2 *Measurement and Data*

2.1 Introduction

Our aim is to discover relationships that exist in the "real world," where this may be the physical world, the business world, the scientific world, or some other conceptual domain. However, in seeking such relationships, we do not go out and look at that domain firsthand. Rather, we study data describing it. So first we need to be clear about what we mean by *data*.

Data are collected by mapping entities in the domain of interest to symbolic representation by means of some measurement procedure, which associates the value of a variable with a given property of an entity. The relationships between objects are represented by numerical relationships between variables. These numerical representations, the data items, are stored in the data set; it is these items that are the subjects of our data mining activities.

Clearly the measurement process is crucial. It underlies all subsequent data analytic and data mining activities. We discuss this process in detail in section 2.2.

We remarked in chapter 1 that the notion of "distance" between two objects is fundamental. Section 2.3 outlines distance measures between two objects, based on the vectors of measurements taken on those objects. The raw results of measurements may or may not be suitable for direct data mining. Section 2.4 briefly comments on how the data might be transformed before analysis.

We have already noted that we do not want our data mining activities simply to discover relationships that are mere artifacts of the way the data were collected. Likewise, we do not want our findings to be properties of the way the data are defined: discovering that people with the same surname often live in the same household would not be a major breakthrough. In section 2.5 we briefly introduce notions of the *schema* of data—the a priori structure imposed on the data.

No data set is perfect, and this is particularly true of large data sets. Measurement error, missing data, sampling distortion, human mistakes, and a host of other factors corrupt the data. Since data mining is concerned with detecting unsuspected patterns in data, it is very important to be aware of these imperfections—we do not want to base our conclusions on patterns that merely reflect flaws in data collection or of the recording processes. Section 2.6 discusses quality issues in the context of measurements on cases or records and individual variables or fields. Section 2.7 discusses the quality of aggregate collections of such individuals (i.e., samples).

Section 2.8 presents concluding remarks, and section 2.9 gives pointers to more detailed reading.

2.2 Types of Measurement

Measurements may be categorized in many ways. Some of the distinctions arise from the nature of the properties the measurements represent, while others arise from the use to which the measurements are put.

To illustrate, we will begin by considering how we might measure the property WEIGHT. In this discussion we will denote a property by using uppercase letters, and the variable corresponding to it (the result of the mapping to numbers induced by the measurement operation) by lowercase letters. Thus a measurement of WEIGHT yields a value of weight. For concreteness, let us imagine we have a collection of rocks.

The first thing we observe is that we can rank the rocks according to the WEIGHT property. We could do this, for example, by placing a rock on each pan of a weighing scale and seeing which way the scale tipped. On this basis, we could assign a number to each rock so that larger numbers corresponded to heavier rocks. Note that here only the ordinal properties of these numbers are relevant. The fact that one rock was assigned the number 4 and another was assigned the number 2 would not imply that the first was in any sense twice as heavy as the second. We could equally have chosen some other number, provided it was greater than 2, to represent the WEIGHT of the first rock. In general, any monotonic (order preserving) transformation of the set of numbers we assigned would provide an equally legitimate assignment. We are only concerned with the order of the rocks in terms of their WEIGHT property.

We can take the rocks example further. Suppose we find that, when we place a large rock on one pan of the weighing scale and two small rocks on the other pan, the pans balance. In some sense the WEIGHT property of the two small rocks has combined to be equal to the WEIGHT property of the large rock. It turns out (this will come as no surprise!) that we can assign numbers to the rocks in such a way that not only does the order of the numbers correspond to the order observed from the weighing scales, but the sum of the numbers assigned to the two smaller rocks equals the number assigned to the larger rock. That is, the total weight of the two smaller rocks equals the weight of the larger rock. Note that even now the assignment of numbers is not unique. Suppose we had assigned the numbers 2 and 3 to the smaller rocks, and the number 5 to the larger rock. This assignment satisfies the ordinal and additive property requirements, but so too would the assignment of 4, 6, and 10 respectively. There is still some freedom in how we define the variable weight corresponding to the WEIGHT property.

The point of this example is that *our numerical representation reflects the empirical properties of the system we are studying*. Relationships between rocks in terms of their WEIGHT property correspond to relationships between values of the measured variable weight. This representation is useful because it allows us to make inferences about the physical system by studying the numerical system. Without juggling sacks of rocks, we can see which sack contains the largest rock, which sack has the heaviest rocks on average, and so on.

The rocks example involves two empirical relationships: the order of the rocks, in terms of how they tip the scales, and their *concatenation* property—the way two rocks together balance a third. Other empirical systems might involve less than or more than two relationships. The order relationship is very common; typically, if an empirical system has only one relationship, it is an order relationship. Examples of the order relationship are provided by the SEVERITY property in medicine and the PREFERENCE property in psychology.

Of course, not even an order relationship holds with some properties, for example, the properties HAIR COLOR, RELIGION, and RESIDENCE OF PROGRAMMER, do not have a natural order. Numbers can still be used to represent "values" of the properties, (blond = 1, black = 2, brown = 3, and so on), but the only empirical relationship being represented is that the colors are different (and so are represented by different numbers). It is perhaps even more obvious here that the particular set of numbers assigned is not unique. Any set in which different numbers correspond to different values of the property will do.

Given that the assignment of numbers is not unique, we must find some way to restrict this freedom—or else problems might arise if different researchers use different assignments. The solution is to adopt some convention. For the rocks example, we would adopt a basic "value" of the property WEIGHT, corresponding to a basic value of the variable weight, and defined measured values in terms of how many copies of the basic value are required to balance them. Examples of such basic values for the WEIGHT/weight system are the gram and pound.

Types of measurement may be categorized in terms of the empirical relationships they seek to preserve. However, an important alternative is to categorize them in terms of the transformations that lead to other equally legitimate numerical representations. Thus, a numerical severity scale, in which only order matters, may be represented equally well by any numbers that preserve the order—numbers derived through a monotonic or ordinal transformation of the original ones. For this reason, such scales are termed *ordinal scales*.

In the rocks example, the only legitimate transformations involved multiplying by a constant (for example, converting from pounds to grams). Any other transformation (squaring the numbers, adding a constant, etc.) would destroy the ability of the numbers to represent the order and concatenation property by addition. (Of course, other transformations may enable the empirical relationships to be represented by different mathematical operations. For example, if we transformed the values 2, 3, and 5 in the rocks example to e^2, e^3, and e^5, we could represent the empirical relationship by multiplication: $e^2 e^3 = e^5$. However, addition is the most basic operation and is a favored choice.) Since with this type of scale multiplying by a constant leaves the ratios of values unaffected, such scales are termed *ratio scales*.

In the other case we outlined above (the hair color example) any transformation was legitimate, provided it preserved the unique identity of the different numbers—it did not matter which of two numbers was larger, and addition properties were irrelevant. Effectively, here, the numbers were simply used as labels or names; such scales are termed *nominal scales*.

