Topics

• **glmpath**: R package for fitting the $\ell_1$ regularization path for generalized linear models and the Cox PH model.

• **stepPLR**: a regularized forward-stepwise logistic regression package for fitting gene/gene and gene/environment interaction models in studies of genetic diseases.

• **gbm**: Greg Ridgeway’s package for fitting gradient boosted models, including logistic regression.
Background

Logistic regression is a heavily used tool in statistics. Here are some applications we have been involved in:

• Risk modeling — e.g. risk factors for heart disease, risk of insurance fraud of payment default on credit card.

• Prediction models, e.g. QSAR (Quantitative Structure-Activity Relationship). Use compound’s chemical and structural attributes to predict its biological activity, toxicity etc.

• Discovering genes and their interactions in genetic studies of diseases, based on measurements on a large number of SNPs.
Problems with (stepwise) logistic regression

Stepwise logistic regression is popular in the biosciences (SAS community), because it automatically builds a model (with interactions). It has a number of failings:

- It can overfit and perform (predict) poorly.
- With mult-level factors and smallish datasets (genetic studies), empty/sparse cells cause instability.
- It is difficult to assess results — how do we assign a p-value to selected variables?
GLMs with $\ell_1$ (lasso) regularization

For logistic regression we fit the linear model

$$\log \frac{\Pr(Y = 1|X)}{\Pr(Y = 0|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

via regularized maximum likelihood:

$$\max_{\beta} \ell(y; \beta) - \lambda ||\beta||_1$$

This is the lasso (Tibshirani, 1996) for logistic regression, and is well known:

- Does variable selection and shrinkage.
- Smoother path than forward stepwise.
- Select $\lambda$ by AIC, BIC or $k$-fold CV.
**glmpath package**

- Computes *entire* $\ell_1$ path for GLMs and Cox model.
- Uses Predictor-corrector ideas of convex optimization.
- Computes exact path at a sequence of index points $t$.
- Can approximate the junctions (in $\lambda$) where the active set changes.
Fit a logistic regression path, and use 10-fold CV to select $\lambda$.

- For SA heart disease data, red lines indicate fitted values.
- Histograms represent distribution obtained when repeating this procedure on bootstrapped datasets.
- Something like a Bayesian posterior distribution.
- Pairwise plots are useful too; often a selected variable has a correlated cousin, and either could be important.

- Can be used with stepwise logistic regression as well.
• GLMpath paper on my website: Park & Hastie (2006), *An \(l_1\) regularization-path algorithm for generalized linear models.*

• Yuan and Li (2006) extend lasso to deal with groups of variables (e.g. dummy variables for factors):

\[
\max_{\beta} \ell(y; \beta) - \lambda \sum_{m=1}^{M} \gamma_m \|\beta_m\|_2.
\]

The \(l_2\) norm ensures the vector of coefficients \(\beta_m\) are all zero or non-zero together.

• Using predictor-corrector methods we can construct the path for this criterion: Park & Hastie (2006), *Regularization path algorithms for detecting gene interactions.*
Stepwise Penalized Logistic Regression

- In genetic disease studies, we are often faced with modest sample sizes, binary (case-control) responses, and many candidate genes, each a 3-level factor (AA, Aa, aa).
- Usually the wild-type is prevalent, so the factor levels are unevenly populated.
- Two-way interactions have 9 cells, many of which have zero or very low counts.
- Logistic regression does not do well in these scenarios: exact aliasing, high-variance coefficients, convergence problems and overfitting.
• We work with the \( \ell_2 \) penalized log-likelihood:

\[
\max_{\beta} \ell(y; \beta) - \lambda \| \beta \|_2^2,
\]

with typically a small value for \( \lambda \).

• Benefits of the \( \ell_2 \) penalty:
  
  – Can code factors with a saturated set of dummy variables; the natural “summation to 0 constraints” are automatically maintained.
  
  – Coefficients for sparsely populated cells are shrunk more to zero than well-populated cells.
  
  – Coefficients for empty cells are set to zero automatically.

• Can calibrate \textit{effective df} for such a fit from the trace of the weighted ridge operator of the final IRLS step.

• We then run forward stepwise logistic regression using this fitting engine.
• Hierarchy rule: we allow interactions to enter if either of the main effects are present.

• Use AIC/BIC to guide the forward growing/backward deletion.

• Small/large $\lambda$ allows less/more complex models. We estimate $\lambda$ by cross-validation.

