1 *Introduction*

1.1 Introduction to Data Mining

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavor, from the mundane (such as supermarket transaction data, credit card usage records, telephone call details, and government statistics) to the more exotic (such as images of astronomical bodies, molecular databases, and medical records). Little wonder, then, that interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database. The discipline concerned with this task has become known as *data mining*.

Defining a scientific discipline is always a controversial task; researchers often disagree about the precise range and limits of their field of study. Bearing this in mind, and accepting that others might disagree about the details, we shall adopt as our working definition of data mining:

> Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

The relationships and summaries derived through a data mining exercise are often referred to as *models* or *patterns*. Examples include linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series.

The definition above refers to "observational data," as opposed to "experimental data." Data mining typically deals with data that have already been collected for some purpose other than the data mining analysis (for example, they may have been collected in order to maintain an up-to-date record of all the transactions in a bank). This means that the objectives of the data mining exercise play no role in the data collection strategy. This is one way in which data mining differs from much of statistics, in which data are often collected by using efficient strategies to answer specific questions. For this reason, data mining is often referred to as “secondary” data analysis.

The definition also mentions that the data sets examined in data mining are often large. If only small data sets were involved, we would merely be discussing classical exploratory data analysis as practiced by statisticians. When we are faced with large bodies of data, new problems arise. Some of these relate to housekeeping issues of how to store or access the data, but others relate to more fundamental issues, such as how to determine the representativeness of the data, how to analyze the data in a reasonable period of time, and how to decide whether an apparent relationship is merely a chance occurrence not reflecting any underlying reality. Often the available data comprise only a sample from the complete population (or, perhaps, from a hypothetical superpopulation); the aim may be to *generalize* from the sample to the population. For example, we might wish to predict how future customers are likely to behave or to determine the properties of protein structures that we have not yet seen. Such generalizations may not be achievable through standard statistical approaches because often the data are not (classical statistical) “random samples,” but rather “convenience” or “opportunity” samples. Sometimes we may want to summarize or *compress* a very large data set in such a way that the result is more comprehensible, without any notion of generalization. This issue would arise, for example, if we had complete census data for a particular country or a database recording millions of individual retail transactions.

The relationships and structures found within a set of data must, of course, be novel. There is little point in regurgitating well-established relationships (unless, the exercise is aimed at “hypothesis” confirmation, in which one was seeking to determine whether established pattern also exists in a new data set) or necessary relationships (that, for example, all pregnant patients are female). Clearly, novelty must be measured relative to the user’s prior knowledge. Unfortunately few data mining algorithms take into account a user’s prior knowledge. For this reason we will not say very much about novelty in this text. It remains an open research problem.

While novelty is an important property of the relationships we seek, it is not sufficient to qualify a relationship as being worth finding. In particular, the relationships must also be understandable. For instance simple relationships are more readily understood than complicated ones, and may well be preferred, all else being equal.

Data mining is often set in the broader context of *knowledge discovery in databases,* or KDD. This term originated in the artificial intelligence (AI) research field. The KDD process involves several stages: selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures. Once again the precise boundaries of the data mining part of the process are not easy to state; for example, to many people data transformation is an intrinsic part of data mining. In this text we will focus primarily on data mining algorithms rather than the overall process. For example, we will not spend much time discussing data preprocessing issues such as data cleaning, data verification, and defining variables. Instead we focus on the basic principles for modeling data and for constructing algorithmic processes to fit these models to data.

The process of seeking relationships within a data set— of seeking accurate, convenient, and useful summary representations of some aspect of the data—involves a number of steps:

- determining the nature and structure of the representation to be used;
- deciding how to quantify and compare how well different representations fit the data (that is, choosing a "score" function);
- choosing an algorithmic process to optimize the score function; and
- deciding what principles of data management are required to implement the algorithms efficiently.

The goal of this text is to discuss these issues in a systematic and detailed manner. We will look at both the fundamental principles (chapters 2 to 8) and the ways these principles can be applied to construct and evaluate specific data mining algorithms (chapters 9 to 14).

Example 1.1 Regression analysis is a tool with which many readers will be familiar. In its simplest form, it involves building a predictive model to relate a *predictor* variable, X, to a *response* variable, Y, through a relationship of the form $Y = aX + b$. For example, we might build a model which would allow us to predict a person's annual credit-card spending given their annual income. Clearly the model would not be perfect, but since spending typically increases with income, the model might well be adequate as a rough characterization. In terms of the above steps listed, we would have the following scenario:

- The representation is a model in which the response variable, spending, is linearly related to the predictor variable, income.

- The score function most commonly used in this situation is the sum of squared discrepancies between the predicted spending from the model and observed spending in the group of people described by the data. The smaller this sum is, the better the model fits the data.
- The optimization algorithm is quite simple in the case of linear regression: a and b can be expressed as explicit functions of the observed values of spending and income. We describe the algebraic details in chapter 11.
- Unless the data set is very large, few data management problems arise with regression algorithms. Simple summaries of the data (the sums, sums of squares, and sums of products of the X and Y values) are sufficient to compute estimates of a and b. This means that a single pass through the data will yield estimates.

Data mining is an interdisciplinary exercise. Statistics, database technology, machine learning, pattern recognition, artificial intelligence, and visualization, all play a role. And just as it is difficult to define sharp boundaries between these disciplines, so it is difficult to define sharp boundaries between each of them and data mining. At the boundaries, one person's data mining is another's statistics, database, or machine learning problem.

1.2 The Nature of Data Sets

We begin by discussing at a high level the basic nature of data sets.

