Analysis of a complexity based pruning scheme for classification trees

Andrew B. Nobel *

February 15, 2002

Abstract

A complexity based pruning procedure for classification trees is described, and bounds on its finite sample performance are established. The procedure selects a subtree of a (possibly random) initial tree in order to minimize a complexity penalized measure of empirical risk. The complexity assigned to a subtree is proportional to the square root of its size. Two cases are considered. In the first the growing and pruning data sets are identical, and in the second they are independent. Using the performance bound, the Bayes risk consistency of pruned trees obtained via the procedure is established when the sequence of initial trees satisfies suitable geometric and structural constraints. The pruning method and its analysis are motivated by work on adaptive model selection using complexity regularization.

Appears in the IEEE Transactions on Information Theory, vol. 48, pp.2362-2368, 2002.

Key words and phrases: Classification trees, pruning, complexity regularization, Bayes risk consistency, tree structured partitions.

*Andrew Nobel is with the Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. Email: nobel@stat.unc.edu. His work was supported in part by NSF grants DMS-9501926 and DMS-9971964.

1 Introduction

Let $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ be a jointly distributed pair of random variables, where the covariate vector X contains the outcomes of a sequence of experiments, and the binary response variable Y is an associated class label of interest. For example, X may contain the results of d diagnostic tests performed on a patient, and Y might indicate whether or not the patient has a particular disease. A classification rule is a deterministic map $\phi : \mathbb{R}^d \to \{0, 1\}$ that assigns a class label to each possible value of X. The performance of ϕ is measured by its probability of error, or risk,

$$R(\phi) = \mathbb{P}\{\phi(X) \neq Y\}\,.$$

(We assume throughout this paper that classes zero and one have equal prior probabilities and identical misclassification costs.) The best achievable risk of any prediction rule is given by the Bayes probability of error

$$R^* = \inf_{\phi} R(\phi)\,,$$

where the infimum is taken over all measurable functions $\phi : \mathbb{R}^d \to \{0, 1\}$. The infimum is achieved by the Bayes rule $\phi^*(x) = I\{E(Y|X = x) > 1/2\}$, which can be deduced from the joint distribution of (X, Y). A comprehensive treatment of probabilistic pattern recognition can be found in [6, 13].

Histogram classification rules are defined by partitioning the space $\mathbb{R}^d$ of the covariates into disjoint regions, and then assigning a class label to each region. Binary classification trees, also known as decision trees, are a widely used family of histogram rules. A binary classification tree is described by a labeled binary tree, each of whose leaves corresponds to a unique cell of a partition of $\mathbb{R}^d$. The tree structure makes computation of the corresponding classification rule fast, and provides a ready interpretation of the rule. A systematic account of classification and regression trees can be found in the book of Breiman, Friedman, Olshen, and Stone [3]. There the authors propose and study the well-known Classification and Regression Tree (CART) algorithm for growing and pruning classification trees. For a description and discussion of the related C4.5 algorithm, see Quinlan [24]. An overview of tree structured classification and pruning can be found in [21, 13].

The general classification problem can be stated as follows: given a data set D_n containing n i.i.d. replicates of the pair (X, Y), produce a classification rule $\hat{\phi}_n$ whose probability of error is close to R^*. In CART and related algorithms, classification trees are produced from D_n in two stages. In the first stage a large initial tree is produced one node at a

time in an iterative, greedy fashion. In the second stage, a small subtree of the initial tree is selected, again using the data set D_n. Whereas the growing procedure proceeds in a top-down fashion, the second stage, known as pruning, proceeds from the bottom-up by successively removing nodes from the initial tree.

The CART pruning procedure selects a subtree of the initial tree that minimizes a weighted sum of performance, measured by the number of misclassifications, and complexity, measured by number of nodes. An appropriate weighting factor is chosen from the data using a resampling scheme (V-fold cross validation) that involves the growing and pruning of auxiliary trees.

In this paper two related pruning schemes for classification trees are described. The first is analyzed under the assumption that the data used to grow and prune the tree are the same. The second is analyzed under the assumption that independent data sets are used to grow and prune the tree. The pruning schemes are, like CART pruning, based on minimizing a complexity penalized empirical risk over all subtrees of an initial tree. However, they differ from CART pruning in two important respects. First, the complexity of a tree is measured not by the number of its nodes, but by the square root of that number. Second, the weighting factor relating performance and classification is given explicitly as a function of the sample size n, and is not obtained via resampling. Upper bounds on the expected risk of each procedure are established. In each case the expected performance of the pruning scheme is comparable to a penalized search among a sequence of idealized pruning schemes, where the k'th such scheme selects a subtree of size k having minimal probability of error.

1.1 Outline

A precise definition of binary classification trees and their associated partitions is given in the next section. In Section 3 the pruning schemes are described, and are compared briefly with the pruning method of the CART algorithm. Upper bounds on the expected performance of the pruning schemes are given in Theorem 1. The Bayes risk consistency of pruned subtrees is investigated in Section 4. Proofs of the principal results are given in Section 5.

2 Preliminaries

2.1 Tree-Structured Partitions

A test tree is a pair (Γ, τ), where Γ is a finite, rooted binary tree such that every non-terminal node has two descendants, and $\tau : \Gamma \to \mathbb{R}^d$ assigns a *test vector* in $\mathbb{R}^d$ to every node $t \in \Gamma$. Every vector $x \in \mathbb{R}^d$ is associated with a descending path in Γ through a sequence of binary comparisons: beginning at the root, and at each subsequent internal node of Γ, x moves to that child of its current node whose test vector is nearest to x in Euclidean distance. In case of ties, x moves to the left child of its current node.

