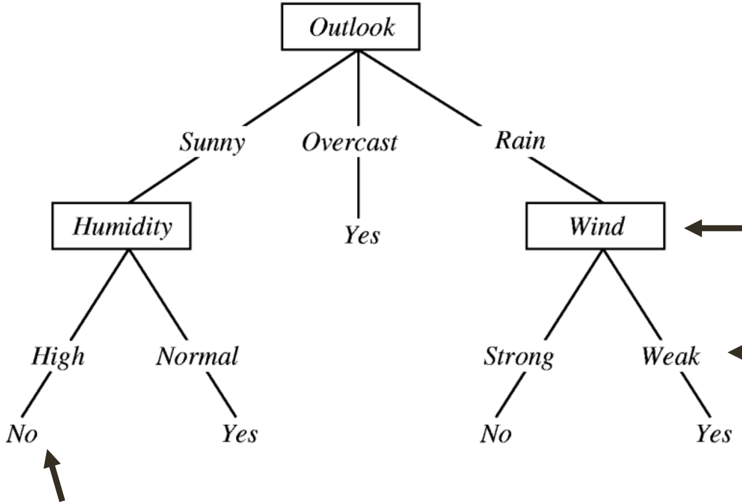
Decision Trees

Abdessalam Bouchekif
abdessalam.bouchekif@epita.fr

Decision Tree representation

- **Decision Trees** are supervised learning method used for classification and regression.

- Learning simple decision rules inferred from the data features.



Node test of some attribute

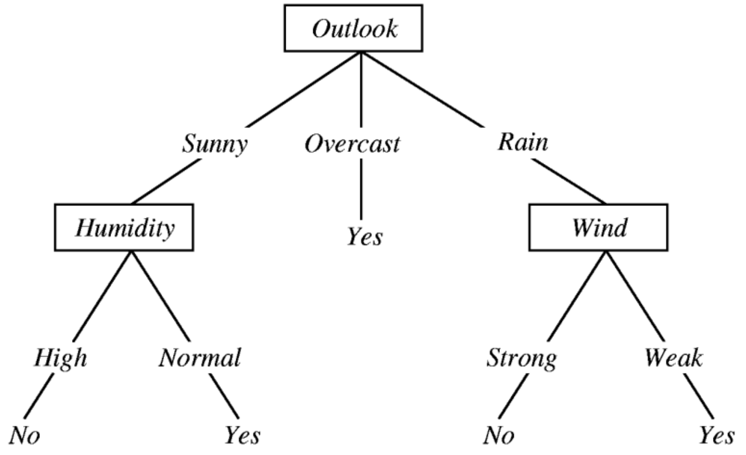
Branch : one of the possible values for the attribute

each branch represents one if-then-rule

leaf node has a class label

Decision Tree as Set of Rules

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



This decision is equivalent to:

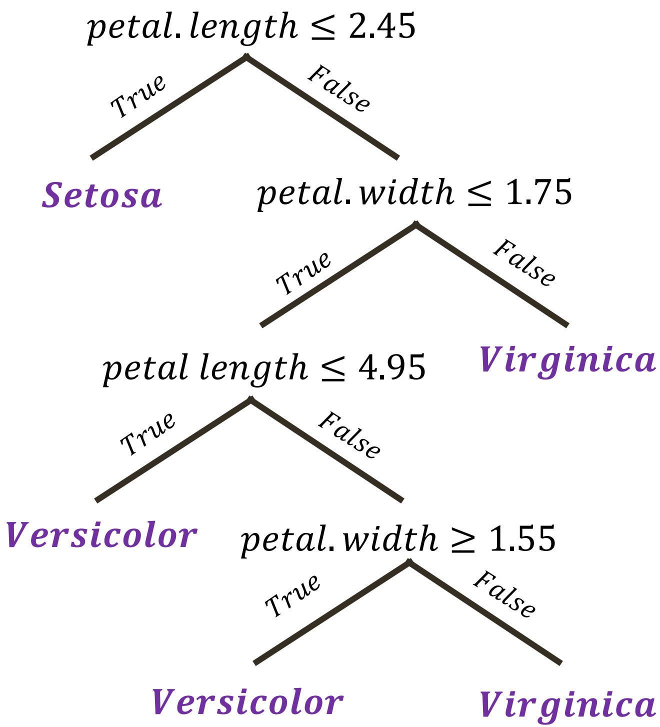
if (Outlook == "Sunny")∧(Humidity=="*Normal*")
then Yes

