BOOK REVIEW

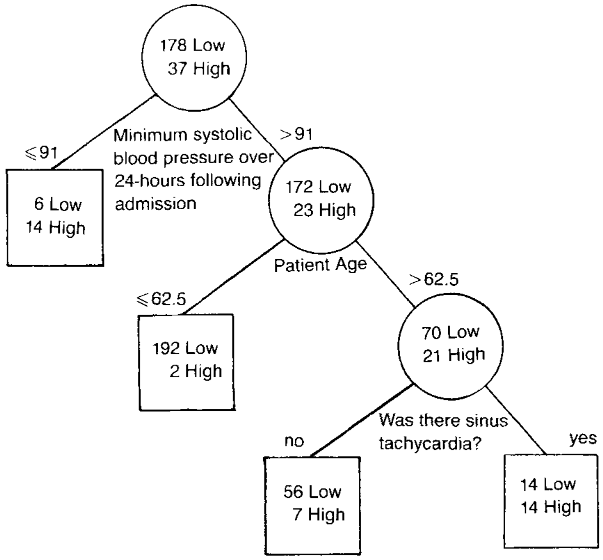
Classification and Regression Trees, by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Brooks/Cole Publishing, Monterey, 1984, 358 pages, $27.95.

This paperback book describes a relatively new, computer based method for deriving a classification rule for assigning objects to groups. As the authors state in their preface:

> Binary trees give an interesting and often illuminating way of looking at data in classification or regression problems. They should not be used to the exclusion of other methods. We do not claim that they are always better. They do add a flexible nonparametric tool to the data analyst's arsenal.

The authors use the acronym CART (derived from the book title) in referring to their method and take pains to compare results using CART with those obtained by more conventional types of classification, such as linear discriminant analysis and nearest neighbor classification.

An example of classifying heart attack patients into two groups, those who will survive 30 days (low risk) and those who will not (high risk), is used in the first chapter to introduce the method. After examining 19 variables, including age and blood pressure, the following classification tree was produced:



Using this simple procedure 89% of low risk and 75% of high risk patients were correctly classified. Linear stepwise discriminant analysis required 12 variables to achieve slightly less accurate results. Logistic regression required 10 variables, including 3 interactions suggested by the CART program, to achieve comparable results. Nearest neighbor procedures were unsuccessful because of their poor performance and excessive computational requirements, according to the authors. Thus, with this easily understood result, the reader is enticed to read further in hopes of understanding how the method works.

The book is organized into four parts: chapters 1–5: Description of tree classification methodology; chapters 6–7: Examples of classification by CART; chapter 8: Use of tree methodology in regression; chapters 9–12: Theoretical framework for CART.

In general the book is easy to read but suffers from the (probably unavoidable) necessity of having to introduce a plethora of symbols, some capitalized, some asterisked, some script and others Greek. Fortunately, a 7 page notation index is supplied as an appendix.

The first chapter deals with the purpose of classification analysis, problems in estimating accuracy of a classification scheme and a description of the Bayes rule for optimum classification. In the second chapter the construction of a tree is described. Basically there are three elements: 1) the selection of the splits; 2) the decisions about when to declare a node terminal or to continue splitting it; and 3) the assignment of each terminal node to a group. A nodal impurity function is defined both algebraically and descriptively. It is stated that the goodness of a split will be measured by the decrease in impurity. The tree is made by selecting, at each stage, that split, from among all possible splits (hence the reason that a computer must be used since the total number is often quite large), which leads to the greatest decrease in impurity. Unfortunately, no examples are given of this key concept which is essential to tree formation. The reader is left with only a fuzzy idea of what is meant. Because of the algebraic complexity of the definitions, it requires much effort to work out, with pencil and paper, the examples necessary to gain a clearer idea of how the process works. Surely the authors could have spared the reader this effort.

Chapter three is concerned with two main issues: pruning the tree and getting more accurate estimates of the true probabilities of misclassification. It can be shown algebraically that, in the absence of a stopping rule, splitting will continue to reduce the impurity until each node is pure (i.e., only one group is represented). Such a procedure would be unrealistic since application of the same procedure to another sample of data could lead to a different set of terminal nodes. The authors' solution is to generate the entire tree and then prune

Address reprint requests to Dr. Dan H. Moore II, Biomedical Sciences L-452, Lawrence Livermore National Laboratory, PO Box 5507, Livermore CA 94550.

upward according to some criterion. The criterion chosen is the sum of the cost of misclassification and a measure of tree complexity. Tree complexity is defined as a multiple of the total number of nodes. The cost of misclassification is the sum of products: cost of misclassification of type k times the probability of making that type of misclassification. This leads to concern over "honest" estimates of misclassification probabilities. The authors suggest cross-validation as a method for obtaining nearly unbiased estimates. Under this scheme the data are divided into several equal partitions. The classification tree is derived leaving out, in succession, each partition. Once the tree is derived the omitted partition is classified according to the tree obtained by using the others. In this way nearly unbiased estimates of misclassification probabilities can be obtained. Again these methods are described verbally and algebraically but no concrete examples are worked through. The results of applying the procedures to real data are shown but that this is, unfortunately, insufficient for understanding how the method works.

Splitting rules are discussed in Chapter 4. The reasons behind the authors suggestion for using either of two possible rules, the Gini diversity index and the "twoing" criterion are discussed. Both criteria have been implemented in the CART computer program. After applying both criteria to simulated data the authors conclude that Gini tends to favor a split into one small, pure node and a large, impure mode while twoing favors splits that tend to equalize populations.

Chapter 5 treats ancillary topics such as methods for using combinations of variables rather than one at a time, the growing of exploratory trees, and methods for reducing computations by subsampling. The examples of application of CART to data provided in chapters 6 and 7 are imaginative and the results are interesting. They should arouse an interest in the reader in applying the method to his data. Regression trees are described in chapter 8, but I feel the authors chose a poor example for illustration of the method. The data were originally used to study the relationship between air pollution concentration (NOX) and housing values. Unfortunately, the NOX variable was not used in the trees drawn by the CART program. Thus it is difficult to measure the size and strength of the relationship which can be accomplished quite easily by linear regression. Finally, chapters 9–12 take care of many of the theoretical details and appear to be competently written.

In conclusion, this book provides an overview of the CART program, which can be obtained by writing to one of the authors. It is written well enough to make the interested reader willing to try the method on his or her data but, unfortunately, not enough examples are provided to enable him to thoroughly understand the details of the methodology. The CART program will be of interest to readers of this journal who are seeking an alternative to standard statistical methods (i.e. linear discriminant analysis and linear regression) which is applicable to nonstandard data sets (e.g., those with incomplete observations and those in which the relationships among the variables differ from group to group).

Dan H. Moore II
Biomedical Sciences
Lawrence Livermore National Laboratory
Livermore, California