1
Introduction

Many scientific problems reduce to modeling the relationship between two sets of variables. Regression methodology is designed to quantify these relationships. Due to their mathematical simplicity, linear regression for continuous data, logistic regression for binary data, proportional hazard regression for censored survival data, and mixed-effect regression for longitudinal data are among the most commonly used statistical methods. These parametric (or semiparametric) regression methods, however, may not lead to faithful data descriptions when the underlying assumptions are not satisfied. As remedies, extensive literature exists to perform diagnosis of parametric or semiparametric regression models, but the practice of the model diagnosis is uneven at best. A common practice is the visualization of the residual plots, which is a straightforward task for a simple regression model, but can be highly sophisticated as the model complexity grows. Furthermore, model interpretation can be problematic in the presence of higher-order interactions among potent predictors. Nonparametric regression has evolved to relax or remove the restrictive assumptions.

In many cases, recursive partitioning provides a useful alternative to the parametric regression methods. The theme of this book is to describe nonparametric regression methods built on recursive partitioning. Importantly, recursive partitioning is a statistical technique that forms the basis for two classes of nonparametric regression methods: Classification and Regression Trees (CART) and Multivariate Adaptive Regression Splines (MARS). In the last two decades, many methods have been developed on the basis of or inspired by CART and MARS, and some of them are described in this book. Although relatively new, the applications of these methods are far

H. Zhang and B.H. Singer, *Recursive Partitioning and Applications*,
Springer Series in Statistics, DOI 10.1007/978-1-4419-6824-1_1,

reaching, as a result of increasing complexity of study designs and the massive size of many data sets (a large number of observations or variables).

Although most commercial applications of recursive partitioning-based methods have not been well-documented through peer-reviewed publications, there is no doubt about their extensive use. For example, they have been used by financial firms [banking crises (Cashin and Duttagupta 2008), credit cards (Altman 2002; Frydman, Altman and Kao 2002; Kumar and Ravi 2008), and investments (Pace 1995 and Brennan, Parameswaran et al. 2001)], manufacturing and marketing companies (Levin, Zahavi, and Olitsky 1995; Chen and Su 2008), and pharmaceutical industries (Chen et al. 1998). They have also been applied in engineering research. Bahl and colleagues (1989) introduced a tree-based language model for natural language speech recognition, and Wieczorkowska (1999) used decision trees to classify musical sounds. Desilva and Hull (1994) used the idea of decision trees to detect proper nouns in document images. Geman and Jedynak (1996) used a related idea to form an active testing model for tracking roads in satellite images. In addition, decision trees have been used in scientific, social, and musical studies including astronomy (Owens, Griffiths, and Ratnatunga 1996), computers and the humanities (Shmulevich et al. 2001), chemistry (Chen, Rusinko, and Young 1998), environmental entomology (Hebertson and Jenkins 2008), forensics (Appavu and Rajaram 2008), and polar biology (Terhune et al. 2008).

The best-documented, and arguably most popular uses of tree-based methods are in biomedical research for which classification is a central issue. For example, a clinician or health scientist may be very interested in the following question (Goldman et al. 1982 and 1996; Zhang et al. 1998): Is this patient with chest pain suffering a heart attack, or does he simply have a strained muscle? To answer this question, information on this patient must be collected, and a good diagnostic test utilizing such information must be in place. Tree-based methods provide one solution for constructing the diagnostic test.

To help readers understand the methods and appreciate the applications, while explaining the methodology in its entirety, we emphasize the applications of these methods. Moreover, it should become apparent from those applications that the resulting models have very natural and useful interpretations, and the computation will be less and less an issue. Specifically, we will see that the tree representations can be stated as a string of hierarchal Boolean statements, facilitating conversion of complex output to narrative form.

In Section 1.1 we give a number of examples for which recursive partitioning has been used to investigate a broad spectrum of scientific problems. In Section 1.2 we formulate these scientific problems into a general regression framework and introduce the necessary notation. To conclude this chapter, we outline the contents of the subsequent chapters in Section 1.3.

1.1 Examples Using CART

Recursive partitioning has been applied to understand many problems in biological, physical, and social science. The examples selected below are not necessarily fully representative, but they give us some idea about the breadth of applications.