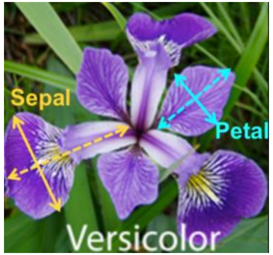
if (Outlook == "Overcast") **then** Yes

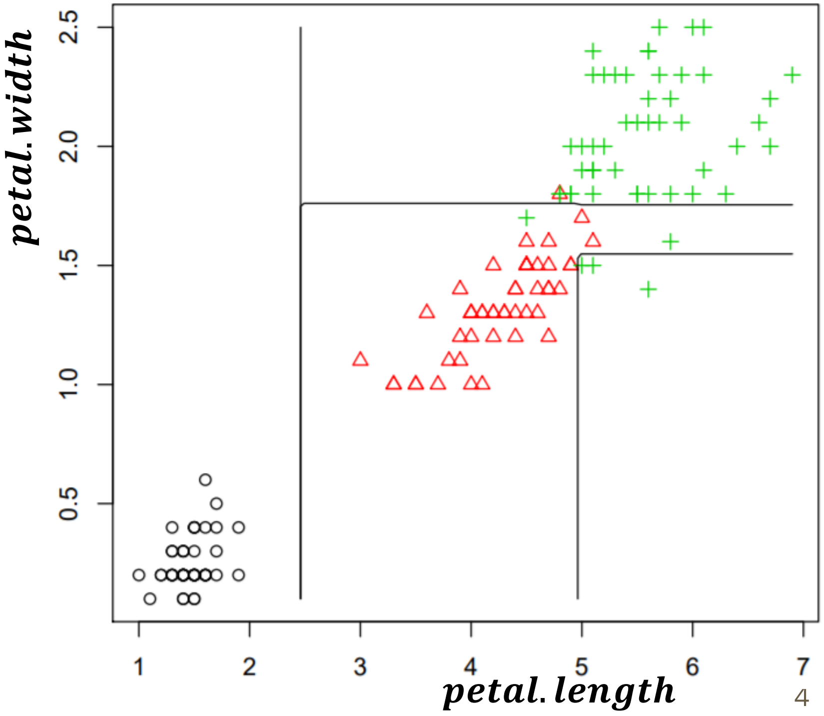
if (Outlook=="Rain") ∧ (*Wind* == Weak)
then Yes

… ….

Decision boundaries







Why interesting?

- What we can do:
  - Given a set training examples
  - Find the general classification rules

- The rules can used to classify future examples

- Which is useful in many situations:
  - Medical diagnosis
  - Credit application scoring: grant a loan or not?
  - Fraud detection: is the transaction suspicious or not?
  - Identify groups of similar credit card users
  - Modeling calendar scheduling preferences
  - …

Decision trees

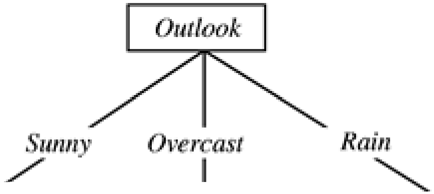
- Algorithms used
  - ID3
  - C4.5
  - CART

- Basic idea of $ID3$ algorithm: A decision tree can be constructed by considering attributes of instances one by one
  - The height of decision tree depends on the order attributes that are considered
  - Which attribute should be considered first?

How to build decison trees (ID3 algorithm)?

