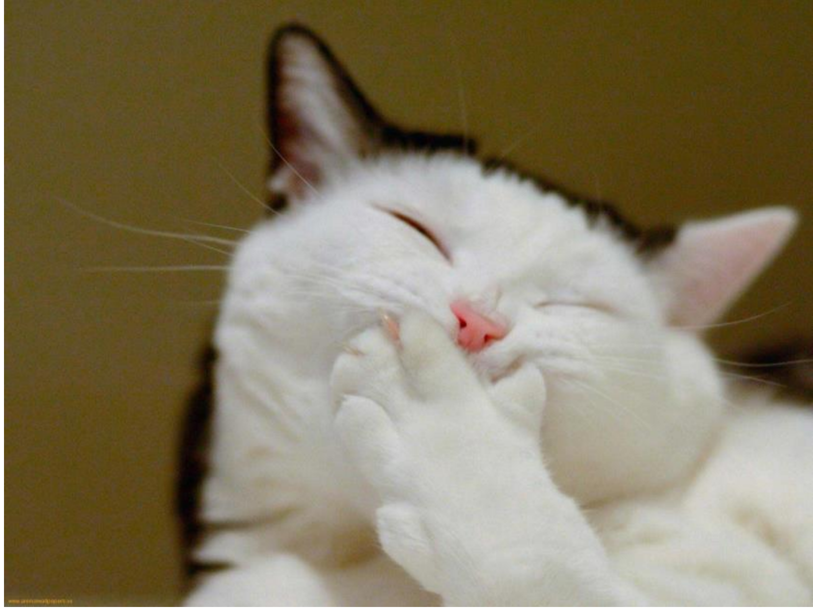
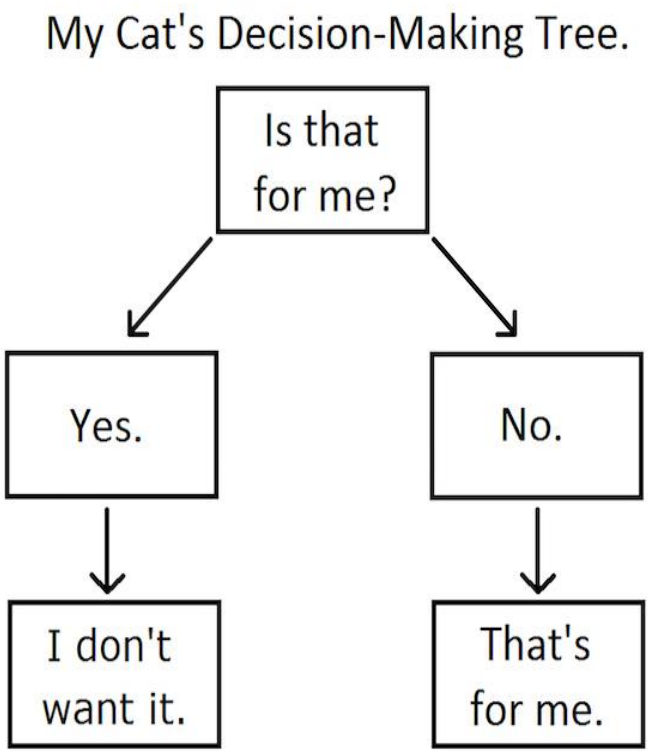
Decision Trees

- Bhavesh Bhatt





Decision Tree

- A decision tree is a supervised machine learning algorithm used for predicting outcomes based on certain rules and is done by partitioning the data into subsets.
- The partitioning process starts with a binary split and continues until no further splits can be made.
- Various branches of variable length are formed.

