Brandeis

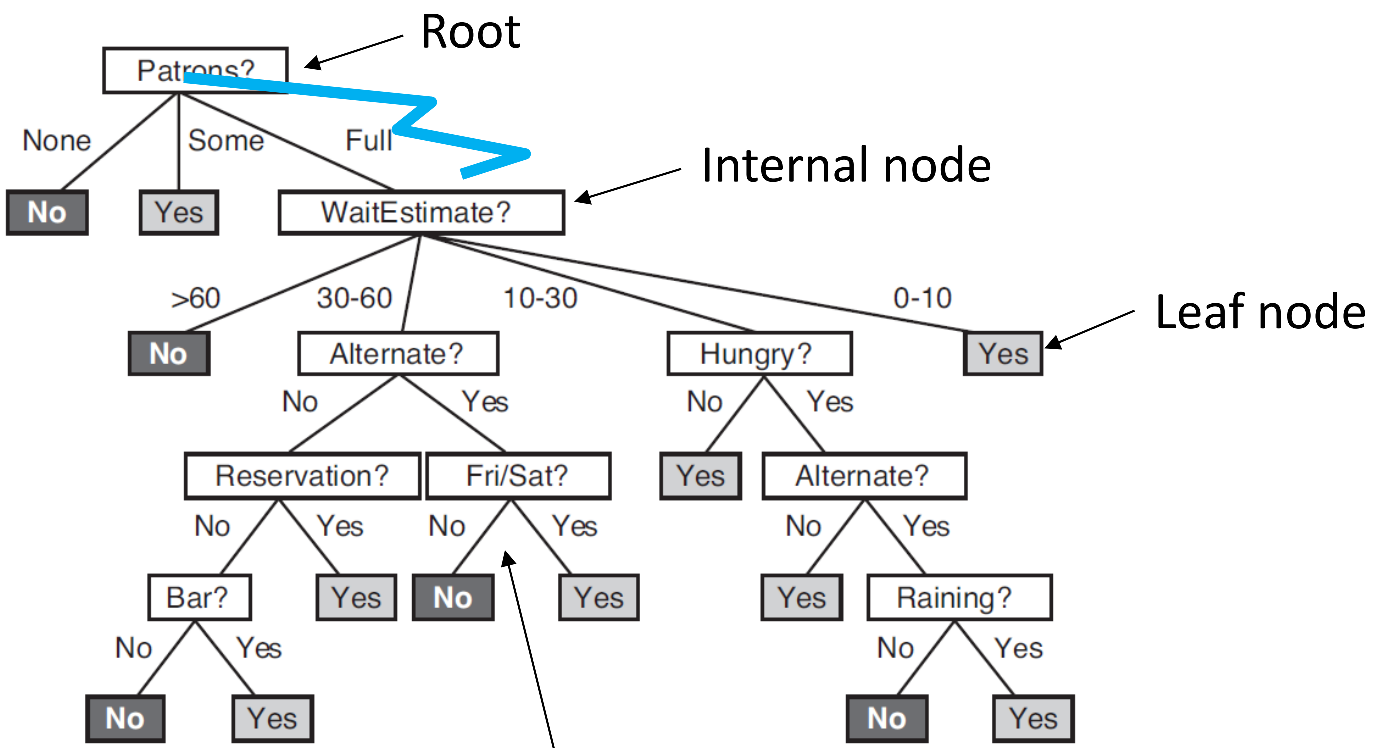
COSI 104a Introduction to machine learning

Chapter 9 – Tree

Instructor: Dr. Hongfu Liu
Email: hongfuliu@brandeis.edu

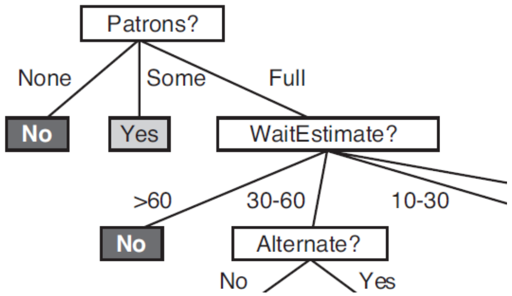
Main source: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