- Suppose first attribute (root) chosen is "Outlook"

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



Outlook = Sunny

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

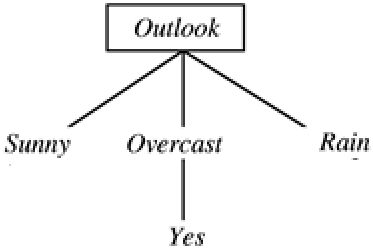
Outlook = Overcast

D3	Overcast	Hot	High	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

Outlook = Rain

D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D10	Rain	Mild	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

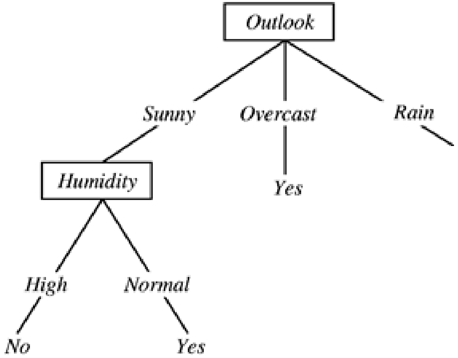
How to build decison trees ?



- For the node "$outlook = Overcast$", all example are labeled "yes"

$\Longrightarrow$ hence it becomes a leaf node with classification "$PlayTennis = yes$"

- For the node "$Outlook = sunny$" need to select another attribute
  - Suppose "$Humidity$" is chosen.
  - Get left-lower part of tree.
  - Split data



"$Humidity = High$"

Day	Outlook	T	Humidity	W	P
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No

All are labeled "No" becomes leaf.

"$Humidity = Normal$"

D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

All are labeled "Yes", becomes leaf

How to build decison trees ?

- For the node "$Outlook = rain$" need to select another attribute.
  - Suppose "$Wind$" is chosen. Get right-lower part of tree. Split data:

"$Wind = Strong$"

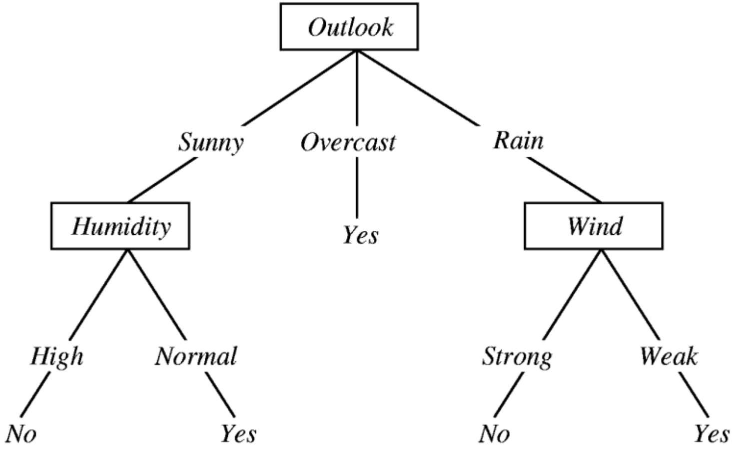
Day	Outlook	T	H	Wind	P
D6	Rain	Cool	Normal	Strong	No
D14	Rain	Mild	High	Strong	No

All are labeled "No" becomes leaf.

"$wind = Weak$"

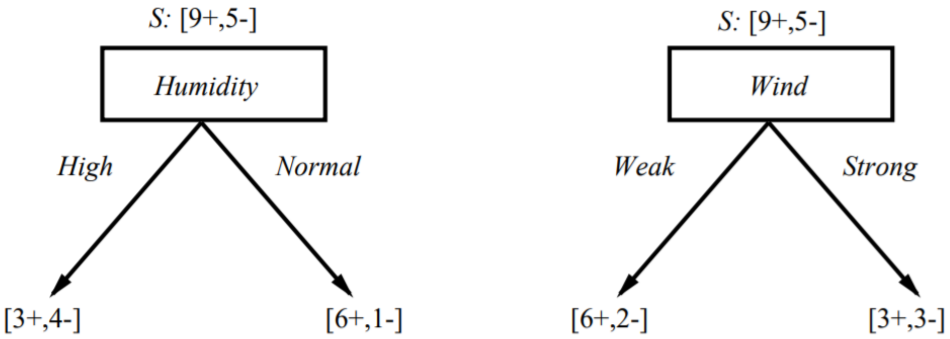
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes

All are labeled "Yes", becomes leaf



End of tree construction

Which attribute is best?



