CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

A growing number of fields, in particular the fields of business and science, are turning to data mining to make sense of large volumes of data. Financial institutions, manufacturing companies, and government agencies are just a few of the types of organizations using data mining. Data mining is also being used to address a wide range of problems, such as managing financial portfolios, optimizing marketing campaigns, and identifying insurance fraud. The adoption of data mining techniques is driven by a combination of competitive pressure, the availability of large amounts of data, and ever increasing computing power. Organizations that apply it to critical operations achieve significant returns. The use of a process helps ensure that the results from data mining projects translate into actionable and profitable business decisions. The following chapter summarizes four steps necessary to complete a data mining project: (1) definition, (2) preparation, (3) analysis, and (4) deployment. The methods discussed in this book are reviewed within this context. This chapter concludes with an outline of the book's content and suggestions for further reading.

1.2 DEFINITION

The first step in any data mining process is to define and plan the project. The following summarizes issues to consider when defining a project:

- *Objectives*: Articulating the overriding business or scientific objective of the data mining project is an important first step. Based on this objective, it is also important to specify the success criteria to be measured upon delivery. The project should be divided into a series of goals that can be achieved using available data or data acquired from other sources. These objectives and goals should be understood by everyone working on the project or having an interest in the project's results.
- *Deliverables*: Specifying exactly what is going to be delivered sets the correct expectation for the project. Examples of deliverables include a report outlining the results of the analysis or a predictive model (a mathematical model that estimates critical data) integrated within an operational system. Deliverables also

Making Sense of Data II. By Glenn J. Myatt and Wayne P. Johnson

identify who will use the results of the analysis and how they will be delivered. Consider criteria such as the accuracy of the predictive model, the time required to compute, or whether the predictions must be explained.

- *Roles and Responsibilities*: Most data mining projects involve a cross-disciplinary team that includes (1) experts in data analysis and data mining, (2) experts in the subject matter, (3) information technology professionals, and (4) representatives from the community who will make use of the analysis. Including interested parties will help overcome any potential difficulties associated with user acceptance or deployment.
- *Project Plan*: An assessment should be made of the current situation, including the source and quality of the data, any other assumptions relating to the data (such as licensing restrictions or a need to protect the confidentiality of the data), any constraints connected to the project (such as software, hardware, or budget limitations), or any other issues that may be important to the final deliverables. A timetable of events should be implemented, including the different stages of the project, along with deliverables at each stage. The plan should allot time for cross-team education and progress reviews. Contingencies should be built into the plan in case unexpected events arise. The timetable can be used to generate a budget for the project. This budget, in conjunction with any anticipated financial benefits, can form the basis for a cost–benefit analysis.

1.3 PREPARATION

1.3.1 Overview

Preparing the data for a data mining exercise can be one of the most time-consuming activities; however, it is critical to the project's success. The quality of the data accumulated and prepared will be the single most influential factor in determining the quality of the analysis results. In addition, understanding the contents of the data set in detail will be invaluable when it comes to mining the data. The following section outlines issues to consider when accessing and preparing a data set. The format of different sources is reviewed and includes data tables and nontabular information (such as text documents). Methods to categorize and describe any variables are outlined, including a discussion regarding the scale the data is measured on. A variety of descriptive statistics are discussed for use in understanding the data. Approaches to handling inconsistent or problematic data values are reviewed. As part of the preparation of the data, methods to reduce the number of variables in the data set should be considered, along with methods for transforming the data that match the problem more closely or to use with the analysis methods. These methods are reviewed. Finally, only a sample of the data set may be required for the analysis, and techniques for segmenting the data are outlined.

1.3.2 Accessing Tabular Data

Tabular information is often used directly in the data mining project. This data can be taken directly from an operational database system, such as an ERP (enterprise resource planning) system, a CRM (customer relationship management) system, SCM (supply chain management) system, or databases containing various transactions. Other common sources of data include surveys, results from experiments, or data collected directly from devices. Where internal data is not sufficient for the objective of the data mining exercise, data from other sources may need to be acquired and carefully integrated with existing data. In all of these situations, the data would be formatted as a table of observations with information on different variables of interest. If not, the data should be processed into a tabular format.

Preparing the data may include joining separate relational tables, or concatenating data sources; for example, combining tables that cover different periods in time. In addition, each row in the table should relate to the entity of the project, such as a customer. Where multiple rows relate to this entity of interest, generating a summary table may help in the data mining exercise. Generating this table may involve calculating summarized data from the original data, using computations such as sum, mode (most common value), average, or counts (number of observations). For example, a table may comprise individual customer transactions, yet the focus of the data mining exercise is the customer, as opposed to the individual transactions. Each row in the table should refer to a customer, and additional columns should be generated by summarizing the rows from the original table, such as total sales per product. This summary table will now replace the original table in the data mining exercise.

