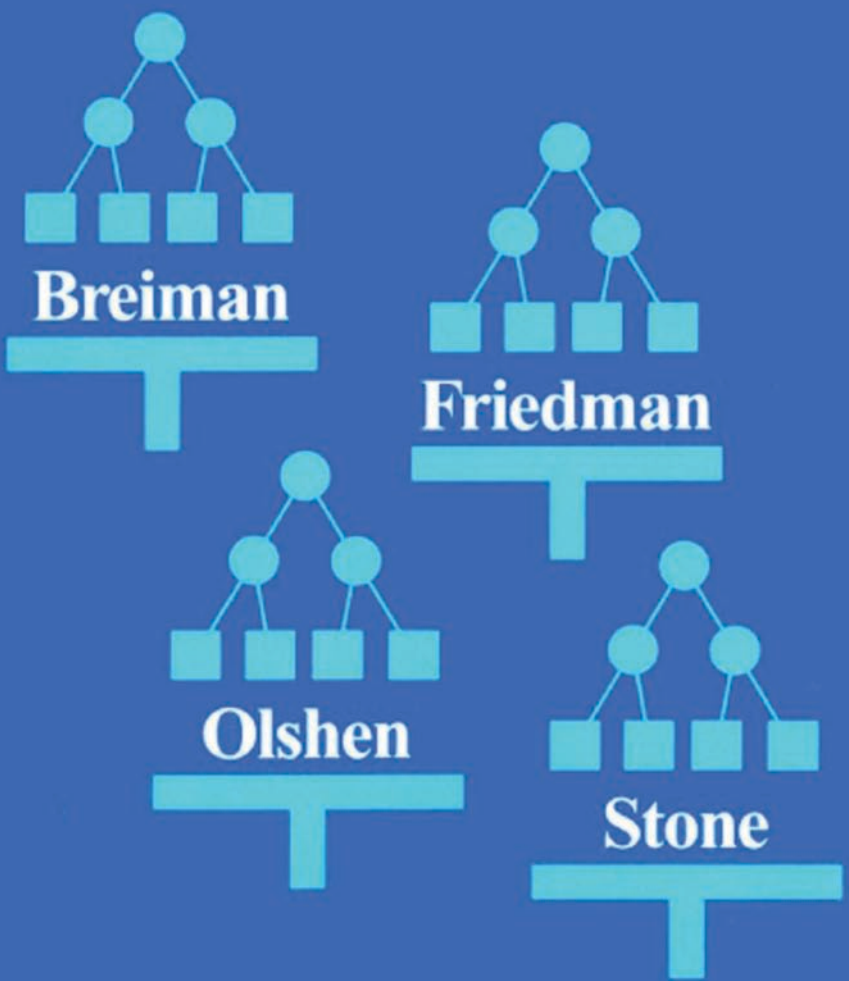
CLASSIFICATION AND REGRESSION TREES



CLASSIFICATION AND REGRESSION TREES

Lovingly dedicated to our children
Jessica, Rebecca, Kymm;
Melanie;
Elyse, Adam, Rachel, Stephen;
Daniel and Kevin

CLASSIFICATION AND REGRESSION TREES

Leo Breiman
University of California, Berkeley

Jerome H. Friedman
Stanford University

Richard A. Olshen
Stanford University

Charles J. Stone
University of California, Berkeley

CHAPMAN & HALL/CRC

Boca Raton London New York Washington, D.C.

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 1984 by Taylor & Francis Group, LLC
Chapman & Hall/CRC is an imprint of Taylor & Francis Group
Originally published by Chapman & Hall

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
20 19 18 17 16 15 14 13 12 11
International Standard Book Number-13: 978-0-412-04841-8 (Softcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Except as permitted by U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Catalog record is available from the Library of Congress

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

CONTENTS

PREFACE

The tree methodology discussed in this book is a child of the computer age. Unlike many other statistical procedures which were moved from pencil and paper to calculators and then to computers, this use of trees was unthinkable before computers.

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.

Both practical and theoretical sides have been developed in our study of tree methods. The book reflects these two sides. The first eight chapters are largely expository and cover the use of trees as a data analysis method. These were written by Leo Breiman with the exception of Chapter 6 by Richard Olshen. Jerome Friedman developed the software and ran the examples.

Chapters 9 through 12 place trees in a more mathematical context and prove some of their fundamental properties. The first three of these chapters were written by Charles Stone and the last was jointly written by Stone and Olshen.

Trees, as well as many other powerful data analytic tools (factor analysis, nonmetric scaling, and so forth) were originated

by social scientists motivated by the need to cope with actual problems and data. Use of trees in regression dates back to the AID (Automatic Interaction Detection) program developed at the Institute for Social Research, University of Michigan, by Morgan and Sonquist in the early 1960s. The ancestor classification program is THAID, developed at the institute in the early 1970s by Morgan and Messenger. The research and developments described in this book are aimed at strengthening and extending these original methods.

Our work on trees began in 1973 when Breiman and Friedman, independently of each other, "reinvented the wheel" and began to use tree methods in classification. Later, they joined forces and were joined in turn by Stone, who contributed significantly to the methodological development. Olshen was an early user of tree methods in medical applications and contributed to their theoretical development.

Our blossoming fascination with trees and the number of ideas passing back and forth and being incorporated by Friedman into CART (Classification and Regression Trees) soon gave birth to the idea of a book on the subject. In 1980 conception occurred. While the pregnancy has been rather prolonged, we hope that the baby appears acceptably healthy to the members of our statistical community.

The layout of the book is

Chapters 1 to 5	Tree structured methodology in classification
Chapters 6, 7	Examples of trees used in classification
Chapters 8	Use of trees in regression
Chapters 9 to 12	Theoretical framework for tree structured methods.

Readers are encouraged to contact Richard Olshen regarding the availability of CART software.

ACKNOWLEDGMENTS

Three other people were instrumental in our research: William Meisel, who early on saw the potential in tree structured methods and encouraged their development; Laurence Rafsky, who participated in some of the early exchanges of ideas; and Louis Gordon, who collaborated with Richard Olshen in theoretical work. Many helpful comments were supplied by Peter Bickel, William Eddy, John Hartigan, and Paul Tukey, who all reviewed an early version of the manuscript.

