CONTENTS

CHAPTER 1

Introduction

Today's statistics applications involve enormous data sets: many cases (rows of a data spreadsheet, with a row representing the information on a studied case) and many variables (columns of the spreadsheet, with a column representing the outcomes on a certain characteristic across the studied cases). A case may be a certain item such as a purchase transaction, or a subject such as a customer or a country, or an object such as a car or a manufactured product. The information that we collect varies across the cases, and the explanation of this variability is central to the tools that we study in this book. Many variables are typically collected on each case, but usually only a few of them turn out to be useful. The majority of the collected variables may be irrelevant and represent just noise. It is important to find those variables that matter and those that do not.

Here are a few types of data sets that one encounters in data mining. In marketing applications, we observe the purchase decisions, made over many time periods, of thousands of individuals who select among several products under a variety of price and advertising conditions. Social network data contains information on the presence of links among thousands or millions of subjects; in addition, such data includes demographic characteristics of the subjects (such as gender, age, income, race, and education) that may have an effect on whether subjects are "linked" or not. Google has extensive information on 100 million users, and Facebook has data on even more. The recommender systems developed by firms such as Netflix and Amazon use available demographic information and the detailed purchase/rental histories from millions of customers. Medical data sets contain the outcomes of thousands of performed procedures, and include information on their characteristics such as the type of procedure and its outcome, and the location where and the time when the procedure has been performed.

While traditional statistics applications focus on relatively small data sets, data mining involves very large and sometimes enormous quantities of information. One talks about megabytes and terabytes of information. A megabyte represents a million bytes, with a byte being the number of bits needed to encode a single character of text. A typical English book in plain text format (500 pages with 2000

Data Mining and Business Analytics with R, First Edition. Johannes Ledolter.

characters per page) amounts to about 1 MB. A terabyte is a million megabytes, and an exabyte is a million terabytes.

Data mining attempts to extract useful information from such large data sets. Data mining explores and analyzes large quantities of data in order to discover meaningful patterns. The *scale* of a typical data mining application, with its large number of cases and many variables, exceeds that of a standard statistical investigation. The analysis of millions of cases and thousands of variables also puts pressure on the *speed* that is needed to accomplish the search and modeling steps of the typical data mining application. This is why researchers refer to data mining as statistics at scale and speed. The large scale (lots of available data) and the requirements on speed (solutions are needed quickly) create a large demand for automation. Data mining uses a combination of pattern-recognition rules, statistical rules, as well as rules drawn from machine learning (an area of computer science).

Data mining has wide applicability, with applications in intelligence and security analysis, genetics, the social and natural sciences, and business. Studying which buyers are more likely to buy, respond to an advertisement, declare bankruptcy, commit fraud, or abandon subscription services are of vital importance to business.

Many data mining problems deal with categorical outcome data (e.g., no/yes outcomes), and this is what makes machine learning methods, which have their origins in the analysis of categorical data, so useful. Statistics, on the other hand, has its origins in the analysis of continuous data. This makes statistics especially useful for correlation-type analyses where one sifts through a large number of correlations to find the largest ones.

The analysis of large data sets requires an efficient way of storing the data so that it can be accessed easily for calculations. Issues of data warehousing and how to best organize the data are certainly very important, but they are not emphasized in this book. The book focuses on the analysis tools and targets their statistical foundation.

Because of the often enormous quantities of data (number of cases/replicates), the role of traditional statistical concepts such as confidence intervals and statistical significance tests is greatly reduced. With large data sets, almost any small difference becomes significant. It is the problem of overfitting models (i.e., using more explanatory variables than are actually needed to predict a certain phenomenon) that becomes of central importance. Parsimonious representations are important as simpler models tend to give more insight into a problem. Large models overfitted on training data sets usually turn out to be extremely poor predictors in new situations as unneeded predictor variables increase the prediction error variance. Furthermore, overparameterized models are of little use if it is difficult to collect data on predictor variables in the future. Methods that help avoid such overfitting are needed, and they are covered in this book. The partitioning of the data into training and evaluation (test) data sets is central to most data mining methods. One must always check whether the relationships found in the training data set will hold up in the future.

Many data mining tools deal with problems for which there is no designated response that one wants to predict. It is common to refer to such analysis as *unsupervised learning*. Cluster analysis is one example where one uses feature (variable) data on numerous objects to group the objects (i.e., the cases) into a

smaller number of groups (also called *clusters*). Dimension reduction applications are other examples for such type of problems; here one tries to reduce the many features on an object to a manageable few. Association rules also fall into this category of problems; here one studies whether the occurrence of one feature is related to the occurrence of others. Who would not want to know whether the sales of chips are being "lifted" to a higher level by the concurrent sales of beer?

