Lecture 4: Bootstrap and Random forest

Instructor: Prof. Shuai Huang

Industrial and Systems Engineering

University of Washington

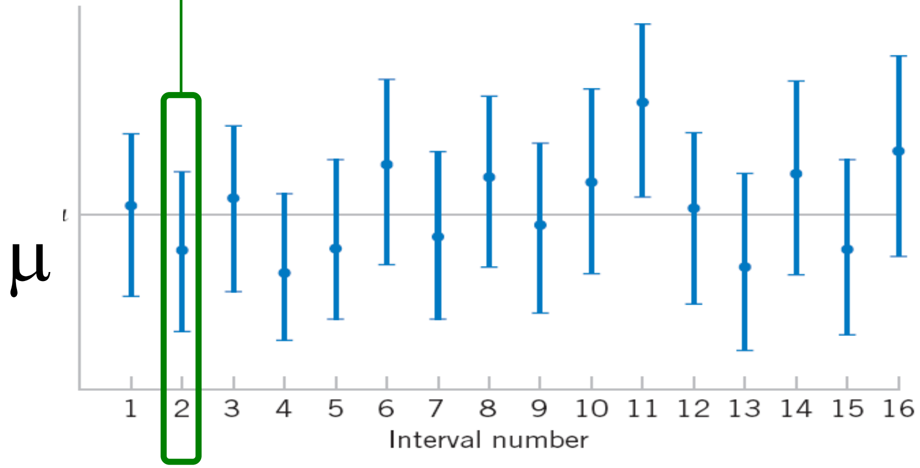
Review the rationale of hypothesis testing and confidence interval

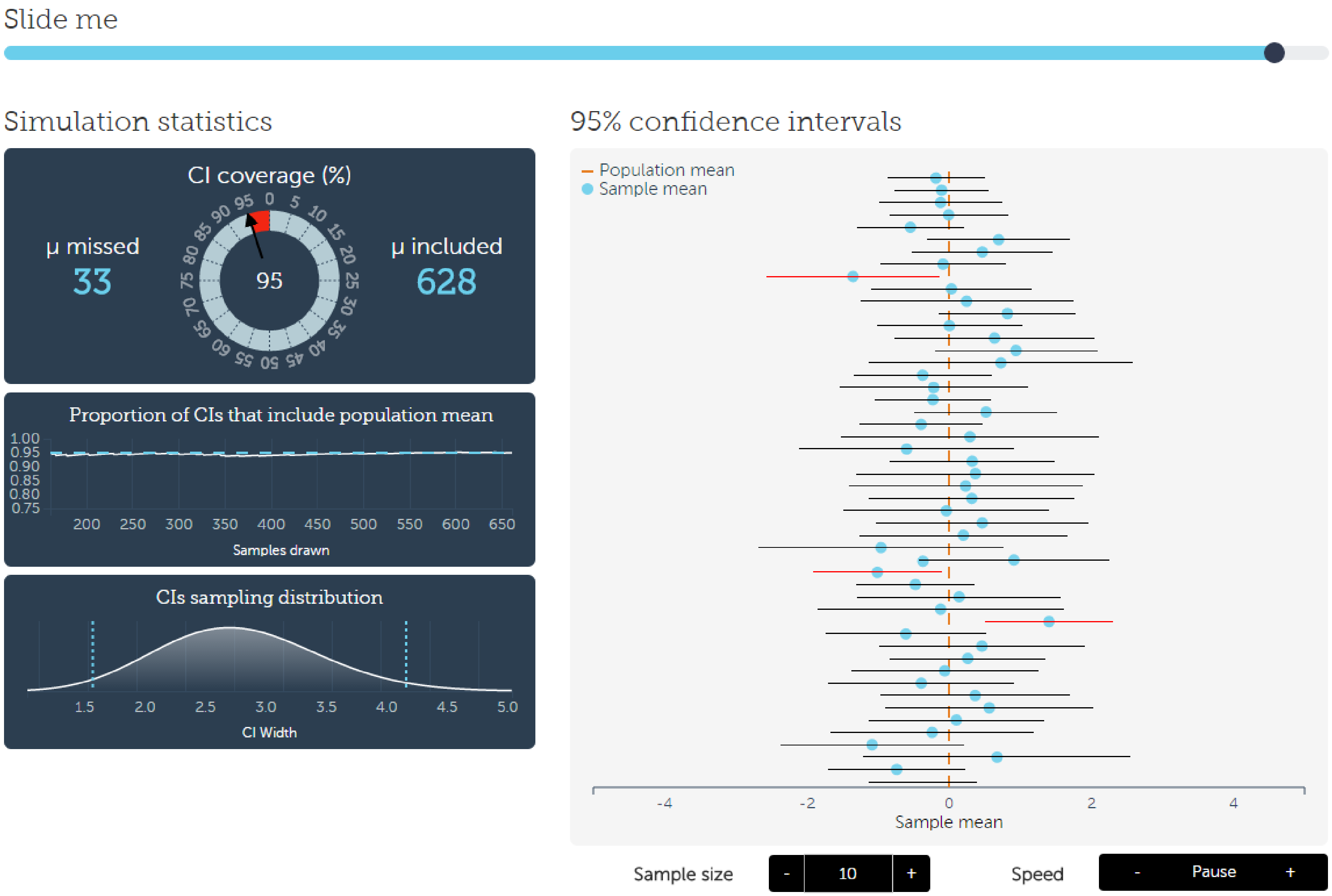
21 19 17 19 19 25 24 20 23 18 $\bar{x}_2 = 20.5$