The cell U_t associated with a node $t \in \Gamma$ is the set of vectors x whose path contains t. Thus, the cell of the root node is $\mathbb{R}^d$; the cell of an internal node is split between its children by the hyperplane that forms the perpendicular bisector of their test vectors.

The cell of a node t at distance k from the root is a (possibly unbounded) polytope having at most k faces. Let $t_0, t_1, \ldots, t_k$ be a descending path in Γ from the root node t_0 to another node $t = t_k$. For $j = 1, \ldots, k$ let $u_j = \tau(t_j)$, and let u'_j be the test vector assigned to the sibling of t_j. Then

$$U_t = \bigcap_{j \in A} \{x : \|x - u_j\| \leq \|x - u'_j\|\} \cap \bigcap_{j \in B} \{x : \|x - u_j\| < \|x - u'_j\|\}\,.$$

Here A contains those indices j for which t_j is the left sibling of its parent, and B contains those indices for which it is the right sibling.

Denote the terminal nodes (leaves) of Γ by $\tilde{\Gamma}$. The cells $\{U_t : t \in \tilde{\Gamma}\}$ associated with the terminal nodes of Γ form a partition of $\mathbb{R}^d$, the *tree-structured partition* defined by (Γ, τ).

2.2 Classification Trees

A classification tree T is a triple (Γ, τ, α) where (Γ, τ) is a test tree and $\alpha : \Gamma \to \{0, 1\}$ assigns a class label to each node of Γ. In this case, T acts as a classification rule if one defines

$$T(x) = \sum_{t \in \tilde{\Gamma}} \alpha(t) I\{x \in U_t\}$$

for each $x \in \mathbb{R}^d$. Let $T[x]$ be the unique cell of the partition defined by (Γ, τ) that contains x. Thus $T[x] = U_t$ if $t \in \tilde{T}$ and $x \in U_t$. Let $|T| = |\Gamma|$ denote the number of nodes in Γ.

Definition: A classification tree $T' = (\Gamma', \tau', \alpha')$ is called a subtree of T, written $T' \leq T$, if

(1) Γ' is a subtree of Γ sharing the same root node

(2) τ' is the restriction of τ to Γ'

(3) α' is the restriction of α to Γ'

If conditions (1) and (2) hold then T' is a called a weak subtree of T, written $T' \preceq T$. The difference between ordinary and weak subtrees of T lies in their compatibility with the labeling of T.

Proposition 1 *Let T be a classification tree and let $1 \leq k \leq |T|$. Then*

$$|\{T' : T' \leq T,\ |T'| = k\}| \ \leq\ 2^k \quad \textit{and} \quad |\{T' : T' \preceq T,\ |T'| = k\}| \ \leq\ 2^{2k}\,.$$

Proof: One may establish an injective correspondence between k-node binary trees T and binary k-tuples as follows. First, partition the nodes of T according to their depth from the root. Denote the root node by a 1, and at each subsequent layer of T encode the nodes in that layer by scanning them from left to right, writing 0 for each leaf and 1 for each internal node. It can be verified by induction that this correspondence is one to one (though it is not onto), and the first claim follows. To establish the second claim, note that there are 2^k ways of assigning class labels to the k nodes of T.

3 Complexity Penalized Pruning

3.1 Pruning

In designing a classification tree, the ultimate goal is to produce from the available data a tree T whose probability of error $R(T)$ is as close to R^* as possible. The CART algorithm and related procedures produce a tree T in two stages. In the first stage a large initial tree T_n is produced from a data set D_n^G of size n by means of a greedy growing algorithm. Greedy growing algorithms are iterative procedures that produce classification trees one node at a time. At each iteration the algorithm divides a single terminal region of the current tree by a plane that is perpendicular to one of the coordinate axes. The algorithm selects a region whose division promises the greatest reduction in the number of misclassifications, or some other empirical impurity measure. When the growing procedure terminates, each terminal region of the tree is assigned a class label according to a majority vote. It is assumed in what follows that $|T_n| \leq n$. It often happens that $|T_n|$ is close to n; in these cases T_n overfits the available data and $R(T_n)$ is typically large.

In the second stage of classification tree design the initial tree T_n is "pruned back" to produce a subtree whose expected performance is (hopefully) superior to that of T_n. If the distribution of (X, Y) is known, the best classification tree that can be obtained by relabeling any subtree of T_n is

$$T_n^* = \arg\min_{T \preceq T_n} R(T) \tag{1}$$

In practice, when exact information about the joint distribution of (X, Y) is usually unavailable, pruning is carried out on the basis of a data set

$$D_n^P = (X_1, Y_1), \ldots, (X_n, Y_n)$$

that contains n i.i.d. replicates of (X, Y). In seeking the optimal tree (1), it is natural to replace $R(T)$ by the empirical risk

$$\hat{R}_n(T) = \frac{1}{n} \sum_{i=1}^{n} I\{T(X_i) \neq Y_i\}\,.$$

If the same data are used to grow and to prune the initial tree, i.e. $D_n^G = D_n^P$, then $\hat{R}_n(\cdot)$ will underestimate the risk of large subtrees. In particular, the empirically optimal subtree

$$\hat{T}_n^* = \arg\min_{T \preceq T_n} \hat{R}_n(T) \tag{2}$$

is usually equal to T_n. On the other hand, using separate data sets for growing and pruning is not feasible when the amount of available data is limited, and additional data are expensive or difficult to obtain.