Other data mining tools deal with problems for which there is a designated response, such as the volume of sales (a quantitative response) or whether someone buys a product (a categorical response). One refers to such analysis as *supervised learning*. The predictor variables that help explain (predict) the response can be quantitative (such as the income of the buyer or the price of a product) or categorical (such as the gender and profession of the buyer or the qualitative characteristics of the product such as new or old). Regression methods, regression trees, and nearest neighbor methods are well suited for problems that involve a continuous response. Logistic regression, classification trees, nearest neighbor methods, discriminant analysis (for continuous predictor variables) and naïve Bayes methods (mostly for categorical predictor variables) are well suited for problems that involve a categorical response.

Data mining should be viewed as a *process*. As with all good statistical analyses, one needs to be clear about the purpose of the analysis. Just to "mine data" without a clear purpose, without an appreciation of the subject area, and without a modeling strategy will usually not be successful. The data mining process involves several interrelated steps:

1. Efficient data storage and data preprocessing steps are very critical to the success of the analysis.
2. One needs to select appropriate response variables and decide on the number of variables that should be investigated.
3. The data needs to be screened for outliers, and missing values need to be addressed (with missing values either omitted or appropriately imputed through one of several available methods).
4. Data sets need to be partitioned into training and evaluation data sets. In very large data sets, which cannot be analyzed easily as a whole, data must be sampled for analysis.
5. Before applying sophisticated models and methods, the data need to be visualized and summarized. It is often said that a picture is worth a 1000 words. Basic graphs such as line graphs for time series, bar charts for categorical variables, scatter plots and matrix plots for continuous variables, box plots and histograms (often after stratification on useful covariates), maps for displaying correlation matrices, multidimensional graphs using color, trellis graphs, overlay plots, tree maps for visualizing network data, and geo maps for spatial data are just a few examples of the more useful graphical displays. In constructing good graphs, one needs to be careful about the right scaling, the correct labeling, and issues of stratification and aggregation.
6. Summary of the data involves the typical summary statistics such as mean, percentiles and median, standard deviation, and correlation, as well as more advanced summaries such as principal components.

7. Appropriate methods from the data mining tool bag need to be applied. Depending on the problem, this may involve regression, logistic regression, regression/classification trees, nearest neighbor methods, k-means clustering, and so on.
8. The findings from these models need to be confirmed, typically on an evaluation (test or holdout) data set.
9. Finally, the insights one gains from the analysis need to be implemented. One must act on the findings and spring to action. This is what W.E. Deming had in mind when he talked about process improvement and his Deming (Shewhart) wheel of "plan, do, check, and act" (Ledolter and Burrill, 1999).

Some data mining applications require an enormous amount of effort to just collect the relevant information. For example, an investigation of Pre-Civil War court cases of Missouri slaves seeking their freedom involves tedious study of handwritten court proceedings and Census records, electronic scanning of the records, and the use of character-recognition software to extract the relevant characteristics of the cases and the people involved. The process involves double and triple checking unclear information (such as different spellings, illegible entries, and missing information), selecting the appropriate number of variables, categorizing text information, and deciding on the most appropriate coding of the information. At the end, one will have created a fairly good master list of all available cases and their relevant characteristics. Despite all the diligent work, there will be plenty of missing information, information that is in error, and way too many variables and categories than are ultimately needed to tell the story behind the judicial process of gaining freedom.

Data preparation often takes a lot more time than the eventual modeling. The subsequent modeling is usually only a small component of the overall effort; quite often, relatively simple methods and a few well-constructed graphs can tell the whole story. It is the creation of the master list that is the most challenging task. The steps that are involved in the construction of the master list in such problems depend heavily on the subject area, and one can only give rough guidelines on how to proceed. It is also difficult to make this process automatic. Furthermore, even if some of the "data cleaning" steps can be made automatic, the investigator must constantly check and question any adjustments that are being made. Great care, lots of double and triple checking, and much common sense are needed to create a reliable master list. But without a reliable master list, the findings will be suspect, as we know that wrong data usually lead to wrong conclusions. The old saying "garbage in–garbage out" also applies to data mining.

Fortunately many large business data sets can be created almost automatically. Much of today's business data is collected for transactional purposes, that is, for payment and for shipping. Examples of such data sets are transactions that originate from scanner sales in super markets, telephone records that are collected by mobile telephone providers, and sales and rental histories that are collected by companies such as Amazon and Netflix. In all these cases, the data collection effort is minimal,