BINARY TREE-STRUCTURED PARTITION AND CLASSIFICATION SCHEMES

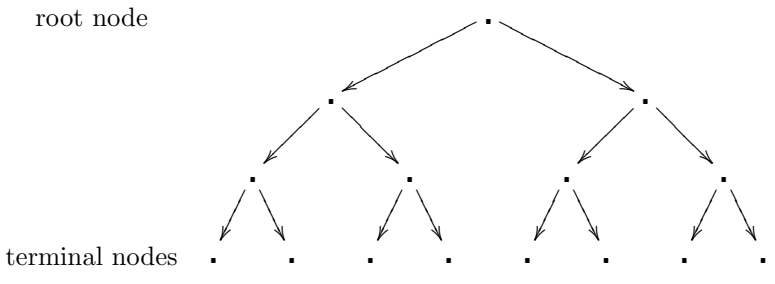
DAVID MCDIARMID

Abstract. Binary tree-structured partition and classification schemes are a class of non-parametric tree-based approaches to classification problems. After developing an original definition of tree-structured partition and classification schemes, we introduce a standard example of such a scheme and apply it to a well-known data set. We then introduce a set of regularity conditions and prove that tree-structured partition and classification schemes are risk consistent when these conditions are satisfied.

Contents

1. Introduction

Tree-structured partition and classification schemes are one of the myriad statistical techniques motivated by a desire to use certain known characteristics of an object to predict some other characteristic of that object. Such schemes address this problem by constructing binary decision trees, which at each node contain a simple rule for determining in which of two child nodes an object with given known characteristics belongs . The end user determines the predicted class of a new object by beginning at a "root node" and moving down the tree until he finds the "terminal node" in which the object belongs. Each object belonging in a given terminal node is predicted to be in the class to which the node is associated.



Date: October 3, 2011.

Binary tree-structured partition and classification schemes, often referred to more simply as classification tree processes, can be applied to any problem to which more common parametric techniques such as logistic or probit regression can be, and classification tree processes have a number of advantages over the latter two techniques. Trees often predict more accurately than parametric methods do, particularly when applied to problems in which the data have non-linear structure. Unlike regression models, tree-structured classifiers do not depend on strong assumptions about the data such as normality and homoscedasticity of the errors. Classification trees are also more easily interpretable than logistic or probit regression functions; non-experts applying the results of statistical analyses often understand tree charts better than they do linear functions of many variables. Finally, under certain regularity conditions, tree structured-classifiers possess a highly desirable property called risk consistency, which means that as the amount of data used to train the tree goes to infinity, the accuracy of the tree in predicting the response characteristic converges to the accuracy with which one could predict the response characteristic if one were to know the explicit conditional distribution of the response characteristic given the predictor variables.

The second section of this paper introduces the terminology needed to rigorously describe binary tree-structured partition and classification schemes. The third section introduces a simplified version of CART[1], the canonical example of a binary tree-structured partition and classification scheme, and the fourth section demonstrates the application of CART to a real-world data set. The fifth section presents a number of theoretical results regarding the risk consistency of partition and classification schemes.

2. Terminology

In order to rigorously describe the theoretical properties and practical performance of classification trees, we first need to introduce a number of definitions. We will start by describing the terminology used to describe the data set which trains a tree. Many of the below definitions are adapted slightly from those used in the 1984 monograph *Classification and Regression Trees* (Breiman, Friedman, Olshen, and Stone), while others are original to this paper.

For a given observation with p predictor variables, we call the tuple containing the values of the predictor variables the **measurement vector**, and denote it $\mathbf{x} = (x_1, ...x_p)$. We call the space of all possible measurement vectors for the p variables the **measurement space**.

If C is the set of all possible values the response variable can take, then for a given observation we call the value of the response variable $c \in C$ the **class** of the observation. For a classification problem, C will always be finite.

The three above definitions allow us to define the collection of observations used to train the tree. A **learning sample** with n observations is a set of pairs of measurement tuples and classes. We write a learning sample as $L = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\}$ where $\mathbf{x}_i = \{x_{i1}, \ldots, x_{ip}\} \in \chi$ and $c_i \in C$.

If $\tilde{T} = \{t_1, \ldots, t_k\}$ is a partition of a measurement space χ, we call the elements of $\tilde{T}$ **nodes**. We let $\mathrm{L}_{t_i} = \{(\mathbf{x}_j, c_j) \in L : \mathbf{x}_j \in t_i\}$ denote the subset of L induced by t_i and let $L_{\tilde{T}}$ denote the partition of L induced by $\tilde{T}$.

A **partition and classification scheme** can be thought of as an operation that uses the information in the learning sample to first partition the measurement space into a set $\tilde{T}$ of nodes and then construct a **partition classification function** on $\tilde{T}$, where a partition classification function on $\tilde{T}$ is a function $d : \tilde{T} \to C$ such that d is constant on every node of

[1]CART as an acronym for Classification and Regression Trees is a registered trademark of the San Diego based data-mining software company Salford Systems

$\tilde{T}$. Formally, if we allow $\mathcal{L}$ to be the space of all learning samples and $\mathcal{D}$ to be the space of all partition classification functions, then a partition and classification scheme is a function $\Phi : \mathcal{L} \to \mathcal{D}$ such that $\Phi(L) = (\psi \circ \phi)(L)$, where ϕ maps L to some induced partition $L_{\tilde{T}}$ and ψ is an **assignment rule** which maps $L_{\tilde{T}}$ to a partition classification function d on the partition $\tilde{T}$.

This paper focuses not on partition and classification schemes in general but on a specific class of such schemes called binary tree-structured partition and classification schemes. Before we define this class of schemes explicitly, we need to introduce a number of important features of tree-structured classifiers.

