PROBABILITY ESTIMATION FOR CLASSIFICATION TREES

BY

DANIEL A. BLOCH, RICHARD A. OLSHEN, AND MICHAEL G. WALKER

TECHNICAL REPORT NO. 200

DECEMBER 1998

PREPARED UNDER THE AUSPICES

OF

PUBLIC HEALTH SERVICE GRANT 2 R01 CA59039-21

DIVISION OF BIOSTATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
Probability Estimation for Classification Trees

By

Daniel A. Bloch, Richard A. Olshen, and Michael G. Walker

Technical Report No. 200

December 1998

Prepared Under the Auspices

Of

Public Health Service Grant

2 R01 CA59039-21

Division of Biostatistics

Stanford University

Stanford, California
PROBABILITY ESTIMATION FOR CLASSIFICATION TREES
Daniel A. Bloch, Richard A. Olshen, and Michael G. Walker

Section 5.4 of Classification and Regression Trees has an estimator of risk within terminal nodes of CART\textsuperscript{TM} classification trees that is due to Leo Breiman. The estimator has two free parameters, and an empirical Bayes method is put forth for estimating them. The book gives some evidence that the method is a good one, but only some; what is more, to the best of our knowledge no argument has been advanced, there or elsewhere, as to why the approach is as worthwhile as it is. Here we give an explanation, part heuristic part mathematics, why it should be successful in the many examples for which it is. In addition, we give numerical evidence from simulations in the two-class case. A variation of Breiman's method in which repeated cross-validation is employed to estimate global rates of misclassification seems to be the most accurate overall from among those techniques we studied. Exceptions occur in cases for which the Bayes risk of the Bayes rule is small, for example when the unconditional probability of misclassification is less than .1 and also is much less than half the expected loss computed from prior probabilities and misclassification costs alone (the "no data Bayes rule"). For them, either a local bootstrap .632 estimate or Breiman's method modified to use a bootstrap estimate of the global misclassification rate is most accurate although the Breiman variant using repeated cross-validation is competitive for these data sets as well.

Daniel A. Bloch and Richard A. Olshen are with the Division of Biostatistics, Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305-5405. Bloch is dbloch@leland.stanford.edu; Olshen is olshen@stat.stanford.edu. Michael G. Walker is with the Program in Medical Information Sciences, Stanford University School of Medicine, Stanford, CA 94305-5479. He is mwalker@incyte.com.
INTRODUCTION

Classification trees are formed by successively partitioning a feature space $\mathcal{X}$ of predictors $\mathbf{X}$ of an outcome of $Y$. The range of $Y$ is denoted by $\mathcal{Y}$; it is finite. Partitioning is based upon a learning sample $\mathcal{L}$ of $(\mathbf{X}, Y)$ pairs. The successive application of an algorithm for partitioning $\mathcal{X}$ and its subsets entails that the process can be represented by a tree, $\mathcal{T}$, which is binary if the partitioning is dichotomous. Terminal nodes $\mathcal{T}$, that is to say leaves of $\mathcal{T}$, correspond to undivided members of a partition of $\mathcal{X}$. A Bayes rule is estimated from $\mathcal{L}$ for each node/leaf $t \in \mathcal{T}$. Thereby, $Y$, if unknown, can be predicted for any subsequent $\mathbf{X}$. Implicitly if not explicitly, the process involves estimated or given “prior” probabilities, the marginal distribution of $Y$, and estimates of the conditional distribution of $\mathbf{X}$ given $Y$. Any function that maps $\mathcal{X}$ to $\mathcal{Y}$ is a classifier, or a classification rule, or simply a rule.

Our focus here is upon recursive partitioning CART™ methodology whereby a large binary tree is grown so as to successively “purify” nodes and then pruned back to a smaller one with best possible estimated Bayes risk of the resulting classifier. CART’s splits are by hyperplanes when features are Euclidean. Its rules are constant on terminal nodes. CART is one of many widely used, clearly adaptive techniques for classification. To be precise, a rule is adaptive if it depends functionally upon $\mathcal{L}$ through its $Y$ as well as its $\mathbf{X}$s. At this writing there is increasing availability of large sets of data with their accompanying fertile areas for application, and inexpensive computation. The general acceptance of careful data mining in the scientific, technologic, medical, and business communities is also increasing. To be deemed “careful” it seems mandatory that suppliers of rules also make every effort to supply technologies that allow users to estimate well how accurate the rules will be when applied to a subsequent $\mathbf{X}$ with $Y$ unknown. Adaptiveness entails bias in the resubstitution estimate of risk, that is, the risk estimated by applying the classifier to $\mathcal{L}$ itself. We report here on techniques whereby risk is estimated for CART, especially “locally” in terminal nodes. They can be adapted to allow simple, approximately unbiased estimates of risk for any subsets of $\mathcal{X}$. Our methods include empirical Bayes, repeated cross-validation, and bootstrapping. Monte Carlo simulations with synthetic data sets for which $|\mathcal{Y}| = 2$ show these methods to be substantially more accurate than is resubstitution. Leo Breiman proposed in (Breiman et al. 1984 – see Section 5.4) a particular empirical Bayes-like approach that works very well. Its functional form was unmotivated in the book. One of our contributions here is a new argument, part heuristic part mathematics, from which it is clear just why Breiman’s method should be successful in the many examples for which it is.
Indeed, a variation of it in which repeated cross-validation is employed to estimate global rates of misclassification seems to be most accurate overall. Exceptional cases are those for which the Bayes risk of the Bayes rule is small, for example when the unconditional probability of misclassification is less than .1 and also is much less than half the Bayes risk of the "no data" Bayes rule. For them either a local bootstrap .632 estimate or Breiman's method modified to use a bootstrap estimate of the global misclassification rate is most accurate, although the Breiman variant using repeated cross-validation is competitive for these data sets as well. In recent work Efron and Tibshirani (1997) have shown that in some examples their .632+ rule is slightly more accurate than is the .632 rule. They offer other rules as well. We have not explored their new rules in our contexts.

