UNIVERSITY OF TEXAS AT AUSTIN

# Homework Assignment 8

## Classification trees.

Please, provide your **complete solutions** to the following problems. Final answers only, even if correct will earn zero points for those problems.

**Problem 8.1.** (10 points) Solve Problem **8.4.3** (pp.361-362) in the textbook.

**Problem 8.2.** (10 points) A classification tree is constructed to predict whether a student will pass *Predictive Analytics*. There are two categorical predictors:

- $X_1$ indicating whether the student passed *Linear Algebra* prior to enrolling in *Predictive Analytics*, and
- $X_2$ indicating whether the student passed *Mathematical Statistics* prior to enrolling in *Predictive Analytics*.

Here is the table from a data set of students:

| $X_1$ | $X_2$ | outcome |
|---|---|---|
| 0 | 0 | 5 passed, 20 didn't |
| 1 | 0 | 10 passed, 10 didn't |
| 0 | 1 | 15 passed, 20 didn't |
| 1 | 1 | 15 passed, 5 didn't |

What's the first split made using the Gini index? Be careful to calculate the **weighted** average.

**Problem 8.3.** (30 points) A classification tree is fitted to be used to classify drivers into one of three categories: *Good, Medium* and *Bad*. There are three terminal nodes in the classification tree designating the regions $R_1, R_2$ and $R_3$. Here is the breakdown of categories in each region:

| Region | Good | Medium | Bad |
|---|---|---|---|
| $R_1$ | 70 | 20 | 10 |
| $R_2$ | 30 | 50 | 10 |
| $R_3$ | 20 | 25 | 35 |

Calculate the overall classification error rate, Gini index, and cross-entropy using **weighted averages**.