Decision Trees ~~and Random Forests~~

A decision tree is a hierarchical ('tree-like) classifier containing two kinds of nodes viz., decision nodes (internal nodes) and leaf nodes.

The internal nodes constitute tests on the features of the dataset and the leaf nodes represent the outcomes of the decisions (interms of class labels)

Defn: A decision tree is a graphical representation of all possible solutions to a decision based on certain conditions.

Process

1. The root node contains entire population (dataset)
2. Based on decisions, the root nodes is split into two or more smaller groups.
3. If uniformity (homogeneity) of a group is observed, then the group is made into a leaf node with the common class as its label. Otherwise, the splitting of the dataset continues.
4. Repeat step 4 until all groups are divided ultimately into leaf nodes.

Example:

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Task : Predict whether you can play or not given the day's attributes

?- Which attribute to pick first?
outlook / temperature / humidity / windy?

A- Determine the ~~best~~ attribute that best classifies the training data

-? how to choose the best attribute
(or)
-? how does the tree decision where to split

Different attribute selection measures

- Information Gain
- Gain Ratio
- Gini Index
- chi-square test

Information Gain

- decrease in entropy after a dataset is split based on an attribute.

So, constructing a decision tree is all about finding the attribute that returns the highest information gain.

Entropy - metric that measures impurity in a given dataset

$$\text{Entropy}(S) = -\sum_{i} P(i) \log_2 P(i)$$

case 1

yes	5
no	5

$$\text{Entropy} = -P(\text{yes}) \log_2 P(\text{yes}) - P(\text{no}) \log_2 P(\text{no})$$

$$= -0.5 \log_2(0.5) - 0.5 \log_2(0.5)$$

$$= 0.5 + 0.5 = 1$$

$$\boxed{\log_2(0.5) = \log_2\left(\tfrac{1}{2}\right) = \log_2(2^{-1}) = -1}$$

case 2

yes	10
no	0

$$\text{Entropy} = -P(\text{yes}) \log_2 P(\text{yes}) - P(\text{no}) \log_2 P(\text{no})$$

$$= -1 \log_2(1) - 0 \log_2 0$$

$$= -1 \times 0 - 0 \times 1 = 0$$

Information Gain

- measures reduction in entropy
- decides which attribute must be selected as decision node.

Information Gain = Entropy(S) - [(Weighted Avg) × Entropy(each feature)]

For the provided dataset (S)

Entropy(S) = $-P(Yes)\log_2 P(Yes) - P(No)\log_2 P(No)$

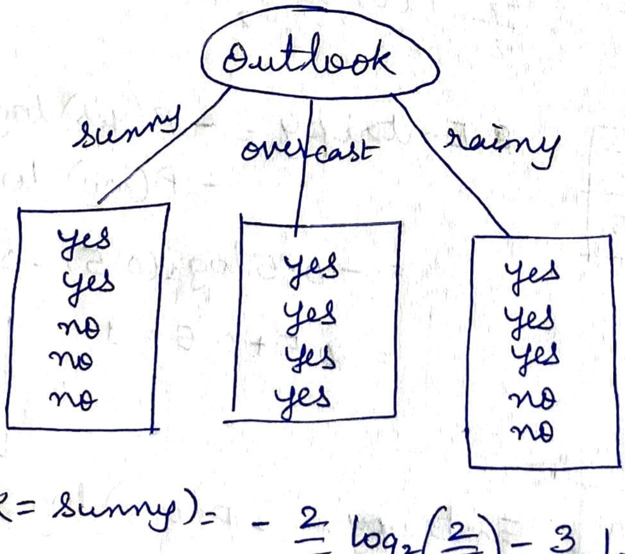
$= -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right)$

$= 0.41 + 0.53 = 0.94$

which node to select as root node

outlook / temperature / humidity / windy?

Outlook ?