Decision Trees

1. Concept learning: an example

Given the data:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

predict the value of PlayTennis for

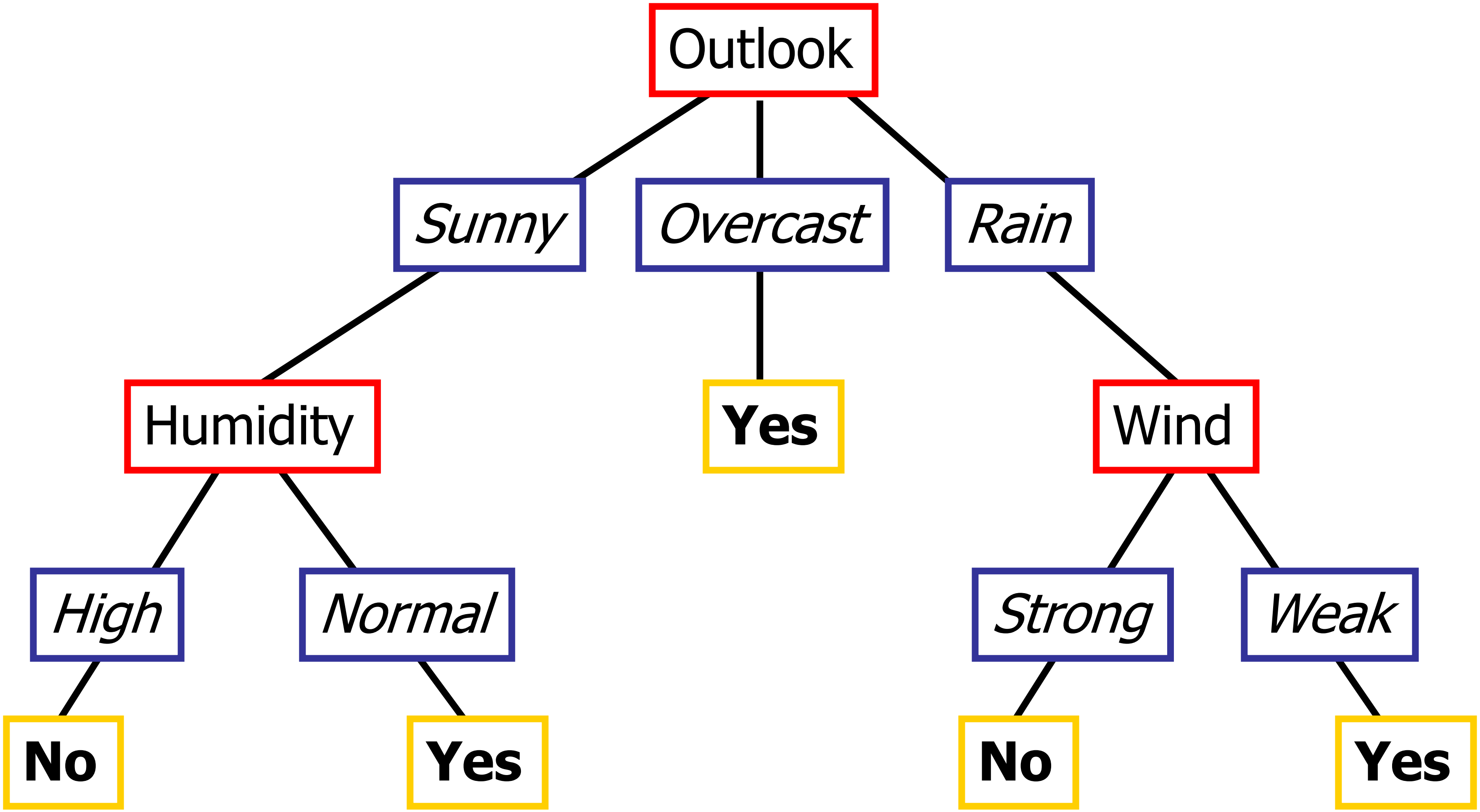
$$\langle Outlook = sunny,\ Temp = cool,\ Humidity = high,\ Wind = strong \rangle$$

Decision Tree for PlayTennis

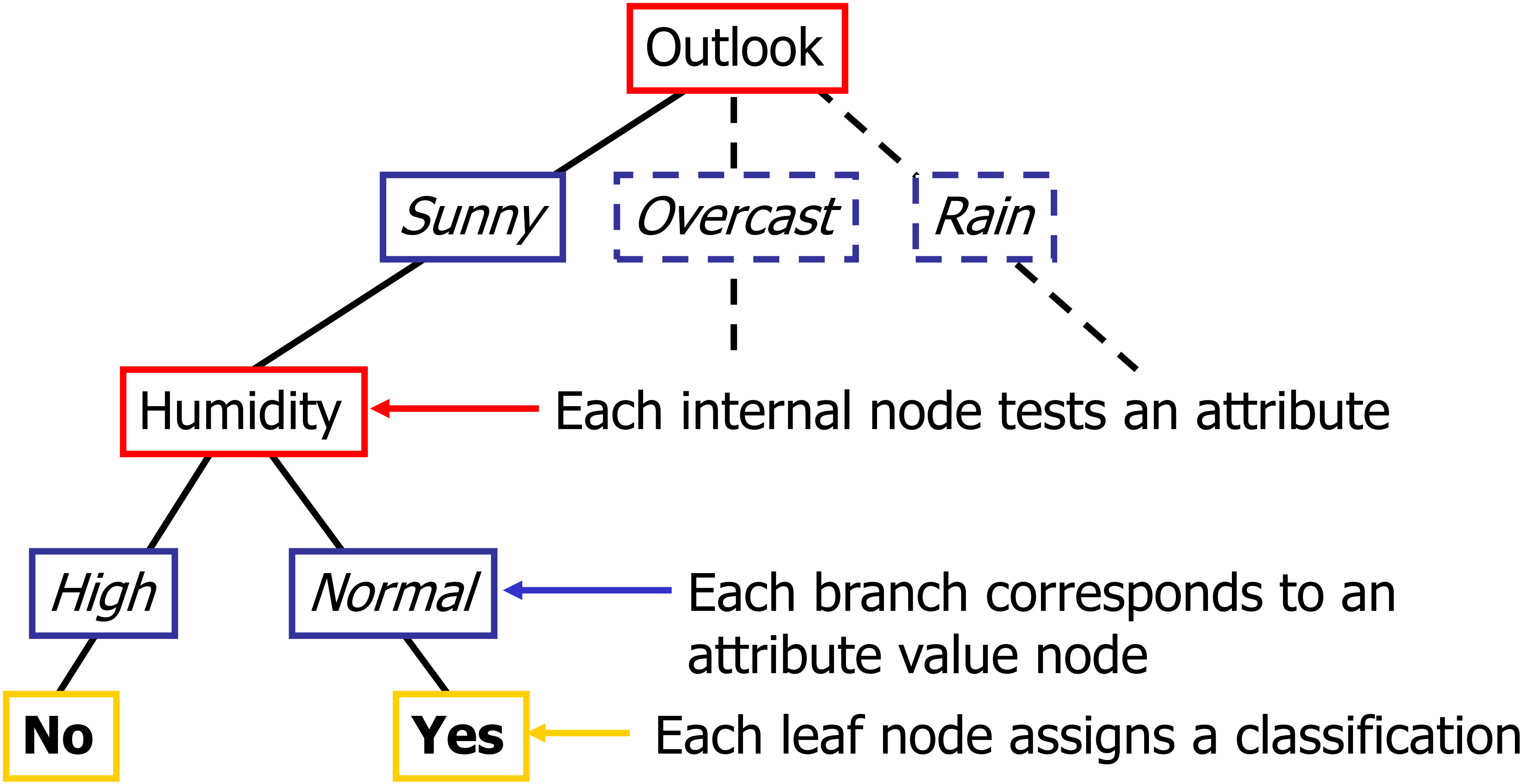
- Attributes and their values:
  - Outlook: *Sunny, Overcast, Rain*
  - Humidity: *High, Normal*
  - Wind: *Strong, Weak*
  - Temperature: *Hot, Mild, Cool*