The CART pruning algorithm seeks to balance optimistic estimates of empirical risk in (2) by adding to $\hat{R}_n$ a complexity term that penalizes larger subtrees. For each complexity cost $\alpha \geq 0$ let $\hat{T}_n(\alpha)$ be the smallest tree S such that

$$S \;=\; \arg\min_{T \leq T_n} [\, \hat{R}_n(T) + \alpha |T| \,]\,. \tag{3}$$

Breiman *et al.* [3] established the following useful result. (See [21] for a somewhat simpler proof.)

Theorem A *For every initial tree T_n there exist constants $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = \infty$ and a nested sequence of trees $T_n = S_1 \geq S_2 \geq \cdots \geq S_m = \{root\}$ such that $\hat{T}_n(\alpha) = S_j$ whenever $\alpha_{j-1} \leq \alpha < \alpha_j$.*

Having identified the subtrees $\{\hat{T}_n(\alpha) : \alpha \geq 0\}$ optimal with respect to the penalized criteria in (3), the CART pruning algorithm searches for an optimal value $\hat{\alpha}_n$ of the complexity cost and outputs the associated subtree $\hat{T}_n(\hat{\alpha}_n)$. The constant $\hat{\alpha}_n$ is chosen by means

of a cross validation procedure that requires growing and pruning trees for each cross validation step. Theorem A and the CART pruning algorithm were later generalized by Chou, Lookabaugh, and Gray [5], who gave an algorithm that finds the subtrees $S_1, \ldots, S_m$ in time $O(|T_n| \log |T_n|)$. They consider applications of cost-complexity pruning to a variety of problems, including data compression and image coding. A pruning scheme based on Akaike's information criterion is suggested in [4].

Donoho [7] establishes an interesting connection between CART pruning of regression trees and best orthonormal basis methods for signal representation. In particular, he considers the problem of reconstructing an unknown function f on $[0,1]^2$ from observations at points $(i/n, j/n)$, $0 \le i, j < n$, in the presence of white Gaussian noise with variance σ^2. Of interest in [7] is a dyadic version of CART, in which the goal is to select a dyadic partition of $[0,1]^2$ that minimizes a weighted sum of empirical squared error and partition size; as noted there, an optimal dyadic partition can be found in time $O(n^2)$. It is shown in Theorem 1 of [7] that, for a complexity cost of order $\sigma^2 \log n/n$, the mean squared error performance of dyadic CART is within a logarithmic factor of minimax for every member of a large family of non-isotropic smoothness classes.

The pruning schemes described below are motivated by recent work [23, 22, 1, 15, 2] on non-parametric model selection using complexity regularization. An application of complexity regularization to the pruning of regression trees can be found in [10]. The schemes here differ from the CART pruning algorithm in two fundamental respects. First, the complexity penalty assigned to a subtree T of T_n depends on $|T|^{1/2}$ rather than $|T|$ as is the case in (3): this is addressed in the discussion below. Second, in the schemes here, a fixed complexity cost is selected, based on the size of the pruning data set and the dimension of the covariate vector X. In particular, no resampling or cross-validation is used.

Case 1: Growing and pruning data sets are the same. Let the initial tree T_n be given. Assign to each subtree $T \le T_n$ the complexity penalty

$$\Delta(|T|, n) \;=\; \sqrt{32 \frac{|T|\, d \log n \;+\; |T| \log 2 \;+\; 2 \log |T|}{n}} \tag{4}$$

and let

$$\tilde{R}_n(T) \;=\; \hat{R}_n(T) \;+\; \Delta(|T|, n)\,, \tag{5}$$

be the penalized empirical risk of T. (The natural logarithm is used here and throughout the paper.) Define

$$\hat{T}_n \;=\; \arg\min_{T \le T_n} \tilde{R}_n(T) \tag{6}$$

to be any subtree of T_n that minimizes $\tilde{R}_n(\cdot)$.

Case 2: Growing and pruning data sets are independent. Let the initial tree T_n be given. Assign to each subtree $T \leq T_n$ the complexity penalty

$$\Delta'(|T|, n) \;=\; \sqrt{\frac{|T| \log 2 + \log |T|}{n}}. \tag{7}$$

Let $\tilde{R}'_n(T) = \hat{R}_n(T) + \Delta'(|T|, n)$ be the penalized empirical risk of T and define

$$\hat{T}'_n = \arg\min_{T \leq T_n} \tilde{R}'_n(T) \tag{8}$$

to be any subtree of T_n minimizing $\tilde{R}'_n(\cdot)$.

Consider for the moment the situation in Case 1 above. Ideally, the complexity penalty assigned to a subtree $T \leq T_n$ would equal the difference $R(T) - \hat{R}_n(T)$ between its true probability of error and its empirical probability of error. As $R(T)$ is not available in practice, one commonly seeks instead a distribution-free bound on the related quantity

$$\mathrm{E}\left[\sup_{S \in \mathcal{G}_k} |\hat{R}_n(S) - R(S)|\right], \tag{9}$$

where $\mathcal{G}_k$ is the family of all classification trees S with $|S| = |T| = k$. As shown in Lemma 3, the above expectation is bounded by $\Delta(|T|, n)$ plus a term of smaller order, the dominant dependence on $|T|$ being through its square root. Support for the use of the penalties $\Delta(|T|, n)$ comes from the expected performance bounds of Theorem 1 below. In particular, use of a larger penalty with dominant term $|T|(\log n/n)^{1/2}$ would lead the pruning scheme to favor undersized subtrees, and would yield performance bounds inferior to those given in the theorem.