Many organizations have invested heavily in creating a high-quality, consolidated repository of information necessary for supporting decision-making. These repositories make use of data from operational systems or other sources. Data warehouses are an example of an integrated and central corporate-wide repository of decision-support information that is regularly updated. Data marts are generally smaller in scope than data warehouses and usually contain information related to a single business unit. An important accompanying component is a metadata repository, which contains information about the data. Examples of metadata include where the data came from and what units of measurements were used.

1.3.3 Accessing Unstructured Data

In many situations, the data to be used in the data mining project may not be represented as a table. For example, the data to analyze may be a collection of documents or a sequence of page clicks on a particular web site. Converting this type of data into a tabular format will be necessary in order to utilize many of the data mining approaches described later in this book. Chapter 5 describes the use of nontabular data in more detail.

1.3.4 Understanding the Variables and Observations

Once the project has been defined and the data acquired, the first step is usually to understand the content in more detail. Consulting with experts who have knowledge

about how the data was collected as well as the meaning of the data is invaluable. Certain assumptions may have been built into the data, for example specific values may have particular meanings. Or certain variables may have been derived from others, and it will be important to understand how they were derived. Having a thorough understanding of the subject matter pertaining to the data set helps to explain why specific relationships are present and what these relationships mean.

(As an aside, throughout this book variables are presented in italics.)

An important initial categorization of the variables is the scale on which they are measured. *Nominal* and *ordinal* scales refer to variables that are *categorical*, that is, they have a limited number of possible values. The difference is that ordinal variables are ordered. The variable *color* which could take values black, white, red, and so on, would be an example of a nominal variable. The variable *sales*, whose values are low, medium, and high, would be an example of an ordinal scale, since there is an order to the values. *Interval* and *ratio* scales refer to variables that can take any *continuous* numeric value; however, ratio scales have a natural zero value, allowing for a calculation of a ratio. Temperature measured in Fahrenheit or Celsius is an example of an interval scale, as it can take any continuous value within a range. Since a zero value does not represent the absence of temperature, it is classified as an interval scale. However, temperatures measured in degrees Kelvin would be an example of a ratio scale, since zero is the lowest temperature. In addition, a bank balance would be an example of a ratio scale, since zero means no value.

In addition to describing the scale on which the individual variables were measured, it is also important to understand the frequency distribution of the variable (in the case of interval or ratio scaled variables) or the various categories that a nominal or ordinal scaled variable may take. Variables are usually examined to understand the following:

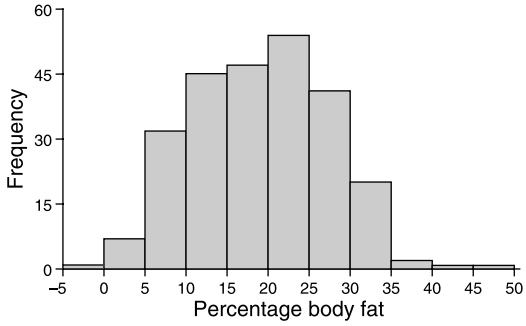
- *Central Tendency*: A number of measures for the central tendency of a variable can be calculated, including the *mean* or average value, the *median* or the middle number based on an ordering of all values, and the *mode* or the most common value. Since the mean is sensitive to outliers, the *trimmed mean* may be considered which refers to a mean calculated after excluding extreme values. In addition, median values are often used to best represent a central value in situations involving outliers or skewed data.
- *Variation*: Different numbers show the variation of the data set's distribution. The minimum and maximum values describe the entire range of the variable. Calculating the values for the different quartiles is helpful, and the calculation determines the points at which 25% (Q1), 50% (Q2), and 75% (Q3) are found in the ordered values. The *variance* and *standard deviation* are usually calculated to quantify the data distribution. Assuming a normal distribution, in the case of standard deviation, approximately 68% of all observations fall within one standard deviation of the mean, and approximately 95% of all observations fall within two standard deviations of the mean.
- *Shape*: There are a number of metrics that define the shape and symmetry of the frequency distribution, including *skewness*, a measure of whether a variable is skewed to the left or right, and *kurtosis*, a measure of whether a variable has a flat or pointed central peak.

Graphs help to visualize the central tendency, the distribution, and the shape of the frequency distribution, as well as to identify any outliers. A number of graphs that are useful in summarizing variables include: frequency histograms, bar charts, frequency polygrams, and box plots. These visualizations are covered in detail in the section on univariate visualizations in Chapter 2.