A *data set* is a set of measurements taken from some environment or process. In the simplest case, we have a collection of objects, and for each object we have a set of the same p measurements. In this case, we can think of the collection of the measurements on n objects as a form of $n \times p$ *data matrix*. The n rows represent the n objects on which measurements were taken (for example, medical patients, credit card customers, or individual objects observed in the night sky, such as stars and galaxies). Such rows may be referred to as *individuals*, *entities*, *cases*, *objects*, or *records* depending on the context.

The other dimension of our data matrix contains the set of p measurements made on each object. Typically we assume that the same p measurements are made on each individual although this need not be the case (for example, different medical tests could be performed on different patients). The p columns of the data matrix may be referred to as *variables*, *features*, *attributes*, or *fields*; again, the language depends on the research context. In all situations the idea is the same: these names refer to the measurement that is represented by each column. In chapter 2 we will discuss the notion of measurement in much more detail.

ID	Age	Sex	Marital Status	Education	Income
248	54	Male	Married	High school graduate	100000
249	??	Female	Married	High school graduate	12000
250	29	Male	Married	Some college	23000
251	9	Male	Not married	Child	0
252	85	Female	Not married	High school graduate	19798
253	40	Male	Married	High school graduate	40100
254	38	Female	Not married	Less than 1st grade	2691
255	7	Male	??	Child	0
256	49	Male	Married	11th grade	30000
257	76	Male	Married	Doctorate degree	30686

Table 1.1 Examples of data in Public Use Microdata Sample data sets.

Example 1.2 The U.S. Census Bureau collects information about the U.S. population every 10 years. Some of this information is made available for public use, once information that could be used to identify a particular individual has been removed. These data sets are called PUMS, for Public Use Microdata Samples, and they are available in 5 % and 1 % sample sizes. Note that even a 1 % sample of the U.S. population contains about 2.7 million records. Such a data set can contain tens of variables, such as the age of the person, gross income, occupation, capital gains and losses, education level, and so on. Consider the simple data matrix shown in table 1.1. Note that the data contains different types of variables, some with continuous values and some with categorical. Note also that some values are *missing*—for example, the Age of person 249, and the Marital Status of person 255. Missing measurements are very common in large real-world data sets. A more insidious problem is that of measurement noise. For example, is person 248's income really \$100,000 or is this just a rough guess on his part?

A typical task for this type of data would be finding relationships between different variables. For example, we might want to see how well a person's income could be predicted from the other variables. We might also be interested in seeing if there are naturally distinct groups of people, or in finding values at which variables often coincide. A subset of variables and records is available online at the Machine Learning Repository of the University of California, Irvine , `www.ics.uci.edu/~mlearn/MLSummary.html`.

Data come in many forms and this is not the place to develop a complete taxonomy. Indeed, it is not even clear that a complete taxonomy can be developed, since an important aspect of data in one situation may be unimportant in another. However there are certain basic distinctions to which we should draw attention. One is the difference between quantitative and categorical measurements (different names are sometimes used for these). A quantitative variable is measured on a numerical scale and can, at least in principle, take any value. The columns Age and Income in table 1.1 are examples of quantitative variables. In contrast, categorical variables such as Sex, Marital Status and Education in 1.1 can take only certain, discrete values. The common three point severity scale used in medicine (mild, moderate, severe) is another example. Categorical variables may be ordinal (possessing a natural order, as in the Education scale) or nominal (simply naming the categories, as in the Marital Status case). A data analytic technique appropriate for one type of scale might not be appropriate for another (although it does depend on the objective—see Hand (1996) for a detailed discussion). For example, were marital status represented by integers (e.g., 1 for single, 2 for married, 3 for widowed, and so forth) it would generally not be meaningful or appropriate to calculate the arithmetic mean of a sample of such scores using this scale. Similarly, simple linear regression (predicting one quantitative variable as a function of others) will usually be appropriate to apply to quantitative data, but applying it to categorical data may not be wise; other techniques, that have similar objectives (to the extent that the objectives can be similar when the data types differ), might be more appropriate with categorical scales.

Measurement scales, however defined, lie at the bottom of any data taxonomy. Moving up the taxonomy, we find that data can occur in various relationships and structures. Data may arise sequentially in time series, and the data mining exercise might address entire time series or particular segments of those time series. Data might also describe spatial relationships, so that individual records take on their full significance only when considered in the context of others.

Consider a data set on medical patients. It might include multiple measurements on the same variable (e.g., blood pressure), each measurement taken at different times on different days. Some patients might have extensive image data (e.g., X-rays or magnetic resonance images), others not. One might also have data in the form of *text*, recording a specialist's comments and diagnosis for each patient. In addition, there might be a hierarchy of relationships between patients in terms of doctors, hospitals, and geographic locations. The more complex the data structures, the more complex the data mining models, algorithms, and tools we need to apply.

For all of the reasons discussed above, the $n \times p$ data matrix is often an oversimplification or idealization of what occurs in practice. Many data sets will not fit into this simple format. While much information can in principle be "flattened" into the $n \times p$ matrix (by suitable definition of the p variables), this will often lose much of the structure embedded in the data. Nonetheless, when discussing the underlying principles of data analysis, it is often very convenient to assume that the observed data exist in an $n \times p$ data matrix; and we will do so unless otherwise indicated, keeping in mind that for data mining applications n and p may both be very large. It is perhaps worth remarking that the observed data matrix can also be referred to by a variety names including *data set, training data, sample, database,* (often the different terms arise from different disciplines).

