Complexity Regularized Dyadic Classification Trees: Efficient Pruning and Rates of Convergence

Clayton Scott and Robert Nowak

Technical Report TREE0201

Department of Electrical and Computer Engineering

Rice University

Email: {cscott, nowak}@rice.edu

April 27, 2002

Abstract

Classification trees are one of the most popular types of classifiers, with ease of implementation and interpretation being among their attractive features. Despite the widespread use of classification trees, theoretical analysis of their performance is scarce. In this paper, we show that a new family of classification trees, called dyadic classification trees (DCTs), are near optimal (in a minimax sense) for a very broad range of classification problems. This demonstrates that other schemes (e.g., neural networks, support vector machines) cannot perform significantly better than DCTs in many cases. We also show that this near optimal performance is attained with linear (in the number of training data) complexity growing and pruning algorithms. Moreover, the performance of DCTs on benchmark datasets compares favorably to that of standard CART, which is generally more computationally intensive and which does not possess similar near optimality properties. Our analysis stems from theoretical results on structural risk minimization, on which the pruning rule for DCTs is based.

1 Introduction

Let $(X, Y) \in \mathbf{R}^d \times \{0, 1\}$ be a jointly distributed pair of random variables. In pattern recognition, X is called an input vector, and contains the measurements from an experiment. The values in X

are referred to as features, attributes, or predictors. Y is called a response variable, and is thought of as a class label associated with X. A classifier is a function $\phi : \mathbf{R}^d \to \{0,1\}$ that attempts to match an input vector with the appropriate class. The performance of ϕ for a given distribution of the data is measured by the probability of error:

$$L(\phi) \stackrel{\text{def}}{=} \mathbf{P}\{\phi(X) \neq Y\}.$$

The classifier with the smallest probability of error, denoted ϕ^*, is called the Bayes classifier. The probability of error for the Bayes classifier is denoted L^*. If we set $\eta(x) = \mathbf{P}\{Y = 1|X = x\} = \mathbf{E}\{Y|x\}$, then the Bayes classifier is given by

$$\phi^*(x) = \begin{cases} 1 & \text{if } \eta(x) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} .$$

The true distribution on the data is generally unknown. In such cases, we may construct a classifier ϕ_n based on a *training dataset* $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of independent, identically distributed samples. The process of constructing ϕ_n is known as supervised learning. We sometimes write $\phi_n(x; D_n)$ to denote the dependency of ϕ_n on D_n. The performance of ϕ_n is measured by the conditional probability of error,

$$L_n = L(\phi_n) = \mathbf{P}\{\phi_n(X; D_n) \neq Y|D_n\}.$$

Note that L_n is random, since D_n is random.

A procedure that selects a classifier for all n is called a discrimination rule. A good discrimination rule satisfies the property that L_n approaches L^* as n tends to infinity. This is formalized by the notion of *consistency*: A discrimination rule is said to be weakly (strongly) consistent if $L_n \to L^*$ in probabilty (with probability one) as $n \to \infty$. If a discrimation rule is consistent for all distributions, it is said to be *universally* (weakly or strongly) consistent. Examples of consistent discrimination rules, as well as a comprehensive introduction to probabilistic pattern recognition, may be found in [1].

In this paper, we examine a family of classifiers called *dyadic classification trees* (DCTs), built by recursive, dyadic partitioning of the input space. The appropriate tree from this family is obtained by building an initial tree (in a data-independent fashion), followed by a data-dependent pruning operation based on structural risk minimization (SRM). The pruning strategy resembles that used by CART, except that the penalty assigned to a subtree is proportional to the square root of its size.

SRM penalized DCTs lead to a strongly consistent discrimination rule for input data X with support in the unit cube $[0,1]^d$. We also derive bounds on the rate of convergence of DCTs to the Bayes error. Under a modest regularity assumption (in terms of the box-counting dimension) on the underlying optimal Bayes decision boundary, we show that complexity-regularized DCTs converge to the Bayes decision at a rate of $n^{-1/(d+1)}$. Moreover, the minimax error rate for this class is at least $n^{-1/d}$. This shows that dyadic classification trees are near minimax-rate optimal, i.e., that no discrimination rule can perform significantly better in this minimax sense.

We also present two efficient algorithms for implementing the square-root pruning strategy, the faster of which leads to an $O(n)$ algorithm for DCT construction. The first pruning algorithm, which is slightly more general than the second, requires $O(M^2)$ operations to prune an initial tree with M terminal nodes. The second pruning algorithm is based on the familiar pruning algorithm used by CART [2], and requires $O(M \log M)$ operations. Using these algorithms, we compare DCTs with a CART-like tree classifier on four common datasets.

The paper is organized as follows: In Section 2, we review structural risk minimization (SRM), a method for selecting classifiers using complexity regularization. In Section 3, we review classification trees. In Section 4, we introduce dyadic classification trees. In Section 5, we apply SRM to DCTs, giving rise to our pruning scheme, and state some properties of the resulting discrimination rule. In Section 6, we present rate of convergence results. In Section 7, we compare the results of Section 5 to recent performance bounds for the case when the initial tree (to be pruned) is grown in a data-dependent fashion [3]. In Section 8, we describe an efficient algorithm to implement the proposed pruning scheme. In Section 9, we compare DCTs to a CART-like classifier on four benchmark datasets. In Section 10, we relate the current work to previous work. In Section 11, we discuss future lines of work, and in Section 12 we provide conclusions.

2 Structural Risk Minimization

Let $\mathcal{C}$ be a collection of classifiers $\phi : \mathbf{R}^d \to \{0,1\}$, and let $z_1, \ldots, z_n \in \mathbf{R}^d$. If each of the 2^n possible labellings of $z_1, \ldots, z_n$ can be correctly classified by some $\phi \in \mathcal{C}$, we say $\mathcal{C}$ *shatters* $z_1, \ldots, z_n$. The *Vapnik-Chervonenkis dimension* (or VC dimension) of $\mathcal{C}$, denoted by V, is the largest integer n for which there exist $z_1, \ldots, z_n \in \mathbf{R}^d$ such that $\mathcal{C}$ *shatters* $z_1, \ldots, z_n$. If $\mathcal{C}$ shatters some n points for every n, then $V = \infty$ by definition. The VC dimension is a measure of the capacity of $\mathcal{C}$. As V increases, $\mathcal{C}$ is able to separate more complex patterns.

