

# Decision Tree I

## Growing a Large Tree

### Contents

<b>1</b>	<b>A Single Split</b>	<b>2</b>
1.1	Node Impurity . . . . .	2
1.2	Computation of $i(t)$ . . . . .	4
1.2.1	In Prospective Studies . . . . .	4
1.2.2	In Retrospective Studies . . . . .	4
1.3	Goodness-of-Split Measure . . . . .	5
1.4	Alternative Splitting Criterion . . . . .	5
<b>2</b>	<b>Induction of a Split</b>	<b>5</b>
2.1	Count the Number of Splits . . . . .	5
2.2	Strategies to Reduce the Number of Possible Partitions . . . . .	7
<b>3</b>	<b>Other Issues</b>	<b>8</b>
3.1	Bias Adjustment for Multi-Way Split . . . . .	8
3.2	Variable Combination . . . . .	10

We will follow the CART methodology to develop tree models, which contains the following three major steps:

1. Grow a large initial tree,  $T_0$ ;
2. Iteratively truncate branches of  $T_0$  to obtain a sequence of optimally pruned (nested) subtrees;
3. Select the best tree size based on validation provided by either the test sample method or cross validation (CV).

To illustrate, we first consider decision trees with binary responses. Namely,

$$y_i = \begin{cases} 1 & \text{when an event of interest occurs to subject } i; \\ 0 & \text{otherwise.} \end{cases}$$

In this section, we focus on how to perform a single split of the data and how to grow a large tree.

## 1 A Single Split

### 1.1 Node Impurity

In general, the impurity  $i(t)$  of node  $t$  can be defined as a nonnegative (concave) function of  $P\{y = 1|t\}$ , which is the occurrence rate in node  $t$ . More formally,

$$i(t) = \phi(P\{y = 1|t\}),$$

where the function  $\phi(\cdot)$ , very intuitively, the impurity is the largest when both classes are equally mixed together and it is the smallest when the node contains only one class. Hence, it has the following properties:

1.  $\phi(p) \geq 0$ ;
2.  $\phi(p)$  attains its minimum 0 when  $p = 0$  or  $p = 1$ .
3.  $\phi(p)$  attains its maximum when  $p = 1 - p = 1/2$ .
4.  $\phi(p) = \phi(1 - p)$ , *i.e.*,  $\phi(p)$  is symmetric to  $p = 1/2$ .

Common choices of  $\phi$  include: the Bayes error or the minimum error, the entropy function, and the Gini index.

- **The Bayes Error**

$$\phi(p) = \min(p, 1 - p) = 1 - \max(p, 1 - p).$$

This measure corresponds to the misclassification rate when majority vote is used. The Bayes error is rarely used in practice due to some undesirable properties as explained by Brieman et al. (1984, p. 99).

- **The Entropy Function**

$$\phi(p) = -p \log(p) - (1 - p) \log(1 - p).$$

Entropy is a measure of variability for categorical data. It was first developed by Shannon in 1984 to measure the uncertainty of a transmitted message. Quinlan (1993) first proposed to use the reduction of Entropy as a measure of split criteria. Ripley (1996) showed the entropy reduction criterion is equivalent to using the likelihood ratio chi-square statistics for association between the branches and the target categories.

- **The Gini Index**

$$\phi(p) = p(1 - p).$$

Gini index is a measure of variability for categorical data developed by Italian statistician Corrado Gini in 1912. Breiman et al. (1984) proposed to use the reduction of Gini index as a measure for split criteria. It has been observed that this rule has an undesirable end-cut preference problem (Breiman et al., 1984, Ch. 11): It gives preference to the splits that result in two daughter nodes of extremely unbalanced sizes. To resolve this problem, a modification called the delta splitting method has been adopted in both the THAID (Morgan and Messenger, 1973) and CART programs.