Readers may ask why we work hard to justify improving estimates of local risk for CART when it is clear by now that with bagging (Breiman, 1996), boosting (Freund and Schapire, 1997), and resampling within nodes (Dannegger, 1999), we can use standard CART as a springboard to producing more accurate classification. Our reply has two aspects. First, CART is widely used and has had its successes (for example, Goldman et al., 1988, Bloch and Segal, 1989, and much else). It is likely to continue to be used not only because it can produce accurate classification when subsets of feature space defined by the Bayes rule can be expressed as unions of polytopes with sides parallel to the coordinate axes, but maybe especially because its rules are so simple to understand. Our experiences are that consumers of computer-intensive statistics find the simplicity of the CART tree – in its construction and its final rendering – so attractive that often they prefer CART even when better accuracy is available with less easily understood techniques. Second, the justification for Breiman's method and its empirical Bayes estimates of local risk does not rely on the classifier coming from CART at all. Any method that is non-randomized in its assignment of observations to classes and that therefore partitions a feature space into subsets within each of which class assignment is constant is a candidate for the methods studied here.

2. RESUBSTITUTION, CROSS-VALIDATION, REPEATED CROSS-VALIDATION, AND THE .632 BOOTSTRAP

2.1 CROSS-VALIDATION AND REPEATED CROSS-VALIDATION

We describe these techniques for two-class problems; they have obvious extensions to N-class problems. Prior probabilities are \( \pi(i), i = 1, 2; \) and misclassification costs are
$C(i|j)$, the cost of classifying an observation to class $i$ given it is of class $j$; $C(i|j) \geq 0; C(i|i) \equiv 0$. The entire set of data from which the tree is developed is the learning sample; it is denoted by $\mathcal{L}$. The cardinality of the learning sample is $n = |\mathcal{L}|$. We study repeated $K$-fold cross validation, where $K < n$; typically, $K = 10$. The resubstitution risk of a decision rule $d$ applied to $\mathcal{L} = \{(X_i, Y_i) : i = 1, \ldots, n\}$ is

$$R_{\text{RES}}(d) = \sum_{i=1}^{2} \pi(i)(n_i)^{-1} \sum_{(j, Y_j = i)} C(d(X_j)|i),$$

(1)

where

$$n_i = \sum_{j=1}^{n} I[Y_j = i],$$

the number of learning sample observations $(X_j, Y_j)$ for which $Y_j = i$. In other contexts $R_{\text{RES}}(d)$ specializes to the apparent error rate. See (Efron, 1983, Efron and Tibshirani, 1993; Efron and Tibshirani, 1997). With ordinary $K$-fold cross-validation, $\mathcal{L} = \bigcup_{k=1}^{K} \mathcal{L}_k$, $\mathcal{L}_k \cap \mathcal{L}_l = \emptyset$ for $k \neq l$, and $\max\{|\mathcal{L}_k|\} - \min\{|\mathcal{L}_k|\} \leq 1$. We write $d^{(k)}$ for the decision rule based on $\bigcup_{l \neq k} \mathcal{L}_l$, and the same algorithm used in computing $d$. Then, $(X_j, Y_j) \in \mathcal{L}_k$ with $Y_j = i$ incurs loss $C(d^{(k)}(X_j)|i)$, and we write

$$R^{(k)} = \sum_{i=1}^{2} \pi(i)(n_i^{(k)})^{-1} \sum_{\mathcal{L}_k \cap \{j: Y_j = i\}} C(d^{(k)}(X_j)|i),$$

(2)

where

$$n_i^{(k)} = \sum_{\mathcal{L}_k} I[Y_j = i].$$

Then the cross-validatory estimate of risk, $R_{\text{CV}}(d) = R_{\text{CV}}$ is

$$R_{\text{CV}} = K^{-1} \sum_{k=1}^{K} R^{(k)}.$$  

(3)

For $|\mathcal{L} \setminus \mathcal{L}_k| < |\mathcal{L}|$, $R_{\text{CV}}$ is (but for differences in sample size) an unbiased estimate of the risk of $d$. Note that $(X_j, Y_j) \in \mathcal{L}$ figures in only one $\mathcal{L}_k$, so in only one $R^{(k)}$, so only once in $R_{\text{CV}}$. In particular, if $(X_j, Y_j) \in \mathcal{L}_k$ and $Y_j = i$, then $(X_j, Y_j)$ contributes $K^{-1} \pi(i)(n_i^{(k)})^{-1} C(d^{(k)}(X_j)|i) = \hat{r}((X_j, Y_j)) = \hat{r}_j$ to $\hat{R}_{\text{CV}}$. We can rewrite $R_{\text{CV}}$ as

$$R_{\text{CV}} = \sum_{j=1}^{n} \hat{r}_j.$$  

(4)