Part of the research, especially that of Breiman and Friedman, was supported by the Office of Naval Research (Contract No. N00014-82-K-0054), and we appreciate our warm relations with Edward Wegman and Douglas De Priest of that agency. Stone's work was supported partly by the Office of Naval Research on the same contract and partly by the National Science Foundation (Grant No. MCS 80-02732). Olshen's work was supported by the National Science Foundation (Grant No. MCS 79-06228) and the National Institutes of Health (Grant No. CA-26666).

We were fortunate in having the services of typists Ruth Suzuki, Rosaland Englander, Joan Pappas, and Elaine Morici, who displayed the old-fashioned virtues of patience, tolerance, and competence.

We are also grateful to our editor, John Kimmel of Wadsworth, for his abiding faith that eventually a worthy book would emerge, and to the production editor, Andrea Cava, for her diligence and skillful supervision.

1

BACKGROUND

At the University of California, San Diego Medical Center, when a heart attack patient is admitted, 19 variables are measured during the first 24 hours. These include blood pressure, age, and 17 other ordered and binary variables summarizing the medical symptoms considered as important indicators of the patient's condition.

The goal of a recent medical study (see Chapter 6) was the development of a method to identify high risk patients (those who will not survive at least 30 days) on the basis of the initial 24-hour data.

Figure 1.1 is a picture of the tree structured classification rule that was produced in the study. The letter *F* means not high risk; *G* means high risk.

This rule classifies incoming patients as *F* or *G* depending on the yes-no answers to at most three questions. Its simplicity raises the suspicion that standard statistical classification methods may give classification rules that are more accurate. When these were tried, the rules produced were considerably more intricate, but less accurate.

The methodology used to construct tree structured rules is the major story of this monograph.

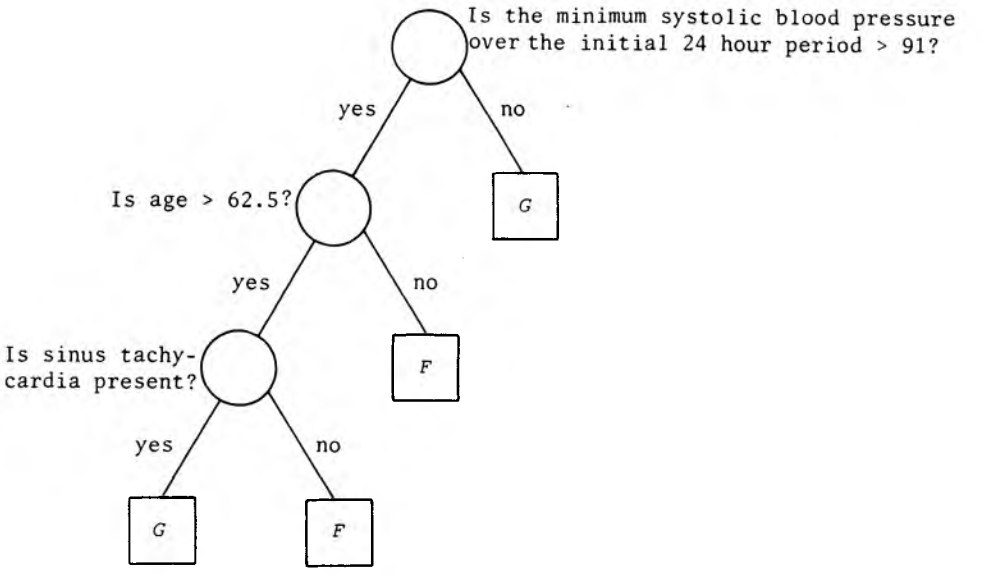


FIGURE 1.1

1.1 CLASSIFIERS AS PARTITIONS

The general classification problem is similar to the medical diagnosis problem sketched above. Measurements are made on some case or object. Based on these measurements, we then want to predict which class the case is in.

For instance, days in the Los Angeles basin are classified according to the ozone levels:

Class 1: nonalert (low ozone)
Class 2: first-stage alert (moderate ozone)
Class 3: second-stage alert (high ozone)

During the current day, measurements are made on many meteorological variables, such as temperature, humidity, upper atmospheric conditions, and on the current levels of a number of airborne pollutants. The purpose of a project funded by the California Air Resources Board (Zeldin and Cassmassi, 1978) was to explore methods for using the current-day measurements to predict the classification of the following day.

An EPA project had this goal: The exact analysis of a complex chemical compound into its atomic constituents is slow and costly. Measuring its mass spectra can be done quickly and at relatively low cost. Can the measured mass spectra be used to accurately predict whether, for example, the compound is in

class 1 (contains one or more chlorine atoms), or

class 2 (contains no chlorine)?

(See Chapter 7 for more discussion.)

In these problems, the goal is the same. Given a set of measurements on a case or object, find a systematic way of predicting what class it is in. In any problem, a *classifier* or a *classification rule* is a systematic way of predicting what class a case is in.

To give a more precise formulation, arrange the set of measurements on a case in a preassigned order; i.e., take the measurements to be x_1, x_2, ..., where, say, x_1 is age, x_2 is blood pressure, etc. Define the measurements $(x_1, x_2, \ldots)$ made on a case as the *measurement vector* $\mathbf{x}$ corresponding to the case. Take the *measurement space* X to be defined as containing all possible measurement vectors.

For example, in the heart attack study, X is a 19-dimensional space such that the first coordinate x_1 (age) ranges, say, over all integer values from 0 to 200; the second coordinate, blood pressure, might be defined as continuously ranging from 50 to 150. There can be a number of different definitions of X. What is important is that any definition of X have the property that the measurement vector $\mathbf{x}$ corresponding to any case we may wish to classify be a point in the space X.

Suppose that the cases or objects fall into J classes. Number the classes 1, 2, ..., J and let C be the set of classes; that is, $C = \{1, \ldots, J\}$.

A systematic way of predicting class membership is a rule that assigns a class membership in C to every measurement vector $\mathbf{x}$ in X. That is, given any $\mathbf{x} \in X$, the rule assigns one of the classes $\{1, \ldots, J\}$ to $\mathbf{x}$.