  - Target concept - Play Tennis: *Yes, No*

Decision Tree for PlayTennis

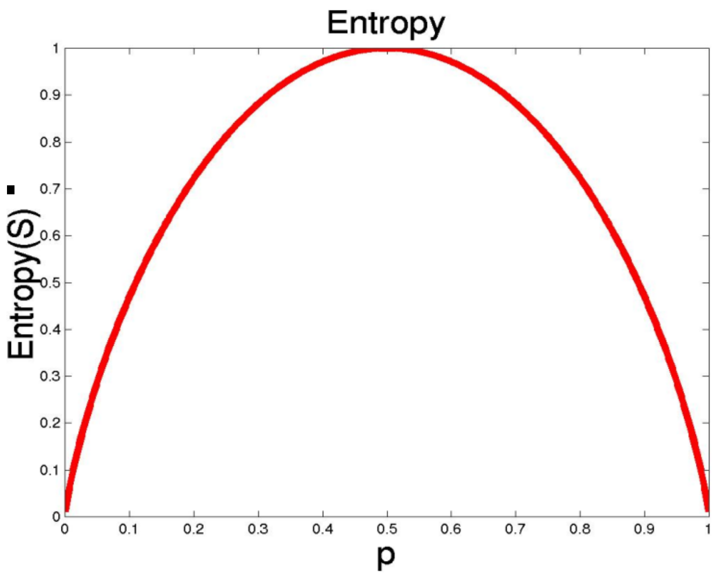


Decision Tree for PlayTennis



Entropy

- Entropy is the measure of the homogeneity of a sample in a node.
- If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.



- S is a sample of training examples
- p_+ is the proportion of positive examples
- p_- is the proportion of negative examples
- Entropy measures the impurity of S

$$\text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Information Gain

- Gain(S,A): expected reduction in entropy due to sorting S on attribute A

- The information gain is based on the decrease in entropy after a dataset is split on an attribute.

- Information Gain = entropy (parent) –
  [Weighted Average] entropy (children)

$$Gain(S,A)=Entropy(S) - \sum_{v \in values(A)} |S_v|/|S| \; Entropy(S_v)$$

Information Gain

Calculate entropy of the target

$\text{Entropy(PlayTennis)} = -p_+ \log_2 p_+ - p_- \log_2 p_-$

➔ (-0.36)log2(0.36) – (0.64)log2(0.64) ➔ 0.94

Choosing an attribute to split on

- Idea: a good attribute should reduce uncertainty and result in “gain in information”

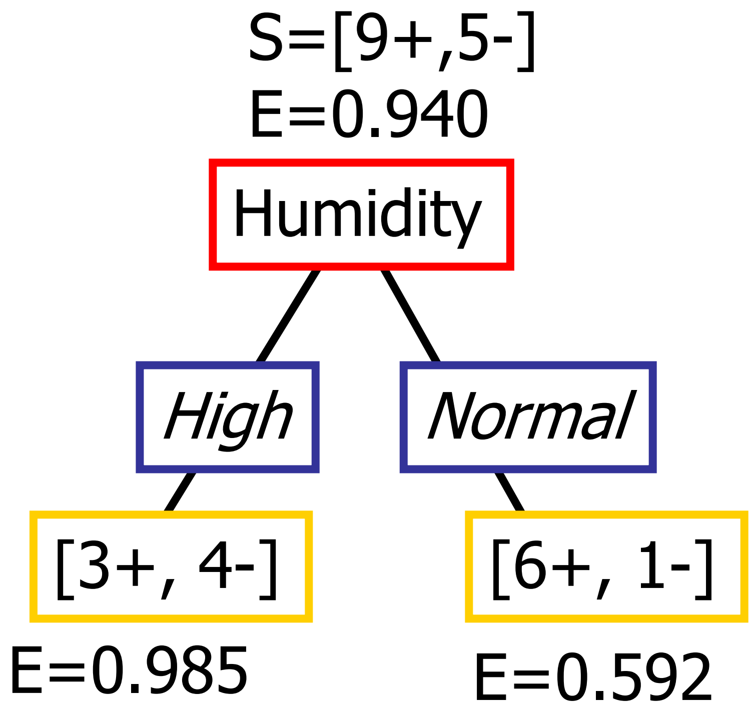
- How much information do we gain if we disclose the value of some attribute?

