DA5030 | Intro to Machine Learning & Data Mining

Northeastern University
Khoury College of
Computer Sciences

Practice Problems 1

An organization has collected data on customer visits, transactions, operating system, and gender and desires to build a model to predict revenue. For the moment, the goal is to prepare the data for modeling. Analyze the data set in the following manner:

1. (0 pts) Either install Base R and R Studio on your computer or create an account at RStudio.cloud and then learn how to build R Markdown Notebooks to execute your code and organize your output into a readable report. For those working on Windows, you may also use Microsoft Open R.
2. (5 pts) Download this data set and then upload the data into RStudio Cloud. Each row represents a customer's interactions with the organization's web store. The first column is the number of visits of a customer, the second the number of transactions of that customer, the third column is the customer's operating system, and the fourth column is the customer's reported gender, while the last column is revenue, *i.e.*, the total amount spent by that customer.
3. (10 pts) Calculate the following summative statistics: total transaction amount (revenue), mean number of visits, median revenue, standard deviation of revenue, most common gender. Exclude any cases where there is a missing value.
4. (10 pts) Create a bar/column chart of gender (x-axis) versus revenue (y-axis). Omit missing values, *i.e.*, where gender is *NA* or missing.
5. (5 pts) What is the Pearson Moment of Correlation between number of visits and revenue? Comment on the correlation.
6. (10 pts) Which columns have missing data? How did you recognize them? How would you impute missing values?
7. (15 pts) Impute missing transaction and gender values. Use the mean for transaction (rounded to the nearest whole number) and the mode for gender.
8. (20 pts) Split the data set into two equally sized data sets where one can be used for training a model and the other for validation. Take every odd numbered case and add them to the training data set and every even numbered case and add them to the validation data set, i.e., row 1, 3, 5, 7, etc. are training data while rows 2, 4, 6, etc. are validation data.
9. (10 pts) Calculate the mean revenue for the training and the validation data sets and compare them. Comment on the difference.
10. (15 pts) For many data mining and machine learning tasks, there are packages in R. Use the *sample()* function to split the data set, so that 60% is used for training and 20% is used for testing, and another 20% is used for validation. To ensure that your code is reproducible and that everyone gets the same

result, use the number 77654 as your seed for the random number generator. Use the code fragment below for reference:

```
set.seed(101) # Set Seed so that same sample can be reproduced in future also
# Now Selecting 75% of data as sample from total 'n' rows of the data
sample <- sample.int(n = nrow(data), size = floor(.75*nrow(data)), replace = F)
train <- data[sample, ]
test  <- data[-sample, ]
```

Submission Details

- Practice Problems are for learning and practice and therefore are not graded and no submission is required. You are encouraged to discuss and review them with your peers. Additionally, they are reviewed during weekly recitations. If you desire, you may ask for individual feedback from the instructional staff during office hours. Completing practice problems will prepare you for the graded practicums and their completion is critical to doing well on the practicums and the final project.

Northeastern University
Khoury College of Computer Sciences

Learning

Blackboard
Lynda.com
Data Camp

Support

Contact Instructor
Virtual Office
Book Appointment

Search