Exploring Interactions in Genomic Data Using Logic Regression

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins University

Email: ingo@jhu.edu

http://biostat.jhsph.edu/∼iruczins

With Charles Kooperberg and Michael LeBlanc, FHCRC

Abstract

Logic Regression is a regression methodology intended for situations where most of the predictors are binary. Logic Regression searches for Boolean combinations of predictors in the entire space of such combinations, while being completely embedded in a regression framework where the quality of the models is determined by the objective function of the regression class. Logic Regression stands apart from most methods in the computer science and machine learning literature in that it uses general Boolean expressions, a non-greedy search algorithm, and that it works in any regression framework. As SNP data is effectively binary (recorded as common/variant nucleotide or two 'dummy variables' coding for the number of variant alleles), Logic Regression is particularly useful for such data. Other genomic applications include data indicating chromosomal deletions in cancer patients, and amino acid variations associated with drug resistance.

Motivation

Lucek and Ott (1997):

"Current methods for analyzing complex traits include analyzing and localizing disease loci one at a time. However, complex traits can be caused by the interaction of many loci, each with varying effect."

"... patterns of interactions between several loci, for example, disease phenotype caused by locus A and locus B, or A but not B, or A and (B or C), clearly make identification of the involved loci more difficult. While the simultaneous analysis of every single two-way pair of markers can be feasible, it becomes overwhelmingly computationally burdensome to analyze all 3-way, 4-way to N-way 'and' patterns, 'or' patterns, and combinations of loci."

Logic Regression

Logic regression tackles the problem stated by Lucek and Ott.

Assume that $X_1, \ldots, X_k$ are binary (0/1) predictors and Y is a response variable (Y does not have to be binary).

Logic regression models are of the form

$$g(E(Y)) = \beta_0 + \sum_{j=1}^{t} \beta_j \times L_j,$$

where L_j is a Boolean combination of the covariates, for example $L_j = (X_1 \vee X_2) \wedge X_4^c$.

The task is to find the best logic regression models, i. e. to determine the logic terms L_j and estimate the β_j simultaneously.

Fine print: There are options to include continuous variables in the above model, as separate variables and in the interaction term. Also, other models (not just generalized linear models) can be implemented as long as a scoring function can be defined. One such example is the Cox proportional hazards model.

An Application

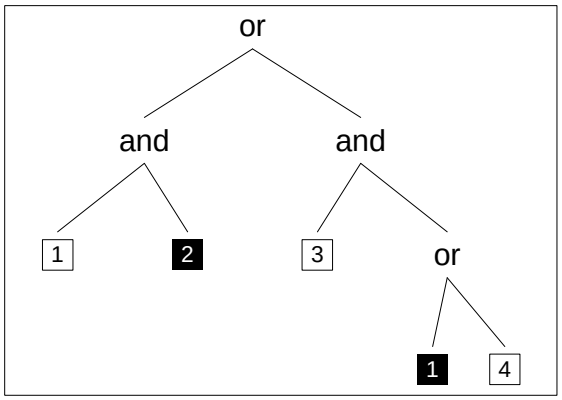
[With Tony Alberg]

Background: Various studies have shown that individuals with non-melanoma skin cancers (NMSC) are at increased risk for subsequent cancers, such as lip and skin cancer. Current estimates are that individuals with a history of NMSC have an about 30% higher risk of developing subsequent cancers compared to people without a history of NMSC. The reasons for this phenomenon are not exactly known, but it is hypothesized that, among other factors, a decreased ability to adequately repair DNA damage plays a significant role. Further, some studies have established an association between SNPs in DNA repair genes that result in variant amino acid substitutions, and an DNA repair capacity. In particular, there is some evidence that failure to adequately repair DNA damage from UV or ionizing radiation in people who develop NMSC and second primary cancers might be due to some genetic factors underlying DNA repair. To investigate the relationship between DNA repair genotypes, NMSC, and the occurrence of subsequent cancers, Anthony Alberg's group in the Department of Epidemiology at the Johns Hopkins University proposed a prospective cohort study, using data from the CLUE II specimen bank established in 1989. Dr. Alberg suggested to screen for 28 polymorphisms in 19 candidate genes involved in five major DNA repair pathways.

Application: Two different types of analysis are carried out. An evaluation of differences in DNA repair genotypes with respect to NMSC prevalence in 1989 is be carried out first (cross-sectional). Then, following the individuals without cancer history in 1989, the association of genotypes to the risk of developing NMSC will be evaluated (prospective). An estimated 493 patients will have developed the disease by 2007. The CLUE II cohort has been followed since 1989, and environmental variables such as smoking status and history, blood pressure, cholesterol, height, weight, family history of cancer, medication history, diet, exercise and sleep habits were recorded at various points in time. Adjusting for some of these potential confounders, the predictor of interests are the SNPs, and in particular, their interactions.

