15.095: Machine Learning Under a Modern Optimization Lens **Fall 2021**

Homework 4: Due October 29

Hand in: **pdf** upload to Canvas. Please append any Julia code **at the end** of the whole pdf.

4.1 Question 1: True or False (30 marks)

Please classify the following statements as true or false and justify your answer. We will assign 1 mark for correctly classifying the answer, and 2 marks for the validity of the justification/counterexample.

1. (3 marks) You can have the same variable appear in two splits in a row in an optimal classification tree.
2. (3 marks) Two trees with the same misclassification objective value may have different gini impurity objective values.
3. (3 marks) When people want to train tree-based models to maximize AUC, they use the misclassification loss function.
4. (3 marks) The misclassification error of a classification tree for two classes that has zero splits is always 0.5.
5. (3 marks) Decision Trees are not sensitive to outliers.
6. (3 marks) The randomness of each tree in Random Forests only comes from using different subsets of observations in the training set to build the tree.
7. (3 marks) OCT is a nonlinear prediction method, whereas logistic regression is a linear prediction method.
8. (3 marks) Standardising or normalizing the input features for a tree-based model does not matter to improve the model's predictive performance.
9. (3 marks) A trained XGBoost model always achieves better in-sample AUC than the first tree from the boosted tree sequence.
10. (3 marks) Random Forest and Boosted Trees (e.g., XGBoost) are tree-based methods that improve the performance by aggregating the results of weak learners.

4.2 Question 2: Investigating Justice Stevens Decisions with Tree-Based Models (70 Points)

For this part of the homework, we are supplying you with the Supreme Court data we covered in Lecture 8. The variables we use are as follows:

- Independent Variables (Features X):

  - Circuit court of origin (1st – 11th, DC, FED).
  - Issue area of the case (e.g., civil rights, federal taxation).
  - Type of petitioner, type of respondent (e.g., US, an employer).
  - Ideological direction of the lower court decision (conservative or liberal).
  - Whether petitioner argued that a law/practice was unconstitutional.

- Dependent variable (Target y):
  - Did Justice Stevens vote to reverse the lower court decision? A 1 is *Yes*, a 0 is *No*.

We will start by investigating how different choices of parameters for Optimal Classification Trees (OCTs) can influence the final tree model. Feel free to explore the IAI software documentation here. Please start by splitting your dataset into a training and test set in a stratified fashion.

(a) (10 Points) Train OCTs with depth=3 and minbucket=5, using each of the loss functions (misclassification, gini, and entropy). Discuss the differences in the resulting trees and report the performance (AUC + Accuracy) on the training and test set.

(b) (10 Points) Train OCTs using the gini loss function and minbucket=5, for varying depths. Plot the test scores (AUC + accuracy) as a function of the tree depth. Plot the complexity parameter selected by the algorithm as a function of the tree depth (note: the IAI software autotunes the cp parameter for you. You just have to extract it from the final model). Discuss the results.

(c) (10 Points) Similarly, fix the depth and loss function, and plot the test scores as function of different minbuckets. As before, discuss the results.

(d) (40 Points) Now, we investigate the dataset further and build the best machine learning model we can for it. As a reminder, you should appropriately validate the hyperparameters of your models and have a robust procedure for fairly deciding which model performs best. Please implement and tune the following methods (any package is welcome but IAI implements all of them):

- CART,
- OCT,
- Random Forest,
- Boosted Trees (e.g., XGBoost),
- Sparse Logistic Regression (reminder, IAI implements it).

Implement one additional technique or model of your choice that could improve the results (it may not improve in the end, but motivate your initial attempt). This could include, but is not limited to:

- Sparse feature selection on the data before inputting it to a Random Forest or Boosted-Tree model,
- Optimal Classification Trees with Hyperplanes,
- Ensembling of your models (e.g., majority vote of all your best models).

Overall:

(i) (5 marks) Explain your methodology to tune the hyperparameters of your models.

(ii) (5 marks) Explain the additional technique or model you implemented and why you chose it.

(iii) (10 marks) Report the out-of-sample AUC and accuracy of each model on the test set.

(iv) (10 marks) Comment about the interpretability/explainability of each model and discuss about which variables were used by the different models.

(v) (10 marks) Write an executive summary about your findings, in particular which model(s) you find the most appropriate for a practical use in a consulting company, what are the trade-offs, and how reliable/interpretable/explainable you think they are.