Given a training dataset D_n, and a classifier ϕ, the *empirical risk* (or empirical error) of ϕ is

$$\widehat{L}_n(\phi) = \frac{1}{n} \sum_{i=1}^{n} I(\phi(X_i) \neq Y_i),$$

where I is the indicator function. Given $\mathcal{C}$, the method of *empirical risk minimization* selects the classifier $\tilde{\phi}_n$ minimizing the empirical risk error over all $\phi \in \mathcal{C}$. Vapnik and Chervonenkis [4, 5] showed that the error probability of $\tilde{\phi}_n$ approaches the error probability of the best classifier in $\mathcal{C}$ as n grows. What follows is one of several variations on their original result [1].

Theorem 1. *For any distribution on (X, Y), and for all n and $\epsilon > 0$,*

$$\mathbf{P}\left\{ \sup_{\phi \in \mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)| > \epsilon \right\} \leq 8n^V e^{-n\epsilon^2/32},$$

and in particular, for empirical risk minimization,

$$\mathbf{P}\left\{ L(\tilde{\phi}_n) - \inf_{\phi \in \mathcal{C}} L(\phi) > \epsilon \right\} \leq 8n^V e^{-n\epsilon^2/128}.$$

The theorem is of practical importance because (a) the bound holds for all n, not just for n sufficiently large, (b) the bound does not depend on the distribution of the data, and (c) V can be computed, or at least bounded, for several common choices of $\mathcal{C}$. Note that the inequality is useful only if $V < \infty$.

If V is too small, then the estimation error $L(\tilde{\phi}_n) - \inf_{\phi \in \mathcal{C}} L(\phi)$ may be small, but the approximation error $\inf_{\phi \in \mathcal{C}} L(\phi) - L^*$ will be large. On the other hand, if V is too large, the approximation error may be small, but the estimation error could be large. Structural risk minimization (SRM) is an inductive principle for choosing a class $\mathcal{C}$ with the appropriate VC dimension so that the estimation and approximation errors are balanced, in a sense that is made precise in Theorem 2. SRM was originally proposed by Vapnik and Chervonenkis (see [5]). Lugosi and Zeger later analyzed the perfomance of SRM [6], and their analysis is central to this work. We present SRM as laid out in Chapter 18 of [1], which is based on [6]. The latter work uses a different penalty, leading to bounds with slightly smaller constants.

SRM operates as follows. Suppose we have for each $k = 1, 2, \ldots$, a collection $\mathcal{C}^{(k)}$ of classifiers, with corresponding VC dimension $V_k < \infty$. Let $\phi_n^{(k)}$ be a classifier in $\mathcal{C}^{(k)}$ selected by empirical risk minimization. Define the penalty term

$$\Delta(k, n) = \sqrt{\frac{32 V_k \log(en)}{n}},$$

and for $\phi \in \mathcal{C}^{(k)}$, define the penalized risk

$$\tilde{L}_n(\phi) = \widehat{L}_n(\phi) + \Delta(k, n).$$

Structural risk minimization selects the classifier ϕ_n^* minimizing $\tilde{L}_n(\phi_n^{(k)})$ over $k \geq 1$. The strength of structural risk minimization comes from the following results of Lugosi and Zeger [1, 6].

Theorem 2. *Let $\Lambda = \sum_{i=1}^{\infty} e^{-V_i}$, and assume $\Lambda < \infty$. For all n and k, and for all $\epsilon \geq 4\Delta(k, n)$,*

$$\mathbf{P}\left\{ L(\phi_n^*) - \inf_{\phi \in \mathcal{C}^{(k)}} L(\phi) > \epsilon \right\} \leq \Lambda e^{-n\epsilon^2/128} + 8 n^{V_k} e^{-n\epsilon^2/512},$$

and in particular, for all n,

$$\mathbf{E}\{L(\phi_n^*)\} - L^* \leq \inf_{k \geq 1} \left(16 \sqrt{\frac{V_k \log n + 4}{2n}} + \left(\inf_{\phi \in \mathcal{C}^{(k)}} L(\phi) - L^* \right) \right). \tag{1}$$

Corollary 1. *If the Bayes classifier is in $\mathcal{C}^{(k)}$ for some k, then*

$$\mathbf{E}\{L(\phi_n^*)\} - L^* = O\left(\left(\frac{\log n}{n} \right)^{\frac{1}{2}} \right).$$

Corollary 2. *If for all distributions on* (X, Y)

$$\lim_{k\to\infty} \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) = L^*,$$

then $L(\phi_n^*) \to L^*$ *with probability one, i. e.,* ϕ_n^* *is stongly universally consistent.*

The first corollary follows from the second part of Theorem 2, since the approximation error is 0 for some k. The second corollary follows from the first part of Theorem 2 and the Borel-Cantelli lemma (see Appendix A of [1]).

The first part of Theorem 2 comes from the inequality

$$\mathbf{P}\left\{L(\phi_n^*) - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) > \epsilon\right\} \le \mathbf{P}\left\{L(\phi_n^*) - \tilde{L}_n(\phi_n^*) > \frac{\epsilon}{2}\right\} + \mathbf{P}\left\{\tilde{L}_n(\phi_n^*) - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) > \frac{\epsilon}{2}\right\},$$

and bounding both terms on the right using the Vapnik-Chervonenkis inequality. The second part results from writing the expected error as the expected estimation error plus the approximation error,

$$\mathbf{E}\{L(\phi_n^*)\} - L^* = \left(\mathbf{E}\{L(\phi_n^*)\} - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi)\right) + \left(\inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) - L^*\right),$$

and using the first part to bound the first term on the right (see Problem 12.1 in [1]).

Suppose that

$$\lim_{k\to\infty} \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) = L^*.$$

Then, as $n \to \infty$, the bound 1 decays most rapidly for a choice of $k = k(n)$ such that the two terms on the right hand side of 1 have the same size. Even though the penalized DCT does not know the value of k that optimally balances the two terms, it performs as though it does, because of the "inf" in the expression. This supports the assertion that SRM chooses $\mathcal{C}^{(k)}$ so that the estimation and approximation errors are balanced.

