Logic Regression

Ingo Ruczinski
Department of Biostatistics
Johns Hopkins University

Email: ingo@jhu.edu

http://biosun01.biostat.jhsph.edu/~iruczins

With Charles Kooperberg and Michael LeBlanc, FHCRC

Introduction and Motivation

$X_1, \ldots, X_k$ are 0/1 (False/True) predictors.

$Y$ is a response variable.

Fit a model $g(F(Y)) = \beta_0 + \sum_{j=1}^{t} \beta_j \times L_j$, where $L_j$ is a Boolean combination of the covariates, e.g. $L_j = (X_1 \lor X_2) \land X_4^c$.

Determine the logic terms $L_j$ and estimate the $\beta_j$ simultaneously.
An equivalent representation of \((X_1 \land X_2^c) \lor (X_3 \land (X_1^c \lor X_4))\) is the following:

```
1 2 3 or
and and
```

This is a Logic Tree!

## Comparison to Decision Trees

A Decision Tree (CART) is something different!
An Example

\[
\text{logit(death)} = -9.01 + 0.06 \times \text{age} + 1.07 \times L
\]

The Move Set

<table>
<thead>
<tr>
<th>Possible Moves</th>
<th>Initial Tree</th>
<th>Prune Branch</th>
<th>Split Leaf</th>
<th>Delete Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>and 1</td>
<td>or 2 3</td>
<td>or 1 3</td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td>or 1 3</td>
<td>and 2 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c)</td>
<td>and 1</td>
<td>or 2 3</td>
<td>and 1 2</td>
<td></td>
</tr>
<tr>
<td>(d)</td>
<td>and 1</td>
<td>or 2 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(e)</td>
<td>or 1 3</td>
<td>and 2 3</td>
<td>and 3 6</td>
<td></td>
</tr>
<tr>
<td>(f)</td>
<td>and 1</td>
<td>or 2 3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alternate Leaf

Alternate Operator

Grow Branch

Delete Leaf
Simulated Annealing for Logic Regression

We try to fit the model \( g(F(Y)) = \beta_0 + \sum_{j=1}^{t} \beta_j \times L_j \).

- Select a scoring function (RSS, log-likelihood, \ldots).
- Pick the maximum number of Logic Trees.
- Pick the maximum number of leaves in a tree.
- Initialize the model with \( L_j = 0 \) for all \( j \).
- Carry out Simulated Annealing Algorithm:
  - Propose a move.
  - Accept or reject the move, depending on scores and temperature scheme.

Another Example

The Cardiovascular Health Study
(Fried et. al., Annals of Epidemiology, 1991).

- The Cardiovascular Health Study (CHS) is a study of coronary heart disease and stroke in elderly people.
- Between 1989 and 1993, 5888 subjects over the age of 65 were recruited in four communities in the United States.
- During 1992 and 1994, a subset of these patients underwent an MRI scan.
- For 3647 CHS participants, MRI detected strokes (infarcts bigger than 3mm that led to deficits in functioning) were recorded as entries into a 23 region atlas of the brain.
- The mini-mental state examination is a brief screening test for dementia. The response \( Y \) is a variable derived by transforming the mini-mental score.

We investigated models of the form \( Y = \beta_0 + \beta_1 \times L_1 + \cdots + \beta_p \times L_p + \epsilon \).
A Global Randomization Test of Association

Example Cont.

- **X**
- **Y**
- **Perm(Y)**

![Graph showing variables X, Y, and Perm(Y)]

- **a** deviance of lowest scoring model
- **b** deviance of null model

![Histogram showing distribution of scores]
Cross-Validation

Cross-Validation Cont.
A Sequential Randomization Test for Model Size

Randomization Test Cont.
Sequential Randomization Test for 2 Trees:

\[
\begin{array}{cccccc}
X & T1 & T2 & Y & Perm(Y) \\
0 & 0 & 0 & 0 & \\
0 & 1 & 1 & 0 & \\
1 & 0 & 0 & 0 & \\
1 & 1 & 1 & 0 & \\
\end{array}
\]

Results

The model we found was \[ Y = 1.96 + 0.36 \times L, \] with the following Logic Tree:
logit(affected) = $\beta_0 + \beta_1 \times \text{ENV}_1 + \beta_2 \times \text{ENV}_2 + \beta_3 \times \text{GENDER} + \sum_{i=1}^{K} \beta_{i+3} \times L_i$

L_1 = \text{and} \quad L_2 = \text{or} \quad L_3 = \text{or}

GAW12 Cont.