Illustrative Example for ID3 Induction

An Illustrative Example (1)

The dependent variable „Tennis“ determines if the weather is good for tennis („Yes“) or not („No“).

Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

An Illustrative Example (2): Entropy of the Decision Tree

$$\text{Entropy}(S) = -\,9/14 * \log_2(9/14) - 5/14 * \log_2(5/14) = 0{,}94$$

Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

positive frequency (Yes)
negative frequency (No)

Selection of the topmost Node

- In order to determine the attribute that should be tested first in the tree, the information gain for attributes (*Outlook* , *Temperature* ,*Humidity* and *Wind*) are determined.
  - Gain(S,Outlook) = 0.246
  - Gain(S,Humidity) = 0.151
  - Gain(S,Wind) = 0.048
  - Gain(S,Temperature) = 0.029
- Since *Outlook* attribute provides the best prediction, it is selected as the decision attribute for the root node.

An Illustrative Example (3): Selection of the topmost Node

Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

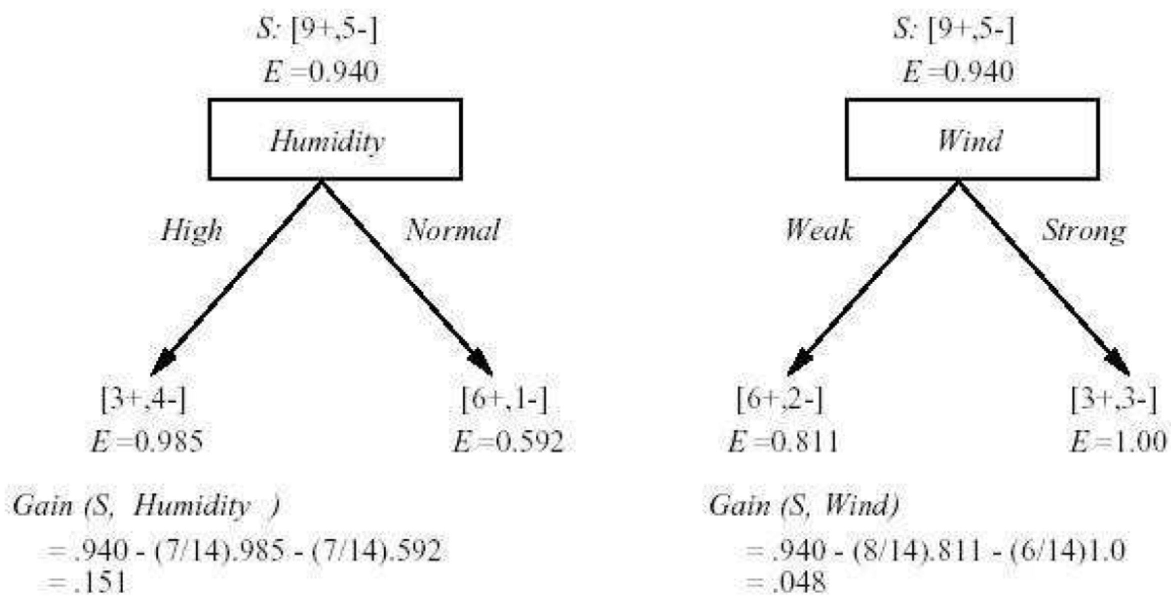
- In order to determine the attribute that should be tested first in the tree, the information gain for each attribute (*Outlook* , *Temperature, Humidity* and *Wind*) is determined.
  - Gain(S,Outlook) = **0.246**
  - Gain(S,Humidity) = **0.151**
  - Gain(S,Wind) = **0.048**
  - Gain(S,Temperature) = **0.029**
- Since *Outlook* attribute provides the best prediction, it is selected as the decision attribute for the root node.

An Illustrative Example (4): Computation of Information Gain

- The computation of Information Gain for Outlook:

$$\begin{aligned} GAIN(S,Outlook) &= Entropy(S) - EE(Outlook) \\ &= 0.94 - 0.694 = \mathbf{0.246} \end{aligned}$$