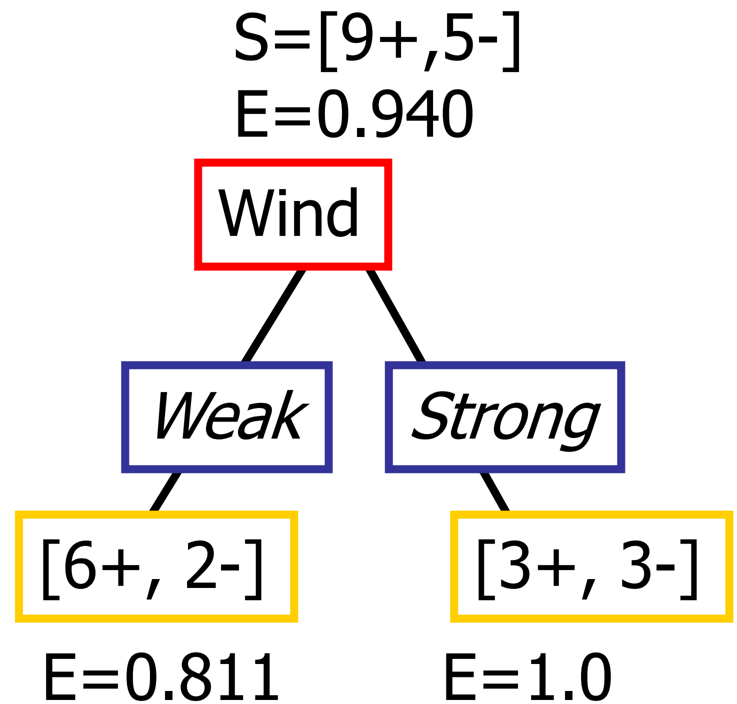
- Answer:
  - Uncertainty before – Uncertainty after

Training Examples

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Selecting the Next Attribute



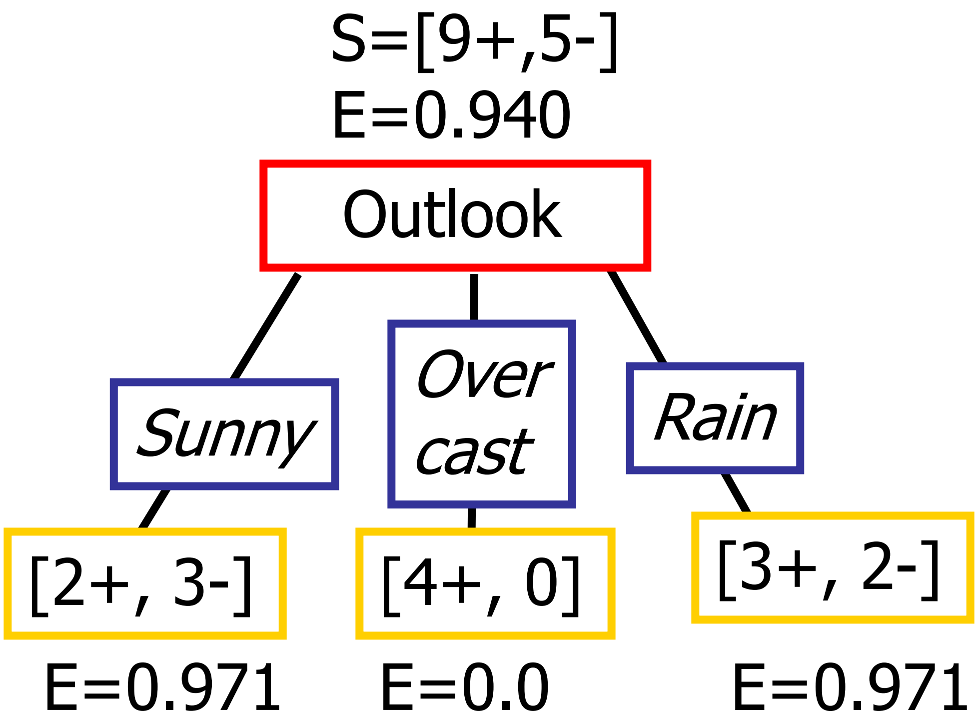


Gain(S,Humidity)
=0.940-(7/14)*0.985
– (7/14)*0.592
=0.151

Gain(S,Wind)
=0.940-(8/14)*0.811
– (6/14)*1.0
=0.048

Humidity provides greater info. gain than Wind, w.r.t target classification.