It should be noted, though, that use of the penalty $\Delta(|T|, n)$ comes at a price. Theorem A, and the accompanying algorithm for finding the optimal subtrees $\{\hat{T}_n(\alpha) : \alpha \geq 0\}$, are at the heart of the CART pruning method. There does not appear to be an analogous result, or a corresponding algorithm, for finding the subtrees minimizing $\hat{R}_n(T) + \alpha|T|^{1/2}$. Thus the exact calculation of the pruned subtrees $\hat{T}_n$ or $\hat{T}'_n$ may be computationally intensive, and in such cases one must rely on heuristics or randomization in order to search for an approximate minimizer of (6) or (8), respectively.

In this regard, we note that the subtrees $S_1, \ldots, S_m$ of Theorem A are still of some use. In particular, the optimality of S_j ensures that S_j has minimal empirical risk among all subtrees $T \leq T_n$ with $|T| = |S_j|$. Moreover, if a subtree S with $S_j \geq S \geq S_{j+1}$ is obtained

by pruning all the descendants of a terminal node $t \in \tilde{S}_j$, then S has minimal empirical risk among all subtrees of T_n with $|T| = |S|$. The same conclusion holds if S is obtained by pruning all the descendants of $t' \in \tilde{S}_j$ with $t' \neq t$. (See Lemma 2 of [5] for the proof.) Let $\mathcal{S}$ be the family consisting of $S_1, \ldots, S_m$ and all the subtrees obtained as above, and let $V = \{|S| : S \in \mathcal{S}\}$. Then the subtree

$$\hat{S}_n \;=\; \arg\min_{S \in \mathcal{S}} [\, \hat{R}_n(S) + \alpha |S|^{1/2} \,] \;=\; \arg\min_{T \preceq T_n, |T| \in V} [\, \hat{R}_n(T) + \alpha |T|^{1/2} \,] \tag{10}$$

may be used as an approximation to $\hat{T}_n$, or as a starting point for a more extensive search.

Gey and Nedelec [10] have analyzed complexity penalized pruning schemes for regression trees under the squared error. In this case, analysis of expectations like those in (9) leads to a complexity whose dominant term is $|T| \log n/n$, in accordance with the linear penalty used in [3]. A similar complexity penalty is used by Donoho [7] in his analysis of dyadic CART regression trees.

In Case 1 and Case 2 above the complexity penalized risk may be minimized by two or more subtrees. The analysis below applies to any of these trees.

Definition: A tree $T = (\Gamma, \tau, \alpha)$ is *compatible* with the pruning data set $D_n^P = \{(X_1, Y_1)\}_{i=1}^n$ if for every $t \in \Gamma$

$$\alpha(t) = \text{ majority-vote } \{Y_j : X_j \in U_t\}$$

If the growing and pruning data are the same, then typically the trees produced by greedy algorithms such as CART will be compatible with D_n^P.

Theorem 1 *Let T_n be a random classification tree produced from D_n^G that has been relabeled if necessary so that it is compatible with the pruning data set D_n^P.*

(A) If $D_n^G = D_n^P$ and $\hat{T}_n$ is given by (6), then

$$\mathrm{E}\, R(\hat{T}_n) - R^* \;\leq\; \min_{1 \leq k \leq |T_n|} \left\{ 2\Delta(k,n) + \frac{6\sqrt{2}}{\sqrt{dkn \log n}} \;+\; \mathrm{E}\left[\min_{T \preceq T_n, |T| \leq k} (R(T) - R^*) \right] \right\}. \tag{11}$$

(B) If D_n^G and D_n^P are independent, and $\hat{T}_n$ is given by (8), then

$$\mathrm{E}\, R(\hat{T}'_n) - R^* \leq \min_{1 \leq k \leq |T_n|} \left\{ 2\Delta'(k,n) + \frac{3}{2\sqrt{kn \log 2}} \;+\; \mathrm{E}\left[\min_{T \preceq T_n, |T| \leq k} (R(T) - R^*) \right] \right\}.$$

While a variety of rigorous results on the convergence and structural properties of pruning methods have appeared in the literature (see for example [3, 5, 8]), Theorem 1 appears

to be the first result giving bounds on the expected performance of a complexity based pruning scheme. Gey and Nedelec [10] have recently obtained results analogous to Theorem 1 for complexity penalized pruning of regression trees, under both bounded and Gaussian regression models.

In CART and related algorithms the terminal regions of the initial tree are typically rectangles with sides parallel to the coordinate axes. The conclusions of Theorem 1 are also valid when the terminal regions of the initial tree are polytopes. This is the case, for instance, when the regions associated with internal nodes of T_n are split by halfspaces that are not perpendicular to one of the coordinate axes. This type of linear splitting is used in some variants of CART, and in multivariate clustering schemes used for data compression (see [9]). In Theorem 1 no assumptions are placed on the joint distribution of the labeled samples (X, Y).

