

# STA6703 SML HW8

## Directions

Please submit **ONE PDF** file including all your reports (answer + code + figures + comments; must be easily readable and have file size under a few megabytes) and **ONE R code script**. The R script is supplementary material to ensure that your code runs correctly. If you are using RMarkdown, please also include your `.Rmd` file. If using Python, please submit the Python notebook/script in lieu of the R script.

Place these two (or three) files in a folder, make a zip or rar archive, and submit the archive electronically via Dropbox file request at [tinyurl.com/nbliznyuk-submit-files](https://tinyurl.com/nbliznyuk-submit-files) (on the landing page, enter your name so that we know it is you and email so that you get a confirmation).

Please **submit only ONE solution** on behalf of the entire work group, **NOT separate/individual solutions** by different group members. You can have multiple submissions, in which case only the most recent will be graded.

**Deadline:** 22-Nov-2022, 10:00 PM EST.

## Practice/Optional Problems (do not submit)

1. Complete the R tutorial for the ISLR chapter 8. You may find the Youtube videos by Trevor Hastie helpful; for links, see file `!_youtube_lab_links.txt` in the subfolder "`[2].code/islr_labs/`"
2. ISLR ch.8: 1,6
3. Typed problem 1: gaining intuition behind the classification trees. Assume the same two-dimensional continuous features (as in class), i.e., iid Uniform on the square  $[-1, 1] \times [-1, 1]$ .

Consider the following two scenarios:

Scenario A: binary classification, where  $Y = 1$  if the values  $(x_1, x_2)$  are both positive or both negative;  $Y=0$  otherwise. (This is a 2x2 piece of a chessboard.)

Scenario B: classification with 4 categories, where each of the four quadrants corresponds to 4 distinct categories.

For each scenario, use the node purity criterion (Gini) presented in class in order to grow a classification tree (no pruning).

1.1. Suppose you grow the trees "by hand" (like we did in class) or "in your head" (no need to document this process in every detail; focus on the "big picture"), and  $n$  is very large. Can we answer the following question (without resorting to R): If  $n$  is very large, which scenario, A or B, would likely lead to a solution closer to the truth (e.g., smaller trees with low test data misclassification)? Briefly explain why.

1.2. Confirm your intuition using R and briefly discuss your findings. For reproducibility, generate the matrix of features  $X$  as follows:

```
set.seed(0); n = 1000; # you can try larger values of n as well
```

```
M = matrix(runif(n*2),ncol=2); X = 2*M - 1;
```

A clarification: in class, we looked at the definition of node/leaf purity (Gini) for the  $m$ th leaf; call it  $G_m = \sum_{k=1}^K \hat{p}_{km}(1 - \hat{p}_{km})$ . The aggregate Gini criterion (for the entire tree  $T$ ) needs to weigh  $G_m$  by the number of data points in the  $m$ th leaf (call it  $n_m$ ), i.e.,  $G(T) = \sum_m n_m \cdot G_m$ .

## Required Problems (for submission)

ISLR ch.8: 3,8,9,10