Machine Learning, 24, 41–47 (1996)

Technical Note: Some Properties of Splitting Criteria

LEO BREIMAN leo@stat.berkeley.edu
Statistics Department, University of California, Berkeley, CA 94720

Editor: Paul Utgoff

Abstract. Various criteria have been proposed for deciding which split is best at a given node of a binary classification tree. Consider the question: given a goodness-of-split criterion and the class populations of the instances at a node, what distribution of the instances between the two children nodes maximizes the goodness-of-split criterion? The answers reveal an interesting distinction between the gini and entropy criterion.

Keywords: Trees, Classification, Splits

1. Introduction

There are different splitting criteria in use for growing binary decision trees. The CART program offers the choice of the gini or twoing criteria. Many other programs use the entropy criterion. Recently Fayyad (1991) and Fayyad and Irani (1990, 1992, 1993) proposed other criteria, which give improved accuracy on a number of data sets. Taylor and Silverman (1993) also explore alternative criteria, and Buntine and Niblet (1992) compare various splitting rules.

To be more specific, suppose that a class of splits $\{s\}$ is defined on the data in a node t. A "goodness-of-split" function $\theta(s,t)$ is defined and the best split taken as the maximizer of $\theta(s,t)$. Let there be J classes numbered $1,\ldots,J$, and denote the proportions of the classes in t by $\mathbf{p} = p_1,\ldots,p_J$. If s sends a proportion P_L of the t population left and $P_R = 1 - P_L$ right, then assume

$$\theta(s,t) = f(P_L, P_R, \mathbf{p}_L, \mathbf{p}_R)$$

where $\mathbf{p}_L = (p_{1,L},\ldots,p_{J,L})$ is the proportion of the J classes in the left node t_L and similarly for $\mathbf{p}_R$.

Equivalently, for every split s, there are numbers α_j, $0 \le \alpha_j \le 1$, and $\beta_j = 1 - \alpha_j$ such that $P_L = \sum_j \alpha_j p_j$, $P_R = \sum \beta_j p_j$, $p_{j,L} = \alpha_j p_j / P_L, p_{j,R} = \beta_j p_j / P_R$ and $\theta(s,t) = f(\boldsymbol{\alpha}, \boldsymbol{p})$. In practice, the set of splits is restricted, e.g. univariate, but what we explore here is the question of what happens if all possible splits are allowed. That is, over the set of *all* $\boldsymbol{\alpha} \in [0,1]^J$, which $\boldsymbol{\alpha}$ maximizes $\theta(s,t)$? We answer this question for goodness-of-split criteria generated by impurity functions (Breiman, et al., 1984). We call the split corresponding to the maximizing $\boldsymbol{\alpha}$ the optimum split even though *it may not be realizable in terms of splits on the input variables.*

If $\boldsymbol{p} = (p_1,\ldots,p_J)$ are the node proportions, then $\phi(\boldsymbol{p})$ is an impurity function if it is convex in $\boldsymbol{p}$, has a maximum when all p_j are equal and is a minimum when one of the

$p_j = 1$. For $\phi(\boldsymbol{p})$ an impurity function the associated goodness-of-split is defined as

$$\theta(s,t) = \phi(\boldsymbol{p}) - P_L\phi(\boldsymbol{p}_L) - P_R\phi(\boldsymbol{p}_R).$$

The most commonly encountered impurity functions are the gini:

$$\phi(\boldsymbol{p}) = \sum_j p_j(1-p_j)$$

and the entropy

$$\phi(\boldsymbol{p}) = -\sum_j p_j \log p_j.$$

Another criterion discussed in Breiman et al. (1984) (pp. 104-106) is twoing. The idea is to find that grouping of all J classes into two superclasses so that considered as a two-class problem, the greatest decrease in node impurity is realized. If the gini impurity measure is used in the two class problem, then it is shown that the best twoing split at a node maximizes