From (4) it follows that, with $K$-fold cross-validation, we have implicitly imputed a risk $(\hat{r}_j)$ to each member of the learning sample. Because $\hat{r}_j$ is based on only a single number, its sample variance is “large.” Since $K < n = |\mathcal{L}|$, if we again randomly subdivide $\mathcal{L}$ into $K$ disjoint subsets as equal as possible in size, with probability near 1 we will obtain different subdivisions, and a different estimate $\hat{r}_j$; so we will obtain a different $R_{\text{CV}}$. By repeating this process and averaging, we diminish the (conditional) variance (given $\mathcal{L}$) of each $\hat{r}_j$ (by order $1/\#$ repetitions), and of $R_{\text{CV}}$, without changing the bias. What is more, for any (measurable) subset $S$ of the feature space $\mathcal{X}$, the common range of the $X_j$, we can estimate the contribution to the risk of $d$ from $S \subset \mathcal{X}$ by $\sum_{X_j \in S} \hat{r}_j$. $S$s of special interest
are those that correspond to terminal nodes of the tree associated with CART rule \( d \), and \( X \) itself.

With repeated cross-validation, all the original data points are used to build the tree; however, the validation is based on repeatedly dividing the entire data set to create separate learning and test sets. For related procedures on different problems see Zhang 1993, Burman 1989, and Burman 1990.

### 2.2 Trimmed Repeated Cross-Validation

The procedure for the trimmed repeated cross-validation method of estimation is the same as that for repeated cross-validation, except that the most extreme 10% or 50% of the single cross-validations are trimmed before the repeated cross-validation estimate is calculated.

### 2.3 Bootstrap 0.632

We turn now to a description of the bootstrap .632 approach. Our presentation is different from the usual one because its formulation allows user specified or empirical prior distributions. Beyond that, our algorithm differs ever so slightly from the usual in a way we make precise. See Efron (1982, 1983), Efron and Gong (1983), Efron and Tibshirani (1993, 1997). Our .632 methodology closely resembles cross-validation and repeated cross-validation in their creation of separate learning and test samples from the original learning sample \( \mathcal{L} \). However, there are differences in their procedures for creating these samples and also in how risk is estimated. We describe one iteration of our bootstrap .632, in practice what one computes on the basis of a single bootstrap sample. What one reports is the numerical average of the results of multiple bootstrap samples.

We require some further notation. Thus, by \( \mathcal{L}^* \) we mean a bootstrap sample taken with replacement from \( \mathcal{L} \) and of the same cardinality as that of \( \mathcal{L} \). It is easy to compute that for \( |\mathcal{L}| \) large, about 100(1 - \( e^{-1} \))% \( \approx 63.2\% \) of \( \mathcal{L} \) will appear at least once in \( \mathcal{L}^* \). Those \( (X, Y) \) pairs in both \( \mathcal{L} \) and \( \mathcal{L}^* \) comprise \( \mathcal{L}^{LS} \), the bootstrap learning sample. If \( (X, Y) \in \mathcal{L} \setminus \mathcal{L}^* \), then \( (X, Y) \in \mathcal{L}^{TS} \), the bootstrap test sample. For \( i = 1, 2 \) define

\[
n_i^{LS} = \sum_{L^{LS}} I_{[Y=\hat{Y}]}(X)
\]

and define \( n_i^{TS}, i = 1, 2 \), by analogy. Create a CART tree \( \mathcal{T}^* \) from \( \mathcal{L}^* \) according to the same algorithm by which \( \mathcal{T} \) was created from \( \mathcal{L} \). Denote its rule by \( d^* \). Then by analogy
with (2), the global estimate of risk, $R^*$, for this single bootstrap sample $\mathcal{L}^*$ is

$$(.632)(R^{TS}) + (.368)(R^{LS}),$$

where

$$R^{TS} = \sum_{i=1}^{2} \pi(i) (n^{TS}_i)^{-1} \sum_{\mathcal{L}^{TS} \cap \{j:y_j=i\}} C(d^*(X_j|i)).$$

$R^{LS}$ is defined likewise, with $\mathcal{L}^{LS}$ substituted for $\mathcal{L}^{TS}$ and $n^{LS}_i$ for $n^{TS}_i$. As was mentioned, in practice repeated bootstrap samples are drawn. For each, a global estimate of risk is computed. These numbers are averaged to obtain the estimate that is reported.

We said that our bootstrap .632 differs slightly from the usual. This is because the bootstrap expectation of $R^{LS}$ is simply $R^{RES}$. It is customary and sensible, though not obviously more accurate, to substitute $(.368)R^{RES}$ for repeatedly averaged values of $(.368)R^{LS}$. The strong law of large numbers and related probability inequalities imply that in practice the distinction is irrelevant for all practical purposes. Finally, though our exposition is focused on estimating global risk, bootstrap .632 applies as does repeated cross-validation to estimating the contribution to the risk of any (measurable) subset $S$ of $\mathcal{X}$.

