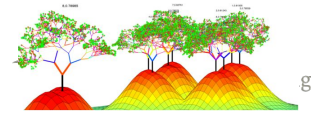


Random Forests Intro

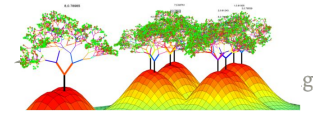
What happens when our lonely tree, grows into a mighty forest?



Objectives

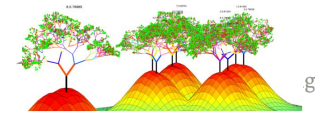
In the next 45 minutes, students will be able to...

1. **explain and build** a classification random forest
2. **discuss** the differences between bagging and a random forest
3. **interpret** how tuning “n_estimators” will effect the random forest model’s variance



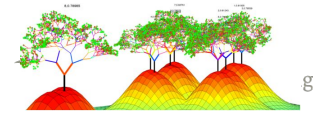
Ensemble Methods

- **Combination** of many weak models
- **Example:** Jellybeans in a Jar
 - Individuals all have poor guesses
 - Average of poor guesses turns out to be a great guess
- Works for **Classification** or **Regression**



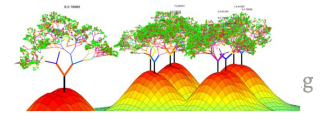
Decision Trees Review

- Strengths of an individual Tree
 - —
 - —
 - —
- Weaknesses of an individual Tree
 - —
 - —



Decision Trees Review

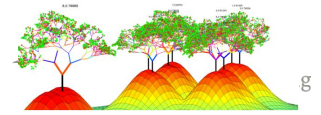
- Strengths of an individual Tree
 - No scaling/normalization necessary
 - Useful for various data types
 - Easy to explain
- Weaknesses of an individual Tree
 - High variance
 - Propensity to overfit
 - Small change in data can cause instability



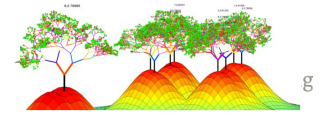
Decision Trees Cont...

- How is a split determined for an individual tree?
- What would be the difference between two decision trees trained with the same data?

Decision Trees Cont...

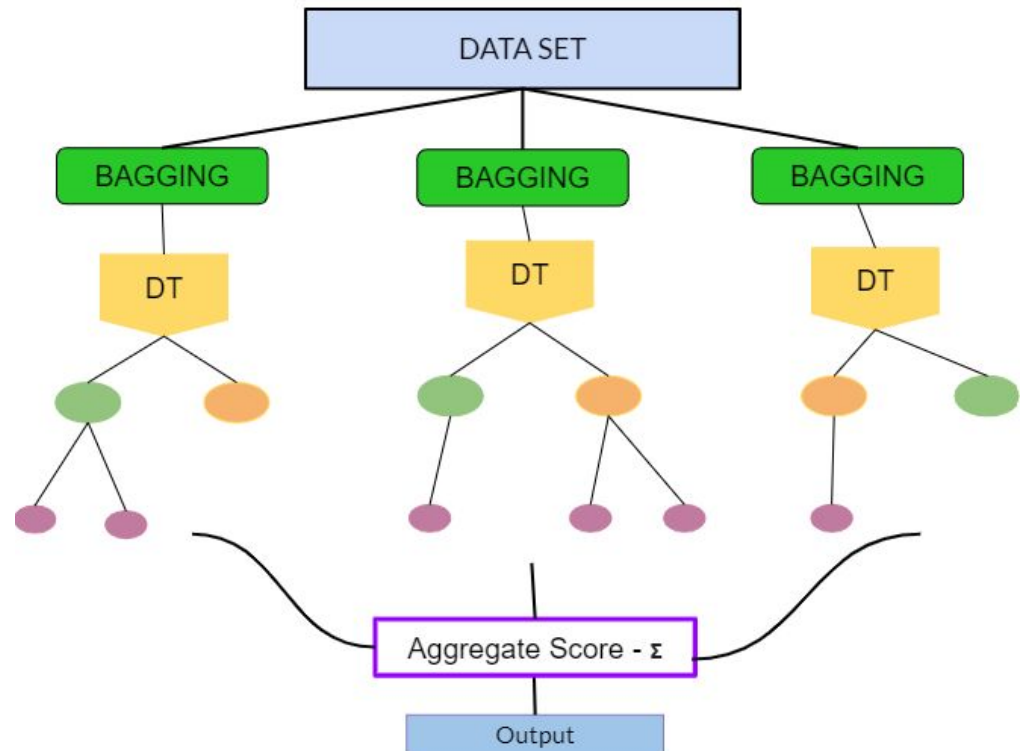


- How is a split determined for an individual tree?
 - Numerical feature:
 - Split at a threshold (like a percentile or value)
 - Categorical feature:
 - Split on value (is or is not value)
 - Information Gain
- What would be the difference between two decision trees trained with the same data?
 - Since each split is mathematically determined and all features are considered for each split, there would be no difference

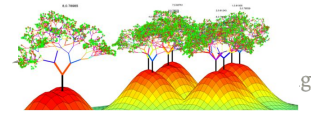


Bagging

- Bagging:
 - “bootstrap” + “aggregation”
- procedure used to reduce variance of a statistical learning method

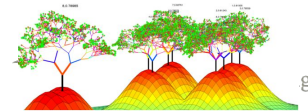


Bagging



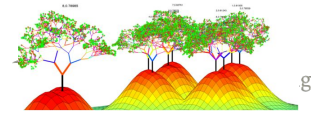
- Term Bagging?
- How does Bagging accuracy compare to Decision Tree accuracy?
- What is an Ensemble method?
 - Example?

A Random Forest



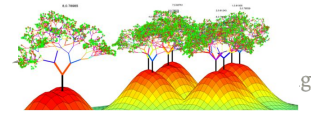
- “ensemble” aka “forest” of decision trees
- Each tree gets a vote
- Bagging combined with random feature subsets considered
 - higher decorrelation with individual tree
 - Decrease variance

Random Forest vs. Bagging



- Bagging
 - Bagging decision trees are pretty cool, but the trees still tend to look pretty similar
 - all features are considered for splitting a node
- Random Forest
 - Bootstrapped datasets
 - Only a random selection of features are chosen for each split in each decision tree

Check for Success



- You are successful today if you can ...
 - Explain Bagging in 1 - 2 sentences.
 - Express why Random Forests work better than traditional Bagging.
 - Explain how changing $n_estimators$ will affect the Random Forest model's variance.