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\text{for } \alpha = 0.05: 20.5 - (1.96)\frac{2.5}{\sqrt{10}} \le \mu \le 20.5 + (1.96)\frac{2.5}{\sqrt{10}}$$

$$18.95 \le \mu \le 22.05$$





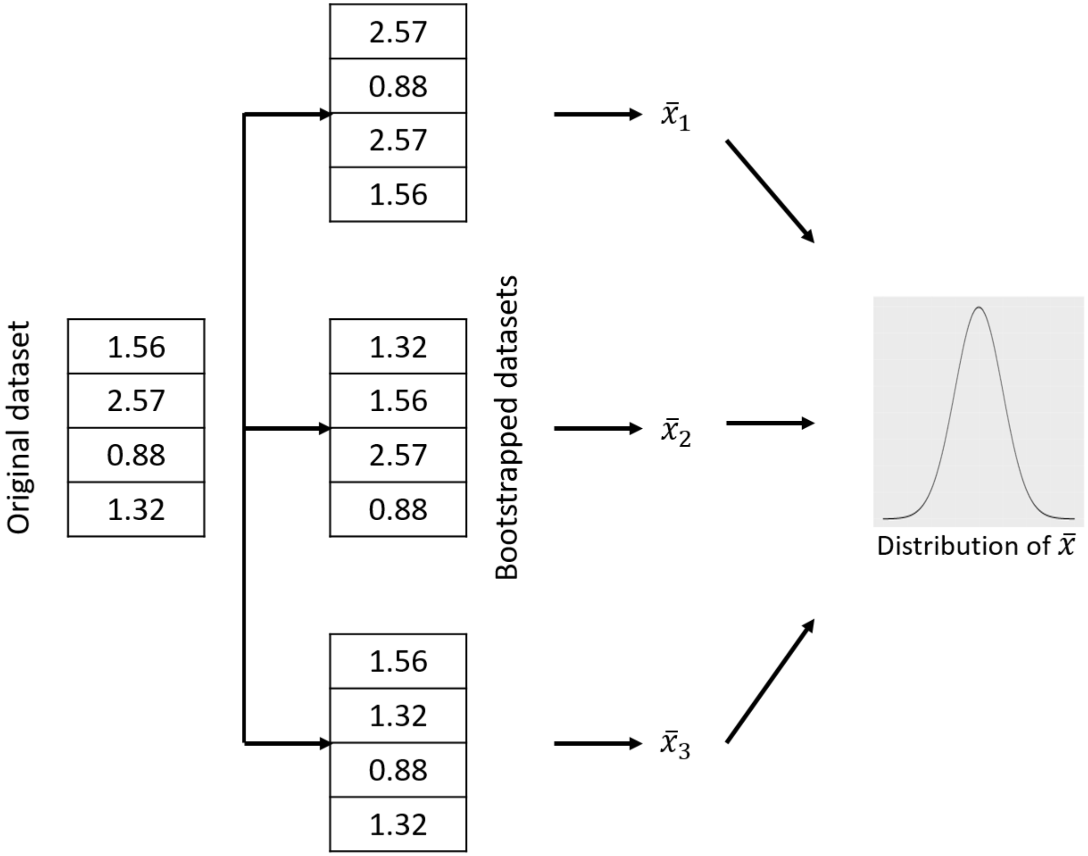
http://rpsychologist.com/d3/CI/

Without analytical tractability?

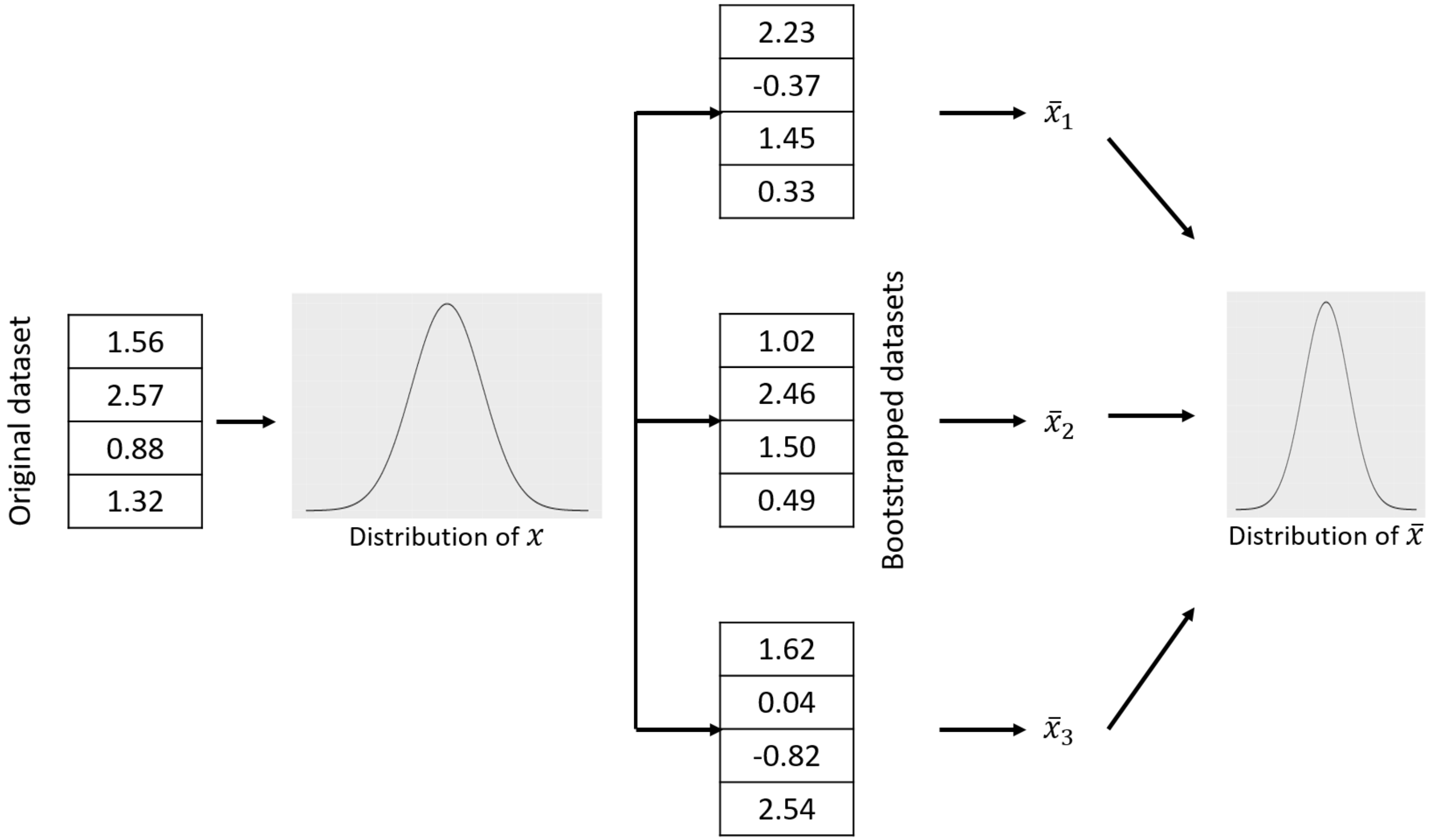
- The idea of Bootstrap to computationally mimic the sampling process