Example 1.3 Text documents are important sources of information, and data mining methods can help in retrieving useful text from large collections of documents (such as the Web). Each document can be viewed as a sequence of words and punctuation. Typical tasks for mining text databases are classifying documents into predefined categories, clustering similar documents together, and finding documents that match the specifications of a query. A typical collection of documents is "Reuters-21578, Distribution 1.0," located at `http://www.research.att.com/~lewis`. Each document in this collection is a short newswire article.

A collection of text documents can also be viewed as a matrix, in which the rows represent documents and the columns represent words. The entry (d, w), corresponding to document d and word w, can be the number of times w occurs in d, or simply 1 if w occurs in d and 0 otherwise.

With this approach we lose the ordering of the words in the document (and, thus, much of the semantic content), but still retain a reasonably good representation of the document's contents. For a document collection, the number of rows is the number of documents, and the number of columns is the number of distinct words. Thus, large multilingual document collections may have millions of rows and hundreds of thousands of columns. Note that such a data matrix will be very sparse; that is, most of the entries will be zeroes. We discuss text data in more detail in chapter 14.

Example 1.4 Another common type of data is *transaction data,* such as a list of purchases in a store, where each purchase (or transaction) is described by the date, the customer ID, and a list of items and their prices. A similar example is a Web transaction log, in which a sequence of triples (user id, web page, time), denote the user accessing a particular page at a particular time. Designers and owners of Web sites often have great interest in understanding the patterns of how people navigate through their site.

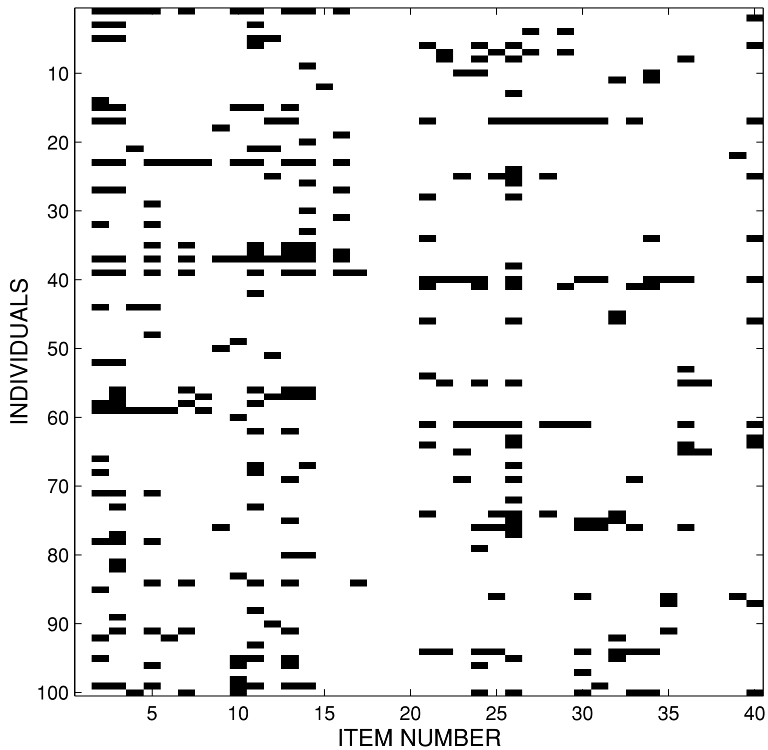

Figure 1.1 A portion of a retail transaction data set displayed as a binary image, with 100 individual customers (rows) and 40 categories of items (columns).

As with text documents, we can transform a set of transaction data into matrix form. Imagine a very large, sparse matrix in which each row corresponds to a particular individual and each column corresponds to a particular Web page or item. The entries in this matrix could be binary (e.g., indicating whether a user had ever visited a certain Web page) or integer-valued (e.g., indicating how many times a user had visited the page).

Figure 1.1 shows a visual representation of a small portion of a large retail transaction data set displayed in matrix form. Rows correspond to individual customers and columns represent categories of items. Each black entry indicates that the customer corresponding to that row purchased the item corresponding to that column. We can see some obvious patterns even in this simple display. For example, there is considerable variability in terms of which categories of items customers purchased and how many items they purchased. In addition, while some categories were purchased by quite a few customers (e.g., columns 3, 5, 11, 26) some were not purchased at all (e.g., columns 18 and 19). We can also see pairs of categories which that were frequently purchased together (e.g., columns 2 and 3).

Note, however, that with this "flat representation" we may lose a significant portion of information including sequential and temporal information (e.g., in what order and at what times items were purchased), any information about structured relationships between individual items (such as product category hierarchies, links between Web pages, and so forth). Nonetheless, it is often useful to think of such data in a standard $n \times p$ matrix. For example, this allows us to define distances between users by comparing their p-dimensional Web-page usage vectors, which in turn allows us to cluster users based on Web page patterns. We will look at clustering in much more detail in chapter 9.

1.3 Types of Structure: Models and Patterns

The different kinds of representations sought during a data mining exercise may be characterized in various ways. One such characterization is the distinction between a global *model* and a local *pattern*.

A *model structure,* as defined here, is a *global* summary of a data set; it makes statements about any point in the full measurement space. Geometrically, if we consider the rows of the data matrix as corresponding to p-dimensional vectors (i.e., points in p-dimensional space), the model can make a statement about any point in this space (and hence, any object). For example, it can assign a point to a cluster or predict the value of some other variable. Even when some of the measurements are missing (i.e., some of the components of the p-dimensional vector are unknown), a model can typically make some statement about the object represented by the (incomplete) vector.