Figure 1.1 illustrates a series of statistics calculated for a particular variable (*percentage body fat*). In this example, the variable contains 251 observations, and the most commonly occurring value is 20.4 (mode), the median is 19.2, and the average or mean value is 19.1. The variable ranges from 0 to 47.5, with the point at which 25% of the ordered values occurring at 12.4, 50% at 19.2 (or median), and 75% at 25.3. The variance is calculated to be 69.6, and the standard deviation at 8.34, that is, approximately 68% of observations occur ± 8.34 from the mean (10.76–28.44), and approximately 95% of observations occur ± 16.68 from the mean (2.42–35.78).

At this point it is worthwhile taking a digression to explain terms used for the different roles variables play in building a prediction model. The *response* variable, also referred to as the *dependent* variable, the *outcome*, or *y-variable*, is the variable any model will attempt to predict. *Independent* variables, also referred to as *descriptors, predictors*, or *x-variables*, are the fields that will be used in building the model. *Labels*, also referred to as *record identification*, or *primary key*, is a unique value corresponding to each individual row in the table. Other variables may be present in the table that will not be used in any model, but which can still be used in explanations.

During this stage it is also helpful to begin exploring the data to better understand its features. Summary tables, matrices of different graphs, along with interactive techniques such as brushing, are critical data exploration tools. These tools are described in Chapter 2 on data visualization. Grouping the data is also helpful to understand the general categories of observations present in the set. The visualization of groups is presented in Chapter 2, and an in-depth discussion of clustering and grouping methods is provided in Chapter 3.



Descriptive statistics for: percentage body fat

	All observations
Number of observations	251
Central tendency	
Mode	25.4
Median	19.2
Mean	19.1
Variation	
Minimum	0
Maximum	47.5
Quartiles	Q1: 12.4 Q2: 19.2 Q3: 25.3
Variance	69.6
Standard deviation	8.34
Shape	
Skewness	0.148
Kurtosis	–0.351

Figure 1.1 Descriptive statistics and a histogram

1.3.5 Data Cleaning

Having extracted a table containing observations (represented as rows) and variables (represented as columns), the next step is to clean the data table, which often takes a considerable amount of time. Some common cleaning operations include identifying (1) errors, (2) entries with no data, and (3) entries with missing data. Errors and missing values may be attributable to the original collection, the transmission of the information, or the result of the preparation process.

Values are often missing from the data table, but a data mining approach cannot proceed until this issue is resolved. There are five options: (1) remove the entire observation from the data table; (2) remove the variable (containing the missing values) from the data table; (3) replace the missing value manually; (4) replace the value with a computed value, for example, the variable's mean or mode value; and (5) replace the entry with a predicted value based on a generated model using other fields in the data table. Different approaches for generating predictions are described in Chapter 4 on Predictive Analytics. The choice depends on the data set and the problem being addressed. For example, if most of the missing values are found in a single variable, then removing this variable may be a better option than removing the individual observations.

A similar situation to missing values occurs when a variable that is intended to be treated as a numeric variable contains text values, or specific numbers that have specific meanings. Again, the five choices previously outlined above may be used; however, the text or the specific number value may suggest numeric values to replace them with. Another example is a numeric variable where values below a threshold value are assigned a text string such as "$<10^{**}-9$." A solution for this case might be to replace the string with the number 0.000000001.

Another problem occurs when values within the data tables are incorrect. The value may be problematic as a result of an equipment malfunction or a data entry error. There are a number of ways to help identify errors in the data. Outliers in the data may be errors and can be found using a variety of methods based on the variable, for example, calculating a *z-score* for each value that represents the number of standard deviations the value is away from the mean. Values greater than plus or minus three may be considered outliers. In addition, plotting the data using a box plot or a frequency histogram can often identify data values that significantly deviate from the mean. For variables that are particularly noisy, that is they contain some degree of errors, replacing the variable with a binned version that more accurately represents the variation of the data may be necessary. This process is called *data smoothing*. Other methods, such as data visualization, clustering, and regression models (described in Chapters 2–4) can also be useful to identify anomalous observations that do not look similar to other observations or that do not fit a trend observed for the majority of the variable's observations.

Looking for values that deviate from the mean works well for numeric variables; however, a different strategy is required to handle categorical data, especially where all data values are nonnumeric. Looking at the list of all possible values a variable can take helps to eliminate and/or consolidate values where more than one value has the same meaning, which might happen, for example, in a categorical variable. Even though a

data value may look different from other values in the variable, the data may, in fact, be correct, so it is important to consult with an expert.