Example 1.1 *Chest Pain*

Goldman et al. (1982, 1996) provided a classic example of using CART. Their purpose was to build an expert computer system that could assist physicians in emergency rooms to classify patients with chest pain into relatively homogeneous groups within a few hours of admission using the clinical factors available. This classification can help physicians to plan for appropriate levels of medical care for patients based on their classified group membership. The authors included 10,682 patients with acute chest pain in the derivation data set and 4,676 in the validation data set. The derivation data were used to set up a basic model frame, while the validation data were utilized to justify the model and to conduct hypothesis testing.

Example 1.2 *Coma*

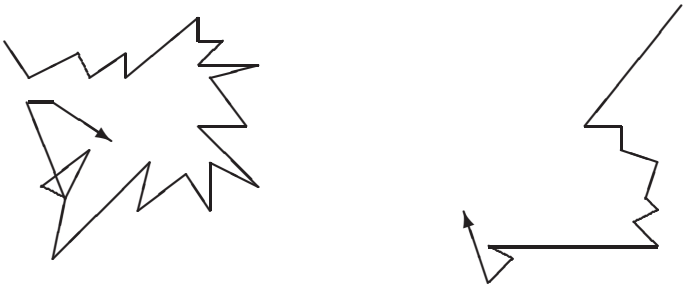
Levy et al. (1985) carried out one of the early applications of CART. To predict the outcome from coma caused by cerebral hypoxia-ischemia, they studied 210 patients with cerebral hypoxia-ischemia and considered 13 factors including age, sex, verbal and motor responses, and eye opening movement. Several guidelines were derived to predict within the first few days which patients would do well and which would do poorly.

Example 1.3 *Mammalian Sperm*

Mammalian sperm move in distinctive patterns, called hyperactivated motility, during capacitation. Figure 1.1(a) is a circular pattern of hyperactivated rabbit spermatozoa, and Figure 1.1(b) displays a nonhyperactivated track. In general, hyperactivated motility is characterized by a change from progressive movement to highly vigorous, nonprogressive random motion. This motility is useful for the investigation of sperm function and the assessment of fertility. For this reason, we must establish a quantitative criterion that recognizes hyperactivated sperm in a mixed population of hyperactivated and nonhyperactivated sperm. After collecting 322 hyperactivated and 899 nonhyperactivated sperm, Young and Bod (1994) derived a classification rule based on the wobble parameter of motility and the curvilinear velocity, using CART. Their rule was shown to have a lower misclassification rate than the commonly used ones that were established by linear discriminant analysis.

Example 1.4 *Infant Fever*

Important medical decisions are commonly made while substantial uncertainty remains. Acute unexplained fever in infants is one such frequently



a b

FIGURE 1.1. Motility patterns for mammalian sperm. (a) Hyperactivated and (b) nonhyperactivated

encountered problem. To make a correct diagnosis, it is critical to utilize information efficiently, including medical history, physical examination, and laboratory tests. Using a sample of 1,218 childhood extremity injuries seen in 1987 and 1988 by residents in family medicine and pediatrics in the Rochester General Hospital Emergency Department, McConnochie, Roghmann, and Pasternack (1993) demonstrated the value of the complementary use of logistic regression and CART in developing clinical guidelines.

Example 1.5 *Pregnancy Outcome*

Birth weight and gestational age are strong predictors for neonatal mortality and morbidity; see, e.g., Bracken (1984). In less developed countries, however, birth weight may not be measured for the first time until several days after birth, by which time substantial weight loss could have occurred. There are also practical problems in those countries in obtaining gestational age because many illiterate pregnant women cannot record the dates of their last menstrual period or calculate the duration of gestational age. For these considerations, Raymond et al. (1994) selected 843 singleton infants born at a referral hospital in Addis Ababa, Ethiopia, in 1987 and 1988 and applied CART to build a practical screening tool based on neonatal body measurements that are presumably more stable than birth weight. Their study suggests that head and chest circumferences may adequately predict the risk of low birth weight (less than 2,500 grams) and preterm (less than 37 weeks of gestational age) delivery.

