CORT: Classification Or Regression Trees

Clayton Scott, Rebecca Willett and Robert Nowak *

March 18, 2004

Abstract

In this paper we challenge three of the underlying principles of CART, a well known approach to the construction of classification and regression trees. Our primary concern is with the penalization strategy employed to prune back an initial, overgrown tree. We reason, based on both intuitive and theoretical arguments, that the pruning rule for classification should be different from that used for regression (unlike CART). We also argue that growing a tree-structured partition that is specifically fitted to the data is unnecessary. Instead, our approach to tree modeling begins with a non-adapted (fixed) dyadic tree structure and partition, much like that underlying multiscale wavelet analysis. We show that dyadic trees provide sufficient flexibility, are easy to construct, and produce near-optimal results when properly pruned. Finally, we advocate the use of a negative log-likelihood measure of empirical risk. This is a more appropriate empirical risk for non-Gaussian regression problems, in contrast to the sum-of-squared errors criterion used in CART regression.

1 Introduction

In regression, the objective is to estimate a function $f^* : \mathbf{R}^d \to \mathbf{R}$ based on a random sample of input-output pairs $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathbf{R}^d$ and $y_i \in \mathbf{R}$. In classification, the objective is to construct a classifier whose performance is close to the Bayes-optimal classifier $\phi^* : \mathbf{R}^d \to \{0, 1, \ldots, M-1\}$, also based on a random sample of points $\{(x_i, y_i)\}_{i=1}^n$, where now $y_i \in \{0, 1, \ldots, M-1\}$ represents class label associated with the input x_i. In this paper we consider the two-class problem $M = 2$.

A common approach to solving classification and regression problems is to partition the input space in a tree-structured fashion, and construct an estimator $\widehat{f}$ or a classifier $\widehat{\phi}$ by fitting to the data in each cell of the partition. The first such tree-based method to gain wide recognition was CART (Classification and Regression Trees) [1]. After nearly two decades, the techniques presented in that seminal work continue to influence the design of new tree-based algorithms.

In this paper we challenge several of the underlying principles of CART. Our first concern is with the penalization strategy employed to prune back an overgrown tree. We reason, based on both intuitive and theoretical arguments, that the pruning rule for classification should be different from that used for regression

*R. Nowak was supported by the National Science Foundation, grants CCR-0310889 and ANI-0099148, the Office of Naval Research grant N00014-00-1-0390, and the State of Texas ATP, grant 003604-0064-2001. R. Willett and R. Nowak are with the Departments of Electrical and Computer Engineering at the University of Wisconsin-Madison and Rice University.

(unlike CART). Hence, our title, Classification *or* Regression Trees. Second, we argue that growing trees that are adapted to fit the data is an unnecessary step, and advocate the use of dyadic trees instead. We show that dyadic trees provide sufficient flexibility, are easy to construct, and produce near-optimal results when properly pruned. Third, in the regression setting, we replace the usual sum-of-squared errors criterion with the negative log-likelihood function, which more accurately reflects the randomness in the data and leads to near-optimal theoretical performance.

2 Review of CART

CART traditionally involves two phases: growing and pruning. In the growing phase, the input domain is recursively partitioned into cells. Each cell corresponds to a leaf of a large initial tree. The partitioning is often done to fit the data as closely as possible, although as we will discuss later non-adaptive initial trees/partitions have certain advantages. The initial tree usually provides a very good, perhaps perfect, fit to the data. Unfortunately, this can mean that the tree is overfitting, and that its true predictive capabilities may be very sub-optimal. To avoid overfitting, the initial tree is pruned. Let $\mathcal{T}$ denote the set consisting of the initial tree and all possible prunings of this tree. CART selects the tree in $\mathcal{T}$ that minimizes

$$C(T) \;=\; \widehat{L}_n(T) \,+\, \alpha\,|T|, \tag{1}$$

where $\widehat{L}_n(T)$ is the *empirical risk* (estimation or classification error on the training data) using the tree T, $|T|$ is the cardinality of the tree (i.e., the number of leaf nodes or partition cells), and $\alpha > 0$ is a constant that controls the trade-off between fidelity to the training data and the complexity of the tree. For regression the empirical risk is typically of the form

$$\widehat{L}_n(T) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_T(x_i) - y_i)^2.$$

Recent results discussed herein and in [2, 3] demonstrate that this criterion is appropriate for Gaussian regression problems. For classification, the empirical risk is

$$\widehat{L}_n(T) = \frac{1}{n}\sum_{i=1}^{n} I(\widehat{\phi}_T(x_i) \neq y_i),$$

where I denotes the indicator function.