3. BREIMAN'S ESTIMATOR

3.1 A CONJUGATE PRIOR BAYES APPROACH TO JUSTIFYING THE FUNCTIONAL FORM

For simplicity we study a two-class problem, with misclassification costs $C(i|j) = 1 - \delta_{ij}$, and given prior probabilities $\{\pi(i) : i = 1, 2\}$ that are known to the statistician. As usual, $\delta_{ij}$ is 1 if $i = j$; otherwise it is 0. We write $F_i$ for the true probability distribution of the predictors when the correct class is $i$ and $F$ for the unconditional distribution. $N$ is the total learning sample size. We take the learning sample to be iid. The probability mechanism by which it is generated can be thought of thus: Generate a $Y$ from the prior; on $\{Y = i\}$, sample $X$ from $F_i$ independent of other learning sample pairs. Write $p_N(\cdot|\cdot)$ $(p_{N}(\cdot))$ for empirical conditional (respectively unconditional) probabilities when $|\mathcal{L}| = N_i$; $N_i(t)$ is the number of class $i$ observations at $t$ and $N_i$ the total number of class $i$ observations. Here $t$ is used interchangeably as a subset of $X$ and as a node of a CART tree. With this formulation, the resubstitution estimate of the misclassification cost at
node $t$ is the min\{$p_N(1|t), p_N(2|t)$\}. Bayes’ formula gives

$$p_N(t|t) = \frac{p_N(t|1)\pi(1)}{p_N(t|1)\pi(1) + p_N(t|2)\pi(2)}.$$ 

We suppose here that the estimated Bayes rule at terminal node $t \in \tilde{T}$ chooses Class 2. Our heuristics apply as well when Class 1 is its choice. It follows that the resubstitution estimate of misclassification cost at $t$ is $p_N(1|t)$.

Were we in the standard Bayesian beta-binomial situation, then for some positive numbers $a$ and $b$ we would replace $p_N(t|1)$ by

$$\frac{N_i(t) + a}{N_i + a + b}$$

$$= \frac{(N_i(t)/N_i) + (a/N_i)}{1 + ((a + b)/N_i)}$$

$$= \frac{N_i(t)}{N_i}(1 - \frac{a + b}{N_i} + \frac{(a + b)^2}{N_i^2} - \ldots) + \frac{a}{N_i}(1 - \frac{a + b}{N_i} + \frac{(a + b)^2}{N_i^2} - \ldots)$$

$$= \frac{N_i(t)}{N_i} + \frac{a}{N_i} - \frac{(a + b)}{N_i} \cdot \frac{N_i(t)}{N_i} + O\left(\frac{1}{N_i^2}\right).$$  \hspace{1cm} (3)

Strictly speaking, one of recursive partitioning’s several goals, that of rendering terminal nodes homogeneous in their class memberships, precludes the “binomial” part of the “standard beta-binomial situation.” However, our exposition here is somewhat informal, and rests inevitably in part upon large sample considerations. A mathematical analysis that included uniformity of the asymptotics over candidate terminal nodes might allay such concerns only at the price of “losing sight of the forest for the trees.” For present purposes we focus upon the case where $F$ is nonatomic and the algorithm is implemented so that

$$F\{N_i(t)/N_i \to 0\} = 1, \text{ and}$$  \hspace{1cm} (4)

$$F\{N_i(t) \to \infty\} = 1, \; i = 1, 2.$$  \hspace{1cm} (5)

See Breiman et al., Chapter 12 and Lugosi and Nobel. It follows from (4) that (3) can be written

$$\frac{N_i(t)}{N_i} + \frac{a}{N_i} + o\left(\frac{1}{N_i}\right).$$
With beta-binomial substitutions, \( p_N(1|t) \) becomes \( \tilde{p}_N(1|t) \)

\[
\tilde{p}_N(1|t) = \left[ \frac{N_1(t) + a_1}{N_1 + a_1 + b_1} \right] \pi(1) \left( \left[ \frac{N_1(t) + a_1}{N_1 + a_1 + b_1} \right] \pi(1) + \left[ \frac{N_2(t) + a_2}{N_2 + a_2 + b_2} \right] \pi(2) \right)
\]

(6)

It is crucial to our heuristic that

\[
\tilde{p}_N(1|t) > p_N(1|t).
\]

(7)

The mathematical development requires it; and in any case, the research we summarize has as its motivation that CART’s resubstitution estimates of risk are too low. Furthermore, (7) is only mildly restrictive. The \( a_s \) and \( b_s \) are fixed nonnegative numbers. Simple algebra shows, for example, that when \( a_2 = b_2 = 0 \), then (7) is equivalent to \( N_1(t)/N_1 < a_1/(a_1 + b_1) \), which is ultimately automatic in view of (4). We write \( \tilde{p}_N(1|t) \)

\[
\tilde{p}_N(1|t) = \left[ \frac{N_1(t)}{N_1} + \frac{a_1}{N_1} + o\left( \frac{1}{N_1} \right) \right] \pi(1) \left( \left[ \frac{N_1(t)}{N_1} + \frac{a_1}{N_1} + o\left( \frac{1}{N_1} \right) \right] \pi(1) + \left[ \frac{N_2(t)}{N_2} + \frac{a_2}{N_2} + o\left( \frac{1}{N_2} \right) \right] \pi(2) \right).
\]

(8)

The next step is to divide numerator and denominator of (8) by

\[
p_N(t) = p_N(1|t) \pi(1) + p_N(2|t) \pi(2)
\]

\[
= \left( \frac{N_1(t)}{N_1} \right) \pi(1) + \left( \frac{N_2(t)}{N_2} \right) \pi(2).
\]

The quotient in the numerator of (8) is thus

\[
\frac{p_N(1|t) \pi(1)}{p_N(t)} + \frac{a_1 \pi(1)}{N_1 p_N(t)} + o\left( \frac{1}{N_1 p_N(t)} \right)
\]

\[
= \left[ \frac{N_1(t)}{N_1} \right] \pi(1) + \left( \frac{a_1}{N_1} \right) \pi(1) + o\left( \frac{1}{N_1} \right) \pi(1)
\]

