Practical 4

Statistical Learning, 2011

Niels Richard Hansen
September 21, 2011

Microarray Classification

This data set comes from a study of gene expression in two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays (HU6800 chip) containing probes for approximately 6,800 human genes and ESTs. The chip actually contains 7,129 different probe sets; some of these map to the same genes and others are there for quality control purposes. The data comprise 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL) and 25 cases of AML. Samples are divided into a learning set with 38 observations and a test set of 34 observations. The data have been "cleaned" and non-specifically filtered to avoid e.g. many genes that are either not expressed at all or expressed very little. The data were first considered by Golub et al. *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.* Science, 286:531-537, 1999, and has become a standard tutorial microarray data set.

The data are available for download from the course homepage.

1. Implement the *nearest shrunken centroid* method such that you can easily specify a g of your wish to be used for thresholding.

2. Use the method just implemented on the Golub data. Consider soft as well as hard thresholding and compute the training error and the test error as a function of Δ in both cases. Make a plot. How should Δ be chosen?

3. Use logistic regression with the lasso penalty and compute for different choices of λ the training error and test error. Compute these using the likelihood loss as well as the 0-1-loss and make a plot.

4. Write a function that does forward stepwise inclusion of terms in a logistic regression such that each term *decreases* AIC the most. Compute the training error and the test error and plot against the number of variables included. Use the likelihood loss as well as the 0-1-loss.

5. Construct a t-test for the difference in mean expression between the two groups for each gene and sort the data according to the t-test statistic. Starting with one gene, which is the one with the largest t-test statistic, construct sequentially LDA classifiers using in the k'th iteration the set of k genes with the largest t-test statistics. Compute the training error and the test error and plot against k.