Because of the above concerns, from now on the impurity refers to the entropy criterion unless stated otherwise. The following SAS program plots the three criteria described above in one graph.

```
%let n = 1000;
data temp;
  do i = 1 to (&n-1);
    p = i/&n;
    misclass = min(p, 1-p);
    Entropy= - p*log(p) - (1-p)*log(1-p);
    Gini= p *(1-p);
    output;
  end;
run;

goptions reset=global gunit=pct border cback=pink
  colors=(black blue green red)
  ftitle=centb ftext=swiss htitle=6 htext=4;
proc gplot data=temp;
title font=SWISSB 'Figure 5.1 Node Impurity Measures';
symbol1 c=b v=none i=join;
symbol2 c=bl v=none i=join;
symbol3 c=r v=none i=join;
```

```

legend1 across=3 position=(top inside right) label=none mode=share;
plot misclass*p=1 Entropy*p=2 Gini*p=3 / frame overlay legend=legend1;
run;
quit;

```

## 1.2 Computation of $i(t)$

The computation of impurity is simple when the occurrence rate  $P\{y = 1|t\}$  in node  $t$  is available. In many applications such as prospective studies, this occurrence rate can be estimated empirically from the data. At other times, additional prior information may be required to estimate the occurrence rate.

For a given split  $s$ , we have the following  $2 \times 2$  table according to the split and the response.

node	response		
	0	1	
left ( $t_L$ )	$n_{11}$	$n_{12}$	$n_{1\cdot}$
right ( $t_R$ )	$n_{21}$	$n_{22}$	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n$

### 1.2.1 In Prospective Studies

In prospective studies,  $P\{y = 1|t_L\}$  and  $P\{y = 0|t_L\}$  can be estimated by  $n_{12}/n_{1\cdot}$  and  $n_{22}/n_{1\cdot}$ . And hence

$$i(t_L) = -\frac{n_{12}}{n_{1\cdot}} \log\left(\frac{n_{12}}{n_{1\cdot}}\right) - \frac{n_{11}}{n_{1\cdot}} \log\left(\frac{n_{11}}{n_{1\cdot}}\right).$$

In fact, it can be shown that the above entropy criterion is proportional to the maximized log-likelihood score associated with  $t_L$ . In light of this fact, many node-splitting criteria originate from the maximum of certain likelihood functions. The importance of this observation will be appreciated later.

### 1.2.2 In Retrospective Studies

In retrospective studies, Bayes' theorem can be used to compute  $P\{y = 1|t_L\}$ :

$$\begin{aligned}
P\{y = 1|t_L\} &= \frac{P\{y = 1, t_L\}}{P\{t_L\}} \\
&= \frac{P\{y = 1\} P\{t_L|y = 1\}}{P\{y = 1\} P\{t_L|y = 1\} + P\{y = 0\} P\{t_L|y = 0\}},
\end{aligned}$$

where  $P\{y = 1\} = 1 - P\{y = 0\}$  is the prior occurrence rate of the event of interest.

### 1.3 Goodness-of-Split Measure

Suppose  $s$  be any candidate split and  $s$  divides  $t$  into  $t_L$  and  $t_R$  such that the proportions of the cases in  $t$  go into  $t_L$  and  $t_R$  are  $P\{t_L\}$  and  $P\{t_R\}$ , respectively. Define the reduction in node impurity as

$$\Delta I(s, t) = i(t) - [P\{t_L\}i(t_L) + P\{t_R\}i(t_R)],$$

which provides a goodness-of-split measure for  $s$ . The best split  $s^*$  for node  $t$  provides the maximum impurity reduction, i.e.,

$$\Delta I(s^*, t) = \max_{s \in \mathcal{S}} I(s, t).$$

Then  $t$  will be split into  $t_L$  and  $t_R$  according to the split  $s^*$  and the search procedure for the best split repeated on  $t_L$  and  $t_R$  separately. A node becomes to a terminal node when the impurity cannot decrease any further based some terminal conditions specified.

### 1.4 Alternative Splitting Criterion

There are two alternative splitting criteria: the twoing rule (CART, Brieman, et al., 1984) and the  $\chi^2$  test.

