Decision Trees and Ensemble Models

Harvard | Spring 2022 | Anna Trella

Original Material Provided by Bill Zhang

The Machine Learning Pipeline

1. Define the problem.
2. Acquire data.
3. Examine the data.
4. Create a specification.
5. **Build the model.**
6. Measure performance. (Repeat)
7. Deploy!

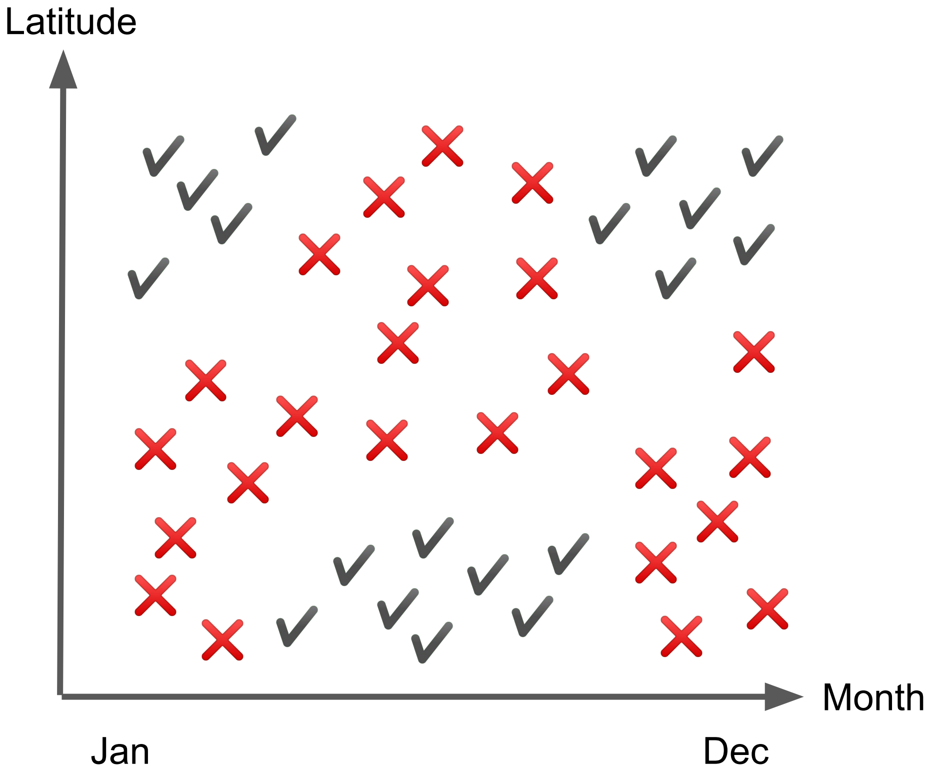
Credit: Stanford CS 229

Decision Trees

A Motivating Example

Problem: Predict if skiing is possible, given the time of year and latitude

- You collected the data to the right

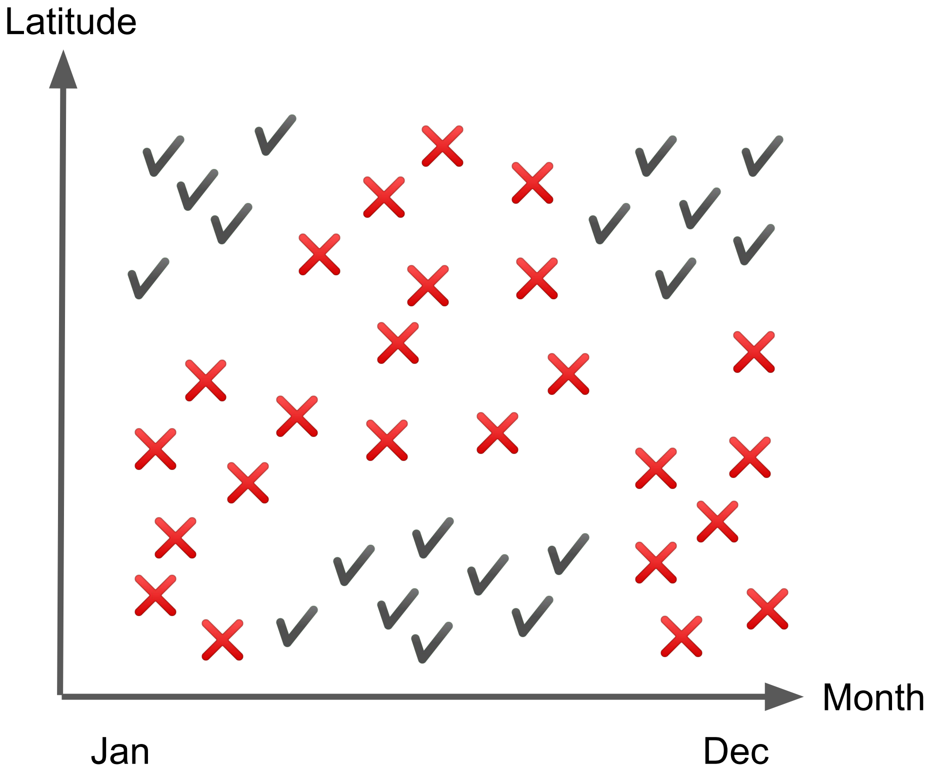


Credit: Stanford [CS 229](CS 229) (Fall 2018, “Tree Ensembles”)