A simple model might take the form $Y = aX + c$, where Y and X are variables and a and c are parameters of the model (constants determined during the course of the data mining exercise). Here we would say that the functional form of the model is *linear,* since Y is a linear function of X. The conventional statistical use of the term is slightly different. In statistics, a model is linear if it is a linear function of the *parameters*. We will try to be clear in the text about which form of linearity we are assuming, but when we discuss the *structure* of a model (as we are doing here) it makes sense to consider linearity as a function of the variables of interest rather than the parameters. Thus, for example, the model structure $Y = aX^2 + bX + c$, is considered a linear model in classic statistical terminology, but the functional form of the model relating Y and X is *nonlinear* (it is a second-degree polynomial).

In contrast to the global nature of models, *pattern structures* make statements only about restricted regions of the space spanned by the variables. An example is a simple probabilistic statement of the form `if` $X > x_1$ `then prob`$(Y > y_1) = p1$. This structure consists of *constraints* on the values of the variables X and Y, related in the form of a probabilistic rule. Alternatively we could describe the relationship as the conditional probability $p(Y > y_1|X > x_1) = p1$, which is semantically equivalent. Or we might notice that certain classes of transaction records do not show the peaks and troughs shown by the vast majority, and look more closely to see why. (This sort of exercise led one bank to discover that it had several open accounts that belonged to people who had died.)

Thus, in contrast to (global) models, a (local) pattern describes a structure relating to a relatively small part of the data or the space in which data could occur. Perhaps only some of the records behave in a certain way, and the pattern characterizes which they are. For example, a search through a database of mail order purchases may reveal that people who buy certain combinations of items are also likely to buy others. Or perhaps we identify a handful of "outlying" records that are very different from the majority (which might be thought of as a central cloud in p-dimensional space). This last example illustrates that global models and local patterns may sometimes be regarded as opposite sides of the same coin. In order to detect unusual behavior we need a description of usual behavior. There is a parallel here to the role of *diagnostics* in statistical analysis; local pattern-detection methods have applications in anomaly detection, such as fault detection in industrial processes, fraud detection in banking and other commercial operations.

Note that the model and pattern structures described above have parameters associated with them; a, b, c for the model and x_1, y_1 and $p1$ for the pattern. In general, once we have established the structural form we are interested in finding, the next step is to estimate its parameters from the available data. Procedures for doing this are discussed in detail in chapters 4, 7, and 8. Once the parameters have been assigned values, we refer to a particular model, such as $y = 3.2x + 2.8$, as a "fitted model," or just "model" for short (and similarly for patterns). This distinction between model (or pattern) structures and the actual (fitted) model (or pattern) is quite important. The structures represent the general functional forms of the models (or patterns), with unspecified parameter values. A fitted model or pattern has specific values for its parameters.

The distinction between models and patterns is useful in many situations. However, as with most divisions of nature into classes that are convenient for human comprehension, it is not hard and fast: sometimes it is not clear whether a particular structure should be regarded as a model or a pattern. In such cases, it is best not to be too concerned about which is appropriate; the distinction is intended to aid our discussion, not to be a proscriptive constraint.

1.4 Data Mining Tasks

It is convenient to categorize data mining into types of *tasks*, corresponding to different objectives for the person who is analyzing the data. The categorization below is not unique, and further division into finer tasks is possible, but it captures the types of data mining activities and previews the major types of data mining algorithms we will describe later in the text.

1. **Exploratory Data Analysis (EDA)** (chapter 3): As the name suggests, the goal here is simply to explore the data without any clear ideas of what we are looking for. Typically, EDA techniques are *interactive* and *visual*, and there are many effective graphical display methods for relatively small, low-dimensional data sets. As the dimensionality (number of variables, p) increases, it becomes much more difficult to visualize the cloud of points in p-space. For p higher than 3 or 4, projection techniques (such as principal components analysis) that produce informative low-dimensional projections of the data can be very useful. Large numbers of cases can be difficult to visualize effectively, however, and notions of scale and detail come into play: "lower resolution" data samples can be displayed or summarized at the cost of possibly missing important details. Some examples of EDA applications are:

    - Like a pie chart, a *coxcomb* plot divides up a circle, but whereas in a pie chart the angles of the wedges differ, in a coxcomb plot the radii of the wedges differ. Florence Nightingale used such plots to display the mortality rates at military hospitals in and near London (Nightingale, 1858).
    - In 1856 John Bennett Lawes laid out a series of plots of land at Rothamsted Experimental Station in the UK, and these plots have remained

untreated by fertilizers or other artificial means ever since. They provide a rich source of data on how different plant species develop and compete, when left uninfluenced. Principal components analysis has been used to display the data describing the relative yields of different species (Digby and Kempton, 1987, p. 59).

- More recently, Becker, Eick, and Wilks (1995) described a set of intricate spatial displays for visualization of time-varying long-distance telephone network patterns (over 12,000 links).

2. **Descriptive Modeling** (chapter 9): The goal of a descriptive model is describe all of the data (or the process generating the data). Examples of such descriptions include models for the overall probability distribution of the data (*density estimation*), partitioning of the p-dimensional space into groups (*cluster analysis and segmentation*), and models describing the relationship between variables (*dependency modeling*). In segmentation analysis, for example, the aim is to group together similar records, as in market segmentation of commercial databases. Here the goal is to split the records into homogeneous groups so that similar people (if the records refer to people) are put into the same group. This enables advertisers and marketers to efficiently direct their promotions to those most likely to respond. The number of groups here is chosen by the researcher; there is no "right" number. This contrasts with cluster analysis, in which the aim is to discover "natural" groups in data—in scientific databases, for example. Descriptive modelling has been used in a variety of ways.

