Stats315B – Homework 2

Spring 2016

Due: May 22, 2016

Please keep answers short. Verbosity will not be rewarded.

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?

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?

3. Let

$$\hat{R}(\mathbf{a}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i,\, a_0 + \sum_{j=1}^{n} a_j x_{ij})$$

be the (convex) empirical risk in a linear regression problem. Show that the convex members of the power family of penalties, except for the lasso,

$$P_\gamma(\mathbf{a}) = \sum_{j=1}^{n} |\, a_j \,|^\gamma \qquad (\gamma > 1)$$

have the property that solutions to

$$\hat{\mathbf{a}}(\lambda) = \arg\min_{\mathbf{a}} \hat{R}(\mathbf{a}) + \lambda \cdot P_\gamma(\mathbf{a})$$

have nonzero values for all coefficients at each path point indexed by λ. By contrast the convex members of the elastic net (except ridge)

$$P_\gamma(\mathbf{a}) = \sum_{j=1}^{n} (\gamma - 1)\, a_j^2/2 + (2 - \gamma)\,|\, a_j \,| \qquad (1 \le \gamma < 2)$$

can produce solutions with many zero valued coefficients at various path points.

4. Consider an outcome variable y and predictor variables $\{x_j\}_{j=1}^J$ with $E[x_j] = 0$ and $E[x_j^2] = 1$ for all x. Show that the variable x_{j^*} that has the maximum absolute correlation with y

$$j^* = \arg\max_{1 \le j \le J} |\, E(y \cdot x_j)\,|$$

is the same as the one that best predicts y using squared–error loss

$$j^* = \arg\min_{1 \le j \le J} \min_{\rho} E\,[y - \rho \cdot x_j]^2.$$

This shows that the base learner most correlated with the generalized residual is the one that best predicts it with squared–error loss.

5. Let $\mathbf{z}_l = \{z_1, \cdots, z_l\}$ be a subset of the predictor variables $\mathbf{x} = \{x_1, \cdots, x_n\}$ and $\mathbf{z}_{\setminus l}$ the complement subset $\mathbf{z}_l \cup \mathbf{z}_{\setminus l} = \mathbf{x}$. Show that if a function $F(\mathbf{x})$ is additive in $\mathbf{z}_l$ and $\mathbf{z}_{\setminus l}$

$$F(\mathbf{x}) = F_l(\mathbf{z}_l) + F_{\setminus l}(\mathbf{z}_{\setminus l})$$

then the partial dependence of $F(\mathbf{x})$ on $\mathbf{z}_l$ is $F_l(\mathbf{z}_l)$ up to an additive constant. This is the dependence of $F(\mathbf{x})$ on $\mathbf{z}_l$ accounting for the effect of the other variables $\mathbf{z}_{\setminus l}$. Show that this need not be the case for $E\,[F(\mathbf{x}) \,|\, \mathbf{z}_l]$ which is the dependence of $F(\mathbf{x})$ on $\mathbf{z}_l$ ignoring the other variables $\mathbf{z}_{\setminus l}$. Under what conditions would the two be the same?

For this homework, as in Homework 1, you will be using the R. The rest of this homework involves becoming familiar with the R package gbm (gradient boosting machine). Gradient boosting is covered in Sections 10.8 – 10.14.3 of the text. Further information can be found in the papers: *Greedy function approximation: a gradient boosting machine* and *Stochastic gradient boosting*. Both of these papers are available at

http://www-stat.stanford.edu/~jhf/#reports.

The first step is to install the gbm package with the R command: **install.packages("gbm")** and follow the instructions. This requires an internet connection. It need be done only once at the first R session. Next the package must be loaded with the R command **library(gbm)**. This must be done in every R session before using any gbm procedures. The R documentation for gbm *Package 'gbm'* (gbm_doc.pdf), a guide (gbm_guide.pdf) and a tutorial *Boosting with R Programming* (gbm_tutorial.pdf) are posted in the *R and Splus Tutorials* section of the *Materials* section of the class web site. Study the tutorial (gbm_tutorial.pdf) carefully as it describes the necessary information to perform the homework.

The data sets Spam_Data, Income_Data, California_Data and Occupation_Data along with documentation files Spam_Info, Spam_Names, Income_Info, California_Info can be found in the class web page in the Data section. Note that the results from the computational problems may not look exactly like the ones in the textbook due to plotting options and slight possible differences in the data sets used.

6. Binary classification: Spam Email. The data set for this problem is Spam_Data, with documentation files Spam_Info and Spam_Names. 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_Names 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_Train represents a subsample of these emails randomly selected from Spam_Data to be used for training. Spam.Test contains the remaining emails to be used for evaluating results.

(a) Based on the training data Spam_Train, 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* Spam.Test what percentage was misclassified, and of all the non-spam emails in the *test set* what percentage was misclassified?

(b) Your classifier in part (a) can be used as a spam filter. One of the possible disadvantages of such a spam filter is that it might filter out too many good (non-spam) emails. Therefore, a better spam filter might be the one that penalizes misclassifying non-spam emails more heavily than the spam ones. Suppose that you want to build a spam filter that "throws out" no more that 0.3% of the good (non-spam) emails. You have to find and use a cost matrix that penalizes misclassifying "good" emails as"spam" more than misclassifying "spam" emails as "good" by the method of trial and error. Once you have constructed your final spam filter with the property described above, answer the following questions:

(i) What is the overall misclassification error of your final filter and what is the percentage of good emails and spam emails that were misclassified respectively?

(ii) What are the important variables in discriminating good emails from spam for your spam filter?

(iii) Using the interpreting tools provided by gbm, describe the dependence of the response on the most important attributes.

7. Regression: California Housing. The data set California_Data consists of aggregated data from 20,640 California census blocks (from the 1990 census). The goal is to predict the median house value in each neighborhood from the others described in California_Info. Fit a gbm model to the data and write a short report that should include *at least*

(a) The prediction accuracy of gbm on the data set.

(b) Identification of the most important variables.

(c) Comments on the dependence of the response on the most important variables (you may want to consider partial dependence plots (*plot*) on single and pairs of variables, etc.).

8. Regression: Marketing data. The data set Income_Data was already used in Homework 1. Review Income_Info for the information about order of attributes etc.

(a) Fit a gbm model for predicting income form the other demographic attributes and compare the accuracy with the accuracy of your best tree from Homework 1.

(b) Identify the most important variables. Note that here "sex" appears to be one of the *least* influential variables in predicting income. Yet it is well know that on average in the United States women receive considerably less pay than men. Assuming that the result from this analysis is correct, is it inconsistent with the national average result. If not, why not.

9. Multiclass classification: marketing data. The data set Occupation_Data comes from the same marketing database used in Homework 1. The description of the attributes can be found in Occupation_Info. The goal in this problem is to fit a gbm model to predict the

type of occupation from the 13 other demographic variables.

(a) Report the test set misclassification error for gbm on the data set, and also the misclassification error *for each class*.

(b) Identify the most important variables.