [0, 0]^T$, $\boldsymbol{x}_2 = [0, 2]^T$, $\boldsymbol{x}_3 = [2, 0]^T$, $\boldsymbol{x}_4 = [2, 2]^T$, $\boldsymbol{y}_1 = [-3, 0]^T$, $\boldsymbol{y}_2 = [-5, 0]^T$, $\boldsymbol{y}_3 = [-3, -2]^T$, $\boldsymbol{y}_4 = [-5, -2]^T$. The Euclidean distance is employed as the distance between two vectors. The distances between D_1 and D_2 according to the proximity functions just defined are $d_{\min}^{ss}(D_1, D_2) = 3$, $d_{\max}^{ss}(D_1, D_2) = 8.06$, $d_{\text{avg}}^{ss}(D_1, D_2) = 5.57$, $d_{\text{mean}}^{ss}(D_1, D_2) = 5.39$, $d_{e}^{ss}(D_1, D_2) = 7.62$.

[4] This definition is a generalization of that given in [Ward 63] (see Chapter 13).

(b) Consider now the set $D_2^{'} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4\}$, with $\boldsymbol{z}_1 = [1, 1.5]^T$, $\boldsymbol{z}_2 = [1, 0.5]^T$, $\boldsymbol{z}_3 = [0.5, 1]^T$, $\boldsymbol{z}_4 = [1.5, 1]^T$. Notice that the points of D_1 and $D_2^{'}$ lie in two concentric circles centered at $[1, 1]^T$. The radius corresponding to D_1 ($D_2^{'}$) is $\sqrt{2}$ (0.5). The distances between D_1 and $D_2^{'}$ according to the proximity functions are $d_{\min}^{ss}(D_1, D_2^{'}) = 1.19$, $d_{\max}^{ss}(D_1, D_2^{'}) = 1.80$, $d_{\text{avg}}^{ss}(D_1, D_2^{'}) = 1.46$, $d_{\text{mean}}^{ss}(D_1, D_2^{'}) = 0$, $d_e^{ss}(D_1, D_2^{'}) = 0$.

Notice that in the last case, in which one of the sets lies in the convex hull of the other, some proximity measures may not be appropriate. For example, the measure based on the distances between the two means of the clusters gives meaningless results. However, this distance is well suited for cases in which the two sets are compact and well separated, especially because of its low computational requirements.

Notice that the proximities between two sets are built on proximities between two points. *Intuitively, one can understand that different choices of proximity functions between sets may lead to totally different clustering results.* Moreover, if we use different proximity measures between points, the same proximity function between sets will lead, in general, to different clustering results. *The only way to achieve proper clustering of the data is by trial and error and, of course, by taking into account the opinion of an expert in the field of application.*

Finally, proximity functions between a vector $\boldsymbol{x}$ and a set D_i may also be derived from the functions defined here, if we set $D_j = \{\boldsymbol{x}\}$.

11.3 PROBLEMS

11.1 Let s be a metric similarity measure on X with $s(\boldsymbol{x}, \boldsymbol{y}) > 0$, $\forall \boldsymbol{x}, \boldsymbol{y} \in X$ and $d(\boldsymbol{x}, \boldsymbol{y}) = a/s(\boldsymbol{x}, \boldsymbol{y})$, with $a > 0$. Prove that d is a metric dissimilarity measure.

11.2 Prove that the Euclidean distance satisfies the triangular inequality.
Hint: Use the Minkowski inequality, which states that for a positive integer p and two vectors $\boldsymbol{x} = [x_1, \ldots, x_l]^T$ and $\boldsymbol{y} = [y_1, \ldots, y_l]^T$ it holds that

$$\left(\sum_{i=1}^{l} |x_i + y_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^{l} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{l} |y_i|^p\right)^{1/p}$$

11.3 Show that:

a. if s is a metric similarity measure on a set X with $s(\boldsymbol{x}, \boldsymbol{y}) \geq 0, \forall \boldsymbol{x}, \boldsymbol{y} \in X$, then $s(\boldsymbol{x}, \boldsymbol{y}) + a$ is also a metric similarity measure on $X, \forall a \geq 0$.

b. If d is a metric dissimilarity measure on X, then $d + a$ is also a metric dissimilarity measure on $X, \forall a \geq 0$.

11.4 Let $f : \mathcal{R}^+ \rightarrow \mathcal{R}^+$ be a continuous monotonically increasing function such that

$$f(x) + f(y) \geq f(x + y), \quad \forall x, y \in \mathcal{R}^+$$

and let d be a metric dissimilarity measure on a set X with $d_0 \geq 0$. Show that $f(d)$ is also a metric dissimilarity measure on X.