DEFINITION 1.1. *A classifier or classification rule is a function* $d(\mathbf{x})$ *defined on* X *so that for every* $\mathbf{x}$, $d(\mathbf{x})$ *is equal to one of the numbers* $1, 2, \ldots, J$.

Another way of looking at a classifier is to define A_j as the subset of X on which $d(\mathbf{x}) = j$; that is,

$$A_j = \{\mathbf{x};\ d(\mathbf{x}) = j\}.$$

The sets $A_1, \ldots, A_j$ are disjoint and $X = \bigcup_j A_j$. Thus, the A_j form a partition of X. This gives the equivalent

DEFINITION 1.2. *A classifier is a partition of* X *into* J *disjoint subsets* $A_1, \ldots, A_j$, $X = \bigcup_j A_j$ *such that for every* $\mathbf{x} \in A_j$ *the predicted class is* j.

1.2 USE OF DATA IN CONSTRUCTING CLASSIFIERS

Classifiers are not constructed whimsically. They are based on past experience. Doctors know, for example, that elderly heart attack patients with low blood pressure are generally high risk. Los Angelenos know that one hot, high pollution day is likely to be followed by another.

In systematic classifier construction, past experience is summarized by a *learning sample*. This consists of the measurement data on N cases observed in the past together with their actual classification.

In the medical diagnostic project the learning sample consisted of the records of 215 heart attack patients admitted to the hospital, all of whom survived the initial 24-hour period. The records contained the outcome of the initial 19 measure-

ments together with an identification of those patients that did not survive at least 30 days.

The learning sample for the ozone classification project contained 6 years (1972-1977) of daily measurements on over 400 meteorological variables and hourly air pollution measurements at 30 locations in the Los Angeles basin.

The data for the chlorine project consisted of the mass spectra of about 30,000 compounds having known molecular structure. For each compound the mass spectra can be expressed as a measurement vector of dimension equal to the molecular weight. The set of 30,000 measurement vectors was of variable dimensionality, ranging from about 50 to over 1000.

We assume throughout the remainder of this monograph that the construction of a classifier is based on a learning sample, where

DEFINITION 1.3. *A learning sample consists of data* $(\mathbf{x}_1, j_1)$, ..., $(\mathbf{x}_N, j_N)$ *on* N *cases where* $\mathbf{x}_n \in \mathcal{X}$ *and* $j_n \in \{1, \ldots, J\}$, $n = 1, \ldots, N$. *The learning sample is denoted by* $\mathcal{L}$; i.e.,

$$\mathcal{L} = \{(\mathbf{x}_1, j_1), \ldots, (\mathbf{x}_N, j_N)\}.$$

We distinguish two general types of variables that can appear in the measurement vector.

DEFINITION 1.4. *A variable is called ordered or numerical if its measured values are real numbers. A variable is categorical if it takes values in a finite set not having any natural ordering.*

A categorical variable, for instance, could take values in the set {red, blue, green}. In the medical data, blood pressure and age are ordered variables.

Finally, define

DEFINITION 1.5. *If all measurement vectors* $\mathbf{x}_n$ *are of fixed dimensionality, we say that the data have standard structure.*

In the medical and ozone projects, a fixed set of variables is measured on each case (or day); the data have standard structure. The mass spectra data have nonstandard structure.

1.3 THE PURPOSES OF CLASSIFICATION ANALYSIS

Depending on the problem, the basic purpose of a classification study can be either to produce an accurate classifier or to uncover the predictive structure of the problem. If we are aiming at the latter, then we are trying to get an understanding of what variables or interactions of variables drive the phenomenon—that is, to give simple characterizations of the conditions (in terms of the measurement variables $\mathbf{x} \in X$) that determine when an object is in one class rather than another. These two are not exclusive. Most often, in our experience, the goals will be both accurate prediction and understanding. Sometimes one or the other will have greater emphasis.

In the mass spectra project, the emphasis was on prediction. The purpose was to develop an efficient and accurate on-line algorithm that would accept as input the mass spectrum of an unknown compound and classify the compound as either chlorine containing or not.

The ozone project shared goals. The work toward understanding which meteorological variables and interactions between them were associated with alert-level days was an integral part of the development of a classifier.

The tree structured classification rule of Figure 1.1 gives some interesting insights into the medical diagnostic problem. All cases with blood pressure less than or equal to 91 are predicted high risks. For cases with blood pressure greater than 91, the classification depends only on age and whether sinus tachycardia is present. For the purpose of distinguishing between high and low

risk cases, once age is recorded, only two variables need to be measured.

An important criterion for a good classification procedure is that it not only produce accurate classifiers (within the limits of the data) but that it also *provide insight and understanding into the predictive structure of the data.*

Many of the presently available statistical techniques were designed for small data sets having standard structure with all variables of the same type; the underlying assumption was that the phenomenon is homogeneous. That is, that the same relationship between variables held over all of the measurement space. This led to models where only a few parameters were necessary to trace the effects of the various factors involved.

With large data sets involving many variables, more structure can be discerned and a variety of different approaches tried. But largeness by itself does not necessarily imply a richness of structure.

What makes a data set interesting is not only its size but also its complexity, where complexity can include such considerations as:

High dimensionality
A mixture of data types
Nonstandard data structure

and, perhaps most challenging, nonhomogeneity; that is, different relationships hold between variables in different parts of the measurement space.

Along with complex data sets comes "the curse of dimensionality" (a phrase due to Bellman, 1961). The difficulty is that the higher the dimensionality, the sparser and more spread apart are the data points. Ten points on the unit interval are not distant neighbors. But 10 points on a 10-dimensional unit rectangle are like oases in the desert.

For instance, with 100 points, constructing a 10-cell histogram on the unit interval is a reasonable procedure. In M dimen-

sions, a histogram that uses 10 intervals in each dimension produces 10^M cells. For even moderate M, a very large data set would be needed to get a sensible histogram.

