Homework 4

April 27, 2018

Data Analysis 1. Bagging Linear Predictors

The purpose of this exercise is to apply bagging to linear predictors.

Let us recall the forward stepwise procedure. Consider training data $\{(\mathbf{x}_i, y_i)\}_1^n$, where $\mathbf{x} = (x_1, \ldots, x_p)$ and let $b_i(\mathbf{x})$ be a linear predictor of $\mathbf{x}$ that depends only on i of the p predictors. It is constructed as follows. If the model $b_{i-1}(\mathbf{x})$ is the linear predictor given by

$$b_{i-1}(\mathbf{x}) = a_0 + a_1 x_{j_1} + \ldots + a_{i-1} x_{j_{i-1}}$$

the model $b_i(\mathbf{x})$ is obtained by adding to $b_{i-1}(\mathbf{x})$ the best variable chosen among the variables $x \notin \{x_{j_1}, \ldots, x_{j_{i-1}}\}$. The best variable is the variable x_k such that the RSS of the linear regression of y vs $\{x_{j_1}, \ldots, x_{j_{i-1}}, x_k\}$ is the least. [You can use a criterion other than RSS, if you wish, as we discussed when we considered model selection procedures in linear regression]. Let $\{b_1, \ldots, b_p\}$ be the sequence of models built according to the procedure (forward stepwise selection). Generally one then selects the best of these p models as the one model that is thought of as fitting the data best. This procedure (and similar variable selection methods such as best subset selection, backward stepwise selection) are very unstable since variables are competing to enter the models, and small changes in the training data will result in very different models.

We want to see the advantages of bagging such predictors and compare the accuracy of the models obtained by forward stepwise selection with that of the models obtained by bagging such a linear predictors. The comparison is to be carried out via simulations. You may want to follow these steps

1) Simulate the training data T. Assume $p = 30$, $n = 60$. Draw $(X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \Sigma)$ a multivariate normal with $\Sigma_{ij} = \rho^{|i-j|}$. Then simulate Y from the model $Y_i = \beta_0 + \sum_{i=1}^p X_{ij}\beta_j + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, 1)$. The following code will work as a building block for the simulation for an arbitrary choice on non-zero coefficients β, but you should choose a different vectors of coefficients. Namely, you should consider the (two) cases in which 5 and 25 of the regression coefficients are non zero (choose yourselves values for the non-zero coefficients). You may also want to consider a different value of ρ among $\{0.3, 0.4, 0.6, 0.7, 0.8\}$ (one only).

```
library(MASS)
n=60; p=30
rho=0.6
betazero=-1.2
beta=c(-1,1.3,4,1.2,5,-2,0.34)
beta=c(beta, rep(0, p-length(beta))

Sigma=matrix(ncol=p, nrow=p,0)
for(i in 1:p){
            for(j in 1:i){Sigma[i,j]=rho^{abs(i-j)}
                          Sigma[j,i]=Sigma[i,j]
              }
          }
X=mvrnorm(n=60, mu=rep(0,p), Sigma)
Y=betazero+ X%*%beta +rnorm(n,0,1)
```

2) Compute by forward stepwise selection the sequence $\{b_1(\mathbf{x}, T), \ldots, b_p(\mathbf{x}, T)\}$ and their mean-squared prediction errors $\{e_1, \ldots, e_p\}$

3) Draw $B = 50$ bootstrap samples T_b and for each determine by forward stepwise selection the set of models $\{b_1(\mathbf{x}, T_b), \ldots, b_p(\mathbf{x}, T_b)\}$

4) Consider the bagged sequence $\{\bar{b}_1(\mathbf{x}), \ldots, \bar{b}_p(\mathbf{x})\}$ and the prediction error $\{e_1^{bag}, \ldots, e_p^{bag}\}$

5) Repeat step 1 to 4, 300 times, average the errors and plot the the curves of resulting mean errors as a function of the number of predictors in the model. (In step one, you need to draw new X's and new errors, to get the new Y, but Σ, β, p, n should not change)

6) Explain the curves. Are the errors equal? Is the minimum achieved at the same point? Is one of the two error curves always lower than the other? If not, why not?

Notice that since these data are simulated, you can simulate an additional data set as your validation set to compute your error.

Data Analysis 2. Support Vector Machine and RF

Consider the attached simulated data. The file train.txt contains the training data ($p = 21$, $N = 300$, the last column being the class label) and the file test.txt contains the testing data ($N = 200$). Compare the predictive performance of Random Forest, SVM, boosted trees and, if you wish, the Lasso on these data.

Data Analysis 3. Clustering

For the data set train.txt, consider clustering the X using a) K-means (you can choose $K = 2$ since you already know there are 2 classes), b) hierarchical clustering with the following distances: Euclidean, correlation-based distance and using the proximity measure as obtained by RF (in the previous exercise). Comment your findings.

Theoretical Questions

1) Exercise 12.1 of ESL page 455

2) Exercise 12.2 of ESL page 455