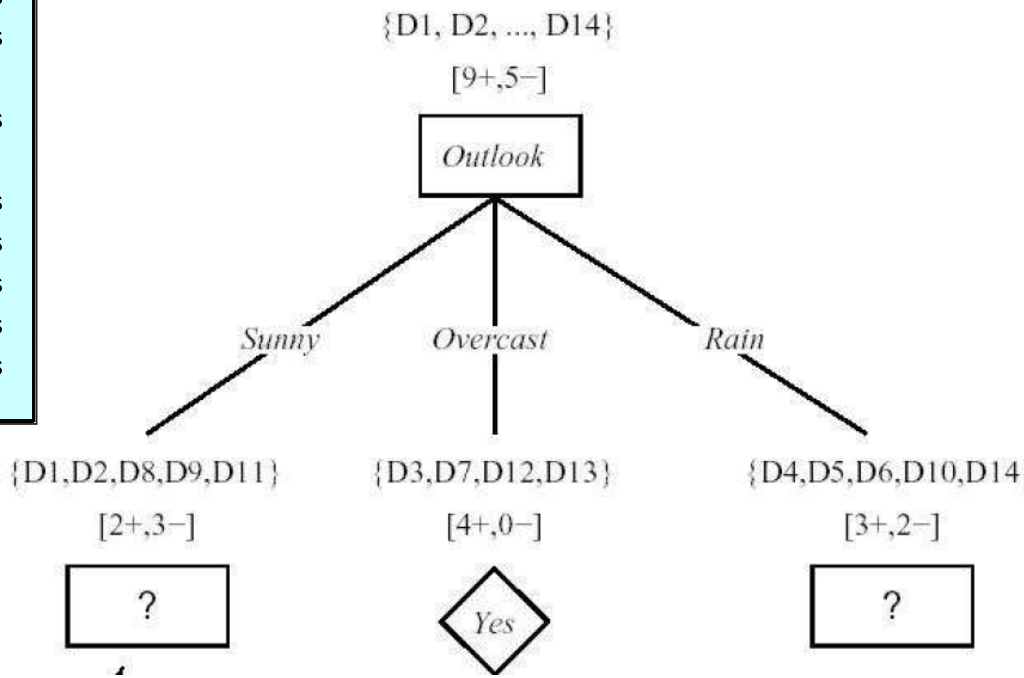
- The computation of information gain for *Humidity* and *Wind:*



An Illustrative Example (5): Resulting Subtree

Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

- The partially learned decision tree resulting from the first step of ID3:

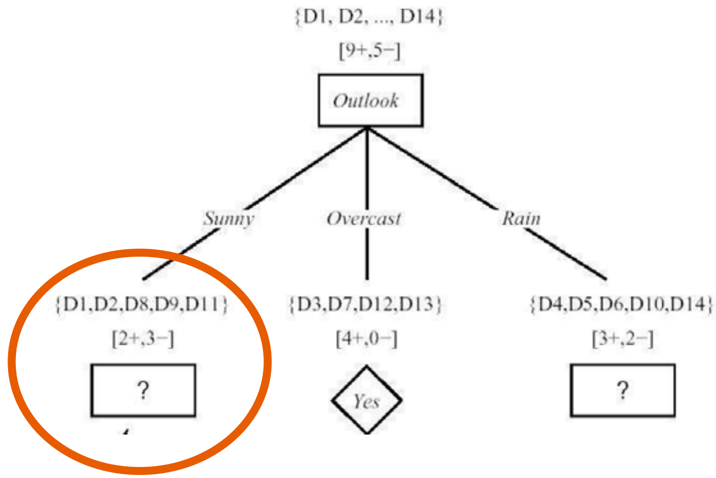


An Illustrative Example (6): Entropie of a Subtree

The subtree with root Sunny:

$$\text{Entropy(Sunny)} = -\,2/5 \log_2 (2/5) - 3/5 \log_2 (3/5)$$
$$= 0{,}970$$

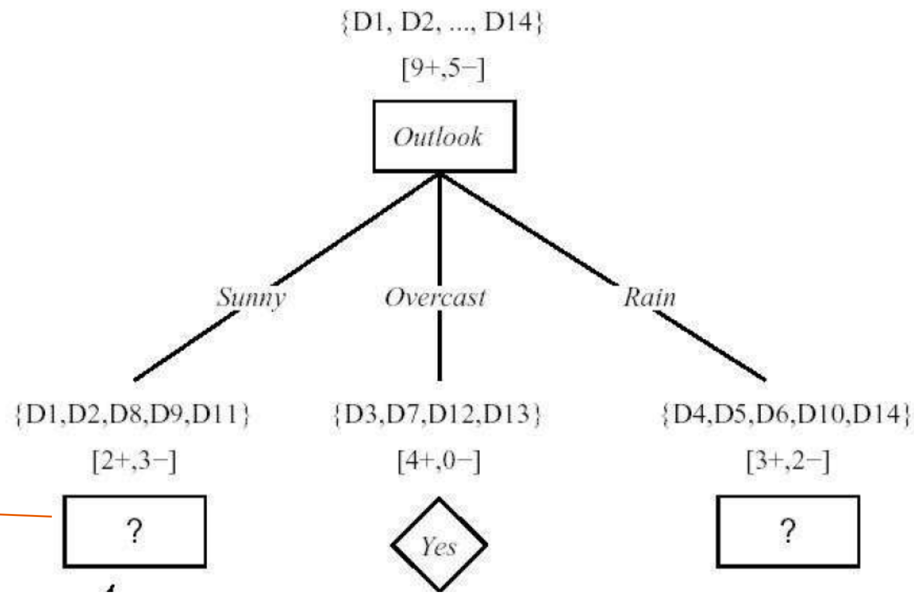
Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