Two problems with SRM are: (i) we require $\phi^* \in \mathcal{C}^{(k)}$ for some k in order to deduce the rate of convergence to the Bayes error L^*, and (ii) it is often difficult to find ϕ_n^* efficiently. We will address both of these issues in this paper. With regards to (i), we will define a very general class of distributions on the data, and a certain family of classifiers (dyadic classification trees), so that the approximation error goes to zero at a certain rate without ever having to reach zero. This will allow us to establish rate of convergence results for a very large class of distributions, albeit at slower rates than $O((\log n/n)^{1/2}))$ (see Section 6). Regarding (ii), we also demonstrate two algorithms for finding ϕ_n^* efficiently (Section 8).

3 Classification Trees

Definition 1. *A* classification tree *is a finite, binary tree T such that*

1. *T has a distinguished root node that corresponds to the full input space,*
2. *If $t \in T$ is a non-terminal node with children t_L and t_R, then these three nodes correspond to subsets (called cells) A, A_L, and A_R (respectively) of the input space such that A_L and A_R form a partition of A,*
3. *Every node in T (and the corresponding cell) is assigned a class label.*

If T' is a subtree of T that inherits the partitioning scheme of T, we way that T' is *weak* subtree of T, and write $T' \preceq T$. If in additon T' inherits the labelling from T, we say T' is a *strong* subtree of T, and write $T' \leq T$. Clearly, if $T' \leq T$, then T' is another classification tree.

The *depth* of a node is the length of the path from that node to the root. The root node has depth 0. The depth of a classification tree is the depth of the deepest node. The *size* of T is the number of terminal nodes of T, denoted $|T|$. The cells associated with the terminal nodes of T form a *tree structured partition* of the input space into $|T|$ cells. T corresponds to a classifier ϕ_T in an obvious way: If $A(x)$ denotes the terminal cell of T containing x, and $\alpha(A)$ denotes the class assigned to cell A, then $\phi_T(x) = \alpha(A(x))$. We will refer to T and ϕ_T interchangeably. We will also refer to a node of T and the cell associated with that node interchangeably.

4 Dyadic Classification Trees

Throughout this work we assume that the input data is restricted to the unit hypercube, $[0,1]^d$. This is a realistic assumption for real-world data, provided appropriate translation and scaling has occured. Let $\mathcal{P} = \{R_1, \ldots, R_k\}$ be a tree-structured partition of the input space, where each R_i is a hyperrectangle with sides parallel to the coordinate axes. Given an integer ℓ, let $[\ell]_d$ denote the element of $\{1, \ldots, d\}$ that is congruent to ℓ modulo d. If $R_i \in \mathcal{P}$ is a cell at depth j in the tree, let $R_i^{(1)}$ and $R_i^{(2)}$ be the rectangles formed by splitting R_i at its midpoint along coordinate $[j+1]_d$. As a convention, assume $R_i^{(1)}$ contains those points of R_i that are less than or equal to the midpoint along the dimension being split.

Definition 2. *A* sequential dyadic partition *(SDP) is any partition of* $[0,1]^d$ *that can be obtained by applying the following rules recursively:*

1. *The trivial partition* $\mathcal{P} = \{[0,1]^d\}$ *is an SDP,*

2. *If* $\mathcal{P} = \{R_1, \ldots, R_k\}$ *is an SDP, then so is* $\{R_1, \ldots, R_{i-1}, R_i^{(1)}, R_i^{(2)}, R_{i+1}, \ldots, R_k\}$*, where* i *may be any integer,* $1 \leq i \leq k$*.*

We define a dyadic classification tree *(DCT) to be a sequential dyadic partition with a class label assigned to each node in the tree.*

The partitions are sequential because children must be split along the next coordinate after the coordinate where their parent was split. Such splits are referred to as *forced* splits, as opposed to *free* splits, in which any coordinate may be split. The partitions are dyadic because we only allow midpoint splits.

By a *complete* DCT of depth L, we mean a DCT such that every possible split up to depth L has been made. In a complete DCT, every cell has volume 2^{-L}. If L is a multiple of d, then the cells of a complete DCT are hypercubes of sidelength $2^{-L/d}$.

5 SRM for DCTs

In this section we formulate structural risk minimization for sets of classifiers organized in a nested, tree structured fashion, like DCTs. The material here is adapted from Section 2.

For a positive integer m, let $\mathcal{P}_m$ denote the partition of $[0,1]^d$ into m^d hypercubes of sidelength $1/m$. If $m = 2^J$ for some integer $J \geq 0$, we say m is *dyadic*. In that case, $\mathcal{P}_m$ is the partition formed by a complete DCT of depth dJ. For dyadic m, and for $1 \leq k \leq m^d$, let $\mathcal{D}_m^{(k)}$ denote the collection of all DCTs with k terminal nodes and depth not exceeding dJ.

Observe that the VC dimension of $\mathcal{D}_m^{(k)}$ is k. To see this, suppose $\phi \in \mathcal{D}_m^{(k)}$. Consider $z_1, \ldots, z_k \in \mathbf{R}^d$, where each z_i is in a different terminal node of ϕ. It is clear that any labelling of $z_1, \ldots, z_k$ can be correctly classified by reassigning the labels of ϕ. Thus, $\mathcal{D}_m^{(k)}$ shatters $z_1, \ldots, z_k$. On the other hand, if $z_1, \ldots, z_{k+1} \in \mathbf{R}^d$, then for any $\phi \in \mathcal{D}_m^{(k)}$, at least two of the z_i belong to the same terminal node of ϕ, and so cannot be assigned different labels. Thus, $\mathcal{D}_m^{(k)}$ does not shatter $z_1, \ldots, z_{k+1}$.

Given a dyadic integer m, and training data $\{(X_i, Y_i)\}_{i=1}^n$, for $k = 1, 2, \ldots, m^d$, define

$$\phi_{n,m}^{(k)} = \arg\min_{\phi \in \mathcal{D}_m^{(k)}} \widehat{L}_n(\phi).$$

Define the penalty term

$$\Delta(k, n) = \sqrt{\frac{32k \log(en)}{n}}, \tag{2}$$

and for $\phi \in \mathcal{D}_m^{(k)}$, define the penalized risk

$$\tilde{L}_n(\phi) = \widehat{L}_n(\phi) + \Delta(k, n), \tag{3}$$

as before. The SRM procedure selects the classifier $\phi_{n,m}^*$ from among $\phi_{n,m}^{(k)}, k = 1, 2, \ldots, m^d$, that minimizes $\tilde{L}_n(\phi_{n,m}^{(k)})$. We refer to $\phi_{n,m}^*$ as a *penalized dyadic classification tree.* Observe that the penalty is the same as in Section 2, with $V_k = k$. We have the following analog to Theorem 2.