Complete dataset	X_1	X_2	X_3	X_4	X_5
Bootstrapped dataset 1	X_3	X_1	X_3	X_3	X_5
Bootstrapped dataset 2	X_5	X_5	X_3	X_1	X_2
Bootstrapped dataset 3	X_5	X_5	X_1	X_2	X_1
...					
Bootstrapped dataset K	X_4	X_4	X_4	X_4	X_1

A nonparametric Bootstrap scheme



A parametric Bootstrap scheme



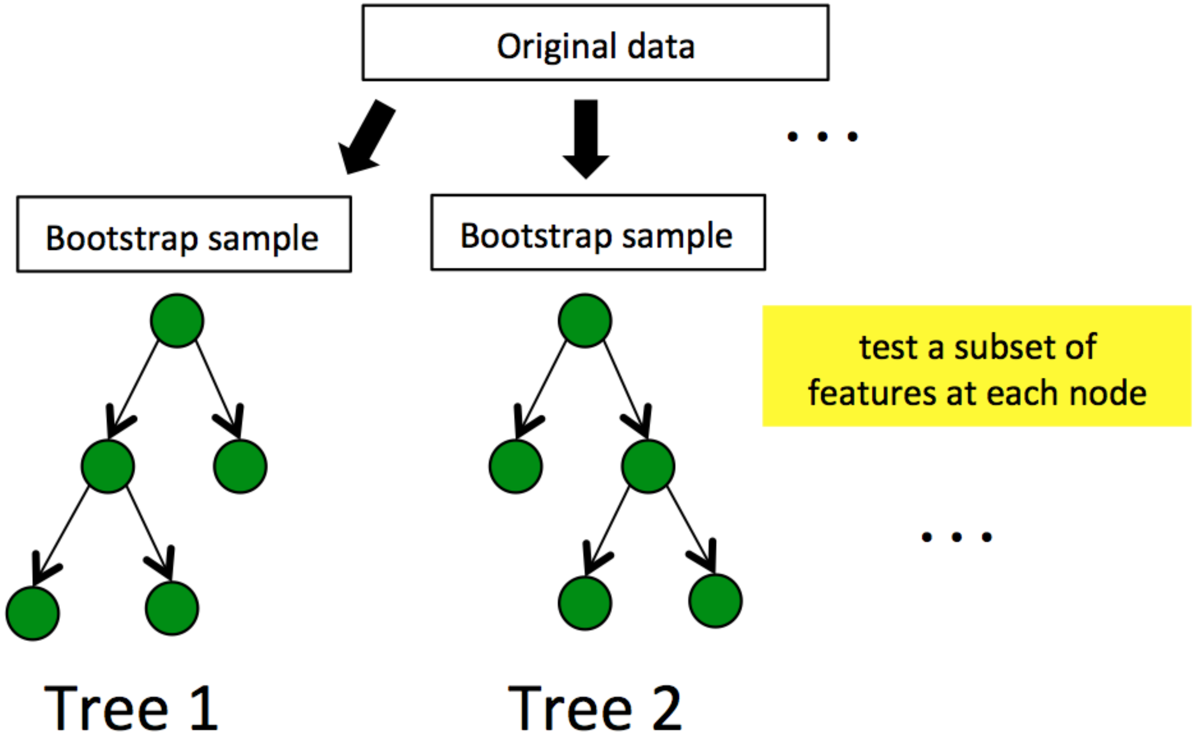
Bootstrap for regression models

- Option 1: we could simply resample the data points (i.e., the $(\boldsymbol{x}, y)$ pairs) similarly as the nonparametric Bootstrap scheme. Then, for each sampled dataset, we can fit a regression model and obtain the fitted regression parameters.
- Option 2: we could simulate new samples of X using the nonparametric Bootstrap method on the samples of X only. Then, for the new samples of X, we draw samples of Y using the fitted conditional distribution model $P(Y|X)$.