Selecting the Next Attribute



Gain(S,Outlook)
=0.940-(5/14)*0.971
-(4/14)*0.0 – (5/14)*0.0971
=0.247

Selecting the Next Attribute

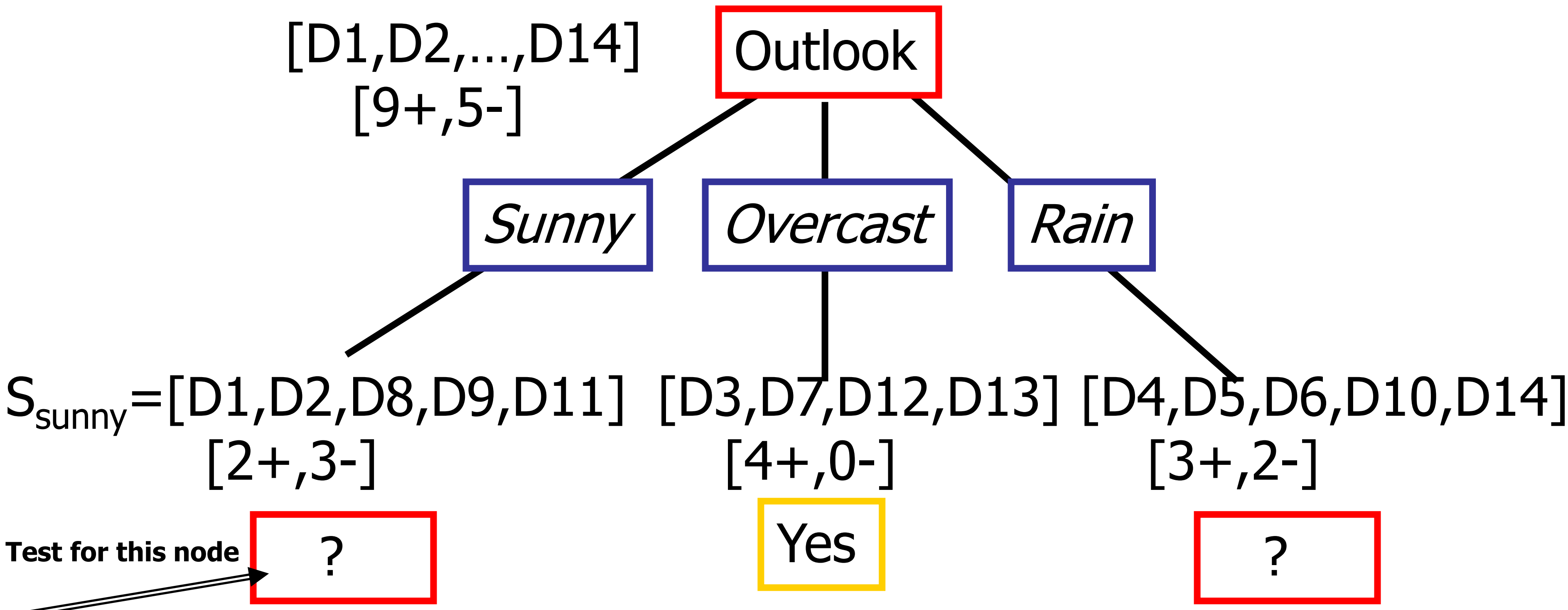
The information gain values for the 4 attributes are:

- Gain(S,Outlook) =0.247
- Gain(S,Humidity) =0.151
- Gain(S,Wind) =0.048
- Gain(S,Temperature) =0.029