   - Segmentation has been extensively and successfully used in marketing to divide customers into homogeneous groups based on purchasing patterns and demographic data such as age, income, and so forth (Wedel and Kamakura, 1998).
   - Cluster analysis has been used widely in psychiatric research to construct taxonomies of psychiatric illness. For example, Everitt, Gourlay and Kendell (1971) applied such methods to samples of psychiatric inpatients; they reported (among other findings) that "all four analyses produced a cluster composed mainly of patients with psychotic depression."
   - Clustering techniques have been used to analyze the long-term climate variability in the upper atmosphere of the Earth's Northern hemisphere. This variability is dominated by three recurring spatial pressure patterns (clusters) identified from data recorded daily since 1948

(see Cheng and Wallace [1993] and Smyth, Ide, and Ghil [1999] for further discussion).

3. **Predictive Modeling: Classification and Regression** (chapters 10 and 11): The aim here is to build a model that will permit the value of one variable to be predicted from the known values of other variables. In classification, the variable being predicted is categorical, while in regression the variable is quantitative. The term "prediction" is used here in a general sense, and no notion of a time continuum is implied. So, for example, while we might want to predict the value of the stock market at some future date, or which horse will win a race, we might also want to determine the diagnosis of a patient, or the degree of brittleness of a weld. A large number of methods have been developed in statistics and machine learning to tackle predictive modeling problems, and work in this area has led to significant theoretical advances and improved understanding of deep issues of inference. The key distinction between prediction and description is that prediction has as its objective a unique variable (the market's value, the disease class, the brittleness, etc.), while in descriptive problems no single variable is central to the model. Examples of predictive models include the following:

   - The SKICAT system of Fayyad, Djorgovski, and Weir (1996) used a tree-structured representation to learn a classification tree that can perform as well as human experts in classifying stars and galaxies from a 40-dimensional feature vector. The system is in routine use for automatically cataloging millions of stars and galaxies from digital images of the sky.
   - Researchers at AT&T developed a system that tracks the characteristics of all 350 million unique telephone numbers in the United States (Cortes and Pregibon, 1998). Regression techniques are used to build models that estimate the probability that a telephone number is located at a business or a residence.

4. **Discovering Patterns and Rules** (chapter 13): The three types of tasks listed above are concerned with model building. Other data mining applications are concerned with pattern detection. One example is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest. Another use is in astronomy, where detection of unusual stars

or galaxies may lead to the discovery of previously unknown phenomena. Yet another is the task of finding combinations of items that occur frequently in transaction databases (e.g., grocery products that are often purchased together). This problem has been the focus of much attention in data mining and has been addressed using algorithmic techniques based on *association rules*.

A significant challenge here, one that statisticians have traditionally dealt with in the context of outlier detection, is deciding what constitutes truly unusual behavior in the context of normal variability. In high dimensions, this can be particularly difficult. Background domain knowledge and human interpretation can be invaluable. Examples of data mining systems of pattern and rule discovery include the following:

- Professional basketball games in the United States are routinely annotated to provide a detailed log of every game, including time-stamped records of who took a particular type of shot, who scored, who passed to whom, and so on. The *Advanced Scout* system of Bhandari et al. (1997) searches for rule-like patterns from these logs to uncover interesting pieces of information which might otherwise go unnoticed by professional coaches (e.g., "When Player X is on the floor, Player Y's shot accuracy decreases from 75% to 30%.") As of 1997 the system was in use by several professional U.S. basketball teams.
- Fraudulent use of cellular telephones is estimated to cost the telephone industry several hundred million dollars per year in the United States. Fawcett and Provost (1997) described the application of rule-learning algorithms to discover characteristics of fraudulent behavior from a large database of customer transactions. The resulting system was reported to be more accurate than existing hand-crafted methods of fraud detection.

5. **Retrieval by Content** (chapter 14): Here the user has a pattern of interest and wishes to find similar patterns in the data set. This task is most commonly used for text and image data sets. For text, the pattern may be a set of keywords, and the user may wish to find relevant documents within a large set of possibly relevant documents (e.g., Web pages). For images, the user may have a sample image, a sketch of an image, or a description of an image, and wish to find similar images from a large set of images. In both cases the definition of similarity is critical, but so are the details of the search strategy.

There are numerous large-scale applications of retrieval systems, including:

- Retrieval methods are used to locate documents on the Web, as in the Google system (`www.google.com`) of Brin and Page (1998), which uses a mathematical algorithm called PageRank to estimate the relative importance of individual Web pages based on link patterns.
- QBIC ("Query by Image Content"), a system developed by researchers at IBM, allows a user to interactively search a large database of images by posing queries in terms of content descriptors such as color, texture, and relative position information (Flickner et al., 1995).

Although each of the above five tasks are clearly differentiated from each other, they share many common components. For example, shared by many tasks is the notion of *similarity* or *distance* between any two data vectors. Also shared is the notion of score functions (used to assess how well a model or pattern fits the data), although the particular functions tend to be quite different across different categories of tasks. It is also obvious that different model and pattern structures are needed for different tasks, just as different structures may be needed for different kinds of data.

1.5 Components of Data Mining Algorithms

In the preceding sections we have listed the basic categories of tasks that may be undertaken in data mining. We now turn to the question of how one actually accomplishes these tasks. We will take the view that data mining algorithms that address these tasks have four basic components:

1. **Model or Pattern Structure**: determining the underlying structure or functional forms that we seek from the data (chapter 6).

2. **Score Function**: judging the quality of a fitted model (chapter 7).

3. **Optimization and Search Method**: optimizing the score function and searching over different model and pattern structures (chapter 8).

4. **Data Management Strategy**: handling data access efficiently during the search/optimization (chapter 12).

We have already discussed the distinction between model and pattern structures. In the remainder of this section we briefly discuss the other three components of a data mining algorithm.