Minimizing (1) produces a tree $\widehat{T}$ and a corresponding estimator $\widehat{f} = f(\widehat{T})$ or classifier $\widehat{\phi} = \phi(\widehat{T})$. Specifically, a model (regression function or classification label) is fitted to each cell of the partition associated with $\widehat{T}$ to minimize the empirical risk. For example, the estimator $\widehat{f}$ could be constant on each cell of the partition, with the constant value equal to the average of y_i in that cell. Similarly, the classifier $\widehat{\phi}$ is constant on each cell, with the classification label determined by a majority vote of the training data in the cell.

We argue here that the CART criterion is natural and in a certain sense optimal for regression problems, but that it tends to penalize large trees too aggressively in the classification context. Instead, for classification we show that an alternative criterion of the form

$$C(T) \;=\; \widehat{L}_n(T) \,+\, \alpha\,|T|^{1/2}, \tag{2}$$

is appropriate and optimal for the classification problem. Remarkably, both (1) and (2) can be solved by the efficient, bottom-up pruning process traditionally used in CART. We also suggest that in non-Gaussian regression problems (e.g., Poisson, multinomial) it is more appropriate to employ the negative log-likelihood function as the empirical risk instead of the usual sum of squared errors (note the two are the same in the Gaussian case). In fact, the theoretical performance bounds discussed later will only hold with the negative log-likelihood measure of empirical risk.

3 Bias-Variance Trade-off in Classification and Regression

The theoretical performance of the estimator $\widehat{f}$ or classifier $\widehat{\phi}$ is measured in terms of a *risk function*, denoted by R. The empirical risk described in the preceding section, $\widehat{L}_n$, is an estimate of the true risk, R. For regression, the risk is typically a mean square error (MSE). If the observations $\{y_i\}$ are Gaussian distributed, then

$$R(\widehat{f}, f^*) = \mathbf{E}\left[(f^* - \widehat{f})^2\right],$$

where $\mathbf{E}$ denotes the expectation operator. In the case where the observations are Poisson or multinomial distributed, as arises in intensity or density estimation,

$$R(\widehat{f}, f^*) = \mathbf{E}\left[\left(\sqrt{f^*} - \sqrt{\widehat{f}}\right)^2\right].$$

This "square-root" scale MSE is necessary to stabilize the density-dependent variance of Poisson or multinomial processes. For classification, the risk function is

$$R(\widehat{\phi}, \phi^*) = \mathbf{E}\left[L(\widehat{\phi})\right] - L(\phi^*).$$

Here, $L(\phi) = \mathbf{P}\{\phi(X) \neq Y\}$ is the probability of error for the classifier ϕ, and $L(\phi^*)$ is the Bayes error, which is the minimum probability of error among all possible classifiers.

While the risk functions for classification and regression are different, both risk functions have a decomposition of the form $R = R_1 + R_2$, such that R_1 decreases and R_2 increases as the complexity of the classifier/estimator increases. For regression in the Gaussian case, we have the familiar bias-variance decomposition:

$$R(\widehat{f}, f^*) = \left(f^* - \mathbf{E}[\widehat{f}]\right)^2 + \mathbf{E}\left[(\widehat{f} - \mathbf{E}[\widehat{f}])^2\right].$$

Thus, we have $R_1 = (f^* - \mathbf{E}[\widehat{f}])^2$ as the squared bias term and $R_2 = \mathbf{E}\left[(\widehat{f} - \mathbf{E}[\widehat{f}])^2\right]$ as the variance term; a similar decomposition holds in the Poisson/multinomial cases.

In classification, the risk is written in terms of the *approximation error* and *estimation error*:

$$R(\widehat{\phi}, \phi^*) = (L_{\mathcal{C}} - L(\phi^*)) + \left(\mathbf{E}\left[L(\widehat{\phi})\right] - L_{\mathcal{C}}\right),$$

where $L_{\mathcal{C}} = \inf_{\phi \in \mathcal{C}} L(\phi)$. For example, $\mathcal{C}$ might be the collection of all tree classifiers with no more than 10 leaf nodes. Here the approximation error, $R_1 = (L_{\mathcal{C}} - L(\phi^*))$, functions as the bias term, while the estimation error, $R_2 = \left(\mathbf{E}\left[L(\widehat{\phi})\right] - L_{\mathcal{C}}\right)$ functions as the variance. For convenience, we use the terms "bias" and "variance" to refer to R_1 and R_2, respectively, for both regression and classification problems.

