Seminararbeit

Conditional tress

Christoph Molnar

Supervisor:
Stephanie MÃ¶st

1. June 2012
Department of Statistics
University of Munich

Contents

1. Introduction

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: $I(X_{ji} \in A) = \begin{cases} 1, & if X_{ij} \in A \\ 0 & X_{ij} \notin A \end{cases}$, 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.

A. Computational details

B. Digital Appendix

List of Figures