Another way of looking at the "curse of dimensionality" is the number of parameters needed to specify distributions in M dimensions:

Normal: $O(M^2)$
Binary: $O(2^M)$

Unless one makes the very strong assumption that the variables are independent, the number of parameters usually needed to specify an M-dimensional distribution goes up much faster than $O(M)$. To put this another way, *the complexity of a data set increases rapidly with increasing dimensionality.*

With accelerating computer usage, complex, high dimensional data bases, with variable dimensionality or mixed data types, nonhomogeneities, etc., are no longer odd rarities.

In response to the increasing dimensionality of data sets, the most widely used multivariate procedures all contain some sort of dimensionality reduction process. Stepwise variable selection and variable subset selection in regression and discriminant analysis are examples.

Although the drawbacks in some of the present multivariate reduction tools are well known, they are a response to a clear need. To analyze and understand complex data sets, methods are needed which in some sense select salient features of the data, discard the background noise, and feed back to the analyst understandable summaries of the information.

1.4 ESTIMATING ACCURACY

Given a classifier, that is, given a function $d(\mathbf{x})$ defined on X taking values in C, we denote by $R^*(d)$ its "*true misclassifica-*

tion rate." The question raised in this section is: What is truth and how can it be estimated?

One way to see how accurate a classifier is (that is, to estimate $R^*(d)$) is to test the classifier on subsequent cases whose correct classification has been observed. For instance, in the ozone project, the classifier was developed using the data from the years 1972-1975. Then its accuracy was estimated by using the 1976-1977 data. That is, $R^*(d)$ was estimated as the proportion of days in 1976-1977 that were misclassified when $d(\mathbf{x})$ was used on the previous day data.

In one part of the mass spectra project, the 30,000 spectra were randomly divided into one set of 20,000 and another of 10,000. The 20,000 were used to construct the classifier. The other 10,000 were then run through the classifier and the proportion misclassified used as an estimate of $R^*(d)$.

The value of $R^*(d)$ can be conceptualized in this way: Using $\mathcal{L}$, construct d. Now, draw another very large (virtually infinite) set of cases from the same population as $\mathcal{L}$ was drawn from. Observe the correct classification for each of these cases, and also find the predicted classification using $d(\mathbf{x})$. The proportion misclassified by d is the value of $R^*(d)$.

To make the preceding concept precise, a probability model is needed. Define the space $X \times C$ as a set of all couples $(\mathbf{x}, j)$ where $\mathbf{x} \in X$ and j is a class label, $j \in C$. Let $P(A, j)$ be a probability on $X \times C$, $A \subset X$, $j \in C$ (niceties such as Borel measurability will be ignored). The interpretation of $P(A, j)$ is that a case drawn at random from the relevant population has probability $P(A, j)$ that its measurement vector $\mathbf{x}$ is in A and its class is j. Assume that the learning sample $\mathcal{L}$ consists of N cases $(\mathbf{x}_1, j_1)$, $\ldots$, $(\mathbf{x}_N, j_N)$ independently drawn at random from the distribution $P(A, j)$. Construct $d(\mathbf{x})$ using $\mathcal{L}$. Then define $R^*(d)$ as the probability that d will misclassify a new sample drawn from the same distribution as $\mathcal{L}$.

DEFINITION 1.6 *Take* $(\mathbf{X}, Y)$, $\mathbf{X} \in \mathcal{X}$, $Y \in \mathcal{C}$, *to be a new sample from the probability distribution* $P(A, j)$; *i.e.*,

(i) $P(\mathbf{X} \in A, Y = j) = P(A, j)$,

(ii) $(\mathbf{X}, Y)$ *is independent of* $\mathcal{L}$.

Then define

$$R^*(d) = P(d(\mathbf{X}) \neq Y).$$

In evaluating the probability $P(d(\mathbf{X}) \neq Y)$, the set $\mathcal{L}$ is considered fixed. A more precise notation is $P(d(\mathbf{X}) \neq Y|\mathcal{L})$, the probability of misclassifying the new sample given the learning sample $\mathcal{L}$.

This model must be applied cautiously. Successive pairs of days in the ozone data are certainly not independent. Its usefulness is that it gives a beginning conceptual framework for the definition of "truth."

How can $R^*(d)$ be estimated? There is no difficulty in the examples of simulated data given in this monograph. The data in $\mathcal{L}$ are sampled independently from a desired distribution using a pseudo-random number generator. After $d(\mathbf{x})$ is constructed, 5000 additional cases are drawn from the same distribution independently of $\mathcal{L}$ and classified by d. The proportion misclassified among those 5000 is the estimate of $R^*(d)$.

In actual problems, only the data in $\mathcal{L}$ are available with little prospect of getting an additional large sample of classified cases. Then $\mathcal{L}$ must be used both to construct $d(\mathbf{x})$ and to estimate $R^*(d)$. We refer to such estimates of $R^*(d)$ as *internal estimates*. A summary and large bibliography concerning such estimates is in Toussaint (1974).

Three types of internal estimates will be of interest to us. The first, least accurate, and most commonly used is the *resubstitution* estimate.

After the classifier d is constructed, the cases in $\mathcal{L}$ are run through the classifier. The proportion of cases misclassified is the resubstitution estimate. To put this in equation form:

DEFINITION 1.7. *Define the indicator function* $\chi(\cdot)$ *to be 1 if the statement inside the parentheses is true, otherwise zero.*

The resubstitution estimate, denoted $R(d)$, is

$$R(d) = \frac{1}{N} \sum_{n=1}^{N} \chi(d(\mathbf{x}_n) \neq j_n). \tag{1.8}$$

The problem with the resubstitution estimate is that it is computed using the same data used to construct d, instead of an independent sample. All classification procedures, either directly or indirectly, attempt to minimize $R(d)$. Using the subsequent value of $R(d)$ as an estimate of $R^*(d)$ can give an overly optimistic picture of the accuracy of d.

As an exaggerated example, take $d(\mathbf{x})$ to be defined by a partition $A_1, \ldots, A_j$ such that A_j contains all measurement vectors $\mathbf{x}_n$ in $\mathcal{L}$ with $j_n = j$ and the vectors $\mathbf{x} \in \mathcal{X}$ not equal to some $\mathbf{x}_n$ are assigned in an arbitrary random fashion to one or the other of the A_j. Then $R(d) = 0$, but it is hard to believe that $R^*(d)$ is anywhere near zero.