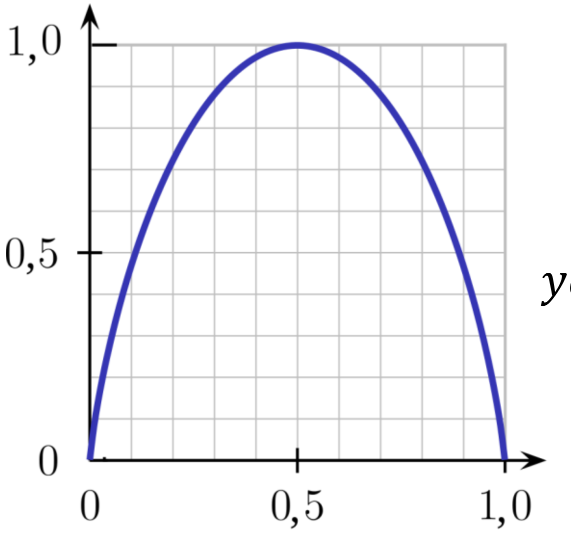
- Intuitively, we want a test attribute that **separates** the training set as well as possible
- Need a measure of node impurity
- $ID3$ uses the entropy and information gain

Entropy

- Given probabilities $p_1, p_2, \dots, p_c$ whose *sum* is 1, Entropy is defined as:

$$E(p_1, p_2, \dots, p_c) = \sum_{i=1}^{c} -p_i log_2\, p_i$$

- *All samples belong to the same class* $\Rightarrow \boldsymbol{E = 0}$
- *Samples are equally mixed* for binary classification $\Rightarrow \boldsymbol{E = 1}$
- *Samples are equally mixed* for multiclass classification $\Rightarrow \boldsymbol{E = log_2\, c}$



PlayTennis

$yes = 9$ $\quad$ $no = 5$

$$E(play\ tennis) = -\left(\frac{9}{14}\right) log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) log_2\left(\frac{5}{14}\right) = 0.96$$

Information gain

- We want to determine which attribute is most useful for discriminating between the classes to be learned

$\Longrightarrow$ Select the attribute with the highest information gain

- $ID3$ chooses to split on an attribute that gives the highest information gain:

$$Gain(S,A) = Entropie(S) - \sum_{v \in Valeurs(A)}^{s} \frac{|S_v|}{|S|} Entropie(S_v)$$

Attribute Selection: An Example

S: $[8+, 8-]$

A_1splits S into S_{11}: $[8+, 0-]$ and S_{12}: $[0+, 8-]$

A_2 splits S into S_{21}: $[4+, 4-]$ and S_{12}: $[4+, 4-]$

$$Entropy(S) = -0.5log_2(0.5) - 0.5log_2(0.5) = 1$$

$$Gain(S, A_1) = -Entropy(S) - 0.5\ Entropy([8+, 0\ -]) - 0.5\ Entropy([0+, 8\ -])$$
$$= 1 - 0 - 0 = 0$$

$$Gain(S, A_2) = -Entropy(S) - 0.5\ Entropy([4+, 4\ -]) - 0{,}5\ Entropy([4+, 4\ -])$$

$$= 1 - 0.5 - 0.5 = 1$$

ID3 algorithm

$\boldsymbol{Input}$: Example set S

$\boldsymbol{Output}$: Decesion Tree DT

- $\boldsymbol{if}$ all examples in S belong to the same class c

  return a new leaf and label it label it with c

- $\boldsymbol{else}$ Select the best atribute A

  Generate a new note DT with A as test

  for each value v_i of A

  - Let S_i= all examples in S with $A = v_i$
  - Use $ID3$ to construct a decision tree DT_i for example set S_i