$$\theta(s,t) = \frac{P_L P_R}{4}[\sum_j |p_{j,L} - p_{j,R}|]^2$$

and that when the split maximizing θ is used, the two superclasses are

$$\mathcal{C}_1 = \{j; p_{j,L} \geq p_{j,R}\}$$

$$\mathcal{C}_2 = \{j; p_{j,L} < p_{j,R}\}.$$

For splitting criteria generated by impurity functions, our approach reveals interesting differences. For example, the optimum split for the gini criterion sends all data in the class with the largest p_j to t_L and all other classes to t_R. Thus the best gini splits try to produce pure nodes. But the optimal split under the entropy criterion breaks the classes up into two disjoint subsets $\mathcal{C}_1, \mathcal{C}_1^c \subset \{1,\ldots,J\}$ such that $\mathcal{C}_1$ minimizes $|\sum_{j\in\mathcal{C}} p_j - .5|$ among all subsets $\mathcal{C} \subset \{1,\ldots,J\}$. Thus, optimizing the entropy criterion tends to equalize the sample sizes in t_L, t_R. The twoing criterion also tries to equalize.

The outline is as follows: in Section 2 we show that the split optimizing $\theta(s,t)$ has the property that all α_j are zero or one. That is, no classes have parts both in t_L and t_R. In Section 3 we find the optimal splits under the gini, entropy, and twoing measures. Section 4 gives conclusions. In particular, the results for the entropy measure suggest use of a partial look-ahead strategy.

2. Optimal Splits Do Not Split Classes

Let $\phi(\mathbf{x})$ be defined and twice differentiable for $\mathbf{x} \in [0,1]^J$. Assume that $\phi(\mathbf{x})$ is convex, i.e. the matrix $(\partial^2\phi/\partial x_i \partial x_j)$ is non-positive definite for all $\boldsymbol{x} \in [0,1]^J$. Let the impurity of t be $\phi(\mathbf{p})$ and the goodness-of-split be the decrease in impurity, i.e.

$$\theta(s,t) = \phi(\mathbf{p}) - P_L\phi(\mathbf{p}_L) - P_R\phi(\mathbf{p}_R).$$

THEOREM 1 *Let* $P_L = \sum \alpha_j p_j$, $p_{j,L} = \alpha_j p_j / P_L$, $P_R = 1 - P_L$, $p_{j,R} = (1 - \alpha_j)\, p_j / P_R$. *Then the maximum impurity decrease over* $\boldsymbol{\alpha} \in [0,1]^J$ *is achieved at a vertex of* $[0,1]^J$.

Proof: Suppose $P_L\phi(\mathbf{p}_L) + P_R\phi(\mathbf{p}_R)$ is convex in α. Then its minimum over $[0,1]^J$ is at an extreme point of $[0,1]^J$, i.e. a vertex. It is sufficient to show that $P_L\phi(\mathbf{p}_L)$ is convex in $\boldsymbol{\alpha}$, since $P_R\phi(\mathbf{p}_R)$ is the same function of $\boldsymbol{\beta} = \mathbf{e} - \boldsymbol{\alpha}$ ($\mathbf{e} = (1, \ldots, 1)$) as $P_L\phi(\mathbf{p}_L)$ is of $\boldsymbol{\alpha}$ and the sum of convex functions is convex. ■

The rest of the proof comes from using the result that

$$\frac{\partial^2}{\partial\alpha_i \partial\alpha_j}(P_L\phi(\mathbf{p}_L)) = \frac{1}{P_L} p_i p_j \sum_{\ell,h} \phi_{\ell h}(\delta_{i\ell} - \frac{\alpha_\ell p_\ell}{P_L})(\delta_{jh} - \frac{\alpha_h p_h}{P_L}) \tag{2.1}$$

where

$$\phi_{\ell h} = \frac{\partial^2 \phi(\mathbf{x})}{\partial x_\ell \partial x_h}|_{\mathbf{x}=\mathbf{p}_L}.$$

Equation (2.1) is derived in the Appendix. To show $P_L\phi(\boldsymbol{p}_L)$ convex in $\boldsymbol{\alpha}$, it is sufficient to show that for any J-vector $\boldsymbol{u}$,

$$\sum_{ij} u_i u_j \frac{\partial^2}{\partial\alpha_i \partial\alpha_j}(P_L\theta(\boldsymbol{p}_L)) \leq 0.$$

For any J-vector $\mathbf{u}$, define the J-vector $\boldsymbol{v}$ by

$$v_\ell = \sum_i u_i p_i(\delta_{i\ell} - \frac{\alpha_\ell p_\ell}{P_L}) = u_\ell p_\ell - \frac{\alpha_\ell p_\ell}{P_L}(\sum_i u_i p_i).$$

Then

$$\sum u_i u_j \frac{\partial^2}{\partial\alpha_i \partial\alpha_j}(P_L\phi(\mathbf{p}_L)) = \frac{1}{P_L} \sum_{\ell,n} \mathrm{v}_\ell \mathrm{v}_h \phi_{\ell h}.$$

Since ϕ is convex, this last term is non-positive, and thus, $P_L\phi(\mathbf{p}_L)$ is convex in α.