A Decision Tree Example



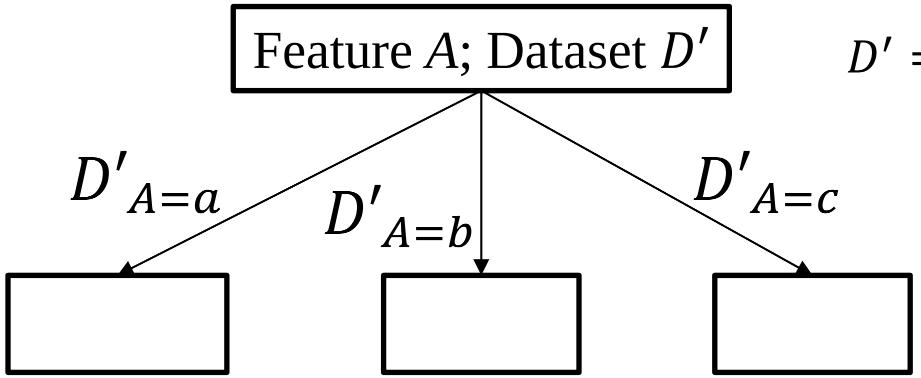
(Patrons = Full) ^ (WaitEstimate = 30-60) ^ (Alternate = Yes) ^ (Fri/Sat = No)

Basic Idea of Decision-Tree Learning Algorithms



1. Recursively builds a tree, starts with a root node and the full dataset *D*.

2. At each node $\boldsymbol{v}$ and the associated dataset $D' \subseteq D$
   - Select a feature *A*.
   - Uses *A* to make local decision at $\boldsymbol{v}$. If decide $\boldsymbol{v}$ as a leaf node, stop.
   - Divide D' into subsets accordingly to *A*, and create one child node of v for each subset.
   - Repeat 2 for each child node