(9)

Now,

\[
N_1 p_N(t) = N_1 \left\{ \frac{N_1(t)}{N_1} \pi(1) + \frac{N_2(t)}{N_2} \pi(2) \right\}
\]

\[
= N_1(t) \pi(1) + (N_1/N_2) N_2(t) \pi(2).
\]

(10)

Because our learning sample is assumed to be an iid sample from the unconditional distribution of \((X, Y)\) pairs, it follows that (10) is almost surely of exact order

\[
N_1(t) \pi(1) + \left( \frac{\pi(1)}{\pi(2)} \right) N_2(t) \pi(2)
\]
\[ = \pi(1)(N_1(t) + N_2(t)) = \pi(1)N(t). \]  

(11)

Therefore, (9) is almost surely of exact order

\[
p_N(1|t) + \frac{a_1\pi(1)}{\pi(1)N(t)} + o\left(\frac{1}{N(t)}\right)
= p_N(1|t) + \frac{a_1}{N(t)} + o\left(\frac{1}{N(t)}\right).
\]

Entirely analogous considerations suffice to show that when the denominator of (8) is divided by \( p_N(t) \), the resulting quotient is almost surely of exact order

\[
p_N(1|t) + \frac{a_1}{N(t)} + p_N(2|t) + \frac{a_2}{N(t)} + o\left(\frac{1}{N(t)}\right)
= 1 + \left(\frac{a_1 + a_2}{N(t)}\right) + o\left(\frac{1}{N(t)}\right).
\]

As a consequence, (8) is seen to be almost surely of exact order

\[
\{p_N(1|t) + (a_1/N(t)) + o(1/N(t))\}/\{1 + ((a_1 + a_2)/N(t)) + o(1/N(t))\},
\]

(12)

which is almost surely ultimately less than

\[
p_N(1|t) + \left(\frac{a_1 + a_2}{N(t)}\right)
\]

(13)

Write

\[
\varepsilon = \varepsilon_N = \frac{a_1}{N_1} \pi(1) + \frac{a_2}{N_2} \pi(2).
\]

From (8), the argument immediately subsequent to it, and (9) it follows from (11) that \( P_N(1|t) \) is almost surely of exact order

\[
p_N(1|t) + \frac{\varepsilon}{p_N(t)},
\]

and that for \( N \) large enough

\[
\hat{p}_N(1|t) < p_N(1|t) + (\varepsilon/p_N(t)).
\]

(14)

When (12) holds, (7) allows us to define \( \lambda = \lambda_N \) implicitly so that

\[
\hat{p}_N(1|t) = p_N(1|t) + \frac{\varepsilon}{p_N(t) + \lambda},
\]

(15)
the Breiman estimate. See Section 5.4 of Breiman et al. (1984). Breiman's estimate applies to any of a sequence of subtrees of a large initial tree, and even more generally to rules that need not be tree-based at all. However, our interest is in the tree produced by the CART algorithm. Given the outcome of the sample reuse method by which \( R^{CV} \) or an analogue is computed, the tree is unique. Note that \( \varepsilon \) is (almost surely) of order \( \max(a_1, a_2) / N \). Since (12) and (14) are different expressions for \( \tilde{p}_N(1|t) \), simple algebra reveals that \( \lambda \) is allowed to vary with \( t \) is of order \( p_N(t) \). From Lugosi and Nobel (1996) and Breiman et al., we see that large sample considerations beyond (4) and (5) require that \( \log(N) / N = o(p_N(t)) \), though \( N(t) \) need not be larger than what this requirement entails. We do not repeat Breiman's empirical Bayes argument here, but we do note that it rests upon two key ideas. The first is that of Cover and Hart (1967); namely, for samples large enough, the Bayes risk of a single nearest neighbor rule is at most twice the Bayes risk of the Bayes rule. This has been extended in its domain of applicability by Stone (1977). Samples are assumed large enough that the Cover-Hart-Stone result applies. The second idea is that to an adequate degree of approximation, \( R^{CV} \) is about the Bayes risk of CART. For a detailed exposition see Walker (1992). The result of the heuristics are that \( \varepsilon \) and \( \lambda \) be estimated by the unique simultaneous solution to these equations

\[
\varepsilon = 2\lambda R^{CV}(d) \tag{16}
\]

\[
\lambda \sum_{\mathcal{I}} \frac{p(t)}{p(t) + \lambda} = \max \left\{ \frac{R^{CV}(d) - R^{RES}(d)}{R^{CV}(d)}, 0 \right\} \tag{17}
\]

The left hand side of (17) is readily seen to be monotonic in \( \lambda \); so its unique solution can be found by, for example, Brent's algorithm (1971). Because \( R^{CV} \) is typically (almost surely) bounded away from 0 as \( N \) grows without bound, \( \varepsilon \) is seen to be of the same order as \( \lambda \). In order that (16) and (17) be consistent with the previous discussion, at least one of \( a_1 \) and \( a_2 \) must be an unbounded function of \( N \). Almost surely \( \tilde{p}_N(1|t) - p_N(1|t) \to 0 \). Discussion of the uniformity of the convergence of both to their obvious limiting value is implicit in Lugosi and Nobel, or more simply in Chapter 12 of Breiman et al. depending upon the allowable geometry of terminal regions. Theorem 12.2 of the latter reference can be exploited to show that in problems where densities are suitably smooth and bounded away from 0 and \( \infty \), the ratio of largest to smallest value of \( p_N(t) \) tends almost surely to a finite constant; in this case requiring \( \lambda \) to be constant across terminal nodes \( t \) matters little or not at all.
3.2 BREIMAN’S METHOD WITH ALTERNATE GLOBAL ERROR RATES

Breiman’s method uses the single cross-validation estimate of the global error rate, $R^{CV}(T)$, provided by CART. Crawford (1989) showed that the cross-validation estimate is not necessarily the best estimate of global error rate, and that, in some situations, the bootstrap estimate is superior. Another alternative is to use repeated cross-validation in place of single cross-validation.