Problems can also arise when data from multiple sources is integrated and inconsistencies are introduced. Different sources may have values for the same variables; however, the values may have been recorded using different units of measurement and hence must be standardized to a single unit of measurement. Different sources of data may contain the same observation. Where the same observation has the same values for all variables, removing one of the observations is the most straightforward approach. Where the observations have different values, choosing which observation to keep is more challenging and best decided by someone who is able to assess the most trusted source. Other common problems when dealing with integrated data concern assessing how up-to-date the observations are and whether the quality is the same across different sources of data. Where observations are taken from different sources, retaining information on the source for future reference is prudent.

1.3.6 Transformation

In many situations, it is necessary to create new variables from existing columns of data to reflect more closely the purpose of the project or to enhance the quality of the predictions. For example, creating a new column *age* from an existing column *date of birth*, or computing an average from a series of experimental runs might be helpful. The data may also need to be transformed in order to be used with a particular analysis technique. There are six common transformations:

1. *Creating Dummy Variables*: A variable measured on a nominal or ordinal scale is usually converted into a series of *dummy* variables for use within data mining methods that require numbers. Each category is usually converted to a variable with one of two values: a one when the value is present in the observation and a zero when it is absent. Since this method would generate a new variable for each category, care should be taken when using all these columns with various methods, such as multiple linear regression or logistic regression (discussed in Chapter 4). These methods are sensitive to issues relating to *colinearity* (a high degree of correlation between variables), and hence including all variables would introduce a problem for these methods. When a final variable can be deduced from the other variables, there is no need to include the final variable. For example, the variable *color* whose values are black, white, and red could be translated into three dummy variables, one for each of the three values. Each observation would have a value one for the color corresponding to the row, and zero corresponding to the other two colors. Since the red column can be derived from the other two columns, only black and white columns are needed. The use of dummy variables is illustrated in the case studies in Chapter 5.
2. *Reducing the Number of Categories*: A categorical variable may be comprised of many different values, and using the variable directly may not draw any meaningful conclusions; however, generalizing the values may generate useful conclusions. This can be achieved through a manual definition of a

concept hierarchy or assisted using automated approaches. References in the further readings section of this chapter discuss this further, along with Appendix B (Software). For example, a variable comprising street names may be more valuable if it is generalized to the town containing those streets. This may be achieved through the construction of a concept hierarchy, where individual street names map on to the town names. In this case, there will be more observations for a particular town which hopefully result in more interesting conclusions.

3. *Create Bins for Continuous Variables*: To facilitate the use of a continuous variable within methods that require categorical variables (such as the association rules method), or to perform data smoothing, a continuous variable could be divided into a series of contiguous ranges or *bins*. Each of the observation's values would then be assigned to a specific bin, and potentially assigned a value such as the bin's mean. For example, a variable *temperature* with values ranging from 0 to 100, may be divided into a series of bins: 0–10, 10–20, and so on. A value could be assigned as each bin's mid-point. There are a variety of manual or automated approaches, and references to them are provided in the further readings section of this chapter, as well as in cases in Chapter 5 (Applications) and Appendix B (Software).
4. *Mapping the Data to a Normal Distribution*: Certain modeling approaches require that the frequency distribution of the variables approximate a normal distribution, or a bell-shaped curve. There are a number of common transformations that can be applied to a variable to achieve this. For example, a *Box-Cox* transformation or a *log* transformation may be used to generate a new variable where the data more closely follows the bell-shaped curve of a normal distribution. The Further Reading section, as well as Appendix B, provide more details related to this subject.
5. *Standardizing the Variables to a Consistent Range*: In order to treat different variables with the same weight, a scheme for normalizing the variables to the same range is often used, such as between zero and one. *Min–max*, *z-score*, and *decimal scaling* are examples of approaches to normalizing data to a specific, common range. As an example, a data set containing the variables *age* and *bank account balance* may be standardized using the *min–max* normalization to a consistent range of zero to one. These new variables make possible the consistent treatment of variables within methods, such as clustering, which utilizes distances between variables. If these two variables were not on a standard range, the *bank account balance* variable would, for the most part, be more influential than the *age* variable.
6. *Calculating Terms to Enhance Prediction*: To improve prediction, certain variables may be combined, or the variables may be transformed using some sort of mathematical operation. This may, for example, allow the more accurate modeling of nonlinear relationships. Some commonly used mathematical operations include square, cube, and square root. Appendix B and the Further Reading section of this chapter provide more details and references on this subject.