There are other scale types, corresponding to different families of legitimate (or admissible) transformations. One is the *interval scale*. Here the family of legitimate transformations permit changing the units of measurement by multiplying by a constant, plus adding an arbitrary constant. Thus, not only is the unit of measurement arbitrary, but so also is the origin. Classic examples of such scales are conventional measures of temperature (Fahrenheit, Centigrade, etc.) and calendar time.

It is important to understand the basis for different kinds of measurement scale so we can be sure that any patterns discovered during mining operations are genuine. To illustrate the dangers, suppose that two groups of three patients record their pain on an ordinal scale that ranges from 1 (no pain) to 10 (severe pain); one group of patients yields scores of 1, 2, and 6, while the other yields 3, 4, and 5. The mean of the first three is $(1 + 2 + 6)/3 = 3$, while that of the second three is 4. The second group has the larger mean. However, since the scale is purely ordinal any order-preserving transformation will yield an equally legitimate numerical representation. For example, a transformation of the scale so that it ranged from 1 to 20, with (1, 2, 3, 4, 5, 6) transformed to (1, 2, 3, 4, 5, 12) would preserve the order relationships between the different levels of pain—if a patient A had worse pain than a patient B using the first scale, then patient A would also have worse pain than patient B using the second scale. Now, however, the first group of patients would have a mean score $(1 + 2 + 12)/3 = 5$, while the second group would still have a mean score 4. Thus, two equally legitimate numerical representations have led to opposite conclusions. The pattern observed using the first scale (one mean being larger than the other) was an artifact of the numerical representation adopted, and did not correspond to any true relationship among the objects (if it had, two equally legitimate representations could not have led to opposite conclusions). To avoid such problems we must be sure to only make statistical statements for which the truth value will be invariant under legitimate transformations of the measurement scales. In this example, we could make the statement that the median of the scores of the second group is larger than the median of the scores of the first group; this would remain true, whatever order-preserving transformation we applied.

Up to this point, we have focussed on measurements that provide mappings in which the relationships between numbers in the empirical system being studied correspond to relationships between numbers in a numerical system. Because the mapping serves to represent relationships in an empirical system, this type of measurement is called *representational*.

However, not all measurement procedures fit easily into this framework. In some situations, it is more natural to regard the measurement procedure as *defining* a property in question, as well as assigning a number to it. For example, the property QUALITY OF LIFE in medicine is often measured by identifying those components of human life that one regards as important, and then defining a way of combining the scores corresponding to the separate components (e.g., a weighted sum). EFFORT in software engineering is sometimes defined in a similar way, combining measures of the number of program instructions, a complexity rating, the number of internal and external documents and so forth. Measurement procedures that define a property as well as measure it are called *operational* or *nonrepresentational* procedures. The operational perspective on measurement was originally conceived in physics, around the start of the century, amid uneasiness about the reality of concepts such as atoms. The approach has gone on to have larger practical implications for the social and behavioral sciences. Since in this method the measurement procedure also defines the property, no question of legitimate transformations arises. Since there are no alternative numerical representations any statistical statements are permissible.

Example 2.1 One early attempt at measuring programming effort is given by Halstead (1977). In a given program if a is the number of unique operators, b is the number of unique operands, n is the number of total operator occurrences, and m is the total number of operand occurrences, then the programming effort is

$$e = am(n + m)\log(a + b)/2b.$$

This is a nonrepresentational measurement, since it defines programming effort, as well as providing a way to measure it.

One way of describing the distinction between representational and operational measurement is that the former is concerned with *understanding* what is going on in a system, while the latter is concerned with *predicting* what is going on. The difference between understanding (or describing) a system and predicting its behavior crops up elsewhere in this book. Of course, the two aims overlap, but the distinction is a useful one. We can construct effective and valuable predictive systems that make no reference to the mechanisms underlying the process. For instance most people successfully drive automobiles or operate video recorders, without any idea of their inner workings.

In principle, the mappings defined by the representational approach to measurement, or the numbers assigned by the operational approach, can take any values from the continuum. For example, a mapping could tell us that the length of the diagonal of a unit square is the square root of 2. However, in practice, recorded data are only approximations to such mathematical ideals. First, there is often unavoidable error in measurement (e.g., if you repeatedly measure someone's height to the nearest millimeter you will observe a distribution of values). Second, data are recorded to a finite number of decimal places. We might record the length of the diagonal of a unit square as 1.4, or 1.41, or 1.414, or 1.4142, and so on, but the measure will never be exact. Occasionally, this kind of approximation can have an impact on an analysis. The effect is most noticeable when the approximation is crude (when the data are recorded to only very few decimal places).

The above discussion provides a theoretical basis for measurement issues. However, it does not cover all descriptive measurement terms that have been introduced. Many other taxonomies for measurement scales have been described, sometimes based not on the abstract mathematical properties of the scales but rather on the sorts of data analytic techniques used to manipulate them. Examples of such alternatives include counts versus measurements; nominal, ordinal, and numerical scales; qualitative versus quantitative measurements; metrical versus categorical measurements; and grades, ranks, counted fractions, counts, amounts, and balances. In most cases it is clear what is intended by these terms. Ranks, for example, correspond to an operational assignment of integers to the particular entities in a given collection on the basis of the relative "size" of the property in question: the ranks are integers which preserve the order property.

In data mining applications (and in this text), the scale types that occur most frequently are categorical scales in which any one-to-one transformation is allowed (nominal scales), ordered categorical scales, and numerical (quantitative or real-valued) scales.

2.3 Distance Measures

Many data mining techniques (for example, nearest neighbor classification methods, cluster analysis, and multidimensional scaling methods) are based on similarity measures between objects. There are essentially two ways to obtain measures of similarity. First, they can be obtained directly from the objects. For example, a marketing survey may ask respondents to rate pairs of objects according to their similarity, or subjects in a food tasting experiment may be asked to state similarities between flavors of ice-cream. Alternatively, measures of similarity may be obtained indirectly from vectors of measurements or characteristics describing each object. In the second case it is necessary to define precisely what we mean by "similar," so that we can calculate formal similarity measures.

Instead of talking about how similar two objects are, we could talk about how dissimilar they are. Once we have a formal definition of either "similar" or "dissimilar," we can easily define the other by applying a suitable monotonically decreasing transformation. For example, if $s(i, j)$ denotes the similarity and $d(i, j)$ denotes the dissimilarity between objects i and j, possible transformations include $d(i, j) = 1 - s(i, j)$ and $d(i, j) = \sqrt{2(1 - s(i, j))}$. The term *proximity* is often used as a general term to denote either a measure of similarity or dissimilarity.

Two additional terms—*distance* and *metric*—are often used in this context. The term distance is often used informally to refer to a dissimilarity measure derived from the characteristics describing the objects—as in *Euclidean distance*, defined below. A *metric*, on the other hand, is a dissimilarity measure that satisfies three conditions:

1. $d(i, j) \geq 0$ for all i and j, and $d(i, j) = 0$ if and only if $i = j$;
2. $d(i, j) = d(j, i)$ for all i and j; and
3. $d(i, j) \leq d(i, k) + d(k, j)$ for all i, j, and k.

