Decision Trees II

Conditional Inference Trees

Introduction

Extending Decision Trees

- Conditional Inference Trees (Hothorn et al. 2006)
  - Addresses selection bias for variables with many potential split points
  - Separates variable and split point decision
  - Variable selection and stopping criterion based on statistical test
- Model-based Recursive Partitioning (Zeileis et al. 2008)
  - Connects recursive partitioning with fitting parametric (regression) models
  - Approach to fitting "homogeneous" models in tree nodes

Conditional Inference Trees

Algorithm 1: Grow a CTREE

Parameter : p-value threshold
Initialization: Assign training data to root node

Perform permutation tests for each covariate (H_0: Y and X_j independent);
if *minimum p-value exceeds threshold* **then**
  end splitting (global H_0 not rejected);
else
  select covariate with strongest association (smallest p-value);
  find the optimal split point for the selected variable;
  split node into two subnodes at this split point;
  for *each node of the current tree* **do**
    continue tree growing process;
  end
end

Conditional Inference Trees

General test statistic for **variable selection** with weights w, transformation g and influence function h:

$$\mathbf{T}_j = vec\left(\sum_{i=1}^{n} w_i g_j(X_{ij}) h(Y_i, (Y_1, \ldots, Y_n))^T\right)$$

Continuous case: $\mathbf{T}_j = \sum_{i \in node} X_{ji} Y_i$

Standardized test statistic:

$$c(\mathbf{t}, \mu, \Sigma) = \max_{k=1,\ldots,pq} \left| \frac{(\mathbf{t} - \mu)_k}{\sqrt{(\Sigma)_{kk}}} \right|$$

Continuous case: $c \propto$ Pearson's r

Conditional Inference Trees

Permutation tests

- Unconditional/ parametric tests involve distribution assumptions
- Conditional tests: Consider distribution of test statistic given the observed data
- Idea: Infer null distribution from randomly shuffled data
- General procedure
  1. Calculate test statistic c_{j0}
  2. For all possible permutations
     1. Permute values of variables
     2. Calculate test statistic c
  3. Count number of c which are more extreme than c_{j0}, $n_{extreme}$
  4. $p = \frac{n_{extreme}}{n_{permutations}}$

Conditional Inference Trees

General test statistic for **split point selection** with A denoting a possible partition:

$$\mathbf{T}_{j*}^{A} = vec\left(\sum_{i=1}^{n} w_i I(X_{j*i} \in A) \cdot h(Y_i, (Y_1, \dots, Y_n))^T\right)$$

Continuous case: $\mathbf{T}_{j*}^{A} = n_A \bar{Y}_A$

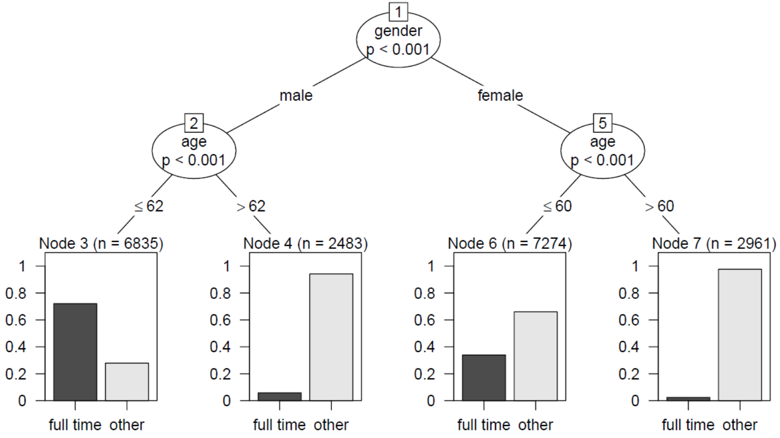
Search for partition A which maximizes the (standardized) test statistic c:

$$A^* = argmax_A c(\mathbf{t}_{j*}^{A}, \mu_{j*}^{A}, \Sigma_{j*}^{A})$$

Continuous case: Maximize difference between $\bar{Y}_A$ and $\bar{Y}_{node}$

Conditional Inference Trees

Figure: Conditional Inference Tree of employment status with SOEP (2008) data



Kopf et al. (2010)