Introduction to Decision Tree

Wednesday, 17 April 2024 3:48 PM

Decision Tree

A Decision Tree takes as input an object or situation described by a set of properties, and output a yes/no "decision".

Entropy: Entropy is measure of the degree of randomness or uncertainty in the dataset, It measures the randomness based on the distribution of class labels in the dataset

$$\boxed{\text{Entropy} = \frac{-m}{m+n}\log_2\left(\frac{m}{m+n}\right) + \frac{-n}{m+n}\log_2\left(\frac{n}{m+n}\right)}$$

$m \Rightarrow$ no. of Red balls
$n \Rightarrow$ no. of Blue balls

Entropy for Bucket 1: [4R 0B] $\frac{-4}{4+0}\log_2\left(\frac{4}{4+0}\right) + \frac{-0}{0+4}\log_2\left(\frac{0}{4+0}\right) = 0+0 = 0$

Entropy for Bucket 2: [3R 1B] $\frac{-3}{3+1}\log_2\left(\frac{3}{3+1}\right) + \frac{-1}{3+1}\log_2\left(\frac{1}{3+1}\right) = 0.81125$

Entropy for Bucket 3: [2R 2B] $\frac{-2}{2+2}\log_2\left(\frac{2}{2+2}\right) + \frac{-2}{2+2}\log_2\left(\frac{2}{2+2}\right) = \frac{1}{2}+\frac{1}{2} = 1$

Information Gain: Information Gain measures the reduction in entropy or variance that results from splitting a dataset based on a specific property. It is used in decision tree algorithms to determine the usefulness of a feature by partitioning the dataset into more homogeneous subsets with respect to the class labels or target variable.

$$\boxed{\text{Information Gain}(H,A) = H - \sum \frac{(H_v)}{|H|} H_v}$$

A: is the specific attribute or class label
(H) is the entropy of dataset sample S
(H_v) is the number of instances in the subset S that have the value v for attribute A

For the Given Dataset form a decision Tree

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	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

Play Tenis	
Yes	NO
9	5

$\Rightarrow$ Entropy(H) $= -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$

Outlook

Sunny		Overcast		Rain	
Yes	No	Yes	No	Yes	No
2	3	4	0	3	2

$$\text{Entropy}[H_{sunny}] = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

$$\text{Entropy}[H_{overcast}] = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

$$\text{Entropy}(H_{Rain}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$$

$$\text{Gain}(H, H_v) = \text{Entropy}(H) - \sum_{v \in \{Sunny, Overcast, Rain\}} \frac{|H_v|}{|H|}\text{Entropy}(H_v)$$

$$= 0.94 - \frac{5}{14}\text{Entropy}(H_{sunny}) - \frac{4}{14}(H_{overcast}) - \frac{5}{14}\text{Entropy}(H_{rain})$$

$$= 0.94 - \frac{5}{14}(0.971) - \frac{4}{14}(0) - \frac{5}{14}(0.971) = 0.2464$$

$$\Rightarrow 0.2464$$

$$\boxed{\text{Inf. Gain}(H, \text{outlook}) = 0.2464}$$

Attribute : Temperature

Hot		Mild		Cool	
Yes	No	Yes	No	Yes	No
2	2	4	2	3	1

$$\text{Entropy}(H_{Hot}) = 1$$

$$\text{Entropy}(H_{mild}) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.9183$$

$$\text{Entropy}(H_{cool}) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$$

$$\text{Inf. Gain}(H, \text{Temp}) = \text{Entropy}(H) - \frac{2}{14}\text{Entropy}(H_{Hot}) - \frac{6}{14}\text{Entropy}(H_{mild}) - \frac{4}{14}\text{Entropy}(H_{cool})$$

$$\text{Inf Gain}(H, \text{Temp}) = \underline{0.0289}$$

Attribute (Humidity) = (High, Normal)

Humidity			
High		Normal	
Yes	No	Yes	No
3	4	6	1

$$\text{Gain}(H, \text{Humidity}) = 0.1516$$

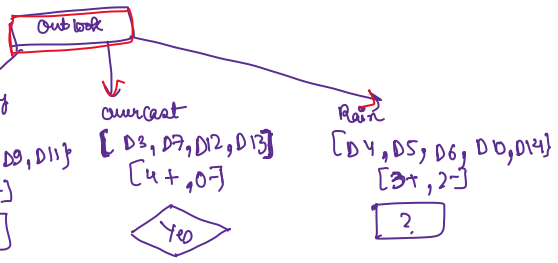
Attribute Wind

Wind			
Strong		Weak	
Yes	No	Yes	No
3	3	6	2

$$\text{Gain}(H, \text{wind}) = 0.0478$$

Attribute	I-Gain
Gain(H, outlook)	0.2464
Gain(H, Temp)	0.0289
Gain(H, Humidity)	0.1516
Gain(H, wind)	0.0478