Theorem 3. *For all n and $k \leq m^d$, and for all $\epsilon \geq 4\Delta(k, n)$,*

$$\mathbf{P}\left\{L(\phi_{n,m}^*) - \inf_{\phi \in \mathcal{D}_m^{(k)}} L(\phi) > \epsilon\right\} \leq e^{-n\epsilon^2/128} + 8n^k e^{-n\epsilon^2/512},$$

and in particular, for all n and m,

$$\mathbf{E}\{L(\phi_{n,m}^*)\} - L^* \leq \min_{1 \leq k \leq m^d} \left(16\sqrt{\frac{k \log n + 4}{2n}} + \left(\inf_{\phi \in \mathcal{D}_m^{(k)}} L(\phi) - L^*\right)\right). \tag{4}$$

Proof: Apply Theorem 2 with $\mathcal{C}^{(k)} = \mathcal{D}_m^{(k)}$ and $V_k = k$ for $k = 1, 2, \ldots, m^d$. $\square$

The next result demonstrates strong "universal" consistency for the penalized DCT, where the "universe" is the set of all distributions supported on the unit hypercube.

Theorem 4. *Suppose $n, m \to \infty$, with $m = m(n)$ assuming only dyadic integer values. If $m^d = o(n/\log n)$, then the penalized dyadic classification tree is strongly consistent for all distributions supported on the unit hypercube.*

Proof: For notational convenience, let $\mathcal{E}_m$ denote $\mathcal{D}_m^{(k)}$ when $k = m^d$. Then $\mathcal{E}_m$ is the collection of regular histogram classifiers that partition the unit hypercube into smaller cubes of sidelength $1/m$. Since $m^d = o(n/\log n)$, for any $\epsilon > 0$, we have $\epsilon \geq 4\Delta(m^d, n)$ for n sufficiently large. Applying the previous theorem, we have

$$\sum_{n=1}^{\infty} \mathbf{P}\left\{ L(\phi^*_{n,m}) - \inf_{\phi \in \mathcal{E}_m} L(\phi) > \epsilon \right\} < \infty.$$

Therefore, the estimation error $L(\phi^*_{n,m}) - \inf_{\phi \in \mathcal{E}_m} L(\phi)$ converges to 0 with probability one (see Appendix A of [1]). The corresponding approximation error is the approximation error of the regular histogram rule mentioned above. It is known that the regular histogram rule is strongly universally consistent when $nm^{-d} \to \infty$ [1]. This is indeed the case under the hypothesis $m^d = o(n/\log n)$, so the approximation error tends to zero as $n \to \infty$, and hence $L(\phi^*_{n,m}) - L^* \to 0$ with probability one. $\square$

The classifier that achieves this result is obtained by growing an initial tree, and pruning this initial tree to obtain a subtree minimizing the penalized risk in Equation 3. The initial tree ϕ_c is a complete DCT of depth $d \log_2 m$ having m^d terminal nodes. Since $m^d = o(n/\log n)$ by assumption, construction of the initial tree requires $m^d = O(n)$ splits. By a result from [7], since ϕ_c is balanced, it may be pruned in $O(m^d \log m^d) = O(n)$ operations.

If d is large, and $m \geq 2$, the prospect of growing an initial tree with m^d terminal nodes may seem daunting. Fortunately, it is not necessary to grow the complete tree ϕ_c before pruning. Observe that for all k, each terminal node of $\phi^{(k)}_{n,m}$ is labelled by majority vote. Also note that each internal node of $\phi^*_{n,m}$ must contain data points from both classes, otherwise it could be pruned back to that node, leading to a lower penalty without increasing the empirical risk. From these remarks, we see that it is not necessary to compute the complete DCT of depth $d \log_2 m$. It suffices to construct an initial classification tree $\phi_0 \leq \phi_c$ by recursively applying the sequential dyadic subdivision scheme until each terminal node of ϕ_0 is either empty or contains data belonging to the same class. The nodes of ϕ_0 are labelled by majority vote, and the search for the penalized risk minimizing classifier $\phi^*_{n,m}$ may be restricted to strong subtrees of ϕ_0.

6 Rate of Convergence to Bayes Error

In this section, we investigate the rate at which the error of the penalized DCT converges to the Bayes error for a general class of distributions. We also derive a lower bound for the minimax rate of convergence for the same class.

Most rate of convergence studies in pattern recognition place a constraint on the regression function $\eta(x) = \mathbf{P}\{Y = 1|X = x\}$ by requiring it to belong to a certain smoothness class (e.g. Lipschitz, Besov, bounded variation). In contrast, the class we study is defined in terms of the regularity of the Bayes decision boundary, denoted $\mathcal{B}$. We allow $\eta(x)$ to be arbitrarily irregular away from $\mathcal{B}$, so long as it is well behaved near $\mathcal{B}$. The Bayes decision boundary is informally defined as $\mathcal{B} = \{x : \eta(x) = 1/2\}$. A more rigorous definition, given below, should take into account the fact that η might not take on the value 1/2. Let $B(x, \epsilon)$ denote the open ball (with respect to Euclidean norm) of radius ϵ centered at x.

Definition 3. *Let $(X,Y) \in \mathbf{R}^d \times \{0,1\}$ be a random pair, and let $\phi^* : \mathbf{R}^d \to \{0,1\}$ be the corresponding Bayes classifier. The* Bayes decision boundary $\mathcal{B}$ *for this problem is the set of all $\mathbf{x} \in \mathbf{R}^d$ such that for all $\epsilon > 0$, there exists sets $A_0, A_1 \subseteq B(x,\epsilon)$, with $P(A_)) > 0$ and $P(A_1) > 0$, such that $\eta \leq 1/2$ on A_0 and $\eta \geq 1/2$ on A_1.*

We now define a class of distributions. Let (X,Y) denote a random pair, as before, where X takes on values in $[0,1]^d$.

