Random Forest

Random Forest Classifier

- Why a ***forest***?

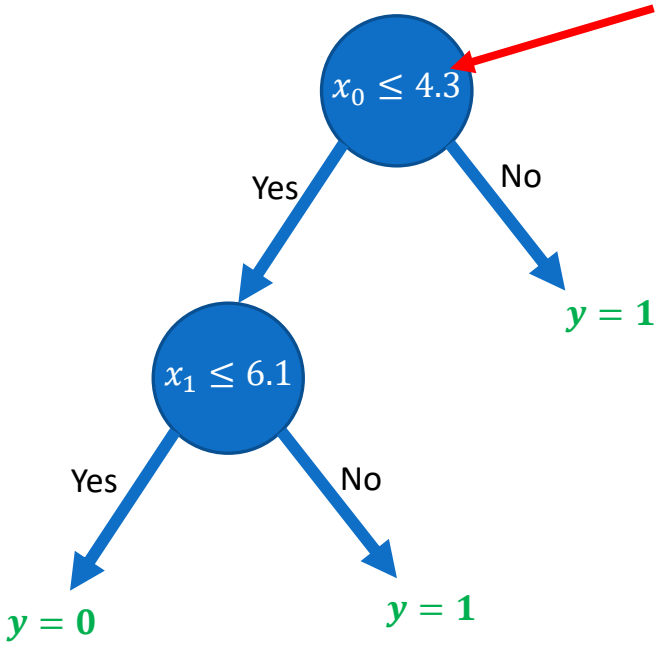
- Why ***Random***?

Random Forest Classifier

- Consider the following dataset.

ID	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

Side Note: How did we know this threshold?



Idea: Build many more trees from the data…

A Slight change/error in the input data can cause a very different decision tree, hence, affecting its generalizability.

Random Forest Classifier

- Consider the following dataset.

Step 1: Create more than one datasets by picking up instances at random

ID	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

ID	ID	ID	ID
2	2	4	3
0	1	1	3
2	3	3	2
4	1	0	5
5	4	0	1
5	4	2	2

Bootstrapped Datasets

Records are being repeated in a dataset because we are using random sampling with replacement.

In each dataset, there should be equal number of records.

Random Forest Classifier

- Consider the following dataset.

Step 2: Select a subset of features randomly for every dataset

ID	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

ID	ID	ID	ID
2	2	4	3
0	1	1	3
2	3	3	2
4	1	0	5
5	4	0	1
5	4	2	2
x_0, x_1	x_2, x_3	x_2, x_4	x_1, x_3

Random Forest Classifier

- Consider the following dataset.

Step 2: Select a subset of features randomly for every dataset

ID	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

ID	x_0	x_1	y
2	2.7	4.8	0
0	4.3	4.9	0
2	2.7	4.8	0
4	6.5	2.9	1
5	2.7	6.7	1
5	2.7	6.7	1

ID	x_2	x_3	y
2	4.1	5.0	0
1	5.9	5.5	0
3	4.5	3.9	1
1	5.9	5.5	0
4	4.7	4.6	1
4	4.7	4.6	1

ID	x_2	x_4	y
4	4.7	6.1	1
1	5.9	5.9	0
3	4.5	5.9	1
0	4.1	5.5	0
0	4.1	5.5	0
2	4.1	5.6	0

ID	x_1	x_3	y
3	4.4	3.9	1
3	4.4	3.9	1
2	4.8	5.0	0
5	6.7	5.3	1
1	6.1	5.5	0
2	4.8	5.0	0

Random Forest Classifier

- Consider the following dataset.

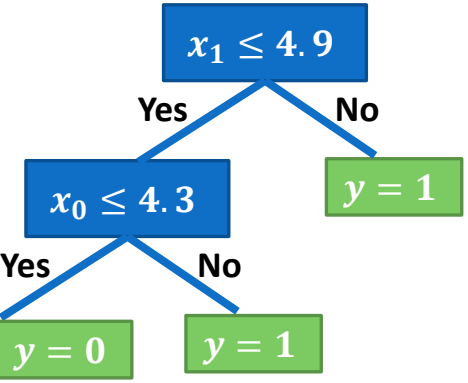
Step 3: Build a tree for each bootstrapped dataset

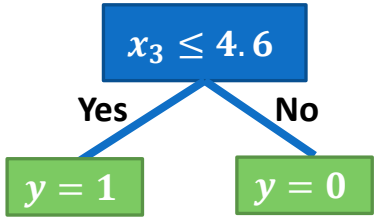
ID	x_0	x_1	y
2	2.7	4.8	0
0	4.3	4.9	0
2	2.7	4.8	0
4	6.5	2.9	1
5	2.7	6.7	1
5	2.7	6.7	1

