Chapter 2: The Nature of the World and Its Impact on Data Preparation

Overview

Data is explored to discover knowledge about the data, and ultimately, about the world. There are, however, some deep assumptions underlying this idea. It presupposes that knowledge is discoverable. In the case of using data mining as a tool for discovering knowledge, it presupposes that knowledge is discoverable in a collection of data. A reasonable assumption is that the discovered knowledge is to be usefully applied to the real world. It is therefore also assumed that the data to be mined does in fact have some persistent relationship to the world from which it was drawn. It is also assumed that any relationships that happen to be present in the assembled data can be meaningfully related back to real-world phenomena.

These crucial assumptions underpinning data exploration and data mining are usually unstated. They do, however, have a major impact on the actual process of mining data, and they affect how data is prepared for mining. Any analysis of data that is made in the hope of either understanding or influencing the world makes these assumptions. To better understand why data is manipulated in the way that it is during data preparation, and to understand the effects of the manipulations, we need to closely examine the assumptions, the nature of what data is measuring, and define the terms “data,” “information,” and “knowledge.”

Chapter 1 provided an overall framework for data exploration and put all of the components into perspective. This chapter will focus on the nature of the connection between the experiential world and the measurements used to describe it, how those measurements are turned into data, and how data is organized into data sets. Having created an organized representation of part of the world in a data set, we will also look at the nature and reasons for some of the adjustments, alterations, and reformatting that have to be applied to the data sets to prepare them for mining.

2.1 Measuring the World

The world is a place of unbelievable complexity. No matter how closely we look at some facet of the world, there is an infinite depth of detail. Yet our brains and minds construct meaningful (for us) simplicities from the stunning complexity that surrounds us. Using these simplicities we make representations of the world that we find useful, like lunch and banks. And using these simplicities, we can collect and record impressions about various facets of them, which we call data. It is this data that we then explore, at least with data mining, to understand something about the reality of the world—to discover information.

The data itself from which information is to be discovered, however rich and copious, is but a pale reflection of the real world. It doesn't matter how much care is taken in examining the world and collecting data about it, reality is always more fluid, rich, and complex than any human can comprehend. Data never provides more than a pale and hazy shadow, a murky outline, of the true workings of the world. And yet this gossamer wisp is just enough for us to grasp at the edges of understanding. We may imagine that we control and manipulate the firm reality, but it is no more than a shadow of reality that is in our grasp. Understanding this, and understanding too the way that data connects to the world, is crucial for any data explorer. However powerful the exploring tools, or aggressive the explorer, nothing can be discovered that is beyond the limits of the data itself.

2.1.1 Objects

This is not a philosophical treatise, and I will leave discussing the true nature of the world to philosophers. The world exists in a way that humans generally agree on. It consists of *objects* that we can identify, such as cars, trees, cost-of-living adjustments, cartons of milk, beams of light, gross national products, beauty, truth, and justice. For data exploration through data mining it is these objects that form the basic material of the world to be explored. These objects actually comprise the fundamental underpinning, or the interface, that connects the activities of mining to the real world. Data mining explores the relationships that exist between these objects.

The precise definition of objects is another philosophical issue that need not concern miners. It is almost, if not actually, impossible to define what an object "really" is. It is also difficult or impossible to define the limits of an object precisely and unambiguously, since the world at very fine scale seems to appear as "shades of gray." The miner takes a pragmatic view of the objects in the world, finding it unnecessary to define the actual objects and instead regarding an object as a collection of *features* about which measurements can be taken.

A car, for instance, is accepted by the miner as a defined object. The car possesses certain measurable features, such as the number of wheels, number of seats, color, weight, number of cylinders, fuel consumption, and a host of others. These measurements are not necessarily fixed; for instance, weight will change if fuel is added. However, they can be defined and measured with sufficient accuracy for any particular purpose, and the features can be specified as needed, such as "weight of vehicle empty."

Clearly, objects do not have to be physical. The cost of living is a non-physical object. It has a definition and features. The features of the cost of living can be measured, such as what it may be in dollars, its rate of change per month or year, what percentage of the mean or median income it represents, and so on.

Objects in the real world relate to and interact with each other. Living objects interact with

the world in noticeable and familiar ways, such as eating and breathing. Even inanimate objects interact with the world. Rocks, for example, interact with the ground on which they rest at an atomic level, and so do not sink into it. Mountains are worn down by weather, and even continents interact with the core of the earth and drift about. The cost of living changes, as does the unemployment level—driven (we say) by the economy and marketplace. All of these interactions form what philosophers have called “the great system of the world.” The features of objects captured as data form a reflection of this great system of the world. If the reflection is accurate, the features themselves, to a greater or lesser degree, represent that system. It is in this sense that data is said to represent or, sometimes, to form a system.

2.1.2 Capturing Measurements

For the data miner, objects actually consist of measurements of features. It is the groups of features that are taken as the defining characteristics of the objects, and actual instance measurements of the values of those features are considered to represent instances of the object. For instance, my car is a dark blue, two-door, six-cylinder, five-passenger vehicle. That is to say, for this particular instance of “car,” considering five features—ownership, color, door count, cylinder count, and passenger capacity—the measurements are Dorian Pyle, dark blue, 2, 6, 5.

These measurements are all taken in such a way that they have a particular type of validity. In this particular case, they were all taken at the same time, which is to say that they were true at the instant of my writing. The *validating feature* here, then, is a timestamp. This need not be the case, of course, although timestamps are very often used. To continue using cars as an example, other validating stamps might be “all 18-year-old males,” or “all Ford Escorts,” or “all red cars with four cylinders.” This would mean collecting measurements about all vehicles owned by 18-year-olds, all Ford Escorts, or all red cars with four cylinders, for instance.

There is an assumption here, then, that measurements are taken about objects under some validating circumstance. In effect, the world state is “frozen” by the validating circumstance and the measurements taken yielding a particular value. This idea of “freezing” the world’s state while taking measurements is an important one, particularly for miners. There are a variety of factors involved in taking measurements that can make the measurements seem inconsistent. Since it is very often part of mining to understand and estimate where the variability in a particular measurement comes from, as well as how reliable the measurement is, we need to look at some sources of variability.

2.1.3 Errors of Measurement

Measurement implies that there is some quantity to measure, and some device to calibrate the measurement against. A simple illustration of such a physical measurement is measuring a distance with a ruler. A nonphysical measurement might be of an opinion

poll calibrated in percentage points of one opinion or another.

There are several ways in which a measurement may be in error. It may be that the quantity is not correctly compared to the calibration. For instance, the ruler may simply slip out of position, leading to an inaccurate measurement. The calibration device may be inaccurate—for instance, a ruler that is longer or shorter than the standard length. There are also inevitable errors of precision. For example, measurements of distance simply have to be truncated at some point, whether measuring to the nearest mile, foot, meter, centimeter, or angstrom unit.

Some of these errors, such as incorrect comparison, lead to a sort of “fuzz” in the measurement. Since there are likely to be as many measurements short as there are long, such errors also tend to cluster about the “correct” point. Statisticians have devised many ways to characterize this type of error, although the details are not needed here. If the calibration is in error—say, wrong ruler length—this leads to a systematic error, since all measurements made with a given ruler tend to be “off” the mark by the same amount. This is described as a *bias*.

Figure 2.1 shows the distortion, or error, that might be caused by the “fuzz” in such measurements. It shows what unbiased error might do to a measurement. Figure 2.2 shows what bias added to unbiased error might look like. These types of measurements are showing “point” measurements, so called because if taken without any error they appear as points on a graph.

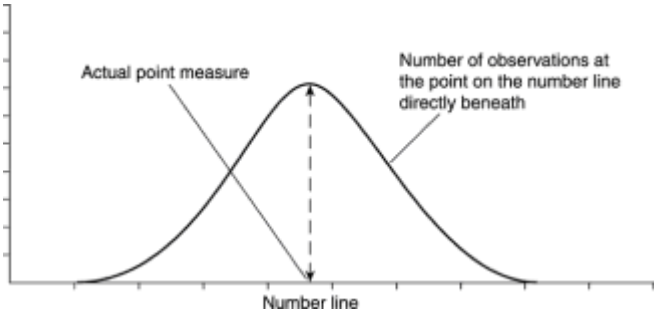

Figure 2.1 Unbiased noise spreads the measurements evenly around the measurement point. Most cluster near the actual value.

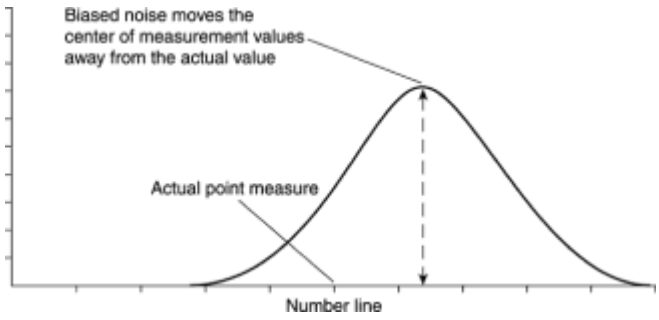

Figure 2.2 Biased noise makes most of the measurements cluster around a point that is not the true measurement.

Environmental errors are rather different in nature, but of particular importance in mining. *Environmental errors* express the uncertainty due to the nature of the world. Another way of looking at this interaction is that it expresses uncertainty due to the nature of the interactions between variables. These between-variable interactions are critically important to miners. Since there is some level of uncertainty in these interactions, they warrant a much closer inspection.

Suppose a particular potential purchaser of products from a catalog has actually made a previous purchase. The catalog company wants to measure several features of the object "purchaser" to combine them with measurements about other purchasers and create a general purchaser profile. There are many circumstances in the world that surround and influence purchasers. To make the required measurements, the world is "frozen" in its state for the particular purchaser and the surrounding circumstances captured. Several variables are measured. Each measurement is, of course, subject to the point distortion, or error, described previously.

Each fuzzy circle in Figure 2.3 represents such a single measurement. The central point of each circle represents the idealized point value, and the surrounding circle represents the unavoidable accompanying fuzz or error. Whatever the value of the actual measurement, it must be thought of as being somewhere in this fuzzy area, near to the idealized point value.

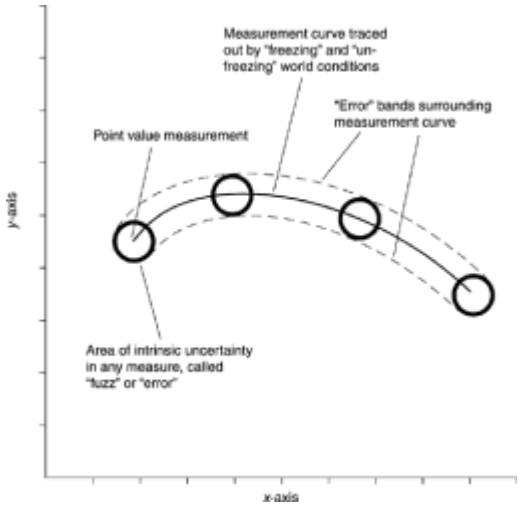

Figure 2.3 Taking several point measurement values with uncertainty due to error outlines a measurement curve surrounded by an error band.

Suppose now that the world is unfrozen, conditions allowed to change minutely, and then refrozen. What happens to the measurement? If the driving factors are linearly related to the measurement, then a minute change in circumstances makes a minute change in measurement. The measurement taken under the slightly changed circumstances is slightly changed in direction and distance from the first measurement. Other minute changes in the world's state make similar minute changes in measurement. Such a series of measurements traces out the fuzzy line shown in Figure 2.3. This is the sort of change in measurement that might be traced out in the value of your bank account, say, if income varied by some small amount. The small change in income represents a change in the state of the world. The error represents the general fluctuation in bank account level due to the normal uncertainties of life. Perhaps if your income were slightly lower, the bank balance would be a little lower. An increase in income might raise the bank balance a little, but a further increase might lower it as you might then choose to put money into another account. This small change in your bank account that is associated with a small change in income demonstrates the effect of a linear relationship.

But perhaps the relationship is not linear, at least locally. What does this mean? It might mean that a minute change in some other circumstance would persuade you to use a completely different bank. Perhaps a better interest rate paid by another bank might be enough. This could mean that the overall shape of the curves would be the same, but their height would change, indicating the influence of interest rate changes. Figure 2.4 shows what this might look like.

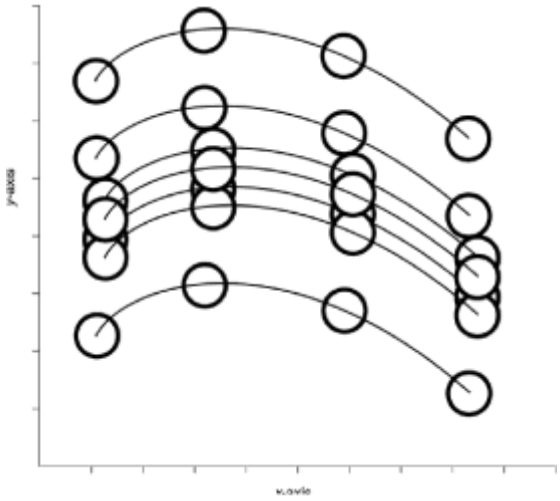

Figure 2.4 Groups and clusters of curves that result when a small change in world conditions makes a nonlinear or “step” change in the measured values.

What this figure might mean is that a small change in interest rate persuaded you to take all your money out of one bank and deposit it in another bank. For one bank the minute change in interest rate means that you completely and totally disappear as a customer, while appearing as a new customer for some other bank. The world change is small, but instead of slightly changing the bank balance up or down in the first bank, it made it disappear! The various lines in Figure 2.4 might then represent different banks, with the curves representing different balances. This “curve bundle” represents where and how the point measurements might map onto the world under the slightly different circumstances.

Some conditions, then, would be very sensitive to a minute change in circumstances due to the unfreezing/minute change/refreezing cycle, while other conditions were not so sensitive, or might be even completely unaffected by small changes. What this means for actual measurements is that, even for minute changes in the circumstances surrounding measurements, there are a variety of possible results. The changes may be undetectably small, given the nature of truncating the measurement accuracy discussed above. So the measurements traced out in state space, due possibly to the miniscule perturbations that are unavoidable in the real world, trace out not fuzzy points, but fuzzy curves. (State space will be covered in more detail later in this chapter. For now, very briefly, state space is the space in which measurement values can be plotted, like the space on a graph.)

A very important point to note for the miner here is that while many of the environmental factors may be unknowable, and certainly uncontrollable, they are subject to some limitations. For instance, it is very unlikely that any minute change in world conditions would change your deposit in a bank account into your ownership of a Swiss bank! Defining the limits, and determining the shape and size of the measurement curves, can be a critical factor in building models.

Determining the extent of the error is not so important to data preparation. What is important, and the reason for the discussion, is that during preparation it may be possible to determine where some of the components in the overall error come from, and to explore its shape. Mining to build models is concerned with addressing and, if possible, understanding the nature of the error; data preparation, with exposing and, perhaps, ameliorating it.

2.1.4 Tying Measurements to the Real World

Sometimes measurements are described as consisting of two components: the actual absolute perfect value, plus distortion. The distortion is often referred to as error. However, the distortion is actually an integral part of the measurement. Use of the term "error" has unfortunate connotations, as if there is somehow something wrong with the measurement. There seems to be an implication that if only the measurer had been more careful, the error could have been eliminated. While some part of the distortion may indeed result from a mistake on the part of the measurer, and so truly is an error in the sense of a "mistake," much of the distortion is not only unavoidable, but is actually a critical part of what is being measured.

This use of the term "error" is emotionally loaded in ways that do a disservice to the miner. For all of the reasons discussed above, actual measurements are better envisioned as represented by curve bundles drawn in some state space. Some part of the curve will represent error in the sense of mistakes on the part of the measurer—whether human or machine. However, most of the range of the curve represents the way the feature maps onto an uncertain world. Even a perfect measurer, should such a thing exist, would still not be able to squeeze the various curves into a single point—nor even into a single curve.

Contrary to the view that there exists some perfect measurement with error, the more realistic situation is one that includes a distributed mapping of the measurement onto the real world, plus some single estimate of the location of some particular instance on the bundle of curves. Nonetheless, the term "error" is the one in general use in measurement and must be accepted. Remember that it is not to be thought of as some sort of mistake to be corrected, but as representing an essential and unavoidable part of the measurement that is integral with mapping the feature measured to the real world. Moreover, it is often the job of the miner to discover the shape of the measurement bundle. In mining, the error is often included in a feature called "noise" (looked at in detail later), although noise also includes other components.

2.2 Types of Measurements

So far this chapter has discussed measurements in general, and problems and limitations with making the measurements. Looked at more closely, there is an intuitive difference

between different types of measurements. For instance, the value “1.26 feet” is obviously of a different type than the value “green.” This difference has a major impact on both the way data is prepared and the way it is modeled. Since these differences are important, the different types of measurements need to be examined in some detail.

All measurements have one feature in common: they are all made on some scale. Measurements in general map onto the world in ways represented by the measurement curve bundles. Individual instance values are not curves, but point measurements. It is usual to speak of the measurements of a particular feature of an object as a variable. A variable represents a measurement that can take on a number of particular values, with a different value possible for each instance. Conceptually, a variable is a container holding all of the measurements of a particular feature of some specific object. But different types of containers are needed to hold different types of measurements, just as tomatoes and soda both need different types of containers to hold them. The “containers” for variables are a way to classify them using descriptions such as “nominal” and “ratio” that will be discussed in a moment. Some variables consist of two components—the scale on which they are measured and the measured value itself—and others require more components. The class of variables that can be indicated by the position of a single point (value) on some particular scale are called *scalar variables*. There are other types of variables that require more than one value to define them; they are often called *vector variables*. Most of the work of the miner considers scalar variables, and these need to be examined in detail. So first, we will look at the different types of containers, and then what is in each of them.

2.2.1 Scalar Measurements

Scalar measurements come in a variety of types. Different types of measurements inherently carry different amounts of information. You can intuitively see this: just think about measuring the temperature of your coffee. By limiting the measurement to just “hot” or “cold,” you will see that this measurement contains less information than the measurements “scalding,” “too hot,” “nice and hot,” “hot,” “not hot,” “warm,” “cool,” and “cold.” The idea of information content is a very useful way to order the types of scalar measurements.

Nominal Scale Measurements

Values that are nominally scaled carry the least amount of information of the types of measurements to be considered. Nominal values essentially just name things. There is a notable difference in type or identity, but little or nothing more can be said if the scale of measurement is actually nominal. A nominal measurement is little more than a label used for purposes of identification. There is no inherent order in the nominal measurements. Nor indeed can nominally measured values even be meaningfully grouped together. They do, nonetheless, carry definite information, little though it might be.

