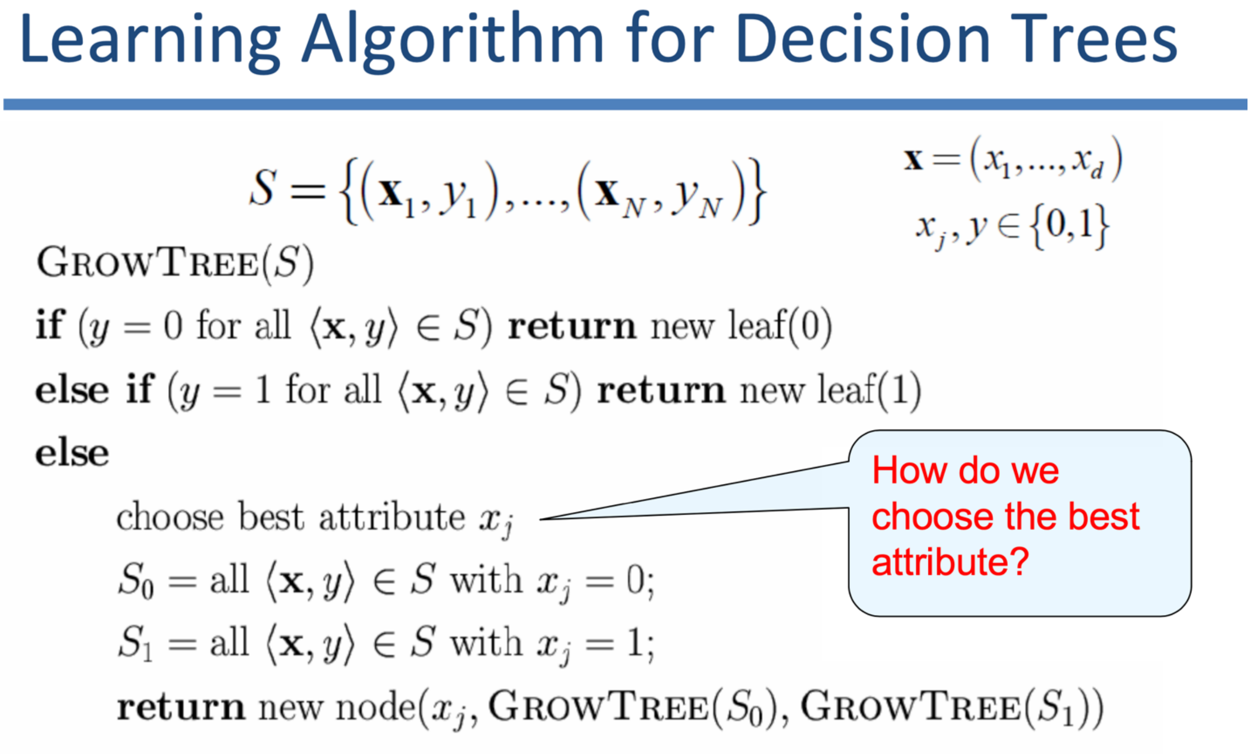
Decision Tree Tutorial

1. Question and Answer for Decision Trees (Lectures and Homework)

- There is a 1-to-1 map between decision trees and Boolean functions.
- Decision trees are not unique. Finding the simplest (optimal) decision tree is NP-complete.

2. Decision Tree Algorithm

1. Begin with the original set S as the root node of the tree.
2. On each iteration of the algorithm, iterate through the attributes in set S and calculate the Information Gain (IG) of each attribute.
3. Select the attribute which has the largest Information Gain.
4. Split the set S by the selected attribute to produce a subset of the data.
5. Repeat on each subset.
6. End when the subset at a node is pure or when splitting is no longer effective.