The more **up** in the decision tree, the smaller the entropy of the subtree

An Illustrative Example (7): Selectiong Next Attribute

Which attribute should be tested here?



$$S_{sunny} = \{D1,D2,D8,D9,D11\}$$

$$Gain\ (S_{sunny}, Humidity) = .970 - (3/5)\ 0.0 - (2/5)\ 0.0 = .970$$

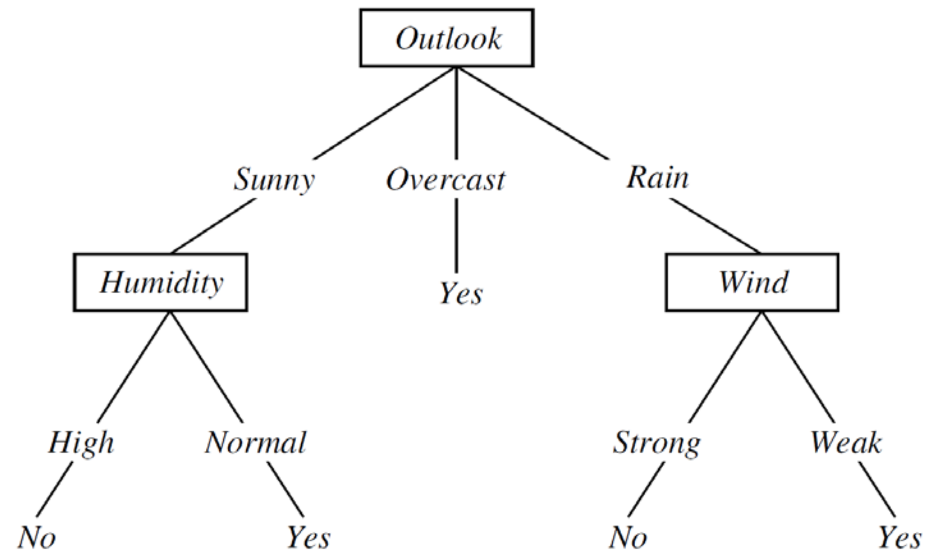
$$Gain\ (S_{sunny}, Temperature) = .970 - (2/5)\ 0.0 - (2/5)\ 1.0 - (1/5)\ 0.0 = .570$$

$$Gain\ (S_{sunny}, Wind) = .970 - (2/5)\ 1.0 - (3/5)\ .918 = .019$$

An Illustrative Example (8): The Resulting Decision Tree

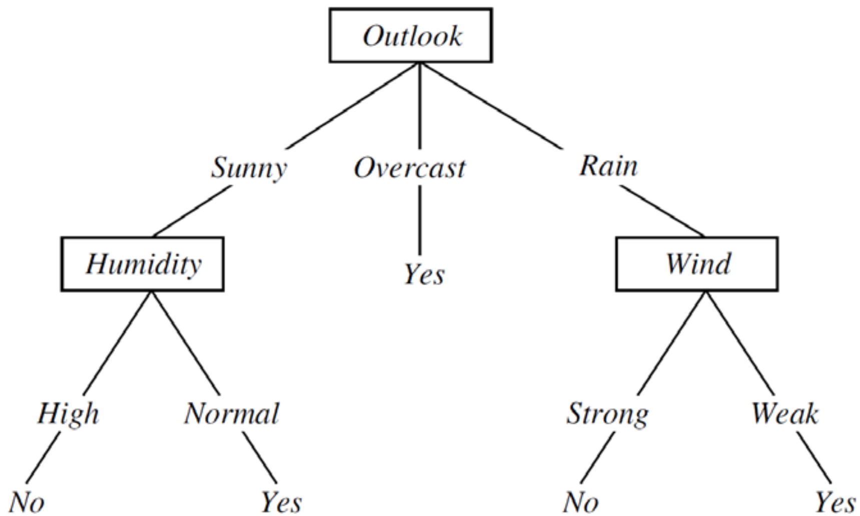
The dependent variable „Tennis“ determines if the weather is good for tennis („Yes“) or not („No“).

Element	Outlook	Temperature	Humidity	Wind	Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



The result of the induction algorithms classifies the data with only three of the four attributes into the classes „Yes“ and „No“.

An Illustrative Example (9): Decision Tree represented as Decision Table



Playing Tennis

	Outlook	Humidity	Wind	Tennis
	Sunny, Overcast, Rain	*High, Normal*	*Strong, Weak*	*Yes, No*
1	Sunny	High		No
2	Sunny	Normal		Yes
3	Overcast			Yes
4	Rain		Strong	No
5	Rain		Weak	Yes