Interactions and Variable Importance in Genomic Data

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Email: ingo@jhu.edu

Joint work with Charles Kooperberg, Fred Hutchinson Cancer Research Center, Seattle.

The publications and software used for this poster are available at http://biostat.jhsph.edu/∼iruczins

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

- is a regression methodology intended for situations where most of the predictors are binary,
- 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,
- 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,
- tackles the problem stated by Lucek and Ott!

Logic Regression

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 defi ned. One such example is the Cox proportional hazards model.

Applications

Most of the applications are in studies trying to reveal the association between single nucleotide polymorphisms (SNPs) and disease. SNP data are effectively binary, recorded as common and variant nucleotide or two 'dummy variables' coding for the number of variant alleles. Outcomes of interest in current SNP association studies include for example bladder, breast, and skin cancer.

Other applications with genomic data involve

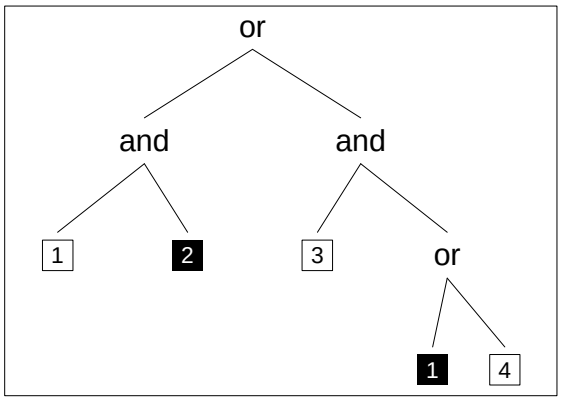
- amino acid variations and drug resistance,
- chromosomal deletions in cancer patients,
- haplotype analyses.

Applications with non-genomic data concern for example

- risk factors for lymphoma patients,
- risk factors for frailty and death in the elderly.

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:

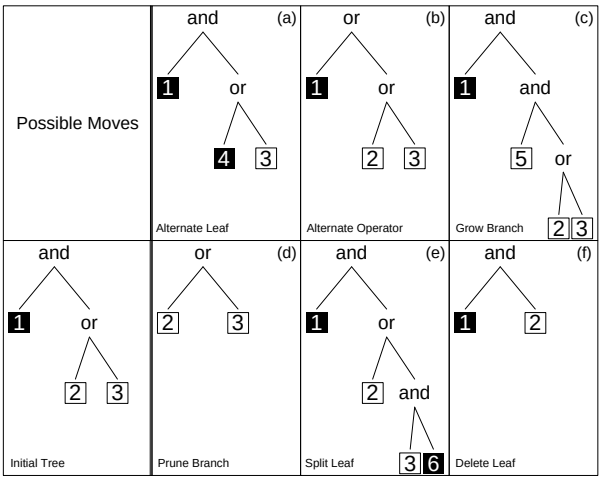


This is a logic tree!

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, classifi cation 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 (using the CART terminology). In the Figure on the right we show the different types of moves that are currently implemented in the software. These moves define a neighborhood system for the logic trees, which is the basis for the probabilistic search algorithm (simulated annealing) introduced on the next panel.



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

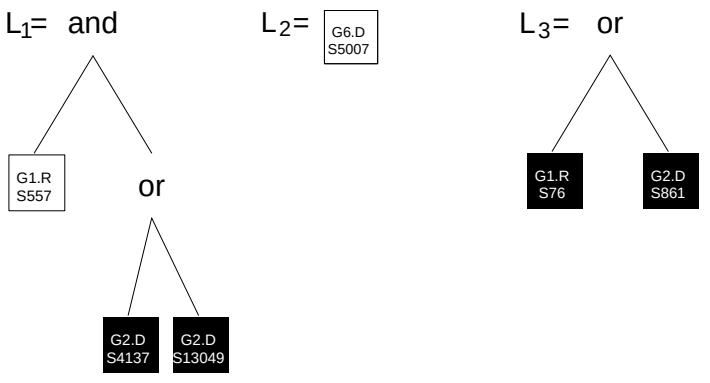
Using model selection in addition to a stochastic model building strategy is of critical importance, as the logic tree with the best score typically over-fits the data. A variety of methods of model selection using cross-validation and randomization tests exist. If we have an abundance of data, we typically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

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



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. ENV denotes environmental factors.