where S denotes the collection of training examples

ID3 Algorithm

Note: $0\text{Log}_2 0 = 0$

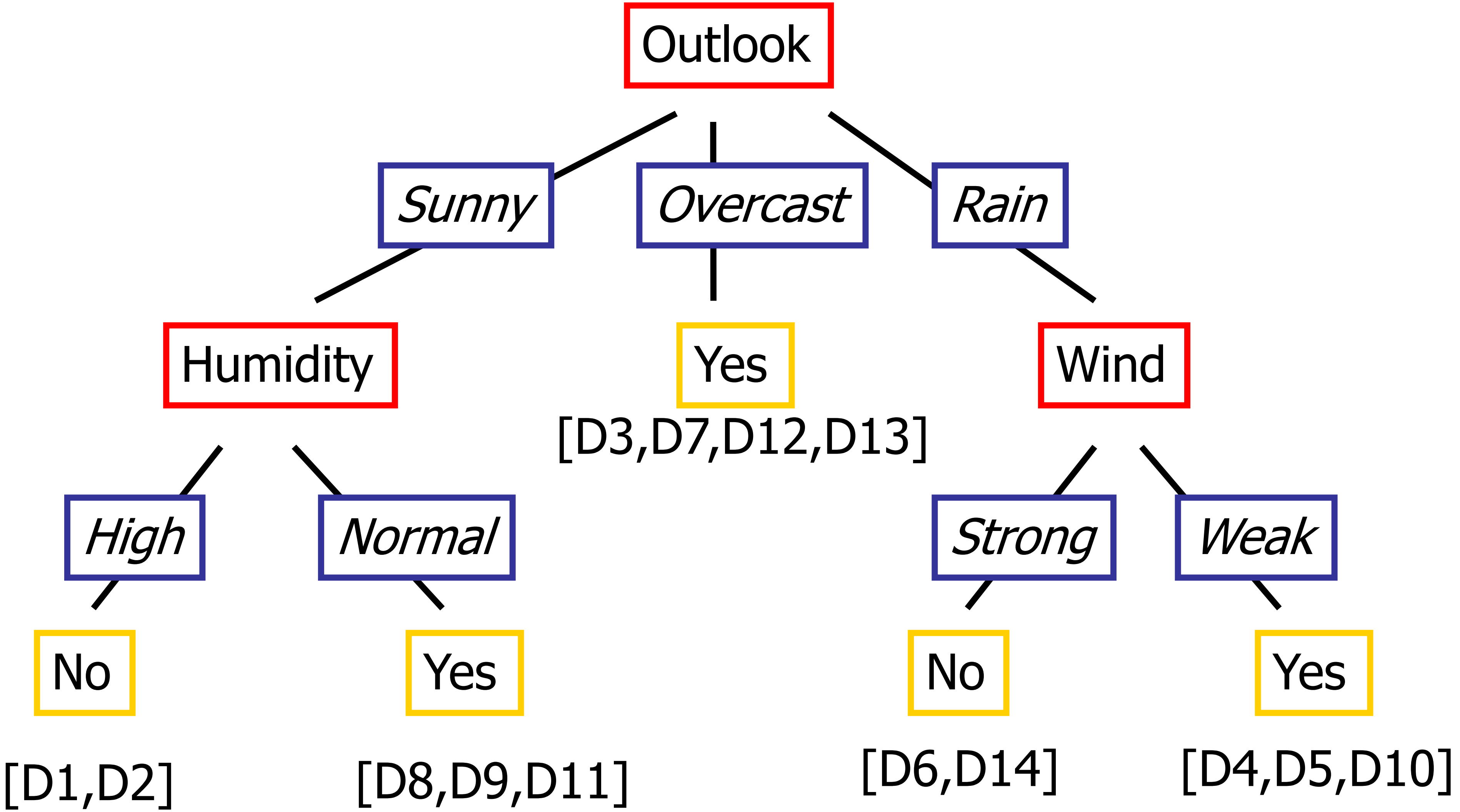


$\text{Gain}(S_{sunny}, \text{Humidity}) = 0.970 - (3/5)0.0 - 2/5(0.0) = 0.970$

$\text{Gain}(S_{sunny}, \text{Temp.}) = 0.970 - (2/5)0.0 - 2/5(1.0) - (1/5)0.0 = 0.570$

$\text{Gain}(S_{sunny}, \text{Wind}) = 0.970 = -(2/5)1.0 - 3/5(0.918) = 0.019$

ID3 Algorithm



Gini Index (CART)

- A Gini score gives an idea of how good a split is by how mixed the classes are in the two groups created by the split.
- A perfect separation results in a Gini score of 0, whereas the worst case split that results in 50/50 classes.

Gini Index

- If a data set D contains examples from n classes, gini index, $gini(D)$ is defined as:

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

where p_j is the relative frequency of class j in D

- If a data set D is split on A into two subsets D_1 and D_2, the $gini$ index $gini(D)$ is defined as

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

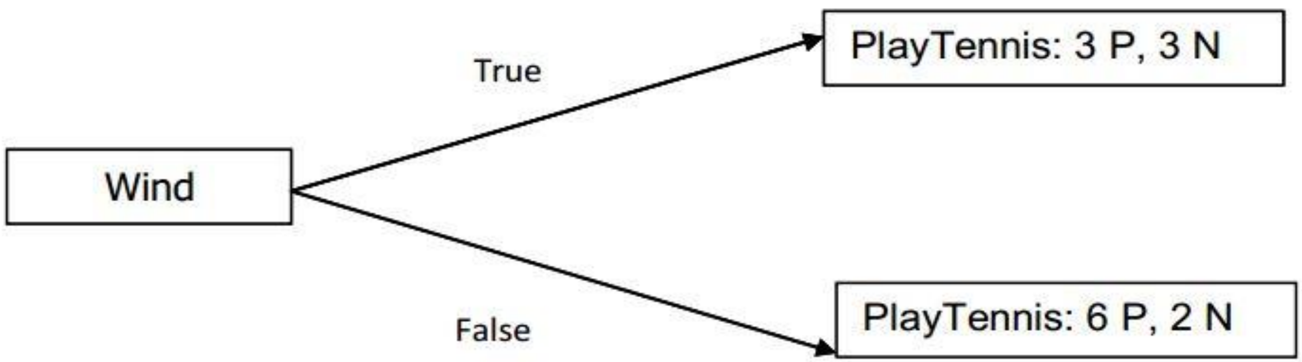
- Reduction in Impurity: $\Delta gini(A) = gini(D) - gini_A(D)$

Gini Index I

Gini index calculation:

There are 5 Ns and 9 Ps, so the

- Calculate the information gain after the Wind test is applied:



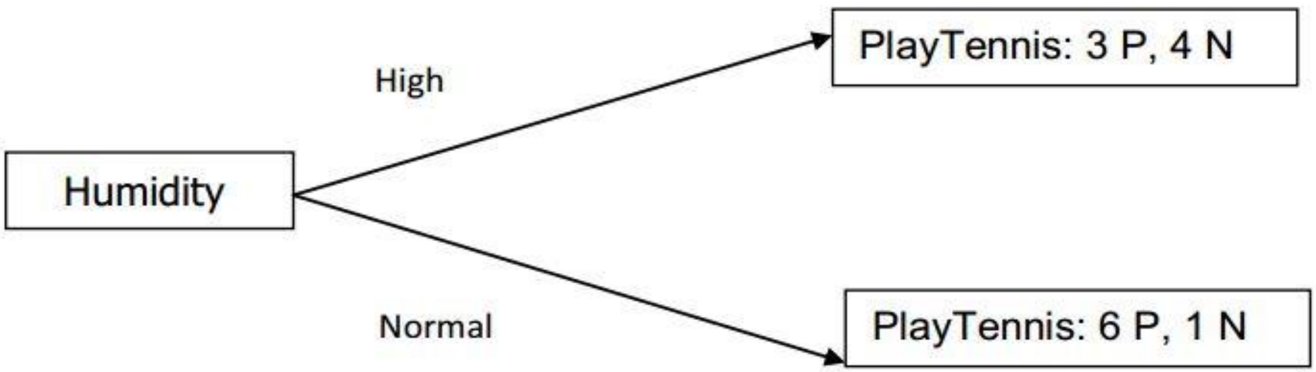
Gini (PlayTennis|Wind=True) = $1- (3/6)^2 – (3/6)^2 = 0.5$
Gini (PlayTennis|Wind=False) = $1- (6/8)^2 – (2/8)^2 = 0.375$