Both the gini and entropy criteria are of the form

$$\phi(\mathbf{x}) = \sum_j f(x_j)$$

with $f(x)$ convex implying ϕ convex. The gini f is $x(1-x)$ and entropy f is $-x \log x$.

The twoing criterion is

$$\theta(s,t) = \frac{P_L P_R}{4}[\sum_j |p_{j,L} - p_{j,R}|]^2. \tag{2.2}$$

This is not given by a difference in impurities, so the theorem above does not directly apply. Recall that the twoing criterion is derived from dividing the classes into two superclasses, finding the best gini split in this two class problem, and then optimizing the decrease in impurity over all divisions into two superclasses. If all splits are allowed, then the above theorem implies that each optimum two class split sends all of one class to t_L and all of the other to t_R. Thus, the best twoing split is also at a vertex of $[0,1]^J$.

3. Specific Optima

This section answers the question of which vertex of $[0,1]^J$ is optimum for the entropy, gini, and twoing criteria. For the entropy measure, we want to maximize

$$P_L \sum_j p_{j,L} \log p_{j,L} + P_R \sum_j p_{j,R} \log p_{j,R}.$$

For a given vertex, let $\mathcal{C}_0 = \{j; \alpha_j = 0\}$, $\mathcal{C}_1 = \{j; \alpha_j = 1\}$. The above expression becomes

$$P_L \sum_{j \in \mathcal{C}_1} (p_j/P_L) \log(p_j/P_L) + P_R \sum_{j \in \mathcal{C}_0} (p_j/P_R) \log(p_j/P_R)$$

$$= \sum_j p_j \log p_j - P_L \log P_L - P_R \log P_R.$$

The optimum vertex maximizes

$$-P_L \log P_L - P_R \log P_R \tag{3.3}$$

So at the best vertex $|P_L - .5|$ is minimized.

With the gini measure, the best vertex minimizes

$$P_L \sum p_{j,L}(1 - p_{j,L}) + P_R \sum p_{j,R}(1 - p_{j,R}) =$$

$$P_L \sum_{j \in \mathcal{C}_1} (p_j/P_L)(1 - p_j/P_L) + P_R \sum_{j \in \mathcal{C}_0} (p_j/P_R)(1 - (p_j/P_R).$$

Equivalently, choose that vertex which maximizes

$$\frac{1}{P_L} \sum_{j \in \mathcal{C}_1} p_j^2 + \frac{1}{P_R} \sum_{j \in \mathcal{C}_0} p_j^2. \tag{3.4}$$

PROPOSITION *Let $p_i = \max_j(p_j)$. Then the best gini vertex sends all of class i to t_L and the remainder to t_R.*

The proof of this proposition involves some algebraic manipulation and is deferred to the appendix. Finally, note that on any vertex, the twoing measure (2) equals $P_L P_R/4$. Thus, the best vertex minimizes $|P_L - .5|$.

4. Discussion and Conclusions

The above shows the difference between the best splits selected using the gini criterion versus the entropy and twoing criteria. The gini prefers splits that put the largest class into one pure node, and all others into the other. Entropy and twoing put their emphasis on

balancing the sizes at the two children nodes. These theoretical conclusions get support in the simulations in Breiman et. al (1984) (see pp. 111).

In problems with a small number of classes, all criteria should produce similar results. The differences appear in data where J is larger. Here, high up in the tree, gini may produce splits that are too unbalanced. On the other hand, the above results show a disturbing facet of the entropy and twoing criterion, i.e. a lack of uniqueness. If J is moderate to large, there are usually many vertices such that $P_L \simeq .5$. For instance, in a little simulation, we took $J = 10$ and selected the $\{p_j\}$ to be uniform random numbers, suitably normalized. On the average, for each set of $\{p_j\}$ about 40 vertices gave P_L values between .49 and .51 with 4 vertices such that $.499 \leq P_L \leq .501$. These vertices often differed in the distribution of both the larger and smaller p_j values.

Since many vertices have similar goodness-of-split values, selecting the best split is a bit arbitrary. Which split is best depends on the future evolution of the tree. This suggests that use of the entropy or twoing criteria be combined with a limited two step look-ahead. For instance, one could set an integer N, and for each of the N best splits of a node t compute the total decrease in impurity following the splits of t_L into t_{LL}, t_{LR} and t_R into t_{RL}, t_{RR}. Then use the best of the N. One must take care to ensure that if some of the N splits are on the same variable, they are sufficiently different.

Appendix

Derivation of (2.1)

Let $x_j = \alpha_j p_j / P_L$. Then for any $g(\boldsymbol{x})$

$$\begin{aligned} \frac{\partial g(x)}{\partial \alpha_j} &= \sum_\ell \frac{\partial g}{\partial x_\ell}\left(-\frac{\alpha_\ell p_\ell p_j}{P_L^2} + \delta_{\ell j}\frac{p_j}{P_L}\right) \\ &= \frac{p_j}{P_L}\left(\frac{\partial g}{\partial x_j} - \sum_\ell \frac{\partial g}{\partial x_\ell} x_\ell\right) \end{aligned} \tag{A.1}$$

Using the notation $\phi_i = \partial\phi/\partial x_i, \phi_{ij} = \partial^2\phi/\partial x_i \partial x_j$ and applying (A.1) gives

$$\frac{\partial}{\partial \alpha_j}(P_L \phi(\boldsymbol{x})) = p_j[\phi + \phi_j - \sum_\ell \phi_\ell x_\ell].$$

Take $H(\boldsymbol{x})$ to be the term in brackets in the above equation and note that

$$\frac{\partial H(\boldsymbol{x})}{\partial x_i} = \phi_{ij} - \sum_\ell \phi_{i\ell} x_\ell.$$

Using (A.1) again

$$\frac{\partial}{\partial \alpha_i}(p_j H) = \frac{p_i p_j}{P_L}\left(\frac{\partial H}{\partial x_i} - \sum_h \frac{\partial H}{\partial x_h} x_h\right)$$

$$= \frac{p_i p_j}{P_L}[\phi_{ij} - \sum_{\ell} \phi_{i\ell} x_\ell - \sum_{h} \phi_{hj} x_h + \sum_{\ell,h} \phi_{h\ell} x_h x_\ell]$$

$$= \frac{p_i p_j}{P_L} \sum_{\ell,h} \phi_{\ell h}(\delta_{i\ell} - x_\ell)(\delta_{jh} - x_h)$$

Proof of the proposition: For any set of indices $\mathcal{C} \subset \{1, \ldots, J\}$, let $Q(\mathcal{C}) = \sum_{j \in \mathcal{C}} p_j^2$, $P(\mathcal{C}) = \sum_{j \in \mathcal{C}} p_j$, and $\lambda(\mathcal{C}) = Q(\mathcal{C})/P(\mathcal{C})$. We want to maximize $G(\mathcal{C}) = \lambda(\mathcal{C}) + \lambda(\mathcal{C}^c)$. If $\mathcal{C}$ maximizes G, so does $\mathcal{C}^c$. Take as $\mathcal{C}$ whichever one satisfies $\lambda(\mathcal{C}) \geq \lambda(\mathcal{C}^c)$. Let $p_i = \max(p_j, j \in \mathcal{C})$, and take $\mathcal{C}_1 = \mathcal{C} - \{i\}$. We will show that

$$G(\{i\}) \geq G(\mathcal{C}) \tag{A.2}$$

so that a maximizer of G sends all cases in one class to one child node, and all other classes to the other child. ■

The inequality A.2 follows from the identity

$$G(\{i\}) - G(\mathcal{C}) = \frac{P(\mathcal{C}_1)}{p_i(1-p_i)}[(1-p_i)\lambda(\mathcal{C}) - p_i\lambda(\mathcal{C}^c) - (1-2p_i)\lambda(\mathcal{C}_1)] \tag{A.3}$$