1.5.1 Score Functions

Score functions quantify how well a model or parameter structure fits a given data set. In an ideal world the choice of score function would precisely reflect the utility (i.e., the true expected benefit) of a particular predictive model. In practice, however, it is often difficult to specify precisely the true utility of a model's predictions. Hence, simple, "generic" score functions, such as least squares and classification accuracy are commonly used.

Without some form of score function, we cannot tell whether one model is better than another or, indeed, how to choose a good set of values for the parameters of the model. Several *score functions* are widely used for this purpose; these include likelihood, sum of squared errors, and misclassification rate (the latter is used in supervised classification problems). For example, the well-known squared error score function is defined as

$$\sum_{i=1}^{n} \left(y(i) - \hat{y}(i)\right)^2 \tag{1.1}$$

where we are predicting n "target" values $y(i)$, $1 \leq i \leq n$, and our predictions for each are denoted as $\hat{y}(i)$ (typically this is a function of some other "input" variable values for prediction and the parameters of the model).

Any views we may have on the theoretical appropriateness of different criteria must be moderated by the practicality of applying them. The model that we consider to be most likely to have given rise to the data may be the ideal one, but if estimating its parameters will take months of computer time it is of little value. Likewise, a score function that is very susceptible to slight changes in the data may not be very useful (its utility will depend on the objectives of the study). For example if altering the values of a few extreme cases leads to a dramatic change in the estimates of some model parameters caution is warranted; a data set is usually chosen from a number of possible data sets, and it may be that in other data sets the value of these extreme cases would have differed. Problems like this can be avoided by using *robust* methods that are less sensitive to these extreme points.

1.5.2 Optimization and Search Methods

The score function is a measure of how well aspects of the data match proposed models or patterns. Usually, these models or patterns are described in terms of a structure, sometimes with unknown parameter values. The goal of optimization and search is to determine the structure and the parameter values that achieve a minimum (or maximum, depending on the context) value of the score function. The task of finding the "best" values of parameters in models is typically cast as an optimization (or estimation) problem. The task of finding interesting patterns (such as rules) from a large family of potential patterns is typically cast as a combinatorial search problem, and is often accomplished using heuristic search techniques. In linear regression, a prediction rule is usually found by minimizing a least squares score function (the sum of squared errors between the prediction from a model and the observed values of the predicted variable). Such a score function is amenable to mathematical manipulation, and the model that minimizes it can be found algebraically. In contrast, a score function such as misclassification rate in supervised classification is difficult to minimize analytically. For example, since it is intrinsically discontinuous the powerful tool of differential calculus cannot be brought to bear.

Of course, while we can produce score functions to produce a good match between a model or pattern and the data, in many cases this is not really the objective. As noted above, we are often aiming to generalize to new data which might arise (new customers, new chemicals, etc.) and having too close a match to the data in the database may prevent one from predicting new cases accurately. We discuss this point later in the chapter.

1.5.3 Data Management Strategies

The final component in any data mining algorithm is the data management strategy: the ways in which the data are stored, indexed, and accessed. Most well-known data analysis algorithms in statistics and machine learning have been developed under the assumption that all individual data points can be accessed quickly and efficiently in random-access memory (RAM). While main memory technology has improved rapidly, there have been equally rapid improvements in secondary (disk) and tertiary (tape) storage technologies, to the extent that many massive data sets still reside largely on disk or tape and will not fit in available RAM. Thus, there will probably be a price to pay for accessing massive data sets, since not all data points can be simultaneously close to the main processor.

Many data analysis algorithms have been developed without including any explicit specification of a data management strategy. While this has worked in the past on relatively small data sets, many algorithms (such as classification and regression tree algorithms) scale very poorly when the "traditional version" is applied directly to data that reside mainly in secondary storage.

The field of databases is concerned with the development of indexing methods, data structures, and query algorithms for efficient and reliable data retrieval. Many of these techniques have been developed to support relatively simple counting (aggregating) operations on large data sets for reporting purposes. However, in recent years, development has begun on techniques that support the "primitive" data access operations necessary to implement efficient versions of data mining algorithms (for example, tree-structured indexing systems used to retrieve the neighbors of a point in multiple dimensions).

1.6 The Interacting Roles of Statistics and Data Mining

Statistical techniques alone may not be sufficient to address some of the more challenging issues in data mining, especially those arising from massive data sets. Nonetheless, statistics plays a very important role in data mining: it is a necessary component in any data mining enterprise. In this section we discuss some of the interplay between traditional statistics and data mining.

With large data sets (and particularly with very large data sets) we may simply not know even straightforward facts about the data. Simple eyeballing of the data is not an option. This means that sophisticated search and examination methods may be required to illuminate features which would be readily apparent in small data sets. Moreover, as we commented above, often the object of data mining is to make some inferences beyond the available database. For example, in a database of astronomical objects, we may want to make a statement that "all objects like this one behave thus," perhaps with an attached qualifying probability. Likewise, we may determine that particular regions of a country exhibit certain patterns of telephone calls. Again, it is probably not the calls in the database about which we want to make a statement. Rather it will probably be the pattern of future calls which we want to be able to predict. The database provides the set of objects which will be used to construct the model or search for a pattern, but the ultimate objective will not generally be to describe those data. In most cases the objective is to describe the general process by which the data arose, and other data sets which could have arisen by the same process. All of this means that it is necessary to avoid models or patterns which match the available database too closely: given that the available data set is merely one set from the sets of data which could have arisen, one does not want to model its idiosyncrasies too closely. Put another way, it is necessary to avoid *overfitting* the given data set; instead one wants to find models or patterns which *generalize* well to potential future data. In selecting a score function for model or pattern selection we need to take account of this. We will discuss these issues in more detail in chapter 7 and chapters 9 through 11. While we have described them in a data mining context, they are fundamental to statistics; indeed, some would take them as the defining characteristic of statistics as a discipline.