Definition 4. *Let $C_1, C_2 > 0$. Define $\mathcal{F}(C_1, C_2)$ to be the collection of all distributions on (X,Y) such that for all dyadic integers m, if we subdivide the unit cube into cubes of side length $1/m$,*

A1 (Bounded marginal): *For any such cube A intersecting the Bayes decision boundary, $\mathbf{P}\{X \in A\} \leq C_1\lambda(A) = C_1/m^d$, where λ denotes the Lebesgue measure.*

A2 (Regularity): *The Bayes decision boundary passes through at most $C_2 m^{d-1}$ of the resulting m^d cubes.*

Define $\mathcal{F}$ to be the class of all (X, Y) belonging to $\mathcal{F}(C_1, C_2)$ for some C_1, C_2.

The first condition holds, for example, if the density of X is essentially bounded with respect to the Lebesgue measure, with essential supremum $\leq C_1$. The second condition can be shown to hold when one coordinate of the Bayes decision boundary is a Lipschitz function of the others. If we require only that A1 and A2 hold for m sufficiently large, the upper bound presented below still holds, but only for n sufficiently large.

The regularity condition A2 is closely related to the notion of box-counting dimension of the Bayes decision boundary [8]. Roughly speaking, A2 holds for some C_2 if and only if the Bayes decision boundary has box-counting dimension $d-1$. The box-counting dimension is an upper bound on the Hausdorff dimension, and the two dimensions are equal for most "reasonable" sets. For example, if $\mathcal{G}$ is a smooth k-dimensional submanifold of $\mathbf{R}^d$, then $\mathcal{G}$ has box-counting dimension k.

6.1 Upper Bound on DCT Rate of Convergence

Proof of the following result is given in Appendix A.

Theorem 5. *Assume the distribution of (X, Y) belongs to $\mathcal{F}(C_1, C_2)$. Let $\phi^*_{n,m}$ be the penalized dyadic classification tree, as described in Section 5. If $m \sim (n/\log n)^{1/(d+1)}$, then there exists a constant $C_3 > 0$ such that for all $n > 2$,*

$$\mathbf{E}[L(\phi^*_{n,m})] - L^* \leq C_3(\log n/n)^{1/(d+1)}.$$

When we write $m \sim (n/\log n)^{1/(d+1)}$, we mean $\log_2 m = \lfloor \log_2((n/\log n)^{1/(d+1)}) + C_0 \rfloor$, where $C_0 \in \mathbf{R}$ is arbitrary. By an argument similar to the one given after Theorem 4, the classifier that achieves the result in this theorem may be constructed in $O(n)$ operations.

The minimal value of C_3 in the above theorem tends to $2C_1C_2$ as $d \to \infty$. Note that similar rate of convergence results for data-grown trees would be more difficult to establish, since the approximation error is random in those cases.

It is possible to eliminate the log factor in the upper bound by means of Alexander's inequality, as discussed in [1, Ch. 12]. This leads to a much larger value of C_3, but an improved asymptotic rate.

To illustrate the significance of Theorem 5, consider a penalized histogram classifer, with bin width determined adaptively by structural risk minimization, as described in [1, Problem 18.6]. For that rule, the best exponent on the rate of convergence for our class is $1/(d+2)$, compared with $1/(d+1)$ for our rule. Intuitively, this is because the adaptive resolution of dyadic classification trees enables them to focus on the $d-1$ dimensional decision boundary, rather than the d dimensional regression function.

In the event that the data X occupies a $d' < d$ dimensional subset of $[0,1]^d$, the proof of Theorem 5 follows through as before, but with an exponent of $d'+1$ instead of $d+1$. Thus, the penalized DCT is able to automatically adapt to the dimensionality of the input data.

6.2 Minimax Lower Bound

The next result demonstrates that complexity-regularized DCTs nearly achieve the minimimax rate for our class of distributions.

Theorem 6. *Let δ_n denote any discrimination rule based on training data. There exists a constant $C > 0$ such that for n sufficiently large,*

$$\inf_{\delta_n} \sup_{\mathcal{F}} E[L(\delta_n)] - L^* \quad \geq \quad C\, n^{-1/d}.$$

Proof: Earlier we remarked that condition A2 is satisfied when one coordinate of the Bayes decision boundary is a Lipschitz function of the others. From this, it follows that our class of distributions contains the set $\mathcal{F}_{frag}$ studied in [9] (with $\gamma = \kappa = 1$ therein). The theorem follows imediately from Theorem 2 in [9]. The proof of that result is in turn based on Assouad's lemma.

Theorems 5 and 6, together with the above remark on Alexander's inequality, show that complexity-regularized DCTs are close to minimax-rate optimal (within a small factor of $n^{1/(d(d+1))}$ for the class $\mathcal{F}$. We suspect that the class studied by Tsybakov [9], used in our minimax proof, is more restrictive than our class. Therefore, it may be that the exponent $1/d$ in the above theorem can be decreased to $1/(d+1)$, in which case we achieve the minimax rate.

Although bounds on the minimax rate of convergence in pattern recognition have been investigated in previous work [10, 11], the focus has been on placing regularity assumptions on the regression function $\eta(x) = \mathbf{P}\{Y = 1|X = x\}$. Yang demonstrates that in such cases, for many common function spaces (e.g. Lipschitz, Besov, bounded variation), classification is not easier than regression function estimation [11]. This contrasts with the conventional wisdom that, in general, classification *is* easier than regression function estimation [1, Ch. 6]. Our approach is to study minimax rates for distributions defined in terms of the regularity of the Bayes decision boundary. With this framework, we see that minimax rates for classification can be orders of magnitude faster than for estimation of $\eta(x)$, since $\eta(x)$ may be arbitrarily irregular away from the decision boundary for distributions in our class. This view of minimax classification has also been adopted by Mammen and Tsybakov [9, 12]. Our contribution with respect to their work is an implementable discrimination rule, with guaranteed computational complexity, that nearly achieves the minimax lower bounds.

7 Comparison with Data-Grown Classification Trees

The structure of the initial DCT (before pruning) does not depend on the data. This allows us to apply SRM, which assumes a fixed sequence of sets of classifiers $\mathcal{C}^{(k)}$. However, recent work by Nobel [3] extends the results of Lugosi and Zeger to classifiers based on initial trees that depend on the data (e.g., trees obtained by a succession of stepwise optimal splits based on impurity functions). We refer to such trees as *data-grown*. We now summarize Nobel's work, and compare with our DCT upper bound in the previous section.

