

Generalization to Multi-Class and Continuous Responses

STA 5703 - Data Mining I

Outline

1. Categorical Responses

(a) Splitting Criterion

- Goodness-of-split Criterion
- Chi-square Tests and Twoing Rule

(b) Pruning and Tree-Size Selection

(c) Ordinal Responses

2. Continuous Responses

(a) Splitting Criterion

- Goodness-of-Split Criterion
- F or t test

(b) Pruning and Tree Size Selection

3. Other Extensions - Maximum Likelihood Framework

Categorical Responses

- Nominal Responses

Which data mining software was used by each surveyed company?

- Ordinal Responses

Symptoms of back pain after treatment: worse - 1, same - 2, slight improvement-3, moderate improvement - 4, marked improvement - 5, complete relief -6.

- Suppose that the response y has J classes: $\{1, 2, \dots, J\}$.

Associated with each response y is a number of K predictors or inputs $\{x_1, \dots, x_K\}$.

Again, we follow the CART methodology to develop the final tree model.

Splitting by Goodness-of-Split Criterion

If the goodness-of-split criterion is used for comparing splits, the best split s^* achieves the greatest reduction in terms of node impurity. Namely,

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

where

$$\Delta i(s, t) = i(t) - \{p(t_L)i(t_L) + p(t_R)i(t_R)\}.$$

The node impurity measure $i(t)$ is a nonnegative concave function of $p(j|t)$ for $j = 1, 2, \dots, J$, which are the probability that an individual falls into class j in node t . Computation of $p(j|t)$ depends on the availability of the prior probabilities.

Entropy-Based Node Impurity

$$i(t) = - \sum_{j=1}^J p(j|t) \log\{p(j|t)\}$$

- The entropy based node impurity is preferable in the splitting stage as it favors balanced splits.
- The above entropy measure has a correspondence with the maximum likelihood score.
- The goodness-of-split criterion based on entropy is equivalent to maximizing the likelihood ratio test statistic.

Gini Index -Based Node Impurity

CART gives lots of credit to Gini index mainly because of its interpretation and easy incorporation of misclassification cost.

$$\begin{aligned}i(t) &= \sum_{j=1}^J p(j|t)(1 - p(j|t)) \\&= 1 - \sum_{j=1}^J p^2(j|t) \\&= \left\{ \sum_{j=1}^J p(j|t) \right\}^2 - \sum_{j=1}^J p^2(j|t) \\&= \sum_{j \neq i} p(j|t)p(i|t).\end{aligned}$$

using $\sum_{j=1}^J p(j|t) = 1$.

Interpretation of Gini Index

- Interpretation: Instead of using the plurality (majority) rule to classify objects in a node t , one may use the rule that assigns an object at random from the node to class i with probability $p(i|t)$. The estimated probability that the item is *actually* in class j is $p(j|t)$. Therefore, the estimated probability of misclassification under this rule is the Gini index

$$\sum_{j \neq i} p(j|t)p(i|t).$$

- The above property allows easy incorporation of misclassification cost into the Gini index and makes the Gini index a useful criterion frequently used in the final tree size selection. (see the Advance tab in Tree Node of Enterprise Miner.)

Other Properties of Gini Index

- Another interpretation for expression

$i(t) = \sum_{j=1}^J p(j|t)(1 - p(j|t))$ can be made in terms of the variances of dummy variables that indicate memberships.

- Gini index as a function $\phi(p_1, \dots, p_J)$ of p_1, \dots, p_J is concave in the sense that for $r + s = 1$, $r, s \geq 0$,

$$\phi(rp_1 + sp'_1, \dots, rp_J + sp'_J) \geq r\phi(p_1, \dots, p_J) + s\phi(p'_1, \dots, p'_J).$$

This ensures that for any split s ,

$$\Delta(s, t) \geq 0.$$

- The Gini index based splitting criterion tends to favor unbalanced splits. Again, the delta splitting rule should be applied: the values of cut points for x_k are restricted to the interval $(\min_i x_{ik} + \delta, \max_i x_{ik} - \delta)$.

Outline (Sign-Posting)

1. Categorical Responses
 - (a) Splitting Criterion
 - Goodness-of-split Criterion
 - **Chi-square Tests and Twoing Rule**
 - (b) Pruning and Tree-Size Selection
 - (c) Ordinal Responses
2. Continuous Responses
 - (a) Splitting Criterion
 - Goodness-of-Split Criterion
 - F or t test
 - (b) Pruning and Tree Size Selection
3. Other Extensions - Maximum Likelihood Framework

Chi Square Tests

A split s induces the following $2 \times J$ contingency table:

node	response				
	1	2	...	J	
left	n_{11}	n_{12}	...	n_{1J}	$n_{1.}$
right	n_{21}	n_{22}	...	n_{2J}	$n_{2.}$
	$n_{.1}$	$n_{.2}$...	$n_{.J}$	n

The chi-square test statistic can be constructed as

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^J \frac{n_{ij} - E_{ij}}{E_{ij}} \sim \chi_{(J-1)}^2,$$

where $E_{ij} = n_{i.}n_{.j}/n$ is the expected count under the null of no association between split and response.