A Motivating Example

Problem: Predict if skiing is possible, given the time of year and latitude

- You collected the data to the right
- **What model is best here?**

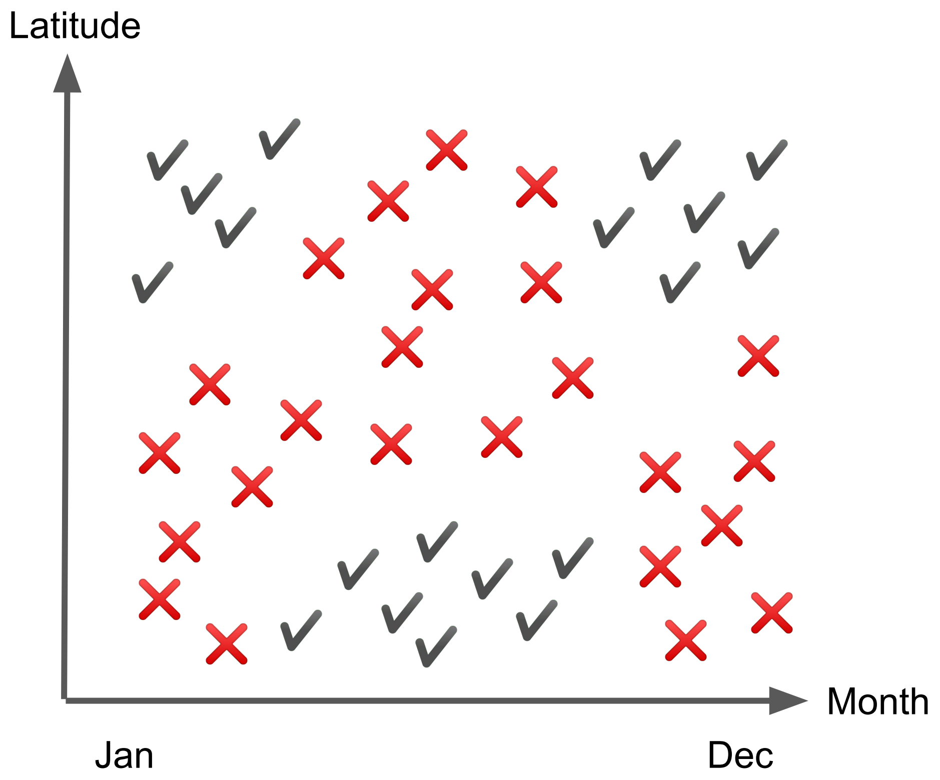


Credit: Stanford CS 229 (Fall 2018, “Tree Ensembles”)

A Motivating Example

Problem: Predict if skiing is possible, given the time of year and latitude

- You collected the data to the right
- What model is best here?
    - Need a non-linear model
- **How would you reason about this problem?**

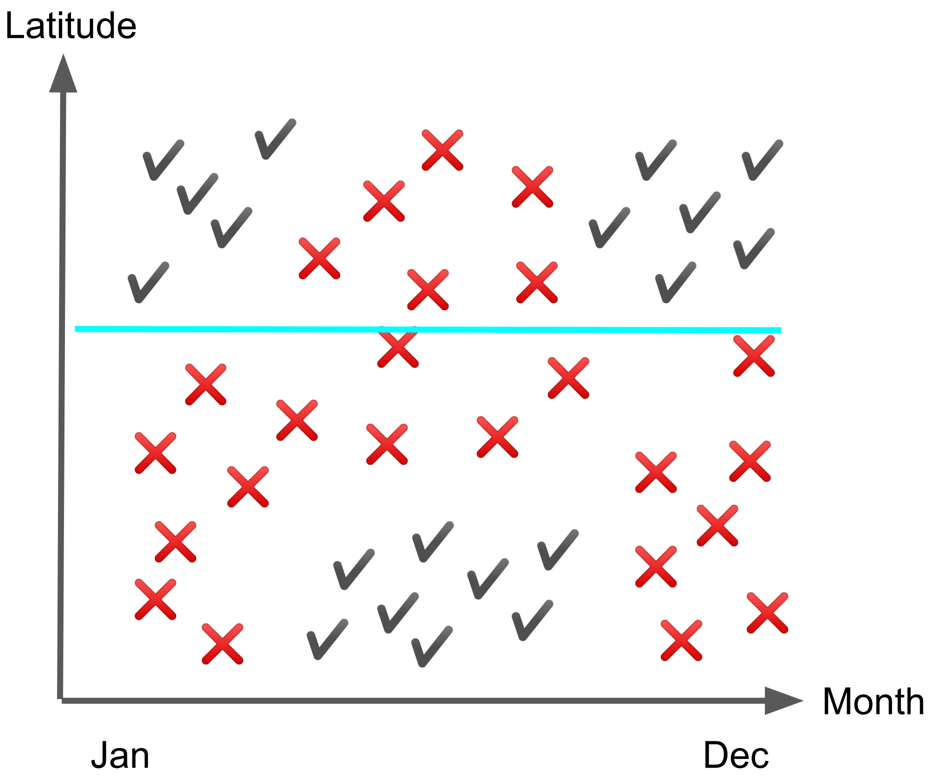


Credit: Stanford CS 229 (Fall 2018, “Tree Ensembles”)

A Motivating Example

Problem: Predict if skiing is possible, given the time of year and latitude

- You collected the data to the right
- What model is best here?
    - Need a non-linear model
- How would you reason about this problem?
    - “Are we far enough North?”

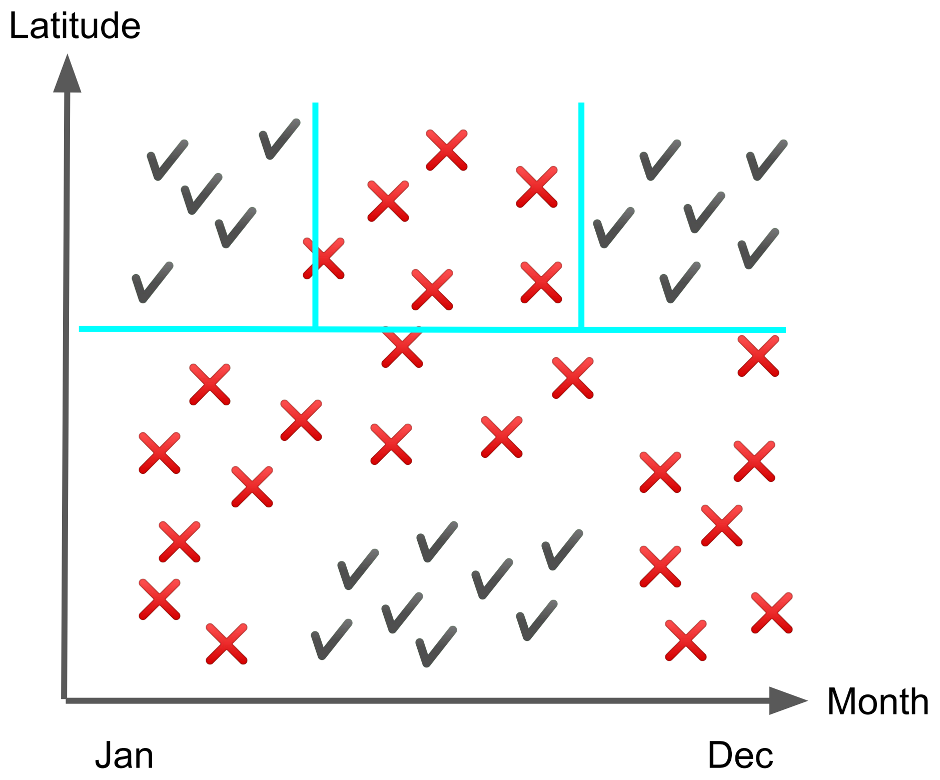


Credit: Stanford [CS 229](CS 229) (Fall 2018, “Tree Ensembles”)