Let T_n be a classification tree that results from an initial growing procedure. For data-grown classification trees, T_n may be constructed with arbitrary hyperplane splits. For dyadic classification trees, T_n is a complete DCT of depth Jd for some integer J. See Section 5 for details. Assume that each node of T_n is labelled according to the majority vote over training samples in that node. (For data-grown trees, if a *pruning dataset* is used to prune the tree, independent from the data used to grow the tree, assume the nodes are labelled according to the pruning set). The next theorem summarizes the previously mentioned results.

Theorem 7. *For dyadic classification trees, let*

$$\Delta(k,n) = \sqrt{\frac{32k\log(en)}{n}},$$

and

$$\Gamma(k,n) = 16\sqrt{\frac{k\log n + 4}{2n}}.$$

For data-grown trees where the growing dataset equals the pruning dataset, let

$$\Delta(k,n) = \sqrt{\frac{32(kd\log n + k\log 2 + 2\log k)}{n}},$$

and

$$\Gamma(k,n) = 2\Delta(k,n) + \frac{6\sqrt{2}}{\sqrt{dkn\log n}}.$$

For data-grown trees with independent pruning and growing datasets, let

$$\Delta(k,n) = \sqrt{\frac{k\log 2 + \log|T|}{n}},$$

and

$$\Gamma(k,n) = 2\Delta(k,n) + \frac{3}{2\sqrt{kn\log 2}}.$$

If T_n^* *is obtained by minimizing* $\tilde{L}_n(T) = \widehat{L}_n(T) + \Delta(|T|,n)$ *over all strong subtrees* $T \leq T_n$*, then*

$$\mathbf{E}\{L(T_n^*)\} - L^* \leq \min_{1\leq k\leq |T_n|}\left\{\Gamma(k,n) + \mathbf{E}\left[\min_{T\preceq T_n, |T|\leq k}(L(T) - L^*)\right]\right\}. \tag{5}$$

$\Gamma(k,n)$ is an upper bound on the expected estimation error, $\mathbf{E}\{L(T_n^*) - \inf_{T\preceq T_n, |T|=k} L(T)\}$. The second term on the righthand side of (5) is the expected approximation error. For dyadic classification trees, the expectation may be dropped, since the initial tree T_n is not data-dependent and thus the approximation error is not random.

Recall that in the case of DCTs, this theorem was used to show strong consistency for data supported in the unit hypercube. For data-grown trees with identical growing and pruning sets, Nobel shows weak consistency of the pruned tree under certain conditions on the initial tree [3].

In particular, observe that $\Gamma(k,n) \to 0$ if $k = o(n/\log n)$. Thus, if the sequence of initial trees T_n contains a subsequence of trees T'_n such that $\mathbf{E}\{L(T'_n)\} \to L^*$ and $\mathbf{E}\{|T'_n|\} = o(n/\log n)$, then ϕ^*_n is weakly consistent. Nobel gives conditions on T_n that guarantee this property.

Similar weak consistency results exist for when the growing and pruning datasets are independent. Note that in this case, $\Gamma(k,n) \to 0$ under the weaker hypothesis $k = o(n)$, reflecting the advantage of having independent data for growing and pruning.

8 Two Efficient Pruning Algorithms

This section has been expanded and is now contained in a separate technical report [7].

9 Experimental Comparison

To gain a rough idea of the usefulness of dyadic classification trees in practice, we compared two DCT based classifiers with a greedy tree growing procedure, similar to that used by CART [2] or C4.5 [13], where each successive split is chosen to maximize an information gain defined in terms of an impurity function. We considered four two-class datasets, available on the web at http://www.ics.uci.edu/~mlearn/MLRepository.html. For each dataset, we randomly split the data into two halves to form training and testing datasets.

For the greedy growing scheme, we used half of the training data to grow the tree, and constructed every possible pruning of the initial tree with an additive penalty. The best pruned tree was chosen to minimize the holdout error on the rest of the training data. We call this classifier CART-HOLD. The second classifier, DCT-HOLD, was constructed in a similar manner, except that the initial tree was a complete DCT, and all of the training data was used for computing the holdout error estimate. Finally, we implemented the complexity-regularized DCT, denoted DCT-SRM, with square-root penalty determined by Equation 2. Table 1 shows the misclassification rate for each algorithm on each dataset.

From these expirements, we might conclude two things: (i) The greedily-grown partition outperforms the dyadic partition; and (ii) Much of the descrepancy between CART-HOLD and DCT-SRM comes from the partitioning, and not from the model selection method (holdout versus SRM). In-

Table 1: Comparison of a greedy tree growing procedure, with model selection based on holdout error estimate, and two DCT based methods. Numbers shown are test errors.

	CART-HOLD	DCT-HOLD	DCT-SRM
Pima Indian Diabetes	26.8 %	27.2 %	33.0 %
Wisconsin Breast Cancer	4.7 %	6.4 %	6.3 %
Ionosphere	12.88 %	18.6 %	18.8 %
Waveform	19.8 %	29.1 %	31.0 %

deed, DCT-SRM beats or nearly equals DCT-HOLD on three of the four datasets. Conclusion (i) may be premature, for it is shown in [1, Ch. 20] that greedy partitioning based on impurity functions can perform arbitrarily poorly for some distributions, while this is never the case for complexity-regularized DCTs. In light of (ii), it may be possible to apply Nobel's pruning rules for data-grown trees [3], which can now be implemented with our algorithm, to equal or surpass the performance of CART, while avoiding the heuristic and computationally expensive cross-validation technique usually employed by CART to determine the appropriately pruned tree.

10 Relations to Previous Work

Classification trees have been a popular approach to supervised learning since the pioneering work of Breiman, et. al. [2]. For a recent survey of techniques for growing and pruning classification trees, see [14]. For an empirical comparison of various classification tree induction strategies, see [15]. Several tree growing strategies leading to consistent classifiers are given in [1]. For a general introduction to classification trees, see [16] or [17].

The pruning strategy presented here is based on recent work on adaptive model selection using complexity regularization [6]. Other classification tree pruning strategies that use complexity regularization are described in [18], [19], and especially [3], to which the present work is most closely related. An application of complexity regularization to the pruning of regression trees may be found in [20]. Other uses of complexity regularization appear in [5, 21–23].