A **binary split function** is a map s that sends one node to a pair of nodes such that if $s(t) = (s_1(t), s_2(t)) = (t_1, t_2)$, then $t_1, t_2 \neq \varnothing$, $t_1 \cap t_2 = \varnothing$, and $t_1 \cup t_2 = t$. Intuitively, a binary split function s partitions a parent node $t \subseteq \chi$ into a non-empty left child node t_1 and a non-empty right child node t_2.

A **question set** is a finite set $S = \{s_1, \ldots, s_m\}$ of binary split functions. We can think of a question set as the collection of all the potential rules we might use to split the measurement space.

A **goodness of split criterion** is a function g which maps each pair (t, s) consisting of a node $t \subset \chi$ and a binary split function $s \in S$ to a real number. For any parent node t, the goodness of split criterion ranks the split functions in the question set based on some measure of the quality of the child nodes the split would produce. This ranking is typically determined by the size of the reduction in some "impurity function." A definition and examples of impurity functions appear in the next section.

A **stop-splitting rule** is a map r from the power set of the measurement space χ to $\{0, 1\}$. If $r(t) = 0$, then t will be split into two child nodes, but if $r(t) = 1$ then t is a terminal node and will not be split.

Now we are finally able to define the class of processes which are the subject of this paper. A **binary tree-structured partition and classification scheme** is a partition and classification scheme Φ which can be written in the form

$$\Phi(L) = (\psi \circ \lim_{i \to \infty} \phi^{(i)})(L), \tag{2.1}$$

where ψ is an assignment rule and

$$\phi^{(i)}(L) = L_{(\phi_i \circ \phi_{i-1} \circ \cdots \circ \phi_1)(\chi)} \tag{2.2}$$

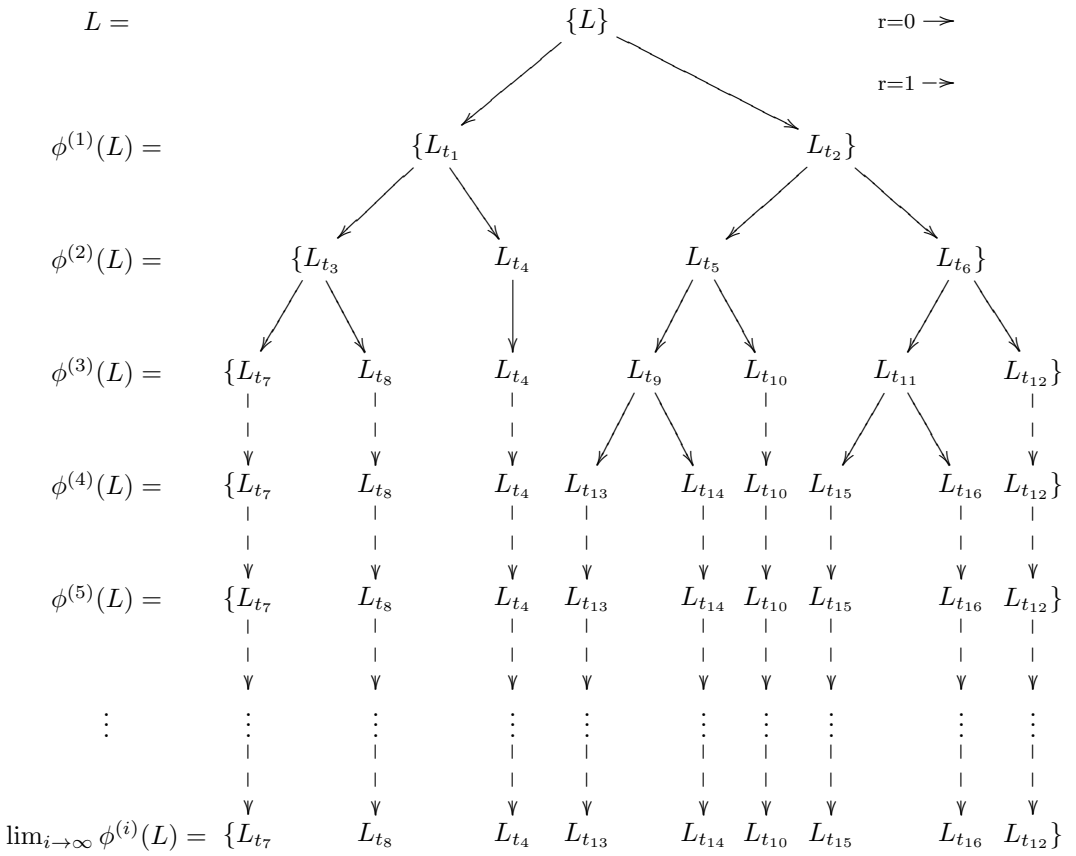
and for any partition $\tilde{T}$ of χ, including χ itself, $\phi_i(\tilde{T}) = \{t' : t' \text{ is in the pair } \nu_i(t) \text{ for some } t \in \tilde{T}\}$ where

$$\nu_i(t) = \begin{cases} (\arg\max_{s \in S} g(L_t, s))(t) & \text{if } r = 0 \\ (t, t) & \text{if } r = 1. \end{cases} \tag{2.3}$$

Remark 2.4. If $g(L_t, s_i) = g(L_t, s_j) = \max_{s \in S} g(L_t, s)$ and $i < j$, then s_i is used.

A binary tree-structured partition and classification scheme Φ is therefore defined as an assignment rule applied to the limit of a sequence of induced partitions $\phi^{(i)}(L)$, where $\phi^{(i)}(L)$ is the partition of the learning sample L induced by the partition $(\phi_i \circ \phi_{i-1} \circ \cdots \circ \phi_1)(\chi)$. For every node t in a partition $\tilde{T}$ such that the stopping rule $r(t) = 0$, the function $\phi(\tilde{T})$ splits each node into two child nodes using the best binary split in the question set as determined by the goodness of split criterion. For every node $t \in \tilde{T}$ such that the stopping rule $r(t) = 1$, $\phi(\tilde{T})$ leaves t unchanged.

