Background
Objective
Proposed statistical methods
Summary

Logistic and LASSO regression models in predicting cancer event with significant analysis of microarray experiments

Presented by Chongshu Chen
Project Proposal
BST550

Department of Biostatistics and Computational Biology
University of Rochester

OCT 5th, 2015

Table of Contents

1. Background
2. Objective
3. Proposed statistical methods
   - Logistic regression model
   - Lasso regression model
4. Summary

Key facts

- A gene is the basic physical and functional unit of heredity
- Genes vary in size from a few hundred DNA bases to more than 2 million bases in humans.
- The Human Genome Project has estimated that humans have between 20,000 and 25,000 genes.

The Path to Genomic Medicine

Human Genome Project (2003)

Sequence More Genomes

Interpret Genome Data

Identify Functions

Realization of Genomic Medicine (20XX?)

Data Links: Gene Expression Omnibus (GEO)

NCBI

DATASET BROWSER (CURATED)

GEO Gene Expression Omn

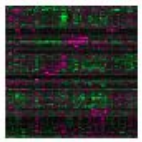
Search for GDS807[ACCN] | Search | Clear | Show All | Advanced Search

DataSet Record GDS807: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Estrogen positive breast cancer recurrence during tamoxifen therapy: microdissected tumor
Summary:	Expression profiling of microdissected estrogen positive primary breast cancer tumors from 60 patients. Patients later treated with tamoxifen for 5 years, and tumors grouped according to whether cancer recurred. Results identify markers of disease-free survival that include HOXB13 and IL17BR.
Organism:	*Homo sapiens*
Platform:	GPL1223: Arcturus 22k human oligonucleotide microarray
Citations:	Ma XJ, Wang Z, Ryan PD, Isakoff SJ et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 2004 Jun;5(6):607-16. PMID: 15193263 Loi S, Haibe-Kains B, Desmedt C, Wirapati P et al. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* 2008 May 22;9:239. PMID: 18498629

Reference Series:	GSE1378	**Sample count:**	60
Value type:	log ratio	**Series published:**	2004/06/07

Cluster Analysis



Download

- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

Data Analysis Tools

Find genes

- Compare 2 sets of samples
- Cluster heatmaps
- Experiment design and value distribution

Find gene name or symbol: Go

Find genes that are up/down for this condition(s): ☑ disease state Go

Description of data set

Estrogen positive breast cancer recurrence during tamoxifen therapy:

- probeID is the slide probe ID
- ControlType identifies the probe as positive control, negative control or experimental
- GeneName is the gene accession number for the (non-control) probes
- disease.state indicates whether or not the cancer recurred for each subject during that period:
- Includes 21901 experimental features (22575 total features/attributes)

Preprocessing

The platform for this data set is GPL1223, a custom-designed spotted human oligonucteotide microarray (designed by Arcturus and fabricated by Agilent).

- The data in the ExpressionSet can be taken to be normalized
- Converted from a two-color micro array to a single gene expression value

A classification problem in data mining

A scientific objective: identifying biomarkers (genes) that are significant in predicting disease-free or cancer recurred event

A technical issue: p $>$ n

	Y	genes	Models
disease free/cancer		??	??

Proposed statistical methods

- Logistic regression models with significant analysis of microarray experiments
- A Lasso model: a regression shrinkage and selection method
- Implement CV/nested CV procedures to evaluate the "TRUE" performance of models

Logistic regression model

Model recurrence/non-recurrence as the outcome and selected the most 5 differential gene expression levels as predictors. We will have the linear predictor term

$$\eta = \beta_0 + \beta_1 x_1 + ... + \beta_5 x_5 \quad (1)$$

Then, suppose Y is a Bernoulli random variable for which Y = 1 if the cancer recurs, we have the logistic function.

$$\phi(t) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{e^{-\eta}+1}$$

According to the logistic regression model, we assume that the probability of recurrence give gene expression level x is

