IS 621: Business Analytics and Data Mining

Assignment 4: Entropy and Decision Trees I

Instructions:

- Soft deadline for this assignment (recommended due date): **Wednesday, March 18, 2015 at 11:59 p.m. EST**.
- Hard deadline for this assignment (penalties apply if late): **Wednesday, March 25, 2015 at 11:59 p.m. EST**.
- Late assignments will receive 50% of credit earned and can be submitted until 11:59 p.m. Friday, March 27, 2015.
- Solutions should be typed. Your submission should be electronic and through Blackboard. Multiple files, clearly named, are acceptable.
- Assignments that either cannot be opened correctly or are illegible will receive no credit. If you have any concerns about this, please ask me before the deadline.
- The quality of your solution presentation is as important as the correct answer. For full credit you must show your steps and give clear, thorough answers. In code, this means good commenting.

Background

In this assignment we will focus on another classification technique: decision trees. We will focus in this assignment on the concept of entropy. (We'll use a Gini index approach in the next assignment.) You will be implementing entropy solutions by hand and applying your code to the juror problem we saw in the last assignment.

Data

Data for the assignment are attached in Blackboard as CSV files. These include…

- Training, public testing, and private testing data sets for the Juror Problem

Restrictions

No outside packages should be necessary for problems 1-4. For problem 5 you have instructions on which packages to use. If you know others, you can use them as well.

If you wrote code for these functions in a previous course, you may reuse the code. (Please indicate if this is the case.)

Tasks to Complete:

1. (Programming) Implement a function that takes a categorical vector (in character or factor form) and calculates the entropy for that vector.

   Your code should accept as input a single vector and output a single number.

   For this task, you should submit

   - Your R code that implements the Entropy function

2. (Programming) Implement a function that calculates the information gain of one categorical vector when partitioned according to another categorical vector.

   For this task, you should submit

   - Your R code that implements the Information Gain function

3. (Programming) Implement a function that takes as its input a data frame of categorical variables, one of which is identified as the target variable, and outputs the following in list format:

   - The information gain on the target column when partitioning according to each of the remaining columns
   - The identity of the column that provides the highest information gain

4. (Analysis) Using your custom functions above, build by hand a decision tree on the jury data contained in the file jury-training-data.csv. Document the final set of rules you come up with for the data set. (You should end up with a series of if/then statements, one for each final branch of your tree. Be sure to indicate the support and the probability at the end of each branch. You need not worry about a Laplace correction at this point.

   Once you have built your decision tree, use it to classify all of the observations in the public and private testing data sets for the jury problem. Comment on how well your tree performs on the public data set. If you "prune your tree" can you get better results on the testing data? Explain.

   For this task, you should submit:

   - Your set of decision rules
   - The R code that you used to build the tree (this need not be polished, I just want to see how you made your tree)
   - Your commentary on the classification success on the public learning data set
   - Your classifications (as a CSV file or similar) of the private learning data set

5. (R Package Exploration) Investigate the C5.0() function in the **C50** package as well as the rpart() function in the **rpart** package. There is no deliverable here (there will be in the next assignment), but start trying to use these packages to build a decision tree on the jury data. We'll look at these in our meetup.