4 Proper penalties for tree pruning

For tree-based methods, the complexity of a classifier or estimator is quantified in terms of the number of leaf nodes of the tree. Let $R(k)$ denote the risk associated with a tree estimator or classifier based on a tree with k leaf nodes. Let $R_1(k)$ and $R_2(k)$ denote the corresponding bias and variance, respectively. Generally, the bias cannot be gauged without some knowledge of the true function or Bayes optimal classifier. The variance, however, can be assessed in both cases, without knowledge of the underlying functions or distributions, as we will see below. Thus, assume the variance $R_2(k)$ grows like (or is bounded by) a certain function $g(k)$ depending on the number of leaf nodes k. For simplicity, assume that $R_2(k) = \alpha g(k)$, for some $\alpha > 0$.

Since the risk is the sum of $R_1(k)$ and $R_2(k)$, two positive quantities, it is clear that no k-leaf tree can achieve a risk lower than $R_2(k)$. Therefore, if a tree has an empirical error that falls below this lower bound, then the empirical error is no longer an accurate estimate of the true error, and one can infer that the tree is overfitting the training data, as depicted in Figure 1. This reasoning forms the intuitive basis for tree pruning strategies. As shown in the next section, for regression problems the function $R_2(k)$ is linear in k. This agrees with the usual CART penalty. For classification, $R_2(k)$ is sublinear and behaves like $k^{1/2}$, as shown below in Figure 1. This suggests the modified CART criterion in (2).

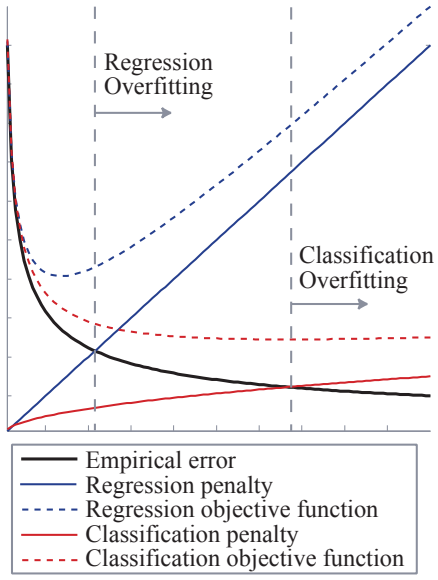

Figure 1: Penalty structures for classification and regression. When the empirical risk is less than the variance, overfitting occurs. By choosing a penalty proportional to the variance term (which is different for classification than for regression), CORT produces a pruned tree near the beginning of the overfitting region.

5 Bounding the Variance Term

In this section we derive bounds on the variance $R_2(k)$ for the classification and regression problems. These bounds will help us to establish the proper penalties for pruning.

5.1 Regression

The variance term $R_2(k) \propto k$, the number of leaves (degrees of freedom) in the tree-based estimator. To see this, consider the case in which are data are n samples of a signal contaminated with Gaussian white noise with power σ^2. Let T be a tree with k leaves, and define the estimator $\widehat{f}$ as the sample-average over each of the cells in the partition defined by T. The average of the samples in each cell is Gaussian with variance σ^2, and thus the total variance of the estimator is $k\sigma^2$. Similar conclusions can be made for other data types (e.g., Poisson, multinomial), and more sophisticated regression models (e.g., polynomial fits in each cell).

5.2 Classification

Let us first consider a simple case in which $\mathcal{C} = \mathcal{C}_k$ is the collection of all tree classifiers corresponding to the different possible labelings of a fixed tree-structured partition of $\mathbf{R}^d$ having k cells. In this case, there are 2^k different classification trees in $\mathcal{C}_k$, and $R_2(k) = \mathbf{E}\left[L(\widehat{\phi})\right] - L_{\mathcal{C}_k}$. The classifier $\widehat{\phi}$ is chosen to minimize the empirical risk, in which case each cell of the partition is labelled by majority vote.

The derivation of a bound on $R_2(k)$ proceeds in three steps. In the first step, we observe

$$\begin{aligned}
& L(\widehat{\phi}) - L_{\mathcal{C}_k} \\
&= L(\widehat{\phi}) - \min_{\phi \in \mathcal{C}_k} L(\phi) \\
&= L(\widehat{\phi}) - \widehat{L}_n(\widehat{\phi}) + \widehat{L}_n(\widehat{\phi}) - \min_{\phi \in \mathcal{C}_k} L(\phi) \\
&\leq L(\widehat{\phi}) - \widehat{L}_n(\widehat{\phi}) + \max_{\phi \in \mathcal{C}_k} \left| \widehat{L}_n(\phi) - L(\phi) \right| \\
&\leq 2 \max_{\phi \in \mathcal{C}_k} \left| \widehat{L}_n(\phi) - L(\phi) \right|.
\end{aligned}$$

