Homework 2

Yu-Ru Lin

University of Pittsburgh
INFSCI 2160: Data Mining

yurulin@pitt.edu

2018-01-24

Homework 2 I

You will use the dataset “German Credit Data” for this assignment.

- Data csv available at: german_credit_data.csv
- Data description on UCI
- Submit your report using R Markdown and save into *.html file. Submit your *.html and your original*.Rmd via courseweb.
- Due: 2018-02-06 11.59pm

Task: analyze dataset “German Credit Data.” The objective is to predict whether the client’s credit is good or not (the binary variable `V21`).

Homework 2 II

1. Read the data description. Identify and report response variable and predictors.
2. Explore the dataset, and generate both statistical and graphical summary with respect to the numerical and categorical variables. (Consider only the following variables in this exploration: V1, V2, V5, V7, V10, V11, V13, V15, V16, V19 and V21.)
   a) Generate a summary table for the data. For each numerical variable, list: variable name, mean, median, 1st quartile, 3rd quartile, and standard deviation.
   b) For numerical variables, plot the density distribution. Describe whether the variable has a normal distribution or certain type of skew distribution.
   c) For each categorical predictor, generate the conditional histogram plot of response variable.
   hint: E.g., you can use `facet_grid` in `ggplot`.

Homework 2 III

3. Apply logistic regression analysis to predict `V21`. Evaluate the models through cross-validation and on holdout samples. Interpret the effect of the predictors.

   a) Implement a 10-fold cross-validation scheme by splitting the data into training and testing sets. Use the training set to train a logistic regression model to predict the response variable. Examine the performance of different models by varing the number of predictors. Report the performance of the models on testing set using proper measures (accuracy, precision, recall, F1) and plots (ROC, lift).

   b) For the best model, compute the odds ratio and interpret the effect of each predictors.