$$D' = D'_{A=a} \cup D'_{A=b} \cup D'_{A=c}$$

An Example Dataset

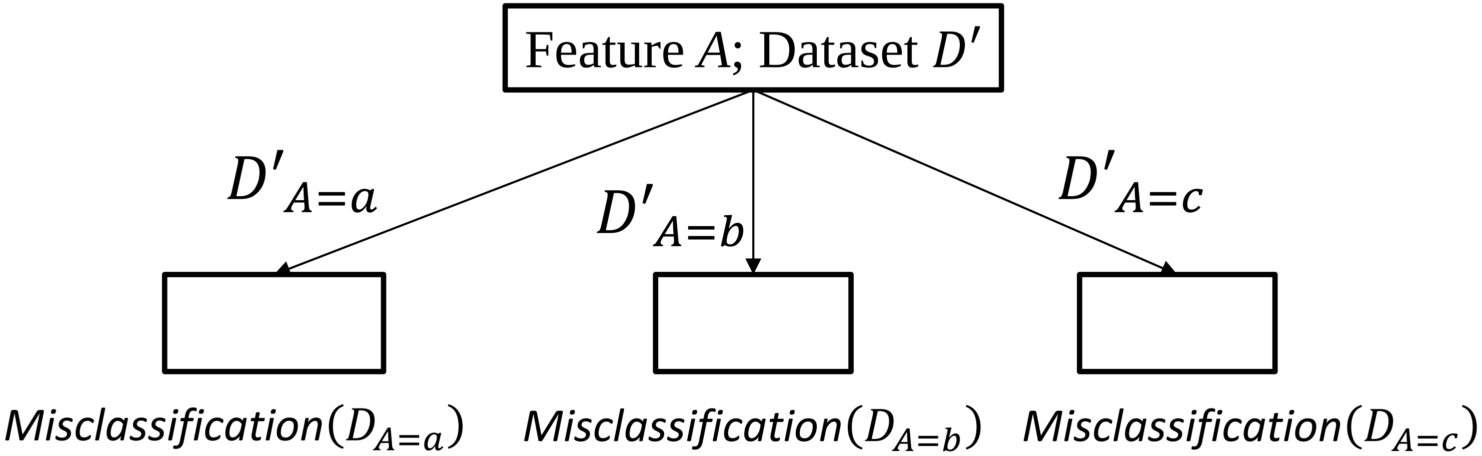
Example	Attributes										Wait
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	
1	*Yes*	*No*	*No*	*Yes*	*Some*	$$$	*No*	*Yes*	*French*	0-10	*Yes*
2	*Yes*	*No*	*No*	*Yes*	*Full*	$	*No*	*No*	*Thai*	30-60	*No*
3	*No*	*Yes*	*No*	*No*	*Some*	$	*No*	*No*	*Burger*	0-10	*Yes*
4	*Yes*	*No*	*Yes*	*Yes*	*Full*	$	*Yes*	*No*	*Thai*	10-30	*Yes*
5	*Yes*	*No*	*Yes*	*No*	*Full*	$$$	*No*	*Yes*	*French*	> 60	*No*
6	*No*	*Yes*	*No*	*Yes*	*Some*	$$	*Yes*	*Yes*	*Italian*	0-10	*Yes*
7	*No*	*Yes*	*No*	*No*	*None*	$	*Yes*	*No*	*Burger*	0-10	*No*
8	*No*	*No*	*No*	*Yes*	*Some*	$$	*Yes*	*Yes*	*Thai*	0-10	*Yes*
9	*No*	*Yes*	*Yes*	*No*	*Full*	$	*Yes*	*No*	*Burger*	> 60	*No*
10	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	$$$	*No*	*Yes*	*Italian*	10-30	*No*
11	*No*	*No*	*No*	*No*	*None*	$	*No*	*No*	*Thai*	0-10	*No*
12	*Yes*	*Yes*	*Yes*	*No*	*Full*	$	*No*	*No*	*Burger*	30-60	*Yes*

