STAT 530 Bioinformatics: Homework 7

Due May 3, 2017

For problems using R, turn in your answers in the form of a compiled R notebook PDF.

Problem 1

This problem is based on the MAQC-II multiple myeloma data, which you can download from `ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE24nnn/GSE24080/matrix/`. Preprocess the data using `process.R`.

Part (a) (5 points)

Build a logistic regression model on the training data to predict the probability of surviving past 24 months using age, sex, and the gene expression data. Use lasso and choose the λ that gives the smallest 5-fold CV misclassification error. Use the R package `glmnet`. How many probesets have non-zero estimated coefficients? Also report the misclassification rate on the test data.

Part (b) (1 points)

Judging by your reported misclassification rate above, do you think your classifier has done a good job or not? Why or why not? Hint: think about what misclassification rate would be achieved by the dumbest classifier you can think of.

Part (c) (extra credit: 5 points)

Fit a random forest using the first 2,000 covariates in the training data (using all covariates will crash R). Use the R package `ranger`, with the tuning parameters set to their defaults. Report the misclassification rate on the test data, also using the first 2,000 covariates. Use 0.5 for the cutoff c when defining $\hat{Y}^0$ (in the notation of slide 12).