Dyadic thinking has also been applied to a variety of statistical problems, especially density estimation and function estimation. A form of dyadic classification trees was introduced in [24], but only for two-dimensional problems. In that approach, free splits are allowed, whereas we require forced (sequential) splits. Extending the free-split approach to d dimensions would require $O(n^d)$ operations for optimal growing and pruning with an additive penalty term, and even greater computational complexity for optimal growing and pruning with a square-root penalty.

Rate of convergence results have not been common in the pattern recognition community until recently [6,9–12,24]. The main reason for this is that any rate of convergence result must necessarily place a constraint on the distribution of the data, a result due to Devroye [25], [1, Ch. 7]. Whether such constraints are satisfied in practice is impossible to verify given only a finite collection of training samples.

We feel that rate of convergence studies can aid in comparing the relative performance of two classifiers, and in motivating new discrimination rules. The constraints considered in this investigation are quite mild, and may hold in a wide array of practical situations. Furthermore, rate of convergence and approximation-theoretic results have been very useful in other areas, for example in connection with wavelets, where they have led to an increased understanding of denoising and compression algorithms [26–29]. We feel that similar analysis can lead to insights into the strengths and weaknesses of various classification tools, from trees to support vector machines. The rate of convergence results presented here were inspired by Donoho [30].

11 Future Work

Our first priority is to improve our minimax lower bound so that it matches the upper bound, or else understand why this is not possible.

For large values of the d, the marginal improvement of the penalized dyadic classification tree over the penalized histogram classifier becomes negligible, and both error rates decay too slowly to be useful. We intend to pursue tree-based classifiers that give better performance in high dimensions (say $d \geq 100$). One possible approach is to classify terminal nodes using polynomial classifiers of degree p. As the regularity condition on the data is strengthened, and p is increased appropriately, it should be possible to obtain rates of convergence approaching $O(\log n/n)$. Although minimizing the empirical risk over polynomial classifiers is computationally infeasible, it may be sufficient to use polynomial support vector machines instead (see [31]).

Another high priority is to more thoroughly evaluate the usefulness of our pruning algorithms on real-world data. One interesting question is, when selecting a tree by cross-validation, whether it is sufficient to look at the sequence of trees produce by CART, or whether is pays to look at all of the trees produced by the main loop of the first pruning algorithm, of which the CART trees are a subset. Another interesting question is whether the theoretical penalties in Theorem 7 are appropriate for small sample sizes, or whether different weighting factors should be used.

12 Conclusions

Dyadic classification trees exhibit desirable theoretical properties (finite sample risk bounds, consistency, near minimax-rate optimality) and can be trained extremely rapidly. The minimax result demonstrates that other discrimination rules, such as neural networks or support vector machines, cannot significantly outperform DCTs (in this minimax sense). This minimax result is asymptotic, and considers worst-case distributions. From a practical standpoint, with finite samples and non-worst-case distributions, other rules may beat DCTs, which our experiments on benchmark datasets confirm. The sequential dyadic partitioning scheme is especially susceptible when many of the features are irrelevant, since it must cycle through all features before splitting a feature again. Several modifications to the current dyadic partitioning scheme may be envisioned, such as free dyadic or median splits.

Such modified tree induction strategies would still possess many of the desirable theoretical properties of DCTs, especially those falling under Nobel's framework [3]. Our square-root pruning algorithms now provide a means of implementing his pruning schemes for comparison with other model selection techniques (e.g., holdout or cross-validation). It remains to be seen whether the rate of convergence analysis presented here extends to his work.

Acknowledgments

The authors would like to thank Ben Graf for his assistance in collecting the experimental results of Section 9. This work was partially supported by the National Science Foundation, grant no. MIP–9701692, the Army Research Office, grant no. DAAD19-99-1-0349, and the Office of Naval Research, grant no. N00014-00-1-0390.

A Proof of Theorem 5

Assume the distribution of (X, Y) belongs to $\mathcal{F}(C_1, C_2)$. Let $m = 2^J$. Recall that $\mathcal{P}_m$ is the partition of $[0,1]^d$ into cubes of sidelength $1/m = 2^{-J}$. It is also the partition formed by a complete DCT of depth dJ. Let $\mathcal{B}_m$ be the collection of cells in $\mathcal{P}_m$ that intersect the Bayes decision boundary for (X, Y). Then $|\mathcal{B}_m| \leq C_2 m^{d-1}$, by property A2. The following lemma expresses the adaptive resolution of the penalized DCT, which allows the faster rate of convergence.

Lemma 1. *Suppose the distribution of* (X, Y) *belongs to* $\mathcal{F}(C_1, C_2)$. *For all dyadic integers* m,

$$\inf_{\phi \in \mathcal{D}_m^{(k)}} L(\phi) - L^* \leq \frac{C_1 C_2}{m}$$

for some $k \leq 4dC_2 m^{d-1}$.

Proof: Let ϕ_c be the complete DCT of depth dJ with terminal nodes labelled so as to minimize the probability of error. There are 2^j nodes at depth j in ϕ_c. A classifier ϕ that satisfies the stated bound is the strong subtree of ϕ_c consisting of all those nodes in ϕ_c that are ancestors of nodes in $\mathcal{B}_m$, together with their children. Observe that only those nodes in ϕ_c that intersect the Bayes decision boundary can be ancestors of nodes in $\mathcal{B}_m$. By property A2, there are at most $C_2 2^{\lceil \ell/d \rceil (d-1)}$ nodes of ϕ_c at depth ℓ that can intersect the Bayes decision boundary. Hence, there are at most

$$\sum_{\ell=0}^{Jd} C_2 2^{\lceil \ell/d \rceil (d-1)} \leq \sum_{j=0}^{J} dC_2 2^{j(d-1)} \leq 2dC_2 2^{J(d-1)} = 2dC_2 m^{d-1}$$

ancestors of cells in $\mathcal{B}_m$. Since the terminal nodes of ϕ are the children of ancestors of cells in $\mathcal{B}_m$, ϕ has size $k \leq 4dC_2 m^{d-1}$.

