Homework 1

Yu-Ru Lin

University of Pittsburgh
INFSCI 2160: Data Mining

yurulin@pitt.edu

2017-01-11

Homework 1 I

You will use the dataset D2 described on p.294 in DMR (online access via Pitt network) Appendix A for this assignment.

- Submit your report in PDF, and your code in *.R, via courseweb.
  - It is ok to generate your PDF report using "R Markdown." Make sure your code is reproducible, and you provide clear answers (elaborate description) to the questions.
- Due: 2017-01-17 11.59pm

Task: analyze dataset D2 DirectMarketing.csv
The objective is to explain `AmountSpent` in terms of the provided customer characteristics.

Ref. DMR p.294

Homework 1 II

1. Read the data description on DMR p.294. Identify and report response variable and predictors (also called explanatory variables or features).
2. Explore the dataset and generate both statistical and graphical summary.
   a) There are missing values in data. Describe how you deal with them. **hint:** Check the data description to see what the missingness mean in this dataset.
   b) Generate a summary table for the data. For each numerical variable, list: variable name, mean, median, 1st quartile, 3rd quartile, and standard deviation.
   c) For numerical variables `AmountSpent` and `Salary`, plot the density distribution. Describe whether the variable has a normal distribution or certain type of skew distribution.
   d) For each numerical predictor, describe its relationship with the response variable through correlation and scatterplot.

Homework 1 III

e) For each categorical predictor, generate the conditional density plot of response variable.
hint: Plot the density of response variable into multiple distributions separated by the predictor's categories, on the same figure. Use different colors or line shapes to differentiate categories

f) (extra points) For each categorical predictor, compare and describe whether the categories have significantly different means.

3. Apply regression analysis on D2. Evaluate the model as well as the impact of different predictors.

a) Use all predictors in a standard linear regression model to predict the response variable. Report the model performance using R^2, adjusted R^2 and RMSE. Interpret the regression result.

b) Use different combination of predictors in standard linear and non-linear regression models to predict the response variable. (Here we don't consider interaction terms.) Evaluate which model performs better using out-of-sample RMSE.

Homework 1 IV

hint:: Use `poly`, `locfit` or other appropriate packages for non-linear regression models.

hint: Implement leave-one-out cross-validation for out-of-sample evaluation.

c) From the best model, identify the most important predictor in the model, and explain how you determine the importance of the predictors.

hint: Consider "variable selection" in the out-of-sample evaluation setting.