A Motivating Example

Problem: Predict if skiing is possible, given the time of year and latitude

- You collected the data to the right
- What model is best here?
  - Need a non-linear model
- How would you reason about this problem?
  - “Are we far enough North?”
    - “Is it between Oct and Mar?"

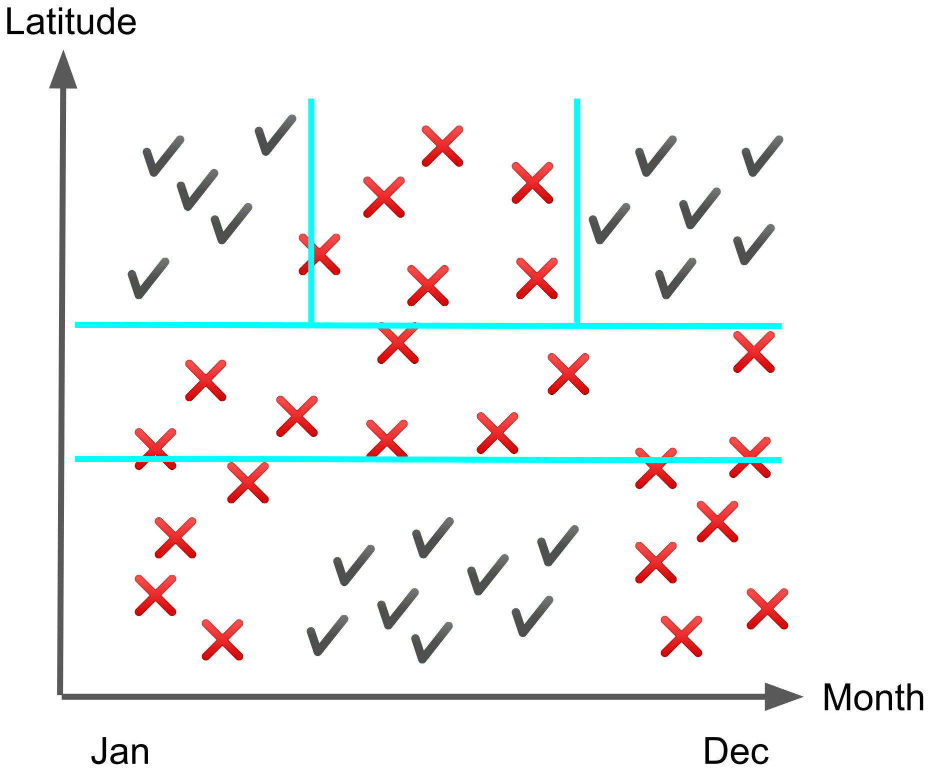


Credit: Stanford CS 229 (Fall 2018, “Tree Ensembles”)

A Motivating Example

Problem: Predict if skiing is possible, given the time of year and latitude

- You collected the data to the right
- What model is best here?
  - Need a non-linear model
- How would you reason about this problem?
  - “Are we far enough North?”
    - “Is it between Oct and Mar?"
  - “Are we far enough South?”

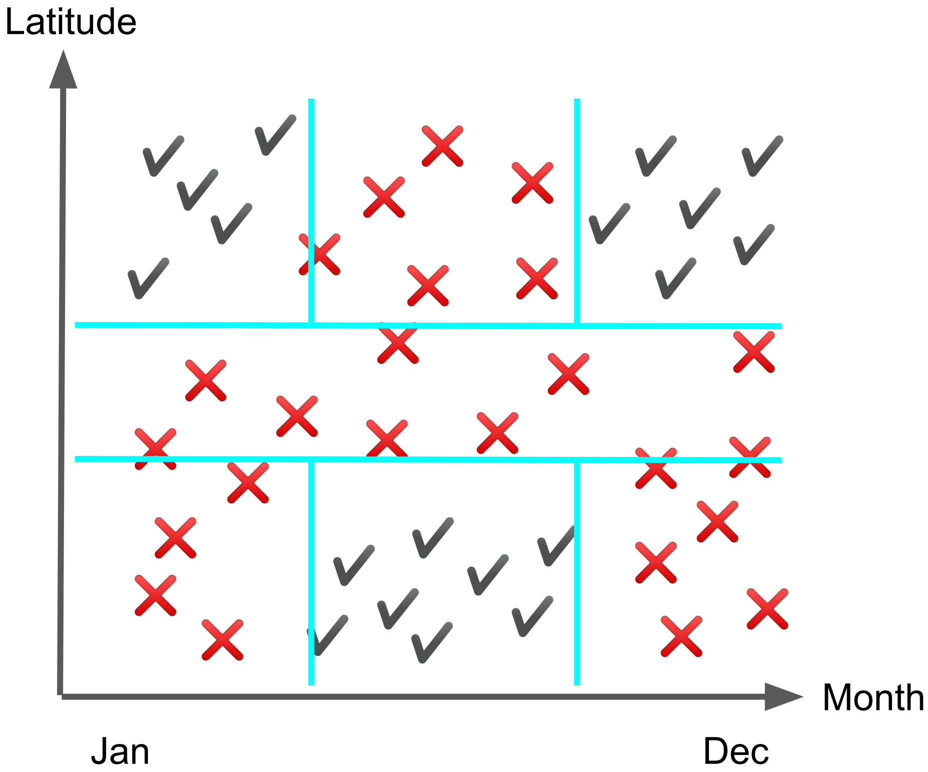


Credit: Stanford CS 229 (Fall 2018, “Tree Ensembles”)