- The twoing rule is a different measure for the goodness of a split as follows:

$$\frac{P\{t_L\}P\{t_R\}}{4} \left[ \sum_{j=0,1} |P\{y = j | t_L\} - P\{y = j | t_R\}| \right]^2.$$

For a binary response, this twoing rule coincides with the use of the Gini index, which has the end-cut preference problem.

- The Pearson chi-square test statistics can be used to measure the difference between the observed cell frequencies and the expected cell frequencies (under the independent assumption). We can use the p-value to make the judgement. When p-value is too small, we can use logworth:

$$\text{logworth} = -\log_{10}(\text{p-value}).$$

## 2 Induction of a Split

### 2.1 Count the Number of Splits

The next problem in tree construction is how to determine the number of partitions needed to examine at each node. An exhaustive (greedy) search algorithm considers all possible partitions of all input variables at every node in the tree. However, the number of child

nodes tends to increase significantly when either too many levels in one variable or too many variables. This makes an exhaustive search algorithm prohibitively expensive.

Multi-way splits are not more flexible than binary splits. Actually, we can always find a binary tree that is equivalent to any multi way tree (see Figure 5.1). Multi-way splits often give a more interpretable tree because split variable tends to be used fewer times. However, the number of possible partitions for a binary split tree is much less than a multi-way split tree and exhaustive search is more feasible in binary split tree.

## Examples

1. Suppose  $x$  is an ordinal variable with four levels 1, 2, 3, and 4. What is the total number of possible splits including both binary and multiway ones?

**Solution:**

2 way split: 1-234, 12-34, 123-4  
 3 way split: 1-2-34, 1-23-4, 12-3-4  
 4 way split: 1-2-3-4  
 Total number of splits = 3+3+1=7

Note that there are  $2^{L-1} - 1$  possible splits for an ordinal variable with  $L$  levels. Splits in a tree only depend on the order statistics of the numerical variables. The same formula can be used to compute the number of possible partitions for both ordinal and numerical variables.

2. Suppose  $x$  is a numerical variable with 100 distinct values. What is the total number of possible splits?

**Solution:**

Total number of splits =  $2^{100-1} - 1 \approx 6.34 \times 10^{29}$ .

3. Suppose  $x$  is a nominal variable with four categories a, b, c, d. What is the total number of possible splits?

**Solution:**

2 way split: ab-cd, ac-bd, ad-bc, abc-d, abd-c, acd-b, a-bcd  
 3 way split: a-b-cd, a-bc-d, ac-b-d, ab-c-d, a-c-bd, ad-b-c  
 4 way split: a-b-c-d  
 Total number of splits = 7+6+1=14

Note that the total number of possible partitions for a nominal variable with  $L$  levels is called the *Bell number*

$$B_L = \sum_{i=2}^L S(L, i),$$

where  $S(L, i)$  is the *Stirling number of the second kind* or also called a Stirling set number, which corresponds to the number of ways of partitioning a set of  $L$  elements into  $i$  nonempty

sets (i.e.,  $i$  set blocks). Special values of  $S(L, i)$  include

$$\begin{aligned}S(L, L) &= S(L, 1) = 1 \\S(L, 2) &= 2^{n-1} - 1 \\S(L, L - 1) &= \binom{n}{2}.\end{aligned}$$

$S(L, i)$  satisfies

$$S(L, i) = i \cdot S(L - 1, i) + S(L - 1, i - 1).$$

The Stirling numbers of the second kind can be computed from the sum

$$S(L, i) = \frac{1}{i!} \sum_{j=0}^{i-1} (-1)^j \binom{i}{j} (i - j)^n.$$

For more information, visit

<http://mathworld.wolfram.com/StirlingNumberoftheSecondKind.html>

## 2.2 Strategies to Reduce the Number of Possible Partitions

Since the exhaustive search is too expensive, many methods that can reduce the total number of partitions needed at each node have been developed. Some of these methods are listed below:

- Binary Splits Exclusive (CART like tree)
- Agglomerative Clustering of Levels (CHAID like tree)
- Variable Selection First
- Within-Node Sampling Approach
- Minimum Child Size

### 1. Binary Splits Exclusive

One way to reduce the number of possible partitions is to consider only binary splits. If  $x_1$  is an ordinal variable with  $L$  levels, then the total number of possible binary splits is  $L - 1$ ; if  $x_2$  is an ordinal variable with  $L$  levels, then the total number of possible binary splits is  $2^{L-1} - 1$ .

### 2. Agglomerative Clustering of Levels

The procedure for clustering levels is as follows:

Step 1 : Start with an L-way split

- Step 2 : Collapse the two levels that are closest based on some splitting criterion
- Step 3 : Repeat Step 1 and Step 2 on the set of L-1 consolidated levels if  $L - 1 \geq 2$ .
- Step 4 : Step 1 to Step 3 will give the best split for each size and can choose the best split for this variable.
- Step 5 : Repeat this process for all other input variables and choose the best possible split for all variables

### 3. Variable Selection First

We can choose only these variables that are highly correlated to the target variable. Suppose there are 2000 binary variables and only 200 variables are highly correlated to the target variables, the number of possible partitions to be considered are reduced by 90%.

### 4. Within-Node Sampling

The number of possible partitions to be considered depends on the sample size in a node. For example, there are 10000 observations in one node, the maximum possible distinct values for a numerical variable is 10000. Thus, the possible partitions for this numerical variable are 4999 if we use binary split exclusive if we only consider a sample of 5000 observations.

### 5. Set Minimum Number of Observations in Each Node

Suppose the minimum number of observation for a child node is set to 50. We will not split this node further if the number of the observations in this node is below 50.

Besides, for categorical predictors that has many levels  $\{B_1, \dots, B_L\}$ , one way to reduce the number of splits is to rank the levels as  $\{B_{l_1}, \dots, B_{l_L}\}$  according to the occurrence rate within each level

$$P\{1|B_{l_1}\} \leq P\{1|B_{l_2}\} \leq \dots \leq P\{1|B_{l_L}\}$$

and then treat it as an ordinal input. (See CART, p. 101).

## 3 Other Issues

### 3.1 Bias Adjustment for Multi-Way Split

For the  $\chi^2$  test with multi-way splits, the large table has higher chance to produce a large chi-square statistics because the expectation of the chi-square statistics with  $\gamma$  degrees of freedom is  $\gamma$  and the large table has large degrees of freedom. However, the p-value does not depend on the size of the contingency table.

Both the Gini Index and the Entropy tend to increase as the number of branches increased. For example, the maximum possible value for a Gini index with 20 branches is 0.95

and the maximum possible value for a Gini index with two branches is 0.5. In addition, there is not any analogous adjustment similar to p-value in chi-square test statistics that can be used.

The use of splitting criterion can be thought as a two-step process. First, we can use split criterion to choice the best split among all possible splits for a given input variable. Then, we can use the split criterion to choice the best split among all the “best” split from each input variable. Both steps required some adjustment to ensure the fair comparison.

### 1. Adjustment Within the Same Input Variable

- Adjustment is not necessary if only binary split is considered.
- Adjustment can not be done if either Gini or Entropy split criterion is used
- P-value (or logworth) adjustment can be used if the chi-square split criterion is used
- Entropy tends to favor balance branches (Breiman, 1996)
- Gini index tends to favor isolating the largest target class (Breiman, 1996)

### 2. Adjustment among Best Splits from Different Input Variables

There are more splits to consider on a variable with more levels. Therefore, the maximum possible value for Gini index, Entropy, and logworth tends to become large as the number of possible splits,  $m$ , increases. For example, there is only one split for a binary input variable and there are 511 possible binary splits for a nominal input variable with 10 levels. Thus, all three split-criteria are favor variables with large number of levels. Nominal variables are favored than ordinal variables with the same number of levels. This problem has been identified as ‘*variable selection bias*’ problem by Loh, WY.