Part A of Theorem 1 has the following interpretation. Let $\mathcal{P}_k$, $k \geq 1$, be a pruning scheme that, knowing the distribution of (X, Y), always selects the best k-node weak subtree of T_n, *i.e.* $\mathcal{P}_k(T_n) = \arg\min\{P(T) : T \preceq T_n, |T| \leq k\}$. Then the expectation appearing on the right hand side of (11) is the expected performance of $\mathcal{P}_k$ relative to the Bayes probability of error, and is equivalently the expected approximation error of the family of k-node weak subtrees of T_n. No matter what the initial tree T_n may be, the family of k-node weak subtrees of T_n is contained in the family $\mathcal{G}_k$ of all k-node classification trees. The first two terms in curly braces in (11) are an upper bound on the value of

$$\mathrm{E}\left[\sup_{T \in \mathcal{G}_k} |\hat{R}_n(T) - R(T)|\right],$$

which governs the estimation error of the family $\mathcal{G}_k$. Together these terms reflect the fact that, as k increases, it becomes increasingly difficult to select a tree $T \in \mathcal{G}_k$ with small probability of error $R(T)$ on the basis of the finite data set D_n^P. For fixed k, the quantity in curly braces is typically a good upper bound on the expected performance of data-driven schemes, such as empirical risk minimization [22], that search among the k-node weak subtrees of T_n for a tree with small probability of error. Theorem 1 says that the expected performance of $\hat{T}_n$ is no greater than the best expected performance among n such schemes, one for each value of k. Part B of the theorem may be interpreted similarly.

As one might expect, the absolute performance of the pruning procedure depends critically on the initial tree, and on the growing procedure that produced it. Conditions for the consistency of the procedure are discussed below.

4 Bayes Risk Consistency of Pruned Classification Trees

A sequence of random classification trees $\{T_n\}$ is said to be weakly Bayes risk consistent if $\mathrm{E}\, R(T_n) \to R^*$. Gordon and Olshen [11] and Breiman *et al.* [3] established the Bayes risk consistency of supervised greedy growing algorithms for classification trees based on axis-parallel splits. Extensions of these results to oblique hyperplane and more general splits were given by Lugosi and Nobel [16], see also [6]. The Bayes risk consistency of unsupervised greedy growing algorithms based on hyper-rectangular splits was established by Devroye *et al.* [6]. The structural consistency and shrinking cell properties of greedy growing algorithms for tree-structured clustering schemes were studied by Nobel and Olshen [17], and Nobel [18, 20]. Sufficient conditions for the consistency of tree-structured density and regression estimates produced via recursive partitioning can be found in [12, 3, 25, 16, 19, 6].

The work cited above does not address the consistency of pruned subtrees $\hat{T}_n$ of the initial trees T_n. In general, searching among the subtrees of T_n for one achieving a good performance-complexity tradeoff makes pruning attractive, even when the initial trees are themselves consistent, or when they contain consistent subtrees whose identities may be known. In what follows it is assumed that the growing and pruning data sets are the same. Analogous results may be established in the independent case. Sufficient conditions for the weak (in probability) consistency of pruned classification trees follow as an immediate corollary to Theorem 1.

Corollary 1 *If there exist a sequence of random trees T'_n such that $T'_n \preceq T_n$ with probability one, $\mathrm{E}\, R(T'_n) \to R^*$ and $\mathrm{E}\, |T'_n| = o(n/\log n)$, then the pruned subtrees $\hat{T}_n$ are weakly Bayes risk consistent.*

Proof: The assumptions ensure that there exist integers k_n such that $\mathbb{P}\{|T'_n| \leq k_n\} \to 1$ and $k_n = o(n/\log n)$. By part (A) of Theorem 1,

$$\begin{aligned}
\mathrm{E}\, R(\hat{T}_n) - R^* &\leq 2\Delta(k_n, n) + \frac{6\sqrt{2}}{\sqrt{d k_n n \log n}} \;+\; \mathrm{E}\left[\min_{T \preceq T_n, |T| \leq k_n} (R(T) - R^*)\right] \\
&\leq 2\Delta(k_n, n) + \frac{6\sqrt{2}}{\sqrt{d k_n n \log n}} \;+\; (\mathrm{E}\, R(T'_n) - R^*) \;+\; \mathbb{P}\{|T'_n| > k_n\}.
\end{aligned}$$

The definition of k_n ensures that the each term in the final inequality tends to zero as n tends to infinity.

Remark: Note that the identity and labeling of the trees T'_n need not be known. Since the pruning scheme effectively searches among the labeled subtrees of T_n, the existence of such trees is all that is required.

4.1 Existence of Consistent Subtrees

When the initial trees T_n are not known to contain consistent subtrees of suitable size, the existence of such subtrees can be established by purely analytical means. To do this, some constraints must be placed on the structure of the initial trees, and on the coarseness of their partitions.

Definition: For a given classification tree T let T_t denote the binary tree consisting of the node t and all of its descendants. Call T α-balanced, $\alpha \in [0, 1/2]$, if for every internal node t with children t_1 and t_2,

$$|T_{t_1}|,\ |T_{t_2}| \ \geq\ \alpha |T_t|.$$

A binary tree satisfies the condition above with $\alpha = 1/2$ if and only if it is balanced.

Recall that the diameter of a set $A \subseteq \mathbb{R}^d$ is given by $\mathrm{diam}(A) = \sup_{u,v \in A} \|u - v\|$. The following Lemma can be proved by standard approximation arguments, see for example [16].

Lemma 1 *Let $T_1, T_2, \ldots$ be any sequence of random classification trees and suppose the covariate vector X has distribution μ. If the cells of T_n shrink, in the sense that*

$$\mu\{x : \mathrm{diam}(T_n[x]) > \epsilon\} \to 0 \quad \textit{wp1}$$

for every $\epsilon > 0$, then there exist trees T_n', formed by relabeling the nodes of T_n, such that $R(T_n') \to R^*$ with probability one.

