STAT 575
Homework 9 Problems
due Wednesday April 19

2 Problems. Show all work.

This is a Predictive Analytics HW using R. Include your R code used to answer the questions. Please follow the lab report directions linked off the Homework page.

This HW will use the `randomForest` package and function. Please see Example 4 in the Lab and Presentation linked off the course Calendar for additional information.

1. We will use Random Forest for classification. We will use a German Credit Approval dataset. There is link to the description of the data off the course HW page. A description of the data can also be found at:

`https://onlinecourses.science.psu.edu/stat857/node/222`

We will need to read in the data and make sure that the variables are of the correct type. There is a link to the file `hw9_dataprep.r` off the course HW page. It can also be found at:

`http://www-rohan.sdsu.edu/~babailey/stat575/hw9_dataprep.r`

You should copy and paste the lines of the R code (The > indicates R commands!) , so that you are able to fit a Random Forest.

(a) Use `randomForest` to predict the `Creditability` in the `credit` dataset, using all the remaining predictor variables, with the default settings in R, except use the option `importance=TRUE`, so that we can make the mean decrease in accuracy variable importance measure plot. Use the `set.seed(6)` command before you call the `randomForest` function. Call your fitted object `credit.rf`

Note: You will have to make sure that you have installed the `randomForest` package and have loaded the library in order to use the function. After installing the package, use the command in R:

```
> library(randomForest)
```

(b) For your fit to part (a), print to the screen the fit. Include this in your report.

(c) Examine the output from (b). What is the out of bag (OOB) error rate? If you look at the confusion matrix, where are most of the missclassification being made? (Hint: The rows of the confusion matrix are the truth and the columns are the Random Forest predictions.)

(d) Make a variable importance plot of your fit. Include this in your report. What are the top 4 most important variable in predicting Creditability using the Mean Decrease in Accuracy as the importance measure? Describe these 4 variables?

2. We will now use Random Forest for regression. We will use the dataset `Boston` in the `MASS` library. You should be able to access this dataset with the R command `library(MASS)`. We will use this dataset to predict the median value of owner-occupied homes in the suburbs of Boston. The help file will describe the 13 predictor variables.

(a) Use `randomForest` to predict the `medv` using all the remaining predictor variables, with the default settings in R, except use the option `importance=TRUE`, so that we can make the mean decrease in accuracy variable importance measure plot. Use the `set.seed(6)` command before you call the `randomForest` function. Call your fitted object `BH.rf`

(b) For your fit to part (a), print to the screen the fit. Include this in your report.

(c) Make a variable importance plot of your fit. Include this in your report. What are the top 2 most important variable in predicting home values? Describe these 2 variables?

(d) For fun, let's compare the above fit to a linear model fit. We will use the `lm` function, which uses the same formula syntax as `randomForest`. Call your fitted object `BH.lm` (There is no importance option for the linear model, see the help file.) There is no need to set a seed for a linear model!

(e) Use the R `summary` command with your linear model fitted object. Include this in your report.

(f) Examine the summary in (b), which 2 variables have the smallest p-values? Describe these 2 variables?

(g) Let's make a plot of the fitted values versus the actual data values for both the Random Forest and the Linear Model. Include this in your report. You can do this by

```
> plot(Boston$medv, BH.rf$predicted, main="Random Forest Predictions")
> abline(a=0, b=1)
> plot(Boston$medv, BH.lm$fitted.values, main="Linear Model Predictions")
> abline(a=0, b=1)
```

(h) Using all of the above information, which of the models fits "best"? Explain.