A Motivating Example

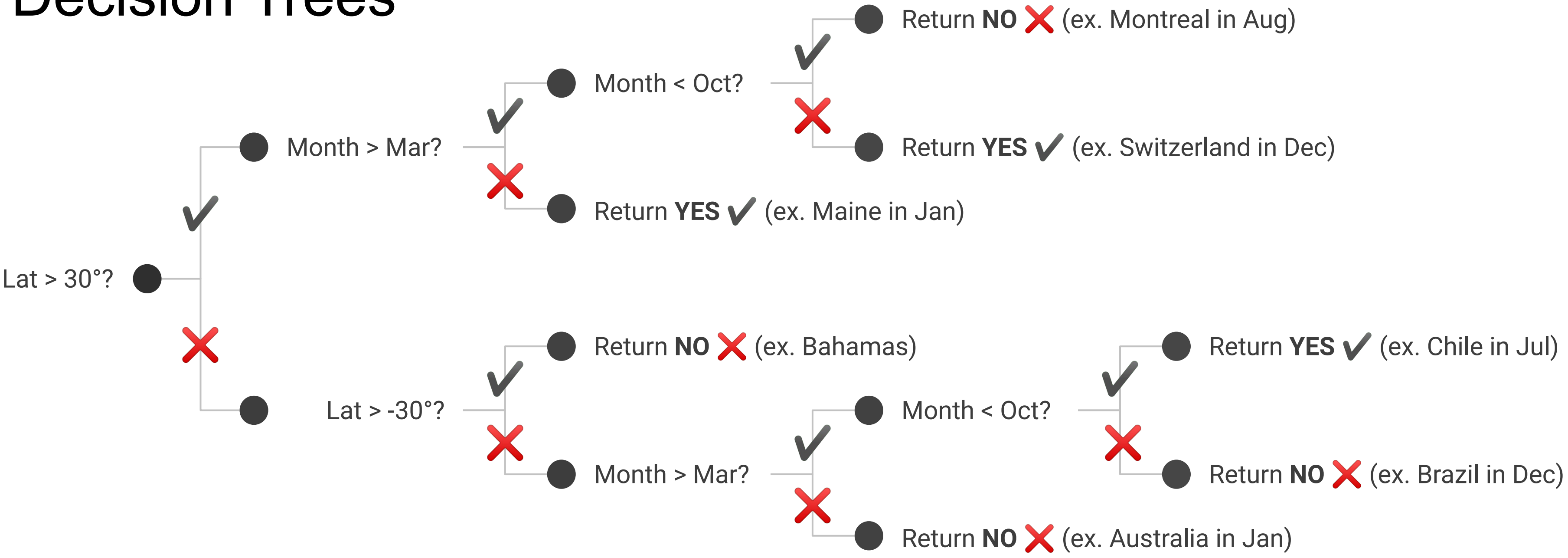
Problem: Predict if skiing is possible, given the time of year and latitude

- You collected the data to the right
- What model is best here?
  - Need a non-linear model
- How would you reason about this problem?
  - “Are we far enough North?”
    - “Is it between Oct and Mar?"
  - “Are we far enough South?”
    - “Is it between Mar and Oct?”



Credit: Stanford CS 229 (Fall 2018, “Tree Ensembles”)

Decision Trees



The structure follows how we would intuitively model the problem!

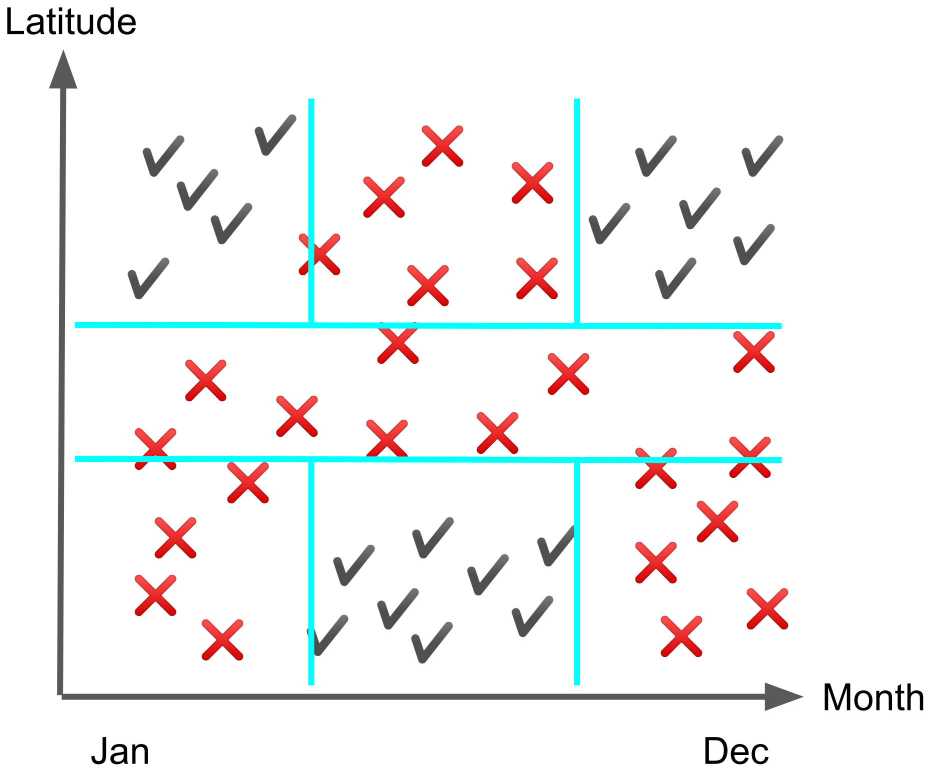
Nodes = questions　　Branches = partitions (Y/N)　　Leaves = final classifications

Credit: Stanford [CS 229](CS 229) (Fall 2018, “Tree Ensembles”)