Therefore, the Gini index after the Wind test is applied is

$6/14 \times 0.5 + 8/14 \times 0.375 = 0.4286$

Gini Index II

- Calculate the information gain after the Humidity test is applied:



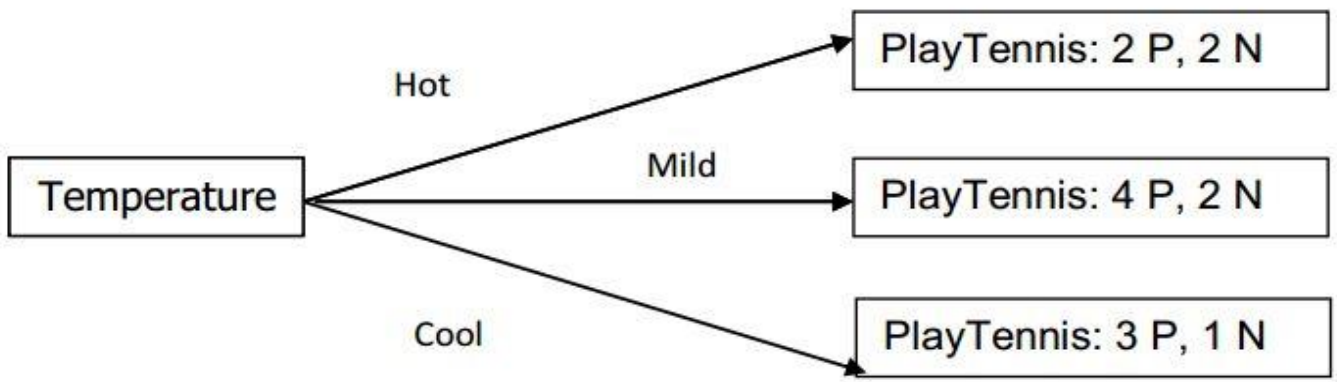
Gini (PlayTennis|Humidity=High) = $1- (3/7)^2 - (4/7)^2 = 0.4898$
Gini (PlayTennis|Humidity=Normal) = $1- (6/7)^2 - (1/7)^2 = 0.2449$

Therefore, the Gini index after the Wind test is applied is

$7/14 \times 0.4898 + 7/14 \times 0.2449 = 0.3674$

Gini Index III

- Calculate the information gain after the Temperature test is applied:



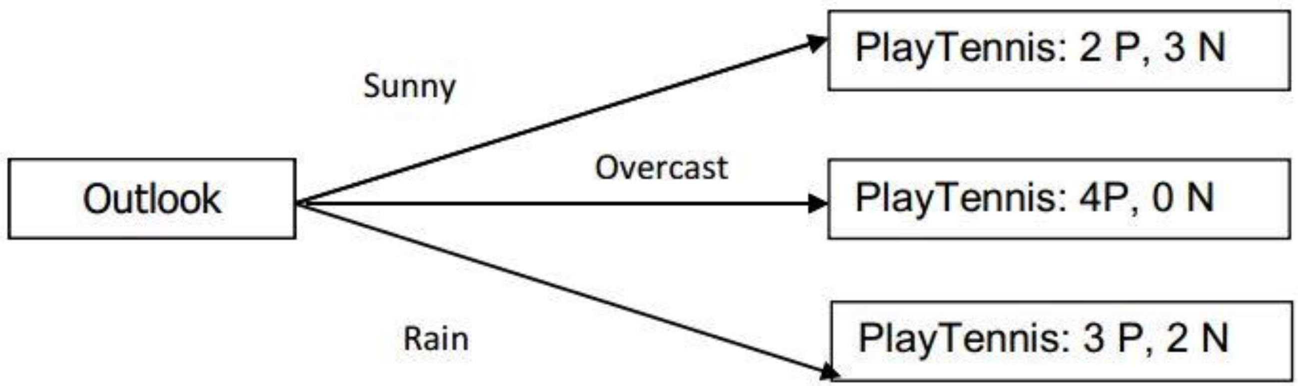
Gini (PlayTennis| Temperature =Hot) = $1- (2/4)^2 - (2/4)^2 = 0.5$
Gini (PlayTennis| Temperature =Mild) = $1- (4/6)^2 - (2/6)^2 = 0.4444$
Gini (PlayTennis| Temperature =Cool) = $1- (3/4)^2 - (1/4)^2 = 0.375$

Therefore, the Gini index after the Temperature test is applied is

$4/14 \times 0.5 + 6/14 \times 0.4444 + 4/14 \times 0.375 = 0.4405$

Gini Index IV

- Calculate the information gain after the Outlook test is applied:



Gini (PlayTennis| Outlook =Sunny) = 1- $(2/5)^2 - (3/5)^2 = 0.48$
Gini (PlayTennis| Outlook =Overcast) = 1- $(4/4)^2 - (0/4)^2 = 0$
Gini (PlayTennis| Outlook =Rain) = 1- $(3/5)^2 - (2/5)^2 = 0.48$

Therefore, the Gini index after the Temperature test is applied is

$5/14 \times 0.48 + 4/14 \times 0 + 5/14 \times 0.48 = 0.3429$

Gini Index V

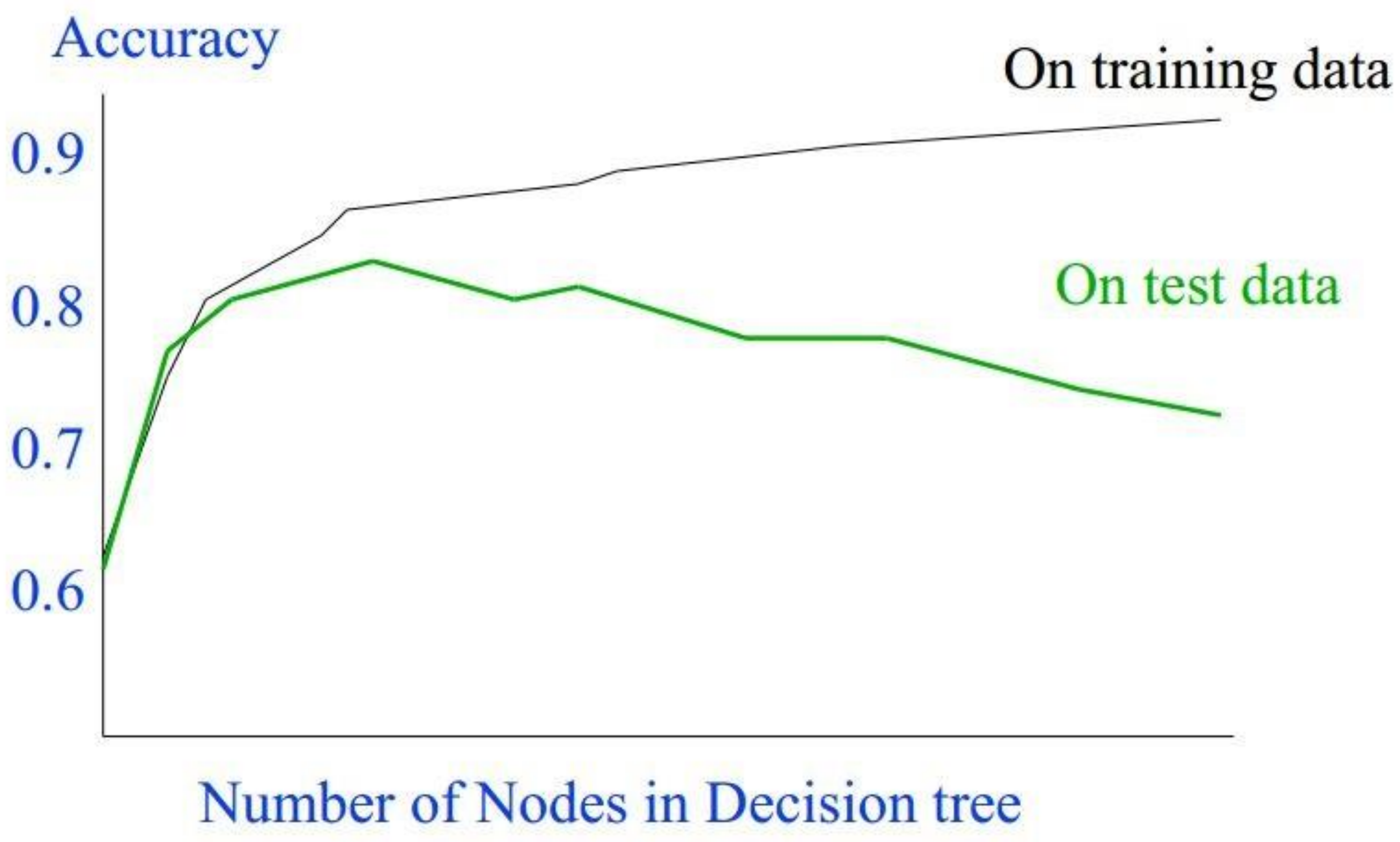
After calculating all attributes:

- gain(outlook) = 0.3429
- gain(temperature) = 0.4405
- gain(humidity) = 0.3674
- gain(windy) = 0.4286

Overfitting

- One of the biggest problems with decision trees is Overfitting
- Overfitting is a modelling error which occurs when a function is too closely fit to a limited set of training data points
- The training error is less but the testing error is really high

Overfitting



Avoid Overfitting

How can we avoid overfitting?

- Prepruning
  - Limit Tree Depth
  - Minimum node size
  - Neglegible change in classification error

- Postpruning
  - Remove sections of the tree which provide little or no prediction power

Decision Trees - Strengths

- Very Popular Technique
- Fast
- Useful when Target Function is discrete

Decision Trees - Weakness

- Less useful for continuous outputs
- Can have difficulty with continuous input features as well.


Thank you!