The third condition is called the *triangle inequality*.

Suppose we have n data objects with p real-valued measurements on each object. We denote the vector of observations for the ith object by $\mathbf{x}(i) = (x_1(i), x_2(i), \ldots, x_p(i)), 1 \leq i \leq n$, where the value of the kth variable for the ith object is $x_k(i)$. The *Euclidean distance* between the ith and jth objects is defined as

$$d_E(i, j) = \left(\sum_{k=1}^{p} (x_k(i) - x_k(j))^2 \right)^{\frac{1}{2}}. \tag{2.1}$$

This measure assumes some degree of *commensurability* between the different variables. Thus, it would be effective if each variable was a measure of length (with the number p of dimensions being 2 or 3, it would yield our standard physical measure of distance) or a measure of weight, with each variable measured using the same units. It makes less sense if the variables are noncommensurate. For example, if one variable were length and another were weight, there would be no obvious choice of units; by altering the choice of units we would change which variables were most important as far as the distance was concerned.

Since we often have to deal with data sets in which the variables are not commensurate, we must find some way to overcome the arbitrariness of the choice of units. A common strategy is to standardize the data by dividing each of the variables by its sample standard deviation, so that they are all regarded as equally important. (But note that this does not resolve the issue—treating the variables as equally important in this sense is still making an arbitrary assumption.) The standard deviation for the kth variable X_k can be estimated as

$$\hat{\sigma}_k = \left(\frac{1}{n} \sum_{i=1} (x_k(i) - \mu_k)^2 \right)^{\frac{1}{2}} \tag{2.2}$$

where μ_k is the mean for variable X_k, which (if unknown) can be estimated using the *sample mean* $\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_k(i)$. Thus, $x_k' = x_k/\hat{\sigma}_k$ removes the effect of scale as captured by $\hat{\sigma}_k$.

In addition, if we have some idea of the relative importance that should be accorded to each variable, then we can weight them (after standardization), to yield the weighted Euclidean distance measure

$$d_{WE}(i,j) = \left(\sum_{k=1}^{p} w_k \left(x_k(i) - x_k(j)\right)^2 \right)^{\frac{1}{2}} . \tag{2.3}$$

The Euclidean and weighted Euclidean distances are both additive, in the sense that the variables contribute independently to the measure of distance. This property may not always be appropriate. To take an extreme case, suppose that we are measuring the heights and diameters of a number of cups. Using commensurate units, we could define similarities between the cups in terms of these two measurements. Now suppose that we measured the height of each cup 100 times, and the diameter only once (so that for any given cup we have 101 variables, 100 of which have almost identical values). If we combined these measurements in a standard Euclidean distance calculation, the height would dominate the apparent similarity between the cups. However, 99 of the height measurements do not contribute anything to what we really want to measure; they are very highly correlated (indeed, perfectly, apart from measurement error) with the first height measurement. To eliminate such redundancy we need a data-driven method. One approach is to standardize the data, not just in the direction of each variable, as with weighted Euclidean distance, but also taking into account the *covariances* between the variables.

Example 2.2 Consider two variables X and Y, and assume we have n objects, with X taking the values $x(1), \ldots, x(n)$ and Y taking the values $y(1), \ldots, y(n)$.

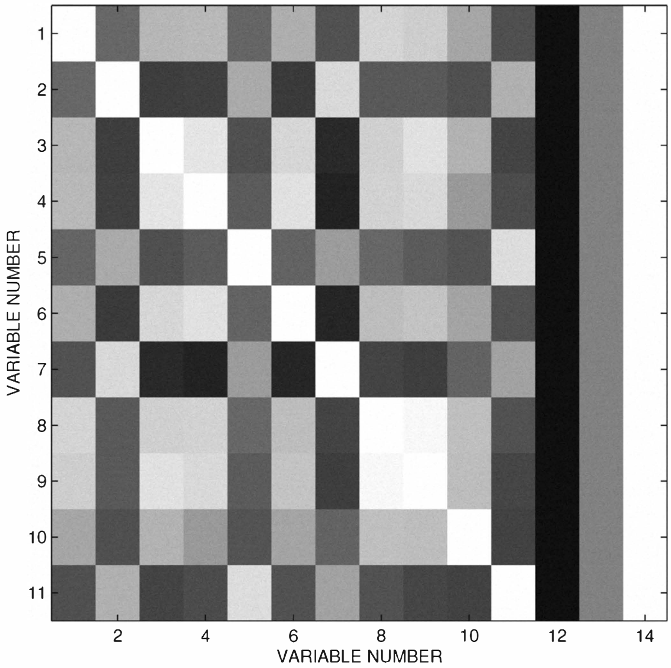

Figure 2.1 A sample correlation matrix plotted as a pixel image. White corresponds to +1 and black to -1. The three rightmost columns contain values of -1, 0, and +1 (respectively) to provide a reference for pixel intensities. The remaining 11 × 11 pixels represent the 11 × 11 correlation matrix. The data come from a well-known data set in the regression research literature, in which each data vector is a suburb of Boston and each variable represents a certain general characteristic of a suburb. The variable names are (1) per-capita crime rate, (2) proportion of area zoned for large residential lots, (3) proportion of non-retail business acres, (4) nitric oxide concentration, (5) average number of rooms perdwelling, (6) proportion of pre-1940 homes, (7) distance to retail centers index, (8) accessibility to highways index, (9) property tax rate, (10) pupil-to-teacher ratio, and (11) median value of owner-occupied homes.

Then the *sample covariance* between X and Y is defined as

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y}), \tag{2.4}$$

where $\bar{x}$ is the sample mean of the X values and $\bar{y}$ is the sample mean of the Y values.

The covariance is a measure of how X and Y vary together: it will have a large positive value if large values of X tend to be associated with large values of Y and small values of X with small values of Y. If large values of X tend to be associated with small values of Y, it will take a negative value.

More generally, with p variables we can construct a $p \times p$ matrix of covariances, in which the element (k,l) is the covariance between the kth and lth variables. From the definition of covariance above, we can see that such a matrix (a covariance matrix) must be symmetric.

The value of the covariance depends on the ranges of X and Y. This dependence can be removed by standardizing, dividing the values of X by their standard deviation and the values of Y by their standard deviation. The result is the *sample correlation coefficient* $\rho(X,Y)$ between X and Y:

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y})}{\left(\sum_{i=1}^{n}(x(i)-\bar{x})^2\sum_{i=1}^{n}(y(i)-\bar{y})^2\right)^{\frac{1}{2}}}. \tag{2.5}$$

In the same way that a covariance matrix can be formed if there are p variables, a $p \times p$ correlation matrix can be formed in the same manner. Figure 2.1 shows a pixel image of a correlation matrices for an 11-dimensional data set on housing-related variables across different Boston suburbs. From the matrix we can clearly see structure in terms of how different variables are correlated. For example, variables 3 and 4 (relating to business acreage and presence of nitrous oxide) are each highly negatively correlated with variable 2 (the percent of large residential lots in the suburb) and positively correlated with each other. Variable 5 (average number of rooms) is positively correlated with variable 11 (median home value) (i.e., larger houses tend to be more valuable). Variables 8 and 9 (tax rates and highway accessibility) are also highly correlated.