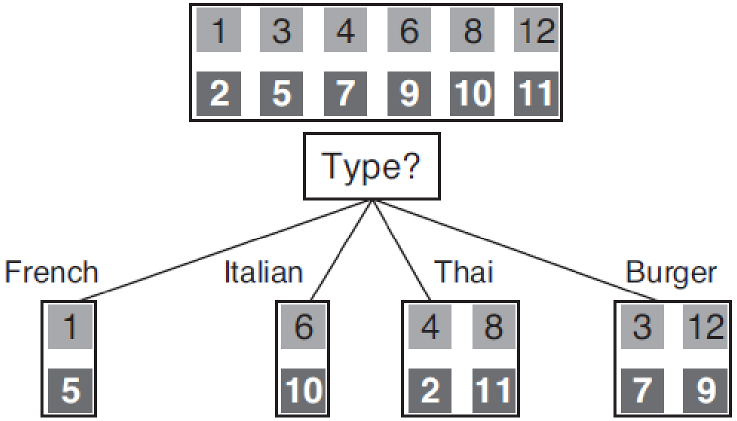
Select the Best Feature I

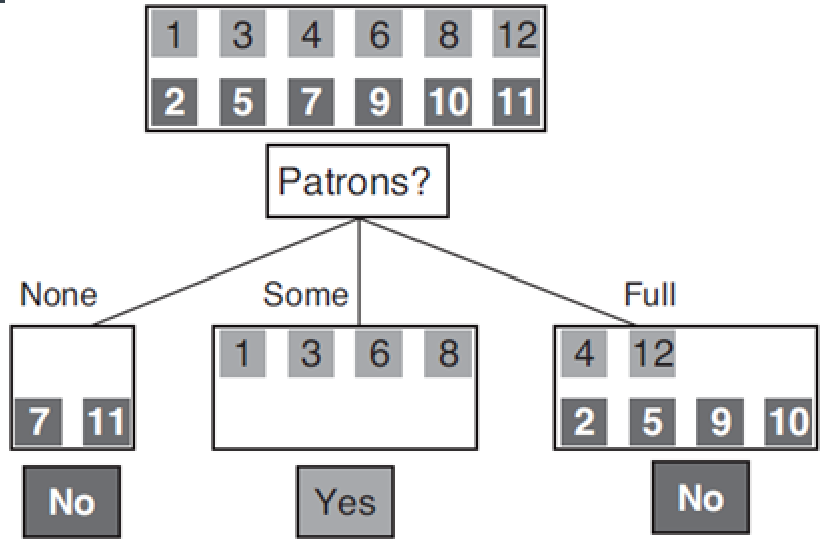
Misclassification error

$$\sum_{a \in Values(A)} \frac{|D_{A=a}|}{|D|} Misclassification(D_{A=a})$$

$Misclassification(D_{A=a})$ can be as simple as the error of majority voting







$$Misclassification([]) = 0.5$$

$$Misclassification(Type) = \frac{2}{12} \times 0.5 + \frac{2}{12} \times 0.5 + \frac{4}{12} \times 0.5 + \frac{4}{12} \times 0.5 = 0.5$$

$$Misclassification(Patrons) = \frac{2}{12} \times 0 + \frac{4}{12} \times 0 + \frac{6}{12} \times \frac{2}{6} = \frac{1}{6}$$

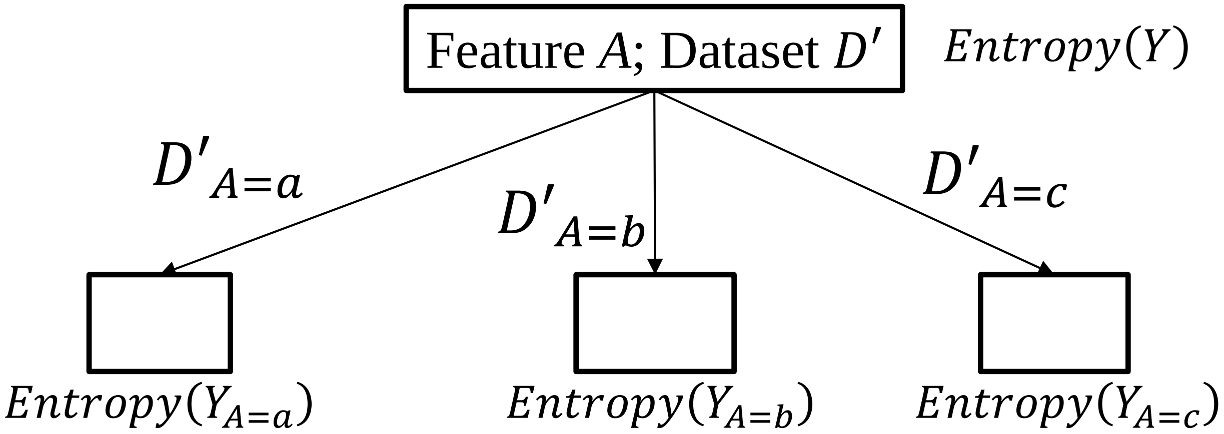
Select the Best Feature II

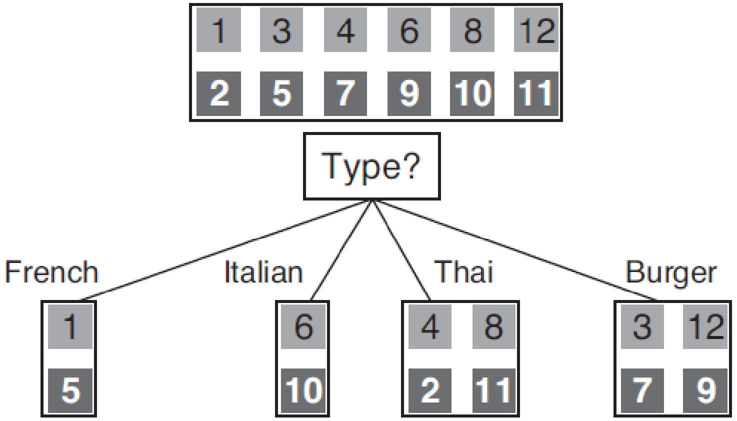
- Entropy – characterizes the amount of uncertainty.

