Homework 3

Yu-Ru Lin

University of Pittsburgh
INFSCI 2160: Data Mining

yurulin@pitt.edu

2018-02-14

Homework 3 I

You will use the dataset `Pokemon` for this assignment.

Data csv: pokemon.csv
Data description: pokemon_description.txt
Original source

- Submit your report in .html, and your code in *.Rmd, via courseweb.
- Due: 2018-02-28 8am (two weeks from today)
- You can extend the sample R code for this assignment: hw3sample.R

Homework 3 II

Task: analyze dataset `Pokemon`

The objective is to predict the binary target variable `Total` (> 500 or not).

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: Coerce the categorical variables into discrete numbers because some of the techniques (e.g., kNN) cannot take categorical variables as input.