Learning Algorithm for Decision Trees

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$$

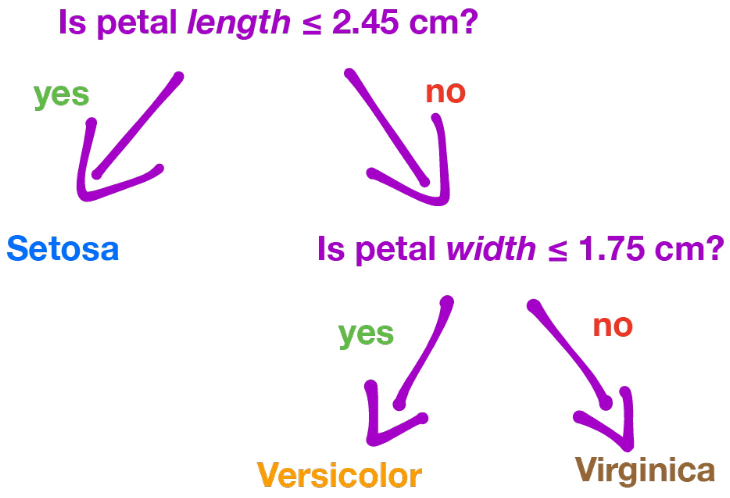
$$\mathbf{x} = (x_1, \ldots, x_d)$$

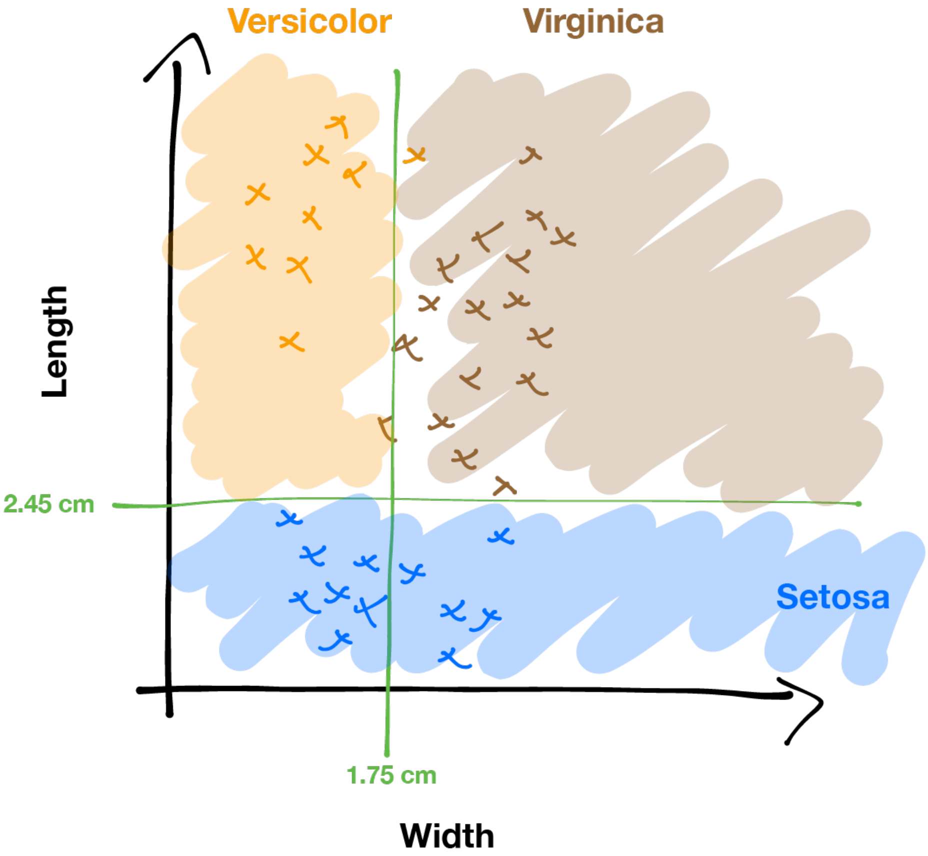
$$x_j, y \in \{0,1\}$$

```
GrowTree($S$)
if ($y = 0$ for all $\langle \mathbf{x}, y \rangle \in S$) return new leaf(0)
else if ($y = 1$ for all $\langle \mathbf{x}, y \rangle \in S$) return new leaf(1)
else
    choose best attribute $x_j$
    $S_0$ = all $\langle \mathbf{x}, y \rangle \in S$ with $x_j = 0$;
    $S_1$ = all $\langle \mathbf{x}, y \rangle \in S$ with $x_j = 1$;
    return new node($x_j$, GrowTree($S_0$), GrowTree($S_1$))
```

How do we choose the best attribute?