The process defined above is binary in the sense that each application of the function ϕ splits each node in a partition into two or fewer child nodes. It is tree-structured in the sense that the sequence $\phi^{(i)}$ can easily be envisioned as an expanding tree, as illustrated below.



It remains to be shown of, of course, that binary tree-structured partition and classification schemes are well-defined, which is true only if there always exists some induced partition L' such that

$$\lim_{i\to\infty} \phi^{(i)}(L) = L'. \tag{2.5}$$

In fact a stronger result is true.

Proposition 2.6. *If L is a learning sample and $\phi^{(i)}$ is as above, then there exists some $N \in \mathbb{N}$ such that*

$$n \geq N \implies \phi^{(n)}(L) = \lim_{i\to\infty} \phi^{(i)}(L) \tag{2.7}$$

Proof. Let $\{\mathrm{L}_{\tilde{T}_i}\}$ denote the sequence $\{L, \phi^{(1)}(L), \phi^{(2)}(L), \dots\}$ Define $t_i^{max} = \max\{\ t \in \tilde{T}_i : r(t) = 0\}$ as the size of the largest non-terminal node (or nodes) in $\tilde{T}_i$. Notice that if there

exists $N \in \mathbb{N}$ such that t_N^{max} does not exist, then every node in $\tilde{T}_N$ is terminal, which means that for all $n > N$, $\tilde{T}_n = \tilde{T}_N$, in which case (2.7) holds. Also notice that the sequence $\{|t_i^{max}|\}$ is strictly decreasing for as long as it exists, and further that if t_{i+1}^{max} exists then $|t_{i+1}^{max}| \leq |t_i^{max}| - 1$. But since 1 is a lower bound for $\{|t_i^{max}|\}$ and $|t_1^{max}| = |L|$, this means that $t_{|L|}^{max}$ cannot exist, so (2.7) always holds with $N \leq |L|$.

□

3. The CART Process

The canonical example of a binary tree-structured partition and classification scheme is the Classification and Regression Trees (CART) process outlined by Breiman et al. in 1984. Many if not most of the tree-structured classification algorithms available today are variations on the CART process. For the sake of brevity, we consider a somewhat simplified version of CART.

For a learning sample $L = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\}$ where $\mathbf{x}_i = \{x_{i1}, \ldots, x_{ip}\} \in \chi$ and $c_i \in C$, CART's question set S_C is the set $\{s_{ij} : 1 \leq i \leq n, 1 \leq j \leq p\}$ such that if $t \subseteq \chi$ is a node, then $s_{ij}(t) = (t_1, t_2)$ where

$$t_1 = \{\mathbf{x}' = (x'_1, \ldots, x'_p) \in \chi : x'_j < x_{ij}\} \tag{3.1}$$

and

$$t_2 = t \backslash t_1. \tag{3.2}$$

Thus S_C is the set of all univariate splits of the measurement space χ which induce distinct splits of the learning sample L.

CART's goodness of split criterion is

$$g_C(L_t, s) = i(L_t) - \frac{|L_{s_1(t)}|}{|L_t|} i(L_{s_1(t)}) - \frac{|L_{s_2(t)}|}{|L_t|} i(L_{s_2(t)}) \tag{3.3}$$

where i is some **impurity function**. This criterion assesses the quality of a split s by subtracting the average impurity of the child nodes t_1, t_2 from the impurity of the parent node t.

An impurity function is a map i from a subset of L induced by a node t to a real number where for $L_t \subseteq L$ such that $L_t \neq \varnothing$, the following conditions hold: (a) $i(L_t)$ achieves its maximum only when L_t contains all classes $c \in C$ in equal proportions, (b) $i(L_t)$ achieves its minimum only when L_t contains only one class, and (c) $i(L_t)$ is symmetric, meaning that if $f : C \to C$ is a bijection and $L'_t = \{(\mathbf{x}_i, f(c_i)) : (\mathbf{x}_i, c_i) \in L_t\}$, then $i(L_t) = i(L'_t)$. Such functions provide a reasonable measure of the uniformity of the classes in the points $\{(\mathbf{x}, c) : \mathbf{x} \in L_t\}$.

The most commonly used impurity function for CART is the **Gini Index**. If we let $L_{t,c} = \{(\mathbf{x}_i, c_i) \in L_t : c_i = c\}$, then the Gini Index is defined as

$$i_{gini}(t) = \sum_{c \neq c'} \frac{|L_{t,c}|}{|L_t|} \frac{|L_{t,c'}|}{|L_t|}. \tag{3.4}$$

The reader can very easily check that the Gini Index is in fact an impurity function. Another common impurity function is **information entropy**,

$$i_{ent}(t) = -\sum_{c \in C} \frac{|L_{t,c}|}{|L_t|} \log_b\left(\frac{|L_{t,c}|}{|L_t|}\right) \tag{3.5}$$

where $b > 0$.

The stop-splitting rule r for our simplified version of CART is

$$r(t) = \begin{cases} 1 & \text{if } c = c' \text{ for all } (\mathbf{x}, c), (\mathbf{x}', c') \in L_t \\ 1 & \text{if } \mathbf{x} = \mathbf{x}' \text{ for all } (\mathbf{x}, c), (\mathbf{x}', c') \in L_t \\ 1 & \text{if } |L_t| < N_{stop} \text{ for some fixed } N_{stop} \in \mathbb{N} \\ 1 & \text{if } \max_{s \in S} g(L_t, s) < \alpha_{stop} \text{ for some fixed } \alpha_{stop} \in \mathbb{R} \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

So CART continues to split a node until either all the points in the induced subset L_t have the same class, all the points in L_t have the same measurement vectors, the number of observations in L_t is less than some pre-defined number N, or the goodness of the best possible split is below some pre-defined threshold α.