Note that covariance and correlation capture *linear dependencies* between variables (they are more accurately termed *linear covariance* and *linear correlation*). Consider data points that are uniformly distributed around a circle in two dimensions (X and Y), centered at the origin. The variables are clearly *dependent*, but in a nonlinear manner and they will have zero linear correlation. Thus, independence implies a lack of correlation, but the reverse is not generally true. We will have more to say about independence in chapter 4.

Recall again our coffee cup example with 100 measurements of height and one measurement of width. We can discount the effect of the 100 correlated variables by incorporating the covariance matrix in our definition of distance. This leads to the Mahalanobis distance between two p-dimensional measurements $\mathbf{x}(i)$ and $\mathbf{x}(j)$, defined as:

$$d_{MH}(i,j) = \left(\left(\mathbf{x}(i) - \mathbf{x}(j) \right)^T \Sigma^{-1} \left(\mathbf{x}(i) - \mathbf{x}(j) \right) \right)^{\frac{1}{2}} \tag{2.6}$$

where T represents the transpose, Σ is the $p \times p$ sample covariance matrix, and Σ^{-1} standardizes the data relative to Σ. Note that although we have been thinking about our p-dimensional measurement vectors $\mathbf{x}(i)$ as *rows* in our data matrix, the convention in matrix algebra is to treat these as $p \times 1$ *column vectors* (we can still visualize our data matrix as being an $n \times p$ matrix). Entry (k, l) of Σ is defined between variable X_k and X_l, as in equation 2.5. Thus, we have a $p \times 1$ vector transposed (to give a $1 \times p$ vector), multiplied by the $p \times p$ matrix Σ^{-1}, multiplied by a $p \times 1$ vector, yielding a scalar distance. Of course, other matrices could be used in place of Σ. Indeed, the statistical frameworks of canonical variates analysis and discriminant analysis use the average of the covariance matrices of different groups of cases.

The Euclidean metric can also be generalized in other ways. For example, one obvious generalization is to the Minkowski or L_λ metric:

$$\left(\sum_{k=1}^{p} (x_k(i) - x_k(j))^\lambda) \right)^{\frac{1}{\lambda}}, \tag{2.7}$$

where $\lambda \geq 1$. Using this, the Euclidean distance is the special case of $\lambda = 2$. The L_1 metric (also called the *Manhattan* or *city-block metric*) can be defined as

$$\sum_{k=1}^{p} \mid x_k(i) - x_k(j)) \mid . \tag{2.8}$$

The case $\lambda \to \infty$ yields the L_∞ metric

$$\max_k |x_k(i) - x_k(j)| .$$

There is a huge number of other metrics for quantitative measurements, so the problem is not so much defining one but rather deciding which is most appropriate for a particular situation.

For multivariate *binary data* we can count the number of variables on which two objects take the same or take different values. Consider table 2.1, in

	$j = 1$	$j = 0$
$i = 1$	$n_{1,1}$	$n_{1,0}$
$i = 0$	$n_{0,1}$	$n_{0,0}$

Table 2.1 A cross-classification of two binary variables.

which all p variables defined for objects i and j take values in $\{0, 1\}$; the entry $n_{1,1}$ in the box for $i = 1$ and $j = 1$ denotes that there are $n_{1,1}$ variables such that i and j both have value 1.

With binary data, rather than measuring the dissimilarities between objects, we often measure the similarities. Perhaps the most obvious measure of similarity is the simple matching coefficient, defined as

$$\frac{n_{1,1} + n_{0,0}}{n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0}}, \tag{2.9}$$

the proportion of the variables on which the objects have the same value, where $n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0} = p$, the total number of variables. Sometimes, however, it is inappropriate to include the (0,0) cell (or the (1,1) cell, depending on the meaning of 0 and 1). For example, if the variables are scores of the presence (1) or absence (0) of certain properties, we may not care about all the irrelevant properties had by neither object. (For instance, in vector representations of text documents it may be not be relevant that two documents *do not* contain thousands of specific terms). This consideration leads to a modification of the matching coefficient, the *Jaccard coefficient*, defined as

$$\frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}}. \tag{2.10}$$

The *Dice coefficient* extends this argument. If (0,0) matches are irrelevant, then (0,1) and (1,0) mismatches should lie between (1,1) matches and (0,0) matches in terms of relevance. For this reason the number of (0,1) and (1,0) mismatches should be multiplied by a half. This yields $2n_{1,1}/(2n_{1,1}+n_{1,0}+n_{0,1})$. As with quantitative data, there are many different measures for multivariate binary data—again the problem is not so much defining such measures but choosing one that possesses properties that are desirable for the problem at hand.

For categorical data in which the variables have more than two categories, we can score 1 for variables on which the two objects agree and 0 otherwise, expressing the sum of these as a fraction of the possible total p. If we know about the categories, we might be able to define a matrix giving values for the different kinds of disagreement.

Additive distance measures can be readily adapted to deal with mixed data types (e.g., some binary variables, some categorical, and some quantitative) since we can add the contributions from each variable. Of course, the question of relative standardization still arises.

2.4 Transforming Data

Sometimes raw data are not in the most convenient form and it can be advantageous to modify them prior to analysis. Note that there is a duality between the form of the model and the nature of the data. For example, if we speculate that a variable Y is a function of the square of a variable X, then we either could try to find a suitable function of X^2, or we could square X first, to $U = X^2$, and fit a function to U. The equivalence of the two approaches is obvious in this simple example, but sometimes one or other can be much more straightforward.

Example 2.3 Clearly variable V_1 in figure 2.2 is nonlinearly related to variable V_2. However, if we work with the reciprocal of V_2, that is, $V_3 = 1/V_2$, we obtain the linear relationship shown in figure 2.3.

Sometimes, especially if we are concerned with formal statistical inferences in which the shape of a distribution is important (as when running statistical tests, or calculating confidence intervals), we might want to transform the data so that they approximate the requisite distribution more closely. For example, it is common to take logarithms of positively skewed data (such as bank account sizes or incomes) to make the distribution more symmetric (so that it more closely approximates a normal distribution, on which many inferential procedures are based).

Example 2.4 In figure 2.4 not only are the two variables nonlinearly related, but the variance of V_2 increases as V_1 increases. Sometimes inferences are based on an assumption that the variance remains constant (for example, in the basic model for regression analysis). In the case of these (artificial) data, a square root transformation of V_2 yields the transformed data shown in figure 2.5.