3. Decision Tree Decision Regions/Boundaries





4. Are all decision trees equal? Consider the following dataset.

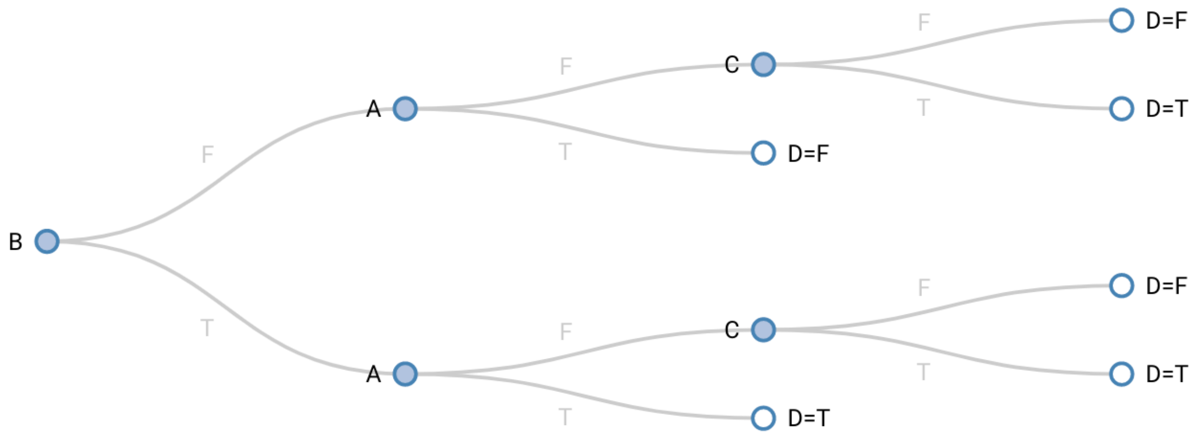
A	B	C	D
F	F	F	**F**
F	F	T	**T**
F	T	F	**F**
F	T	T	**T**
T	F	F	**F**
T	F	T	**F**
T	T	F	**T**
T	T	T	**T**

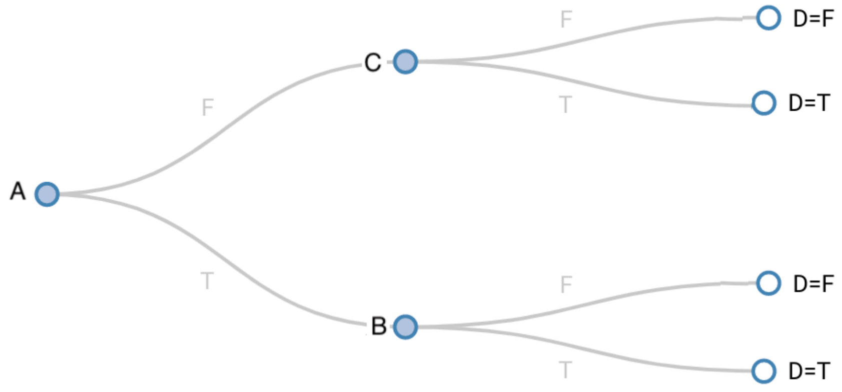
Note that D=**((A and B) or (not A and C))**

Both Decision Trees below represent this dataset.