The most basic reasonable assignment rule ψ is the **plurality rule** $\psi_{pl}(L_{\tilde{T}}) = d$ such that if $\mathbf{x} \in t$, then

$$d(\mathbf{x}) = \arg\max_{c \in C} |L_{c,t}|. \tag{3.7}$$

The plurality rule classifies each new point in t as belonging to the class most common in L_t. This rule will be important in the risk consistency discussion in section 5.

Information about more advanced variations of CART is abundant in the literature. In particular, multivariate question sets as well as more complex assignment rules are discussed in Lugosi and Nobel (1996), Nobel (2002), and of course Breiman et al.

4. Iris Classification

We now demonstrate a practical use of a binary tree-structured partition and classification scheme by applying the simplified CART process to the famous Iris Flower data set first introduced by Fisher in 1936 and now widely used to test pattern recognition. This data set is available online through the UC Irvine Machine Learning Repository.

The Iris Flower data is a collection of observations regarding 150 iris flowers drawn in equal proportion from three distinct species: *Iris Setosa*, *Iris Versicolour*, and *Iris Virginica*. Each observation contains measurements in centimeters of four characteristics of the flower as well as the flower's species. We can frame the Iris Flower data set as a learning sample $L^{Iris} = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_{150}, c_{150})\}$ where $\mathbf{x}_i = \{x_{i1}, \ldots, x_{i4}\} \in \chi$. The measurement variables are sepal length (x_1), sepal width (x_2), petal length (x_3), and petal width (x_4). The set of classes is $C = \{$*Iris Setosa*, *Iris Versicolour*, *Iris Virginica*$\}$.

Our goal in applying a partition and classification scheme to the Iris Flower data set is to develop a tree that accurately predicts the species of an iris flower based on the length and width of its sepal and petals. We use the simplified CART scheme Φ_C presented above and let our minimum number of observations in a node to be split be $N_{stop} = 10$ and our minimum goodness of split be $\alpha_{stop} = .025$. We find by analyzing the data using **Python** that the best initial split is $s^{(1)}(\chi) = (t_a, t_b)$ where $t_a = \{\mathbf{x} \in t : x_4 < .95\}$ and $t_b = t \backslash t_a$. The goodness of the first split is $g(L^{Iris}, s^{(1)}) = .1667$.

We let t_2 denote $s_1^{(1)}(\chi)$ and t_2 denote $s_2^{(1)}(\chi)$. Then the induced node $L_{t_2}^{Iris}$ is pure, consisting entirely of the *Setosa* class, so we split that node no further. The best split for node t_3 is $s^{(3)}(t) = (t_a, t_b)$ such that $t_a = \{\mathbf{x} \in t : x_4 < 1.75\}$ and $t_b = t \backslash t_a$. This split has goodness $g(L_{t_3}^{Iris}, s^{(3)}) = .1948$.

Letting $t_4 = s_1^{(3)}(t_3)$ and $t_5 = s_2^{(3)}(t_3)$, we find that for the induced node $L_{t_5}^{Iris}$, the maximum goodness of split is

$$\max_{s \in S} g(L_{t_5}^{Iris}, s) = .0068 < \alpha_{stop}, \tag{4.1}$$

so we split the node t_5 no further. However, we split the node t_4 one more time, using the split $s^{(4)}(t) = (t_a, t_b)$ such that $t_a = \{\mathbf{x} \in t : x_3 < 4.95\}$ and $t_b = t \backslash t_a$. The goodness of this split is $g(L_{t_4}^{Iris}, s^{(4)}) = .0412$.

We let $t_6 = s_1^{(4)}(t_4)$ and $t_7 = s_2^{(4)}(t_4)$. Then the maximum goodness of split for the induced node $L_{t_6}^{Iris}$ is

$$\max_{s \in S} g(L_{t_6}^{Iris}, s) = .0204 < \alpha_{stop}, \tag{4.2}$$

and the number of observations in the induced node $L_{t_7}^{Iris}$ is $6 < N_{stop}$, so we are finished splitting the space.

In terms of the definition of binary-tree structured classification schemes developed in section 2, we have found that

$$\phi^{(1)}(L^{Iris}) = L_{\phi_1(\chi)}^{Iris} = L_{\{t_2, t_3\}}^{Iris}, \tag{4.3}$$

$$\phi^{(2)}(L^{Iris}) = L_{(\phi_2 \circ \phi_1)(\chi)}^{Iris} = L_{\{t_2, t_4, t_5\}}^{Iris}, \tag{4.4}$$

$$\lim_{n \to \infty} \phi^{(n)}(L^{Iris}) = \phi^{(3)}(L^{Iris}) = L_{(\phi_3 \circ \phi_2 \circ \phi_1)(\chi)}^{Iris} = L_{\{t_2, t_5, t_6, t_7\}}^{Iris}, \tag{4.5}$$

and

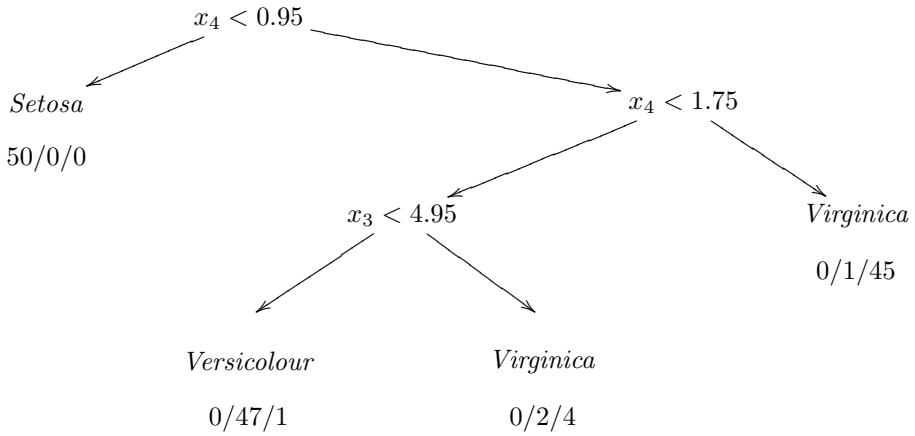
$$\Phi_C(L^{Iris}) = (\nu \circ \lim_{n \to \infty} \phi^{(n)})(L^{Iris}) = d_{C,Iris} \tag{4.6}$$