We created two new estimation methods, based on the Breiman estimate, but using respectively a bootstrap and a repeated cross-validation estimate of the global error rate in (17) in place of the $R^{CV}(d)$. As well, the repeated cross-validation estimates were studied when the mean of the computed values was for samples trimmed at 5% or 25% at both ends.

4. DATA SETS FOR THE EXPERIMENTS

For the Monte Carlo simulations, we use data from 14 distributions, with one to eight dimensions (features), continuous and categorical variables, noise variables, sample sizes from 50 to 10,000 data points, and Bayes error rates ranging from 0.01 to 0.4. We also vary both the minimum node size for CART and the number of repeated cross-validations and bootstraps. In all the experiments, $C(i|j) = 1 - \delta_{ij}$. All the distributions are defined to contain two classes (Class 1 and Class 2).

Distribution 1

1. Divide the two-dimensional square with vertices $(1, 1), (1, -1), (-1, 1), \text{ and } (-1, -1)$ into four quadrants, using the $x$ and $y$ axes as the dividing lines. The quadrants are unconditionally equally likely.
2. Given that an observation is in quadrant 1, it is uniformly distributed in the quadrants, with probability 0.9 of being Class 1 and 0.1 of being Class 2.
3. Given quadrant 2, points are uniformly distributed with probability 0.4 of being Class 1 and 0.6 of Class 2.
4. Given quadrant 3, points are uniformly distributed with probability 0.6 of being Class 1 and 0.4 of Class 2.
5. Given quadrant 4, points are uniformly distributed with probability 0.1 of being Class 1 and probability 0.9 of Class 2.

Distribution 1 is favorable to the CART partitioning algorithm (hence, the name procart)
in that the most accurate partitioning is perpendicular to the axes. In fact the Bayes rule can be represented as a binary tree-structured rule for which the tree has depth 2.

**Distribution 2**

Distribution 2 (anticart) is defined as follows. Both class 1 and class 2 are drawn from two-dimensional independent normally distributed variables. The mean vector for class 1 objects is \((1,1)\); the mean vector for class 2 objects is \((-1,-1)\). The variance of the two classes is identical, and is adjusted to yield a targeted Bayes' error for the sample (typically 0.25). Distribution 2 is unfavorable to the CART partitioning algorithm (hence, the name anticart), in that the most accurate partitioning is a linear discriminant function; that is, a single line at 45 degrees to the axes.

**Distribution 3**

Distribution 3 (\textit{E. coli} promoters) consists of 576 DNA sequences from the bacterium \textit{E. coli} (Harley and Reynolds 1987), for which we previously constructed a classifier to identify functional regions (promoters) in the DNA. The sequences and results of the experiments are described elsewhere (Walker 1992). The statistical results are similar to those reported here for the synthetic data sets.

**Distribution 4**

Distribution 4 (procart + noise) is the same as Distribution 1 (procart) with the addition of a third variable, which is Gaussian noise with mean 0 and standard deviation 1.

**Distribution 5**

Distribution 5 (anticart + noise) is the same as Distribution 2 (anticart) with the addition of a third variable, which is Gaussian noise with mean 0 and standard deviation 1.

**Distribution 6**

Distribution 6 (blip) is one-dimensional. Class 1 objects are normally distributed with mean 0, and standard deviation specified by the user. For these experiments, the default standard deviation was usually set equal to 0.1. Class 2 objects are uniformly distributed over the interval \((-1,+1)\). Thus, Class 1 objects may be viewed as a normally distributed blip in a uniform background of Class 2 objects.
Distribution 7

Distribution 7 (categorical) is a one-dimensional distribution composed of categorical variables. It is designed to simulate a single base position in a DNA sequence such as an *E. coli* promoter. There are four possible categories (corresponding to the four DNA bases: A, C, G, and T). The two classes differ in the relative occurrence of each base. Class 1 has 80% A, 20% C, 0% G, and 0% T. Class 2 has 40% A, 20% C, 20% G, and 20% T.

Distribution 8

Distribution 8 (rotated procart) is identical to Distribution 1 (procart) except that the entire distribution is rotated around the origin by a user-specified angle. In the experiments reported here the angle of rotation was chosen as 45 degrees. The optimal partition is two lines (at right angles to each other) passing through the origin at 45 degrees from horizontal and vertical.

Distribution 9

Distribution 9 (categorical + blip) is a two-dimensional distribution defined as a combination of Distribution 7 (categorical) and Distribution 6 (blip).

Distribution 10

Distribution 10 (categorical + anticart) is a three-dimensional distribution defined as a combination of Distribution 7 (categorical) and Distribution 2 (anticart).

Distribution 11

Distribution 11 (categorical + procart) is a three-dimensional Distribution defined as a combination of Distribution 7 (categorical) and Distribution 1 (procart).

Distribution 12

Distribution 12 (anticart + blip) is a three-dimensional distribution defined as a combination of Distribution 2 (anticart) and Distribution 6 (blip).