Example 1.6 *Head Injury*

Head injuries cause about a half million patient hospitalizations in the United States each year. As a result of the injury, victims often suffer from persistent disabilities. It is of profound clinical importance to make early prediction of long-term outcome so that the patient, the family, and the physicians have sufficient time to arrange a suitable rehabilitation plan. Moreover, this outcome prediction can also provide useful information for

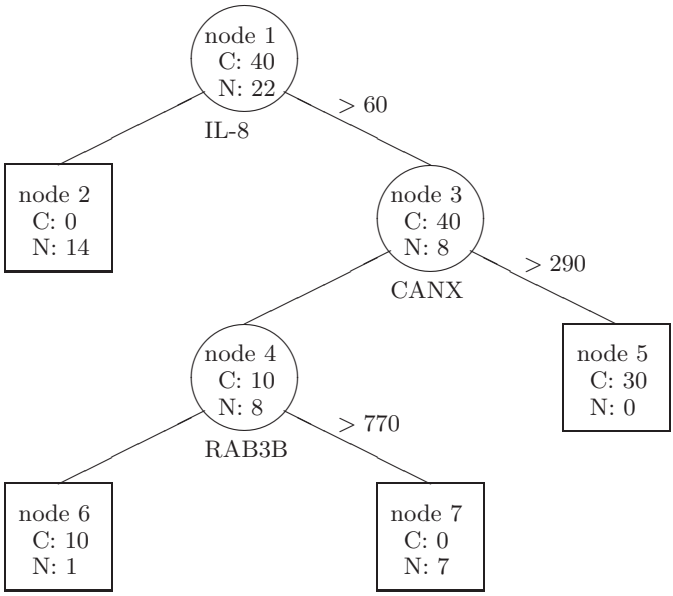


FIGURE 1.2. Classification tree for colon cancer diagnosis based on gene expression data. Inside each node are the number of tumor (C) and normal (N) tissues

assessing the treatment effect. Using CART, Choi et al. (1991) and Temkin et al. (1995) have developed prediction rules for long-term outcome in patients with head injuries on the basis of 514 patients. Those rules are simple and accurate enough for clinical practice.

Example 1.7 *Gene Expression*

As we mentioned earlier, there has been great interest in applying tree-based methods to understand high-throughput genomic data. Zhang et al. (2001) analyzed a data set from the expression profiles of 2,000 genes in 22 normal and 40 colon cancer tissues (Alon et al. 1999). Figure 1.2 is a classification tree constructed from that data set for the diagnosis of colon cancer based on gene expression profiles.

Example 1.8 *Marketing and Management*

Not only have the recursive partitioning-based methods been used in scientific research, but they have also been used in commercial applications. Levin et al. (1995) developed a customer-oriented decision support system for the marketing decisions of the Franklin Mint, a leading Philadelphia-based worldwide direct response marketer of quality collectibles and luxury home decor products. Based on the customers' attributes and characteristics, the system finds the "right" audience for promotion. In another appli-

cation, Alfaro Cortés et al. (2007) used classification trees and an AdaBoost algorithm to predict corporate failure.

Example 1.9 *Chemical Compounds*
Recursive partitioning has been employed to aid drug development. To screen large chemical databases in corporate collections and chemical libraries, Chen et al. (1998) used recursive partitioning to develop three-dimensional pharmacophores that can guide database screening, chemical library design, and lead optimization. They encoded the three-dimensional features of chemical compounds into bit strings, which were then selected to predict the biological activities of the compounds.

Example 1.10 *Musical Audio*
As a part of a large-scale interdisciplinary MAMI project (Musical Audio MIning project) conducted at Ghent University, Martens (2002) attempted to extract the tonal context from a polyphonic musical audio signal and to convert this information into a meaningful character sequence. First, a musical signal is decomposed in different sub-bands and represented as neural patterns by an auditory peripheral module. This process converts the musical signal eventually into a real vector in a 69-dimensional space, which is the predictor space. The class label represents one of the 24 keys (the 12 major and 12 minor keys in the so-called Shepard chords). We used 120 synthesized sounds, 5 from each of the following: Shepard sequences; Bass sequences, sampled from a Yamaha QS300 synthesizer; Piano sequences, sampled from a Yamaha QS300 synthesizer; Strings sequences, sampled from a Yamaha QS300 synthesizer; and Dance-lead sequences, sampled from a Waldorf Micro Q synthesizer. Then, a classification tree is used in the conversion of 120 synthesized sounds into a character string.

