PubH 7475/8400 Homework 2 (Spring 2007)

Due on Feb 22, 2007

1. Apply at least five of the methods discussed in class: linear regression (with variable selection), ridge regression, LASSO, Elastic net, adaptive LASSO, PCR, PLS, LDA, QDA, RDA, (penalzied) logistic regression (with variable selection), and nearest shrunken centroids to one of the two following data sets: (50 pts)

   - NCI (or NCI60) microarray data: there are $p = 6830$ predictors (i.e. genes). By ignoring a few classes with only few samples, we only consider 5 CNS, 9 renal, 7 breast, 9 NSCLC, 8 melanoma, 6 ovarian, 6 leukemia and 7 colon tumor samples. The predictors are in a file called Data, and the class labels in Info. *Use $LOOCV$ to evaluate a classifier.*

     This dataset is one of the three used by Dudoit et al (JASA, 2002, p.77-87) to evaluate several classification methods.

   - Spam data: there are $p = 57$ variables (in the Data file) to distinguish two classes, spam (coded as 1) and email (coded as 0) (in the Indicator file). There are total 1813 spams and 2788 emails. As done in the textbook (p.262-263), we take a random subset with 3065 observations as a training set, and the remaining ones as a test set. *Use the test set to evaluate a classifier.*

     You may want to save your random seed so that in the future you can use the same training/test data to evaluate other methods.

   The data and some information on the data are available from the Data link on our course homepage.

2. **(PubH 8400)** Choose two papers from the lists given under Week 3 and Week 4 on the course Updates page: summarize the main points of each paper and comment. (20 pts)

Please attach your computer program and relevant output.