CSCI 452 (Data Mining)
Dr. Schwartz
Decision Trees
100 pts

This assignment gives you practice with decision tree algorithms. This is an individual assignment. Note that throughout this lab, both the results and the analysis are graded. Be sure to provide your analysis, hypotheses and conclusions in your writeup. You will create an R Markdown file called dtrees.Rmd and submit both the .Rmd and resulting .html (after knitting) files to autolab. Please be sure to follow the order of the questions and identify your answers/analysis clearly so that I can easily find the individual pieces.

When working with your data files, be sure that you store them in the same directory as your .Rmd file (your working directory). Include a link to where I can obtain your individually chose data set in your report in case I need to duplicate results.

1) Background (applied example)

Read this paper: *Bellaachia and Guvenin, "Predicting Breast Cancer Survivability Rates Using Data Mining Techniques," Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining (SDM 2006), 2006. http://www.siam.org/meetings/sdm06/workproceed/Scientific%20Datasets/bellaachia.pdf*

In your submission, answer the following questions about the paper:

a) Briefly describe the discrepancy between the "not survived" and "not alive" patients.
b) What preprocessing did the authors need to perform?
c) How did the authors rank attributes for their effect on prediction?
d) Which learning algorithm performed best?
e) Were there any records with missing data?

2) **R** Investigate the Car Evaluation (http://archive.ics.uci.edu/ml/datasets/Car+Evaluation)dataset hosted on UCI and do the following, using the rpart library and documenting the process and results in your submission file. Generate visualizations along the way, such as (1) the learned tree and (2) a cp graph. Include these in your submission file. Also report on any additional work that you performed, such as during preprocessing. If results are surprising in any way, include some hypotheses about what might have happened. Include all R code.
   a) Build a classification tree for the target variable acceptability (there are four classes). Use information gain as your split measure. Use a training and test set. Report on the classification error for both the training and test sets.
   b) Now employ pruning. Document your decision-making process on where to prune. Report on the classification error for both the training and test sets.
   c) Repeat a-b but using gini. Are there any substantial differences? (Explain)

3) **R and Your Choice of Data** Obtain a dataset from the UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets.html (Do not use any of the datasets that we explored in class). Choose a classification problem (no ordinal target values). Perform analysis similar to what you did in part 2 a and b – that is, choose a target variable and divide the data into training and test sets. Build the tree, prune the tree, analyze and reflect on your results. You do NOT have to repeat with gini. Remember that you don't need to get great results, you just need to understand the results and the process and to present your analysis.

   Submission instructions: Present your answers to the questions above, analysis, commentary, R code and visualizations in an R Markdown file. Name your file dtrees.Rmd and, when you knit, create an .html file. Zip these two files together (don't zip a directory) and upload your submission to Autolab as the Decision Trees lab.