Stats 315B: Homework 1

Joe Higgins, Austin Wang, Jessica Wetstone

Due 5/19/2018

Question 1

Random forests predict with an ensemble of bagged trees each trained on a bootstrap sample randomly drawn from the original training data. Additional random variation among the trees is induced by choosing the variable for each split from a small randomly chosen subset of all of the predictor variables when building each tree. What are the advantages and disadvantages of this random variable selection strategy? How can one introduce additional tree variation in the forest without randomly selecting subsets of variables?

Question 2

Why is it necessary to use regularization in linear regression when the number of predictor variables is greater than the number of observations in the training sample? Explain how regularization helps in this case. Are there other situations where regularization might help? What is the potential disadvantage of introducing regularization? Why is sparsity a reasonable assumption in the boosting context. Is it always? If not, why not?

Question 3

Show that the convex members of the power family of penalties, except for the lasso, have the property that solutions to $\hat{a}(\lambda) = argmin_a \hat{R}(a) + \lambda P_\gamma(a)$ have nonzero values for all coefficients at each path point indexed by λ. By contrast the convex members of the elastic net (except ridge) can produce solutions with many zero valued coefficients at various path points.

Question 4

Show that the variable x_{j*} that has the maximum absolute correlation with $j^* = argmax_{1 \leq j \leq J} |E(yx_j)|$ is the same as the one that best predicts y using squared—error loss. This shows that the base learner most correlated with the generalized residual is the one that best predicts it with squared—error loss.

Question 5

Binary classification: Spam Email. The data set for this problem is spam__stats315B.csv, with documentation files spam__stats315B__info.txt and spam__stats315B__names,txt. The data set is a collection of 4601 emails of which 1813 were considered spam, i.e. unsolicited commercial email. The data set consists of 58 attributes of which 57 are continuous predictors and one is a class label that indicates whether the email was considered spam (1) or not (0). Among the 57 predictor attributes are: percentage of the word "free" in the email, percentage of exclamation marks in the email, etc. See file spam__stats315B__names.txt for the full list of attributes. The goal is, of course, to predict whether or not an email is "spam". This data set is used for illustration in the tutorial Boosting with R Programming. The data set spam__stats315B__train.csv represents a subsample of these emails randomly selected from spam__stats315B.csv to be used for training. The file spam__stats315B__test.csv contains the remaining emails to be used for evaluating results.

(a) Based on the training data, fit a gbm model for predicting whether or not an email is "spam", following the example in the tutorial. What is your estimate of the misclassification rate? Of all the spam emails of the test set what percentage was misclassified, and of all the non-spam emails in the test set what percentage was misclassified?

```r
rm(list = ls())
data_path <- paste(getwd(),'/data',sep='')
setwd(data_path)
```