- Option 3: we could fix the X, only sample for Y. In this way we implicitly assume that the uncertainty of the dataset mainly comes from Y. To sample Y, we draw samples using the fitted conditional distribution model $P(Y|X)$.

R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets

Random forest



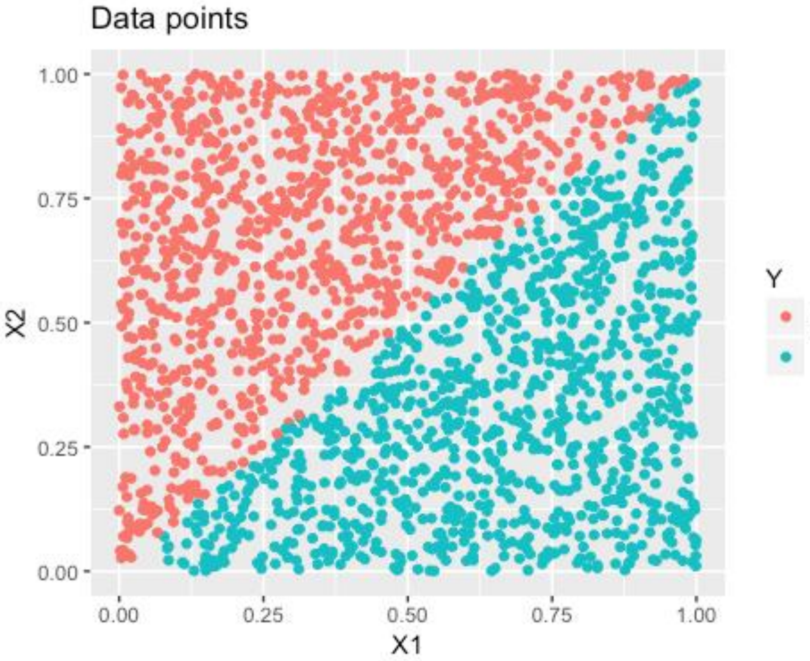
There are two main sources for randomness.

- First, each tree is built on a randomly selected set of samples by applying Bootstrap on the original dataset.
- Second, in building a tree, specifically in splitting a node in the tree, a subset of features is randomly selected to choose the best split.

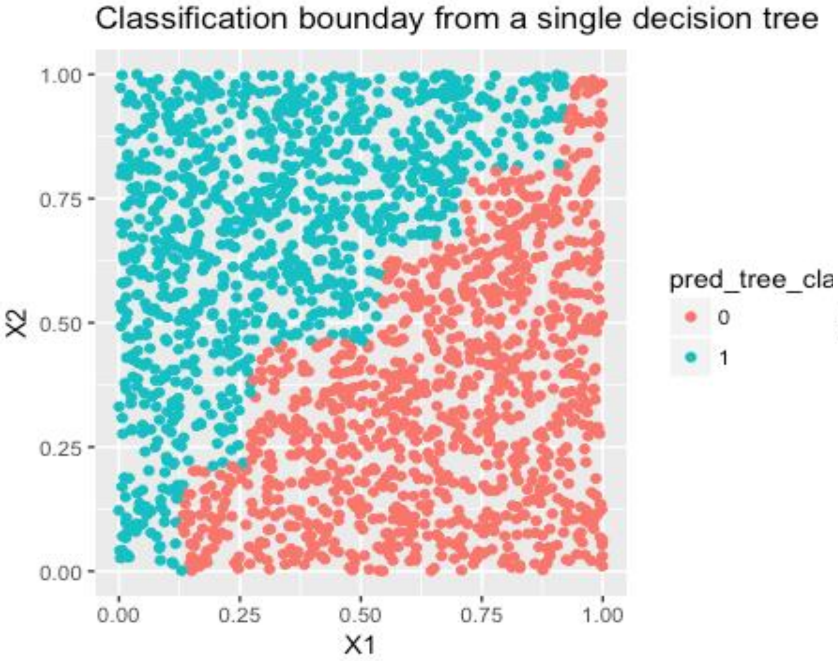
If randomness is troublesome, why we need to ask for it?

