PubH 7475/8475 Homework 3 (Spring 2021)

Due on March 1, 2021

1. Apply 1) a fully grown tree; 2) an optimally pruned tree; 3) random forest; 4) boosting to each of the two following data sets: (4*10*2=80 pts)

   - 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 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). 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. **(Stacking)** Given a dataset $D = \{(Y_i, X_{1i}, X_{2i}) : i = 1, 2, ..., n\}$, we use D and OLS to fit two linear models: $f_1(x_1, x_2) = \alpha x_1$ and $f_2(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$ to obtain the OLSE $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$, and thus $\hat{f}_1(x_1, x_2) = \hat{\alpha} x_1$ and $\hat{f}_2(x_1, x_2) = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$; that is, for example, $\hat{\alpha} = \arg\min_\alpha \sum_{i=1}^n (Y_i - \alpha X_{1i})^2$. Solve

$$(\hat{w}_1, \hat{w}_2) = \arg\min_{(w_1,w_2)} \sum_{i=1}^{n} \Big(Y_i - [w_1 \hat{f}_1(X_{1i}, X_{2i}) + w_2 \hat{f}_2(X_{1i}, X_{2i})]\Big)^2 .$$

   You can impose some reasonable assumptions such as the unique minimizers of the least squared errors and non-zero OLS estimates. (20 pts)

3. Implement the Forward Stagewise Regression in Algorithm 3.4 on p.86. Apply it to a dataset (of your choice) with a small ϵ (say 0.001) and a larger ϵ (say 0.1), then compare their solution paths with that of the Lasso. (30 pts)

4. **(8000) Optional** Consider two models with $p < n$:
   1) true model: $Y = X\beta + \epsilon$, where Y is a n-vector of the response values, $X =$

$(X^{(1)}, X^{(2)}, ..., X^{(p)})$ is the design matrix, and $\beta = (\beta_1, \beta_2, ..., \beta_p)'$ is a p-vector of the unknown regression coefficients;
2) working model: $Y = Z^{(j)} b_j + e$, where $Z^{(j)}$ is a residual vector of regressing $X^{(j)}$ on all other $X^{(k)}$'s with $k \neq j$;
both ϵ and e are the noise vectors with mean 0 and a diagonal covariance matrix.
Prove the OLSEs $\hat{\beta}_j = \hat{b}_j$ and $\hat{b}_j = (Z^{(j)})'Y/(Z^{(j)})'Z^{(j)} = (Z^{(j)})'Y/(Z^{(j)})'X^{(j)}$. (20 pts)

5. **(8000)** Choose two papers from the lists given under Weeks 4-5 on the course Updates page: summarize the main points of each paper and comment. (20 pts)

Please attach your computer program and relevant output.