Decision Tree Models

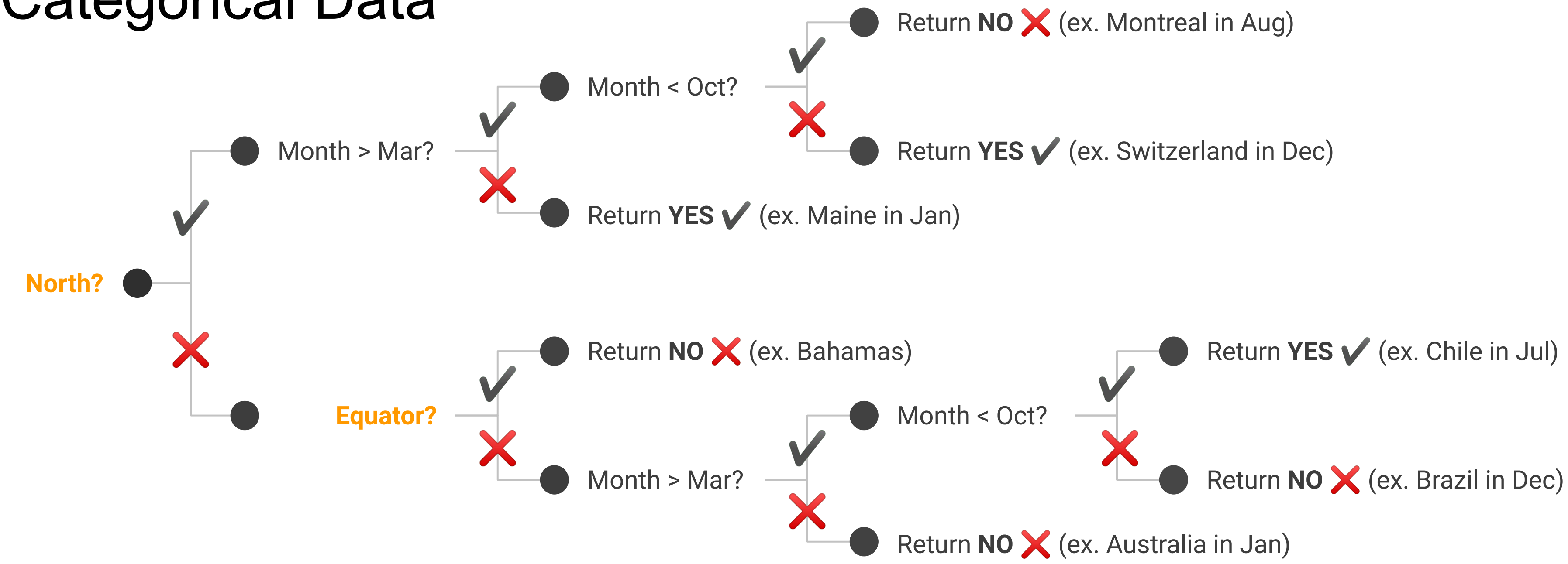
A **decision tree model** does the following:

1. Select some criterion (latitude)
   a. Drawn from features of the data
2. Select some threshold (30°) to split the data up accurately
   a. Minimize a loss function
3. Repeat with each of the split regions to get final tree



Credit: Stanford CS 229 (Fall 2018, “Tree Ensembles”)

Categorical Data



Decision trees work great with categorical data too!

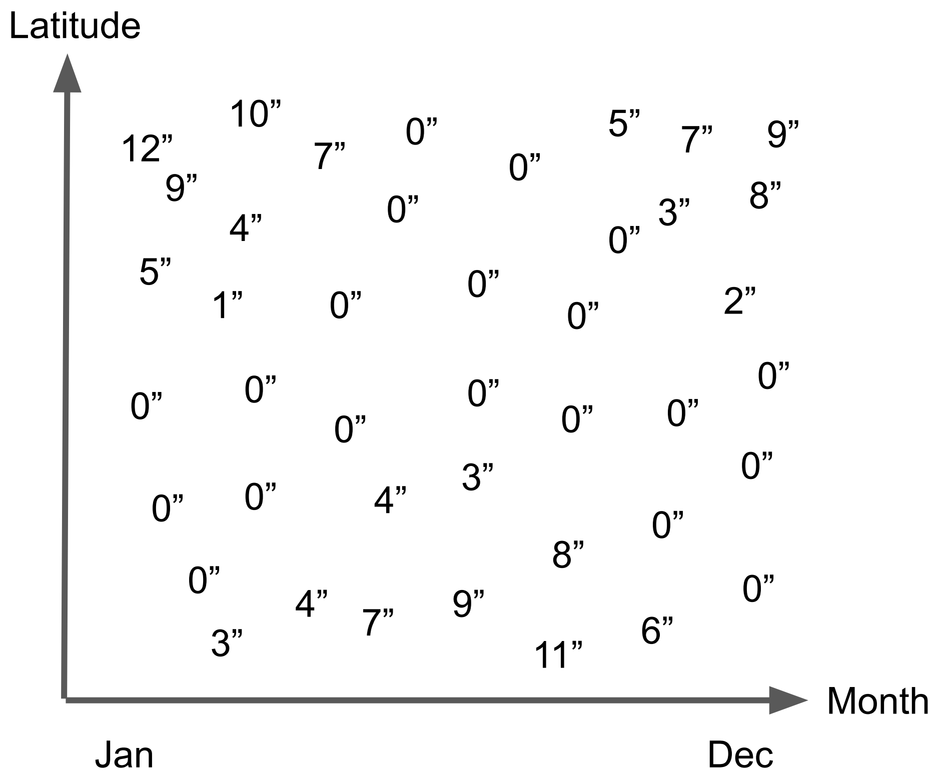
- Ex. {Northern Hemisphere, Southern Hemisphere, Equator} instead of Latitude

Credit: Stanford [CS 229](CS 229) (Fall 2018, “Tree Ensembles”)

Regression Trees

What about a regression task?