Next we show that ϕ approximates the Bayes classifier ϕ^* for (X, Y) to the specified accuracy. Let $\mathcal{A}_m$ be the collection of terminal nodes of ϕ. Observe that ϕ and ϕ^* agree on all cells $A \in \mathcal{A}_m \backslash \mathcal{B}_m$. Then,

$$\begin{aligned} L(\phi) - L(\phi^*) &= \sum_{A \in \mathcal{A}_m} \mathbf{P}\{X \in A\}(\mathbf{P}\{\phi(X) \neq Y | X \in A\} - \mathbf{P}\{\phi^*(X) \neq Y | X \in A\}) \\ &= \sum_{A \in \mathcal{B}_m} \mathbf{P}\{X \in A\}(\mathbf{P}\{\phi(X) \neq Y | X \in A\} - \mathbf{P}\{\phi^*(X) \neq Y | X \in A\}) \end{aligned}$$

$$
\begin{aligned}
&\le \sum_{A\in\mathcal{B}_m} \mathbf{P}\{X\in A\}\\
&\le \sum_{A\in\mathcal{B}_m} \frac{C_1}{m^d}\\
&\le \frac{C_1C_2}{m}.
\end{aligned}
$$

The next to last inequality comes from condition A1, and the fact that $\lambda(A)=1/m^d$ for $A\in\mathcal{B}_m$. The last inequality comes from $|\mathcal{B}_m|\le C_2 m^{d-1}$. □

We are now ready to apply Theorem 3. The first term on the right-hand side of (4) can be bounded by $16\sqrt{k\log n/n}$ for $n\ge 3$ and $k\ge 4$. In that case, by Lemma 1, we have

$$\mathbf{E}\{L(\phi^*_{n,m})\}-L^*\le (32\sqrt{dC_2})\sqrt{\frac{m^{d-1}\log n}{n}}+\frac{C_1C_2}{m}.$$

Now take $m=2^J$ with $J=\lfloor \log_2((n/\log n)^{1/(d+1)})+C_0\rfloor$, where $C_0\in\mathbf{R}$ is arbitrary. Using the inequalities

$$\log_2((n/\log n)^{1/(d+1)})+C_0-1 \;<\; J \;\le\; \log_2((n/\log n)^{1/(d+1)})+C_0,$$

we deduce that

$$m\le 2^{C_0}\left(\frac{n}{\log n}\right)^{\frac{1}{d+1}} \quad\text{and}\quad \frac{1}{m}\le 2^{1-C_0}\left(\frac{\log n}{n}\right)^{\frac{1}{d+1}}.$$

Applying these inequalities to the previous bound, we get

$$\mathbf{E}\{L(\phi^*_{n,m})\}-L^*\le C_3\left(\frac{\log n}{n}\right)^{\frac{1}{d+1}},$$

where $C_3=2^{C_0(d-1)/2}32\sqrt{dC_2}+2^{1-C_0}C_1C_2$. This proves the first part of the theorem. Note that the value of C_3 is minimized when

$$C_0=\frac{2}{d+1}\log_2\left(\frac{C_1\sqrt{C_2}}{8\sqrt{d}(d-1)}\right),$$

in which case

$$C_3=\left(dC_1^{d-1}C_2^d\right)^{\frac{1}{d+1}}\left[32(8(d-1))^{-\frac{d-1}{d+1}}+2(8(d-1))^{\frac{2}{d+1}}\right].$$

As $d \to \infty$, this minimal value of C_3 tends to $2C_1C_2$.

A similar argument can be used to prove the second part of the theorem. For the penalized histogram classifier, in order to achieve the same approximation accuracy as in Lemma 1, we need a partition with m^d cells. The desired result follows from this fact and Theorem 2, in a manner similar to the previous case.

References

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

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

[3] A. Nobel, "Analysis of a complexity based pruning scheme for classification trees," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2362–2368, 2002.

[4] V. Vapnik and C. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

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

[6] G. Lugosi and K. Zeger, "Concept learning using complexity regularization," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 48–54, 1996.

[7] C. Scott, "Tree pruning using a non-additive penalty," Tech. Rep. TREE 0301, Rice University, 2003, available at http://www.dsp.rice.edu/∼cscott.

[8] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, Wiley, West Sussex, England, 1990.

[9] A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," *preprint*, 2001, available at http://www.proba.jussieu.fr/mathdoc/preprints/.

[10] J. S. Marron, "Optimal rates of convergence to Bayes risk in nonparametric discrimination," *Annals of Statistics*, vol. 11, no. 4, pp. 1142–1155, 1983.

[11] Y. Yang, "Minimax nonparametric classification–Part I: Rates of convergence," *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2271–2284, 1999.

[12] E. Mammen and A. B. Tsybakov, "Smooth discrimination analysis," *Annals of Statistics*, vol. 27, pp. 1808–1829, 1999.

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

[14] S. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey," *Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 345–389, 1998.

[15] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," *Machine Learning*, vol. 40, pp. 203–299, 2000.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2001.

[17] B. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, UK, 1996.

[18] J. R. Quinlan and R. L. Rivest, "Inferring decision trees using the minimum description length principle," *Information and Computation*, vol. 80, no. 3, pp. 227–248, 1989.

[19] C. S. Wallace and J. D. Patrick, "Coding decision trees," *Machine Learning*, vol. 11, pp. 7–22, 1993.

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

[21] 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.

[22] A. 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.

[23] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Annals of Statistics*, vol. 11, pp. 416–431, 1983.

[24] D. Donoho, "CART and best-ortho-basis selection: A connection," *Annals of Statistics*, vol. 25, pp. 1870–1911, 1997.

[25] L. Devroye, "Any discrimination rule can have an arbitrarily bad probability of error for finite sample size," *PAMI*, vol. 4, pp. 154–157, 1982.

[26] D. Donoho and I. Johnstone, "Ideal adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[27] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.

[28] R. A. DeVore and B. J. Lucier, "Fast wavelet techniques for near-optimal image processing," *IEEE Communications Society*, pp. 1129–1135, 1992.

[29] R. A. DeVore, B. Jawerth, and B. J. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. Image Processing*, vol. 38, pp. 719–746, 1992.

[30] D. Donoho, "Wedgelets: Nearly minimax estimation of edges," *Annals of Statistics*, vol. 27, pp. 859–897, 1999.

[31] K. Bennett, N. Cristianini, J. Shawe-Taylor, and D. Wu, "Enlarging the margins in perceptron decision trees," *Machine Learning*, vol. 41, pp. 295–313, 2000.