Since our fundamental aim in data mining is exploration, we must be prepared to contemplate and search for the unsuspected. Certain transformations of the data may lead to the discovery of structures that were not at all

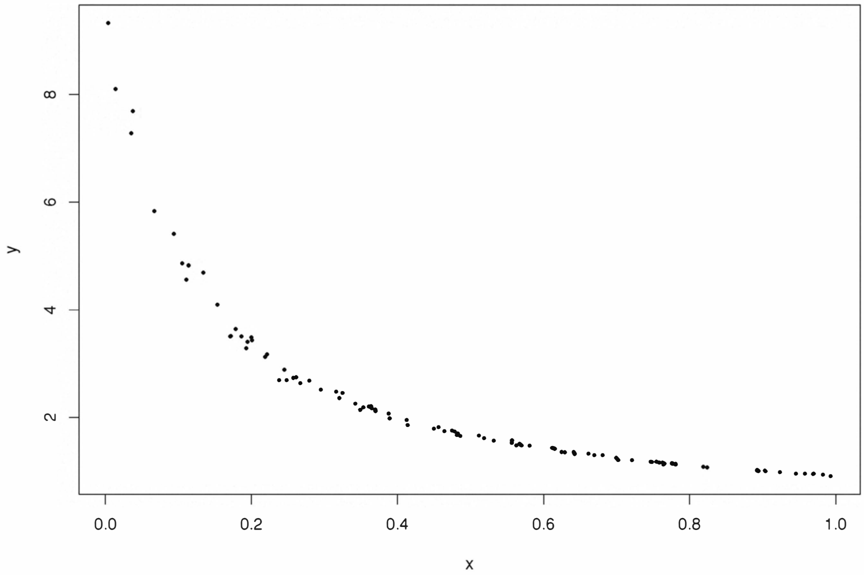

Figure 2.2 A simple nonlinear relationship between variable V_1 and V_2. (In these and subsequent figures V_1 and V_2 are on the X and Y axes respectively).

obvious on the original scale. On the other hand, it is possible to go too far in this direction: we must be wary of creating structures that are simply artifacts of a peculiar transformation of the data (see the example of the ordinal pain scale in section 2.2). Presumably, when this happens in a data mining context, the domain expert responsible for evaluating an apparent discovery will soon reject the structure.

Note also that in transforming data we may sacrifice the way it represents the underlying objects. As described in section 2.2 the standard mapping of rocks to weights maps a physical concatenation operation to addition. If we nonlinearly transform the numbers representing the weights, using logarithms or taking square roots for example, the physical concatenation operation is no longer preserved. Caution—and common sense—must be exercised.

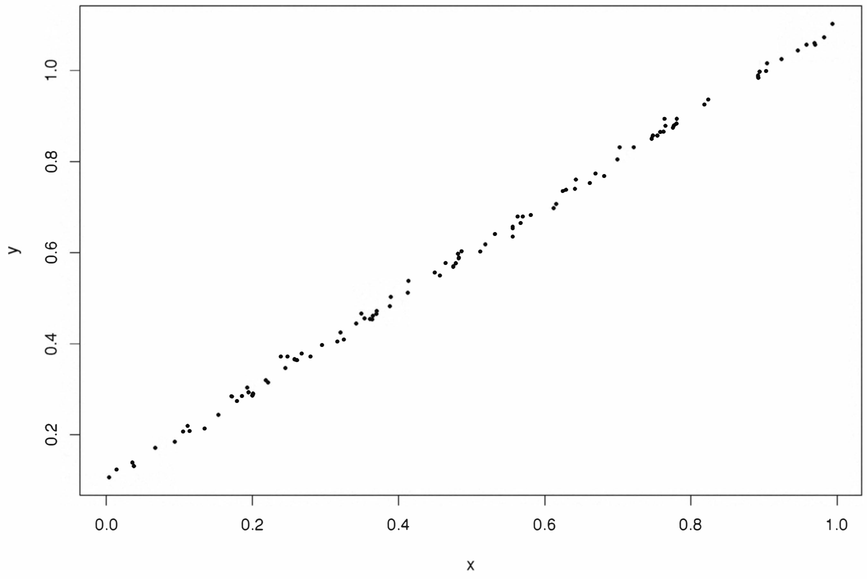

Figure 2.3 The data of figure 2.2 after the simple transformation of V_2 to $1/V_2$.

Common data transformations include taking square roots, reciprocals, logarithms, and raising variables to positive integral powers. For data expressed as proportions, the *logit transformation*, $f(p) = \frac{p}{1-p}$, is often used.

Some classes of techniques assume that the variables are categorical—that only a few (ordered) responses are possible. At an extreme, some techniques assume that responses are binary, with only two possible outcome categories. Of course continuous variables (those that can, at least in principle, take any value within a given interval) can be split at various thresholds to reduce them to categories. This sacrifices information, with the information loss increasing as the number of categories is reduced, but in practice this loss can be quite small.

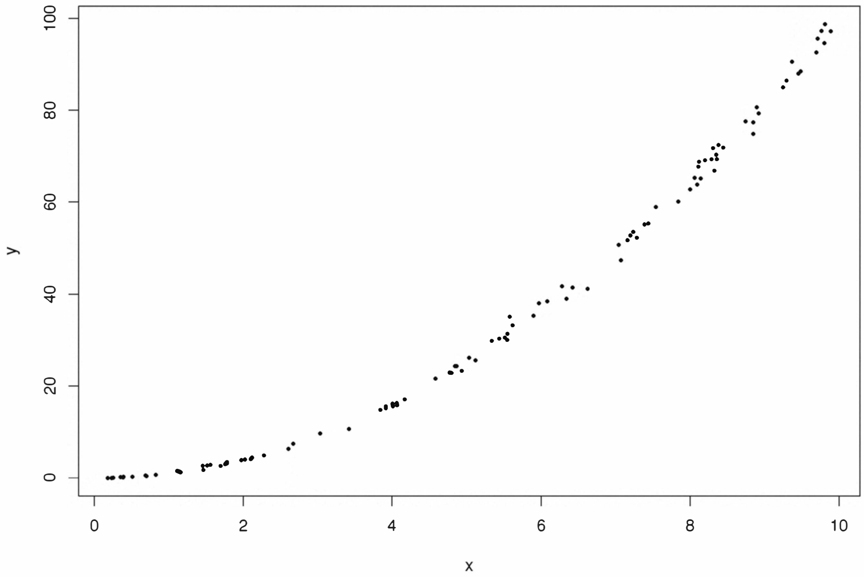

Figure 2.4 Another simple nonlinear relationship. Here the variance of V_2 increases as V_1 increases.

2.5 The Form of Data

We mentioned in chapter 1 that data sets come in different forms; these forms are known as *schemas*. The simplest form of data (and the only form we have discussed in any detail) is a set of vector measurements on objects $o(1), \ldots, o(n)$. For each object we have measurements of p variables $X_1, \ldots, X_p$. Thus, the data can be viewed as a matrix with n rows and p columns. We refer to this standard form of data as a *data matrix*, or simply *standard data*. We can also refer to the data set as a *table*.