Decison Tree Example

- Entropy of S

$$S = \{D_1, \dots, D_{14}\} = [9+, 5-]$$

$$E(S) = \frac{9}{14} log_2\left(\frac{9}{14}\right) - \frac{5}{14} log_2\left(\frac{5}{14}\right) = 0.94$$

- Information gain $(Outlook)$

$$S_{sunny} = [2+, 3-]; E(S_{sunny}) = 0.971$$
$$S_{overcast} = [4+, 0-]; E(S_{overcast}) = 0.0$$
$$S_{rainy} = [3+, 2-]; E(S_{rainy}) = 0.971$$

$$Gain(S, Outlook) = 0.94 - \frac{5}{14}0.971 - \frac{4}{14}0.0 - \frac{5}{14}0.971 = \mathbf{0.246}$$

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	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D6	Rainy	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rainy	Mild	High	Strong	No

Decison Tree Example

- Information gain ($Humidity$)

$$S_{high} = [3+, 4-]; \; E(S_{high}) = 0.985$$

$$S_{normal} = [6+, 1-]; E(S_{normal}) = 0.592$$

$$Gain\,(S, Humidity) = 0.693 - \frac{7}{14}0.985 - \frac{7}{14}0.592 = \mathbf{0.151}$$

- Information gain ($Wind$)

$$S_{weak} = [6+, 2-]; \; E(S_{weak}) = 0.811$$

$$S_{strong} = [3+, 3-]; \; E(S_{strong}) = 1.0$$

$$Gain\,(S, Strong) = 0.940 - \frac{8}{14}0.811 + \frac{6}{14}1.0 = \mathbf{0.048}$$

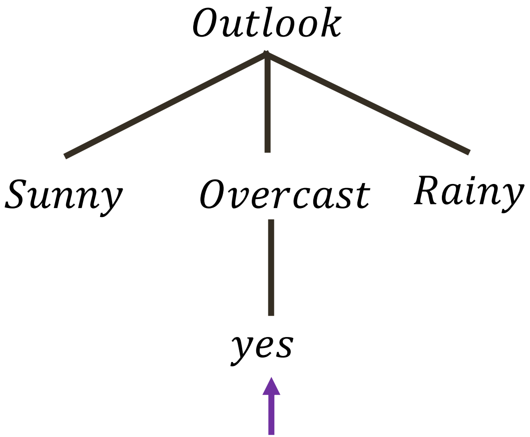
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D6	Rainy	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny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

Decison Tree Example

- Information gain (*Temperature*)

$$Gain\ (S, temperature) = 0.940 - \left(\frac{4}{14}\right)\overbrace{1}^{E([2+,2-])} - \left(\frac{6}{14}\right)\underbrace{0.918}_{E([4+,2-])} - \left(\frac{4}{14}\right)\overbrace{0.811}^{E([3+,1-])} = 0.029$$

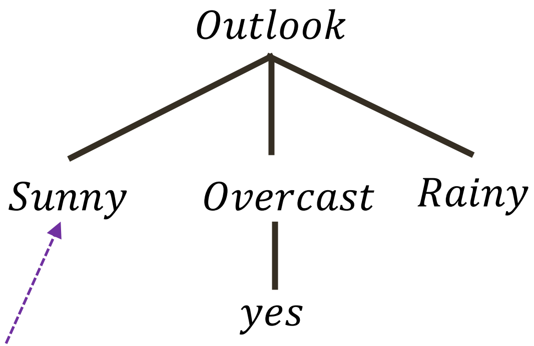
- So start tree construction with *Outlook*



A branch with entropy of 0 is a leaf node.

Decison Tree Example

- A branch with entropy more than 0 needs further splitting.



