Seminararbeit

Conditional tress

Christoph Molnar

Supervisor:
Stephanie MÃ¶st

1. June 2012
Department of Statistics
University of Munich

Contents

1. Introduction

Recursive partitioning is a powerful yet simple tool in predictive and explanatory statistics. Models build by partitioning take on the form of decision trees, which makes the approach easy to understand for everyone without understanding the algorithm behind. The model partitiones the data to reconstruct the relationship

$$Y = f(X)$$

, where Y is called the response variable, which depends on a function f of the covariates matrix X. Partitioning can be done with many different approaches and therefore the landscape of algorithms is very vivid. Recursive partitioning suffers from different problems, some of which are already solved by some approaches:

- Overfitting
- High variance
- Variable selection bias
- Heuristic approach, lack of statistical model behind
- Restriction on possible measurement scales of Y and X

Most approaches suffers at least of one of those problems. CART (Classification And Regression Trees) is a famous and widely used example of partitioning algorithms. Let us take a closer look at the problems with CART as example and how the conditional trees approach solves them.

If a tree is allowed to grow full length, pathological split could happen and if the covariate space is large enough we would end up with a tree, which contains only one observation in each terminal node. This tree would very likely be **overfitting** on the training data and deliver very bad results on new data. Approaches to avoid this problem are techniques called early stopping and pruning. Early stopping forces trees to stop growing when some criterion is not fullfilled. This criterion could be a minimum number of observations in a node. Pruning let's the tree grow at first and prunes the leafs back afterwards. The CART algorithm uses both early stopping and pruning to avoid overfitting.

As the tree strongly depends on the first splits, different variables at for the first split can yield two structurally different trees. Therefore trees (also CART) are sensitive to variance in the data and resulting trees themselves have a **high variance**.

Exhaustive search procedures as used by the CART algorithm tend to choose variables with more possible split points (**variable selection bias**). This is a problem of multiple comparison. Covariates with many possible splits are searched more often for the best split.

The next split is just a heuristic, as the algorithm only searches for the next best split (like CART). Conditional trees algorithm measures the association and uses the covariate with the strongest association with the response variable. The algorithm is embeded in a well-defined framework of hypothesis.

In the family of partitioning algorithm, the CART algorithm is one of the more powerful ones, as it can do regression as well as classification. Many other algorithm are restricted to one of the both tasks. Though, CART still lacks support for other scales of X and Y. Examples are: ordinal regression and censored data, just to name two.

To achieve the above goals, the conditional tree algorithm embeds all decisions into statistical hypothesis tests. The tests are based on conditional inference, i.e. permutation tests. The test statistic used is based on the work of [STRASSER AND WEBER].

2. Algorithm

Steps 1) and 2) of the algorithm are completed. The covariate with the strongest association is chosen for the next partition step. Every covariate (which is not binary) has more than one possible split. To determine where to split, a criterion which measures the goodness of the split has to be applied. The CART algorithm uses Gini for classification and sum of squares for regression. Both of the criteria could be used by the Conditional Tree algorithm as well, but the approach is different. Because of the different types of possible regression-/classification - models (categorial, ordinal, numeric, censored, ...) are more general approach is suitable. Again the test statistic framework from Strasser and Weber [CITE] can be used. A special linear test statistic, of the same kind, which is used for the stop criterion and variable selection, can be used. The formula is:

$$T_j^A(L_n, w) = vec\left(\sum_{i=1}^{n} w_i I(X_{ji} \in A) \cdot h(Y_i, (Y_1, \ldots, Y_n))^T\right)$$

The difference to the test statistic for the association test is the transformation of X. We only look at the different partitions of X. Therefore the scale of X is not of any interest anymore, but the partition which emerges by a certain split point. An appropriate function to capture only the difference in the partition, the transformation of X is the indicator function. It is defined as: $I(X_{ji} \in A) = \begin{cases} 1, & X_{ij} \in A \\ 0 & X_{ij} \notin A \end{cases}$, where A is one possible partition. This results in the statistic

$$c_{max}(\mathbf{t}, \mu, \Sigma) = max_k \left| \frac{(\mathbf{t}^A - \mu)_k}{\sqrt{(\Sigma)_{kk}}} \right|$$

For regression and the identitiy function for the influence function h, the test statistic is the following:

$$c_{max}(\mathbf{t}, \mu, \Sigma) = max_k \left| \frac{(\mathbf{t}^A - \mu)_k}{\sqrt{(\Sigma)_{kk}}} \right| \tag{2.1}$$

$$= \left| \frac{\sum_{i=1}^{n} w_i I(X_{ji} \in A) \cdot Y_i - \sum_{i=1}^{n} w_i I(X_{ji} \in A) \cdot n_{node}^{-1} \sum_{i=1}^{n} w_i Y_i}{\sqrt{\frac{n_{node}}{n_{node}-1} \frac{1}{n_{node}} \sum_{i=1}^{n} w_i (Y_i - \bar{Y}_{node})^2 \cdot \sum_{i=1}^{n} w_i I(X_{ji} \in A)^2 - \frac{1}{n_{node}-1} \frac{1}{n_{node}} \sum_{i=1}^{n} w_i (Y_i - \bar{Y}_{node})^2 \cdot \Bigg(\sum_{i=1}^{n} w_i I(} \tag{2.2}$$

