classmate
Date
Page

Decision Tree

Instance	a_1	a_2	a_3	Classification
1	True	Hot	High	No
2	True	Hot	High	No
3	False	Hot	High	yes
4	False	Cool	Normal	yes
5	False	Cool	Normal	yes
6	True	Cool	High	No
7	True	Hot	High	No
8	True	Hot	Normal	yes
9	False	Cool	Normal	yes
10	False	Cool	High	yes

⇒ For Entropy of All data

Distinct values in classification	Total
Yes	6
NO	4
	10

$$\text{Entropy}(D) = -\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right)$$

$$= 0.9709$$

$$\left[\because \log_y x = \frac{\log_{10} x}{\log_{10} y}\right]$$

⇒ Gain of a_1 :

$$Gain(D, a_1) = Entropy(D) - Entropy(a_1)$$

Entropy(a_1) :=

Distinct values in a_1	Yes	NO	Total
True	1	4	5
False	5	0	$\frac{5}{10}$

$$Entropy(a_1) = \frac{5}{10} \times \left[\frac{-1}{5} \log_2\left(\frac{1}{5}\right) - \frac{4}{5} \log_2\left(\frac{4}{5}\right)\right]$$

$$+$$

$$\frac{5}{10} \times \left[\frac{-5}{5} \log_2\left(\frac{5}{5}\right) - \frac{0}{5} \log_2\left(\frac{0}{5}\right)\right]$$

$$= 0.7219 \times \frac{5}{10} = \boxed{0.3609}$$

$$Gain(D, a_1) = 0.9709 - (0.7219) \times \frac{5}{10}$$

$\therefore$ Gain $(D, a_1) = 0.9709 - 0.3609$

$= \boxed{0.6099}$

$\Rightarrow$ Gain of a_2 :-

Distinct values in a_2	yes	NO	Total
Hot	2	3	5
Cool	4	1	$\frac{5}{10}$

$$\text{Gain}(D, a_2) = 0.9709 - \left\{ \underbrace{\frac{5}{10}\left[\frac{-2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]}_{\text{Hot}} + \underbrace{\frac{5}{10}\left[\frac{-4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right]}_{\text{Cool}} \right\}$$

$= \boxed{0.1245}$

⇒ Gain for a_3 :

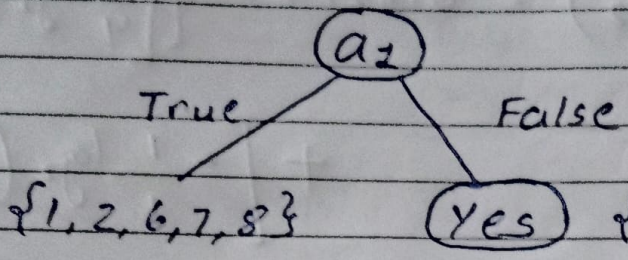
Distinct values in a_3	Yes	No	Total
High	2	4	6
Normal	4	0	4

$$Gain(D, a_3) = 0.9709 - \left\{ \frac{6}{10} \left[\frac{-2}{6} \log_2\left(\frac{2}{6}\right) - \frac{4}{6} \log_2\left(\frac{4}{6}\right) \right] + \frac{4}{10} \left[\frac{-4}{4} \log_2\left(\frac{4}{4}\right) \right] \right\}$$

$$= 0.4200$$

→ $Gain(D, a_1) = 0.6099$ → Maximum
$Gain(D, a_2) = 0.1245$
$Gain(D, a_3) = 0.4200$

=>



New Data

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

Entropy (D)

$$= -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right)$$

$$= 0.7219$$

Dist. value in classification	Count
Yes	1
NO	4
	5

* Gain of a_2

Gain (D, a_2) = Entropy (D) − Entropy (a_2)

Distinct values in a_2	Yes	No	Total
Hot	1	3	4
Cool	1	0	1
			5

$$Gain(D, a_2) = 0.7219 - \left\{\frac{4}{5}\left[-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right] + \frac{1}{5}\left[-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right]\right\}$$

$$= \boxed{0.0729}$$

Gain for a_3:

Distinct values	Yes	No	Total
High	0	4	4
Normal	1	0	1
			5

$$Gain(D, a_3) = 0.7219 - \left\{\frac{4}{5}\left[-\frac{4}{4}\log_2(4/4)\right] + \frac{1}{5}\left[-\frac{1}{1}\log_2(1/1)\right]\right\}$$

$$= \boxed{0.7219}$$ ——— Max.

classmate
Date
Page