where $d_{C,Iris}$ is the partition classification rule defined by

$$d_{C,Iris}(\mathbf{x}) = \begin{cases} \textit{Setosa} & \text{if } x_4 < .95 \\ \textit{Virginica} & \text{if } x_4 \geq 1.75 \\ \textit{Virginica} & \text{if } x_3 \geq 4.95 \text{ and } .95 \leq x_4 < 1.75 \\ \textit{Versicolour} & \text{if } x_3 < 4.95 \text{ and } .95 \leq x_4 < 1.75. \end{cases} \tag{4.7}$$

The partition classification rule $d_{C,Iris}$ correctly classifies 146 out of the 150 observations in the learning sample.

One can also display $d_{C,Iris}$ in tree form, as shown below. The inequality at each non-terminal node is the rule which dictates which observations belong in each of the node's two children. Observations satisfying the rule are assigned to the left child node, while those which do not satisfy the rule are assigned to the right child node. The labels in the terminal nodes are the classes to which these nodes correspond, and the numbers below the labels show the number of learning sample observations of each class which $d_{C,Iris}$ assigned to that node.



The above tree allows even those without a strong understanding of mathematics or statistics to classify the flowers in the data set quite accurately based on only two easily observable measurements. A wealth of other examples of applications of CART and its variants can be found in Breiman et al.

5. Risk Consistency

We now turn to a discussion of risk consistency, an important statistical property of partition and classification schemes which we will define momentarily. Throughout the remainder of this paper, consider the learning sample L to be a collection of n independent and identically distributed random vectors $(\mathbf{X}_i, Y_i)$ such that $\mathbf{X}_i$ takes values in the measurement space χ and Y_i takes values in the class set C.

For a partition classification function d, we say that the **misclassification rate** of d for a random vector $(\mathbf{X}, Y)$ is

$$R_{d,(\mathbf{X},Y)} = \sum_{c \in C} \mathbb{P}(Y \neq c \mid d(\mathbf{X}) = c) \cdot \mathbb{P}(d(\mathbf{X}) = c). \tag{5.1}$$

We say that d is a **Bayes rule** if

$$R_{d,(\mathbf{X},Y)} = \min_{d' \in \mathcal{D}} R_{d',(\mathbf{X},Y)}. \tag{5.2}$$

We denote the Bayes rule $d^{\star}_{(\mathbf{X},Y)}$ and the misclassification rate of the Bayes rule $R^{\star}_{(\mathbf{X},Y)}$. If we let

$$P_{c,(\mathbf{X},Y)}(\mathbf{x}) = \mathbb{P}(Y = c \mid \mathbf{X} = \mathbf{x}), \tag{5.3}$$

then we can also define the Bayes rule explicitly as

$$d^{\star}_{(\mathbf{X},Y)} = \arg\max_{c \in C} P_{c,(\mathbf{X},Y)}(\mathbf{x}). \tag{5.4}$$

Intuitively, the Bayes rule is the most accurate possible predictor of the class given the measurement variables, the rule one would use if one were to know explicitly the conditional distribution of $Y|\mathbf{X}$. The reader should note that while the Bayes rule is a partition classification function for infinitely many partitions of the measurement space χ, it is usually not a partition classification function for all partitions of χ because it will not be constant on all the nodes of some partitions.

Let $L_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a learning sample composed of the first n outcomes of the random vector $(\mathbf{X}, Y)$. Then we say a partition and classification scheme Φ is **risk consistent** if as $n \to \infty$,

$$\Phi(L_n) \to d^{\star}_{(\mathbf{X},Y)}(\mathbf{x}) \textit{ with probability } 1. \tag{5.5}$$

In words, a partition and classification scheme is risk consistent if the partition classification function to which it maps the learning sample L_n converges to the Bayes rule as the number of observations n in the learning sample goes to infinity.

The remainder of this section is dedicated to stating and proving the conditions under which partition and classification schemes in general and binary-tree structured partition and classification schemes in particular are risk consistent. For the most part, we follow the discussion in Lugosi and Nobel, though we use very different terminology and notation.

Before we arrive at our two main results, we need to introduce a bit of notation as well as some definitions and technical lemmas. For any random variable X and set A, let $F_X(A)$ denote $\mathbb{P}(X \in A)$, and let

$$\hat{F}_{n,X}(A) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in A) \tag{5.6}$$

be the empirical probability that $X \in A$ based on n observations. I here is an **indicator function**, meaning

$$I(\cdot) = \begin{cases} 1 & \text{if } \cdot \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

Now, let $\mathcal{T} = \{\tilde{T}_1, \tilde{T}_2, \ldots\}$ be a (possibly infinite) collection of partitions of a measurement space χ. We define the **maximal node count** of $\mathcal{T}$ as the maximum number of nodes in any partition $\tilde{T}$ in $\mathcal{T}$. We write the maximal node count as

$$\lambda(\mathcal{T}) = \sup_{\tilde{T} \in \mathcal{T}} |\tilde{T}| \tag{5.8}$$

We let $\Delta(\mathcal{T}, L_n) = |\{L_{\tilde{T}} : \tilde{T} \in \mathcal{T}\}|$ be the number of distinct partitions of a learning sample of size n induced by by partitions in $\mathcal{T}$. We can then define the **growth function** of $\mathcal{T}$ as

$$\Delta_n(\mathcal{T}) = \sup_{\{L : |L| = n\}} \Delta(\mathcal{T}, L) \tag{5.9}$$

Thus the growth function of $\mathcal{T}$ is the maximum number of distinct partitions $L_{\tilde{T}}$ which partitions $\tilde{T}$ in $\mathcal{T}$ can induce in any learning sample with n observations.