The following proposition is established in Section 5.2 below.

Proposition 2 *Let $T_1, T_2, \ldots$ be a sequence of random classification trees. Suppose that with probability one*

a. there exists $\alpha > 0$ such that each tree T_n is α-balanced,

b. $|T_n| = O(n)$ and $|T_n| / \log n \to \infty$, and

c. $\max\{\mathrm{diam}(T_n[x]) : x \in V\} \cdot \log n \to 0$ for every bounded $V \subseteq \mathbb{R}^d$

Then there exist trees $T_n' \preceq T_n$ such that $|T_n'| = o(|T_n| / \log n)$ and $R(T_n') \to R^*$ with probability one. In particular, the complexity pruned subtrees $\hat{T}_n$ obtained from T_n are weakly Bayes risk consistent.

5 Derivations

5.1 Proof of Theorem 1

The proof of Theorem 1 depends on the following extension of the Vapnik-Chervonenkis inequality to families of partitions. The proof can be found in Lugosi and Nobel [16], with further discussion in [19]. In what follows log denotes the natural logarithm.

Lemma 2 *Let $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \{0, 1\}$ be independent and identically distributed. For each $k \geq 1$, each $s > 0$, and each $n \geq 4$,*

$$\mathbb{P}\left\{ \sup_{T \in \mathcal{G}_k} |\hat{R}_n(T) - R(T)| > s \right\} \leq 2^k n^{kd} \exp\left[\frac{-ns^2}{32} \right],$$

where $\mathcal{G}_k$ is the collection of all k-node classification trees based on tree-structured partitions of $\mathbb{R}^d$.

Lemma 3 *For each $k \geq 1$ and $n \geq 4$,*

$$\mathrm{E}\left[\sup_{T \in \mathcal{G}_k} |\hat{R}_n(T) - R(T)| - \Delta(k, n) \right] \;\leq\; \frac{2\sqrt{2}}{k^2 \sqrt{dkn \log n}}. \tag{12}$$

If D_n^G and D_n^P are independent then for each $k \geq 1$ and $n \geq 4$,

$$\mathrm{E}\left[\max_{T \preceq T_n, |T| = k} |\hat{R}_n(T) - R(T)| - \Delta'(k, n) \right] \;\leq\; \frac{1}{2k^2 \sqrt{kn \log 2}} \tag{13}$$

Proof: Inequality (12) follows from a direct calculation using Lemma 2 and the elementary inequalities

$$\mathrm{E}(Z - u) \;\leq\; \mathrm{E}(Z - u)_+ \;\leq\; \int_u^\infty \mathbb{P}\{Z \geq t\} dt$$

and

$$\int_u^\infty e^{-\alpha t^2} dt \;\leq\; \frac{e^{-\alpha u^2}}{2\alpha u}$$

applied to the random variable $Z = \sup\{|\hat{R}_n(T) - R(T)| : T \in \mathcal{G}_k\}$ and constant $u = \Delta(k, n)$. If D_n^G is independent of D_n^P then the template (Γ_n, τ_n) of T_n is independent of D_n^P, and is fixed when conditioning on D_n^G. It then follows from the union bound, Lemma 1 and Hoeffding's [14] inequality that

$$\mathbb{P}\left\{ \max_{T \preceq T_n, |T| = k} |\hat{R}_n(T) - R(T)| \geq s \;\middle|\; D_n^G \right\} \;\leq\; 2^{2k+1} e^{-2ns^2}$$

Inequality (13) may then be established by a calculation like that yielding inequality (12).

Auxiliary trees: For each random initial tree T_n and each $1 \le k \le |T_n|$ let

$$\hat{T}_{n,k} = \operatorname*{arg\,min}_{T \preceq T_n, |T|=k} \hat{R}_n(T)$$

be the empirically optimal k-node subtree of T_n, and let

$$T_{n,k} = \operatorname*{arg\,min}_{T \preceq T_n, |T|=k} R(T) \tag{14}$$

be the optimum rule that can be obtained by relabeling a k-node subtree of T_n.

Lemma 4 *If $D_n^G = D_n^P$ and $\hat{T}_n$ is defined as in (6), then for every $n \ge 4$*

$$\mathrm{E}\left\{R(\hat{T}_n) - R(T_{n,k})\right\} \;\le\; 2\Delta(k,n) + \frac{6\sqrt{2}}{\sqrt{dkn\log n}}. \tag{15}$$

If D_n^G and D_n^P are independent and $\hat{T}_n$ is given by (8), then

$$\mathrm{E}\left\{R(\hat{T}_n) - R(T_{n,k})\right\} \;\le\; 2\Delta'(k,n) + \frac{3}{2\sqrt{kn\log 2}}. \tag{16}$$

Proof: The proof is based on the argument of Lugosi and Zeger [15]. First consider the decomposition

$$R(\hat{T}_n) - R(T_{n,k}) \;=\; (R(\hat{T}_n) - \tilde{R}_n(\hat{T}_n)) \;+\; (\tilde{R}_n(\hat{T}_n) - \tilde{R}_n(T_{n,k})) \;+\; (\tilde{R}_n(T_{n,k}) - R(T_{n,k}))\,.$$