In the second step, the above fact is used to bound

$$\begin{aligned}
& \mathbf{P}\left(L(\widehat{\phi}) - \min_{\phi \in \mathcal{C}_k} L(\phi) > \epsilon \right) \\
&\leq \mathbf{P}\left(\max_{\phi \in \mathcal{C}_k} \left| \widehat{L}_n(\phi) - L(\phi) \right| > \frac{\epsilon}{2} \right) \\
&\leq \sum_{\phi \in \mathcal{C}_k} \mathbf{P}\left(\left| \widehat{L}_n(\phi) - L(\phi) \right| > \frac{\epsilon}{2} \right) \\
&\leq 2^k e^{-n\epsilon^2/2},
\end{aligned}$$

where in the last step we use Chernoff's bound and the fact that $n\widehat{L}_n(\phi) \sim \text{Binomial}(n, L(\phi))$. For the third step, let $z = L(\widehat{\phi}) - \min_{\phi \in \mathcal{C}_k} L(\phi)$. We want to bound $\mathbf{E}[z] = R_2(k)$. Now observe

$$\begin{aligned}
\mathbf{E}\left[z^2\right] &= \int_0^\infty \mathbf{P}(z^2 > t)\, dt \\
&= \int_0^u \mathbf{P}(z^2 > t)\, dt + \int_u^\infty \mathbf{P}(z^2 > t)\, dt \\
&\leq u + \int_u^\infty 2^k e^{-nt/2}\, dt \\
&= u + \frac{2^{k+1}}{n} e^{-nu/2}.
\end{aligned}$$

Minimizing this bound with respect to u, we obtain

$$\mathbf{E}\left[z^2\right] \leq \frac{2}{n}\left(k \log 2 + 1\right),$$

and thus, by Jensen's inequality, we have

$$R_2(k) = \mathbf{E}[z] \leq \sqrt{\mathbf{E}[z]} \leq \alpha\sqrt{k},$$

where α is a constant not depending on k. The square-root bound on the growth of the variance in classification holds in much more general cases, for both adaptive and non-adaptive tree structures [4, 5]. The key to such results is to replace Chernoff's bound by the Vapnik-Chervonenkis inequality [6, 7] or extensions of the Vapnik-Chervonenkis inequality for data-dependent partitions [5].

6 Risk Bounds for Dyadic Classification and Regression Trees

To characterize the theoretical performance of classification and regression trees, the rate at which the risk converges to zero can be bounded by assuming the true function or Bayes classifier belongs to a certain (smoothness) class and then carefully balancing the bias and variance components of the risk to achieve a minimum. The details behind such bounds are beyond the scope of this paper, and here we simply state the key results. For more information the reader is referred to [8, 9, 4]. A key assumption behind our results is that the class of trees $\mathcal{T}_{Dy}$ considered is the set of all pruned dyadic trees (i.e., dyadic partitions like those underlying conventional wavelet analysis); they are not grown adaptively to fit the data, which makes certain key analysis steps possible.

Classification Risk Bound: Assume that the Bayes boundary is essentially a $d-1$ dimensional manifold in the original d dimensional feature space; a very reasonable assumption for most practical situations. Technically, we require that the Bayes optimal decision boundary has a box-counting dimension of $d-1$ [4]. Select a classification tree in $\mathcal{T}_{Dy}$ that minimizes the criterion

$$C(T) \;\; = \;\; \widehat{L}_n(T) \; + \; \sqrt{\frac{32(\log(n)+1)}{n}}\,|T|^{1/2}, \tag{3}$$

where n is the number of label training data. Then the risk of the corresponding classifier $\widehat{\phi}$ is bounded according to

$$R(\widehat{\phi}, \phi^*) \;\; \leq \;\; C\left(\frac{\log n}{n}\right)^{1/(d+1)}, \tag{4}$$

where $C > 0$ is a constant. It can also be shown that this upper bound is close to the minimax lower bound for this classification problem [4], which demonstrates that dyadic classification trees cannot be significantly outperformed by other methods (e.g., neural networks, support vector machines, standard CART, etc.) under the stated assumptions.

