Homework 3

Yu-Ru Lin

University of Pittsburgh
INFSCI 2160: Data Mining

yurulin@pitt.edu

February 1, 2017

Homework 3 I

You will use the dataset D8 described on DMR (online access via Pitt network) Appendix A for this assignment (the same dataset you used in Homework 2).

- Submit your report in PDF, and your code in *.R, via courseweb.
  - If you use "R Markdown," make sure your code is reproducible, and you provide clear answers (elaborate description) to the questions.
  - Follow the homework guideline posted on Piazza.
- Due: 2017-02-14 11.59pm (two weeks from today)

Ref. DMR Appendix A

Homework 3 II

Task: analyze dataset D8 audit.csv
The objective is to predict the binary (TARGET_Adjusted) target variables.

Apply different classification techniques (incl. logistic regression, kNN, Naive Bayesian, decision tree, SVM, and Ensemble methods) on this dataset. Use all available predictors in your models.

1. Use a 10-fold cross-validation to evaluate different classification techniques. Report your 10-fold CV classification results in a performance table. In the table, report the values of different performance measures for each classification technique. For example, you will generate a table like:

Homework 3 III

	logistic	kNN	NB	Decision tree	SVM	...
accuracy						
precision						
recall						
F-score						
AUC						

Generate two bar charts, one for F-score and one for AUC, that allow for visually comparing different classification techniques.

Homework 3 IV

2. Report at least two variants for techniques with parameters and incorporate them into your table. For examples, for kNN, you may include kNN-1, kNN-3, kNN-5. For decision tree, you may include the default tree, and a tree after pruning. For SVM, you may include different kernels and gamma/cost parameters.
3. Generate an ROC plot that plot the ROC curve of each model into the same figure and include a legend to indicate the name of each curve. For techniques with variants, plot the best curve that has the highest AUC.
4. Summarize the model performance based on the table and the ROC plot in one or two paragraphs.

Homework 3 V

hint: Remove the ID column and rows with missing data before you do the classification. Coerce the categorical variables into discrete numbers because some of the techniques (e.g., kNN) cannot take categorical variables as input.