Problem: Predict snowfall, given the time of year and latitude



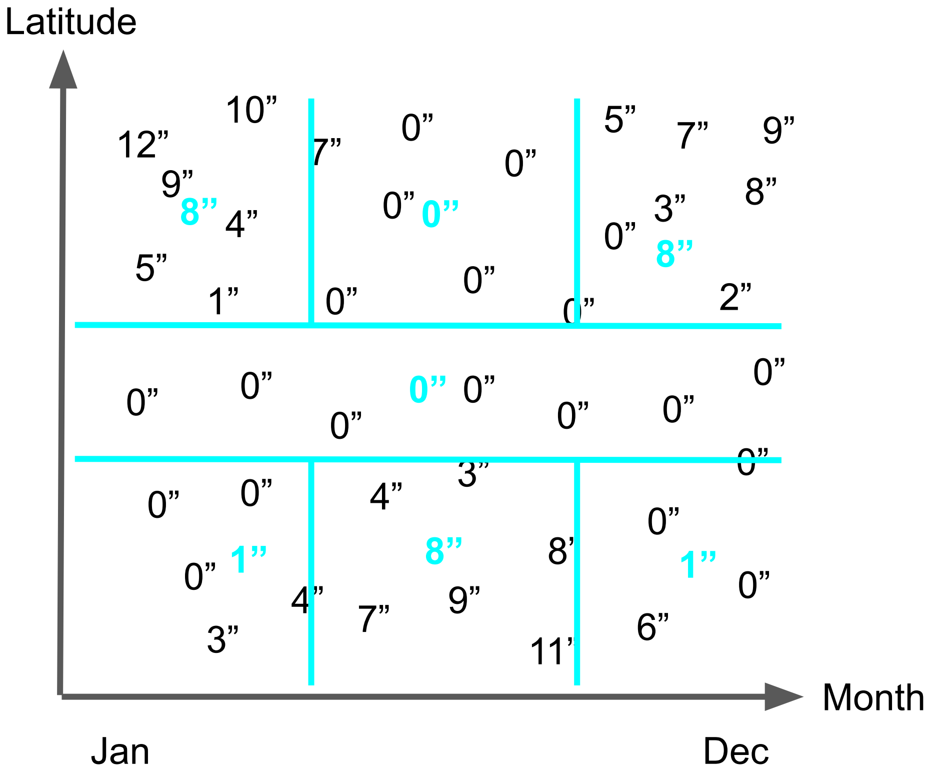
Credit: Stanford [CS 229](CS 229) (Fall 2018, “Tree Ensembles”)

Regression Trees

What about a regression task?

Problem: Predict snowfall, given the time of year and latitude

Idea: Use a decision tree, except return the **average** instead of ✔️❌!

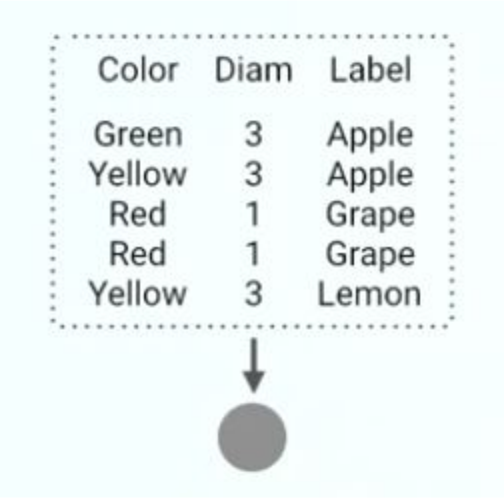


Credit: Stanford [CS 229](CS 229) (Fall 2018, “Tree Ensembles”)

Training

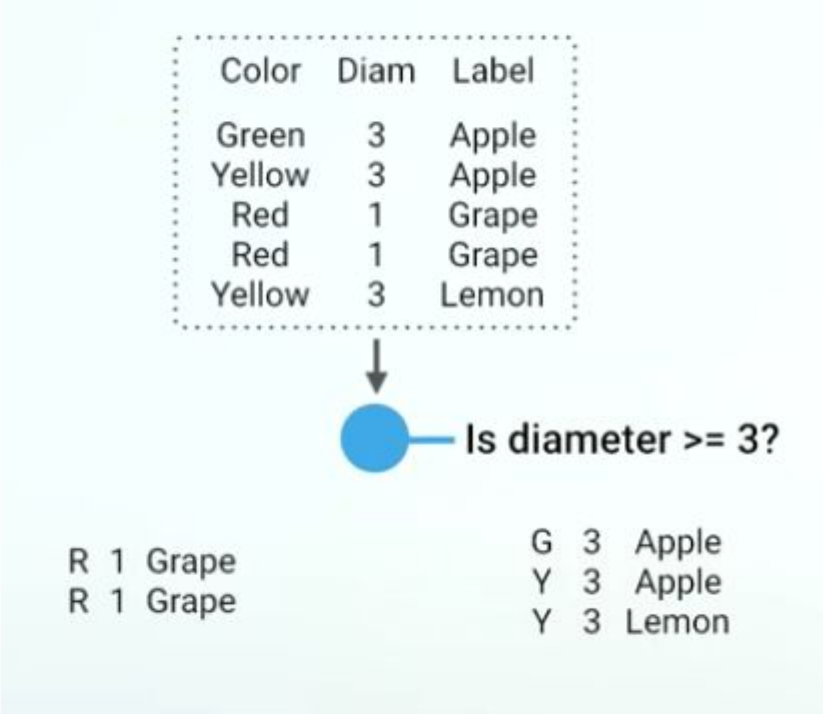
- Greedy, recursive way, downward from the root.
- One way:
  - Choose a dimension
  - Perform a 1D sweep: sort the data points by their values in that dimension and consider those values as split candidates
  - Using some metric of quality, choose the split that has the highest metric value

Training

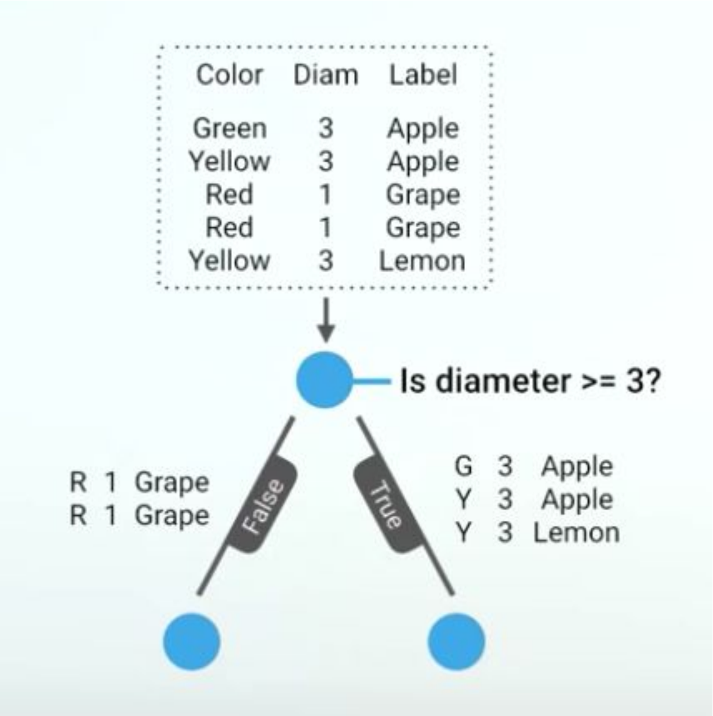


Color	Diam	Label
Green	3	Apple
Yellow	3	Apple
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon

Training

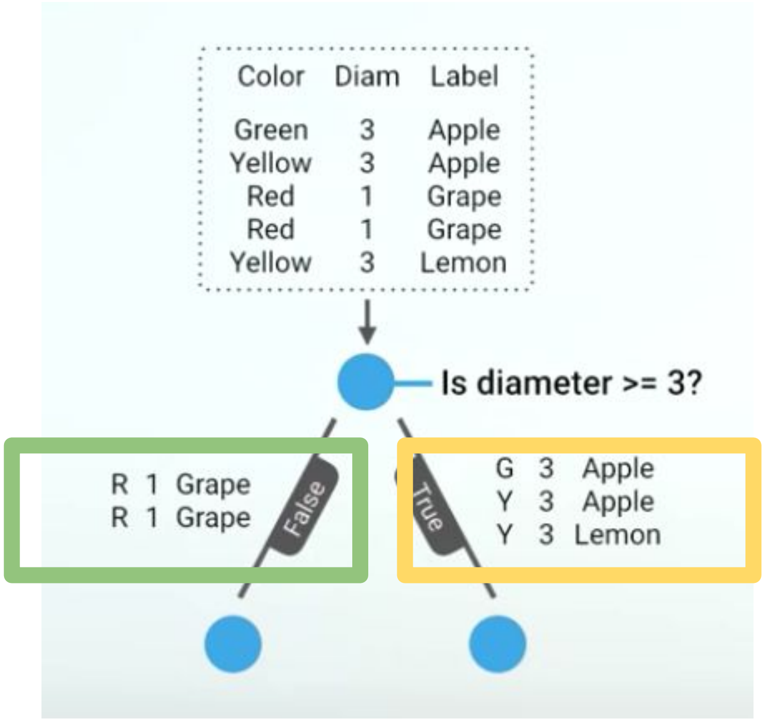

Color Diam Label
Green 3 Apple
Yellow 3 Apple
Red 1 Grape
Red 1 Grape
Yellow 3 Lemon
Is diameter >= 3?
R 1 Grape
R 1 Grape
G 3 Apple
Y 3 Apple
Y 3 Lemon

Training



Color Diam Label
Green 3 Apple
Yellow 3 Apple
Red 1 Grape
Red 1 Grape
Yellow 3 Lemon
Is diameter >= 3?
False
True
R 1 Grape
R 1 Grape
G 3 Apple
Y 3 Apple
Y 3 Lemon

Training



We want to unmix the labels!

Tuning Parameters

- What questions (features splits) should I ask?
- Metric for evaluating how good a question is (how well a node split will lead to unmixed labels in its children).
    - E.g. : using Gini impurity to measure the uncertainty of a node and choose the one that has the most uncertainty.

Discussion

- **What are the pros or cons of decision tree models?**

Discussion

Pros:

- **Popular:** VERY popular model (why don't we learn this in CS 181?)
- **Explainable:** Easy to explain to others and mirrors human decision-making
- **Robust to Data Type and Task:** Works for classification and regression tasks, categorical and numerical data
- **Interpretable:** Can be complex/non-linear but still interpretable

Discussion

Cons:

- **Easily Overfit:** (high variance)
  - Consider creating a very deep tree where each leaf is a single data point
- **Unstable:** sensitive to small changes / perturbations in data
- Low predictive accuracy

How would you address overfitting through regularization?

Discussion

Cons:

- Can overfit easily (high variance)
  - Consider creating a very deep tree where each leaf is a single data point
- Can be unstable, sensitive to small changes in data
- Low predictive accuracy

Regularization techniques:

- Set a minimum number of data points in each leaf
- Set a maximum depth of the tree
- Set a maximum number of nodes

Ensemble Models

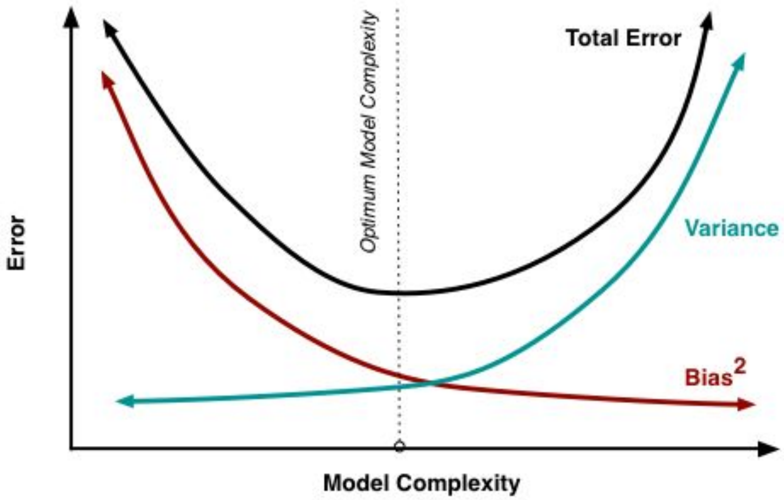
“The interests of truth require a diversity of opinions.” —J. S. Mill

Bias-Variance Tradeoff

Recall from lecture the relationship:

$$E[(\bar{y} - \hat{y})^2] = \text{noise} + \text{bias}^2 + \text{variance}$$

Where do decision tree models fall on this graph?