Regression Risk Bound: For the sake of simplicity, let us assume that the function f is one-dimensional. Extensions to multidimensional function estimation (e.g., images) can be made. Assume that the true function f belongs to a Besov space with smoothness parameter β. This space includes functions that are generally smooth, but may have discontinuities as well; again a very reasonable assumption for most practical problems. A regression tree is based on a pruned dyadic partition, with polynomials of degree $r \geq \beta$ fitted to the data on each cell of the partition. Select such a regression tree from $\mathcal{T}_{Dy}$ according to the criterion

$$C(T) \;\; = \;\; \widehat{L}_n(T) \; + \; \frac{3+r}{2} \log n \, |T|, \tag{5}$$

where n is the number of data. Then the risk of the resulting estimator $\widehat{f}$ is bounded according to

$$R(\widehat{f}, f^*) \;\; \leq \;\; C \left(\frac{\log^2 n}{n} \right)^{2\beta/(2\beta+1)}, \tag{6}$$

where $C > 0$ is a constant. It can also be shown that this upper bound is within a logarithmic factor of the minimax lower bound for this regression problem [9], which demonstrates that dyadic regression trees cannot be significantly outperformed by other methods (e.g., splines, radial basis functions, standard CART, etc.).

7 Conclusions

The work summarized in this paper pointed to three key modifications of the classical CART program.

1. **Different penalties for classification and regression:** The variance component of the risk functions grows differently for classification and regression; like $|T|$ for regression and $|T|^{1/2}$ for classification. This implies that the proper penalization for tree pruning should be modified to account for this distinction. With the appropriate penalties, it is shown that the risks of classification and regression trees converge rapidly to zero, at near minimax-optimal rates in a broad range of cases. Without the proper penalty, CART is well known to "overprune" in classification problems.

2. **Dyadic trees:** CART usually begins by growing a tree (and corresponding partition) to fit the data as closely as possible. This can be a computationally expensive process, and based on our theoretical analysis, as well as our practical experience, appears to be unnecessary. Pruned dyadic trees perform about as well, and sometimes better than, grown and pruned trees. A classification benchmark study [10, 4] demonstrates that the pruning criterion in (3) produces classification trees that perform nearly as well, and sometimes better than, the standard CART-based trees. Moreover, adaptively grown trees are very difficult to analyze, and risk bounds analogous to those presented here for dyadic trees are not currently known. Our use of dyadic partitions for classification here complements the connection made between CART and wavelet-based regression in [3].

3. **Likelihood-based regression:** CART usually employs a sum-of-squared errors criterion for regression. Here we advocate a negative log-likelihood criterion which enables us to devise near-optimal estimators for Gaussian and non-Gaussian data types. The theoretical findings described in this paper

are also supported by empirical evidence in a number of different non-Gaussian regression problems, including applications in astronomical data analysis, density estimation, and medical imaging [8, 9]. These results demonstrate that the theoretical and practical benefits associated with wavelet analysis in Gaussian noise problems can be obtained in a much broader class of problems, through the use of dyadic partitions and piecewise polynomial data fitting.

In addition to the desirable flexibility and theoretical performance characteristics of our tree-based methods, these classifiers and estimators can be constructed very efficiently [8, 9, 4]. The criterion in (1) or (2) can be evaluated for every possible pruning of a initial N leaf tree in $O(N \log N)$ operations. In practice, it is unnecessary to consider initial trees with more leaves than available data n, and thus $N \leq n$. Moreover, our penalties are set according to theoretical bounds, and no tuning or adjustment is required. CART usually employs computationally demanding cross-validation procedures to select a good pruning. Therefore, the overall computational cost of our methods is $O(n)$, which may be much less than that required by traditional CART.

References

[1] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1983.

[2] S. Gey and E. Nedelec. Model selection for cart regression trees. Preprint available at http://www.math.u-psud.fr/ biblio/ppo/2001/ppo2001-56.html.

[3] D. Donoho. Cart and best-ortho-basis selection: A connection. *Annals of Stat.*, 25:1870–1911, 1997.

[4] C. Scott and R. Nowak. Complexity-regularized dyadic classification trees: Efficient pruning and rates of convergence. Technical Report TREE0201, Rice University, 2002.

[5] A. Nobel. Analysis of a complexity based pruning scheme for classification trees. *IEEE Transactions on Information Theory*, 48(8):2362–2368, 2002.

[6] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

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

[8] R. Willett. Multiscale analysis for intensity and density estimation. Master's thesis, Rice University, 2002.

[9] E. Kolaczyk and R. Nowak. Multiscale likelihood analysis and complexity penalized estimation. submitted to Annals of Stat. August, 2001. Available at `http://cmc.rice.edu/docs`.

[10] C. Scott and R. Nowak. Dyadic classification trees via structural risk minimization. In *Neural Information Processing Systems — NIPS 2002*, Dec. 9-14, Vancouver, Canada, 2002.