Since statistical ideas and methods are so fundamental to data mining, it is legitimate to ask whether there are really any differences between the two enterprises. Is data mining merely exploratory statistics, albeit for potentially huge data sets, or is there more to data mining than exploratory data analysis? The answer is yes—there *is* more to data mining.

The most fundamental difference between classical statistical applications and data mining is the size of the data set. To a conventional statistician, a "large" data set may contain a few hundred or a thousand data points. To someone concerned with data mining, however, many millions or even billions of data points is not unexpected—gigabyte and even terabyte databases are by no means uncommon. Such large databases occur in all walks of life. For instance the American retailer Wal-Mart makes over 20 million transactions daily (Babcock, 1994), and constructed an 11 terabyte database of customer transactions in 1998 (Piatetsky-Shapiro, 1999). AT&T has 100 million customers and carries on the order of 300 million calls a day on its long distance network. Characteristics of each call are used to update a database of models for every telephone number in the United States (Cortes and Pregibon, 1998). Harrison (1993) reports that Mobil Oil aims to store over 100 terabytes of data on oil exploration. Fayyad, Djorgovski, and Weir (1996) describe the Digital Palomar Observatory Sky Survey as involving three terabytes of data. The ongoing Sloan Digital Sky Survey will create a raw observational data set of 40 terabytes, eventually to be reduced to a mere 400 gigabyte catalog containing 3×10^8 individual sky objects (Szalay et al., 1999). The NASA Earth Observing System is projected to generate multiple gigabytes of raw data per hour (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). And the human genome project to complete sequencing of the entire human genome will likely generate a data set of more than 3.3×10^9 nucleotides in the process (Salzberg, 1999). With data sets of this size come problems beyond those traditionally considered by statisticians.

Massive data sets can be tackled by sampling (if the aim is modeling, but not necessarily if the aim is pattern detection) or by adaptive methods, or by summarizing the records in terms of *sufficient statistics*. For example, in standard least squares regression problems, we can replace the large numbers of scores on each variable by their sums, sums of squared values, and sums of products, summed over the records—these are sufficient for regression coefficients to be calculated no matter how many records there are. It is also important to take account of the ways in which algorithms scale, in terms of computation time, as the number of records or variables increases. For example, exhaustive search through all subsets of variables to find the "best" subset (according to some score function), will be feasible only up to a point. With p variables there are $2^p - 1$ possible subsets of variables to consider. Efficient search methods, mentioned in the previous section, are crucial in pushing back the boundaries here.

Further difficulties arise when there are many variables. One that is important in some contexts is the *curse of dimensionality*; the exponential rate of growth of the number of unit cells in a space as the number of variables increases. Consider, for example, a single binary variable. To obtain reasonably accurate estimates of parameters within both of its cells we might wish to have 10 observations per cell; 20 in all. With two binary variables (and four cells) this becomes 40 observations. With 10 binary variables it becomes 10240 observations, and with 20 variables it becomes 10485760. The curse of dimensionality manifests itself in the difficulty of finding accurate estimates of probability densities in high dimensional spaces without astronomically large databases (so large, in fact, that the gigabytes available in data mining applications pale into insignificance). In high dimensional spaces, "nearest" points may be a long way away. These are not simply difficulties of manipulating the many variables involved, but more fundamental problems of what can actually be done. In such situations it becomes necessary to impose additional restrictions through one's prior choice of model (for example, by assuming linear models).

Various problems arise from the difficulties of accessing very large data sets. The statistician's conventional viewpoint of a "flat" data file, in which rows represent objects and columns represent variables, may bear no resemblance to the way the data are stored (as in the text and Web transaction data sets described earlier). In many cases the data are distributed, and stored on many machines. Obtaining a random sample from data that are split up in this way is not a trivial matter. How to define the sampling frame and how long it takes to access data become important issues.

Worse still, often the data set is constantly evolving—as with, for example, records of telephone calls or electricity usage. Distributed or evolving data can multiply the size of a data set many-fold as well as changing the nature of the problems requiring solution.

While the size of a data set may lead to difficulties, so also may other properties not often found in standard statistical applications. We have already remarked that data mining is typically a secondary process of data analysis; that is, the data were originally collected for some other purpose. In contrast, much statistical work is concerned with primary analysis: the data are collected with particular questions in mind, and then are analyzed to answer those questions. Indeed, statistics includes subdisciplines of experimental design and survey design—entire domains of expertise concerned with the best ways to collect data in order to answer specific questions. When data are used to address problems beyond those for which they were originally collected, they may not be ideally suited to these problems. Sometimes the data sets are entire populations (e.g., of chemicals in a particular class of chemicals) and therefore the standard statistical notion of inference has no relevance. Even when they are not entire populations, they are often *convenience* or *opportunity* samples, rather than random samples. (For instance, the records in question may have been collected because they were the most easily measured, or covered a particular period of time.)

In addition to problems arising from the way the data have been collected, we expect other distortions to occur in large data sets—including missing values, contamination, and corrupted data points. It is a rare data set that does not have such problems. Indeed, some elaborate modeling methods include, as part of the model, a component describing the mechanism by which missing data or other distortions arise. Alternatively, an estimation method such as the EM algorithm (described in chapter 8) or an imputation method that aims to generate artificial data with the same general distributional properties as the missing data might be used. Of course, all of these problems also arise in standard statistical applications (though perhaps to a lesser degree with small, deliberately collected data sets) but basic statistical texts tend to gloss over them.