$\longrightarrow$ 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.

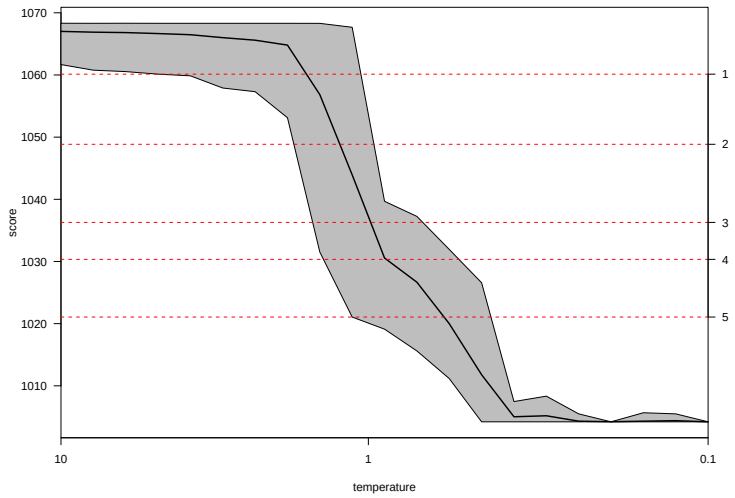
Multiple Models

Often we are interested in more than just one model. For example, in the GAW 12 data several SNPs were almost identical, and therefore there are many plausible models. We can get a list of good scoring models and measures of variable importance by taking advantage of some of the properties of the simulated annealing algorithm.

Let γ_S be the score of a certain state S.

- We use the acceptance function $\alpha(\gamma_{old}, \gamma_{new}, t) = \min\{1, \exp([\gamma_{old} - \gamma_{new}]/t)\}$
- If we keep the temperature constant, this defi nes a homogeneous Markov chain.
- We constructed the move set to be irreducible and aperiodic, therefore each homogeneous Markov chain has a limiting distribution $\pi_t(S)$.

Multiple Models



In this example, the model selection techniques picked a model of size 4. In other words, if we run a homogeneous chain at temperature just below 1 (“crunch time”), we will visit many good scoring, alternate models!

Variable Importance

We can also keep track which variables were in those models we visited, and use the proportion of models that included a certain variable as the measure of "variable importance". The same works for interactions.

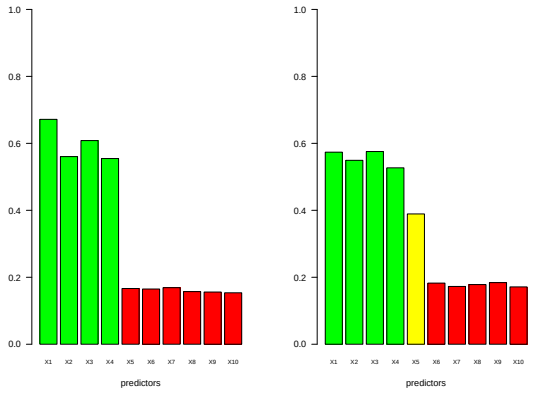
A quick simulation study:

- Simulate 10 binary predictors $X_1, \ldots, X_{10}$.
- Let $Y = 5 + 1 \times L(X_1, X_2, X_3, X_4) + \epsilon, \quad \epsilon \sim \text{N}(0,1)$.
- Run a homogeneous Markov chain during "crunch time" for two separate cases:

  Case 1: All X are independent.

  Case 2: All X are independent, except X_4 (in the signal) and X_5 (not in the signal), which are heavily correlated.

The simulation study:



A SNP example: