Homework 2
Exploratory Data Analysis

Prof: Yannet Interian

Due: July 23th 2015

1 Introduction

For this homework you will write three functions that are meant to interact with the dataset that accompanies this assignment. The dataset is contained in a zip file exp-auc.zip that you can download from canvas. You should submit this homework to github.

1.1 Data

The zip file contains 94 comma-separated-value (CSV) files containing data about tuning parameters for a decision tree model (machine learning algorithm). The model predicts conversion rates (the probability that an impression will convert) for advertising campaigns. The files ending in "_exp.csv" contain the results of various the experiments, each file corresponds to an advertising campaign. The files ending in "_ci.csv" contain confidence intervals for one of the experiments. In this homework you are going to write three functions to analyze the result of this experiment. The main performance matrix is "auc" (area under the curve).

1.2 Exercise 1 (5 points)

Write a function named “mean_auc” that computes the mean “auc” per campaign id **for campaigns specified by the input**. Given a vector of campaing ids, the function reads a subset of the “_exp.csv” files from the directory specified in the “directory” argument and returns the mean auc per campaign id. A prototype of the function is as follows

```
auc_mean <- function(directory, campaings=1:47)
    # directory is a vector of length 1
    # campaings is a numeric vector e.g c(2,4,5)
    # this function outputs a data frame with columns
    # campaign,mean_auc
}
```

Please save your code to a file named mean_auc.R. The function that you write should be able to match this output:

```
> source("mean_auc.R")
> mean_auc("exp-auc", campaings=1:2)
  campaign  mean_auc
1        1 0.6382610
2        2 0.6329466
>
> auc <- mean_auc("exp-auc")
> summary(auc$mean_auc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4631  0.5762  0.6357  0.6441  0.7067  0.8300
```

1.3 Exercise 2 (5 points)

Write a function that given a *directory name* and threshold (a number between 0 and 1). Does the following:

- Reads all files ending in “_exp.csv”
- Computes *mean_auc* for each campaign (call it *mean_auc_campaign*)

- Keeps campaigns with *mean_auc_campaign* > threshold
- For the subset of the campaigns, computes *mean_auc* as a function of *num_trees*

The function should return a data frame with columns *num_trees* and *mean_auc*. If no campaigns meet the threshold requirement, the function should return an empty data frame with the right names. A prototype of this function follows:

```
num_trees_auc <- function(directory, threshold=0.5)
    # this function outputs a data frame with columns
    # num_trees,mean_auc
}
```

hint: list.files, lapply, ave

Please save your code to a file named num_trees_auc.R. The function that you write should be able to match this output:

```
> source("num_trees_auc.R")
> num_trees_auc("exp-auc")
 num_trees  mean_auc
1         2 0.5844875
2         5 0.6263897
3        10 0.6470991
4        20 0.6603226
5        30 0.6627370
6        40 0.6667567
7        50 0.6674781
8        60 0.6688882

num_trees_auc("exp-auc", threshold=0.6)
  num_trees  mean_auc
1         2 0.6170085
2         5 0.6679285
3        10 0.6891289
4        20 0.6989281
5        30 0.7023458
```

```
6         40 0.7063402
7         50 0.7076356
8         60 0.7080126
```

1.4 Exercise 3 (5 points)

In this exercise you are going to use the function you wrote in HW 1 exercise 1. Write a function that computes the minimum num_trees per campaign for experiments with auc in the confidence interval. You can follow the following steps:

- Use the function from HW 1 exercise 1 to compute a data frame for each campaign. (This filters experiments if auc outside the confidence interval).
- Use the "rbind" or similar function to create a data frame with all campaigns.
- Use the function aggregate to compute the minimum num_trees per campaign.

A prototype of this function follows

```
min_num_trees <- function(directory)
    # this function outputs a data frame with columns
    # campaign,num_trees
}
```

Please save your code to a file named min_num_trees.R. The function that you write should be able to match this output:

```
 head(min_num_trees("exp-auc"))
  campaign num_trees
1        1        10
2        2         5
3        3        10
4        4         5
5        5         5
6        6        20
```

2 Testing your code

I will deduct 10% of your grade if your code is not executable exactly in the fashion mentioned in this document. Please use the test code (test_hw2.R) provided with this assignment and make sure your code passes the tests before submitting your homework. I will be testing your code in a unix machine (a mac) please make sure your code runs properly under linux.

Before submitting, place your R files, ”test_hw2.R” and the data dir “exp-auc” in your working directory and run the following commands in R.

```
# if you don’t have it installed
install.packages("testthat")
library("testthat")
test_dir(".")
# You should get this output
Running unit tests

Testing mean_auc.R
...
Testing num_trees_auc.R
...
Testing min_num_trees.R
..
```