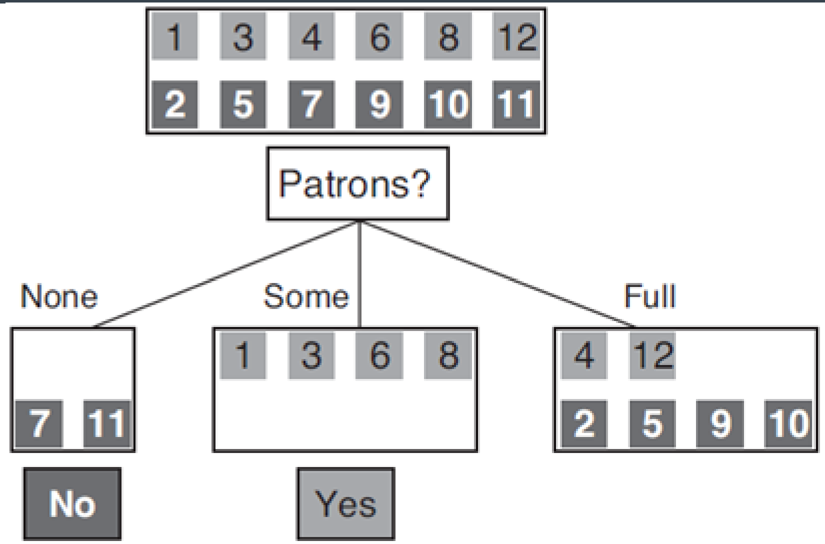
$$Entropy(Y) = \sum_{y} P(y) \log_2 \frac{1}{P(Y)} \quad = P(+) \log_2 \frac{1}{P(+)} + P(-) \log_2 \frac{1}{P(-)}$$

- Information Gain

$$Gain(Y, A) = Entropy(Y) - \sum_{a \in Values(A)} \frac{|D_{A=a}|}{|D|} Entropy(Y_{A=a})$$







$$Gain([]) = 1 - \frac{12}{12} H\left(\frac{1}{2}\right) = 0$$

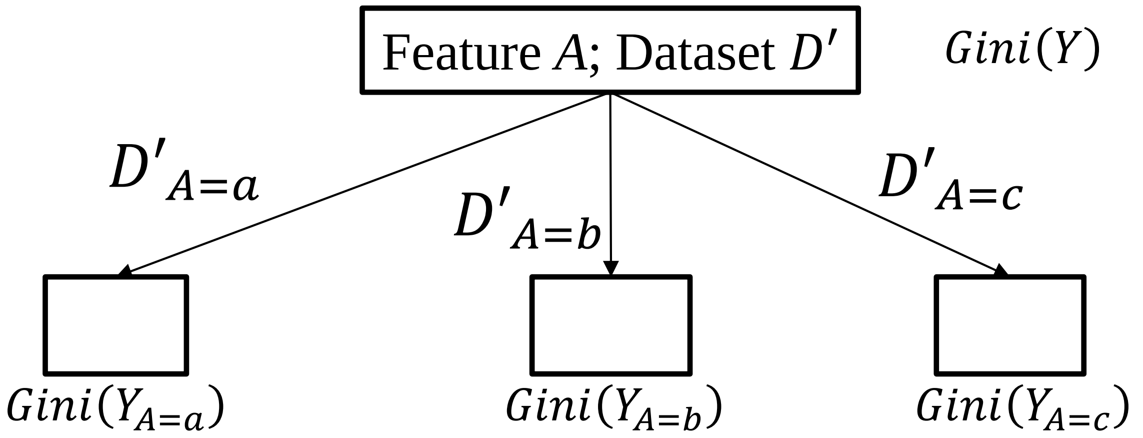
$$Gain(Type) = 1 - \left[\frac{2}{12} H\left(\frac{1}{2}\right) + \frac{2}{12} H\left(\frac{1}{2}\right) + \frac{4}{12} H\left(\frac{1}{2}\right) + \frac{4}{12} H\left(\frac{1}{2}\right)\right] = 0$$