Why we need random forest?

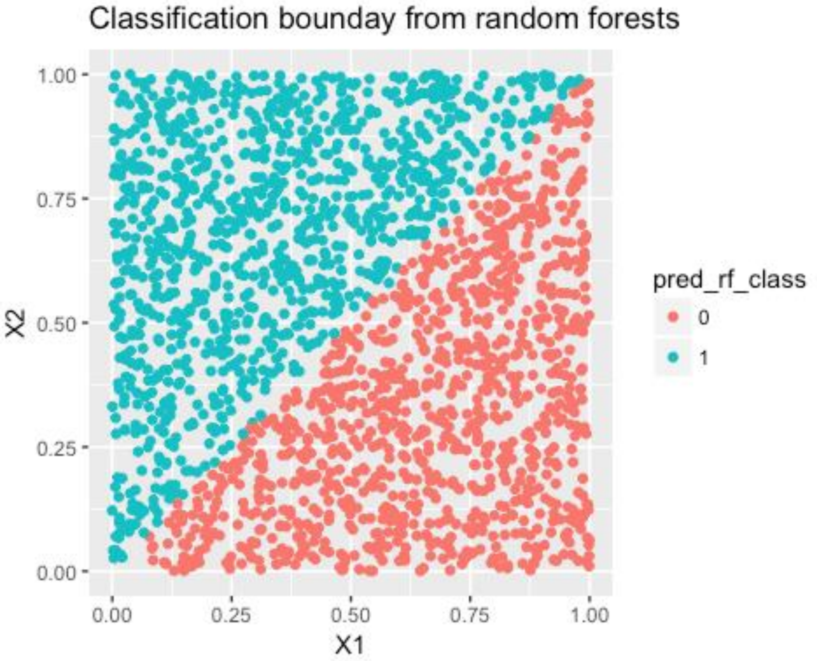
A simulated dataset



Prediction by decision tree



Prediction by random forest



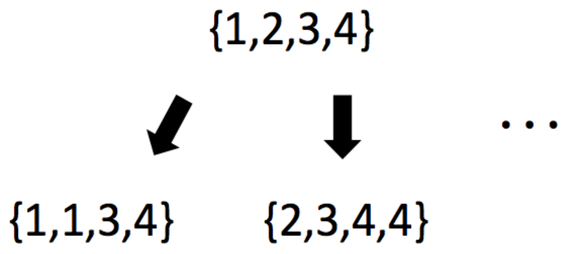
An exemplary data

- Thus, random forest is more of a systematically organized set of heuristics, rather than highly regulated algebraic operations derived from a mathematical characterization.