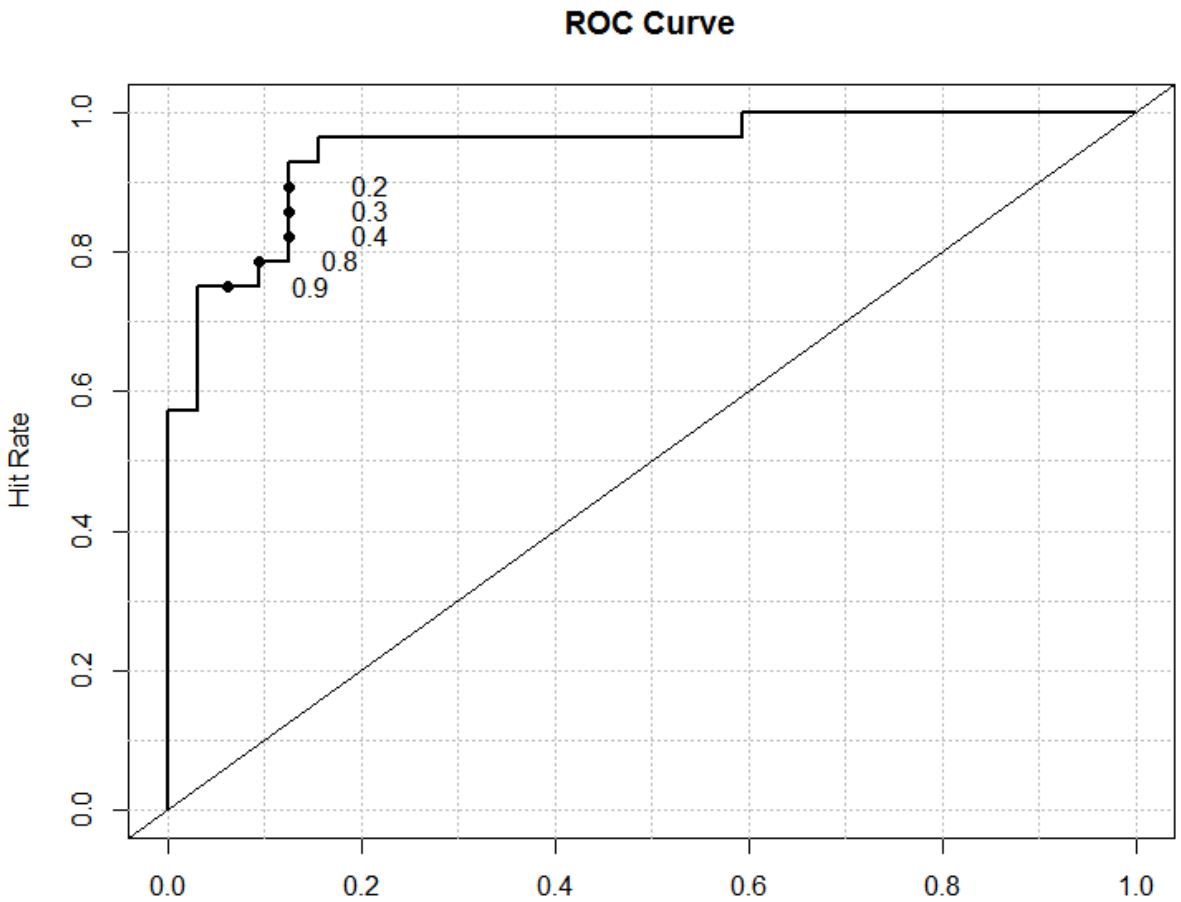
$$Pr(Y = 1 \mid x) = \phi(\eta).$$

CV for logistic regression models

- Select columns as the test data, and the remaining data as the training data.
- Select the 5 genes according to wilcoxon rank sum test from the training data only.
- Evaluate the risk score, the linear predictor term η for subjects making up the test data.
- Estimate the predictive outcomes under the model for all test set and repeat above steps.

Measuring predictability with ROC

AUC statistic is used to evaluate the error rate of prediction



A brief explanation of Lasso

Give a set of input measurements $x_1, x_2, ..., x_k$ and an outcome measurement y, the lasso fits a linear model

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{2}$$

subject to minimize $\sum(y_i - \beta^T x_i)^2 + \lambda \sum \mid \beta_j \mid$

Use CV method to estimate the optimal value for λ

- Extend the cross validation procedure with a method of nested cross-validation approach to evaluate robustness of the LASSO model in predicting the disease state for women
- The nested cross valuation not only selects the optimal lambda, but also evaluates the accuracy of prediction simultaneously

Double CV approach

- Implement a cross validation procedure in an inner loop to choose the optimal λ value
- Make an outer loop function to fit the LASSO model at the optimal λ and validate the error rate with another cross validation procedure.
- Set up a strong penalty for fitting models
- Expect the AUC statistic would be much lower than previous cross validation procedures

Summary

- Identify the genes that are significantly predicting the occurrence/non-occurrence of cancer events
- Evaluate the performance the logit and Lasso models with cross-validation procedures
- Implement the nested cross-validation procedure to evaluate the lasso model

References

[1] Loi S, Haibe-Kains B, Desmedt C, Wirapati P et al. (2008). Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* 9:239

[2] Ma, X., Wang, Z., et al. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen.*Cancer Cell* 5(6):607-16.

[3] NCBI databases http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS807

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.