Categorical Scale Measurements

Categorical measurements name groups of things, not individual entities. This categorization allows values to be grouped in meaningful ways. As with nominal measurements, nothing more can be said about the size or type of the differences. They are no more than labels for different groups.

For instance, ZIP codes, although they look like numbers, are really simply arbitrary labels for postal delivery zones. Listing them in their apparent numerical order is not particularly revealing. Standard industry classification (SIC) codes are very similar to ZIP codes in that, although they categorize different types of business activity, a numerical ordering seems no more nor less reasonable than an alphabetical listing of the activity represented by the number. It should also be clear that any ordering of scales such as marital status, ethnic background, or academic interest seems quite arbitrary.

However, it is possible to use a number as a categorically measured value label in order to more conveniently label and differentiate the category to which it belongs. Although formed using characters that are numerical symbols rather than letters of the alphabet, the labels remain exactly that, labels, and carry no numerical significance. Postal authorities numerically labeled ZIP codes, and the federal government numerically labeled SIC codes. Numerically labeled or not, all that can be said about the categories is that they are different in type. Numbering the measurement values is only a matter of convenience, and there is no implied ordering or ranking.

Not only do the categorical labels have no particular order, there is no information included in the categorization that indicates how different they are from each other. There is no real meaning in saying, for instance, that a plumber is twice a carpenter or three-quarters of a corporate director. You may have some personal feeling as to the amount of difference there is between different brands of boot polish. However, there is no way to determine, simply by knowing the category of the product, if the amount of difference between black and brown polish is more or less than between, say, black and red or red and brown types of polish. All that can be said, simply by knowing its category, is that there is a difference.

Categorical measurements, then, denote that there is a difference in kind or type, but are not able to quantify the difference. The scale used amounts to no more than a comprehensive listing of all of the categories into which the value can fall.

Ordinal Scale Measurements

When something more can be said about the measurement scale used, the additional information gives some sort of order to the categories that are used to label the measurement. Because there is some sort of meaningful order to the listing of the labels, this type of measurement is often known as an *ordinal measurement*.

Ordinal measurements carry far more information than either nominals or categoricals. You may not be surprised to learn that by taking a table of the actual distances between, say, major American cities, the joint distance table alone is enough to re-create the layout of the cities that show up on a map. Surprisingly, instead of the actual distances, just using a simple pairwise ranking of cities by distance is also enough to re-create the layout of the cities as seen on a map almost perfectly.

This example shows that for the purpose of making a schematic map, knowing the actual distances provides little additional information. Most of the information required to accurately create a schematic map of American cities is enfolded in a simple pairwise ranking of cities by their distance apart.

The ranking of the categories must be done subject to a very particular condition, called *transitivity*, which is actually a reasonable notion although of critical importance. Transitivity means that if A is ranked higher than B, and B higher than C, then A must be ranked higher than C. That is: If A > B and B > C, then A > C. While measuring a value using an ordinal scale adds a huge amount of information over that contained in a categorical measurement, the transitivity requirement places some constraints on how the ordinal scale can actually be built. Note that the ordinal scale does not require that anything has to be specified about the amount of the difference between each category.

For instance, at a “blind tasting” for wines, you sample several different types and styles of wine and mark down pairwise combinations of preference. Perhaps you prefer the cabernet to the merlot, and the merlot to the shiraz. If transitivity holds and you prefer the cabernet to the shiraz, the result is an ordinal listing of wine preferences: the favorites, in order, are cabernet, merlot, and shiraz. However, there is no indication of by how much you prefer the cabernet to the merlot. It may be that the difference in preference is slight: you choose cabernet 51% of the time and merlot 49% of the time; the shiraz doesn’t get a look in. On the other hand, given the availability of cabernet, perhaps you will choose that every time, only considering the others when cabernet is unavailable.

The point here is that ordinal measurements do indeed carry a lot of information, but allow for no comparison of the magnitude of the differences between the categories.

Interval Scale Measurements

When there is information available not only about the order for ranking the values measured but also about the differences in size between the values, then the scale is known as an *interval scale*. This means that the scale carries with it the means to indicate the distance that separates the values measured. Interval variables are almost always measured using numbers. Because numbers are almost exclusively used when discussing interval-scaled values, measurements scaled this way are part of the group called *quantitative measurements*—that is, values that capture differences in, changes in, or the amount of the quantity of some attribute of an object.

An interval scale that almost everyone is familiar with is the temperature scale. Every day newspapers, radio, and television provide a forecast of the temperature range for the day's weather. If the low for the day is predicted to be 40 and the high 50, this provides some particular idea of what temperature you will experience. In this case the range through which the temperature is expected to move is 10º. If at some other time of year the low/high is forecast as 80 through 90, you can tell that the expected temperature range is again 10º, the same as earlier in the year. Thus the difference of 10º indicates the same amount of temperature change regardless of where it occurs on the range of the scale.

However, you cannot say, based on the interval scale used, that the low for the two days can be compared using their ratio. That is to say, 80º is not twice as hot as 40º. It is easy to see that there must be something wrong in supposing that the ratios are meaningful if instead of using Fahrenheit, you made the same comparison using the Celsius scale to measure the same temperature range.

Roughly speaking, 80ºF corresponds to 25ºC, while 40ºF corresponds to about 5ºC. However, the ratio of 80 to 40 is 2, but the ratio of 25 to 5 is 5! This means that when measuring the temperature with a Fahrenheit thermometer, you might say that it was twice as hot (ratio of 2), but your Celsius-using neighbor claims that it was five times as hot (ratio of 5)! One of these observations at least must be wrong, and in fact they are both wrong.

What is wrong, of course, is that the zero point, often called the *origin of the scale*, is not at the same temperature for the two scales. This means that the scales have differing ratios at equivalent temperatures. In fact, the zero point is arbitrarily set, which is a characteristic of interval scales. So, as far as temperature goes, scientists use a scale known as the "Absolute" or "Kelvin" scale specifically to overcome this problem. On this scale, the zero point corresponds to a true zero point so that the ratios of numbers compared on this scale have meaning.

Ratio Scale Measurements

The scale that carries the most information content is the *ratio scale.* One ratio scale measure that you are no doubt very familiar with measures the content of your bank account. It starts at a true zero point, which is to say that when the bank balance is 0 it is because there is no money in it. Also, it is denominated in currency units of equal value and size. This means that you can express meaningful ratio values of the state of your finances, knowing, for instance, that $10 is twice as much as $5, and $100 is twice $50. At any position on the scale, for any values, the ratio is a meaningful measure of properties of the scale.

As with the interval scale, ratio-scaled values are also quantitative. It is useful to consider

two types of ratio-scaled measurements: those for which the scale that they are measured on must be named and those for which no scale is named. The characteristics of each type are sufficiently different that it is sometimes important to treat them differently during data preparation.

Usually it is important to know the units of a particular ratio measurement. To measure sales activity as “5” is not useful. Even if you knew that they were “4” last month, there is no reference in the numbers to indicate their significance. Knowledge of the unit of measurement is required. It means something if we stipulate that the units are millions of dollars and something else again if the units are thousands of Russian rubles or numbers of units shipped.

There is a class of ratio-scaled values that is measured *only* as numbers. These numbers are sometimes called *dimensionless*. A dimensionless number expresses a relationship that holds true without reference to the underlying measurements of the scale. For instance, consider a lighthouse standing on a rocky headland. Each lighthouse signals in a particularly distinct way such that any ship that sees the signal knows which particular lighthouse is in view just from the pattern of the signal. The lighthouse signals by showing a light in a unique pattern that is repeated over time. In any time cycle, however long or short, the light is on for a certain duration and off for another duration. Suppose that for a particular lighthouse the light is on for 10 seconds and off for 5. The ratio of on/off is 10/5, which, by division, reduces to 2/1, or 2. This measurement, sometimes known as a *duty cycle*, is dimensionless, and for this particular lighthouse it is 2. That is not 2 per anything, or 2 anythings, simply 2. The lighthouse pattern repeats once in 15 seconds, or four times per minute. So long as we consider only complete cycles, it doesn’t matter at all over how long a period the duty cycle of the lighthouse is measured; it will always be 2.

Care must be taken with measurements over time. Sometimes these measurements are assumed to be dimensionless when in fact they are not. A common discussion of ratio-scale variables discusses the distinction between “how many” and “how much.” The “how much” type is said to require the scale units, as in the sales figures just discussed. “How many” types of measurements are often said not to need such units. For instance, in stock market reports, not only is the market index quoted, but frequently the “advance/decline ratio” is given.

“The stock market was up today,” the news anchor might say, “with advances leading declines 5 to 4.” This means that five stocks went up in price for every four that did not. Now a ratio of 5/4 can be given as 1.25 and this gives the appearance of a dimensionless number. Here is a measurement of “how many” (i.e., 5/4) rather than “how much” (which is measured in “points” or dollars or some other specified unit).

However, the “gotcha” is that the count of advances to declines was taken over today. When considering the example of the lighthouse, a very important point was that, so long as we looked at complete cycles, the length of time of the observation did not matter. The

advance/decline ratio applies only to the specific period over which it was measured. If a period is included that is longer or shorter by only a few minutes, it is possible that the measurement would not be 5/4. Indeed, you can be quite sure that in choosing some other period there is no reason to think that the 5/4 ratio holds true except by coincidence.