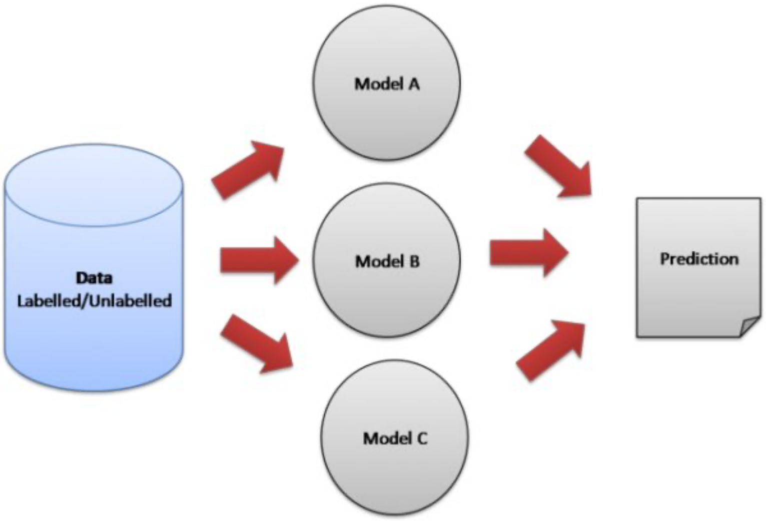
Question: Should we split on A? Why or Why not?

Question: Should we split on B or C first?

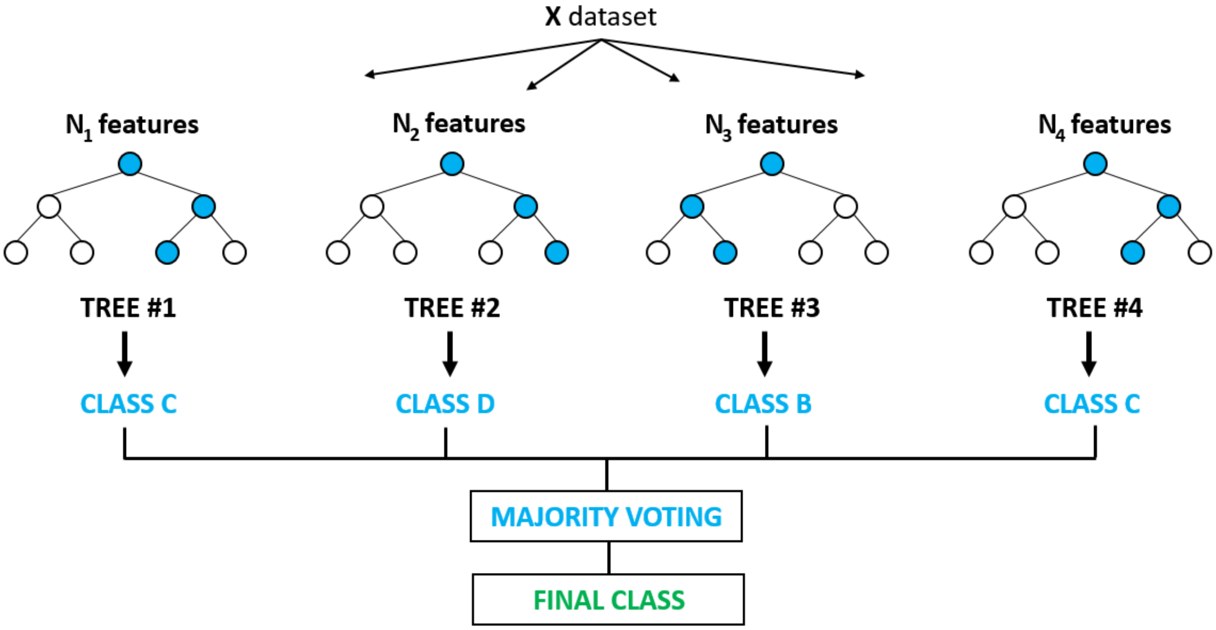




5. Ensemble learning – a powerful technique to improve performance.



Random Forest



- The averaging in Random Forest prevents overfitting. It reduces variance. You can think of variance as the error due to sensitivity to the input signals.
- Random forest does not increase the bias much. Think of bias as the error due to underfitting.
- Underfitting is bad as it means we have not learned enough from our data.
- Overfitting is bad as it means we are too sensitive to our data.
- We want both low bias (no underfitting) and low variance (no overfitting)!