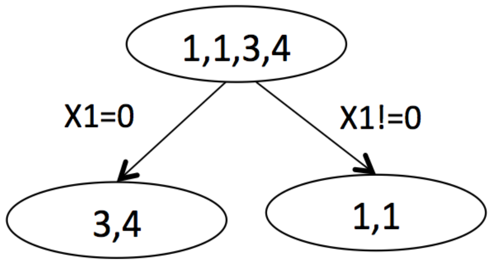
A dataset with 4 instances

ID	X1	X2	Class
1	1	1	C0
2	1	0	C1
3	0	1	C1
4	0	0	C0

Bootstrap the dataset



Build tree on each dataset

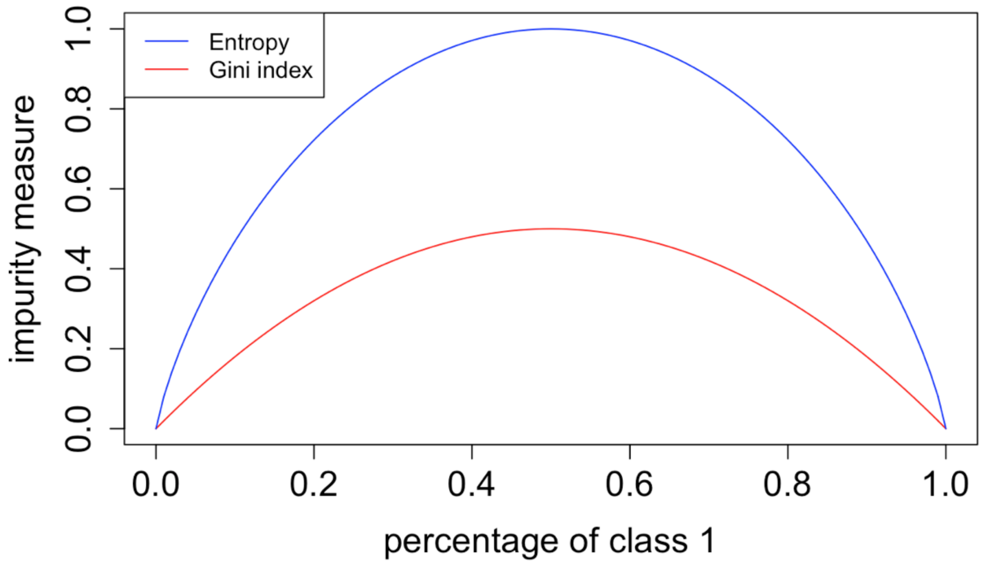


Gini index

- The R package “randomforest” uses the Gini index to measure impurity
- The Gini index is defined as

$$Gini = \sum_{c=1}^{C} p_c(1 - p_c),$$

where C is the number the classes in the dataset, and p_c is the proportion of data instances that come from the class c.

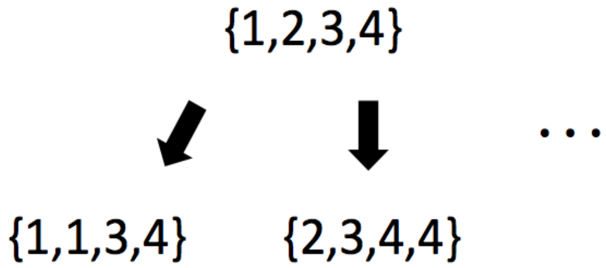


Gini gain

Similar as the information gain, the **Gini gain** can be defined as

$$\nabla Gini = Gini - \ w_i Gini_i,$$

where $Gini$ is the Gini index at the node to be split; w_i and $Gini_i$, are the proportion of samples and the Gini index at the i^{th} children node, respectively.