Split with Chi Square Tests

- The chi square test can be easily extended to split data with multi-class responses. For binary splits, the best split maximizes the chi square test statistic.
- For multi-way splits, one may compute the logworth associated with the chi square test

$$\text{logworth} = -\log(\text{p-value}) = -\log \Pr \left\{ \chi^2_{(J-1)} > \chi^2 \right\}$$

and select the split associated with largest logworth. Note that chi square test is the only way of performing multi-way splits in SAS Enterprise Miner (SEM).

- Another advantage of splitting data with chi square tests is that it allows Bonferroni-type adjustments for number of inputs and depth.

Twoing Rule - Detailed Steps

1. Denote the whole set of all classes by \mathcal{C} , i.e., $\mathcal{C} = \{1, \dots, J\}$. At each node, separate the classes into two superclasses or subsets,

$$\mathcal{C}_1 = \{j_1, \dots, j_{n_1}\} \quad \text{and} \quad \mathcal{C}_2 = \mathcal{C} - \mathcal{C}_1.$$

Assign all objects in \mathcal{C}_1 as level 1 and all objects in \mathcal{C}_2 as level 0.

2. For any given split s of the node t , compute the Gini index based $\Delta i(s, t)$ as if it were a two-class problem. And find the superclass \mathcal{C}_1 that maximizes $\Delta i(s, t, \mathcal{C}_s)$.
3. The best split s^* maximizes $\Delta i(s, t, \mathcal{C}_s)$:

$$\Delta i(s^*, t, \mathcal{C}_{s^*}) = \max_{s \in \mathcal{S}} \Delta i(s, t, \mathcal{C}_s).$$

Twoing Rule - Advantages

- The key idea of the twoing rule is to transfer the multi-class problem to the two-class (binary response) problem.
- Note that one significant advantage of this approach is that it gives “strategic” splits and informs the user of class similarities. Near the top of the tree structure, this criterion attempts to group together large numbers of classes that are similar in some characteristic; near the bottom of the tree it attempts to isolate single classes.
- The twoing rule has particular use for ordinal responses.
- Apparently, the twoing rule has disadvantage in computational efficiency. However, COROLLARY 4.11 of CART (Brieman, et al., 1984) shows, rather surprisingly, that twoing can be reduced to an overall criterion.

Twoing Rule - Advantages

THEOREM For any node t and split s of t into t_L and t_R , define the twoing criterion function $\psi(s, t)$ by

$$\psi(s, t) = \frac{P(t_L)P(t_R)}{4} \left\{ \sum_{j=1}^J |p(j|t_L) - p(j|t_R)| \right\}^2 .$$

Then the best splits s^* (\mathcal{C}_{1s^*}) is given by the split that maximizes $\psi(s, t)$ and \mathcal{C}_{1s^*} is given by

$$\{j : p(j|t_L^*) \geq p(j|t_R^*)\} ,$$

where t_L^*, t_R^* are the nodes given by the best split s^* .

Outline (Sign-Posting)

1. Categorical Responses
 - (a) Splitting Criterion
 - Goodness-of-split Criterion
 - Chi-square Tests and Twoing Rule
 - (b) **Pruning and Tree-Size Selection**
 - (c) Ordinal Responses
2. Continuous Responses
 - (a) Splitting Criterion
 - Goodness-of-Split Criterion
 - F or t test
 - (b) Pruning and Tree Size Selection
3. Other Extensions - Maximum Likelihood Framework

Pruning and Tree-Size Selection

There are two ways of defining the misclassification cost for a particular node t , depending on the classifying criterion.

1. If the plurality or majority rule is used as classification rule, one may define the total misclassification cost for node t as

$$R(t) = \sum_{i \in t} c(j_t | y_i),$$

where j_t is the assigned class for node t by the plurality rule and $c(i|j)$ denotes the cost for misclassifying a class- j object as class i . Note the $c(i|i) = 0$.

2. Alternative, one may use the Gini index to incorporate

misclassification cost

$$R(t) = \sum_{i \neq j} c(j|i)p(i|t)p(j|t),$$

for $i, j = 1, \dots, J$. Recall that $p(i|t)p(j|t)$ is the probability that misclassifies i as j and vice versa.

The total misclassification cost for tree T is then

$$R(T) = \sum_{t \in \tilde{T}} R(t).$$

The cost-complexity pruning can then be readily extended to the multi-class responses. Furthermore, the misclassification cost, after validation, can be used to evaluate the subtrees and select the best tree size.

Ordinal Responses

- For ordinal responses, usually the misclassification cost does the special handling, e.g., assigning class 6 as 1 costs more than assigning class 2 as 1 since class 2 is closer to class 1 than class 6. To incorporate the ordering, one sets $c(1|6) > c(1|2)$.
- The twoing rule can be naturally modified to incorporate ordinal responses. It is natural to consider the *ordered twoing criterion* given by

$$\psi(s, t) = \max_{\mathcal{C}_1} \Delta i(s, t, \mathcal{C}_1),$$

where \mathcal{C}_1 and \mathcal{C}_2 are partitions of $\mathcal{C} = \{1, \dots, J\}$ into two superclasses restricted by the condition that they be of the form

$$\mathcal{C}_1 = \{1, \dots, j_1\} \quad \text{and} \quad \mathcal{C}_2 = \{j_1 + 1, \dots, J\},$$

and $j_1 = 1, \dots, J - 1$.