We have become so culturally accustomed to the idea of fixed and “natural” measurements of time that it is easy to overlook the fact that measurements of duration are arbitrary. By happenstance the Babylonians had a number system based on the number 60. Since it was they who made the original astronomical measurements, and because they thought there were approximately 360 days in a year, we now have 360 degrees in a circle. By a series of what they thought were convenient divisions, they arrived at a 24-hour day as standard. The hours were further divided in smaller parts, and by making a second division of the minute parts of an hour by 60 we get the “seconds,” so called because of the second division involved. Very clever and useful. However, there are alternatives.

Napoleon, in attempting to introduce the metric system, tried to “rationalize” all of the measuring systems then in use. Measures for distance and mass—the meter and gram, respectively—were adopted; however, the division of the year into 10 months and the day into 10 hours, and so on, was not accepted. The point is that all of our measurements of time are arbitrary. Some are arbitrary through human selection, but even the rotation of our planet has varied considerably through the eons. The first creatures out of the primordial soup probably experienced an 18-hour day. The slowing of planetary rotation has brought us to a day of approximately 24 hours. Because we are creatures of planet Earth, there are many cycles that are tied to days, seasons, and years. However, there is nothing inherently special about these scale units any more than any other scale unit.

Measurements in time, then, need to be considered carefully. By identifying and confirming complete cycles, returning to an identifiable identical state from time to time, dimensionless numbers may be useful. Measures based on the “how many/how much” dichotomy are suspect.

2.2.2 Nonscalar Measurements

Scalar values consist of just two component parts, the value of the measurement and the scale against which the measurement was made. In traffic court it is enough to prove that the speed of a vehicle was, or was not, some particular number of miles per hour. The speed is expressed as a number and the scale in miles per hour. Nonscalar measurements need more component parts to capture additional information. Speed is the number of miles traveled in one hour. Velocity, however, is measured as speed in a particular direction. There are at least four components in such a measurement—two scales and two measurements on those scales. Navigation at sea, for instance, is very concerned with velocity—how fast and in which direction the vessel is travelling.

Measurements such as velocity can be plotted on a two-dimensional graph in the form of a point specified by the measurements of speed on one axis, and direction on another. A line drawn on a graph from the common point that was chosen to begin the measurement representation, to the point where it ends up, is known as a *vector*. There is a great deal of literature about such vector quantities, their properties, and how to manipulate them. So far as data preparation is concerned, however, vector quantities are built out of scalar quantities. It is true that the scalar quantities are linked in particular and significant ways, but for the purposes of data preparation, the vectors can be carefully treated as multiple scalar values.

This is not to minimize the importance of vector quantities. Indeed, the concept of state space regards the instance value of multiple features as a multidimensional vector. Each record in a table, in other words, is taken as a vectoral representation. The point here is that most vectors can be thought of as being made up of separate scalar values and can be usefully treated for purposes of data preparation at the scalar level.

2.3 Continua of Attributes of Variables

So far this chapter has addressed the way in which measurements are taken using different types of scale. Collections of measured values of particular features are grouped together into variables. Because the values are collected together, it is possible to look for patterns in the way the values change with changes in the validating feature, or with changes in other variables.

When the measurements are actually taken in practice, certain patterns appear if many instances of values of a variable are considered as a whole. These aggregate collections of values begin to show a variety of different features. It is hard to characterize these elementary patterns as a part of data mining, although they are in truth the surface ripples of the deeper structure that miners will be seeking when the actual data mining tools are applied to the prepared data. Although this discussion only concerns introductory issues about data preparation, it is still true that the data preparation begins with a fairly comprehensive survey of the properties of each of the variables taken individually. It is in the appreciation of the basic types of attributes of variables that data preparation begins. Chapter 4 looks at this issue in considerable detail. Later discussion in this chapter summarizes the methods that will prepare variables for modeling.

Although described as if each of the scales were separate, actually the types blur together into a more continuous spectrum than the separate descriptions seem to imply. It is usual to describe variables as being of the same type as the scale, or features of the scale, on which they are measured. So it is convenient, say, to talk of a categorical variable, or a continuous variable. A measured value on a scale is, of course, a single point and as such cannot show any pattern. It cannot even show any of the “fuzz” of noise discussed earlier. Variables, being collections of instance values of a particular feature, all being made on a common scale, do show recognizable patterns, or attributes. It is these common attributes

of variables that can be described as existing in a continuum.

2.3.1 The Qualitative-Quantitative Continuum

This continuum captures the low to high information content of the different types of scalar variables. Describing variables as qualitative or quantitative might not make it obvious that what is being described is information content. Nominal variables are at the qualitative end of the scale—that is, they separate attributes by a difference in quality. Similarly, at the other end of the scale is the ratio (quantitative) association. Information content varies continuously across the scale. Any sharp division implied by the qualitative-quantitative differentiation is not really present. So this continuum really recapitulates the differences in the scales that were discussed before, except that it considers the impact of the different scales on variables.

2.3.2 The Discrete-Continuous Continuum

This will prove to be a very important distinction about variables. In fact, the discrete-continuous distinction forms a continuum. As was done when considering scales, for ease of explanation it is easiest to look at several points along the continuum. At each of the points viewed, the distinctions are easy to draw. There are, however, no hard and fast boundaries in practice.

As a very brief introduction to the following discussion, discrete variables are considered to have a very limited set of values that they can take on, such as colors of the rainbow. Continuous values can take on any value within a range, like the temperature. To see that this is a continuum, consider your bank account—discrete or continuous? Technically, it is discrete as it is restricted to values to the nearest penny. In practice, however, the quantization, or fineness of division, is such that it would usually be more useful to consider it as a continuous value.

Single-Valued Variables (Constants)

It may seem odd to discuss a “variable” as having only a single value. Strictly speaking, since it is not varying its value, it would seem to be something other than a variable. However, variables that do not vary are often used, and very useful they are, too. Some examples of constants are the number of days in a week, inches in a foot, the distance represented by a light year, and the number of sides in a triangle. These constant values are representative of what we see as invariant, defining characteristics of an object.

They also turn up when modeling variables. Perhaps a marketing organization wants to examine all records for “the gold card upgrade program.” There may be many different marketing programs represented in the original data set. In this original data set, the variable “program name” is variable—it varies by having different values representing the different programs. The indicator for the gold card upgrade program is, say, “G”. Different

letters are used to identify other programs. However, by the time only the records that are relevant to the gold card upgrade program are extracted into a separate file, the variable "program name" becomes a constant, containing only "G" in this data set. The variable is a defining feature for the object and, thus, becomes a constant.

Nonetheless, a variable in a data set that does not change its value does not contribute any information to the modeling process. Since constants carry no information within a data set, they can and should be discarded for the purposes of mining the data.

Two-Valued Variables

At least variables with two values do vary! Actually, this is a very important type of variable, and when mining, it is often useful to deploy various techniques specifically designed to deal with these *dichotomous* variables. An example of a dichotomous variable is "gender." Gender might be expected to take on only values of male and female in normal use. (In fact, there are always at least three values for gender in any practical application: "male," "female," and "unknown.")

Empty and Missing Values: A Preliminary Note

A small digression is needed here. When preparing data for modeling, there are a number of problems that need to be addressed. One of these is missing data. Dealing with the problem is discussed more fully later, but it needs to be mentioned here that even dichotomous variables may actually take on four values. These are the two values it nominally contains and the two values "missing" and "empty."

It is often the case that there will be variables whose values are missing. A *missing* value for a variable is one that has not been entered into the data set, but for which an actual value exists in the world in which the measurements were made. This is a very important point. When preparing a data set, the miner needs to "fix" missing values, and other problems, in some way. It is critical to differentiate, if at all possible, between values that are missing and those that are empty. An *empty* value in a variable is one for which no real-world value can be supposed.

A simple example will help to make the difference clear. Suppose that a sandwich shop sells one particular type of sandwich that contains turkey with either Swiss or American cheese. In order to determine customer preferences and to control inventory, the store keeps records of customer purchases. The data structure contains a variable "gender" to record the gender of the purchaser, and a variable "cheese type" to record the type of cheese in the sandwich. "Gender" could be expected to take the values "M" for male and "F" for female. "Cheese type" could be expected to take the values "S" for Swiss and "A" for American cheese.

Suppose that during the recording of a sale, one particular customer requests a turkey

sandwich with no cheese. In recording the sale the salesperson forgets to enter the customer's gender. This transaction generates a record with both fields "gender" and "cheese type" containing no entry. In looking at the problem, the miner can assume that in the real world in which the measurements were taken, the customer was either male or female, and any adjustment must be made accordingly. As for "cheese type," this value was not measured because no value exists. The miner needs a different "fix" to deal with this situation.

If this example seems contrived, it is based on an actual problem that arose when modeling a grocery store chain's data. The original problem occurred in the definition of the structure of the database that was used to collect the data. In a database, missing and empty values are called *nulls*, and there are two types of null values, one each corresponding to missing and empty values. Nulls, however, are *not* a type of measurement.

Miners seldom have the luxury of going back to fix the data structure problem at the source and have to make models with what data is available. If a badly structured data set is all that's available, so be it; the miner has to deal with it! Details of how to handle empty and missing values are provided in Chapter 8. At this point we are considering only the underlying nature of missing and empty variables.

Binary Variables

A type of dichotomous variable worth noting is the *binary* variable, which takes on only the values "0" and "1." These values are often used to indicate if some condition is true or false, or if something did or did not happen. Techniques applicable to dichotomous variables in general also apply to binary variables. However, when mining, binary variables possess properties that other dichotomous variables may not.