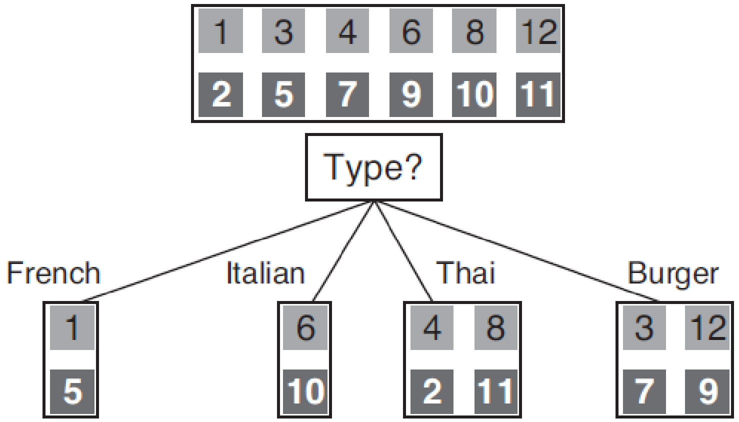
$$Gain(Patrons) = 1 - \left[\frac{2}{12} H\left(\frac{0}{2}\right) + \frac{4}{12} H\left(\frac{4}{4}\right) + \frac{6}{12} H\left(\frac{2}{6}\right)\right] = 0.54$$

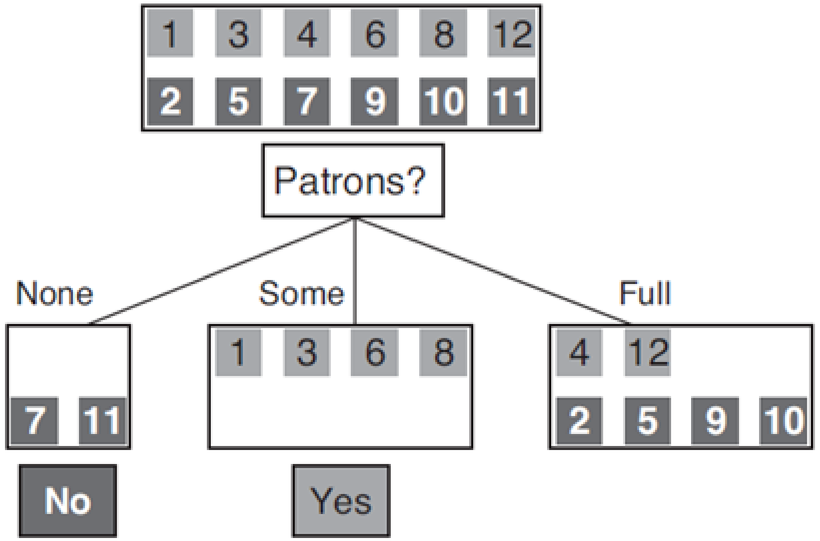
Select the Best Feature III

Impurity $$Gini(Y) = 1 - \sum_{y} P(y)^2$$

Splitting based on attribute *A* $$\sum_{a \in Values(A)} \frac{|D_{A=a}|}{|D|} Gini(Y_{A=a})$$