The second method is *test sample* estimation. Here the cases in $\mathcal{L}$ are divided into two sets $\mathcal{L}_1$ and $\mathcal{L}_2$. Only the cases in $\mathcal{L}_1$ are used to construct d. Then the cases in $\mathcal{L}_2$ are used to estimate $R^*(d)$. If N_2 is the number of cases in $\mathcal{L}_2$, then the test sample estimate, $R^{ts}(d)$, is given by

$$R^{ts}(d) = \frac{1}{N_2} \sum_{(\mathbf{x}_n, j_n) \in \mathcal{L}_2} \chi(d(\mathbf{x}_n) \neq j_n). \tag{1.9}$$

In this method, care needs to be taken so that the cases in $\mathcal{L}_2$ can be considered as independent of the cases in $\mathcal{L}_1$ and drawn from the same distribution. The most common procedure used to help ensure these properties is to draw $\mathcal{L}_2$ at random from $\mathcal{L}$. Frequently, $\mathcal{L}_2$ is taken as 1/3 of the cases in $\mathcal{L}$, but we do not know of any theoretical justification for this 2/3, 1/3 split.

The test sample approach has the drawback that it reduces effective sample size. In a 2/3, 1/3 split, only 2/3 of the data are used to construct d, and only 1/3 to estimate $R^*(d)$. If the sample size is large, as in the mass spectra problem, this is a minor difficulty, and test sample estimation is honest and efficient.

For smaller sample sizes, another method, called *V-fold cross-validation*, is preferred (see the review by M. Stone, 1977). The cases in $\mathcal{L}$ are randomly divided into V subsets of as nearly equal size as possible. Denote these subsets by $\mathcal{L}_1, \ldots, \mathcal{L}_V$. Assume that the procedure for constructing a classifier can be applied to any learning sample. For every v, $v = 1, \ldots, V$, apply the procedure using as learning sample $\mathcal{L} - \mathcal{L}_v$, i.e., the cases in $\mathcal{L}$ not in $\mathcal{L}_v$, and let $d^{(v)}(\mathbf{x})$ be the resulting classifier. Since none of the cases in $\mathcal{L}_v$ has been used in the construction of $d^{(v)}$, a test sample estimate for $R^*(d^{(v)})$ is

$$R^{ts}(d^{(v)}) = \frac{1}{N_v} \sum_{(\mathbf{x}_n, j_n) \in \mathcal{L}_v} \chi(d^{(v)}(\mathbf{x}_n) \neq j_n), \tag{1.10}$$

where $N_v \cong N/V$ is the number of cases in $\mathcal{L}_v$. Now using the same procedure again, construct the classifier d using all of $\mathcal{L}$.

For V large, each of the V classifiers is constructed using a learning sample of size $N(1 - 1/V)$ nearly as large as $\mathcal{L}$. The basic assumption of cross-validation is that the procedure is "stable." That is, that the classifiers $d^{(v)}$, $v = 1, \ldots, V$, each constructed using almost all of $\mathcal{L}$, have *misclassification rates* $R^*(d^{(v)})$ *nearly equal to* $R^*(d)$. Guided by this heuristic, define the V-fold cross-validation estimate $R^{CV}(d)$ as

$$R^{CV}(d) = \frac{1}{V} \sum_{v=1}^{V} R^{ts}(d^{(v)}). \tag{1.11}$$

N-fold cross-validation is the "leave-one-out" estimate. For each n, $n = 1, \ldots, N$, the nth case is set aside and the classifier constructed using the other $N - 1$ cases. Then the nth case is used as a single-case test sample and $R^*(d)$ estimated by (1.11).

Cross-validation is parsimonious with data. Every case in $\mathcal{L}$ is used to construct d, and every case is used exactly once in a test sample. In tree structured classifiers tenfold cross-validation has been used, and the resulting estimators have been satisfactorily close to $R^*(d)$ on simulated data.

The *bootstrap* method can also be used to estimate $R^*(d)$, but may not work well when applied to tree structured classifiers (see Section 11.7).

1.5 THE BAYES RULE AND CURRENT CLASSIFICATION PROCEDURES

The major guide that has been used in the construction of classifiers is the concept of the Bayes rule. If the data are drawn from a probability distribution $P(A, j)$, then the form of the most accurate rule can be given in terms of $P(A, j)$. This rule is called the *Bayes rule* and is denoted by $d_B(\mathbf{x})$.

To be more precise, suppose that $(\mathbf{X}, Y)$, $\mathbf{X} \in \mathcal{X}$, $Y \in \mathcal{C}$, is a random sample from the probability distribution $P(A, j)$ on $\mathcal{X} \times \mathcal{C}$; i.e., $P(\mathbf{X} \in A, Y = j) = P(A, j)$.

DEFINITION 1.12. *$d_B(\mathbf{x})$ is a Bayes rule if for any other classifier $d(\mathbf{x})$,*

$$P(d_B(\mathbf{X}) \neq Y) \leq P(d(\mathbf{X}) \neq Y).$$

Then the Bayes misclassification rate is

$$R_B = P(d_B(\mathbf{X}) \neq Y).$$

To illustrate how $d_B(\mathbf{x})$ can be derived from $P(A, j)$, we give its form in an important special case.

DEFINITION 1.13. *Define the prior class probabilities $\pi(j)$, $j = 1, \ldots, J$, as*

$$\pi(j) = P(Y = j)$$

and the probability distribution of the jth class measurement vectors by

$$P(A|j) = P(A, j)/\pi(j).$$

ASSUMPTION 1.14. *X is M-dimensional euclidean space and for every j, $j = 1, \ldots, J$, $P(A|j)$ has the probability density $f_j(\mathbf{x})$; i.e., for sets $A \subset X$,*

$$P(A|j) = \int_A f_j(\mathbf{x})d\mathbf{x}.$$

Then,

THEOREM 1.15. *Under Assumption 1.14 the Bayes rule is defined by*

$$d_B(\mathbf{x}) = j \text{ on } A_j = \{\mathbf{x};\ f_j(\mathbf{x})\pi(j) = \max_i f_i(\mathbf{x})\pi(i)\} \tag{1.16}$$

and the Bayes misclassification rate is

$$R_B = 1 - \int \max_j [f_j(\mathbf{x})\pi(j)]d\mathbf{x}. \tag{1.17}$$

Although d_B is called the Bayes rule, it is also recognizable as a maximum likelihood rule: Classify $\mathbf{x}$ as that j for which $f_j(\mathbf{x})\pi(j)$ is maximum. As a minor point, note that (1.16) does not uniquely define $d_B(\mathbf{x})$ on points $\mathbf{x}$ such that $\max_j f_j(\mathbf{x})\pi(j)$ is achieved by two or more different j's. In this situation, define $d_B(\mathbf{x})$ arbitrarily to be any one of the maximizing j's.