For instance, it is possible to take the mean, or average, of a binary variable, which measures the occurrence of the two states. In the grocery store example above, if 70% of the sandwich purchasers were female, indicated by the value "1," the mean of the binary variable would be 0.7. Certain mining techniques, particularly certain types of neural networks, can use this kind of variable to create probability predictions of the states of the outputs.

Other Discrete Variables

All of the other variables, apart from the constants and dichotomous variables, will take on at least three or more distinct values. Clearly, a sample of data that contains only 100 instances cannot have more than 100 distinct values of any variable. However, what is important is to understand the nature of the underlying feature that is being measured. If there are only 100 instances available, these represent only a sample of all of the possible measurements that can be taken. The underlying feature has the properties that are

indicated by all of the measurements that could be taken. Much of the full representation of the nature of the underlying feature may not be present in the instance values actually available for inspection. Such knowledge has to come from outside the measurements, from what is known as the *domain of inquiry*.

As an example, the underlying value of a variable measuring "points" on a driving license in some states cannot take on more than 13 discrete values, 0–12 inclusive. Drivers cannot have less than 0 points, and if they get more than 12 their driving licenses are suspended. In this case, regardless of the actual range of values encountered in a particular sample of a data set, the possible range of the underlying variable can be discovered. It may be significant that a sample does, or does not, contain the full range of values available in the underlying attribute, but the miner needs to try to establish how the underlying attribute behaves.

As the density of discrete values, or the number of different values a variable can take on, increases for a given range, so the variable approaches becoming a continuous variable.

In theory, it is easy to determine the transition point from discrete to continuous variables. The theory is that if, between any two measurements, it is inherently possible to find another measurement, the variable is continuous; otherwise not. In practice it is not always so easy, theoretical considerations notwithstanding. The value of a credit card balance, for instance, can in fact take on only a specifically limited number of discrete values within a specified range. The range is specified by a credit limit at the one end and a zero balance (ignoring for the moment the possibility of a credit balance) at the other. The discrete values are limited by the fact that the smallest denomination coin used is the penny and credit balances are expressed to that level. You will not find a credit card balance of "$23.45964829." There is, in fact, nothing that comes between $23.45 and $23.46 on a credit card statement.

Nonetheless, with a modest credit limit of $500 there are 50,000 possible values that can occur in the range of the credit balance. This is a very large number of discrete values that are represented, and this theoretically discrete variable is usually treated for practical purposes as if it were continuous.

On the other hand, if the company for which you work has a group salary scale in place, for instance, while the underlying variable probably behaves in a continuous manner, a variable measuring which of the limited number of group salary scales you are in probably behaves more like a categorical (discrete) variable.

Techniques for dealing with these issues, as well as various ways to estimate the most effective technique to use with a particular variable, are discussed later. The point here is to be aware of these possible structures in the variables.

Continuous Variables

Continuous variables, although perhaps limited as to a maximum and minimum value, can, at least in theory, take on any value within a range. The only limit is the accuracy of representation, which in principle for continuous variables can be increased at any time if desired.

A measure of temperature is a continuous variable, since the “resolution” can be increased to any amount desired (within the limit of instrumentation technology). It can be measured to the nearest degree, or tenth, or hundredth, or thousandth of a degree if so chosen. In practice, of course, there is a limit to the resolution of many continuous variables, such as a limit in ability to discriminate a difference in temperature.

2.4 Scale Measurement Example

As an example demonstrating the different types of measurement scales, and the measurements on those scales, almost anything might be chosen. I look around and see my two dogs. These are things that appear as measurable objects in the real world and will make a good example, as shown in Table 2.1.

TABLE 2.1 Title will go here

Scale Type	Measurement	Measured Value	Note
Nominal	Name	• Fuzzy • Zeus	Distinguishes one from the other.
Categorical	Breed	• Golden Retriever • Golden Retriever	Could have chosen other categories.
Categorical (Dichotomous)	Gender	• Female • Male	
Categorical (Binary)	Shots up to Date (1=Yes;0=No)	• 1 • 1	

Categorical (Missing)	Eye color	• Value exists in real world	
Categorical (Empty)	Drivers License #	• No such value in real world	
Ordinal	Fur length	• Longer • Shorter	Comparative length allowing ranking.
Interval	Date of Birth	• 1992 • 1991	
Ratio	Weight	• 78 lbs • 81 lbs	
Ratio (Dimensionless)	Height / Length	• 0.5625 • 0.625	

2.5 Transformations and Difficulties—Variables, Data, and Information

Much of this discussion has pivoted on information—information in a data set, information content of various scales, and transforming information. The concept of information is crucial to data mining. It is the very substance enfolded within a data set for which the data set is being mined. It is the reason to prepare the data set for mining—to best expose the information contained in it to the mining tool. Indeed, the whole purpose for mining data is to transform the information content of a data set that cannot be directly used and understood by humans into a form that can be understood and used.

Part of Chapter 11 takes a more detailed look at some of the technical aspects of information theory, and how they can be usefully used in the data preparation process. Information theory provides very powerful and useful tools, not only for preparing data, but also for understanding exactly what is enfolded in a data set. However, while within the confines of information theory the term “information” has a mathematically precise definition, Claude Shannon, principal pioneer of information theory, also provided a very apt and succinct definition of the word. In the seminal 1949 work *The Mathematical Theory of Communication*, Claude E. Shannon and Warren Weaver defined information as “that which reduces uncertainty.” This is about as concise and practical a definition of information as you can get.

Data forms the source material that the miner examines for information. The extracted information allows better predictions of the behavior of some aspect of the world. The improved prediction means, of necessity, that the level of uncertainty about the outcome is reduced. Incorporating the information into a predictive or inferential framework provides knowledge of how to act in order to bring about some desired result. The information will usually not be perfect, so some uncertainty will remain, perhaps a great deal, and thus the knowledge will not be complete. However, the better the information, the more predictive or powerfully inferential the knowledge framework model will be.

2.6 Building Mineable Data Representations

In order to use the variables for mining, they have to be in the form of data. Originally the word “datum” was used to indicate the same concept that is indicated here, in part, by “measurement” or “value.” That is, a datum was a single instance value of a variable. Here measurement both signifies a datum, and also is extended to indicate the values of several features (variables) taken under some validating condition.

A collection of data points was called data, and the word was also used as a plural form of datum. Computer users are more familiar with using data as a singular noun, which is the style adopted here. However, there is more to the use of the term than simply a collection of individual measurements. Data, at least as a source for mining, implies that the data points, the values of the measurements, are all related in some identifiable way. One of the ways the variables have to be structured has already been mentioned—they have to have some validating phenomenon associated with a set of measurements. For example, with each instance of a customer of cellular phone service who decides to leave a carrier, a process called *churning*, the various attributes are captured and associated together.

The validating phenomenon for data is an intentional feature of the data, an integral part of the way the data is structured. There are many other intentional features of data, including basic choices such as what measurements to include and what degree of precision to use for the measurements. All of the intentional, underlying assumptions and choices form the *superstructure* for the data set. Three types of structure are discussed in the next chapter. Superstructure, however, is the only one specifically involved in turning variables into data.

Superstructure forms the framework on which the measurements hang. It is the deliberately erected scaffolding that supports the measurements and turns them into data. Putting such scaffolding in place and adding many instances of measured values is what makes a data set. Superstructure plus instance values equals data sets.

2.6.1 Data Representation

The sort of data that is amenable to mining is always available on a computer system.

This makes discussions of data representation easy. Regardless of how the internal operations of the computer system represent the data, whether a single computer or a network, data can almost universally be accessed in the form of a table. In such a table the columns represent the variables, and the records, or rows, represent instances. This representation has become such a standardized form that it needs little discussion. It is also very convenient that this standard form can also easily be discussed as a matrix, with which the table is almost indistinguishable. Not only is the table indistinguishable from a matrix for all practical purposes, but both are indistinguishable from a spreadsheet.

Spreadsheets are of limited value in actual mining due to their limited data capacity and inability to handle certain types of operations needed in data preparation, data surveying, and data modeling. For exploring small data sets, and for displaying various aspects of what is happening, spreadsheets can be very valuable. Wherever such visualization is used, the same row/column assumption is made as with a table.

So it is that throughout the book the underlying assumption about data representation is that the data is present in a matrix, table, or spreadsheet format and that, for discussion purposes, such representation is effectively identical and in every way equivalent. However, it is not assumed that all of the operations described can be carried out in any of the three environments. Explanations in the text of actual manipulations, and the demonstration code, assume only the table structure form of data representation.

2.6.2 Building Data—Dealing with Variables

The data representation can usefully be looked at from two perspectives: as data and as a data set. The terms “data” and “data set” are used to describe the different ways of looking at the representation. Data, as used here, implies that the variables are to be considered as individual entities, and their interactions or relationships to other variables are secondary. When discussing the data set, the implication is that not only the variables themselves are considered, but the interactions and interrelationships have equal or greater import. Mining creates models and operates exclusively on data sets. Preparation for mining involves looking at the variables individually as well as looking at the data set as a whole.

Variables can be characterized in a number of useful ways as described in this chapter. Having described some features of variables, we now turn our attention to the types of actions taken to prepare variables and to some of the problems that need to be addressed.