We now introduce a more general technical concept call the **shatter coefficient**. For any class $\mathcal{A}$ of subsets of $\mathbb{R}^p$, the shatter coefficient

$$\mathcal{S}_n(\mathcal{A}) = \max_{\{B \subset \mathbb{R}^p : |B| = n\}} |\{A \cap B : A \in \mathcal{A}\}| \tag{5.10}$$

is the maximum number of partitions of B induced by $\mathcal{A}$ where B is some n point subset of $\mathbb{R}^p$.

We take as given the following inequality involving shatter coefficients proven by Vapnik and Chervonekis in 1971.

Lemma 5.11. *Suppose $\mathcal{A}$ is a class of subsets of $\mathbb{R}^p$ and $\mathbf{X}$ is a random vector taking values in $\mathbb{R}^p$. Let $\mathcal{S}_n$ be as above. Then for any $n \geq 1$ and $\epsilon > 0$,*

$$\mathbb{P}\{\sup_{A\in\mathcal{A}} |\hat{F}_{n,\mathbf{X}}(A) - F_{\mathbf{X}}(A)| > \epsilon\} \leq 4\mathcal{S}_{2n}(\mathcal{A}) \exp(-n\epsilon^2/8) \tag{5.12}$$

We use this inequality to prove another inequality more directly applicable to partition and classification schemes.

Lemma 5.13. *Suppose $\mathcal{T}$ is a collection of partitions of the measurement space $\mathbb{R}^p$ and $\mathbf{X}$ is a random vector taking values in $\mathbb{R}^p$. Then for every $n \geq 1$ and $\epsilon > 0$,*

$$\mathbb{P}\{\sup_{\tilde{T}\in\mathcal{T}} \sum_{t\in\tilde{T}} |\hat{F}_{n,\mathbf{X}}(t) - F_{\mathbf{X}}(t)| > \epsilon\} \leq 4\Delta_{2n}(\mathcal{T}) 2^{\lambda(\mathcal{T})} \exp(-n\epsilon^2/32) \tag{5.14}$$

Proof. For every partition $\tilde{T} \in \mathcal{T}$, let

$$\mathcal{B}(\tilde{T}) = \{\bigcup_{t\in T} t : T \subseteq \tilde{T}\} \tag{5.15}$$

be the collection of the unions of the nodes in any subset T of $\tilde{T}$. Let $\mathcal{B}(\mathcal{T}) = \{t \in \mathcal{B}(\tilde{T}) : \tilde{T} \in \mathcal{T}\}$ For a given partition $\tilde{T}$, let

$$\ddot{t}_{n,\tilde{T}} = \bigcup_{\{t\in\tilde{T}: F_{n,\mathbf{X}}(t) \geq F_{\mathbf{X}}(t)\}} t \tag{5.16}$$

Then for any partition $\tilde{T}$,

$$\begin{aligned} \sum_{t\in\tilde{T}} |F_{n,\mathbf{X}}(t) - F_{\mathbf{X}}(t)| &= 2(F_{n,\mathbf{X}}(\ddot{t}_{n,\tilde{T}}) - F_{\mathbf{X}}(\ddot{t}_{n,\tilde{T}})) \\ &\leq 2 \sup_{t\in\mathcal{B}(\tilde{T})} |F_{n,\mathbf{X}}(t) - F_{\mathbf{X}}(t)|, \end{aligned} \tag{5.17}$$

and therefore

$$\begin{aligned} \sup_{\tilde{T}\in\mathcal{T}} \sum_{t\in\tilde{T}} |F_{n,\mathbf{X}}(t) - F_{\mathbf{X}}(t)| &\leq 2 \sup_{\tilde{T}\in\mathcal{T}} \sup_{t\in\mathcal{B}(\tilde{T})} |F_{n,\mathbf{X}}(t) - F_{\mathbf{X}}(t)| \\ &= 2 \sup_{t\in\mathcal{B}(\tilde{T})} |F_{n,\mathbf{X}}(t) - F_{\mathbf{X}}(t)| \end{aligned} \tag{5.18}$$

Then Lemma 5.11 gives us

$$\begin{aligned} \mathbb{P}\{\sup_{\tilde{T}\in\mathcal{T}} \sum_{t\in\tilde{T}} |F_{n,\mathbf{X}}(t) - F_{\mathbf{X}}(t)| > \epsilon\} &\leq \mathbb{P}\{\sup_{\tilde{T}\in\mathcal{B}(\mathcal{T})} |F_{n,\mathbf{X}}(t) - F_{\mathbf{X}}(t)| > \epsilon/2\} \\ &\leq 4\Delta_{2n}(\mathcal{T}) 2^{\lambda(\mathcal{T})} \exp(-n\epsilon^2/32). \end{aligned} \tag{5.19}$$

because $\mathcal{S}_{2n}(\mathcal{B}(\mathcal{T})) \leq 2^{\lambda(\mathcal{T})}\Delta_{2n}(\mathcal{T})$. □

Our interest in Lemma 5.13 is due to following corollary of it, which can be proven using the Borel-Cantelli lemma.

Corollary 5.20. *Let $\mathbf{X}$ be a random vector taking values in $\mathbb{R}^p$ and let $\{\mathcal{T}_1, \mathcal{T}_2, \dots\}$ be a sequence of families of partitions of $\mathbb{R}^p$. Suppose that as $n \to \infty$, (a) $n^{-1}\lambda(\mathcal{T}_n) \to 0$, and (b) $n^{-1}log(\Delta_n(\mathcal{T}_n)) \to 0$. Then*

$$\sup_{\tilde{T}\in\mathcal{T}_n} \sum_{t\in\tilde{T}} |F_{n,\mathbf{X}}(t) - F_{\mathbf{X}}(t)| \to 0 \text{ with probability } 1. \tag{5.21}$$

Before stating our first major result, we require one more inequality, proven by Devroye and Gyorfi in 1985.