$$Gini([]) = 1 - 0.5^2 - 0.5^2 = 0.5$$

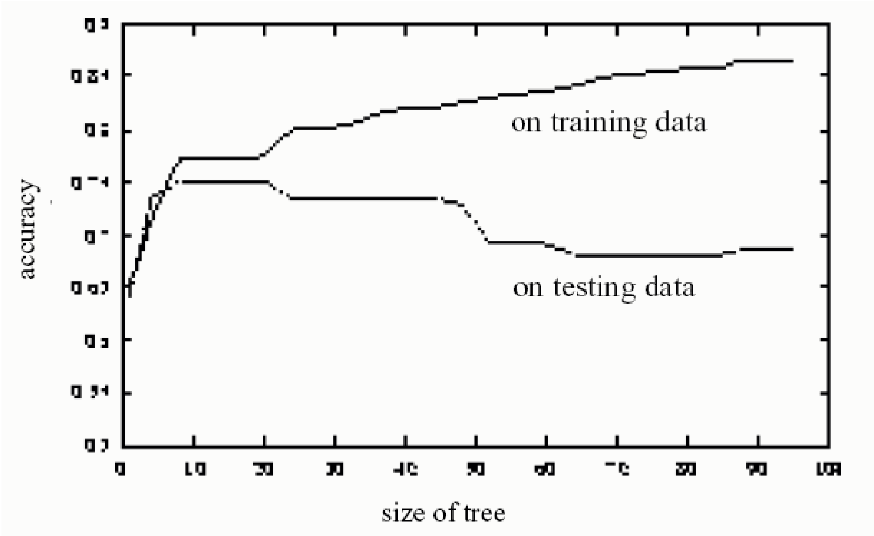
$$Gini(Type) = \frac{2}{12} \times 0.5 + \frac{2}{12} \times 0.5 + \frac{4}{12} \times 0.5 + \frac{4}{12} \times 0.5 = 0.5$$

$$Gini(Patrons) = \frac{2}{12} \times 0 + \frac{4}{12} \times 0 + \frac{6}{12} \times \left(1 - \frac{1}{9} - \frac{4}{9}\right) = \frac{2}{9}$$

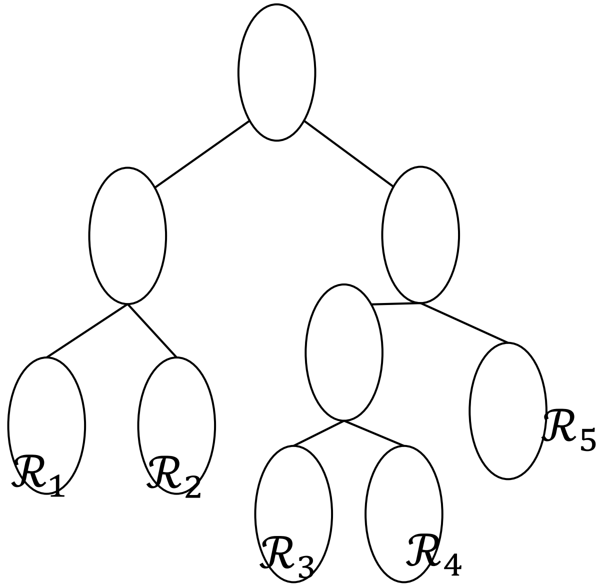
Overfitting Problem

Usually, the training data set is relatively small with respect to the complexity of the model

Large tree, some leaves only represent a tiny set of observations. (not robust to noise)



Pruning by Examining Decision Regions



Each leave represents a decision region $\mathcal{R}_t$

The optimal prediction for $\mathcal{R}_t$ is

$$h_t = \frac{1}{|\mathcal{R}_t|} \sum_{\vec{x}_n \in \mathcal{R}_t} y_n$$

The error of $\mathcal{R}_t$ is

$$Q_t = \sum_{\vec{x}_n \in \mathcal{R}_t} error(h_t, y_n)$$

Objective function:

$$C = \sum_{t=1}^{L} Q_t + \lambda L$$

L is the number of regions

Note: after pruning a node, the corresponding decision region is merged with another region whose Q_t needs to be updated.