As T_n is compatible with D_n^P, it may easily be verified that $\hat{R}_n(\hat{T}_{n,k}) \le \hat{R}_n(T_{n,k})$, and since $|T_{n,k}| = |\hat{T}_{n,k}| = k$,

$$\tilde{R}_n(\hat{T}_n) \;\le\; \tilde{R}_n(\hat{T}_{n,k}) \;\le\; \tilde{R}_n(T_{n,k})\,.$$

Thus the second term in the decomposition is less than or equal to zero. By definition of $\tilde{R}_n(\cdot)$ the third term above is equal to

$$\begin{aligned}
&\mathrm{E}(\hat{R}_n(T_{n,k}) - R(T_{n,k})) \;+\; \Delta(k,n)) \\
&\quad\le\;\; \mathrm{E}\left[\sup_{T\in\mathcal{G}_k} |\hat{R}_n(T) - R(T)| - \Delta(k,n)\right] \;+\; 2\Delta(k,n) \\
&\quad\le\;\; \frac{2\sqrt{2}}{k^{5/2}\sqrt{dn\log n}} \;+\; 2\Delta(k,n),
\end{aligned} \tag{17}$$

where the second inequality follows from (12). As for the first term, note that

$$\begin{aligned}
\mathrm{E}(R(\hat{T}_n) - \tilde{R}_n(\hat{T}_n)) \;&=\; \sum_{k=1}^{|T_n|} \mathrm{E}\left[(R(\hat{T}_n) - \tilde{R}_n(\hat{T}_n))\, I\{|\hat{T}_n| = k\}\right] \\
&=\; \sum_{k=1}^{|T_n|} \mathrm{E}\left[(R(\hat{T}_n) - \hat{R}_n(\hat{T}_n) - \Delta(k,n))\, I\{|\hat{T}_n| = k\}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{|T_n|} \mathrm{E}\left[\sup_{T \in \mathcal{G}_k} |\hat{R}_n(T) - R(T)| - \Delta(k,n) \right] \\
&\leq \sum_{k=1}^{|T_n|} \frac{2\sqrt{2}}{k^{5/2}\sqrt{dn \log n}} \\
&\leq \frac{4\sqrt{2}}{\sqrt{dkn \log n}}, \qquad (18)
\end{aligned}
$$

where the third inequality is a consequence of the fact that $\sum_{k \geq 1} k^{-2} \leq 2$. Combining inequalities (17) and (18) gives the bound (15). The bound (16) may be established in a similar fashion using (13).

Proof of Theorem 1: Performance bounds for $\hat{T}_n$ follow immediately from the inequalities of Lemma 4 and the elementary relation

$$
\mathrm{E}\, R(\hat{T}_n) - R^* \leq \min_{1 \leq k \leq n} \left\{ \mathrm{E}(\, R(\hat{T}_n) - R(T_{n,k})\,) + (\, \mathrm{E}\, R(T_{n,k}) - R^*\,) \right\}.
$$

5.2 Proof of Proposition 2

It follows from the assumptions that there is a sequence of bounded rectangles $V_1 \subseteq V_2 \subseteq \cdots \mathbb{R}^d$ and constants $a_n = \max\{\mathrm{diam}(T_n[x]) : x \in V_n\}$ such that

$$
\cup_{n=1}^{\infty} V_n = \mathbb{R}^d \quad \text{and} \quad a_n \log n \to 0.
$$

Define $b_n = \min\{\sqrt{\log n / a_n},\, |T_n|\}$. Then as n tends to infinity,

$$
a_n b_n \to 0 \quad \text{and} \quad \frac{\log n}{b_n} \to 0\,.
$$

Consider the tree T_n. Letting t' denote the parent of t define

$$
L_n = \{t \in T_n \,:\, |T_{n,t}| \leq b_n \ \text{ and } \ |T_{n,t'}| > b_n\}
$$

If $t \in L_n$ then the subtree rooted at t has at most b_n nodes, while the subtree rooted at its parent has at least $b_n + 1$. By definition, each path from the root of T_n to a terminal node contains exactly one member of L_n, and no member of L_n can be the descendent of another. It follows that there is a unique subtree $T_n' \leq T_n$ having terminal nodes L_n. In particular $|T_n'| \leq 2|L_n|$.

We wish to bound the size of T_n'. Note that if $t \in L_n$ then $|T_{n,t}| \geq \alpha |T_{n,t'}| \geq \alpha b_n$ as T_n is α-balanced. As subtrees rooted at distinct elements of L_n are disjoint, it follows that

$$
|T_n| \ \geq \ \sum_{t \in L_n} |T_{n,t}| \ \geq \ \alpha b_n |L_n|,
$$

and consequently

$$|T_n'| \;\le\; 2|L_n| \;\le\; \frac{2|T_n|}{\alpha b_n} \;=\; o\left(\frac{|T_n|}{\log n}\right)$$

This establishes assertion (i).