Often there are several types of objects we wish to analyze. For example, in a payroll database, we might have data both about employees, with variables name, department-name, age, and salary, and about departments with variables department-name, budget and manager. These data matrices are con-

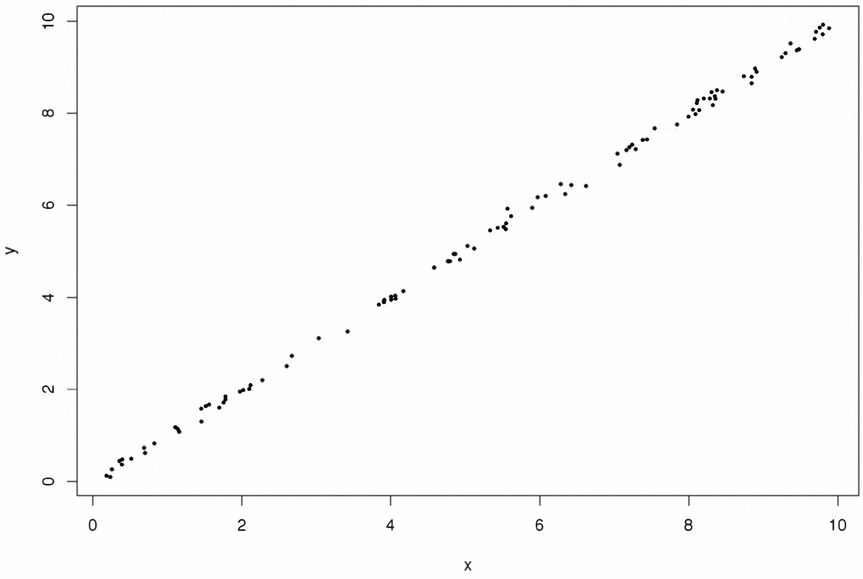

Figure 2.5 The data of figure 2.4 after a simple square root transformation of V_2. Now the variance of V_2 is relatively constant as V_1 increases.

nected to each other by the occurrence of the same (categorical) values in the department-name fields and in the fields name and manager. Data sets consisting of several such matrices or tables are called *multirelational data*.

In many cases multirelational data can be mapped to a single data matrix or table. For example, we could join the two data tables using the values of the variable department-name. This would give us a data matrix with the variables name, department-name, age, salary, budget (of the department), and manager (of the department). The possibility of such a transformation seems to suggest that there is no need to consider multirelational structures at all since in principle we could represent the data in one large table or matrix. However, this way of joining the data sets is not the only possibility: we could also create a table with as many rows as there are departments (this would be useful if we were interested in getting information about the departments, e.g., determining whether there was a dependence between the budget of a department and the age of the manager). Generally no single table best captures all the information in a multirelational data set. More important, from the point of view of efficiency in storage and data access, "flattening" multirelational data to form a single large table may involve the needless replication of numerous values.

Some data sets do not fit well into the matrix or table form. A typical example is a time series, in which consecutive values correspond to measurements taken at consecutive times, (e.g., measurements of signal strength in a waveform, or of responses of a patient at a series of times after receiving medical treatment). We can represent a time series using two variables, one for time and one for the measurement value at that time. This is actually the most natural representation to use for storing the time series in a database. However, representing the data as a two-variable matrix does not take into account the ordered aspect of the data. In analyzing such data, it is important to recognize that a natural order does exist. It is common, for example, to find that neighboring observations are more closely related (more highly correlated) than distant observations. Failure to account for this factor could lead to a poor model.

A *string* is a sequence of symbols from some finite alphabet. A sequence of values from a categorical variable is a string, and so is standard English text, in which the values are alphanumeric characters, spaces, and punctuation marks. Protein and DNA/RNA sequences are other examples. Here the letters are individual proteins (note that a string representation of a protein sequence is a 2-dimensional view of a 3-dimensional structure). A string is another data type that is ordered and for which the standard matrix form is not necessarily suitable.

A related ordered data type is the *event-sequence*. Given a finite alphabet of categorical event types, an event-sequence is a sequence of pairs of the form $\{event,\ occurrence\ time\}$. This is quite similar to a string, but here each item in the sequence is tagged with an occurrence time. An example of an event-sequence is a telecommunication alarm log, which includes a time of occurrence for each alarm. More complicated event-sequences include transaction data (such as records of retail or financial transactions), in which each transaction is time-stamped and the events themselves can be relatively complex (e.g., listing all purchases along with prices, department names, and so forth). Furthermore, there is no reason to restrict the concept of event sequences to categorical data; for example we could extend it to real-valued events occurring asynchronously, such as data from animal behavioral experiments or bursts of energy from objects in deep space.

Of course, order may be imposed simply for logistic convenience: placing patient records in alphabetical order by name assists retrieval, but the fact that Jones precedes Smith is unlikely to have any impact on most data mining activities. Still, care must always be exercised in data mining. For example, records of members of the same family (with the same last name) would probably occur near one another in a data set, and they may have related properties. (We may find that a contagious disease tends to infect groups of people whose names are close together in the data set.)

Ordered data are spread along a unidimensional continuum (per individual variable), but other data often lie in higher dimensions. *Spatial*, *geographic*, or *image* data are located in two and three dimensional spaces. It is important to recognize that some of the variables are part of the defining data schema in these examples: that is, some of the variables merely specify the coordinates of observations in the spaces. The discovery that geographical data lies in a two-dimensional continuum would not be very profound.

A *hierarchical* structure is a more complex data schema. For example, a data set of children might be grouped into classes, which are grouped into years, which are grouped into schools, which are grouped into counties, and so on. This structure is obvious in a multirelational representation of the data, but can be harder to see in a single table. Ignoring this structure in data analysis can be very misleading. Research on statistical models for such multi-level data has been particularly active in recent years. A special case of hierarchical structures arises when responses to certain items on a questionnaire are contingent on answers to other questions: for instance the relevance of the question "Have you had a hysterectomy?" depends on the answer to the question "Are you male or female?"

To summarize, in any data mining application it is crucial to be aware of the schema of the data. Without such awareness, it is easy to miss important patterns in the data or, perhaps worse, to rediscover patterns that are part of the fundamental design of the data. In addition, we must be particularly careful about data schemas when sampling, as we will discuss in more detail in chapter 4.

2.6 Data Quality for Individual Measurements

The effectiveness of a data mining exercise depends critically on the quality of the data. In computing this idea is expressed in the familiar acronym

GIGO—*Garbage In, Garbage Out*. Since data mining involves secondary analysis of large data sets, the dangers are multiplied. It is quite possible that the most interesting patterns we discover during a data mining exercise will have resulted from measurement inaccuracies, distorted samples or some other unsuspected difference between the reality of the data and our perception of it.

It is convenient to characterize data quality in two ways: the quality of the individual records and fields, and the overall quality of the collection of data. We deal with each of these in turn.

No measurement procedure is without the risk of error. The sources of error are infinite, ranging from human carelessness, and instrumentation failure, to inadequate definition of what it is that we are measuring. Measuring instruments can lead to errors in two ways: they can be inaccurate or they can be imprecise. This distinction is important, since different strategies are required for dealing with the different kinds of errors.

A *precise* measurement procedure is one that has small variability (often measured by its variance). Using a precise process, repeated measurements on the same object under the same conditions will yield very similar values. Sometimes the word precision is taken to connote a large number of digits in a given recording. We do not adopt this interpretation, since such "precision" can all too easily be spurious, as anyone familiar with modern data analysis packages (which sometimes give results of calculations to eight or more decimal places) will know.

An *accurate* measurement procedure, in contrast, not only possesses small variability, but also yields results close to what we think of as the true value. A measurement procedure may yield precise but inaccurate measurements. For example repeated measurements of someone's height may be precise, but if these were made while the subject was wearing shoes, the result would be inaccurate. In statistical terms, the difference between the mean of repeated measurements and the true value is the *bias* of a measurement procedure. Accurate procedures have small bias as well as small variance.

Note that the concept of a "true value" is integral to the concept of accuracy. But this concept is rather more slippery than it might at first appear. Take a person's height, for example. Not only does it vary slightly from moment to moment—as the person breathes and as his or her heart beats—but it also varies over the course of a day (gravity pulls us down). Astronauts returning from extended tours in space, are significantly taller than when they set off (though they soon revert to their former height). Mosteller (1968) remarked that "Today some scientists believe that true values do not exist separately from the measuring process to be used, and in much of social science this view can be amply supported. The issue is not limited to social science; in physics, complications arise from the different methods of measuring microscopic and macroscopic quantities such as lengths. On the other hand, because it suggests ways of improving measurement methods, the concept of true value is useful; since some methods come much nearer to being ideal than others, the better ones can provide substitutes for true values."

Other terms are also used to express these concepts. The *reliability* of a measurement procedure is the same as its precision. The former term is typically used in the social sciences whereas the latter is used in the physical sciences. This use of two different names for the same concept is not as unreasonable as it might seem, since the process of determining reliability is quite different from that of determining precision. In measuring the precision of an instrument, we can use that instrument repeatedly: assuming that during the course of the repeated applications the circumstances will not change much. Furthermore, we assume that the measurement process itself will not influence the system being measured. (Of course, there is a grey area here: as Mosteller noted, very small or delicate phenomena may indeed be perturbed by the measurement procedure.) In the social and behavioral sciences, however, such perturbation is almost inevitable: for instance a test asking a subject to memorize a list of words could not usefully be applied twice in quick succession. Effective retesting requires more subtle techniques, such as alternative-form testing (in which two alternative forms of the measuring instrument are used), split-halves testing (in which the items on a single test are split into two groups), and methods that assess internal consistency (giving the expected correlation of one test with another version that contains the same number of items).

Earlier we described two factors contributing to the inaccuracy of a measurement. One was basic precision—the extent to which repeated measurements of the same object gave similar results. The other was the extent to which the distribution of measurements was centered on the true value. While precision corresponds to reliability, the other component corresponds to *validity*. Validity is the extent to which a measurement procedure measures what it is supposed to measure. In many areas—including software engineering and economics—careful thought is required to construct metrics that tap the underlying concepts we want to measure. If a measurement procedure has poor validity, any conclusions we draw from it about the target phenomena will be at best dubious and at worst positively misleading. This is especially true in feedback situations, where action is taken on the basis of measurements. If the measurements are not tapping the phenomenon of interest, such actions could lead the system to depart even further from its target state.

2.7 Data Quality for Collections of Data

In addition to the quality of individual observations, we need to consider the quality of collections of observations. Much of statistics and data mining is concerned with inference from a sample to a population, that is, how, on the basis of examining just a fraction of the objects in a collection, one can infer things about the entire population. Statisticians use the term *parameter* to refer to descriptive summaries of populations or distributions of objects (more generally, of course, a parameter is a value that indexes a family of mathematical functions). Values computed from a sample of objects are called *statistics,* and appropriately chosen statistics can be used as estimates of parameters. Thus, for example, we can use the average of a sample as an estimate of the mean (parameter) of an entire population or distribution.

Such estimates are useful only if they are accurate. As we have just noted, inaccuracies can occur in two ways. Estimates from different samples might vary greatly, so that they are unreliable: using a different sample might have led to a very different estimate. Or the estimates might be biased, tending to be too large or too small. In general, the precision of an estimate (the extent to which it would vary from sample to sample) increases with increasing sample size; as resources permit, we can reduce this uncertainty to an acceptable value. Bias, on the other hand, is not so easily diminished.

Some estimates are intrinsically biased, but do not cause a problem because the bias decreases with increasing sample size. Of more significance in data mining are biases arising from an inappropriate sample. If we wanted to calculate the average weight of people living in New York, it would obviously be inadvisable to restrict our sample to women. If we did this, we would probably underestimate the average. Clearly, in this case, the population from which our sample is drawn (women in New York) is not the population to which we wish to generalize (everyone in New York). Our sampling frame, the list of people from which we will draw our sample, does not match the population about which we want to make an inference. This is a simple example—we were able to clearly identify the population from which the sample was drawn (women in New York). Difficulties arise when it is less obvious what the effect of the incorrect sampling frame will be. Suppose, for example, that we drew our sample from people working in offices. Would this lead to biased estimates? Maybe the sexes are disproportionately represented in offices. Maybe office workers have a tendency to be heavier than average because of their sedentary occupation. There are many reasons why such a sample might not be representative of the population we aim to study. The concept of representativeness is key to the ability to make valid inferences, as is the concept of a random sample. We discuss the need for random samples, as well as strategies for drawing such samples, in chapter 4.

Because we often have no control over the way the data are collected, quality issues are particularly important in data. Our data set may be a distorted sample of the population we wish to describe. If we know the nature of this distortion then we might be able to allow for it in our inferences, but in general this is not the case and inferences must be made with care. The terms *opportunity sample* and *convenience sample* are sometimes used to describe samples that are not properly drawn from the population of interest. The sample of office workers above would be a convenience sample—it is much more convenient to sample from them than to sample from the whole population of New York. Distortions of a sample can occur for many reasons, but the risk is especially grave when humans are involved. The effects can be subtle and unexpected: for instance, in large samples, the distribution of stated ages tends to cluster around integers ending with 0 or 5—just the sort of pattern that data mining would detect as potentially interesting. Interesting it may be, but will probably be of no value in our analysis.

A different kind of distortion occurs when customers are selected through a chain of selection steps. With bank loans, for example, an initial population of potential customers is contacted (some reply and some do not), those who reply are assessed for creditworthiness (some receive high scores and some do not), those with high scores are offered a loan (some accept and some do not), those who take out a loan are followed up (some are good customers, paying the installments on time, and others are not), and so on. A sample drawn at any particular stage would give a distorted perspective on the population at an earlier stage.

In this example of candidates for bank loans, the selection criteria at each step are clearly and explicitly stated but, as noted above, this is not always the case. For example, in clinical trials samples of patients are selected from across the country, having been exposed to different diagnostic practices and perhaps different previous treatments in different primary care facilities.

Here the notion of taking a "random sample from a well-defined population" makes no sense. This problem is compounded by the imposition of inclusion/exclusion criteria: perhaps the patients must be male, aged between 18 and 50, with a primary diagnosis of the disease in question made no longer than two years ago, and so on. (It is hardly surprising in this context, that the sizes of effects recorded in clinical trials are typically larger than those found when the treatments are applied more widely. On the other hand it is reassuring that the *directions* of the effects do normally generalize in this way.)