Outline (Sign-Posting)

1. Categorical Responses

(a) Splitting Criterion

- Goodness-of-split Criterion
- Chi-square Tests and Twoing Rule

(b) Pruning and Tree-Size Selection

(c) Ordinal Responses

2. **Continuous Responses**

(a) Splitting Criterion

- Goodness-of-Split Criterion
- F or t test

(b) Pruning and Tree Size Selection

3. Other Extensions - Maximum Likelihood Framework

Continuous Responses - Splitting

Now we consider regression trees with continuous response.

- A natural choice of node impurity for node t is the within-node variation of the response

$$i(t) = \sum_{i \in t} (y_i - \bar{y}_t)^2,$$

where \bar{y}_t is the average of y_i 's within node t .

- To evaluate a split that bisects a node t into its two child nodes t_L and t_R , we maximize the goodness-of-split criterion

$$\Delta i(s, t) = i(t) - \{i(t_L) + i(t_R)\}.$$

Note that, unlike decision trees, the goodness-of-split criterion does not need weights.

Splitting using F Tests

- One may also use F test to evaluate binary splits

$$F = \frac{SSB/1}{SSW/(n-2)} \sim F(1, n-2),$$

where $SSW = i(t_L) + i(t_R)$ is the total within-node variation and $SSB = n_L \cdot (\bar{y}_L - \bar{y})^2 + n_R \cdot (\bar{y}_R - \bar{y})^2$ is the total between-node variation. Equivalently, the two-sample t test can be used.

- It can be easily shown that maximizing the goodness-of-split criterion is equivalent to maximizing the F test statistic, both also equivalent to maximizing the likelihood ratio test statistic. However, the F test statistic has more advantages because it has a distribution to reference.

- The logworth associated with F is $-\log \Pr \{F(1, n - 2) > F\}$. The best split maximizes the logworth. This is particularly useful to evaluate multi-way splits, in which case

$$F = \frac{SSB/(m - 1)}{SSW/(n - m)} \sim F(m - 1, n - m)$$

for an m -way split.

- The logworth allows for Bonferroni type adjustment.

One Problem with Splitting Statistic

Recall that for the F test to be valid, there are three statistical assumptions:

1. Observations are independent. (independence)
2. Observations within each child node follow a normal distribution. (normality)
3. All observations have the same variance σ^2 . (homoscedasticity)

The independence assumption is usually reasonable to stand. For large observations, the normality assumption is not a big deal. However, the violation of homoscedasticity could seriously bias the split selection.

Detecting and Treating Heteroscedasticity

In order to detect heteroscedasticity, one may plot the residuals versus predicted response and look for systematic patterns. Also, there are various statistical tests available for checking heteroscedasticity such as score tests.

There are two treatments to this problem.

1. Transform the response. For example the logarithm transformation often helps stabilize the variance.
2. Use weighted least squares:

$$i(t) = \sum_{i \in t} w_i \cdot (y_i - \bar{y}_t)^2.$$

Outline (Sign-Posting)

1. Categorical Responses
 - (a) Splitting Criterion
 - Goodness-of-split Criterion
 - Chi-square Tests and Twoing Rule
 - (b) Pruning and Tree-Size Selection
 - (c) Ordinal Responses
2. Continuous Responses
 - (a) Splitting Criterion
 - Goodness-of-Split Criterion
 - F or t test
 - (b) **Pruning and Tree Size Selection**
3. Other Extensions - Maximum Likelihood Framework

Pruning and Tree Size Selection

- To prune, we can make use of $i(t)$ to define the tree cost as

$$R(T) = \sum_{t \in \tilde{T}} i(t) = \sum_{t \in \tilde{T}} \sum_{i \in t} (y_i - \bar{y}_t)^2$$

and then form the cost-complexity measure.

- In size selection, the prediction mean squared error (PMSE) is used

$$PMSE = \sum_{t \in \tilde{T}} \sum_{i \in t \cap \mathcal{L}_2} (y_i - \bar{y}_t)^2,$$

where y_i 's are responses in the test sample \mathcal{L}_2 and \bar{y}_t 's are the node averages computed using the learning sample \mathcal{L}_1 .

Other Extensions - Maximum Likelihood

- Noticing that both Entropy and the F test have correspondence with the likelihood ratio test.
- The Maximum Likelihood (ML) framework is general and flexible to handle various of response types such as count data, censored failure times. In particular, the likelihood ratio test (LRT) statistic can be used to split data. The efficient score test is often a favorable choice since its computation only requires evaluation of the null model. Furthermore, the deviance score can be used for cost and AIC or BIC can be used to select the best tree model. See e.g. Su, Wang, and Fan (JCGS, 2004).
- SAS EM does not implement tree models for other types of responses. See Splus or R implementation - `rpart` for tree models for count and survival data.