Variables as Objects

In order to find out if there are problems with the variables, it is necessary to look at a summary description and discover what can be learned about the makeup of the variable itself. This is the foundation and source material for deciding how to prepare each

variable, as well as where the miner looks at the variable itself as an object and scrutinizes its key features and measurements.

Naturally it is important that the measurements about the variable are actually valid. That is to say, any inferences made about the state of the features of the variable represent the actual state of the variable. How could it be that looking at the variable wouldn't reveal the actual state of the variable? The problem here is that it may be impossible to look at all of the instances of a variable that could exist. Even if it is not actually impossible, it may be impractical to look at all of the instances available. Or perhaps there are not enough instance values to represent the full behavior of the variable. This is a very important topic, and Chapter 5 is entirely dedicated to describing how it is possible to discover if there is enough data available to come to valid conclusions. Suffice it to say, it is important to have enough representative data from which to draw any conclusions about what needs to be done.

Given that enough data is available, a number of features of the variable are inspected. Whatever it is that the features discover, each one inspected yields insight into the variable's behavior and might indicate some corrective or remedial action.

Removing Variables

One of the features measured is a count of the number of instance values. In any sample of values there can be only a limited number of different values, that being the size of the sample. So a sample of 1000 can have at most only 1000 distinct values. It may very well be that some of the values occur more than once in the sample. In some cases—1000 binary variable instances, for example—it is certain that multiple occurrences exist.

The basic information comprises the number of distinct values and the frequency count of each distinct value. From this information it is easy to determine if a variable is entirely empty—that is, that it has only a single value, that of "empty" or "missing." If so, the variable can be removed from the data set. Similarly, constants are discovered and can also be discarded.

Variables with entirely missing values and variables that contain only a single value can be discarded because the lack of variation in content carries no information for modeling purposes. Information is only carried in the pattern of change of value of a variable with changing circumstances. No change, no information.

Removing variables becomes more problematic when most of the instance values are empty, but occasionally a value is recorded. The changing value does indeed present some information, but if there are not many actual values, the information density of the variable is low. This circumstance is described as *sparsity*.

Sparsity

When individual variables are sparsely populated with instance values, the miner needs to decide when to remove them because they have insignificant value. Chapter 5 describes in some detail how to decide when to remove sparse variables. Essentially, the miner has to make an arbitrary decision about confidence levels, that is, how confident the miner needs to be in the model.

There is more to consider about sparsity, however, than can be seen by considering variables individually. In some modeling applications, sparsity is a very large problem. In several applications, such as in telecommunications and insurance, data is collected in ways that generate very sparsely populated data sets. The variable count can be high in some cases, over 7000 variables in one particular case, but with many of the variables very sparsely populated indeed. In such a case, the sparsely populated variables are not removed. In general, mining tools deal very poorly with highly sparse data. In order to be able to mine them, they need to be collapsed into a reduced number of variables in such a way that each carries information from many of the original variables. Chapter 10 discusses collapsing highly sparse data.

Since each of the instances are treated as points in state space, and state space has many dimensions, reducing the number of variables is called *dimensionality reduction*, or *collapsing dimensionality*. Techniques for dealing with less extreme sparsity, but when dimensionality reduction is still needed, are discussed in Chapter 7. State space is described in more detail in Chapter 6.

Note that it has to be the miner's decision if a particular variable should be eliminated when some sparsity threshold is reached, or if the variable should be collapsed in dimensionality with other variables. The demonstration software makes provision for flagging variables that need to be retained and collapsed. If not flagged, the variables are treated individually and removed if they fall below the selected sparsity threshold.

Monotonicity

A *monotonic* variable is one that increases without bound. Monotonicity can also exist in the relationship between variables in which as one variable increases, the other does not decrease but remains constant, or also increases. At the moment, while discussing variable preparation, it is the monotonic variable itself that is being considered, not a monotonic relationship.

Monotonic variables are very common. Any variable that is linked to the passage of time, such as date, is a monotonic variable. The date always increases. Other variables not directly related to time are also monotonic. Social security numbers, record numbers, invoice numbers, employee numbers, and many, many other such indicators are monotonic. The range of such categorical or nominal values increases without bound.

The problem here is that they almost always have to be transformed into some nonmonotonic form if they are to be used in mining. Unless it is certain that every possible value of the monotonic variable that will be used is included in the data set, transformation is required. Transformation is needed because only some limited part of the full range of values can possibly be included in any data set. Any other data set, specifically the execution data set, will contain values of the monotonic variable that were not in the training data set. Any model will have no reference for predicting, or inferring, the meaning of the values outside its training range. Since the mined model will not have been exposed to such values, predictions or inferences based on such a model will at best be suspect.

There are a number of transformations that can be made to monotonic variables, depending on their nature. Datestamps, for instance, are often turned into seasonality information in which the seasons follow each other consecutively. Another transformation is to treat the information as a time series. Time series are treated in several ways that limit the nature of the monotonicity, say, by comparing “now” to some fixed distance of time in the past. Unfortunately, each type of monotonic variable requires specific transformations tailored to best glean information from it. Employee numbers will no doubt need to be treated differently from airline passenger ticket numbers, and those again from insurance policy numbers, and again from vehicle registration numbers. Each of these is monotonic and requires modification if they are to be of value in mining.

It is very hard to detect a monotonic variable in a sample of data, but certain detectable characteristics point to the possibility that a variable is in fact monotonic. Two measures that have proved useful in giving some indication of monotonicity in a variable (described in Chapter 5) are interstitial linearity and rate of detection. *Interstitial linearity* measures the uniformity of spacing between the sampled values, which tends to be more uniform in a monotonic variable than in some nonmonotonic ones. Rate of discovery measures the rate at which new values are experienced during random sampling of the data set. Rate of detection tends to remain uniform for monotonic variables during the whole sampling period and falls off for some nonmonotonic variables.

A problem with these metrics is that there are nonmonotonic variables that also share the characteristics that are used to detect potential monotonicity. Nonetheless, used as warning flags that the variables indicated need looking at more closely for monotonicity or other problems, the metrics are very useful. As noted, automatically modifying the variables into some different form is not possible.

Increasing Dimensionality

The usual problem in mining large data sets is in reducing the dimensionality. There are some circumstances where the dimensionality of a variable needs to be increased. One concern is to increase the dimensionality as much as is needed, but only as little as necessary, by recoding and remapping variables. Chapter 7 deals in part with these

techniques. The types of variables requiring this transformation, which are almost always categorical, carry information that is best exposed in more than one dimension. A couple of examples illustrate the point.

Colors can be represented in a variety of ways. Certainly a categorical listing covers the range of humanly appreciated color through a multitude of shades. Equally well, for some purposes, the spectral frequency might be listed. However, color has been usefully mapped onto a color wheel. Such a wheel not only carries color information, but also describes color as a continuum, carrying information about what other colors are near and far from some selected category. This is very useful information. Since a circle can be drawn on a plane, such as a piece of paper, it is easy to see that any point on the circle's circumference can be unambiguously represented by two coordinates, or numbers. Mapping the color wheel onto a circle on a graph and using the two coordinates for some selected color as the instance values of two variables may form a better description of color than a categorical listing.

ZIP codes form a perennial problem in mining. Sometimes, depending on the application, it is beneficial to translate the ZIP code from the categorical list into latitude and longitude. These values translate the ZIP code into two instance values. The single variable "ZIP" translates into two variables, say, "Lat" and "Lon."

Once again, the decision of whether to expand the dimensionality of a variable must be, in many cases, left up to the miner or domain expert.

Outliers

An *outlier* is a single, or very low frequency, occurrence of the value of a variable that is far away from the bulk of the values of the variable. The question miners always ask is: "Is this a mistake?" As a general rule of thumb, if it can be established that it is a mistake, it can be rectified. (One way to do this, if the correct value cannot be found, is to treat it as a missing value, discussed later in this chapter.) The problem is what to do if it cannot be pinpointed as an error. It is a problem because, for some modeling methods in particular (some types of neural network, for instance), outliers may distort the remaining data to the point of uselessness. Figure 2.5(a) shows this sort of situation.

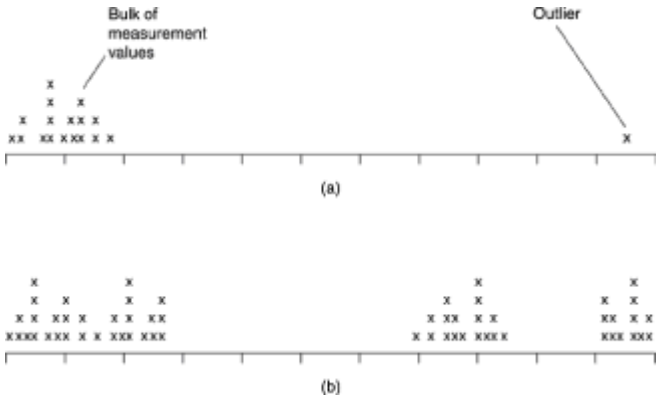

Figure 2.5 Examples of outliers: as an individual value (a) and as clumps of values (b).

Insurance data typically suffers considerably from the problem of outliers. Most insurance claims are small, but occasionally one comes in for some enormous sum. This is no error, and it must be included in modeling. How to do this without distorting the remaining data is a problem.

There is also a problem when the outliers are not individual values but clumps of values, illustrated in Figure 2.5(b). It's actually the gaps between the clumps that can pose problems. Are these clumps, perhaps, valid measurements from differently biased instruments? Once again, it must be determined first that there is not some sort of error. Maybe some measurements were made against an incorrect calibration and are biased. However, again it might not be possible to determine that an error occurred. In general, the miner is constrained to consider that the measurements are not an error until and unless it is possible to definitely show that they are.