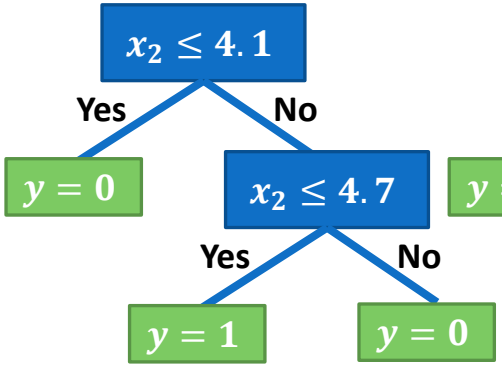
ID	x_2	x_3	y
2	4.1	5.0	0
1	5.9	5.5	0
3	4.5	3.9	1
1	5.9	5.5	0
4	4.7	4.6	1
4	4.7	4.6	1

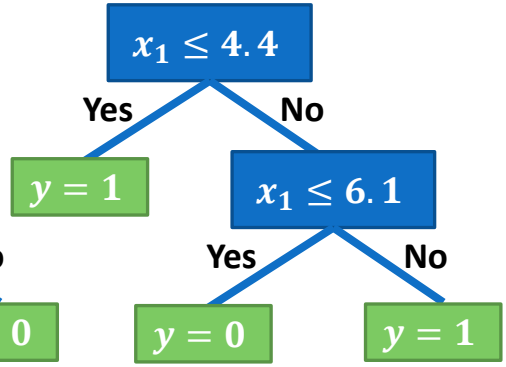
ID	x_2	x_4	y
4	4.7	6.1	1
1	5.9	5.9	0
3	4.5	5.9	1
0	4.1	5.5	0
0	4.1	5.5	0
2	4.1	5.6	0

ID	x_1	x_3	y
3	4.4	3.9	1
3	4.4	3.9	1
2	4.8	5.0	0
5	6.7	5.3	1
1	6.1	5.5	0
2	4.8	5.0	0







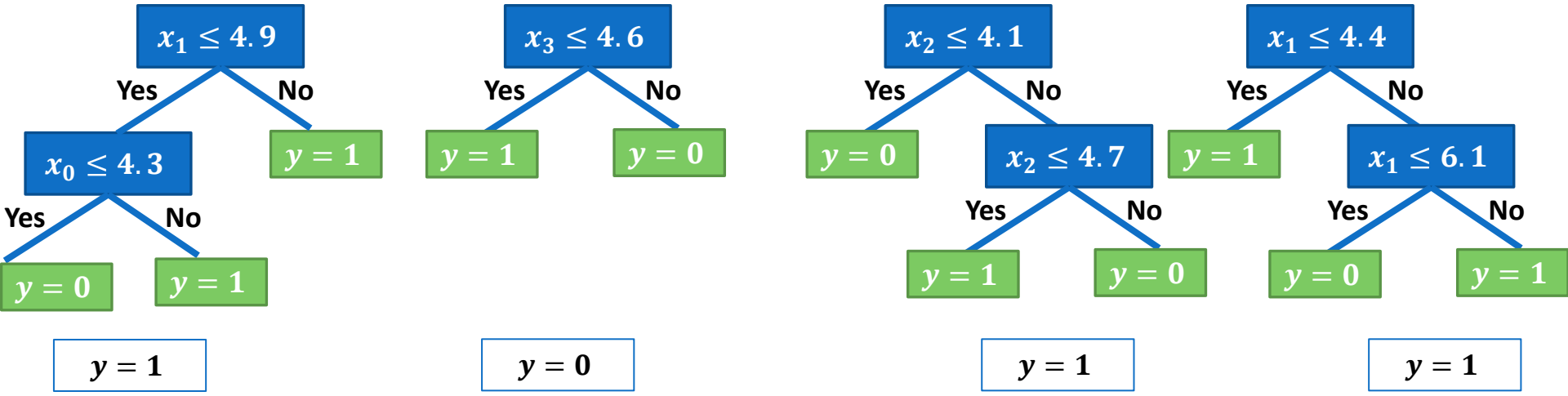


Note: A feature can repeat in a tree if it still has distinct values left (for that path).

Random Forest Classifier

- Consider following test record

Step 4: Use all trees for predictions…



x_0	x_1	x_2	x_3	x_4	y
2.8	6.2	4.3	5.3	5.5	?

Step 5: Consider majority vote to label the test record (Aggregating)

Bootstrapping + Aggregating is called "Bagging"

Random Forest Classifier

- Why a ***forest***?
  - Because there are multiple trees, making one classifier as a whole

- Why ***Random***?
  - Because of bootstrapping (random sampling) and random feature selection

- **Why do Bootstrapping and feature selection?**
  - To ensure that each tree is different from the other in the forest
  - Helps to reduce correlation amongst the trees
  - To ensure that few trees train on less important features, while few trained on best features. At the end, the trees trained on less important features will give bad predictions; thus, their effect will smooth out when taking majority vote.

Random Forest Classifier

- **How many features should be selected for each tree?**
  - We used 2 features
  - Research suggests that using square root of total number of features, or log of the total number of features work well

DT and RF for Regression

- **How to use decision trees and random forest for regression problems?**

Book Reading

- Murphy – Chapter 1, Chapter 14
- Tom Mitchel (TM) – Chapter 3