Lemma 5.22. *Let $(\mathbf{X}, Y)$ be a random vector which takes values in $\mathbb{R}^p \times C$, let $|C| = M$, and let $\beta_1(\mathbf{x}), \ldots, \beta_M(\mathbf{x})$ be real-valued functions on $\mathbb{R}^p$. Define the partition classification function*

$$h(\mathbf{x}) = \arg \max_{1 \leq k \leq M} \beta_k(\mathbf{x}). \tag{5.23}$$

Then

$$R_{h,(\mathbf{X},Y)} - R^{\star}_{(\mathbf{X},Y)} \leq \sum_{k=1}^{M} \int |P_{k,(\mathbf{X},Y)}(\mathbf{x}) - \beta_k(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}). \tag{5.24}$$

Now we are ready to present a set of regularity conditions under which any partition and classification scheme Φ is risk consistent. For a partition $\tilde{T}$ of a measurement space $\chi \ni \mathbf{x}$, let $\tilde{T}[\mathbf{x}] = \{t \in \tilde{T} : \mathbf{x} \in t\}$ be the node t in $\tilde{T}$ which contains $\mathbf{x}$. For a set $A \subseteq \mathbb{R}^p$, let

$$D(A) = \sup_{x,y \in A} \|x - y\| \tag{5.25}$$

be the diameter of A. Let a **hyperplane split** be any binary split function on $\mathbb{R}^p$ that splits a node by dividing it into child nodes consisting of either side of some hyperplane. Then the follow result holds.

Theorem 5.26. *Let $(\mathbf{X}, Y)$ be a random vector taking values in $\mathbb{R}^p \times C$ and let L_n be the set of the first n outcomes of $(\mathbf{X}, Y)$. Suppose that Φ is a partition and classification scheme such that $\Phi(L_n) = (\psi_{pl} \circ \phi)(L_n)$, where ψ_{pl} is the plurality rule and $\phi(L_n) = (L_n)_{\tilde{T}_n}$ for some $\tilde{T}_n \in \mathcal{T}_n$, where*

$$\mathcal{T}_n = \{\phi(l_n) : \mathbb{P}(L_n = l_n) > 0\}. \tag{5.27}$$

Also suppose that all the binary split functions in the question set associated with Φ are hyperplane splits. If as $n \to \infty$, (a) $n^{-1}\lambda(\mathcal{T}_n) \to 0$, (b) $n^{-1}\log(\Delta_n(\mathcal{T}_n)) \to 0$, and (c) for every $\gamma > 0$ and $\delta \in (0, 1)$,

$$\inf_{\{S \subseteq \mathbb{R}^p : F_{\mathbf{X}}(S) \geq 1-\delta\}} F_{\mathbf{X}}(\{\mathbf{x} : D(\tilde{T}_n[\mathbf{x}] \cap S) > \gamma\}) \to 0 \text{ with probability 1}, \tag{5.28}$$

then Φ is risk consistent.

Remark 5.29. We refer to condition (c) of Theorem 5.26 as the **shrinking cell condition**.

Proof. We provide an outline of the proof. The full (and quite long) argument can be found in Lugosi and Nobel (1984).

Let $M = |C|$, and for each $k = 1, \ldots, M$ and $\mathbf{x} \in \mathbb{R}^p$, let $m_k(\mathbf{x}) = P_{k,(\mathbf{X},Y)}(\mathbf{x})$. Let

$$m_{k,n}(\mathbf{x}) = \frac{|L_{k,\tilde{T}_n[\mathbf{x}]}|}{F_{\mathbf{X}}(\tilde{T}_n[\mathbf{x}])}. \tag{5.30}$$

By Lemma 5.22, it suffices to show that as $n \to \infty$,

$$\int |m_k(\mathbf{x}) - m_{k,n}(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) \to 0 \tag{5.31}$$

for each k. Let $\epsilon > 0$ and $r_k : \mathbb{R}^p \to \mathbb{R}$ be some continuous function with compact support and the property that

$$\int |m_k(\mathbf{x}) - r_k(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) < \epsilon. \tag{5.32}$$

Define the auxiliary functions $u_{k,n}$ and $v_{k,n}$ as

$$u_{k,n}(\mathbf{x}) = \frac{E[I\{Y=k\}I\{\mathbf{X} \in \tilde{T}_n[\mathbf{x}]\} \mid L_n]}{F_{\mathbf{X}}(\tilde{T}_n[\mathbf{x}])} \tag{5.33}$$

$$v_{k,n}(\mathbf{x}) = \frac{E[r_k(\mathbf{X})I\{\mathbf{X} \in \tilde{T}_n[\mathbf{x}]\} \mid L_n]}{F_{\mathbf{X}}(\tilde{T}_n[\mathbf{x}])}. \tag{5.34}$$

Notice that

$$\begin{aligned} |m_k(\mathbf{x}) - m_{k,n}(\mathbf{x})| \quad \leq \quad & |m_k(\mathbf{x}) - r_k(\mathbf{x})| + |r_k(\mathbf{x}) - v_{k,n}(\mathbf{x})| \\ & + |v_{k,n}(\mathbf{x}) - u_{k,n}(\mathbf{x})| + |u_{k,n}(\mathbf{x}) - m_{k,n}(\mathbf{x})|, \end{aligned} \tag{5.35}$$

which means

$$\begin{aligned} \int |m_k(\mathbf{x}) - m_{k,n}(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) \quad \leq \quad & \int |m_k(\mathbf{x}) - r_k(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) + \int |r_k(\mathbf{x}) - v_{k,n}(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) \\ & + \int |v_{k,n}(\mathbf{x}) - u_{k,n}(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) \\ & + \int |u_{k,n}(\mathbf{x}) - m_{k,n}(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}), \end{aligned} \tag{5.36}$$

We will now consider separately each of the integrals on the right hand side of (5.36).The first integral is less than ϵ due to (5.32). For the second integral, if we let $K \leq \infty$ be a uniform upper bound for $|r_k|$, it can be shown using Fubini's theorem and shrinking cell condition that

$$\lim_{n \to \infty} \sup \int |r_k(\mathbf{x}) - v_{k,n}(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) \leq \delta(4K+1). \tag{5.37}$$

For the third integral we have

$$\begin{aligned} \int |v_{k,n}(\mathbf{x}) - u_{k,n}(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) \quad &= \quad \sum_{t \in \tilde{T}_n} \left| \int_t m_k(\mathbf{x}) F_{\mathbf{X}}(d\mathbf{x}) - \int_t r_k(\mathbf{x}) F_{\mathbf{X}}(d\mathbf{x}) \right| \\ &\leq \quad \int |m_k(\mathbf{x}) - r_k(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) < \epsilon. \end{aligned} \tag{5.38}$$

Finally, it can be shown using Corollary 5.20 and conditions (a) and (b) that

$$\lim_{n \to \infty} \int |u_{k,n}(\mathbf{x}) - m_{k,n}(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) = 0. \tag{5.39}$$

Together, these bounds and (5.36) give us that

$$\lim_{n \to \infty} \sup \int |m_k(\mathbf{x}) - m_{k,n}(\mathbf{x})| F_{\mathbf{X}}(d\mathbf{x}) \leq 2\epsilon + \delta(4K+1). \tag{5.40}$$

Since δ and ϵ can be arbitrarily close to 0, this proves the theorem.

□

If instead of considering partition and classification schemes in general, we insist that our schemes be binary tree-structured, we can produce a similar result with conditions (a) and (b) of Theorem 5.26 replaced by a single, simpler condition.

Theorem 5.41. *Suppose* $(\mathbf{X}, Y)$ *is a random vector taking values in* $\mathbb{R}^p \times C$ *and* L_n *is the set of the first* n *outcomes of* $(\mathbf{X}, Y)$*. Let* Φ *be a binary tree-structured partition and classification scheme such that*

$$\Phi(L_n) = (\psi_{pl} \circ \lim_{i\to\infty} \phi^{(i)})(L_n) \tag{5.42}$$

where ψ_{pl} *is the plurality rule and*

$$\lim_{i\to\infty} \phi^{(i)}(L_n) = \tilde{T}_n \tag{5.43}$$

for some $\tilde{T}_n \in \mathcal{T}_n$ *where*

$$\mathcal{T}_n = \{\lim_{i\to\infty} \phi^{(i)}(l_n) : \mathbb{P}(L_n = l_n) > 0\}. \tag{5.44}$$

Also suppose that all the binary split functions in the question set associated with Φ *are hyperplane splits. If as* $n \to \infty$*, the shrinking cell condition of Theorem 5.26 is satisfied and for every* n *and* $t \in \tilde{T}_n$*, the induced subset* $(L_n)_t$ *has cardinality at least* k_n *where*

$$\frac{k_n}{\log(n)} \to \infty, \tag{5.45}$$

then Φ *is risk consistent.*

Proof. Since condition (c) of Theorem 5.26 is assumed, it suffices to show that conditions (a) and (b) are satisfied. Because $|t| \geq k_n$ for all $t \in \tilde{T}_n$ we have

$$|\tilde{T}_n| \leq \frac{n}{k_n} \tag{5.46}$$

for every $\tilde{T}_n \in \mathcal{T}_n$, in which case

$$\frac{\lambda(\mathcal{T}_n)}{n} \leq \frac{1}{k_n}. \tag{5.47}$$

And $\dfrac{1}{k_n} \to 0$ as $n \to \infty$, so condition (a) is satisfied. Now, notice that by (2.2) and (2.3), every $\tilde{T}_n \in \mathcal{T}_n$ can be constructed by splitting $\mathbb{R}^p$ using no more than $|\tilde{T}_n|$ hyperplanes. Also notice that any hyperplane split of $\mathbb{R}^p$ can divide n points in $\mathbb{R}^p$ in at most n^p ways. In conjunction with (5.46), this shows that $\Delta_n(\mathcal{T}_n) \leq (n^P)^{n/k_n}$, which means

$$\frac{\log(\Delta_n(\mathcal{T}_n))}{n} \leq p\frac{\log(n)}{k_n}. \tag{5.48}$$

The right hand side of (5.48) goes to 0 as $n \to \infty$ by assumption, so condition (b) of Theorem 5.26 is also satisfied. □

Acknowledgments. It is a pleasure to thank my mentor, Marcelo Alvisio, for his many helpful suggestions regarding the composition of this paper and for his invaluable assistance with the research that went into it. I would also like to thank Peter May, Paul Sally, and all the faculty who made this summer REU possible.

References

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

[2] Chervonekis, A. & Vapnik, V. N. (1971). On the Uniform Convergence of Relative Frequencies of Events to their Probabilities. *Theory of Probability and Its Applications.* 16, 264-280.

[3] Devroye, L. & Gyorfi, L. (1985). Distribution-Free Exponential Bound on the L_1 Error of Partitioning Estimates of a Regression Function. *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, 67-76.

[4] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7(2), 179-188.
[5] Lugosi,G. & Nobel, A. (1996). Consistency of Data-Driven Histogram Methods for Density Estimation and Classification. *The Annuals of Statistics*, 24(2), 687-706.
[6] Nobel, A. (2002). Analysis of a Complexity Based Pruning Scheme for Classification Trees. *IEEE Transactions on Information Theory*, 48, 2362-2368.
[7] University of California at Irvine Machine Learning Repository. (2011). *Iris Flower data set*. Available at http://archive.ics.uci.edu/ml/.