Decision Tree - ID3 Algorithm Mathematical calculation

MONDAY
APR '22
188
108-257 WK-17

Day	Outlook	Temp	Humidity	Wind	PlayTennis
01	Sunny	Hot	High	weak	NO
02	Sunny	Hot	High	Strong	NO
03	Overcast	Hot	High	Weak	yes
04	Rain	Mild	High	weak	yes
05	Rain	Cool	Normal	weak	yes
06	Rain	cool	Normal	Strong	NO
07	Overcast	cool	Normal	Strong	yes
08	Sunny	Mild	High	weak	NO
09	Sunny	Cool	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

Entropy : Determines the uncertainity / impurity in the dataset. $\boxed{Entropy(S) = -\sum_{i=1} p_i \log_2(p_i)}$

Information gain (S, outlook) = $Entropy(S) - \sum_{v \in (Sunny, Overcast, Rain)} \frac{S_v}{S_{Sunny}}$ Entropy

Identify attribute which gives maximum info-gain

For Binary classification: Yes/No $\boxed{Entropy(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-)}$

p_+ : Proportion of +ve class

p_- : Proportion of -ve class

TUESDAY
APR '22
19
WK-17 109-256

Outlook Attribute

$S = [+9, 5-]$ Entropy$(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2 5/14 = 0.94$

$S_{sunny} = [2+, 3-]$ Entropy$(S_{sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2 3/5 = 0.971$

$S_{overcast} = [4+, 0-]$ Entropy$(overcast) = \frac{-4}{5}\log_2\frac{4}{5} - \frac{0}{5}\log_2\frac{0}{5} = 0$

$S_{Rain} = [3+, 1-]$ Entropy$(S_{Rain}) = \frac{-3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.97$

Gain$(S, outlook)$ = Entropy$(S) - \sum_{v \in sunny, overcast, Rain} S_v$ (Entropy) S_v

= Entropy$(S) - \frac{5}{14}$ Entropy$(sunny) - \frac{4}{14}$ Entropy$(overcast)$
$- 5/14$ Entropy(S_{Rain})

$= 0.94 - 5/14 \times 0.971 - 4/14 \times 0 - 5/14 \times 0.971$

$= 0.2464$

March 2022

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30	31										

Temperature Attribute

WEDNESDAY
APR '22
20
110-255 WK-17

$S = [9+, 5-] = \text{Entropy}(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2 5/14$

$= 0.94$

$S_{Hot} = [2+, 2-] = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.0$

$S_{mild} = [4+, 2-] = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.9183$

$S_{cool} = [3+, 1-] = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2 1/4 = 0.8113$

$\text{Gain}(S_{Temp}) = \text{Entropy}(S) - \sum_{v \in (Hot, mild, cool)} \text{Entropy}(Temp)$

$= 0.94 - \frac{4}{14} \times 1.0 - \frac{6}{14} \times 0.9183 - \frac{4}{14} \times 0.8113$

$= 0.0289$

Humidity Attribute

$\text{Entropy}(S) = 0.94$

$S_{High} = [3+, 4-] = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.9852$

$S_{Normal} = [6+, 1-] = -\frac{6}{7}\log_2 6/7 - \frac{1}{7}\log_2\frac{1}{7} = 0.5916$

$\text{Gain}(S_{Humidity})$

$= 0.94 - \frac{7}{14} \times 0.9852 - 7/14 \times 0.5916$

$= 0.1516$

May 2022

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
						1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31					

Wind Attribute.

Entropy(S) = 0.94

$S_{strong}[3+, 3-]$ = Entropy = 1.0

$S_{weak}[6+, 2-] = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$

$= 0.8113$

Gain (wind) = Entropy(S) − $\sum_{v \in strong, weak}$ Entropy(strong, weak)