1.3.7 Variable Reduction

A data set with a large number of variables can present a number of issues within data mining techniques, including the problems of over fitting and model reliability, as well as potential computational problems. In this situation, selecting a subset of the variables will be important. This is sometimes referred to as *feature selection*. An expert with knowledge of the subject matter may be able to identify easily the variables that are not relevant to the problem. Variables that contain the same value for almost all observations do not provide much value and could be removed at this stage. In addition, categorical variables where the majority of observations have different values might not be useful within the analysis, but they may be useful to define the individual observations.

Understanding how the data will be used in a deployment scenario can also be useful in determining which variables to use. For example, the same independent variables must be gathered within a deployment scenario. However, it may be not practical to collect all the necessary data values, so it may be best to eliminate these variables at the beginning. For example, when developing a model to estimate hypertension propensity within a large patient population, a training set may include a variable *percentage body fat* as a relevant variable. The accurate measurement of this variable, however, is costly, and collecting it for the target patient population would be prohibitive. Surrogates, such as a skin-fold measurement, may be collected more easily and could be used instead of *percentage body fat*.

Additionally, examining the relationships between the variables is important. When building predictive models, there should be little relationship between the variables used to build the model. Strong relationships between the independent variables and the response variables are important and can be used to prioritize the independent variables. Bivariate data visualizations, such as scatterplot matrices, are important tools, and they are described in greater detail in Chapter 2. Calculating a correlation coefficient for each pair of continuous variables and presenting these calculations in a table can also be helpful in understanding the linear relationships between all pairs of variables, as shown in Fig. 1.2. For example, there is a strong negative linear relationship between *percentage body fat* and *density* (−0.988), a strong positive linear relationship between *abdomen* (*cm*) and *chest* (*cm*) (0.916), and a lack of a clear linear relationship between *height* (*inches*) and *percentage body fat* since it is close to zero.

	Density	Percentage body fat	Weight, lb	Height, in	Chest, cm	Abdomen, cm	Thigh, cm
Density	1	–0.988	–0.592	0.0375	–0.683	–0.798	–0.547
Percentage body fat	–0.988	1	0.611	–0.0234	0.703	0.813	0.554
Weight, lb	–0.592	0.611	1	0.489	0.894	0.888	0.87
Height, in	0.0375	–0.0234	0.489	1	0.228	0.192	0.344
Chest, cm	–0.683	0.703	0.894	0.228	1	0.916	0.732
Abdomen, cm	–0.798	0.813	0.888	0.192	0.916	1	0.766
Thigh, cm	–0.547	0.554	0.87	0.344	0.732	0.766	1

Figure 1.2 Matrix of correlation coefficients

Other techniques, such as *principal component analysis*, can also be used to reduce the number of continuous variables. The relationships between categorical independent variables can be assessed using statistical tests, such as the chi-square test. Decision trees are also useful for understanding important variables. Those chosen by the method that generates the tree are likely to be important variables to retain. Subsets of variables can also be assessed when optimizing the parameters to a data mining algorithm. For example, different combinations of independent variables can be used to build models, and those giving the best results should be retained. Methods for selecting variables are discussed in Chapter 4 on Predictive Analytics.

1.3.8 Segmentation

Using the entire data set is not always necessary, or even practical, especially when the number of observations is large. It may be possible to draw the same conclusions more quickly using a subset. There are a number of ways of selecting subsets. For example, using a random selection is often a good approach. Another method is to partition the data, using methods such as clustering, and then select an observation from each partition. This ensures the selection is representative of the entire collection of observations.

In situations where the objective of the project is to model a rare event, it is often useful to bias the selection of observations towards incorporating examples of this rare event in combination with random observations of the remaining collection. This method is called *balanced sampling*, where the response variable is used to drive how the partitioning of the data set takes place. For example, when building a model to predict insurance fraud, an initial training data set may only contain 0.1% fraudulent vs 99.9% nonfraudulent claims. Since the objective is the identification of fraudulent claims, a new training set may be constructed containing a better balance of fraudulent to nonfraudulent examples. This approach would result in improved models for identifying fraudulent claims; however, it may reduce the overall accuracy of the model. This is an acceptable compromise in this situation.

When samples are pulled from a larger set of data, comparing statistics of the sample to the original set is important. The minimum and maximum values, along with mean, median, and mode value, as well as variance and standard deviations, are a good start for comparing continuous variables. Statistical tests, such as the *t-test*, can also be used to assess the significance of any difference. When looking at categorical variables, the distribution across the different values should be similar. Generating a contingency table for the two sets can also provide insight into the distribution across different categories, and the chi-square test can be useful to quantify the differences.

Chapter 3 details methods for dividing a data set into groups, Chapter 5 discusses applications where this segmentation is needed, and Appendix B outlines software used to accomplish this.

