Name: Samruddhi Kalsekar
UIN: 678393073

CS 412 - Introduction to Machine Learning

Assignment 08

Q1.)

Example	A_1	A_2	A_3	Output y
x_1	1	0	0	0
x_2	1	0	1	0
x_3	0	1	0	0
x_4	1	1	1	1
x_5	1	1	0	1

Information Gain : IG (A) of attribute A | Entropy of attribute A : E (A)

→ ∴ $IG(y) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)$

$= -(0.6)(-0.73) - (0.4)(-1.32)$

$IG(y) = 0.966$

→ $E(A_1) = \frac{4}{5}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) +$

$\frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1} - 0\right)$

$E(A_1) = \frac{4}{5}\left(-0.5(-1) - 0.5(-1)\right) + \frac{1}{5}(0)$

$E(A_1) = 0.8$

∴ $Gain(A_1) = 0.966 - 0.8 = 0.166$ .. (i)

$$\rightarrow E(A_2) = \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2} - 0\right) + \frac{3}{5}\left(\frac{-1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{1}{3}\right)$$

$$= \frac{2}{5}(0) + \frac{3}{5}\left(\frac{-1}{3}(-1.5) - \frac{2}{3}(-0.5)\right)$$

$$= \frac{3}{5}\left(0.5 + 1/3\right) = 0.5$$

$\therefore$ Gain $= 0.966 - 0.5 = 0.466$ .. (ii)

$$\rightarrow E(A_3) = \frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) +$$

$$\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

$$E(A_3) = \frac{3}{5}(0.389 + 0.52) + \frac{2}{5}(0.5 + 0.5)$$

$$= 0.945$$

Gain $= 0.966 - 0.94 = 0.026$ .. (iii)

$\therefore$ From (i) (ii) & (iii)

A_2 has the highest gain, $\therefore$ A_2 is the root node

A_2

$=0 \swarrow \quad \searrow =1$

$\boxed{y=0}$

→ $IG(A_2=1) = \left(-\frac{1}{3}\log_2\frac{1}{3}\right) - \left(\frac{2}{3}\log_2\frac{2}{3}\right)$

$= 0.914$

→ $E(A_1) = \frac{1}{3}\left(-1\log_2 1\right) + \frac{2}{3}\left(-\frac{2}{2}\log_2\frac{2}{2}\right)$

$= 0+0 = 0$

$E(A_1) = 0$

Gain $= 0.914 - 0 = 0.914$ .. (iv)

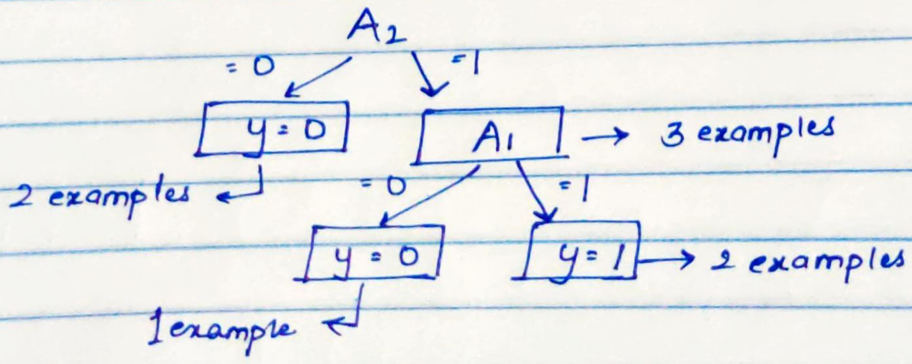
→ $E(A_3) = \frac{1}{3}\left(-1\log_2 1\right) + \frac{2}{3}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right)$

$= 0 + 2/3 = 0.667$

Gain $= 0.914 - 0.667 = 0.248$ .. (v)

From (iv) & (v)

Attribute A_1 has the highest gain



A

Q2. a) $x_6 = (A_1 = 1, A_2 = ?, A_3 = 1)$

Weighted average $= w_0 p_0 + w_1 p_1$

$P_0 = P(A_2 = 0 \mid A_1 = 1, A_3 = 1) = 1/2$
$P_1 = P(A_2 = 1 \mid A_1 = 1, A_3 = 1) = 1/2$

Weighted average $= (1/2)(0) + (1/2)(1)$

$= 1/2 = 0.5$ .. (i)

$\therefore$ if $A_2 = 0$, weighted average < 0.5
and $A_2 = 1$, weighted average ≥ 0.5

$\therefore$ from (i) as weighted average $= 0.5$

$\boxed{A_2 = 1}$

Q2 b) $x_4 = (A_1 = 1, A_2 = ?, A_3 = 1)$; its A_2 attribute is not observed

We cannot calculate the information gain of attribute A_2 due to unobserved data.

∴ There are two cases

$p(A_2 = 0) = 1/2$

$p(A_1 = 0) = 1/2$

→ Calculating information gain for each

- $A_2 = 0$

Entropy on splitting A_1

$$= \frac{4}{5}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)$$

$$+ \frac{1}{5}\left(-\frac{1}{1}\log 1\right) = 0.8$$

Entropy on splitting A_2

$$= \frac{3}{5}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) +$$

$$\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$$

$$= 0.951$$

Entropy on splitting $A_3 = 0.951$

- $A_2 = 1$

Similarly,
Entropy on splitting $A_1 = 0.8$

Entropy on splitting $A_2 = \frac{2}{5}\left(\frac{-2}{2}\log_2 1\right) + \frac{3}{5}\left(\frac{-1}{3}\log_2 \frac{1}{3}\right)$

$= 2/3 \log_2 2/3 = 0.551$

Entropy on splitting $A_3 = 0.951$

To calculate new entropy

$E_1 = \frac{1}{2}(0.8) + \frac{1}{2}(0.8) = 0.8$

$E_2 = \frac{1}{2}(0.951) + \frac{1}{2}(0.551) = 0.751$

$E_3 = \frac{1}{2}(0.951) + \frac{1}{2}(0.951) = 0.951$

Information Gain on splitting by

$A_1 = 0.971 - 0.8 = 0.171$

$A_2 = 0.971 - 0.751 = 0.22$

$A_3 = 0.971 - 0.951 = 0.02$

$\therefore$ information gain is highest for attribute A_2,
Root (decision tree will be split at) $\boxed{A_2}$