Homework 3: Trees

MACS 30100: Perspectives on Computational Modeling
University of Chicago

Overview

For each of the following prompts, produce responses *with* code in-line. While you are encouraged to stage and draft your problem set solutions using any files, code, and data you'd like within the private repo for the assignment, *only the final, rendered PDF with responses and code in-line will be graded.*

A Conceptual Problem

1. (15 points) Of the Gini index, classification error, and cross-entropy in simple classification settings with two classes, which would be best to use when *growing* a decision tree? Which would be best to use when *pruning* a decision tree? Why?

An Applied Problem

For the applied portion, your task is to predict attitudes towards racist college professors using the General Social Survey (GSS) survey data. Each respondent was asked *Should a person who believes that Blacks are genetically inferior be allowed to teach in a college or university?* Given the controversy over Richard J. Herrnstein and Charles Murray's *The Bell Curve* and the ostracization of Nobel laureate James Watson over his controversial views on race and intelligence, this applied task will provide additional insight in the public debate over this issue.

To address this problem, use the `gss_*.csv` data sets, which contain a selection of features from the 2012 GSS. The outcome of interest is `colrac`, which is a binary feature coded as either `ALLOWED` or `NOT ALLOWED`, where 1 means the racist professor *should* be allowed to teach, and 0 means the racist professor *should not* be allowed to teach. Full documentation can be found here. I preprocessed the data for you to ease the model-fitting process:

- Missing values have been imputed
- Categorical features with low-frequency classes collapsed into an "other" category
- Nominal features with more than two classes have been converted to dummy features
- Remaining categorical features have been converted to integer values

Your task is to construct a series of models to accurately predict an individual's attitude towards permitting professors who view Blacks to be racially inferior to teach in a college classroom. The learning objectives are:

- Implement a battery of tree-based learners
- Tune hyperparameters
- Substantively interpret models

2. (35 points) Fit the following four tree-based models predicting `colrac` using the training set (`gss_train.csv`) with 10-fold CV. Remember to tune the relevant hyperparameters for each model as necessary. Only use the tuned model with the best performance for the remaining exercises. **Be sure to leave sufficient *time* for hyperparameter tuning, as grid searches can be quite computationally taxing and take a while.**

   - Decision tree (the rpart algorithm)
   - Bagging
   - Random forest
   - Gradient boosting

3. (20 points) Compare and present each model's (training) performance based on:

   - Cross-validated error rate
   - ROC/AUC

4. (15 points) Which is the best model? Defend your choice.

5. (15 points) Evaluate the performance of the best model selected in the previous question using the test set (`gss_test.csv`) by calculating and presenting the classification error rate and AUC of this model. Compared to the fit evaluated on the training set, does this "best" model generalize well? Why or why not? How do you know?