CSI 5387

Fall 2012

Assignment 1.

Your task is to build two C4.5 decision trees to evaluate the dataset crx available at http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/. One of these trees must be pruned and the other unpruned.

To work on this assignment you will use Weka and R. You can install WEKA, and a subset of UCI repository datasets, on any computer by downloading it from the following links:
Software: http://www.cs.waikato.ac.nz/ml/weka/snapshots/book2ndEd-branch.zip
Datasets: http://prdownloads.sourceforge.net/weka/datasets-UCI.jar

Second step is to install R from http://www.r-project.org/, and then you will use the following R libraries: DMwR and RWeka.

The third step is to download these libraries from CRAN repository using the command install.packages() inside R environment. R will install all necessary libraries that may be necessary.

The fourth step is to call both libraries issuing the commands:
library(DMwR)
library(RWeka).

To learn about the RWeka package, please type vignette("RWeka") and a pdf help file will appear with helpful information. You will also always also get important information from the RWeka website http://cran.r-project.org/web/packages/RWeka/index.html and from the DMwR website http://cran.r-project.org/web/packages/DMwR/index.html.

To evaluate the results from both trees, you will perform cross-validation (hint 1: use the function
`experimentalComparison()` in DMwR package), and execute two statistical tests: the

T-Test and the Wilcoxon test. (hint 2: Wilcoxon test is available in the function compAnalysis() in DMwR package).

Question a) What are the differences between these trees (build them using the graphical interface provided by Weka and include them in your response)? Is there any difference in average accuracy between these classifiers after cross-validation? Explain reasons for the difference if it exists.

Question b) What is the difference between these statistical tests?

Question c) Explain the outputs of the T-Test and Wilcoxon test in R. Are they equal?

Question d) It is possible to repeat the experiments with the Titanic dataset with different results. The dataset is available in the course website at http://www.site.uottawa.ca/~stan/csi5387/titanic.arff

Send the responses in two files: a doc (not docx) file and a text file containing the R-scripts to the email address csi5387.2012@gmail.com.

Due date: Oct. 8. 23:59PM.

Send your questions about this assignment to the same address.