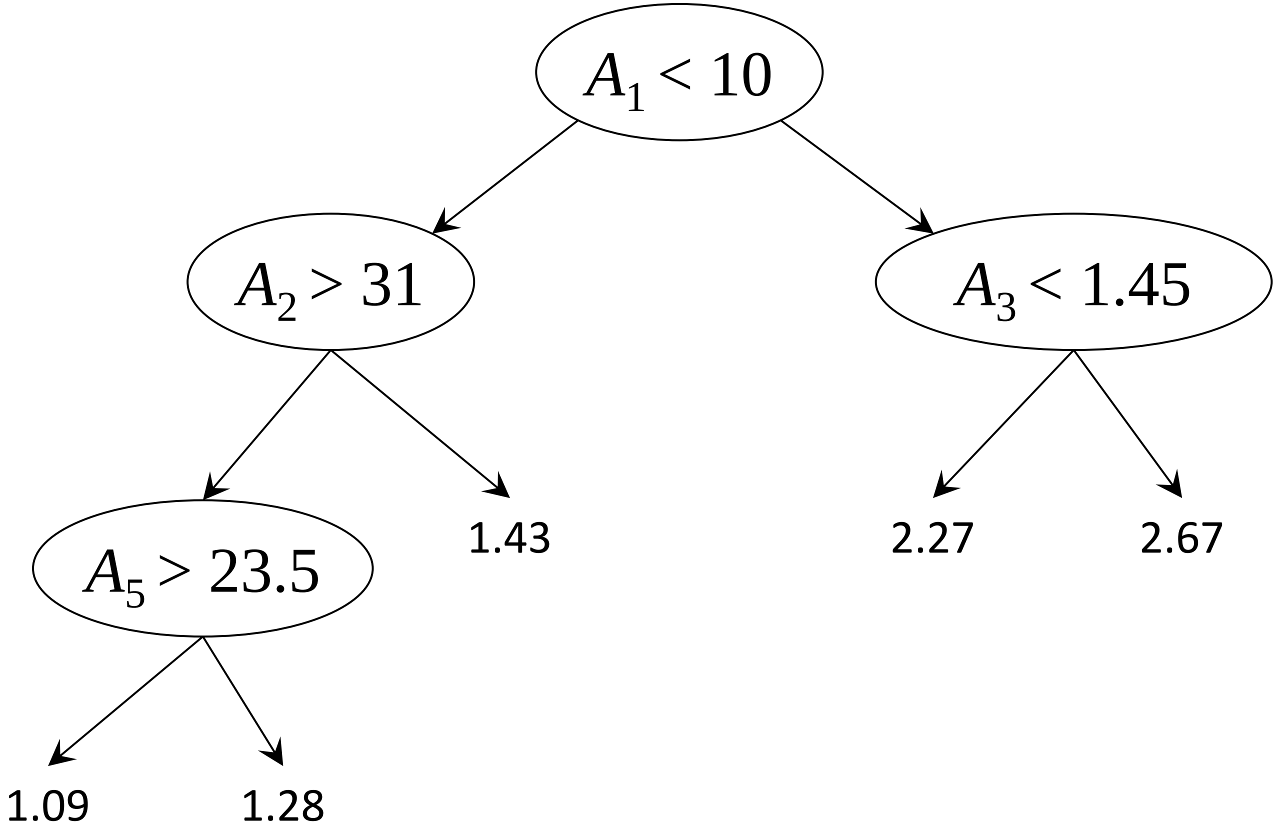
Discussions

- Continuous attributes

- Continuous output (Regression)
  - Leaves will be the regression values rather than the predicted classification values
  - How to choose a feature for an internal node?

    For example, a feature that results in the biggest reduction in regression error.

Discussions



- Missing attributes
  - Impute: fill in the missing values, e.g., using the mean of that attribute over the non-mission observations.
  - Make “missing” a category.
  - Explore the correlations between attributes.

Discussions

- Tree is optimal at each split – it may not be globally optimal.
- Building tree as gradient search.


Q
&
A