ECEN 689–600, Special Topics in Data Mining and Analysis

Assignment 6: due 11:59pm, Wednesday December 6, 2017

General Procedures: Please Read

- *Format*: solutions must be typeset (using e.g. Microsoft Word or LaTex) and rendered in pdf.
- *Transmittal*: email your pdf solutions to me at duffieldng AT tamu DOT edu using the required subject line for the assignment: "DMA Assignment n" where is the number of the assignment (1,2,3, etc).
- *File name*: use file name DMA-n-UIN.pdf where n is the number of the assignment (1,2, etc), UIN is your UIN.
- *Identification*: please include your name and UIN near the top of the first page of your solutions.

Data

- *Data Location.* download the data for this study from https://cesg.tamu.edu/tracedma/
- *Data Description.* The data comprises 10,000 records derived from internet packet measurements. Each record contain 3 fields, separated by the space character.
  - Field 1: PORT is the smallest numeric values of the packet source and destination port.
  - Field 2: SIZE is the packet payload byte size.
  - Field 3: CLASS indicates whether the TCP or UDP network protocol was used.

Decision Trees

This assignment concerns decision trees for predicting CLASS from PORT and SIZE.

1. Create a single plot comprising the scatter of PORT vs. SIZE for each of the two classes. Without doing any computations, annotate where split points could reasonably be located. Use about 10 split points. You may wish to experiment with log-axes.
2. Create a decision tree for predicting CLASS from PORT and SIZE. Use a score function covered in class, i.e., information gain, Gini index, or CART. You may create your own code or use a package, e.g., `rpart` in R
   - Select your own values for purity and size thresholds, and comment on or justify your choice
   - If you create code, include it in your report and cite any sources (e.g. webpages) of code or code snippets you used. Indicate where the thresholds are specified.
   - If you use a package, include your function calls in your report, indicating how the score function and size and purity thresholds are specified. If you use default values, determine these and include in your report.
   - Comment on similarties or differences between these computed results and your choices in item 1 above.

Report

Your report must include the annotated scatter plot (item 1), your code or function calls with parameter details (item 2), and a representation (listed and/or graphical) of the decision tree, e.g., created using tools from a package.