1.3.9 Preparing Data to Apply

Having spent considerable effort preparing a data set ready to be modeled, it is also important to prepare the data set that will be scored by the prediction model in the

same manner. The steps used to access, clean, and transform the training data should be repeated for those variables that will be applied to the model.

1.4 ANALYSIS

1.4.1 Data Mining Tasks

Once a data set is acquired and prepared for analysis, the next step is to select the methods to use for data mining. These methods should match the problem outlined earlier and the type of data available. The preceding exploratory data analysis will be especially useful in prioritizing different approaches, as information relating to data set size, level of noise, and a preliminary understanding of any patterns in the data can help to prioritize different approaches. Data mining tasks primarily fall into two categories:

- *Descriptive*: This refers to the ability to identify interesting facts, patterns, trends, relationships, or anomalies in the data. These findings should be nontrivial and novel, as well as valuable and actionable, that is, the information can be used directly in taking an action that makes a difference to the organization. Identifying patterns or rules associated with fraudulent insurance claims would be an example of a descriptive data mining task.
- *Predictive*: This refers to the development of a model of some phenomena that will enable the estimation of values or prediction of future events with confidence. For example, a prediction model could be generated to predict whether a cell phone subscriber is likely to change service providers in the near future. A predictive model is typically a mathematical equation that is able to calculate a value of interest (response) based on a series of independent variables.

Descriptive data mining usually involves grouping the data and making assessments of the groups in various ways. Some common descriptive data mining tasks are:

- *Associations*: Finding associations between multiple items of interest within a data set is used widely in a variety of situations, including data mining retail or marketing data. For example, online retailers determine product combinations purchased by the same set of customers. These associations are subsequently used when a shopper purchases specific products, and alternatives are then suggested (based on the identified associations). Techniques such as association rules or decision trees are useful in identifying associations within the data. These approaches are covered in Myatt (2007).
- *Segmentation*: Dividing a data set into multiple groups that share some common characteristic is useful in many situations, such as partitioning the market for a product based on customer profiles. These partitions help in developing targeted marketing campaigns directed towards these groups. Clustering methods are widely used to divide data sets into groups of related observations, and different approaches are described in Chapter 3.

- *Outliers*: In many situations, identifying unusual observations is the primary focus of the data mining exercise. For example, the problem may be defined as identifying fraudulent credit card activity; that is, transactions that do not follow an established pattern. Again, clustering methods may be employed to identify groups of observations; however, smaller groups would now be considered more interesting, since they are a reflection of unusual patterns of activity. Clustering methods are discussed in Chapter 3.

The two primary predictive tasks are:

- *Classification*: This is when a model is built to predict a categorical variable. For example, the model may predict whether a customer will or will not buy a particular product. Methods such as logistic regression, discriminant analysis, and naive Bayes classifiers are often used and these methods are outlined in Chapter 4 on Predictive Analytics.
- *Regression*: This is also referred to as *estimation*, *forecasting*, or *prediction*, and it refers to building models that generate an estimation or prediction for a continuous variable. A model that predicts the sales for a given quarter would be an example of a regression predictive task. Methods such as multiple linear regression are often used for this task and are discussed in Chapter 4.

1.4.2 Optimization

Any data mining analysis, whether it is finding patterns and trends or building a predictive model, will involve an iterative process of trial-and-error in order to find an optimal solution. This optimization process revolves around adjusting the following in a controlled manner:

- *Methods*: To accomplish a data mining task, many potential approaches may be applied; however, it is not necessarily known in advance which method will generate an optimal solution. It is therefore common to try different approaches and select the one that produces the best results according to the success criteria established at the start of the project.
- *Independent Variables*: Even though the list of possible independent variables may have been selected in the data preparation step, one way to optimize any data mining exercise is to use different combinations of independent variables. The simplest combinations of independent variables that produced the optimal predictive accuracy should be used in the final model.
- *Parameters*: Many data mining methods require parameters to be set that adjust exactly how the approach operates. Adjusting these parameters can often result in an improvement in the quality of the results.

1.4.3 Evaluation

In order to assess which data mining approach is the most promising, it is important to objectively and consistently assess the various options. Evaluating the different

approaches also helps set expectations concerning possible performance levels during deployment. In evaluating a predictive model, different data sets should be used to build the model and to test the performance of the model, thus ensuring that the model has not overfitted the data set from which it is learning. Chapter 4 on Predictive Analytics outlines methods for assessing generated models. Assessment of the results from descriptive data mining approaches should reflect the objective of the data mining exercise.