Distribution 13

Distribution 13 (procart + anticart + categorical + noise) is an eight-dimensional distribution defined as a combination of Distribution 1 (procart), Distribution 2 (anticart), two categorical dimensions analogous to Distribution 7 (categorical) (relative proportions of A,C,G, and T in Class 1 being 0.6, 0.2, 0.2, 0.0, and in Class 2 0.4, 0.2, 0.2, and 0.2),
and two noise dimensions uniformly distributed on the intervals \((-1, 1)\) and \((0,3)\).

**Distribution 14**

Distribution 14 (procart + blip) is a three-dimensional distribution defined as a combination of Distribution 1 (procart) and Distribution 6 (blip).

5. **RESULTS**

5.1 **RESULTS FOR SAMPLES OF SIZE 200 FROM DISTRIBUTION 2**

So that readers can get the flavor of our simulations we report results for 10 samples of size 200 from Distribution 2, when it was adjusted so that the Bayes risk of the Bayes rule was .25. Remember that the Bayes risk of the "no data" rule is .50. Here the absolute value of the difference between \(p_N(1|t), p_N(2|t)\) and the true risk at node \(t\) is weighted by \(p_N(t)\) and summed over terminal nodes \(\hat{t}\). The results for the 10 separate samples are in Table 1. They are summarized across techniques by mean, standard deviation, and root mean square error of the weighted absolute error in Table 2. The bottom line here is repeated in what follows: Breiman’s method with repeated cross-validation is a bit better than ordinary Breiman, while each dominates other approaches.
Table 1. Results for 10 Samples of Size 200 from Distribution 2 Adjusted so that the Bayes Risk of the Bayes Rule is .25

<table>
<thead>
<tr>
<th>Sample</th>
<th>Resub</th>
<th>0.632</th>
<th>Rep CV</th>
<th>Rep CV</th>
<th>Rep CV</th>
<th>Breiman</th>
<th>Brei RCV</th>
<th>Brei Boot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(5% trim)</td>
<td>(25% trim)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.085</td>
<td>0.068</td>
<td>0.044</td>
<td>0.039</td>
<td>0.045</td>
<td>0.036</td>
<td>0.067</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.067</td>
<td>0.049</td>
<td>0.042</td>
<td>0.048</td>
<td>0.020</td>
<td>0.020</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.017</td>
<td>0.073</td>
<td>0.049</td>
<td>0.045</td>
<td>0.028</td>
<td>0.021</td>
<td>0.017</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.078</td>
<td>0.025</td>
<td>0.024</td>
<td>0.024</td>
<td>0.028</td>
<td>0.028</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.011</td>
<td>0.020</td>
<td>0.021</td>
<td>0.018</td>
<td>0.029</td>
<td>0.020</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.061</td>
<td>0.075</td>
<td>0.052</td>
<td>0.054</td>
<td>0.034</td>
<td>0.029</td>
<td>0.061</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.076</td>
<td>0.036</td>
<td>0.038</td>
<td>0.038</td>
<td>0.033</td>
<td>0.035</td>
<td>0.032</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.065</td>
<td>0.034</td>
<td>0.084</td>
<td>0.086</td>
<td>0.013</td>
<td>0.022</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.090</td>
<td>0.083</td>
<td>0.066</td>
<td>0.065</td>
<td>0.055</td>
<td>0.053</td>
<td>0.083</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.054</td>
<td>0.047</td>
<td>0.080</td>
<td>0.082</td>
<td>0.046</td>
<td>0.054</td>
<td>0.041</td>
<td></td>
</tr>
</tbody>
</table>

Resub = Resubstitution
Rep CV=Repeated cross validation
Brei RCV=Breiman+repeated cross validation
Brei Boot=Breiman + bootstrap
Table 2. Mean, Standard Deviation, and RMS Error for 10 Samples of Size 200 from Distribution 2 Adjusted so that the Bayes risk of the Bayes Rule is .25

<table>
<thead>
<tr>
<th>Estimation method</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Root mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resubstitution</td>
<td>0.0604</td>
<td>0.0266</td>
<td>0.0655</td>
</tr>
<tr>
<td>0.632 Bootstrap</td>
<td>0.0511</td>
<td>0.0225</td>
<td>0.0554</td>
</tr>
<tr>
<td>Repeated CV</td>
<td>0.0499</td>
<td>0.0213</td>
<td>0.0538</td>
</tr>
<tr>
<td>Repeated CV 5% trim</td>
<td>0.0500</td>
<td>0.0216</td>
<td>0.0540</td>
</tr>
<tr>
<td>Repeated CV 25% trim</td>
<td>0.0500</td>
<td>0.0223</td>
<td>0.0543</td>
</tr>
<tr>
<td>Breiman</td>
<td>0.0330</td>
<td>0.0127</td>
<td>0.0351</td>
</tr>
<tr>
<td>Breiman RCV</td>
<td>0.0318</td>
<td>0.0127</td>
<td>0.0340</td>
</tr>
<tr>
<td>Breiman Boot</td>
<td>0.0403</td>
<td>0.0245</td>
<td>0.0465</td>
</tr>
</tbody>
</table>

5.2 EXPERIMENTS VARYING THE SAMPLE SIZE AND DISTRIBUTION

In this section we present the results of many more experiments. Chapter 5 of (Walker 1992) has the complete results in detail for all data sets, showing the weighted and summed absolute error, its standard deviation, and the root mean square of the error in the estimated probability of misclassification for each distribution and sample size, in the same form as that Table 2. We summarize results here.