If indeed the outlying value is not a mistake, or is at least going to be dealt with as if it is not, how is it to be treated? Fortunately there is a way of automatically dealing with the problem if it is not a mistake. This involves remapping the variable's values. Part of Chapter 7 deals with this remapping.

Numerating Categorical Values

Dealing correctly with categorical values is one of the most important functions of data preparation. For many modeling techniques it is necessary to translate categorical values into numbers: they simply cannot deal with untranslated categorical values. Experience shows that even modeling techniques that can deal well with untranslated categorical values benefit from a valid numeration of categoricals.

However, a na‹ve way of making the translation, one that is very commonly done, is terribly destructive of information. Simply assigning numbers to the nominals to create a

numbered list is a disastrous way to proceed! To see why, consider the variable “marital status,” for instance, which might be measured as married, single, widowed, divorced, or never married. To simply assign the numbers 1, 2, 3, 4, and 5 to these is totally destructive of the natural structure of the data. If it turned out, for instance, that the natural order of this variable, when translated, was in fact (on a scale of 0–1)

Never married	0
Single	0.1
Divorced	0.15
Widowed	0.65
Married	1

then the “brute force” assignment of numbers from 1–5 not only destroyed the natural ordering of these measures, but even if they were in the right order, it would have destroyed the interval information. Interval information is contained in the distance between the numbers and may be a significant factor in modeling. Except by pure, unadulterated luck, all of the structure contained in this variable would have been destroyed. Worse than that, some meaningless artificial structure has been forced into the data quite arbitrarily!

But what is the “natural order”? The natural order can be found embedded in the system of variables. Recall that the data set reflects to some degree the system of the world. As such, the data set itself forms a system. Thus, embedded in the data set is a structure that reflects an appropriate ordering and distance for categorical values. Assigning values in accord with the system embedded in the data reveals the natural ordering. Arbitrary assignment not only destroys the order, and any information carried by the variable, but actually introduces an artificial pattern to the data.

It is hard to imagine how more damage can be done to the natural ordering of the data than by arbitrary number assignment to categoricals. If it is not intuitively clear how damaging this might be, imagine working for a company that pays you for set periods of a half-day, a day, a half-week, a week, and a month. Perhaps the rate of pay for these periods in dollars might be

Time period	**Rate of pay ($)**
Half-day	100

Day	200
Half-week	500
Week	1000
Half-month	2000
Month	4000

This gives a natural order to these measures. Now, not knowing the actual numerical values, in building a model the modeler lists these periods alphabetically for convenience and assigns numbers to them:

Day	1
Half-day	2
Half-month	3
Half-week	4
Month	5
Week	6

Would you expect this ranking to accurately reflect anything significant about the categories? Compare the relationship between the arbitrary ordering and the monetary value:

Day	1	200
Half-day	2	100
Half-month	3	2000
Half-week	4	500
Month	5	4000
Week	6	1000

It is clear that the natural order of these ordinal values has been completely destroyed. It would, for instance, be impossible to use the arbitrary value assigned to predict how much is earned in each period. Thus, the arbitrary assignment has destroyed the information carried in the ordinal labels.

Regardless of what arbitrary order is given to the measures—whether it be alphabetic, reverse alphabetic, random selection, or just the order they are encountered in the data set—the arbitrary assignment of values destroys information content at best. *At worst it introduces and creates patterns in the data that are not natural and that reflect throughout the data set, wreaking havoc with the modeling process.*

A data set reflects the real world to some extent. (If not, any model will be useless anyway!) Any variables ordinal or higher in information content are, therefore, in an appropriate ordering to some extent. The variables in a data set, because they form an interlocking system, all have specific relationships to each other. It is quite easy, at least for a computer, to reflect the ordering from the ordinal, or higher variables, into the nominal and categorical variables, thus recovering the natural ordering. This process is not perfect. If domain knowledge is available for a more appropriate ordering, this is preferable. Domain expertise reflects far more knowledge of the real world than is enfolded in any data set for mining! Most often, such domain knowledge is unavailable or unknown. Using the information at hand can help enormously.

So, natural orderings can be recovered, at least to some extent, by looking at the data. In a data set that had the pay periods listed in the above tables as categoricals, but without numeric values, it is usually possible to recover, at least to some degree, the natural order and spacing of the measures. In the event that full recovery cannot be made, it is still possible to assign a ranking and position that turn out to be neutral; that is, even if they don't contribute much information, they at least do not distort the data. One of the key principles in data preparation is to do as little damage as possible to the natural structure in a data set.

Sometimes a nominal variable will fairly easily translate into a single numeric variable. This allows translation of the nominal or categorical, one for one, into a numeric value for modeling. This could have been done in the pay-period example described above if it was possible to recover the value and spacing information. By simply inserting the recovered value, a numeric variable replaces the nominal, one for one, when modeling.

Also note that sometimes a categorical value needs to be expanded into more than one numeric value for reasons similar to those mentioned above during the discussion of increasing the dimensionality of variables. Fortunately, discovering an appropriate numeration of categorical values can be completely automated. Chapter 6 includes a detailed discussion of the technique.

Anachronisms

An *anachronism* is, literally, something out of place in time. Temporal displacement. When mining, an *anachronistic variable* is one that creeps into the variables to be modeled, but that contains information not actually available in the data when a prediction is needed.

For example, in mining a data set to predict people who will take a money market account with a bank, various fields of interest will be set up, one entitled “investor.” This could be a binary field with a “1” indicating people who opened a money market account, and a “0” for the others. Obviously, this is a field to predict. The data set might also include a field entitled “account number” filled in with the issued account number. So far, so good. However, if “account number” is included in the predicting variables, since there is only an account number when the money market account has been opened, it is clearly anachronistic—information not available until after the state of the field to be predicted is known. (Such a model makes pretty good predictions, about 100% accurate—always a suspicious circumstance!)

“Account number” is a fairly straightforward example, but is based on a real occurrence. Easy to spot with hindsight, but when the model has 400–500 variables, it is easy to miss one. Other forms of “leakage” of after-the-fact information can easily happen. It can sometimes be hard to find where the leakage is coming from in a large model. In one telephone company churn application, the variables did not seem to be at all anachronistic. However, the models seemed to be too good to be believed. In order to get information about their customers, the phone company had built a database accumulated over time based on customer interviews. One field was a key that identified which interviewer had conducted the interview. It turned out that some interviewers were conducting general interviews, and others were conducting interviews after the customer had left, or churned. In fact, the interviewer code was capturing information about who had churned! Obviously an anachronistic variable, but subtle, and in this case hard to find.

One of the best rules of thumb is that if the results seem to good to be true, they probably are. Anachronistic variables simply have to be removed.

2.6.3 Building Mineable Data Sets

Looking at data sets involves considering the relationships between variables. There is also a natural structure to the interrelationships between variables that is just as critical to maintain as it is within variables. Mining tools work on exploring the interactions, or relationships, that exist between the collected variables. Unfortunately, simply preparing the variables does not leave us with a fully prepared data set. Two separate areas need to be looked at in the data set: exposing the information content and getting enough data.

A first objective in preparing the data set is to make things as easy as possible for the mining tool. It is to prepare the data in such a way that the information content is best

revealed for the tool to see. Why is it important to make the mining tools' job easier? Actually, there are important reasons. A brief discussion follows in the next section.

Some types of relationships cause problems for modeling tools. A second objective in preparing the data set, then, is to obviate the problems where possible. We will look briefly at some of those. If it is possible to detect such potentially damaging relationships, even without being able to ameliorate them automatically, that is still very useful information. The miner may be able to take corrective or remedial action, or at least be aware of the problem and make due allowance for it. If there is some automatic action that can correct the problem, so much the better.

Exposing the Information Content

Since the information is enfolded in the data, why not let the mining tool find it?

One reason is time. Some data sets contain very complex, involved relationships. Often, these complex relationships are known beforehand. Suppose in trying to predict stock market performance it is believed that the "trend" of the market is important. If indeed that is the case, and the data is presented to a suitable modeling tool in an appropriate way, the tool will no doubt develop a "trend detector." Think, for a moment, of the complexity of calculation involved in creating a trend measurement.

A simple measurement of trend might be to determine that if the mean of the last three days' closing prices is higher than the mean of the previous three days' prices, the trend is "up." If the recent mean is lower than the older mean, the trend is "down." If the means are the same, the trend is "flat." Mathematically, such a relationship might be expressed as

$$t = \frac{p_{i-1} + p_{i-2} + p_{i-3}}{3} - \frac{p_{i-4} + p_{i-5} + p_{i-6}}{3}$$

where t is trend and p is closing price for day i. This is a modestly complex expression yielding a positive or negative number that can be interpreted as measuring trend. For a human it takes insight and understanding, plus a knowledge of addition, subtraction, multiplication, and division, to devise this measure. An automated learning tool can learn this. It takes time, and repeated attempts, but such relationships are not too hard. It may take a long time, however, especially if there are a large number of variables supplied to the mining tool. The tool has to explore all of the variables, and many possible relationships, before this one is discovered.