The proof of Theorem 1.15 is simple. For any classifier d, under Assumption 1.14,

$$\begin{aligned} P(d(\mathbf{X}) = Y) &= \textstyle\sum_{j=1}^{J} P(d(\mathbf{X}) = j|Y = j)\pi(j) \\ &= \textstyle\sum_{j=1}^{J} \int_{\{d(\mathbf{x})=j\}} f_j(\mathbf{x})\pi(j)d\mathbf{x} \\ &= \int [\textstyle\sum_{j=1}^{J} \chi(d(\mathbf{x}) = j) f_j(\mathbf{x})\pi(j)]d\mathbf{x}. \end{aligned}$$

For a fixed value of $\mathbf{x}$

$$\textstyle\sum_{j=1}^{J} \chi(d(\mathbf{x}) = j) f_j(\mathbf{x})\pi(j) \le \max_j [f_j(\mathbf{x})\pi(j)],$$

and equality is achieved if $d(\mathbf{x})$ equals that j for which $f_j(\mathbf{x})\pi(j)$ is a maximum. Therefore, the rule d_B given in (1.16) has the property that for any other classifier d,

$$P(d(\mathbf{X}) = Y) \leq P(d_B(\mathbf{X}) = Y) = \int \max_j [f_j(\mathbf{x})\pi(j)]d\mathbf{x}.$$

This shows that d_B is a Bayes rule and establishes (1.17) as the correct equation for the Bayes misclassification rate.

In the simulated examples we use later on, the data are generated from a known probability distribution. For these examples, d_B was derived and then the values of R_B computed. Since R_B is the minimum misclassification rate attainable, knowing R_B and comparing it with the accuracy of the tree structured classifiers give some idea of how effective they are.

In practice, neither the $\pi(j)$ nor the $f_j(\mathbf{x})$ are known. The $\pi(j)$ can either be estimated as the proportion of class j cases in $\mathcal{L}$ or their values supplied through other knowledge about the problem. The thorny issue is getting at the $f_j(\mathbf{x})$. The three most commonly used classification procedures

Discriminant analysis
Kernel density estimation
Kth nearest neighbor

attempt, in different ways, to approximate the Bayes rule by using the learning sample $\mathcal{L}$ to get estimates of $f_j(\mathbf{x})$.

Discriminant analysis assumes that all $f_j(\mathbf{x})$ are multivariate normal densities with common covariance matrix Γ and different means vectors $\{\mu_j\}$. Estimating Γ and the μ_j in the usual way gives estimates $\hat{f}_j(\mathbf{x})$ of the $f_j(\mathbf{x})$. These are substituted into the Bayes optimal rule to give the classification partition

$$A_j = \{\mathbf{x};\ \hat{f}_j(\mathbf{x})\pi(j) = \max_i \hat{f}_i(\mathbf{x})\pi(i)\}.$$

A stepwise version of linear discrimination is the most widely used method. It is usually applied without regard to lack of normality. It is not set up to handle categorical variables, and these

are dealt with by the artifice of coding them into dummy variables. Our reaction to seeing the results of many runs on different data sets of this program is one of surprise that it does as well as it does. It provides insight into the structure of the data through the use of the discrimination coordinates (see Gnanadesikan, 1977, for a good discussion). However, the form of classifier for the J class problem (which depends on the maximum of J linear combinations) is difficult to interpret.

Density estimation and kth nearest neighbor methods are more recent arrivals generated, in part, by the observation that not all data sets contained classes that were normally distributed with common covariance matrices.

The density method uses a nonparametric estimate of each of the densities $f_j(\mathbf{x})$, most commonly done using a kernel type of estimate (see Hand, 1982). Briefly, a metric $\|\mathbf{x}\|$ on X is defined and a kernel function $K(\|\mathbf{x}\|) \geq 0$ selected which has a peak at $\|\mathbf{x}\| = 0$ and goes to zero as $\|\mathbf{x}\|$ becomes large, satisfying

$$\int K(\|\mathbf{x}\|)d\mathbf{x} = 1.$$

Then $f_j(\mathbf{x})$ is estimated by

$$\hat{f}_j(\mathbf{x}) = \frac{1}{N_j} \sum K(\|\mathbf{x} - \mathbf{x}_n\|),$$

where N_j is the number of cases in the jth class and the sum is over the N_j measurement vectors $\mathbf{x}_n$ corresponding to cases in the jth class.

The kth nearest neighbor rule, due to Fix and Hodges (1951), has this simple form: Define a metric $\|\mathbf{x}\|$ on X and fix an integer $k > 0$. At any point $\mathbf{x}$, find the k nearest neighbors to $\mathbf{x}$ in $\mathcal{L}$. Classify $\mathbf{x}$ as class j if more of the k nearest neighbors are in class j than in any other class. (This is equivalent to using density estimates for f_j based on the number of class j points among the k nearest neighbors.)

The kernal density estimation and kth nearest neighbor methods make minimal assumptions about the form of the underlying distribution. But there are serious limitations common to both methods.

1. They are sensitive to the choice of the metric $\|\mathbf{x}\|$, and there is usually no intrinsically preferred definition.
2. There is no natural or simple way to handle categorical variables and missing data.
3. They are computationally expensive as classifiers; $\mathcal{L}$ must be stored, the interpoint distances and $d(\mathbf{x})$ recomputed for each new point $\mathbf{x}$.
4. Most serious, they give very little usable information regarding the structure of the data.

Surveys of the literature on these and other methods of classification are given in Kanal (1974) and Hand (1981).