1.4.4 Model Forensics

Spending time looking at a working model to understand when or why a model does or does not work is instructive, especially looking at the false positives and false negatives. Clustering, pulling out rules associated with these errors, and visualizing the data, may be useful in understanding when and why the model failed. Exploring this data may also help to understand whether additional data should be collected. Data visualizations and clustering approaches, described in Chapters 2 and 3, are useful tools to accomplish model forensics as well as to help communicate the results.

1.5 DEPLOYMENT

The discussion so far has focused on defining and planning the project, acquiring and preparing the data, and performing the analysis. The results from any analysis then need to be translated into tangible actions that impact the organization, as described at the start of the project. Any report resulting from the analysis should make its case and present the evidence clearly. Including the user of the report as an interested party to the analysis will help ensure that the results are readily understandable and usable by the final recipient.

One effective method of deploying the solution is to incorporate the analysis within existing systems, such as ERP or CRM systems, that are routinely used by the targeted end-users. Examples include using scores relating to products specific customers are likely to buy within a CRM system or using an insurance risk model within online insurance purchasing systems to provide instant insurance quotes. Integrating any externally developed models into the end-user system may require adoption of appropriate standards such as Object Linking and Embedding, Database for Data Mining (Data Mining OLE DB) which is an application programming interface for relational databases (described in Netz et al., 2001), Java Data Mining application programming interface standard (JSR-73 API; discussed in Hornick et al., 2006), and Predictive Model Markup Language (PMML; also reviewed in Hornick et al., 2006). In addition, the models may need to be integrated with current systems that are able to extract data from the current database and build the models automatically.

Other issues to consider when planning a deployment include:

- *Model Life Time*: A model may have a limited lifespan. For example, a model that predicts stock performance may only be useful for a limited time period,

and it will need to be rebuilt regularly with current data in order to remain useful.

- *Privacy Issues*: The underlying data used to build models or identify trends may contain sensitive data, such as information identifying specific customers. These identities should not be made available to end users of the analysis, and only aggregated information should be provided.
- *Training*: Training end-users on how to interpret the results of any analysis may be important. The end-user may also require help in using the results in the most effective manner.
- *Measuring and Monitoring*: The models or analysis generated as a result of the project may have met specific evaluation metrics. When these models are deployed into practical situations, the results may be different for other unanticipated reasons. Measuring the success of the project in the field may expose an issue unrelated to the model performance that impacts the deployed results.

1.6 OUTLINE OF BOOK

1.6.1 Overview

The remainder of this book outlines methods for visual data mining, clustering, and predictive analytics. It also discusses how data mining is being used and describes a software application that can be used to get direct experience with the methods in the book.

1.6.2 Data Visualization

Visualization is a central part of exploratory data analysis. Data analysts use visualization to examine, scrutinize, and validate their analysis before they report their findings. Decision makers use visualization to explore and question the findings before they develop action plans. Each group of people using the data needs different graphics and visualization tools to do its work.

Producing high quality data graphics or creating interactive exploratory software requires an understanding of the design principles of graphics and user interfaces. Words, numbers, typography, color, and graphical shapes must be combined and embedded in an interactive system in particular ways to show the data simply, clearly, and honestly.

There are a variety of tables and data graphics for presenting quantitative data. These include histograms and box plots for displaying one variable (univariate data), scatterplots for displaying two variables (bivariate data), and a variety of multipanel graphics for displaying many variables (multivariate data). Visualization tools like dendrograms and cluster image maps provide views of data that has been clustered into groups. Finally, these tools become more powerful when they include advances from interactive visualization.

1.6.3 Clustering

Clustering is a commonly used approach for segmenting a data set into groups of related observations. It is used to understand the data set and to generate groups in situations where the primary objective of the analysis is segmentation. A critical component in any data clustering exercises is an assessment of the *distance* between two observations. Numerous methods exist for making this determination of distance. These methods are based on the type of data being clustered; that is, whether the data set contains continuous variables, binary variables, nonbinary categorical variables, or a mixture of these variable types. A series of distance calculations are described in detail in Chapter 3.

There are a number of approaches to forming groups of observations. *Hierarchical* approaches organize the individual observations based on their relationship to other observations and groups within the data set. There are different ways of generating this hierarchy based on the method in which observations and groups in the data are combined. The approach provides a detailed hierarchical outline of the relationships in the data, usually presented as a dendrogram. It also provides a flexible way of generating groups directly from this dendrogram. Despite its flexibility, hierarchical approaches are limited in the number of observations they are able to process, and the processing is often time consuming. *Partitioned*-based approaches are a faster method for identifying clusters; however, they do not hierarchically organize the data set. The number of clusters to generate must be known prior to clustering. An alternative method, referred to as *fuzzy* clustering, does not partition the data into mutually exclusive groups, as with a hierarchical or partitioned approach. Instead, all observations belong to all groups to varying degrees. A score is associated with each observation reflecting the degree to which the observation belongs in each group. Like partitioned-based methods, fuzzy clustering approaches require that the number of groups be set prior to clustering.

