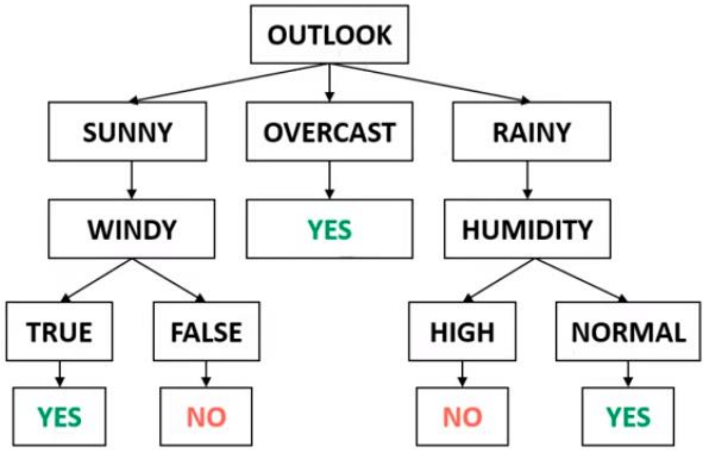
Decision Tree

Decision Tree

Decision Trees

outlook	temperature	humidity	wind	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cold	normal	false	yes
rainy	cold	normal	true	no
overcast	cold	normal	true	yes
sunny	mild	high	false	no
sunny	cold	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



Entropy

Decision Trees

outlook	temperature	humidity	wind	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cold	normal	false	yes
rainy	cold	normal	true	no
overcast	cold	normal	true	yes
sunny	mild	high	false	no
sunny	cold	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

PLAYING GOLF
→ **9** times **YES**
→ **5** times **NO**

We just have to use the Shannon-entropy formula to calculate the **H(x)** values

H(PlayingGolf) = H(9,5) =

$$= -(\underbrace{0.64}_{\frac{9}{14}} \log_2 \underbrace{0.64}_{\frac{9}{14}}) - (\underbrace{0.36}_{\frac{5}{14}} \log_2 \underbrace{0.36}_{\frac{5}{14}}) = 0.94$$

Decision Trees

outlook	temperature	humidity	wind	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cold	normal	false	yes
rainy	cold	normal	true	no
overcast	cold	normal	true	yes
sunny	mild	high	false	no
sunny	cold	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

$$E(T,X) = \sum_{x} P(x)\, E(x)$$

We have to calculate the entropy with respect to a given predictor/feature in order to be able to calculate information gain

		PLAY GOLF	
		YES	**NO**
OUTLOOK	sunny	**2**	**3**
	overcast	**4**	**0**
	rainy	**3**	**2**

E(PlayGolf,Outlook) = P(sunny)E(2,3) + P(overcast)E(4,0) + P(rainy)E(3,2)

$$\frac{5}{14}\,0.971 + \frac{4}{14}\,0 + \frac{5}{14}\,0.971 = 0.6936$$

Information Gain

Decision Trees

outlook	temperature	humidity	wind	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cold	normal	false	yes
rainy	cold	normal	true	no
overcast	cold	normal	true	yes
sunny	mild	high	false	no
sunny	cold	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

Information gain: the decrease in entropy after a dataset is split on an attribute/feature

→ feature/attribute with the highest information gain will be the root node in the tree

Information Gain = H(PlayGolf) - E(PlayGolf,Outlook) =
= 0.94 – 0.693 = 0.247

		PLAY GOLF YES	PLAY GOLF NO
OUTLOOK	sunny	**2**	**3**
	overcast	**4**	**0**
	rainy	**3**	**2**

Decision Trees

outlook	temperature	humidity	wind	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cold	normal	false	yes
rainy	cold	normal	true	no
overcast	cold	normal	true	yes
sunny	mild	high	false	no
sunny	cold	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

Information gain: the decrease in entropy after a dataset is split on an attribute/feature

→ feature/attribute with the highest information gain will be the root node in the tree

Information Gain (outlook) = 0.247

Information Gain (temperature) = 0.029

Information Gain (humidity) = 0.152

Information Gain (wind) = 0.048

Decision Trees

Usually **ID3** algorithm is used to build the decision tree:
~ it is a top-down greedy search of possible branches

→ it uses **entropy** and **information gain** to build the tree

The **H(X)** Shannon-entropy of a dicrete random variable **X** with possible values $x_1 x_2 \ldots x_n$ and probability mass function **P(X)** is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i)$$

Example: https://en.wikipedia.org/wiki/Entropy_(information_theory)

For completely homogoeneous dataset (all TRUE or all FALSE values): entropy is **0**
If the dataset is equally divided (same amount of TRUEs and FALSEs): entropy is **1**

A BRANCH WITH ENTROPY MORE THAN 1 NEEDS SPLITING !!!
+ root node has the maximum information gain (entropy reduction)
+ leaf nodes have entropy 0

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

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

Training Examples

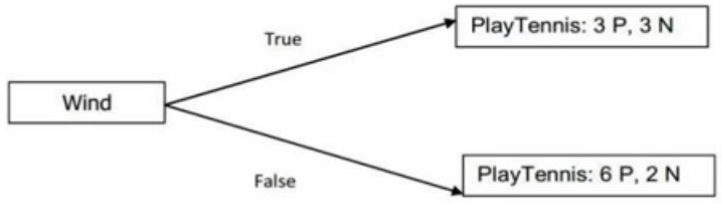
Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes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

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

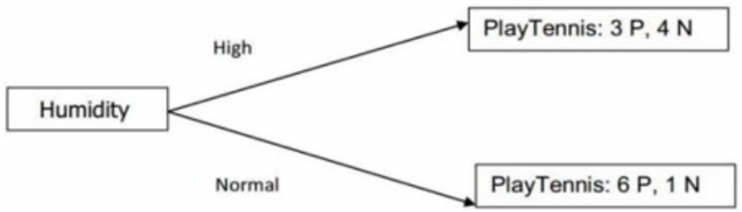
$6/14 \times 0.5 + 8/14 \times 0.375 = 0.4286$

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

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

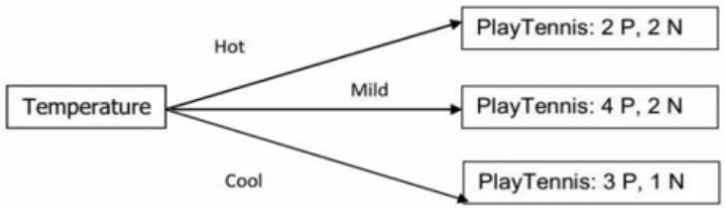
$7/14 \times 0.4898 + 7/14 \times 0.2449 = 0.3674$

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

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

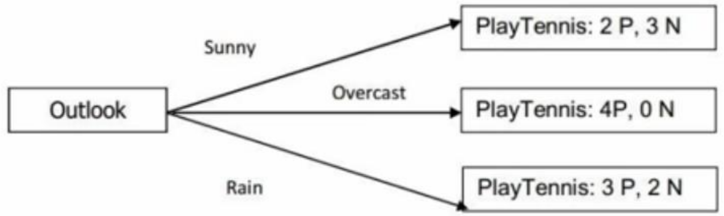
$4/14 \times 0.5 + 6/14 \times 0.4444 + 4/14 \times 0.375 = 0.4405$

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

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

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

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

Gini Index V

After calculating all attributes:

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