1.2 The Statistical Problem

Examples 1.1–1.10 can be summarized into the same statistical problem as follows. They all have an outcome variable, Y, and a set of p predictors, $x_1, \ldots, x_p$. The number of predictors, p, varies from example to example. The x's will be regarded as fixed variables, and Y is a random variable. In Example 1.3, Y is a dichotomous variable representing either hyperactivated or nonhyperactivated sperm. The x's include the wobble parameter of motility and the curvilinear velocity. Obviously, not all predictors appear in the prediction rule. Likewise, the x's and Y can be easily identified for the other examples. The statistical problem is to establish a relationship between Y and the x's so that it is possible to predict Y based on the values of the x's. Mathematically, we want to estimate the conditional probability of the random variable Y,

$$I\!P\{Y = y \,|\, x_1, \ldots, x_p\}, \tag{1.1}$$

or a functional of this probability such as the conditional expectation

$$I\!E\{Y \mid x_1, \ldots, x_p\}. \tag{1.2}$$

Many applications (e.g., Example 1.1) involve dichotomous Y (0 or 1), the conditional expectation in (1.2) coincides with the conditional probability in (1.1) with $y = 1$. In such circumstances, logistic regression is commonly used, assuming that the conditional probability (1.1) is of a specific form,

$$\frac{\exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}, \tag{1.3}$$

where the β's are parameters to be estimated.

In the ordinary linear regression, the conditional probability in (1.1) is assumed to be a normal density function,

$$\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right], \tag{1.4}$$

where the mean, μ, equals the conditional expectation in (1.2) and is of a hypothesized expression

$$\mu = \beta_0 + \sum_{i=1}^{p} \beta_i x_i. \tag{1.5}$$

The σ^2 in (1.4) is an unknown variance parameter. We use $N(\mu, \sigma^2)$ to denote the normal distribution corresponding to the density in (1.4).

In contrast to these models, recursive partitioning is a nonparametric technique that does not require a specified model structure like (1.3) or (1.5). In the subsequent chapters, the outcome Y may represent a censored measurement or a correlated set of responses. We will cite more examples accordingly.

1.3 Outline of the Methodology

In this book, we will describe both classic (mostly parametric) and modern statistical techniques as complementary tools for the analysis of data. The five types of response variables listed in Table 1.1 cover the majority of the data that arise from applications. Table 1.1 is meant to highlight the content of this book. Thus, it is not a complete list of methods that are available in the literature.

Chapter 2 is a practical guide to tree construction, focusing on the statistical ideas and scientific judgment. Technical details are deferred to Chapter 4, where methodological issues involved in classification trees are discussed in depth. We refer to Breiman et al. (1984) for further elaboration. Section

TABLE 1.1. Correspondence Between the Uses of Classic Approaches and Recursive Partitioning Technique in This Book

Type of response	Parametric methods	Recursive partitioning technique
Continuous	Ordinary linear regression	Regression trees and adaptive splines in Chapter 10
Binary	Logistic regression in Chapter 3	Classification trees and forests in Chapters 4 and 6
Censored	Proportion hazard regression in Chapter 8	Survival trees in Chapter 9
Longitudinal	Mixed-effects models in Chapter 11	Regression trees and adaptive splines in Chapter 11
Multiple discrete	Exponential, marginal, and frailty models	Classification trees, all in Chapter 12

4.2.3 on *Nested Optimal Subtrees* is relatively technical and may be difficult for some readers, but the rest of Chapter 4 is relatively straightforward. Technical differences between classification trees and regression trees are very minimal. After elucidating classification trees in Chapter 4, we introduce regression trees briefly, but sufficiently, in Section 10.2, focusing on the differences. To further demonstrate the use of classification trees, we report a stratified tree-based risk factor analysis of spontaneous abortion in Section 5.1.

The most notable addition to this edition is Chapter 6 that introduces forest-based classification and prediction. As a result of many applications where the size of the data is huge, such as high-throughput genomic data, forest-based methods are in high demand.

Chapters 7 to 9 cover the analysis of censored data. The first part is a shortcut to the output of survival trees. We present classical methods of survival analysis prior to the exposition of survival trees in the last compartment of this coverage.

Chapter 12 on classification trees for multiple binary responses is nearly parallel to survival trees from a methodological point of view. Thus, they can be read separately depending on the different needs of readers.

We start a relatively distinct topic in Chapter 10 that is fundamental to the understanding of adaptive regression splines and should be read before Chapter 11, where the use of adaptive splines is further expanded.

Before discussing the trees and splines approaches, we will describe their parametric counterparts and explain how to use these more standard models. We view it as important to understand and appreciate the parametric methods even though the main topic of this book is recursive partitioning.