$= 0.94 - \frac{6}{14} \times 1.0 - \frac{8}{14} \times 0.8113$

$= 0.0478$

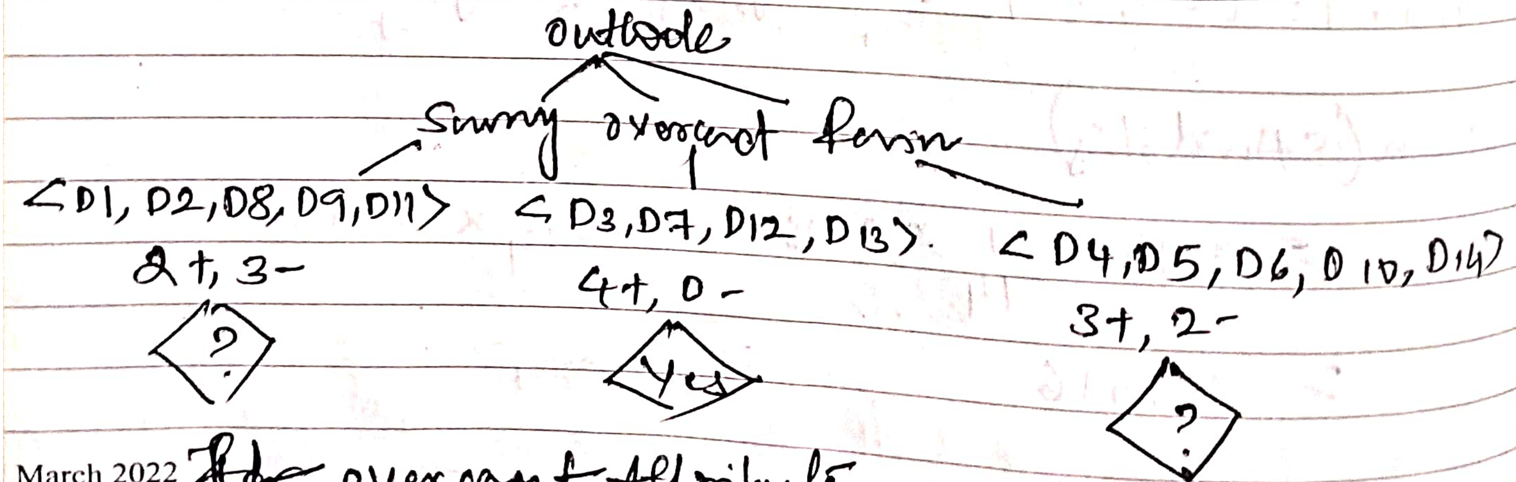
Gain of All Individual Attribute.

Gain (S, Outlook) = 0.2464 ✓ Max gain on root node.

Gain (S, Temp) = 0.0289

Gain (S, Humidity) = 0.1516

Gain (S, wind) = 0.0478



For overcast Attribute, result is yes.
but for sunny & Rain still not got the value.

Sunny Attribute & Play we need to identify the FRIDAY word into gain

APR '22
22
112 253 WK 17

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

Attribute Temp.

value of Temp = Hot, Mild, Cool

$S_{sunny} = [2+, 3-] = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2 3/5 = 0.97$

$S_{hot} = [0+, 2-] =$ Entropy(S_{hot}) $= 0.0$

$S_{mild} = [1+, 1-] =$ Entropy(S_{mild}) $= 1.0$

$S_{cool} = [1+, 0-] =$ Entropy(S_{cool}) $= 0.0$

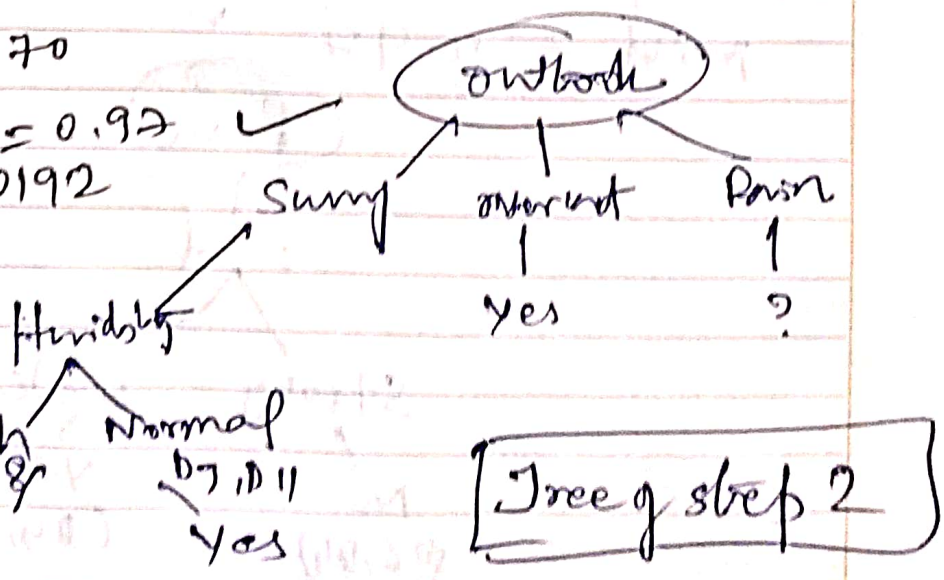
Gain(Ssunny, Temp) = Entropy(S) − $E_{H,M,C}$

$= 0.97 = \frac{2}{5}\times 0.0 - \frac{2}{5}\times 1.0 - \frac{1}{5}\times 0.0$

Attribute Temp = 0.570

Attribute: Humidity = 0.97

Attribute Wind = 0.0192



Tree of step 2

May 2022

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
						1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31					

SATURDAY
APR '22
1233
WK-17 113-252

Attributes				
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

$S_{Rain} = [2+, 2-]$

$= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{3}{5} = 0.97$

$S_{Hot} = [0+, 0-] = 0.0$

$S_{Mild} = [2+, 1-] = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.9183$

$S_{Cool} = [1+, 1-] = S_{Cool} = 1.0$

$S_{Rain} - \sum_{v \in \{Hot, Mild, Cool\}} Entropy(Hot, Mild, Cool)$

$= 0.97 - 0/5 \times 0.0 - 3/5 \cdot 0.918 - 2/5 \cdot X \cdot 1.0$

$= 0.0192$

$S_{Humidity} = 0.0192$

$S_{Mild} = 0.97$ ✓