$E(\text{outlook} = \text{sunny}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971$

$E(\text{outlook} = \text{overcast}) = -1\log_2 1 - 0\log_2 0 = 0$

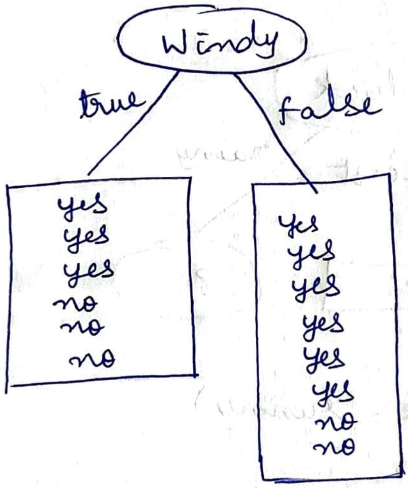
$E(\text{outlook} = \text{rainy}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.971$

Information from outlook

$$I(\text{outlook}) = \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 = 0.693$$

$$\Rightarrow \text{Gain}(\text{outlook}) = E(S) - I(\text{outlook})$$
$$= 0.94 - 0.693$$
$$= 0.247$$

Windy?



$$E(\text{Windy} = \text{true}) = 1$$

$$E(\text{Windy} = \text{false}) = 0.811$$

Information from Windy

$$I(\text{Windy}) = \underbrace{\frac{8}{14} \times 0.811}_{\text{false}} + \underbrace{\frac{6}{14} \times 1}_{\text{true}} = 0.892$$

$$\text{Gain}(\text{Windy}) = E(S) - I(\text{Windy})$$
$$= 0.94 - 0.892$$
$$= 0.048$$

Illy Temperature

~~Info~~ $I(\text{Temperature}) = 0.911$

$$\text{Gain}(\text{Temperature}) = E(S) - I(\text{Temperature})$$
$$= 0.94 - 0.911$$

Humidity

$I(\text{Humidity}) = 0.788$

$\text{Gain}(\text{Humidity}) = \cancel{0.788 - 0}\ \ 0.94 - 0.788$

$= 0.152$

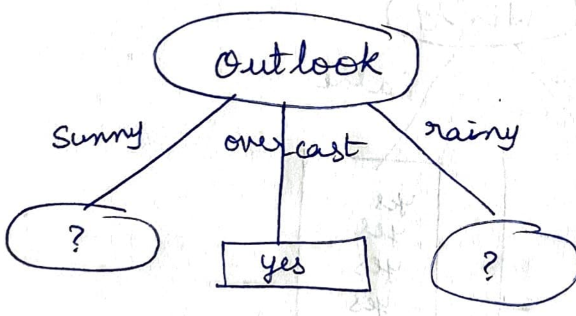
Information Gain

outlook → 0.247 ✓

Temperature → 0.029

Humidity → 0.152

Windy → 0.048



data at

Left subtree (Outlook = sunny)

Temperature	Humidity	Windy	Play
hot	high	false	no
hot	high	true	no
mild	high	false	no
cool	normal	false	yes
mild	normal	true	yes

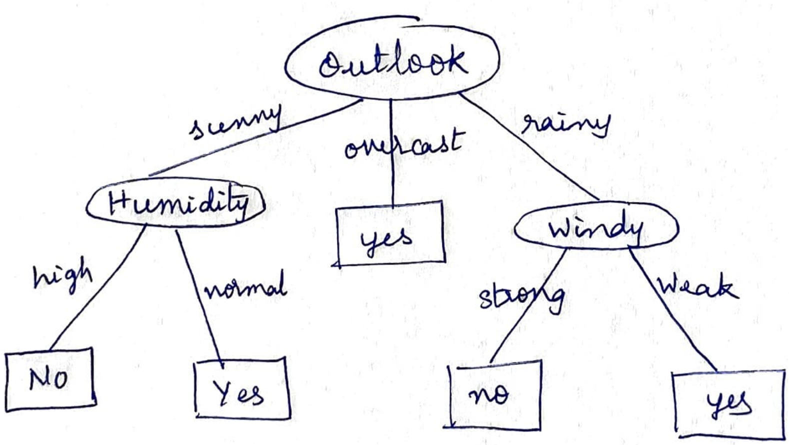
repeat the process

Data at right subtree
(outlook = rainy)

Temperature	Humidity	Windy	Play
mild	high	false	yes
cool	normal	false	yes
cool	normal	true	no
mild	normal	false	yes
mild	high	true	no

repeat the process

<u>complete decision tree</u>



<u>Advantages</u>
- simple to build
- easy to understand the solution (interpretable)

<u>Disadvantages</u>
- susceptible to overfitting
  - soln? pruning

Pruning

- reducing the complexity
- improving the generality

Pruning
- Pre pruning (decide whether or not to split a particular node during model building)
- Post pruning (build the decision tree, then prune the branches to avoid overfitting)