1

DECISION TREE ALGORITHM - ID3 EXAMPLE

A1	A2	A3	Classification	Instance
True	Hot	High	No	1
True	Hot	High	No	2
False	Hot	High	Yes	3
False	Cool	Normal	Yes	4
False	Cool	Normal	Yes	5
True	Cool	High	No	6
True	Hot	High	No	7
True	Hot	Normal	Yes	8
False	Cool	Normal	Yes	9
False	Cool	High	Yes	10

Attribute : A1

Values (A1) = True, False

$S = [6+, 4-]$ → proportion of +ev examples

$$\text{Entropy}(S) = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10}$$

10 → total examples

Entropy(S) = 0.9709

proportion of -ev examples

$S_{True} = [1+, 4-]$

$$\text{Entropy}(S_{True}) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$$

$\text{Entropy}(S_{True}) = 0.7219$

$S_{False} = [5+, 0-]$

$Entropy(S_{False}) = 0$

$$Gain(S, A1) = Entropy(S) - \sum_{v \in \{True, False\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, A1) = Entropy(S) - \frac{5}{10} Entropy(S_{True}) - \frac{5}{10} Entropy(S_{False})$$

$$Gain(S, A1) = 0.9709 - \frac{5}{10}(0.7219) - \frac{5}{10}(0) = 0.6099$$

<u>Attribute: A2</u>

$Values(A2) = Hot, Cool$

$S = [6+, 4-]$

$$Entropy(S) = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10}$$

$$Entropy(S) = 0.9709$$

(3)

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

$$Entropy(S_{Hot}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.9709$$

$Entropy(S_{cool}) = [4+, 1-]$

$$Entropy(S_{cool}) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}$$

$$= 0.7219$$

$$Gain(S, A_2) = Entropy(S) - \sum_{v \in \{Hot, cool\}} \frac{|S_v|}{|S|} Entropy(S_{cool})$$

$$Gain(S, A_2) = Entropy(S) - \frac{5}{10} Entropy(S_{Hot}) - \frac{5}{10} Entropy(S_{cool})$$

$$Gain(S, A_2) = 0.9709 - \frac{5}{10}(0.9709) - \frac{5}{10}(0.7219)$$

$Gain(S, A_2) = 0.1245$

Attribute : A_3

Values (A_3) = High, Normal

$S = [6+, 4-]$

$$Entropy(S) = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10}$$

$Entropy(S) = 0.9709$

$S_{High} = [2+, 4-]$

$$Entropy(S_{High}) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}$$

$= 0.9138$

$S_{Normal} = [4+, 0-]$

$Entropy(S_{Normal}) = 0$

$$Gain(S, A_3) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, A_3) = Entropy(S) - \frac{6}{10} Entropy(S_{High}) - \frac{4}{10} Entropy(S_{Normal})$$

$$Gain(S, A_3) = 0.9709 - \frac{6}{10}(0.9183) - \frac{4}{10}(0) = 0.4199$$

$Gain(S, A_1) = 0.6099$ Max Gain

$Gain(S, A_2) = 0.1245$

$Gain(S, A_3) = 0.4199$

A1

True / False

1, 2, 6, 7, 8   3, 4, 5, 9, 10

Yes

Instance	A_2	A_3	Classification
1	Hot	High	No
2	Hot	High	No
6	Cool	High	No
7	Hot	High	No
8	Hot	Normal	Yes

Attribute: A_2

Values (A_2) = Hot, Cool

$S_{A1} = [1+, 4-]$

$$\text{Entropy}(S_{A1}) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$$
$$= 0.7219$$

$S_{Hot} = [1+, 3-]$

$$\text{Entropy}(S_{Hot}) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$
$$= 0.8112$$

$S_{Cool} = [0+, 1-]$

$\text{Entropy}(S_{Cool}) = 0$

(7)

$\text{Gain}(S, A_2)_{A_1} = \text{Entropy}(S) - \sum_{v \in \{Hot, Cool\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$

$\text{Gain}(S, A_2)_{A_1} = 0.9709 - \frac{4}{5}(0.8112) - \frac{1}{5}(0)$

$\text{Gain}(S, A_2)_{A_1} = 0.3219$

Attribute: A_3

Values (A_3) = High, Normal

$S_{A_1} = [1+, 4-]$

$\text{Entropy}(S_{A_1}) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$

$= 0.7219$

$S_{High} = [0+, 4-]$

$\text{Entropy}(S_{High}) = 0$

$S_{Normal} = [1+, 0]$

$Entropy(S_{Normal}) = 0$

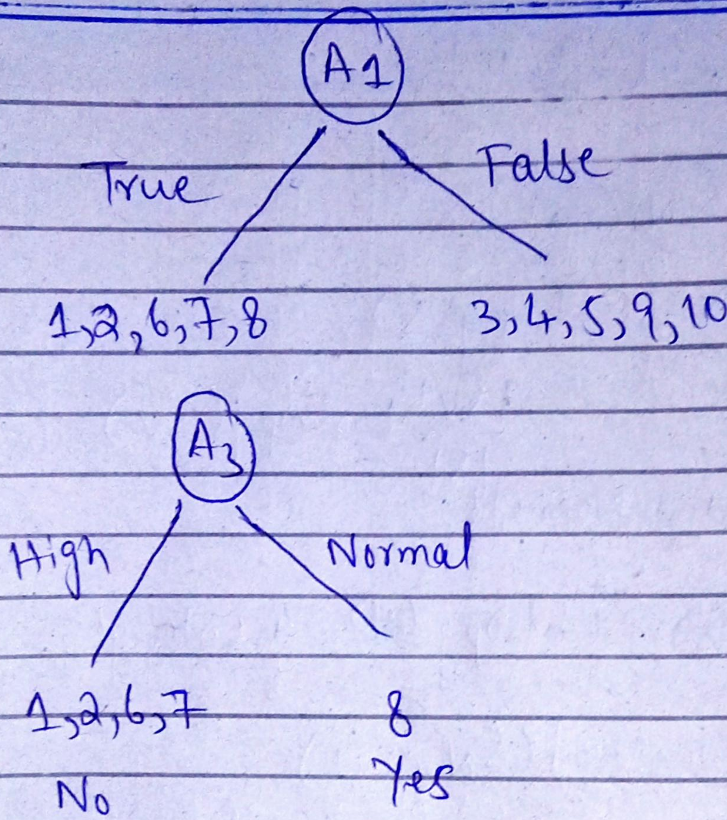
$$Gain(S_{A1}, A_3) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{A1}, A_3) = Entropy(S) - \frac{4}{5} Entropy(S_{High}) - \frac{1}{5} Entropy(S_{Normal})$$

$$Gain(S, A_3) = 0.7219 - \frac{4}{5}(0) - \frac{1}{5}(0)$$
$$= 0.7219$$

$Gain(S_{A1}, A_2) = 0.3219$

$Gain(S_{A1}, A_3) = 0.7219$ - Max Gain



—x—x—