Homework: Ch 08

STAT 4510/7510

Due Thursday, April 21, 11:59 pm

For this homework, we will use the dataset `bike.csv` which is available in the Canvas website under Module 8. This dataset includes 731 observations of daily bike rental count with following additional variables.

- dteday : date
- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit :
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- cnt: count of total rental bikes

Download the dataset to your working directory and load it into `R` by running the following code.

```r
bike <- read.csv("bike.csv")
bike <- bike[,-1] # remove the index column

# some variables are transformed to factors.
bike$season <- as.factor(bike$season)
bike$workingday <- as.factor(bike$workingday)
bike$weathersit <- as.factor(bike$weathersit)

# check the data structure
str(bike)
```

```
## 'data.frame':    731 obs. of  9 variables:
##  $ dteday    : chr  "2011-01-01" "2011-01-02" "2011-01-03" "2011-01-04" ...
##  $ season    : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ workingday: Factor w/ 2 levels "0","1": 1 1 2 2 2 2 2 1 1 2 ...
##  $ weathersit: Factor w/ 3 levels "1","2","3": 2 2 1 1 1 1 2 2 1 1 ...
##  $ temp      : num  0.344 0.363 0.196 0.2 0.227 ...
##  $ atemp     : num  0.364 0.354 0.189 0.212 0.229 ...
##  $ hum       : num  0.806 0.696 0.437 0.59 0.437 ...
##  $ windspeed : num  0.16 0.249 0.248 0.16 0.187 ...
##  $ cnt       : int  985 801 1349 1562 1600 1606 1510 959 822 1321 ...
```

Problem 1

Run following code to create row indices for a training set containing a random sample of 70% observations, and a test set containing the remaining 30% of the observations.

```
set.seed(1)
train <- sample(1:nrow(bike), 0.7*nrow(bike))
bike.test <- bike[-train, ]
```

(a) Fit a regression tree to the training data, with `cnt` as the response and the other variables except `dteday` as predictors. Use the `summary()` function to produce summary statistics for the tree, and describe the results obtained. How many terminal nodes does the tree have? What is the residual mean deviance? (this corresponds to the training MSE for a regression tree)

(b) Create a plot of the tree, and interpret the results.

(c) Make a prediction, manually using the plot you obtained from (c), for the following new observation, and explain this prediction process in plain words.

season	workingday	weathersit	temp	atemp	hum	windspeed
3	0	2	0.426667	0.426737	0.8575	0.146767

(d) Apply the `cv.tree()` function to the training set in order to determine the optimal tree size. Produce a plot with tree size on the x-axis and cross-validated error on the y-axis. Find the best tree size, and prune the tree if necessary.

(e) Using the best tree found in (e), predict the response on the test data, and produce the test MSE.

(f) Run the following code and use bagging to analyze the data. What is the training MSE?

(g) Use the `importance()` function to determine which variables are most important.

(h) Using the trained model in part (f), predict the response on the test data, and produce the test MSE.

(i) Repeat model training and testing using random forests. Set the seed to 1. Specify the number of variables considered at each split as 2.

(j) Now repeat model training and testing using boosting. Set the seed to 1. Specify the distribution as `gaussian`. Use the decision stump (`interaction.depth = 1`) for base tree model. Try four different number of trees with 50, 100, 500, 1000. Descirbe what you observe.