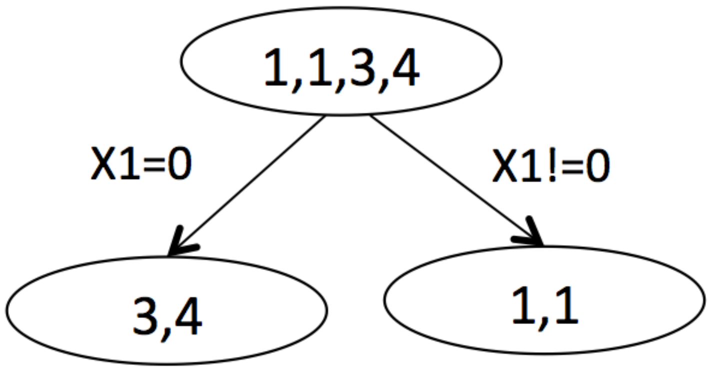


Apply the Gini gain on the exemplary data

- The possible splitting rule candidates include four options: $X_1 = 0$, $X_2 = 0$, $X_1 = 1$ and $X_2 = 1$. Since both variables have two distinct values, both splitting rules $X_1 = 0$ and $X_1 = 1$ will produce the same children nodes, and both splitting rules $X_2 = 0$ and $X_2 = 1$ will produce the same children nodes.
- Therefore, we can reduce the possible splitting rule candidates to two: $X_1 = 0$ and $X_2 = 0$.
- Further, random forest randomly selects variables for splitting a node. In general, for a data set with p predictor variables, $\sqrt{p}$ variables are randomly selected for splitting.
- In our simple example, as there are two variables, we assume that X_1 is randomly selected for splitting the root node.

Apply the Gini gain on the exemplary data – cont'd

- Thus, $X_1 = 0$ is used for splitting the root node



- The Gini index of the root node is calculated as

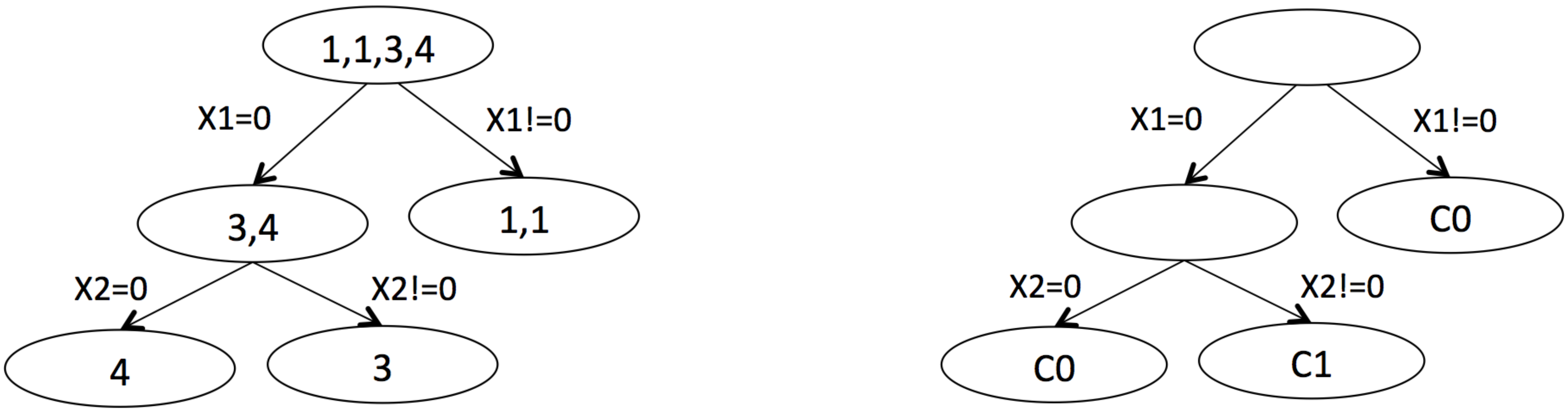
$$\frac{3}{4} * \frac{1}{4} + \frac{1}{4} * \frac{3}{4} = 0.375.$$

- The Gini gain of this split can be calculated as

$$0.375 - 0.5 * 0 - 0.5 * 0.5 = 0.125.$$

Apply the Gini gain on the exemplary data – cont'd

- Let's continue to grow the tree. Now, at the internal node containing data {3,4}, assume that X_2 is randomly selected. The node can be further split



Why randomness?

- The concept as “weak classifier” is very important in understanding random forest
- Assuming that the trees in random forests are independent, and each tree has an accuracy of 0.6.
- For 100 trees, the probability of random forests to make the right prediction reaches as high as 0.97:

$$\sum_{k=51}^{100} C(n,k) * \; 0.6^{k} * 0.4^{100-k}.$$

- Note that, the assumption of the independency between the trees in random forests is the key here. This does not hold in reality in a strict sense. However, the randomness added to each tree makes them less correlated.
- This is probably not the answer for why it has to be this way, but it provides an explanation that why it works!

R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets