PubH 7475/8475 Homework 2 (Spring 2022)

Due on Feb 16, 2022

1. Apply your chosen five methods discussed in class: linear or logistic regression with (sequential or best subset) variable selection, penalzied linear or logistic regression (with any of Ridge, LASSO, Elastic net, group LASSO, adaptive LASSO, SCAD or TLP penalty), PCR, PLS (or sPLS), LDA, QDA, RDA, diagonalized LDA/QDA/RDA, nearest shrunken centroids (NSC), and SIS/ISIS plus any of the above method, to EACH of the following two data sets; your chosen methods can vary between the two datasets. (100 pts: 10 pts for each method-dataset combination; 10 combinations resulting from 5 (possibly different) methods for each of the two datasets.)

   - NCI (or NCI) 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 CV (e.g. LOOCV or 5- or 10-fold CV) 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). 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 (as indicated in the Indicator file). *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 the textbook homepage.

2. **(8000)** Read Friedman et al (2007, Ann Appl Statist) and Boyd et al (2011; sections 2.3, 3.1, 6.3, 6.4). For a given dataset $\{(Y_i, X_i) : i = 1, 2, ..., \}$ with $X_i = (X_{i1}, ..., X_{ip})'$, describe a coordinate-descent algorithm and an ADMM algorithm to compute the Lasso estimates:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

   You can assume that all Y_i's and X_{ij}'s are already centered at 0 (i.e. their sample means are all 0). (30 pts)

3. **(8000)** Choose two papers (other than those in Q 2) from the lists given under Weeks 3 and 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.