- Adjustment for Gini index is unavailable
- The information gain ratio can be used to adjust Entropy (Quinlan, 1993). However, it is not available in Enterprise miner.

$$\text{information gain ratio} = \frac{\Delta\text{Entropy}}{\text{input levels in parent node}}.$$

- Bonferroni type of adjustment can be used to adjust the logworth (Kass, 1980). Kass adjustment is multiply the p-value by  $m$ , the number of possible splits. This adjustment is equivalent to subtract  $\log_{10}(m)$  from the logworth.
- In order to identify the ‘unbiased’ split, Loh (2002) proposed a residual-based method of selecting the most important variable first, and then applying greedy search only on this variable to find the best cutpoint.

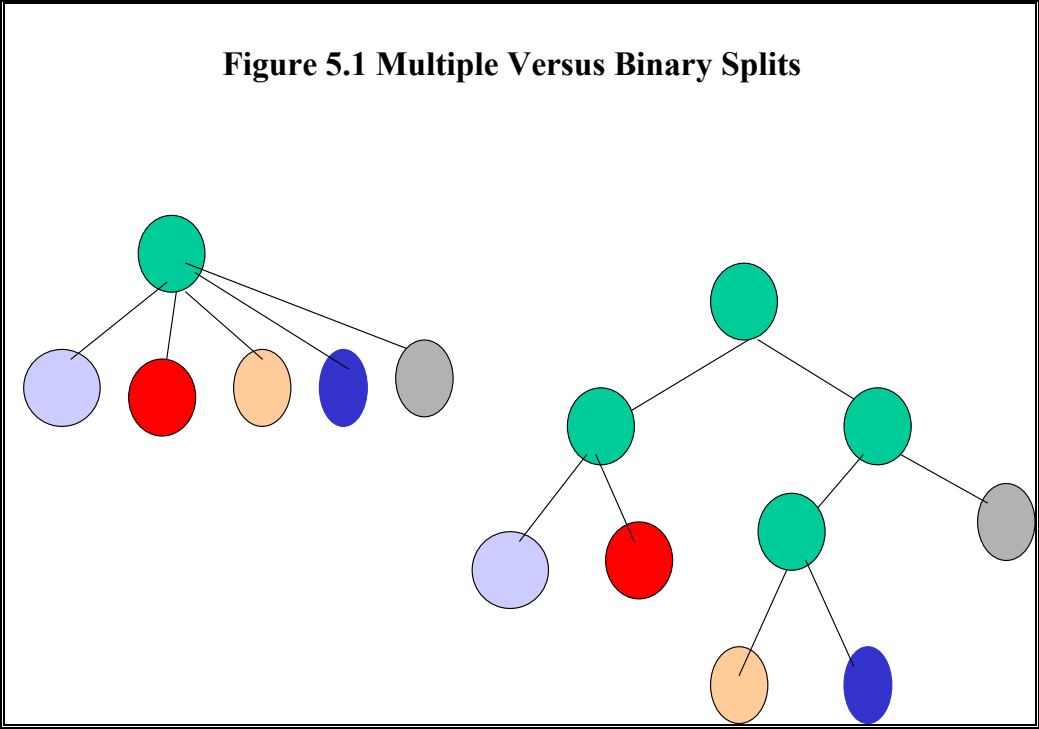
## 3.2 Variable Combination

Another major deficiency for the tree modeling introduced so far is we only consider splits on a single variable. Suppose that there is a linear structure existed in the data, tree structure modeling without considering the linear combination might not be able to pick the best model. Moreover, it might be worse than the existing method such as linear discriminate analysis. We use the next example (figure 5.2) to show illustrate the deficiency of ignoring linear combination of variables.