11.5 Let s be a metric similarity measure on a set X, with $s(\boldsymbol{x},\boldsymbol{y}) > 0, \forall \boldsymbol{x},\boldsymbol{y} \in X$ and $f : \mathcal{R}^+ \rightarrow \mathcal{R}^+$ be a continuous monotonically decreasing function such that

$$f(x) + f(y) \geq f\left(\frac{1}{\frac{1}{x} + \frac{1}{y}}\right), \quad \forall x,y \in \mathcal{R}^+$$

Show that $f(s)$ is a metric dissimilarity measure on X.

11.6 Prove that

$$d_\infty(\boldsymbol{x},\boldsymbol{y}) \leq d_2(\boldsymbol{x},\boldsymbol{y}) \leq d_1(\boldsymbol{x},\boldsymbol{y})$$

for any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in X.

11.7 **a.** Prove that the maximum and the minimum values of $s_F(\boldsymbol{x},\boldsymbol{y})$ given in (11.40) are $l^{1/q}$ and $0.5l^{1/q}$, respectively.

b. Prove that as $q \rightarrow +\infty$, Eq. (11.40) results in $s_F(\boldsymbol{x},\boldsymbol{y}) = \max_{1 \leq i \leq l} s(x_i, y_i)$.

11.8 Examine whether the similarity functions defined by Eqs. (11.32), (11.33) are metric SMs.

11.9 Let d be a dissimilarity measure on X and $s = d_{\max} - d$ a corresponding similarity measure. Prove that

$$s_{\text{avg}}^{ps}(\boldsymbol{x}, C) = d_{\max} - d_{\text{avg}}^{ps}(\boldsymbol{x}, C), \quad \forall \boldsymbol{x} \in X, \quad C \subset X$$

where s_{avg}^{ps} and d_{avg}^{ps} are defined in terms of s and d, respectively. The definition of $\wp_{\text{avg}}^{ps}$ may be obtained from (11.57), where the first set consists of a single vector.

11.10 Let $\boldsymbol{x},\boldsymbol{y} \in \{0,1\}^l$. Prove that $d_2(\boldsymbol{x},\boldsymbol{y}) = \sqrt{d_{Hamming}(\boldsymbol{x},\boldsymbol{y})}$.

11.11 Consider two points in an l-dimensional space, $\boldsymbol{x} = [x_1,\ldots,x_l]^T$ and $\boldsymbol{y} = [y_1,\ldots,y_l]^T$, and let $|x_i - y_i| = \max_{j=1,\ldots,l}\{|x_j - y_j|\}$. We define the distance $d_n(\boldsymbol{x},\boldsymbol{y})$ as

$$d_n(\boldsymbol{x},\boldsymbol{y}) = |x_i - y_i| + \frac{1}{l - [(l-2)/2]} \sum_{j=1, j\neq i}^{l} |x_j - y_j|$$

This distance has been proposed in [Chau 92] as an approximation of the d_2 (Euclidean) distance.

a. Prove that d_n is a metric.

b. Compare d_n with d_2 in terms of computational complexity.

11.12 Let d and s be a dissimilarity and a similarity measure, respectively. Let $d^{ss}_{\min}$ ($s^{ss}_{\min}$), $d^{ss}_{\max}$ ($s^{ss}_{\max}$), d^{ss}_{avg} (s^{ss}_{avg}), d^{ss}_{mean} (s^{ss}_{mean}) be defined in terms of $d(s)$.

a. Prove that $d^{ss}_{\min}$, d^{ss}_{mean} are measures and $d^{ss}_{\max}$, d^{ss}_{avg} are not.

b. Prove that $s^{ss}_{\max}$, s^{ss}_{mean} are measures while $s^{ss}_{\min}$, s^{ss}_{avg} are not.

11.13 Based on Eqs. (11.55),(11.56),(11.57),and (11.58), derive the corresponding proximity functions between a point and a set. Are these proximity functions measures?

REFERENCES

[Ande 73] Anderberg M.R. *Cluster Analysis for Applications*, Academic Press, 1973.

[Ball 71] Ball G.H. "Classification analysis," Stanford Research Institute, *SRI Project 5533*, 1971.

[Broc 81] Brockett P.L., Haaland P.D., Levine A. "Information theoretic analysis of questionnaire data," *IEEE Transactions on Information Theory*, Vol. 27, pp. 438–445, 1981.

[Chau 92] Chaudhuri D., Murthy C.A., Chaudhuri B.B. "A modified metric to compute distance," *Pattern Recognition*, Vol. 25(7), pp. 667–677, 1992.