$$= \left| \frac{\sum_{i:x_{ij} \in A} Y_i - n_A \bar{Y}_{node}}{\sqrt{\frac{1}{n_{node}-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_{node})^2 n_A (1 - \frac{n_A}{n_{node}})}} \right| \tag{2.3}$$

$$= n_A \left| \cdot \frac{\bar{Y}_A - \bar{Y}_{node}}{\sqrt{\frac{1}{n_{node}-1} \sum_{i \in node} (Y_i - \bar{Y}_{node})^2 \cdot n_{node} (\frac{n_A}{n_{node}})(1 - (\frac{n_A}{n_{node}}))}} \right| \tag{2.4}$$

$$= n_A \left| \frac{\bar{Y}_A - \bar{Y}_{node}}{\sqrt{Var(Y) \cdot Var(Z)}} \right| \tag{2.5}$$

$$\tag{2.6}$$

with $Z \sim B(n_{node}, \pi = \frac{n_A}{n_{node}})$ Can be interpretated as the probability that z observations would be assigned to A if the process of aissigning would be random with probability $\frac{n_A}{n_{node}}$. Is maximal for $n_A = \frac{n_{node}}{2}$. The closer $\frac{n_A}{n_{node}}$ to 0.5 the bigger is c (assuming Y_A stays the same). Thus the test statistic favors bigger partitions. $n_{node} := \sum_{i=1}^{n} w_i$ n_{node}: Number of observation in node $\bar{Y}_{node}$: Mean of Y in node n_A: Number of observations in partition A $\bar{Y}_A$: Mean of Y in A

Additional stopping criteria like stopping when the resulting partitions would become to small can be implemented by restricting the searched split points.

3. Example bodyfat

For regression and the identitiy function for the influence function h, the test statistic is the following:

$$c_{max}(\mathbf{t}, \mu, \Sigma) = max_k \left| \frac{(\mathbf{t}^A - \mu)_k}{\sqrt{(\Sigma)_{kk}}} \right| \tag{3.1}$$

$$= \left| \frac{\sum_{i=1}^{n} w_i I(X_{ji} \in A) \cdot Y_i - \sum_{i=1}^{n} w_i I(X_{ji} \in A) \cdot n_{node}^{-1} \sum_{i=1}^{n} w_i Y_i}{\sqrt{\frac{n_{node}}{n_{node}-1} \frac{1}{n_{node}} \sum_{i=1}^{n} w_i (Y_i - \bar{Y}_{node})^2 \cdot \sum_{i=1}^{n} w_i I(X_{ji} \in A)^2 - \frac{1}{n_{node}-1} \frac{1}{n_{node}} \sum_{i=1}^{n} w_i (Y_i - \bar{Y}_{node})^2 \cdot}} \right. \tag{3.2}$$

$$= \left| \frac{\sum_{i:x_{ij} \in A} Y_i - n_A \bar{Y}_{node}}{\sqrt{\frac{1}{n_{node}-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_{node})^2 n_A (1 - \frac{n_A}{n_{node}})}} \right| \tag{3.3}$$

$$= n_A \left| \cdot \frac{\bar{Y}_A - \bar{Y}_{node}}{\sqrt{\frac{1}{n_{node}-1} \sum_{i \in node} (Y_i - \bar{Y}_{node})^2 \cdot n_{node} (\frac{n_A}{n_{node}})(1 - (\frac{n_A}{n_{node}}))}} \right| \tag{3.4}$$

$$= n_A \left| \frac{\bar{Y}_A - \bar{Y}_{node}}{\sqrt{Var(Y) \cdot Var(Z)}} \right| \tag{3.5}$$

$$\tag{3.6}$$

with $Z \sim B(n_{node}, \pi = \frac{n_A}{n_{node}})$ Can be interpretated as the probability that z observations would be assigned to A if the process of aissigning would be random with probability $\frac{n_A}{n_{node}}$. Is maximal for $n_A = \frac{n_{node}}{2}$. The closer $\frac{n_A}{n_{node}}$ to 0.5 the bigger is c (assuming Y_A stays the same). Thus the test statistic favors bigger partitions. $n_{node} := \sum_{i=1}^{n} w_i$ n_{node}: Number of observation in node $\bar{Y}_{node}$: Mean of Y in node n_A: Number of observations in partition A $\bar{Y}_A$: Mean of Y in A

Additional stopping criteria like stopping when the resulting partitions would become to small can be implemented by restricting the searched split points.

A. Computational details

B. Digital Appendix

Bibliography

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.

T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.

H. Strasser and C. Weber. On the asymptotic theory of permutation statistics. 1999.

List of Figures