Logic Trees

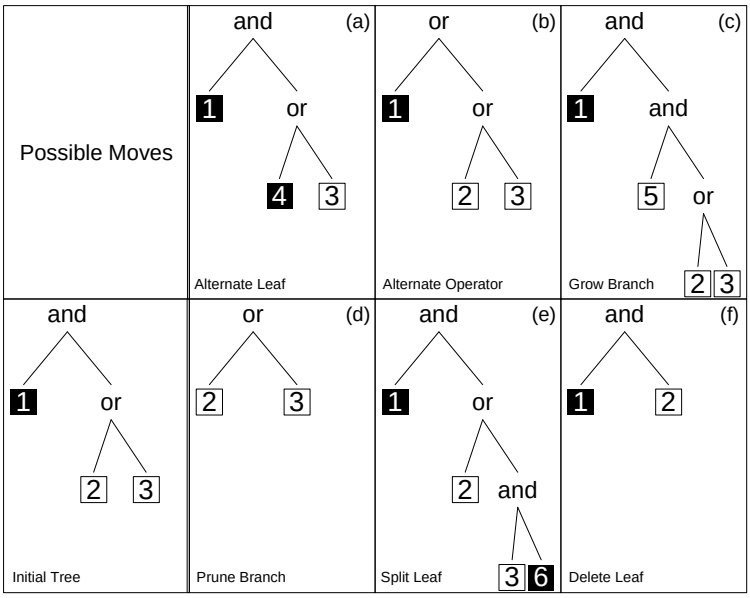
The Boolean terms are coded in a tree format. For example, an equivalent representation of $(X_1 \wedge X_2^c) \vee (X_3 \wedge (X_1^c \vee X_4))$ is:



This is a Logic Tree!

These logic trees are very different from decision (CART) trees. The evaluation of the tree is in a bottom-up fashion, compared to top-down in CART trees. Also, classification trees are in disjunctive normal form, while Logic trees represent general Boolean terms.

The Move Set



Using this logic tree representation, it is possible to obtain any other logic tree by a finite number of operations such as growing of branches, pruning of branches, and changing of leaves (borrowing from CART terminology). In the left Figure we show the different types of moves that are currently implemented in the software. These moves are the basis for the probabilistic search algorithm.

Simulated Annealing for Logic Regression

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

- Select a scoring function (RSS, log-likelihood, ...).
- 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 the Simulated Annealing Algorithm:
  - Propose a move.
  - Accept or reject the move, depending on the scores and the temperature.

Model Selection

Using model selection in addition to a stochastic model building strategy is of critical importance, as the logic tree with the best score typically over-fits the data. A variety of methods of model selection using cross-validation and randomization tests exist. If we have an abundance of data, we typically fit our models on one part of the data, and validate them on the remainder.

Model Size

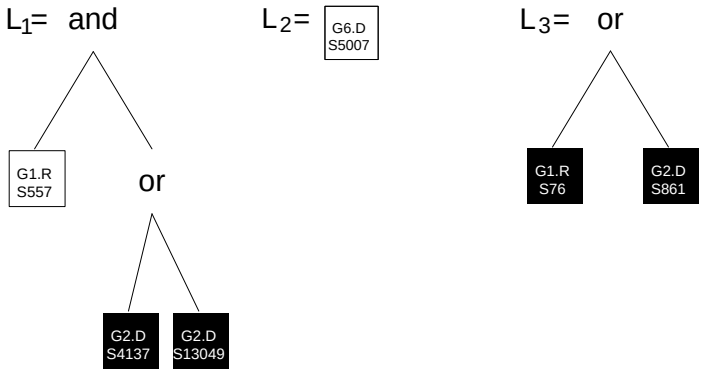
To carry out the model selection techniques mentioned above, we have to define the size of models. Clearly there are many possibilities how to do this. We currently use the total number of leaves for a fixed number of trees as the size of the model.

Genetic Analysis Workshop GAW 12

For the GAW12 workshop the data were simulated under the model of a common disease. Prevalence increased with age, and the disease was more common in females than in males. A total of 50 datasets were generated, each of which consisted of 23 extended pedigrees with 1497 individuals (1000 living). Each living subject had data on affection status, age at last exam, age at onset if affected, marker genotype data for a 1cM autosomal genome screen, and sequence data on six candidate genes, as well as five quantitative traits and two environmental factors. We applied Logic Regression to the sequence data, using one copy of the data set as training data and another copy as test data. In the sequence data, there were a combined total of 694 SNPs on six genes with at least 2% mutations, encoded as 1388 binary predictors. As response we used the affected status. The model selection yielded a model with three trees and six leaves, shown in the next Figure with the Logic Regression model.

Since the GAW data were simulated, we know the correct solution. As it turned out, the Logic Regression algorithm picked exactly those mutational sites on gene 1 and gene 6 that were used in generating the data, and a number of sites on gene 2, where there were multiple mutational hits. It also detected the correct interaction between genes 1 and 2, and it did not include any spurious sites.

$$\text{logit(affected)} = \beta_0 + \beta_1 \times \text{ENV}_1 + \beta_2 \times \text{ENV}_2 + \beta_3 \times \text{GENDER} + \sum_{i=1}^{K} \beta_{i+3} \times L_i$$



Here Gi.D.Sj refers to site j on gene i, using dominant coding, i.e. Gi.D.Sj=1 if at least one variant allele exist. Similarly, Gi.R.Sj refers to site j on gene i, using recessive coding.

Software

The software is freely available from the Logic Regression website at http://bear.fhcrc.org/∼ingor/logic and can be downloaded as a package for R or S-Plus. The core of the Logic Regression code is in Fortran77, which is called by R or S-Plus. Current options include fitting one (large) Logic Regression model, fitting Logic Regression models of pre-specified sizes, carrying out cross-validation, and various randomization tests for model selection. Plotting functions to display the results are available as well. All functions have extensive help files. Currently the Logic Regression methodology has scoring functions for linear regression (residual sum of squares), logistic regression (deviance), classification (misclassification), and proportional hazards models (partial likelihood). A feature of the Logic Regression methodology is that it is possible to extend the method to write ones own scoring function if this is necessary.