In summary, while data mining does overlap considerably with the standard exploratory data analysis techniques of statistics, it also runs into new problems, many of which are consequences of size and the non traditional nature of the data sets involved.

1.7 Data Mining: Dredging, Snooping, and Fishing

An introductory chapter on data mining would not be complete without reference to the historical use of terms such as "data mining," "dredging," "snooping," and "fishing." In the 1960s, as computers were increasingly applied to data analysis problems, it was noted that if you searched long enough, you could always find some model to fit a data set arbitrarily well. There are two factors contributing to this situation: the complexity of the model and the size of the set of possible models.

Clearly, if the class of models we adopt is very flexible (relative to the size of the available data set), then we will probably be able to fit the available data arbitrarily well. However, as we remarked above, the aim may be to generalize beyond the available data; a model that fits well may not be ideal for this purpose. Moreover, even if the aim is to fit the data (for example, when we wish to produce the most accurate summary of data describing a complete population) it is generally preferable to do this with a simple model. To take an extreme, a model of complexity equivalent to that of the raw data would certainly fit it perfectly, but would hardly be of interest or value.

Even with a relatively simple model structure, if we consider enough different models with this basic structure, we can eventually expect to find a good fit. For example, consider predicting a response variable, Y from a predictor variable X which is chosen from a very large set of possible variables, $X_1, \ldots, X_p$, none of which are related to Y. By virtue of random variation in the data generating process, although there are no underlying relationships between Y and any of the X variables, there will appear to be relationships in the data at hand. The search process will then find the X variable that appears to have the strongest relationship to Y. By this means, as a consequence of the large search space, an apparent pattern is found where none really exists. The situation is particularly bad when working with a small sample size n and a large number p of potential X variables. Familiar examples of this sort of problem include the spurious correlations which are popularized in the media, such as the "discovery" that over the past 30 years when the winner of the Super Bowl championship in American football is from a particular league, a leading stock market index historically goes up in the following months. Similar examples are plentiful in areas such as economics and the social sciences, fields in which data are often relatively sparse but models and theories to fit to the data are relatively plentiful. For instance, in economic time-series prediction, there may be a relatively short time-span of historical data available in conjunction with a large number of economic indicators (potential predictor variables). One particularly humorous example of this type of prediction was provided by Leinweber (personal communication) who achieved almost perfect prediction of annual values of the well-known Standard and Poor 500 financial index as a function of annual values from previous years for butter production, cheese production, and sheep populations in Bangladesh and the United States.

The danger of this sort of "discovery" is well known to statisticians, who have in the past labelled such extensive searches "data mining" or "data dredging"—causing these terms to acquire derogatory connotations. The problem is less serious when the data sets are large, though dangers remain even then, if the space of potential structures examined is large enough. These risks are more pronounced in pattern detection than model fitting, since patterns, by definition, involve relatively few cases (i.e., small sample sizes): if we examine a billion data points, in search of an unusual configuration of just 50 points, we have a good chance of detecting this configuration.

There are no easy technical solutions to this problem, though various strategies have been developed, including methods that split the data into subsamples so that models can be built and patterns can be detected using one part, and then their validity can be tested on another part. We say more about such methods in later chapters. The final answer, however, is to regard data mining not as a simple technical exercise, divorced from the meaning of the data. Any potential model or pattern should be presented to the data owner, who can then assess its interest, value, usefulness, and, perhaps above all, its potential reality in terms of what else is known about the data.

1.8 Summary

Thanks to advances in computers and data capture technology, huge data sets—containing gigabytes or even terabytes of data—have been and are being collected. These mountains of data contain potentially valuable information. The trick is to extract that valuable information from the surrounding mass of uninteresting numbers, so that the data owners can capitalize on it. Data mining is a new discipline that seeks to do just that: by sifting through these databases, summarizing them, and finding patterns.

Data mining should not be seen as a simple one-time exercise. Huge data collections may be analyzed and examined in an unlimited number of ways. As time progresses, so new kinds of structures and patterns may attract interest, and may be worth seeking in the data.

Data mining has, for good reason, recently attracted a lot of attention: it is a new technology, tackling new problems, with great potential for valuable commercial and scientific discoveries. However, we should not expect it to provide answers to all questions. Like all discovery processes, successful data mining has an element of serendipity. While data mining provides useful tools, that does not mean that it will inevitably lead to important, interesting, or valuable results. We must beware of over-exaggerating the likely outcomes. But the potential is there.

1.9 Further Reading

Brief, general introductions to data mining are given in Fayyad, Piatetsky-Shapiro, and Smyth (1996), Glymour et al. (1997), and a special issue of the *Communications of the ACM*, Vol. 39, No. 11. Overviews of certain aspects of predictive data mining are given by Adriaans and Zantige (1996) and Weiss and Indurkhya (1998). Witten and Franke (2000) provide a very readable, applications-oriented account of data mining from a machine learning (artificial intelligence) perspective and Han and Kamber (2000) is an accessible textbook written from a database perspective data mining. There are many texts on data mining aimed at business users, notably Berry and Linoff (1997, 2000) that contain extensive practical advice on potential business applications of data mining.

Leamer (1978) provides a general discussion of the dangers of data dredging, and Lovell (1983) provides a general review of the topic. From a statistical perspective. Hendry (1995, section 15.1) provides an econometrician's view of data mining. Hand et al. (2000) and Smyth (2000) present comparative discussions of data mining and statistics. Casti (1990, 192–193 and 439) provides a briefly discusses "common folklore" stock market predictors and coincidences.