The use of classification trees did not come about as an abstract exercise. Problems arose that could not be handled in an easy or natural way by any of the methods discussed above. The next chapter begins with a description of one of these problems.

References

Alexander, K. S. 1983. Rates of growth for weighted empirical processes. Proceedings of the Neyman-Kiefer Conference. In press.

Anderson, J. A. , and Philips, P. R. 1981. Regression, discrimination and measurement models for ordered categorical variables. Appl. Statist., 30: 22–31.

Anderson, T. W. 1966. Some nonparametric multivariate procedures based on statistically equivalent blocks. In Multivariate analysis, ed. P. R. Krishnaiah . New York: Academic Press, 5–27.

Bellman, R. E. 1961. Adaptive control processes. Princeton, N.J.: Princeton University Press.

Belsley, D. A. ; Kuh, E. ; and Welsch, R. E. 1980. Regression diagnostics. New York: Wiley.

Belson, W. A. 1959. Matching and prediction on the principle of biological classification. Appl. Statist., 8: 65–75.

Beta-Blocker Heart Attack Trial Study Group. 1981. Beta-blocker heart attack trial. J. Amer. Med. Assoc., 246: 2073–2074.

Breiman, L. 1968. Probability. Reading, Mass.: Addison-Wesley.

Breiman, L. 1978. Description of chlorine tree development and use. Technical report. Santa Monica, Calif.: Technology Service Corporation.

Breiman, L. 1981. Automatic identification of chemical spectra. Technical report. Santa Monica, Calif.: Technology Service Corporation.

Breiman, L. , and Stone, C. J. 1978. Parsimonious binary classification trees. Technical report. Santa Monica, Calif.: Technology Service Corporation.

Bridges, C. R. 1980. Binary decision trees and the diagnosis of acute myocardial infarction. Masters thesis, Massachusetts Institute of Technology, Cambridge.

Chung, K. L. 1974. A course in probability theory. 2d ed. New York Academic Press.

Collomb, G. 1981. Estimation non parametrique de la regression: review bibliographique. Internat. Statist. Rev., 49: 75–93.

Cover, T. M. , and Hart, P. E. 1967. Nearest neighbor pattern classification. IEEE Trans. Information Theory, IT-13: 21–27.

Darlington, R. B. 1968. Multiple regression in psychological research and practice. Psychological Bull., 69: 161–182.

Dillman, R. O. , and Koziol, J. A. 1983. Statistical approach to immunosuppression classification using lymphocyte surface markers and functional assays. Cancer Res., 43: 417–421.

Dillman, R. O. ; Koziol, J. A. ; Zavanelli, M. I. ; Beauregard, J. C. ; Halliburton, B. L. ; and Royston, I. 1983. Immunocompetence in cancer patients—assessment by *in vitro* stimulation tests and quantification of lymphocyte subpopulations. Cancer. In press.

Doyle, P. 1973. The use of automatic interaction detector and similar search procedures. Operational Res. Quart., 24: 465–467.

Duda, R. O. , and Shortliffe, E. H. 1983. Expert systems research. Science, 220: 261–268.

Dudley, R. M. 1978. Central limit theorems for empirical measures. Ann. Probability, 6: 899–929.

DuMouchel, W. H. 1981. Documentation for DREG. Technical report. Cambridge: Massachusetts Institute of Technology.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. Ann. Statist., 7: 1–26.

Efron, B. 1982. The jackknife, the bootstrap and other resampling plans. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. 1983. Estimating the error rate of a prediction rule: improvements on cross-validation. J. Amer. Statist. Assoc., 78: 316–331.

Einhorn, H. 1972. Alchemy in the behavioral sciences. Pub. Op. Quart., 36: 367–378.

Feinstein, A. 1967. Clinical judgment. Baltimore: Williams and Wilkins.

Fielding, A. 1977. Binary segmentation: the automatic interaction detector and related techniques for exploring data structure. In The analysis of survey data, Vol. I, ed. C. A. O'Muircheartaigh and C. Payne . Chichester: Wiley.

Fisher, W. D. 1958. On grouping for maximum homogeneity. J. Amer. Statist. Assoc., 53: 789–798.

Fix, E. , and Hodges, J. 1951. Discriminatory analysis, nonparametric discrimination: consistency properties. Technical report. Randolph Field, Texas: USAF School of Aviation