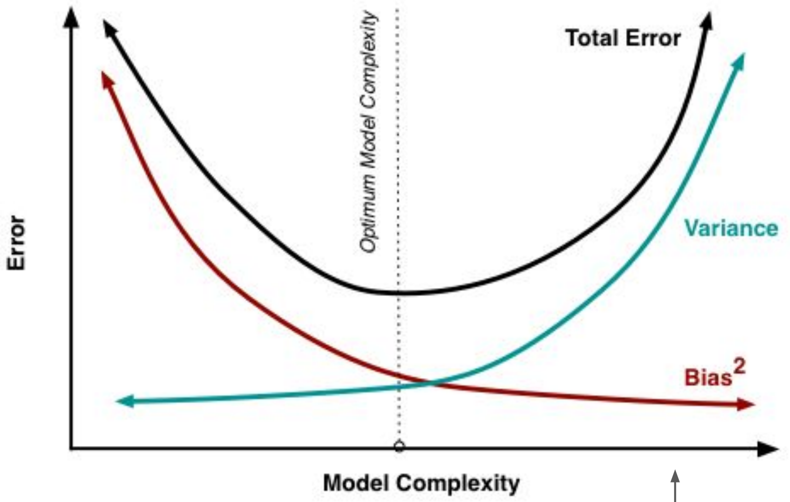


Bias-Variance Tradeoff

Recall from lecture the relationship:

$$E[(\bar{y} - \hat{y})^2] = \text{noise} + \text{bias}^2 + \text{variance}$$

Decision tree models have high variance, but generally low bias



Decision Tree Models

Bias-Variance Tradeoff

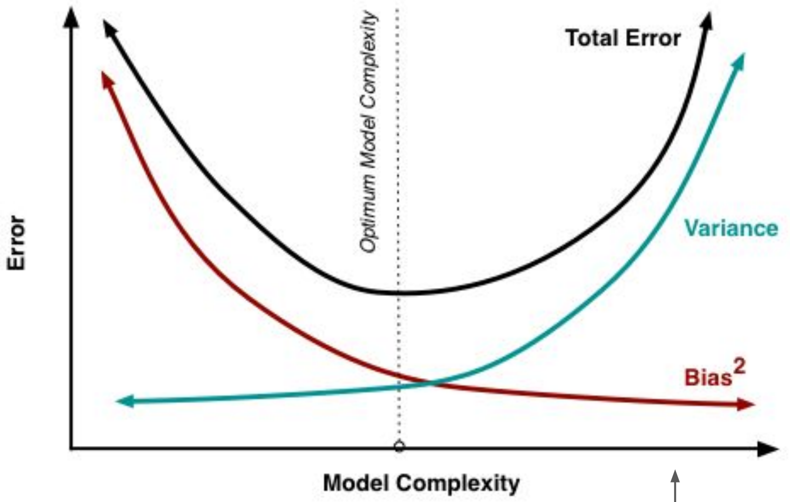
Recall from lecture the relationship:

$$E[(\bar{y} - \hat{y})^2] = \text{noise} + \text{bias}^2 + \text{variance}$$

Decision tree models have high variance, but generally low bias

In lecture, we learned two ways to address models like this:

- Regularization
- **Ensembling**



Decision
Tree Models

Ensembling

An **ensemble model** combines the outputs of multiple models into a final answer

- Classification: cast a majority (plurality) vote
- Regression: take the average

Why does this help us?

Ensembling

An **ensemble model** combines the outputs of multiple models into a final answer

- Classification: cast a majority (plurality) vote
- Regression: take the average

If the outputs of the models are **independent**, the variance can be reduced to **zero** if the number of models is large enough!

If the outputs of the models are **correlated**, the variance can be reduced **somewhat** as the number of models increases…

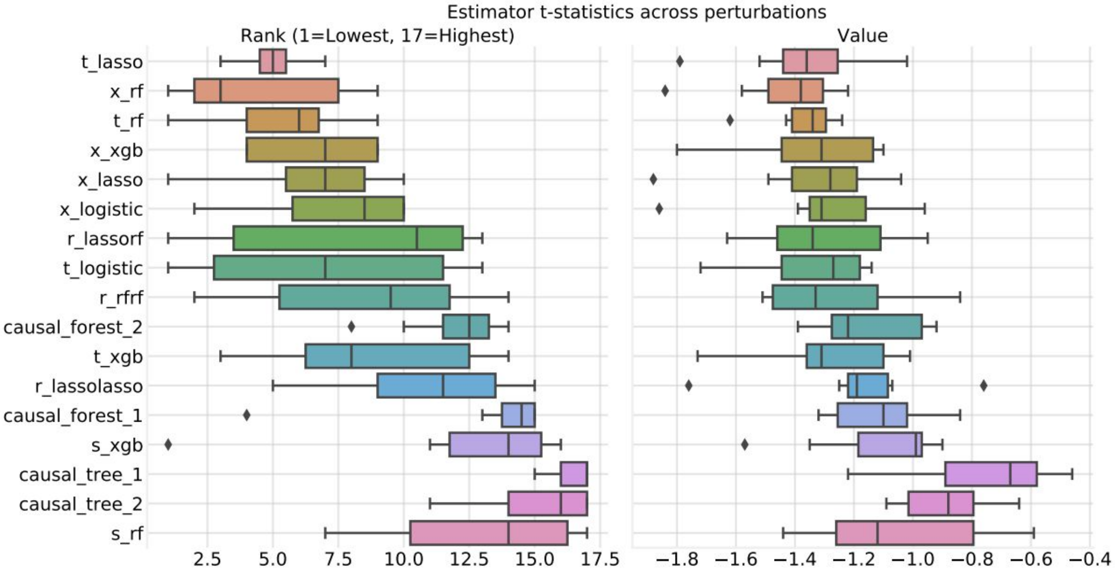
If the outputs of the models are **perfectly correlated**, the variance is **not** reduced.

Credit: Stanford CS 229 (Fall 2018, “Tree Ensembles”) ← more math and justification for why!

Ways to Ensemble

1. Use different algorithms
    a. Outputs of different algorithms are very independent
    b. Tedious to implement
2. Use different training sets
    a. Outputs based on different data can be very different
    b. Collecting more data is hard, otherwise you sacrifice the quantity of data available

Ensemble Example



Ref: [Stable Discovery of Interpretable Subgroups in Causal Studies]

Raaz Dwivedi*, Yan Shuo Tan*, Briton Park, Mian Wei, Kevin Horgan, David Madigan, Bin Yu

International Statistical Review, 2020

Review

- Decision trees are a straightforward and explainable model for regression or classification tasks
- However, they are sensitive to overfitting and high variance
- Some ways to handle overfitting
    - Regularization
    - Ensemble methods
- Ensemble methods combines the outputs of multiple models into a final answer