In the experiments in this section, we compared the accuracy of the estimators on samples of four sizes (usually 50, 200, 500, and 1000) from each of the 13 synthetic data distributions, for a total of 52 experiments.

The experiments show that, for these sample sizes and distributions, Breiman’s method modified to use the repeated cross-validation estimate of the global error rate is the most accurate overall (Table 3) in terms of weighted absolute error.
Table 3. Comparison of Estimators Across Distributions and Sample Sizes.

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Number of times estimator was most accurate</th>
<th>Number of times estimator was second most accurate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resubstitution</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0.632 Bootstrap</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Repeated CV</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Repeated CV 5%</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Repeated CV 25%</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Breiman</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>Breiman rcv</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>Breiman boot</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

The exception to the dominance of the Breiman method using the repeated cross-validation global error rate occurs in distributions with a Bayes risk less than 0.1, for which the bootstrap .632 estimate or Breiman plus bootstrap are most accurate, although the Breiman plus repeated cross-validation method is closely competitive for such distributions.

If all data from all 52 experiments are combined, then in pairwise comparisons of the estimators, using the Bonferroni method to determine the significance level of an approximately Gaussian Behrens-Fisher-Welch $t$ statistic, the three repeated-cross validation estimates (untrimmed, 10% trimmed, 50% trimmed) do not differ significantly at $p < 0.01$; the three Breiman estimators (CV, RCV, and bootstrap global error) also do not differ significantly at $p < 0.01$. All other pairwise comparisons differ significantly at $p < 0.001$. In particular, Breiman’s method using the repeated cross-validation global error rate differs significantly from the resubstitution method, $p << 0.001$. If the pairwise comparisons are made without any correction for their multiplicity, then the Breiman plus repeated cross-validation estimate differs from its nearest competitor, the Breiman estimate, at approximately $p = 0.01$.

5.3 EXPERIMENTS VARYING THE BAYES RISK

In the experiments in this section, we compared the accuracy of the estimators on samples from Distribution 2 (the anticart distribution) that have Bayes risk varying from 0.01 to 0.33, with sample sizes of 50, 200, and 1000.

In samples with Bayes risk of 0.1 or less, the bootstrap 0.632 estimate or the Breiman variant using the bootstrap global error is the most accurate, although Breiman’s method
with repeated cross-validation is closely competitive. For the samples with a Bayes risk greater than 0.1, the results here are the same as the results reported previously. Breiman’s method with repeated cross-validation estimate of global error dominates.

### 5.4 EXPERIMENTS VARYING THE NUMBER OF BOOTSTRAPS AND CROSS-VALIDATIONS

In these experiments, we compared the accuracy of the estimators that use resampling, setting the number of bootstraps and cross-validations to use in the estimators to one of 10, 30, 100, or 1000. As we would expect, we gain some degree of accuracy by taking larger numbers of samples (performing more bootstraps or cross-validations).

As measured with mean squared error, increasing the number of bootstraps or cross-validations from 10 to 30 typically gives an estimate about 1% percent closer to the true value (the weighted-average estimated probability of misclassification is closer to the true probability by about 0.01). Increasing the number of bootstraps or cross-validations to 100 or 1000 for these data sets and samples does not provide appreciable advantage. Using the weighted absolute error, we see a similar but more equivocal gain from additional bootstraps or cross-validations.

### 6. CONCLUSIONS

The resubstitution method for estimating probability in the partitions (terminal nodes) of classification trees is inaccurate. One alternate is an empirical Bayes approach of Breiman, for which we gave new justification. We used Monte Carlo simulations to compare existing and several new methods for estimating probability in the partitions. We evaluated the methods using 14 data distributions, with one to eight dimensions (features), continuous and categorical variables, noise variables, sample sizes from 50 to 1000 data points, and Bayes risk ranging from 0.01 to 0.4. For these sample sizes and distributions, the Breiman method using the repeated cross-validation global error rate is the most accurate overall, as measured by mean absolute error and mean squared error. For data sets with low Bayes risk (less than 0.1 and less than half the Bayes risk of the “no data” rule), either a local bootstrap 0.632 estimate or Breiman’s method modified to use a bootstrap estimate of the global misclassification rate is most accurate, although the Breiman plus repeated cross-validation method is closely competitive for such distributions. That techniques that rely upon the bootstrap should be better at adjusting small to moderate biases than large ones is consistent with Section 11.7 of Breiman et al. (1984). These methods
are applicable to other adaptive-partitioning classification algorithms.

7. ACKNOWLEDGMENTS

This research has taken place over a number of years, during which the authors were supported in party by Grant Y01-DA-40032-0 from the National Institute on Drug Abuse, Grant LM-07033 from the National Library of Medicine, Grants CA55325 and CA59039 from the National Cancer Institute, Grant GM21215 from the Institute of General Medicine, Grant MIT-9706284-001 from the National Science Foundation, and by grants from IBM. Computing facilities were provided to SUMEX-AIM under Grant LM-05208 from the National Library of Medicine. We thank Bonnie Chung and Lyn Dupre for technical help and Ted Shortliffe for support during the performance of this research.

REFERENCES


Walker, M. G. (1992), *Probability Estimation for Classification Trees and Sequence Analysis,* Stan-CS-92-1422, Ph.D. Dissertation, Departments of Computer Science and Medicine, Stanford University, Stanford, CA.