Medicine.
Friedman, J. H. 1977. A recursive partitioning decision rule for nonparametric classification. IEEE Trans. Computers, C-26: 404–408.
Friedman, J. H. 1979. A tree-structured approach to nonparametric multiple regression. In Smoothing techniques for curve estimation, eds. T. Gasser and M. Rosenblatt . Berlin: Springer-Verlag.
Gilpin, E. ; Olshen, R. ; Henning, H. ; and Ross, J., Jr. 1983. Risk prediction after myocardial infarction: comparison of three multivariate methodologies. Cardiology, 70: 73–84.
Glick, N. 1978. Additive estimators for probabilities of correct classification. Pattern Recognition, 10: 211–222.
Gnanadesikan, R. 1977. Methods for statistical data analysis of multivariate observations. New York: Wiley.
Goldman, L. ; Weinberg, M. ; Weisberg, M. ; Olshen, R. ; Cook, F. ; Sargent, R. K. ; Lamas, G. A. ; Dennis, C. ; Deckelbaum, L. ; Fineberg, H. ; Stiratelli, R. ; and the Medical Housestaffs at Yale-New Haven Hospital and Brigham and Women's Hospital. 1982. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. New England J. Med., 307: 588–596.
Gordon, L. , and Olshen, R. A. 1978. Asymptotically efficient solutions to the classification problem. Ann. Statist., 6: 515-533.
Gordon, L. , and Olshen, R. A. 1980. Consistent nonparametric regression from recursive partitioning schemes. J. Multivariate Anal., 10: 611–627.
Hand, D. J. 1981. Discrimination and classification. Chichester: Wiley.
Hand, D. J. 1982. Kernel discriminant analysis. Chichester: Research Studies Press, Wiley.
Harrison, D. , and Rubinfeld, D. L. 1978. Hedonic prices and the demand for clean air. J. Envir. Econ. and Management, 5: 81–102.
Henning, H. ; Gilpin, E. A. ; Covell, J. W. ; Swan, E. A. ; O'Rourke, R. A. ; and Ross, J., Jr. 1976. Prognosis after acute myocardial infarction: multivariate analysis of mortality and survival. Circulation, 59: 1124–1136.
Henrichon, E. G. , and Fu, K. -S. 1969. A nonparametric partitioning procedure for pattern classification. IEEE Trans. Compuputers, C-18: 614–624.
Hills, M. 1967. Discrimination and allocation with discrete data. Appl. Statist., 16: 237–250.
Hooper, R. , and Lucero, A. 1976. Radar profile classification: a feasibility study. Technical report. Santa Monica, Calif.: Technology Service Corporation.
Jennrich, R. , and Sampson, P. 1981. Stepwise discriminant analysis. In BMDP statistical software 1981, ed. W. J. Dixon . Berkeley: University of California Press.
Kanal, L. 1974. Patterns in pattern recognition: 1968–1974. IEEE Trans. Information Theory, IT-20: 697–722.
Kiefer, J. 1961. On large deviations of the empiric d.f. of vector chance variables and a law of iterated logarithm. Pacific J. Math., 11: 649–660.
Light, R. J. , and Margolin, B. H. 1971. An analysis of variance for categorical data. J. Amer. Statist. Assoc., 66: 534–544.
Mabbett, A. ; Stone, M. ; and Washbrook, J. 1980. Cross-validatory selection of binary variables in differential diagnosis. Appl. Statist., 29: 198–204.
McCullagh, P. 1980. Regression models for ordinal data. J. Roy. Statist. Soc. Ser. B, 42: 109–142.
McLafferty, F. W. 1973. Interpretation of mass spectra. Reading, Mass.: Benjamin.
Meisel, W. S. 1972. Computer-oriented approaches to pattern recognition. New York: Academic Press.
Meisel, W. S. , and Michalpoulos, D. A. 1973. A partitioning algorithm with application in pattern classification and the optimization of decision trees. IEEE Trans. Computers, C-22: 93–103.
Messenger, R. C. , and Mandell, M. L. 1972. A model search technique predictive nominal scale multivariate analysis. J. Amer. Statist. Assoc., 67: 768–772.
Morgan, J. N. , and Messenger, R. C. 1973. THAID: a sequential search program for the analysis of nominal scale dependent variables. Ann Arbor: Institute for Social Research, University of Michigan.
Morgan, J. N. , and Sonquist, J. A. 1963. Problems in the analysis of survey data, and a proposal. J. Amer. Statist. Assoc., 58: 415–434.

Narula, S. C. , and Wellington, J. F. 1982. The minimum sum of absolute errors regression: a state of the art survey. Internat. Statist. Rev., 50: 317–326·

Norwegian Multicenter Study Group . 1981. Timolol-induced reduction in mortality and reinfarction in patients surviving acute myocardial infarction. New England J. Med., 304: 801–807.

Pollard, D. 1981. Limit theorems for empirical processes, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 57: 181–195.

Pozen, M. W. ; D'Agostino, R. B. ; and Mitchell, J. B. 1980. The usefulness of a predictive instrument to reduce inappropriate admissions to the coronary care unit. Ann. Internal Med., 92: 238–242.

Ranga Rao, R. 1962. Relations between weak and uniform convergence of measures with applications. Ann. Math. Statist., 33: 659-680.

Rounds, E. M. 1980. A combined nonparametric approach to feature selection and binary decision tree design. Pattern Recognition, 12: 313–317.

Shibata, R. 1981. An optimal selection of regression variables. Biometrika, 68: 45–54.

Sklansky, J. 1980. Locally trained piecewise linear classifiers. IEEE Trans. Pattern Analysis Machine Intelligence, PAMI-2: 101–111.

Sonquist, J. A. 1970. Multivariate model building: the validation of a search strategy. Ann Arbor: Institute for Social Research, University of Michigan.

Sonquist, J. A. ; Baker, E. L. ; and Morgan, J. N. 1973. Searching for structure. Rev. ed. Ann Arbor: Institute for Social Research, University of Michigan.

Sonquist, J. A. , and Morgan, J. N. 1964. The detection of interaction effects. Ann Arbor: Institute for Social Research, University of Michigan.

Steele, J. M. 1978. Empirical discrepancies and subadditive processes. Ann. Probability, 6: 118–127.

Stone, C. J. 1977. Consistent nonparametric regression (with discussion). Ann. Statist., 5: 595–645.

Stone, C. J. 1981. Admissible selection of an accurate and parsimonious normal linear regression model. Ann. Statist., 9: 475–485.

Stone, M. 1977. Cross-validation: a review. Math. Operationforsch. Statist. Ser. Statist., 9: 127–139.

Sutherland, D. H. , and Olshen, R. A. 1984. The development of walking. London: Spastics International Medical Publications. To appear.

Sutherland, D. H. ; Olshen, R. ; Cooper, L. ; and Woo, S. L.-Y. 1980. The development of mature gait. J. Bone Joint Surgery, 62A: 336–353.

Sutherland, D. H. ; Olshen, R. ; Cooper, L. ; Wyatt, M. ; Leach, J. ; Mubarak, S. ; and Schultz, P. 1981. The pathomechanics of gait in Duchenne muscular dystrophy. Developmental Med. and Child Neurology, 39: 598–605.

Szolovits, P. 1982. Artificial intelligence in medicine. In Artificial intelligence in medicine, ed. P. Szolovits . Boulder, Colo.: Westview Press.

Toussaint, G. T. 1974. Bibliography on estimation of misclassification. IEEE Trans. Information Theory, IT-20: 472–479.

Van Eck, N. A. 1980. Statistical analysis and data management highlights of OSIRIS IV. Amer. Statist., 34: 119–121.

Vapnik, V. N. ; and Chervonenkis, A. Ya. 1971. On the uniform convergence of relative frequencies of events to their probabilities. Theor. Probability Appl., 16: 264–280.

Zeldin, M. , and Cassmassi, J. 1978. Development of improved methods for predicting air quality levels in the South Coast Air Basin. Technical report. Santa Monica, Calif.: Technology Service Corporation.