Which attribute should be tested here?

$$S_{sunny} = [2+, 3-]$$

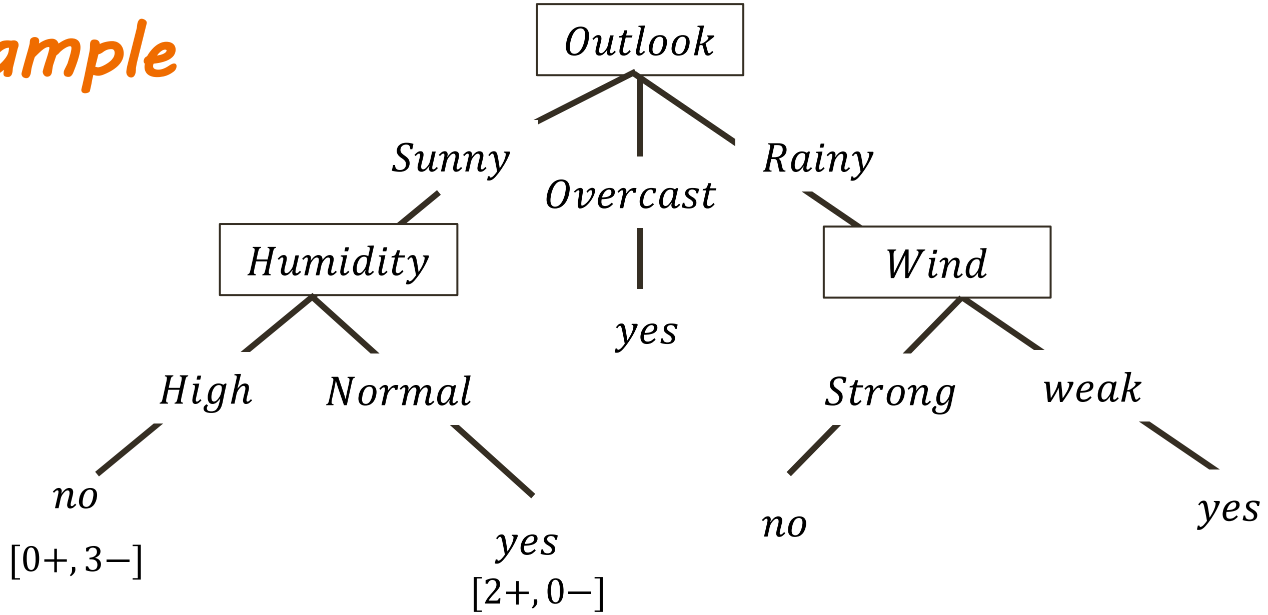
$$Gain(S_{sunny}, Humidity) = 0.97 - \left(\frac{3}{5}\right)0.0 - \left(\frac{2}{5}\right)0.0 = 0.97$$

$$Gain(S_{sunny}, Temperature) = 0.970 - \left(\frac{2}{5}\right)0.0 - \left(\frac{2}{5}\right)1.0 - \left(\frac{1}{5}\right)0.0 = 0.57$$

$$Gain(S_{sunny}, Humidity) = 0.970 - \left(\frac{2}{5}\right)1.0 - \left(\frac{3}{5}\right)0.918 = 0.019$$

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	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D6	Rainy	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rainy	Mild	High	Strong	No

Decison Tree Example



$E([3+, 2-])$

$$Gain(S_{Rainy}, humidity) = 0.970 - \left(\frac{2}{5}\right)1 - \left(\frac{3}{5}\right)0.918 = 0.019$$

$$Gain(S_{Rainy}, temperature\ ) = 0.970 - \left(\frac{2}{5}\right)1 - \left(\frac{3}{5}\right)0.918 = 0.019$$

$$Gain(S_{Rainy}, wind) = 0.970 - \left(\frac{2}{5}\right)0 - \left(\frac{3}{5}\right)0 = 0.970$$