True Root: Outlook (highest Information Gain)



Day	Temp	Humidity	Play Tennis
D_1	Hot	High	No

Play Tennis	
Yes	No
3	3

$$\text{Entropy}[H] = -\frac{3}{6}\log(\frac{3}{6}) + -\frac{2}{6}\log(\frac{2}{6})$$

D2	Hot	High	No
D8	Mild	High	No
D9	Cool	Normal	Yes
D11	Mild	Normal	Yes

$\frac{0}{5}$ sunny $\frac{5}{5}$ $0\log(5)$ $\frac{5}{5}$ $0\log(5)$

$= 0.97$

Temp					
Hot		Mild		Cool	
Yes	No	Yes	No	Yes	No
0	2	1	1	1	0

Entropy(H_Hot) = 0

Entropy[S_mild] = 1

Entropy[S_cool] = 0

$$IGain = Entropy(H) - \sum_{sunny\ v=\{Hot, mild, Cool\}} \frac{|H_v|}{H} H_v$$

$$= 0.97 - \frac{2}{5}0.0 - \frac{2}{5}1 - \frac{1}{5}0.0 = 0.570$$

$$= 0.570$$

Attribute: Humidity

Value (Humidity) = High, Normal

Entropy (H_{High}) = 0

Entropy (S_{normal}) = 0

Humidity			
High		Normal	
Yes	No	Yes	No
0	3	2	0

$$Gain(H_{sunny}, H_{Humidity}) = Entropy(H) - \sum_{v \in \{High, Normal\}} \frac{|H_v|}{H} Entropy(H_v)$$

$$= 0.97 - \frac{3}{5}(0) - \frac{2}{5}(0)$$

$$= 0.97$$

Attribute : Wind

Values (Wind) = Strong, Weak

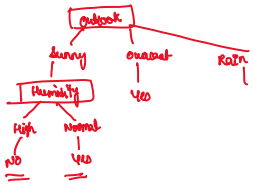
Wind			
Strong		Weak	
Yes	No	Yes	No
1	1	1	2

Entropy (H_{strong}) = 1.0

Entropy (H weak) $= -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.9183$

$$Gain(H_{sunny}, Wind) = 0.97 - \frac{2}{5}1.0 - \frac{3}{5}0.918 = 0.0192$$

$$\boxed{\begin{array}{l} Gain(H_{sunny}, Temp) = 0.570 \\ Gain(H_{sunny}, Humidity) = 0.97 \\ Gain(H_{sunny}, Wind) = 0.192 \end{array}}$$



Day	Temp	Humidity	Wind	Play tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

Rain	
Yes	No
3	2

$$Entropy(H_{Rain}) = \frac{-3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.97$$

Temp					
Hot		Mild		Cool	
Yes	No	Yes	No	Yes	No
0	0	2	1	1	1

Entropy (H_{Hot}) = 0.0

Entropy (H_{mild}) $= -\frac{2}{5}\log_2\frac{2}{5} - \frac{1}{3}\log_2\frac{1}{3} = 0.9183$

Entropy (S_{cool}) = 1.0

$$Gain(H_{Rain}, Temp) = 0.97 - \frac{0}{5}0.0 - \frac{3}{5}0.918 - \frac{2}{5}1.0 = 0.0192$$

Attribute : Humidity

High		Normal	
Yes	No	Yes	No
1	1	2	1

Entropy (H_{High}) = 1.0

Entropy (H_{Normal}) $= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.9183$

$$Gain(H_{rain}, Humidity) = Entropy(H) - \sum_{v \in \{High, Normal\}} \frac{|H_v|}{|H|} Entropy(H_v)$$

$$\boxed{= 0.97 - \frac{2}{5}1.0 - \frac{3}{5}0.918 = 0.192}$$

Attribute wind

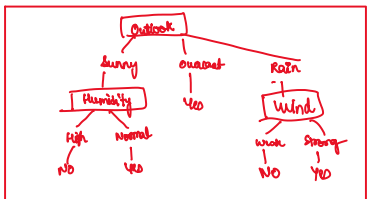
Weak		Strong	
Yes	No	Yes	No
0	2	3	0

$Entropy(H_{weak}) = 0$

$Entropy(H_{strong}) = 0$

$$Inf.\ Gain(H_{rain}, Wind) = Entropy(H_{rain}) - \sum_{v \in \{Strong, Weak\}} \frac{|H_v|}{|H|} Entropy(H_v)$$

$$Inf\ Gain(H_{rain}, Wind) = 0.97 - 0 - 0 = 0.97$$



Final decision tree