[Dave 92] Dave R.N., Bhaswan K. "Adaptive fuzzy c-shells clustering and detection of ellipses," *IEEE Transactions on Neural Networks*, Vol. 3(5), pp. 643–662, 1992.

[Dixo 79] Dixon J.K. "Pattern recognition with partly missing data," *IEEE Transactions on Systems Man and Cybernetics*, Vol. SMC 9, 617–621, 1979.

[Duda 01] Duda R.O., Hart P., Stork D. *Pattern Classification*, 2nd ed., John Wiley & Sons, 2001.

[Eise 98] Eisen M., Spellman P., Brown P., Botstein D. "Cluster analysis and display of genome-wide expression data," *Proceedings of National Academy of Science, USA*, Vol. 95, pp. 14863–14868, 1998.

[Ever 01] Everitt B., Landau S., Leesse M. *Cluster Analysis*, Arnold, 2001.

[Frig 96] Frigui H., Krishnapuram R. "A comparison of fuzzy shell clustering methods for the detection of ellipses," *IEEE Transactions on Fuzzy Systems*, Vol. 4(2), May 1996.

[Fu 93] Fu L., Yang M., Braylan R., Benson N. "Real-time adaptive clustering of flow cytometric data," *Pattern Recognition*, Vol. 26(2), pp. 365–373, 1993.

[Gers 92] Gersho A., Gray R.M. *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[Good 66] Goodall D.W. "A new similarity index based on probability," *Biometrics*, Vol. 22, pp. 882–907, 1966.

[Gowe 67] Gower J.C. "A comparison of some methods of cluster analysis," *Biometrics*, Vol. 23, pp. 623–637, 1967.

[Gowe 71] Gower J.C. "A general coefficient of similarity and some of its properties," *Biometrics*, Vol. 27, pp. 857–872, 1971.

[Gowe 86] Gower J.C., Legendre P. "Metric and Euclidean properties of dissimilarity coefficients," *Journal of Classification*, Vol. 3, pp. 5–48, 1986.

[Hall 67] Hall A.V. "Methods for demonstrating resemblance in taxonomy and ecology," *Nature*, Vol. 214, pp. 830–831, 1967.

[Huba 82] Hubalek Z. "Coefficients of association and similarity based on binary (presence-absence) data—an evaluation," *Biological Review*, Vol. 57, pp. 669–689, 1982.

[Jain 88] Jain A.K., Dubes R.C. *Algorithms for Clustering Data*, Prentice Hall, 1988.

[John 67] Johnson S.C. "Hierarchical clustering schemes," *Psychometrika*, Vol. 32, pp. 241–254, 1967.

[Klir 95] Klir G., Yuan B. *Fuzzy sets and fuzzy logic*, Prentice Hall, 1995.

[Koho 89] Kohonen T. *Self-Organization and Associative Memory*, Springer-Verlag, 1989.

[Kris 95] Krishnapuram R., Frigui H., Nasraoui O. "Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation—Part I," *IEEE Transactions on Fuzzy Systems*, Vol. 3(1), pp. 29–43, February 1995.

[Li 85] Li X., Dubes R.C. "The first stage in two-stage template matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 7, pp. 700–707, 1985.

[Lipp 87] Lippmann R.P. "An introduction to computing with neural nets," *IEEE ASSP Magazine*, Vol. 4(2), April 1987.

[Payk 72] Paykel E.S. "Depressive typologies and response to amitriptyline," *British Journal of Psychiatry*, Vol. 120, pp. 147–156, 1972.

[Snea 73] Sneath P.H.A., Sokal R.R. *Numerical Taxonomy*, W.H. Freeman & Co., 1973.

[Soka 63] Sokal R.R., Sneath P.H.A. *Principles of Numerical Taxonomy*, W.H. Freeman & Co., 1963.

[Spat 80] Spath H. *Cluster Analysis Algorithms*, Ellis Horwood, 1980.

[Tani 58] Tanimoto T. "An elementary mathematical theory of classification and prediction," *Int. Rpt.*, IBM Corp., 1958.

[Wall 68] Wallace C.S., Boulton D.M. "An information measure for classification," *Computer Journal*, Vol. 11, pp. 185–194, 1968.

[Ward 63] Ward J.H., Jr. "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association.*, Vol. 58, pp. 236–244, 1963.

[Wind 82] Windham M.P. "Cluster validity for the fuzzy c-means clustering algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 4(4), pp. 357–363, 1982.

[Zade 65] Zadeh L.A. "Fuzzy sets," *Information and Control*, Vol. 8, pp. 338–353, 1965.

[Zade 73] Zadeh L.A. *IEEE Transactions on Systems Man and Cybernetics SMC-3*, Vol. 28, 1973.