• Package stepPLR in R.

• Method described in Park & Hastie (2006), *Penalized logistic regression for detecting gene interactions*. 
Comparisons

- **MDR: Multi-factor Dimensionality Reduction**, Ritchie et al (2001). This is a shotgun approach that examines all first-order, second-order, third-order, etc tables looking for interactions.
  - The case-control polarity for each cell in a table is determined by the observed majority.
  - This coded table is then uses as a classifier, and the tables are ranked via classification performance using 10-fold CV.

Hypertension Study


![ROC curve for PLR: unequal loss](image1)

![ROC curve for PLR: equal loss](image2)
Comments

- MDR authors claim the binary coding of cells drops the dimension to one. In simulation studies we show the effective dimension is much more than 1; $df = 2, 6$ and $17$ for $3, 9$, and $27$ cell tables. *Df are used up in coding the cells.*

- MDR does not do well (loses power) with additive or low-order multiple effects. It has to see these via a bigger multiway table.

- Flextree models perform almost as well as *stepPLR*, but are hard to interpret.

- *stepPLR* delivers a familiar logistic regression model, suitably tamed via regularization, that so far has performed no worse and often better than all competitors we have tried.
Gradient Boosting

- Adaptive nonparametric method for building powerful predictive models (Friedman, 2001). In our context the model has the form

\[
\log \frac{\Pr(Y = 1 | X)}{\Pr(Y = 0 | X)} = \sum_{m=1}^{M} T_m(X),
\]

where each of the terms is a tree.

- Model is fit sequentially by a form of functional gradient decent; a tree is grown to the current gradient of the log-likelihood, is shrunk down heavily by a shrinkage factor, and added to the current model. See our book *Elements of Statistical Learning, HTF (2001)* for details.

- \( M \) is a tuning parameter, much like \( \lambda \) in lasso.
• The *depth* of the trees is another tuning parameter, and determine the maximum interaction-order of the model. E.g., depth-two trees means second-order models.

• Very nice R package *gbm* by Greg Ridgeway.

• We explored the use of GBM for detecting gene-gene and other interactions in genetic disease studies.
Example: Bladder Cancer

- Data from Hung et al. (2004), Cancer Epidemiology, Biomarkers and Prevention, kindly supplied by Dr John Witte, UCSF.
- Fit interaction order 2 and 3 models, with shrinkage = 0.01
- Performance is similar to stepPLR, and could potentially scale up better to larger problems (many loci).
The tuning parameter $M = \# \text{ of Trees}$ determines complexity. Very similar to Lasso/Cosso (see papers on webpage on Forward Stagewise and Monotone Lasso).
Method for assessing overall contribution of each variable to the model.

Does not treat main effects separately from interactions.

Mixes contributions of correlated variables.

Would need refinements to tease out interaction vs main effect contributions.
Partial Dependence Plots

Give average (main effects here) for important variables:
\[ \bar{f}(x_1) = E_{X_2} f(x_1, X_2). \]

Partial Main Effects

- smoke_3
- mpo_n
- mnsod_n2
Second Order Effects

1.0 1.5 2.0 2.5 3.0
−1.0 −0.5 0.0 0.5

mpo_n / smoke_3

Interaction Effects

1.0 1.5 2.0 2.5 3.0
−1.0 −0.5 0.0 0.5

mpo_n / mnsod_n2

Interaction Effects

1.0 1.5 2.0 2.5 3.0
−1.0 −0.5 0.0 0.5

smoke_3 / mnsod_n2

Interaction Effects

1.0 1.5 2.0 2.5 3.0
−1.0 −0.5 0.0 0.5

mpo_n / gstm1_n

Interaction Effects

1.0 1.5 2.0 2.5 3.0
−1.0 −0.5 0.0 0.5

gstm1_n / mnsod_n2

Interaction Effects
- Splits divide three-level factors into two groups — potentially appropriate for genotype data (dominant vs recessive).
- Potentially useful screening tool for large number of SNPs
- Needs further refinement.
All three methods available in R as packages:

- **glmpath** by Mee Young Park & Hastie. Suitable for automatic variable selection and regularization in linear logistic models.

- **stepPLR** by Mee Young Park & Hastie. Suitable for detecting interactions in logistic regression models.

- **gbm** by Greg Ridgeway. Exploratory tool for screening a large number of variables for main effects and interactions.