Discriminant analysis can do a perfect job to separate these two classes. However, the tree program without considering the linear combination of variable  $x_1$  and  $x_2$  would take many splits to separate these two classes because it tries to separate the plane into rectangular regions. However, this problem can be solved very easily if the tree can consider to separate the plane with linear combination of variables such as  $x' = ax_1 + bx_2$ . In the computer science literature, tree models that allows for linear combinations of inputs are terms as multivariate decision trees. One is referred to Li et al. (PHDRT, *JASA*, 2000) for recent research on this topic.

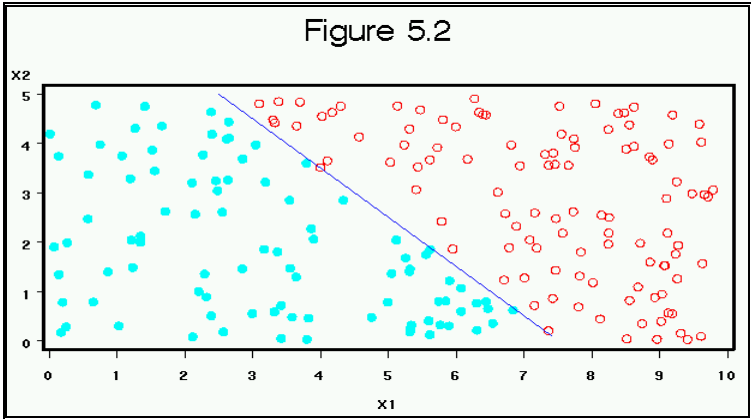
## REFERENCES

- [1] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984) Classification and Regression Trees, Chapman and Hall.
- [2] Kass, G. V. (1980) "An Exploratory Technique for Investigating Large Quantities of Categorical Data," Applied Statistics, Vol. 29, pp. 119-127.
- [3] Loh, W. and Vanichsetakul, N. (1988), "Tree-Structured Classification via Generalized Discriminate Analysis," Vol. 83, pp. 715-727.
- [4] Loh, W. and Shih, Y. (1997) Split Selection Methods for Classification Trees, Statistical sinica, Vol. 7, pp 815-840.
- [5] Quinlan, J. R. (1993), Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann.
- [6] SAS Institute (2000) Decision Tree Modeling Course Notes.
- [7] Zhang, H. and Singer, B. (1999). Recursive Partitioning in the Health Sciences. Springer-Verlag Inc.: New York.



**Table 5.1 The number of partitions on a nominal variable**

Level	Multiway Splits								Total
	2	3	4	5	6	7	8	9	
2	1								1
3	3	1							4
4	7	6	1						14
5	15	25	10	1					51
6	31	90	65	15	1				202
7	63	301	350	140	21	1			876
8	127	966	1701	1050	266	28	1		4139
9	255	3025	7770	6951	2646	462	36	1	21146



**Example**

Complete the following table for split 1, split 2, and split3 if there are 97 distinct values in X.

Split 1:

	X<38.5	X>38.5
1	293	71
7	363	1
9	42	294

Split 2:

	X<17.5	17.5<X<36.5	X>36.5
1	249	42	73
7	338	25	1
9	26	16	294

Split 3:

	X<0.5	0.5<X<41.5	41.5<X<51.5	X>51.5
1	9	143	65	147
7	221	88	1	54
9	1	4	16	315

Table:

	$X_{\gamma}^2$	$\gamma$	$-\log_{10}(\text{P-value})$	m	$-\log_{10}(m*\text{p-value})$
Split 1	640	2	140		
Split 2	660	4	141		
Split 3	814	6	172		

**Solutions:**

	$X_{\gamma}^2$	$\gamma$	$-\log_{10}(\text{P-value})$	m	$-\log_{10}(m*\text{p-Value})$
Split 1	640	2	140	96	138
Split 2	660	4	141	4560	137
Split 3	814	6	172	156849	167

**Note:**  $m = \binom{L-1}{B-1}$  ordinal input variable and  $m = S(L, B)$  for nominal input variable.