For this discussion we assumed that this relationship was in fact a meaningful one, and that after a while, a mining tool could discover it. But why should it? The relationship was already known, and it was known that it was a useful relationship. So the tool would have discovered a known fact. Apart from confirmation (which is often a valid and useful reason for mining), nothing new has yet been achieved. We could have started from this point,

not worked to get here. Giving the tool this relationship to begin with would have sped up the process, perhaps very much. The more complex the relationship, the more the speed is improved.

However, a second reason for providing as much help as possible to the tool is much more important for the end result, but directly related to the time factor. The reason is noise. The longer that training continues, the more likely it is that noise will be learned along with the underlying pattern. In the training set, the noisy relationship is every bit as real as any other. The tool cannot discriminate, inside the training set, between the noise and target patterns.

The relationships in data are known as *features* of the data. The trend that, for this example, is assumed to be a valid relationship is called a *predictive* feature. Naturally, it's desirable for the tool to learn all of the valid predictive features (or *inferential* features if it is an inferential model that is needed) without learning noise features. However, as training continues it is quite possible that the tool learns the noise and thereby misses some other feature. This obscuring of one feature by another is called *feature swamping*.

By including relevant domain knowledge, the mining tool is able to spend its time looking for other features enfolded in the data, and not busy itself rediscovering already known relationships. In fact, there is a modeling technique that involves building the best model prior to overfitting, taking a new data set, using the model to make predictions, and feeding the predictions plus new training data into another round of mining. This is done precisely to give the second pass with the tool a "leg up" so that it can spend its time looking for new features, not learning old ones.

In summary, exposing the information content is done partly to speed the modeling process, but also to avoid feature swamping. Searching for meaningful fine structure involves removing the coarser structure. In other words, if you want to find gold dust, move the rocks out of the way first!

Getting Enough Data

The discussion about preparing variables started with getting sufficient data to be sure that there were enough instance values to represent the variable's actual features. The same is true for data sets. Unfortunately, getting enough of each variable to ensure that it is representative does not also assure that a representative sample of the data set has been captured. Why? Because now we're interested in the interactions between variables, not just the pattern existing within a variable.

Figure 2.6 explains why there is a difference. Consider two variables, instance values of one of them plotted on the vertical axis, and the other on the horizontal axis. The marks on the axes indicate the range of the individual variables. In addition to distributing the individual values on the axes, there is a joint range of values that is shown by the ellipse.

This ellipse shows for this hypothetical example where the actual real-world values might fall. High values of variable 1 always correspond with low values of variable 2, and vice versa. It is quite possible to select joint values that fall only in some restricted part of the joint distribution, and yet still cover the full range of the individual variables. One way in which this could occur is shown in the shaded part of the ellipse. If joint values were selected that fell only inside the shaded area, it would be possible to have the full range of each variable covered and yet only cover part of the joint distribution. In fact, in the example, half of the joint distribution range is not covered at all. The actual method used to select the instance values means that there is only a minute chance that the situation used for the illustration would ever occur. However, it is very possible that simply having representative distributions for individual variables will not produce a fully representative joint distribution for the data set. In order to assure complete coverage of the joint distribution, every possible combination of variables has to be checked, and that can become impossible very quickly indeed!

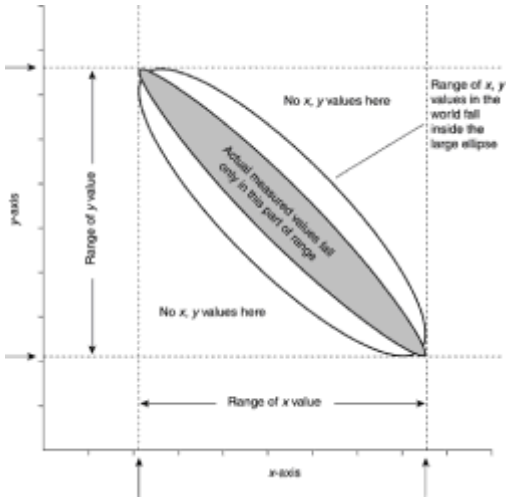

Figure 2.6 Joint occurrence of two variables may not cover the individual range of each. Values falling in only part of the full range, illustrated by half of the ellipse, may cover the full range of each variable, but not the full joint range.

The Combinatorial Explosion

With five variables, say, the possible combinations are shown in Figure 2.7. You can see that the total number of combinations is determined by taking the five variables two at a time, then three at a time, then four at a time. So, for any number of variables, the number of combinations is the sum of all combinations from two to the total number of variables. This number gets very large, very quickly! Table 2.2 shows just how quickly.

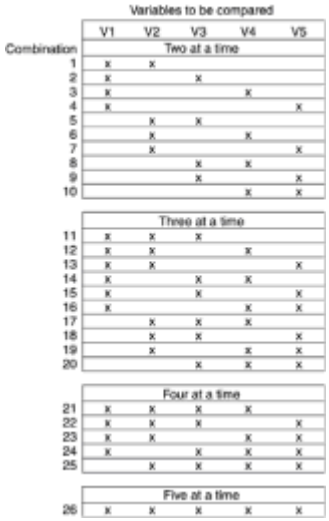

Figure 2.7 Combinations of five variables compared against each other, from two at a time andÿ20increasing to five at a time.

TABLE 2.2 The combinatorial explosion.

Number of variables	Number of combinations
5	26
7	120
9	502
20	1,048,555
25	33554406

This “blowup” in the number of combinations to consider is known as the *combinatorial explosion* and can very quickly defeat any computer, no matter how fast or powerful. (Calculating combinations is briefly described in the Supplemental Material section at the end of this chapter.) Because there is no practical way to check for every combination that the intervariable variability has been captured for the data set, some other method of estimating (technical talk for guessing!) if the variability has been captured needs to be

used. After all, some estimate of variability capture is needed. Without such a measure, there is no way to be certain how much data is needed to build a model.

The expression of certainty is the key here and is an issue that is mentioned in different contexts many times in this book. While it may not be possible to have 100% confidence that the variability has been captured, it reduces the computational work enormously if some lesser degree of confidence is acceptable. Reducing the demanded confidence from 100% to 99%, depending on the number of variables, often changes the task from impossible to possible but time-consuming. If 98% or 95% confidence is acceptable, the estimating task usually becomes quite tractable. While confidence measures are used throughout the preparation process, their justification and use are discussed in Chapter 5. Chapter 10 includes a discussion on capturing the joint variability of multiple variables.

Missing and Empty Values

As you may recall, the difference between “empty” and “missing” is that the first has no corresponding real-world value, while the second has an underlying value that was not captured. Determining if any particular value is empty rather than missing requires domain knowledge and cannot be automatically detected. If possible, the miner should differentiate between the two in the data set. Since it is impossible to automatically differentiate between missing and empty, if the miner cannot provide discriminating information, it is perforce necessary to deal with all missing values in a similar way. In this discussion, they will all be referred to as missing.

Some mining tools use techniques that do not require the replacement of missing values. Some are able to simply ignore the missing value itself, where others have to ignore the instance (record) altogether. Other tools cannot deal with missing values at all, and have to have some default replacement for the missing value. Default replacement techniques are often damaging to the structure of the data set. The discussion on numerating categorical values discusses how arbitrary value replacement can damage information content.

The general problem with missing values is twofold. First, there may be some information content, predictive or inferential, carried by the actual pattern of measurements missing. For example, a credit application may carry useful information in noting which fields the applicant did not complete. This information needs to be retained in the data set.

The second problem is in creating and inserting some replacement value for the missing value. The objective is to insert a value that neither adds nor subtracts information from the data set. It must introduce no bias. But if it introduces no new information, why do it?

First, default replacement methods often do introduce bias. If not correctly determined, a poorly chosen value adds information to the data set that is not really present in the world, thus distorting the data. Adding noise and bias of this sort is always detrimental to

modeling. If a suitable value can be substituted for the missing values, it prevents the distortion introduced by poorly chosen defaults.

Second, for those modeling tools that have to ignore the whole instance when one of the values is missing, plugging the holes allows the instance to be used. That instance may carry important information in the values that are present, and by plugging the holes that information is made available to the modeling tool.

There are several methods that can be used to determine information-neutral values to replace the missing values. Chapter 8 discusses the issues and techniques used. All of them involve caveats and require knowledgeable care in use.

Although the values of individual variables are missing, this is an issue in preparing the data set since it is only by looking at how the variable behaves vis-…-vis the other variables when it is present that an appropriate value can be determined to plug in when it is missing. Of course, this involves making a prediction, but in a very careful way such that no distortion is introduced—at least insofar as that is possible.

The Shape of the Data Set

The question of the shape of the data set is not a metaphorical one. To understand why, we have to introduce the concept of *state space*.

State space can be imagined to be a space like any other—up to a point. It is called state space because of the nature of the instances of data. Recall that each instance captures a number of measurements, one per variable, that were measured under some validating circumstance. An instance, then, represents the state of the object at validation. That is where the “state” part of the phrase comes from. It is a space that reflects the various states of the system as measured and captured in the instance values.

That’s fine, but where does “space” come from? Figure 2.6, used earlier to discuss the variability of two variables, shows a graphical representation of them. One variable is plotted on one axis, and the other variable is plotted on the other axis. The values of the combined states of the two variables can easily be plotted as a single point on the graph. One point represents both values simultaneously. If there were three variables’ values, they could be plotted on a three-dimensional graph, perhaps like the one shown in Figure 2.8. Of course, this three-dimensional object looks like something that might exist in the world. So the two- and three-dimensional representations of the values of variables can be thought of as determining points in some sort of space. And indeed they do—in state space.