1.6.4 Predictive Analytics

The focus of many data mining projects is making predictions to support decisions. There are numerous approaches to building these models, and all can be customized to varying degrees. It is important to understand what types of models, as well as what parameter changes, improve or decrease the performance of the predictions. This assessment should account for how well the different models operate using data separate from the data used to build the model. Dividing the data into sets for building and testing the model is important, and common approaches are outlined in Chapter 4. Metrics for assessment of both regression and classification models are described.

Building models from the fewest number of independent variables is often ideal. Principal component analysis is one method to understand the contribution of a series of variables to the total variation in the data set. A number of popular classification and regression methods are described in Chapter 4, including multiple linear regression, discriminant analysis, logistic regression, and naive Bayes. Multiple

linear regression identifies the linear relationship between a series of independent variables and a single response variable. Discriminant analysis is a classification approach that assigns observations to classes using the linear boundaries between the classes. Logistic regression can be used to build models where the response is a binary variable. In addition, the method calculates the probability that a response value is positive. Finally, naive Bayes is a classification approach that only works with categorical variables and it is particularly useful when applied to large data sets. These methods are described in detail, including an analysis of when they work best and what assumptions are required for each.

1.6.5 Applications

Data mining is being applied to a diverse range of applications and industries. Chapter 5 outlines a number of common uses for data mining, along with specific applications in the following industries: finance, insurance, retail, telecommunications, manufacturing, entertainment, government, and healthcare. A number of case studies are outlined and the process is described in more detail for two projects: a data set related to genes and a data set related to automobile loans. This chapter also outlines a number of approaches to data mining some commonly used nontabular sources, including text documents as well as chemicals. The chapter includes a description of how to extract information from this content, along with how to organize the content for decision-making.

1.6.6 Software

A software program called Traceis (available from http://www.makingsenseofdata.com/) has been created for use in combination with the descriptions of the various methods provided in the book. It is described in Appendix B. The software provides multiple tools for preparing the data, generating statistics, visualizing variables, and grouping observations, as well as building prediction models. The software can be used to gain hands-on experience on a range of data mining techniques in one package.

1.7 SUMMARY

The preceding chapter described a data mining process that includes the following steps:

1. *Definition*: This step includes defining the objectives of the exercise, the deliverables, the roles and responsibilities of the team members, and producing a plan to execute.
2. *Preparation*: The data set to be analyzed needs to be collected from potentially different sources. It is important to understand the content of the variables and define how the data will be used in the final analysis. The data should be cleaned and transformations applied that will improve the quality of the final results. Efforts should be made to reduce the number of variables in the set

TABLE 1.1 Data Mining Tasks

Type of task	Specific task	Description	Example methods
Descriptive	Association	Finding associations between multiple items of interest	Association rules, decision trees, data visualization
	Segmentation	Dividing a data set into groups that share common characteristics	Clustering, decision trees
	Outliers	Identifying usual observations	Clustering, data visualization
Predictive	Classification	A predictive model that predicts a categorical variable	Discriminant analysis, logistic regression, naive Bayes
	Regression	A predictive model that predicts a continuous variable	Multiple linear regression

to analyze. A subset of observations may also be needed to streamline the analysis.

3. *Analysis*: Based on an understanding of the problem and the data available, a series of data mining options should be investigated, such as those summarized in Table 1.1. Experiments to optimize the different approaches, through a variety of parameter settings and variable selections, should be investigated and the most promising one should be selected.
4. *Deployment*: Having implemented the analysis, carefully planning deployment to ensure the results are translated into benefits to the business is the final step.

1.8 FURTHER READING

A number of published process models outline the data mining steps, including CRISP_DM (http://www.crisp-dm.org/) and SEMMA (http://www.sas.com/technologies/analytics/datamining/miner/semma.html). In addition, a number of books discuss the data mining process further, including Shumueli et al. (2007) and Myatt (2007). The following resources provide more information on preparing a data set for data mining: Han and Kamber (2006), Refaat (2007), Pyle (1999, 2003), Dasu and Johnson (2003), Witten and Frank (2003), Hoaglin et al. (2000), and Shumueli et al. (2007). A discussion concerning technology standards for deployment of data mining applications can be found in Hornick et al. (2006).