In addition to sample distortion arising from a mismatch between the sample population and the population of interest other kinds of distortion arise. The aim of many data mining exercises is to make some prediction of what will happen in the future. In such cases it is important to remember that populations are not static. For instance the nature of a customers shopping at a certain store will change over time, perhaps because of changes in the social culture of the surrounding neighborhood, or in response to a marketing initiative, or for many other reasons. Much work on predictive methods has failed to take account of such *population drift*. Typically, the future performance of such methods is assessed using data collected at the same time as the data used to build the model—implicitly assuming that the distribution of objects used to construct the model is the same as that of future objects. Ideally, a more sophisticated model is required that can allow for evolution over time. In principle, population drift can be modeled, but in practice this may not be easy.

An awareness of the risks of using distorted samples is vital to valid data mining, but not all data sets are samples from the population of interest. Often the data set comprises the entire population, but is so large that we wish to work with a sample from it. We can formulate valid descriptions of the population represented in such a data set, to any degree of accuracy, provided the sample is properly chosen. Of course, technical difficulties may arise, as we discuss in more detail in chapter 4, when working with data sets that have complex structures and that might be dispersed over many different databases. In chapter 4, we explain how to draw samples from a data set in such a way that we can make accurate inferences about the overall population of values in the data set, but we restrict our discussion to the cases in which the actual drawing of a sample is straightforward, once we know which cases should be included.

Distortion of samples can be viewed as a special case of incomplete data, one in which entire records are missing from what would otherwise be a representative sample. Data can also be missing in other ways. In particular, individual fields may be missing from records. In some ways this is not as serious as the situation described above. (At least here, one can see that the data are missing!) Still, significant problems may arise from incomplete data. The fundamental question is "Why are the data missing?" Was there information in the missing data that is not present in the data that have been recorded? If so, inferences based on the observed data are likely to be biased. In any incomplete data problem, it is crucial to be clear about the objectives of the analysis. In particular, if the aim is to make an inference only about the cases that have complete records, inferences based only on the complete cases is entirely valid.

Outliers or anomalous observations represent another, quite different aspect of data quality. In many situations the objective of the data mining exercise is to detect anomalies: in fraud detection and fault detection those records that differ from the majority are precisely the ones that are of interest. In such cases we would use a pattern detection process (see chapters 6 and 13). On the other hand, if the aim is model building—constructing a global model to aid understanding of, or prediction from, the data—outliers may simply obscure the main points of the model. In this case we might want to identify and remove them before building our model.

When observing only one variable, we can detect outliers simply by plotting the data—as a histogram, for example. Points that are far from the others will lie out in the tails. However, the situation becomes more interesting—and challenging—when multiple variables are involved. In this case, it is possible that each variable for a particular record has perfectly normal values, but the overall pattern of scores is abnormal. Consider the distribution of points shown in figure 2.6. Clearly there is an unusual point here, one that would immediately arouse suspicion if such a distribution were observed in practice. But the point stands out only because we produced the two dimensional plot. A one dimensional examination of the data would indicate nothing unusual at all about the point in question.

Furthermore, there may be highly unusual cases whose abnormality becomes apparent only when large numbers of variables are examined simultaneously. In such cases, a computer is essential to detection.

Every large data set includes suspect data. Rather than promoting relief, a large data set that appears untarnished by incompleteness, distortion, measurement error, or other problems should invite suspicion. Only when we recognize and understand the inadequacies of the data can we take steps to alleviate their impact. Only then can we be sure that the discovered struc-

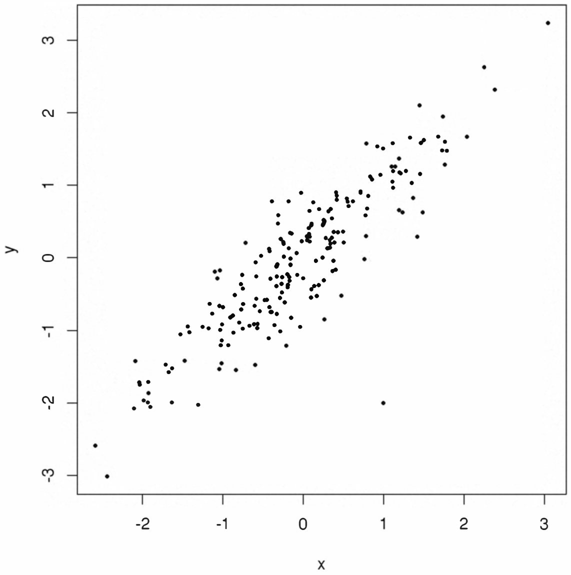

Figure 2.6 A plot of 200 points from highly positively correlated bivariate data (from a bivariate normal distribution), with a single easily identifiable outlier.

tures and patterns reflect what is really going on in the world. Since data miners rarely have control over the data collection processes, an awareness of the dangers that can arise from poor data is crucial. Hunter (1980) stated the risks succinctly:

> Data of a poor quality are a pollutant of clear thinking and rational decisionmaking. Biased data, and the relationships derived from such data, can have serious consequences in the writing of laws and regulations.

And, we might add, they can have serious consequences in developing scientific theories, in unearthing commercially valuable information, in improving quality of life, and so on.

2.8 Conclusion

In this chapter we have restricted our discussion to numeric data. However, other kinds of data also arise. For example, text data is an important class of non-numeric data, which we discuss further in chapter 14. Sometimes the definition of an individual data item (and hence whether it is numeric or non-numeric) depends on the objectives of our analysis: in economic contexts, in which hundreds of thousands of time series are stored in databases, the data items might be entire time series, rather than the individual numbers within those series.

Even with non-numeric data, numeric data analysis plays a fundamental role. Often non-numeric data items, or the relationships between them, are reduced to numeric descriptions, which are subject to standard methods of analysis. For example, in text processing we might measure the number of times a particular word occurs in each document, or the probability that certain pairs of words appear in documents.

2.9 Further Reading

The magnum opus on representational measurement theory is the three volume work of Krantz et al. (1971), Suppes et al. (1989), and Luce et al. (1990). Roberts (1979) also outlines this approach. Dawes and Smith (1985) and Michell (1986, 1990) describe alternative approaches, including the operational approach. Hand (1996) explores the relationship between measurement theory and statistics. Some authors place their discussions of software metrics in a formal measurement theoretical context—see, for example, Fenton (1991). Anderberg (1973) includes a good discussion of similarity and dissimilarity measures.

Issues of reliability and validity are often discussed in treatments of measurement issues in the social, behavioral, and medical sciences—see, for example, Dunn (1989) and Streiner and Norman (1995). Carmines and Zeller (1979) also discuss such issues. A key work on incomplete data and different types of missing data mechanisms is Little and Rubin (1987). The bank loan example of distorted samples is taken from Hand, McConway, and Stanghellini (1997). Goldstein (1995) is a key work on multilevel modeling.