To establish (ii), first fix constants $\epsilon, \delta > 0$. Let $V_a \subseteq \mathbb{R}^d$ be a bounded rectangle such that $P(V_a^c) \le \delta$, and let $V_b \supseteq V_a$ be a closed, bounded rectangle such that

$$\inf\{\|v - v'\| : v \in V_a,\, v' \in V_b^c\} \;\ge 2\,\epsilon. \tag{19}$$

When n is sufficiently large,

$$\max\{\mathrm{diam}(U_t) : t \in \tilde{T}_n \text{ and } U_t \cap V_b \neq \emptyset\} \;\le\; a_n \;\le\; \frac{\epsilon}{b_n}. \tag{20}$$

Fix any such n and consider a terminal node s of the subtree T_n' for which $U_s \cap V_a \neq \emptyset$. If U_s intersects V_b^c then there exist a point $v_a \in V_a$ and a point v_b on the boundary of V_b such that the line segment $L = \{\eta v_a + (1-\eta)v_b : \eta \in [0,1]\}$ is contained in both V_b and U_s. Let $H_1(\cdot)$ denote one-dimensional Hausdorff measure in $\mathbb{R}^d$. By virtue of (19),

$$2\epsilon \;\le\; \|v_a - v_b\| \;=\; H_1(L)\,. \tag{21}$$

The definition of L_n ensures that U_s is the union of $k \le b_n$ disjoint sets $U_1, \ldots, U_k$, each of which is a terminal region of T_n. In conjunction with (20) this implies that

$$H_1(L) \;=\; \sum_{j=1}^{k} H_1(L \cap U_j) \;\le\; \sum_{j=1}^{k} \mathrm{diam}(U_j) \;\le\; k\frac{\epsilon}{b_n} \;\le\; \epsilon\,.$$

However this contradicts (21), so that $U_s = \cup_{j=1}^{k} U_j$ must be contained in V_b. The inequality above then shows that $\mathrm{diam}(U_s) \le \epsilon$, and therefore $\max\{\mathrm{diam}(T_n'[x]) : x \in V_a\} \le \epsilon$. It follows that

$$\limsup_{n\to\infty} P\{x : \mathrm{diam}(T_n'[x]) > \epsilon\} \;\le\; P(V_a^c) \;\le\; \delta$$

for every choice of $\epsilon, \delta > 0$. Relabeling the trees T_n' if necessary, Lemma 1 ensures that $R(T_n') \to 0$. The consistency of the complexity pruned subtrees $\hat{T}_n$ follows immediately from Corollary 1.

References

[1] A. R. Barron, "Complexity regularization with application to artificial neural networks", in *Nonparametric Functional Estimation and Related Topics*, G. Roussas, ed., pp.561–576. NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991.

[2] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization", *Probability Theory and Related Fields*, vol.113, pp.301–413, 1999.

[3] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*. Wadsworth International, Belmont, CA., 1984.

[4] A. Ciampi, C.-H. Chang, S. Hogg and S. McKinney, "Recursive partition: a versatile method for exploratory data analysis in biostatistics", in *Biostatistics*, I.B. MacNeil and G.J. Umphrey eds., pp.23-50. Reidel, Dordrecht, 1987.

[5] P. Chou, T. Lookabaugh, and R.M. Gray, "Optimal pruning with applications to tree-structured source coding and modeling", *IEEE Trans. Inform. Theory*, vol.37, pp.31-42, 1989.

[6] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, New York, 1996.

[7] D.L. Donoho, "CART and best-ortho-basis: A connection", *Ann. Stat.*, vol.25, pp.1870-1911, 1997.

[8] S.B. Gelfand, C.S. Ravishankar, and E.J. Delp, "An iterative growing and pruning algorithm for classification tree design", *IEEE Trans. PAMI*, vol.13, pp.163-174, 1991.

[9] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*. Kluwer, Dordrecht, 1991.

[10] S. Gey and E. Nedelec, "Model selection for CART regression trees". Prépublication 2001-56, Laboratoire de Mathématique, Université Paris-Sud, Orsay, 2001.

[11] L. Gordon and R. Olshen, "Asymptotically efficient solutions to the classification problem", *Annals of Statistics*, vol.6, pp.515-533, 1978.

[12] L. Gordon and R. Olshen, "Almost sure consistent nonparametric regression from recursive partitioning schemes", *Journal of Multivariate Analysis*, vol.15, pp.147-163, 1984.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Verlag, New York, 2001.

[14] W. Hoeffding, "Probability inequalities for sums of bounded random variables", *Journal of the American Statistical Association*, vol.58, pp.13–30, 1963.

[15] G. Lugosi and K. Zeger, "Concept learning using complexity regularization", *IEEE Transactions on Information Theory*, vol.42, pp.48-54, 1996.

[16] G. Lugosi and A.B. Nobel, "Consistency of data-driven histogram methods for density estimation and classification", *Annals of Statistics*, vol.24, pp.687-706, 1996.

[17] A.B. Nobel and R.A. Olshen, "Termination and continuity of greedy growing for tree-structured vector quantizers", *IEEE Transactions on Information Theory*, vol.42, pp.191-205, 1996.

[18] A.B. Nobel, "Vanishing distortion and shrinking cells", *IEEE Transactions on Information Theory*, vol.42, pp.1303-1305, 1996.

[19] A.B. Nobel, "Histogram regression estimation using data-dependent partitions", *Annals of Statistics*, vol.24, pp.1084-1105, 1996.

[20] A.B. Nobel, "Recursive partitioning to reduce distortion", *IEEE Transactions on Information Theory*, vol.43, pp.1122-1133, 1997.

[21] B.D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, Cambridge, 1996.

[22] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

[23] V. N. Vapnik and A. Ya. Chervonenkis, *Theory of Pattern Recognition*. Nauka, Moscow, 1974 (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.

[24] J.R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, 1993.

[25] L. C. Zhao, P. R. Krishnaiah, and X. R. Chen, "Almost sure L_r-norm convergence for data-based histogram density estimates", *Theory of Probability and its Applications*, vol.35, pp.396–403, 1990.