ECEN 758 Data Mining and Analysis
Assignment 6: due 11:59pm, Tuesday December 4, 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 contains 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 plot combining the scatter of PORT vs. SIZE for each of the two classes, using a different color for each class. Create two versions, one with linear axis scaling, and one with log axis scaling. Without doing any computations, annotate where split points could reasonably be located. Use about 10 split points.
2. Select a package to use for desicion tree analysis. Examples include sklearn.tree.DecisionTreeClassifier in python and rpart in R. For your chosen package, determine how to compute prediction accuracy, and how to generate graphical representations of decision trees.
3. Create a decision tree for predicting CLASS from PORT and SIZE.
4. Create a graph of the decision tree.
5. Comment on similarties or differences between this decision tree and your choices for split points in item 1 above.
6. Compute the preduction accuracy for the full dataset, i.e., using the full dataset for both training and testing.
7. Split the data into two parts: A, comprising the first 7,000 records, and B, comprising the last 3,000 records. Train on A then compute the prediction accuracy for B.
8. Comment on two values of prediction accuracy that you computed, relating your answer to the distributiuon of the data and the way the testing-train split was performed.
9. Determine a better split of the data (70% training, 30%) and compute the corresponding prediction accuracy.

Report

In addition to your answers to the questions, your report must include the annotated scatter plots (item 1) and all function calls used in the computation.