

## 1. Summary:

This week we learned about tree-based methods to stratify and segment a predictor space into a number of simple regions. We learned about first decision trees and how they are used but can lead to overfitting. We then moved onto bagging and random forest as alternatives to avoid overfitting both with the data and also with the predictors used. Lastly, we talked about their differences and use cases.

## 2. Concepts:

- For trees, set of splitting rule can be summarized
- They are known as decision trees and can be used for both classification and regression
- Based on rules for each predictor
- Lengths of trees correspond to that decision's impact on the response
- Decisions are yes no answers for value of predictors or  $>$
- Regions are always boxes that are disjoint and cover the entire space
- Convention: left means true and right means false
- Root is top then prompts are internal nodes

Tree Building Process:

- J total regions
- J chosen with CV
- Approach
  - Divide predictors into J distinct non overlapping regions
  - For each pt in  $R_j$ , predict as the mean of response variable among training data in that region

How do you decide on the region:

- Natural objective function is to minimize the differences for each region
- Unfeasible amount of reason split
- Fix to the computational issue
  - Greedy sequential approach
  - Top down recursive splitting
  - At each step pick best split
- Start with mean prediction over entire data
- Select a predictor and cutting points to split
- Do the same for each predictor/region

- However, every data point would have own region and RSS would just be 0

Avoid over-fitting:

- Stop splitting region if it has less than k points
- Pruning a tree - start from original tree then collapse

Classification version:

- Our prediction is majority label for a region
- Error choices: 1 - samples majority level/samples in region
- Gini index based on probability

Fix Overfitting:

- Do not further split if there are a certain number of points
- Max depth of tree
- Choose subset of large tree
- Ensemble methods: Each tree has weak leaves, combine weak leaves for powerful model

Bagging

- Averaging reduced variable by factor B
- Bootstrap B datasets, predictive model for each, avg each model, this dec. variance

Random Forest:

- Small tweaks in bagging to reduce variance further, random selection of predictors considered each time too

### **3. Uncertainties:**

My main uncertainties lie in deciding on when to choose each method. Of course we lower variance with RF and Bagging so does that mean that those are just objectively the best ones to choose? I am sure there are some scenarios when the original tree is perfectly acceptable so I am just curious when we will know and what are things to look for this to be true. Also, in general, at this point we have learned of so many models for so many different situations. Is there a specific guideline on when to choose what? Or, is it typical for data scientists/machine learning engineers to try multiple methods for each situation and see which one yields the best results and just choose that?