This identity can be derived from the simpler identity

$$G(\{i\}) = \frac{1}{p_i}[\lambda(\mathcal{C})p(\mathcal{C}) - Q(\mathcal{C}_1)] + \frac{1}{1-p_i}[\lambda(\mathcal{C}^c)P(\mathcal{C}^c) + Q(\mathcal{C}_1)] \tag{A.4}$$

Subtracting $G(\mathcal{C})$ from A.4 and simplifying gives A.3. Suppose first that $p_i \geq 1/2$. Then to prove A.2 its sufficient to show that

$$(1-p_i)\lambda(\mathcal{C}) \geq p_i\lambda(\mathcal{C}^c). \tag{A.5}$$

For any subset $\mathcal{D}$ of indices

$$\begin{aligned} Q(\mathcal{D}) &\leq (P(\mathcal{D}))^2 \\ \Rightarrow \lambda(\mathcal{D}) &\leq P(\mathcal{D}). \end{aligned}$$

Now

$$(1-p_i)\lambda(\mathcal{C}) = (1-p_i)\frac{Q(\mathcal{C})}{P(\mathcal{C})} \geq \frac{(1-p_i)p_i^2}{P(\mathcal{C})}$$

and $\lambda(\mathcal{C}^c) \leq P(\mathcal{C}^c) = 1 - P(\mathcal{C})$. Hence A.5 will follow from

$$(1-p_i)p_i \geq P(\mathcal{C})(1-P(\mathcal{C})) \tag{A.6}$$

The expression $x(1-x)$ is decreasing for $x \geq 1/2$. Since $P(\mathcal{C}) > p_i$, A.6 is true.

Now assume $p_i \leq 1/2$. Then A.2 again follows from showing that the term in brackets in A.3 is non-negative. Rewrite this term as

$$(1-2p_i)\lambda(\mathcal{C}) + p_i(\lambda(\mathcal{C}) - \lambda(\mathcal{C}^c)) - (1-2p_i)\lambda(\mathcal{C}_1).$$

By assumption, this is greater than or equal to

$$(1 - 2p_i)(\lambda(\mathcal{C}) - \lambda(\mathcal{C}_1)).$$

Since

$$\lambda(\mathcal{C}) - \lambda(\mathcal{C}_1) = p_i \frac{(p_i P(\mathcal{C}) - Q(\mathcal{C}))}{P(\mathcal{C})(P(\mathcal{C}) - p_i)}$$

and $Q(\mathcal{C}) \leq p_i P(\mathcal{C})$, the proof of A.2 is complete, and $G(\{i\})$ is a maximizer of $G(\mathcal{C})$.

Now we show that if $p_i \leq p_j$, then $G(\{i\}) \leq G(\{j\})$. This follows from the identity

$$G(\{j\}) - G(\{i\}) = (p_j - p_i)[(1 - p_i - p_j)^2 + \sum_{h \neq i,j} p_h^2]/(1 - p_i)(1 - p_j)$$

resulting from straightforward algebra.

References

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). "Classification and Regression Trees," Wadsworth.

Buntine, W. & Niblett, T. (1992). "A further comparison of splitting rules for decision tree induction," Machine Learning 8, 75-85.

Fayyad, U.M. (1991). "On the induction of decision trees for multiple concept learning," Ph.D Thesis, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan.

Fayyad, U.M. & Irani, R.B. (1990). "What should be minimized in a decision tree"? Proc. 8th National Conf. on AI, AAAI-90, 749-754, MIT Press.

Fayyad, V.M. & Irani, R.B. (1992). "The attribute selection problem in decision tree generation," Proc. 10th National Conf. on AI, AAAI-92, 104-110, MIT Press.

Fayyad, U.M. & Irani, R.B. (1993). "Multi-interval discretization of continuous valued attribute for classification learning," Proc. 13th International Joint Conf. on AI, 1022-1027, Morgan Kaufmann.

Taylor, P.C. & Silverman, B.W. (1993). "Block diagrams and splitting criteria for classification trees," Statistics and Computing, V 3, p. 147-161.

Received April 29, 1994
Accepted August 26, 1994
Final Maunscript October 2, 1995