Homework 5: Tree-based Inference

Overview

Due Sunday by 11:59 pm.

Fork the `problem-set-5` repository

Conceptual: Cost functions for classification trees

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

Application: Predicting attitudes towards racist college professors

In this problem set, you are going to predict attitudes towards racist college professors, using the GSS survey data. Specifically, 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 kerfuffle 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 analysis will provide further insight into the public debate over this issue.

`gss_*.csv` contain a selection of features from the 2012 GSS. The outcome of interest `colrac` is a binary variable coded as either `ALLOWED` or `NOT ALLOWED`, where 1 = the racist professor should be allowed to teach, and 0 = the racist professor should **not** be allowed to teach. Documentation for the other predictors (if the variable is not clearly coded) can be viewed here. Some data pre-processing has been done in advance for you to ease your model fitting: (1) Missing values have been imputed; (2) Categorical variables with low-frequency classes had those classes collapsed into an "other" category; (3) Nominal variables with more than two classes have been converted to dummy variables; and (4) Remaining categorical variables have been converted to integer values, stripping their original labels

Your mission is to bring trees into the context of other classification approaches, thereby constructing a series of models to accurately predict an individual's attitude towards permitting racist professors to teach. The learning objectives of this exercise are:

```
1. Implement a battery of learners (including trees)
2. Tune hyperparameters
3. Substantively evaluate models
```

Estimate the models

2. (35 points; 5 points/model) Estimate the following models, predicting `colrac` using the training set (the training `.csv`) with 10-fold CV:

    - Logistic regression
    - Naive Bayes

- Elastic net regression
- Decision tree (CART)
- Bagging
- Random forest
- Boosting

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.** Grid searches can be computationally taxing and take quite a while, especially for tree-aggregation methods.

Evaluate the models

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.

Evaluate the *best* model

5. (15 points) Evaluate the *final*, best model's (selected in 4) performance on the test set (the test `.csv`) by calculating and presenting the classification error rate and AUC. Compared to the fit evaluated on the training set in questions 3-4, does the "best" model generalize well? Why or why not? How do you know?

Bonus: PDPs/ICE

6. (Up to 5 extra points) Present and substantively interpret the "best" model (selected in question 4) using PDPs/ICE curves over the range of: `tolerance` and `age`. Note, interpretation must be more than simple presentation of plots/curves. You must sufficiently describe the changes in *probability* estimates over the range of these two features. You may earn *up to* 5 extra points, where partial credit is possible if the solution is insufficient along some dimension (e.g., technically/code, interpretation, visual presentation, etc.).