Assignment - 8

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

There are three binary features A_1, A_2, A_3.

Information gain for A_1.

Group 1 = x_3

Group 2 = $x_1, x_2, x_4 x_5$

Entropy = $-\sum_{i=1}^{2} p_i \log_2 p_i$

$E(\text{Group1}) = -(1 \log_2 1) = 0$

$E(\text{Group 2}) = -(1/2 \log_2 1/2 + 1/2 \log_2 1/2)$

$= 1/2 + 1/2 = 1$

$E(\text{split}) = 1/5 \times 0 + 4/5 \times 1$

$= 4/5$

$I = 97 - E(\text{Split})$

$= 1/5 = 0.2$

Information gain for A_2

group 1 - x_1, x_2
group 2 - x_3, x_4, x_5

$$\text{Entropy} = -\sum_{i=1}^{n} P_i \log_2 P_i$$

$$E(\text{group 1}) = -(0) = 0$$

$$E(\text{group 2}) = -\left(\tfrac{1}{3}\log_2 \tfrac{1}{3} + \tfrac{2}{3}\log_2 \tfrac{2}{3}\right)$$

$$= 0.528 + 0.389$$

$$= 0.917$$

$$E(\text{split}) = \tfrac{2}{5} \times 0 + \tfrac{3}{5} \times 0.917$$

$$= 0.5502$$

$$I = .97 - 0.5502$$

$$= 0.4498$$

Information gain for A_3

group 1 - x_1, x_3, x_5
group 2 - x_2, x_4

$$\text{Entropy} = -\sum_{i=1}^{n} P_i \log_2 P_i$$

$$E(\text{group 1}) = -\left(\tfrac{2}{3}\log_2 \tfrac{2}{3} + \tfrac{1}{3}\log_2 \tfrac{1}{3}\right)$$

$$= 0.917$$

$$E(\text{group 2}) = -\left(\tfrac{1}{2}\log_2 \tfrac{1}{2} + \tfrac{1}{2}\log_2 \tfrac{1}{2}\right)$$

$$= 1$$

$$E(\text{Split}) = \frac{3}{5} \times 0.917 + \frac{2}{5} \times 1$$

$$= 0.5502 + 0.4$$

$$= 0.9502$$

$$I = .97 - 0.9502$$

$$= 0.0498$$

Since A_2 split has the highest information gain we will consider A_2.

Left node $= \{x_1, x_2\}$ Right node $= \{x_3, x_4, x_5\}$

left node

Information gain A_1

group1 = {null} group2 = $\{x_1, x_2\}$

$E(\text{group1}) = 0$

$E(\text{group2}) = -\sum_{i=1}^{n} P_i \log_2 P_i = -(1 \log_2 1) = 0$

$E(\text{Split}) = 0 \times 0 + \frac{2}{2} \times 0 = 0$

$I = 0.97$

Information gain A_2

group1 = $\{x_1, x_2\}$ group2 = {null}

$E(\text{group1}) = -(1 \log_2 1) = 0$

$E(\text{group2}) = 0$

$E(\text{Split}) = \frac{2}{2} \times 0 + 0 \times 0 = 0$

$I = 0.97$

Information gain A_3

group1 = $\{x_1\}$ group2 = $\{x_2\}$

$E(group1) = -(1 \log_2 1) = 0$
$E(group2) = -(1 \log_2 1) = 0$
$E(Split) = 0$
$I = 0.97$

Best split for left node = A_3.

Right node

Information gain A_1

group1 = $\{x_3\}$ group2 = $\{x_4, x_5\}$

$E(group1) = -(1 \log_2 1) = 0$
$E(group2) = -(1 \log_2 1) = 0$
$E(Split) = 0$
$I = 0.97$

Information gain for A_2

group1 = {null} group2 = $\{x_3, x_4, x_5\}$
$E(group1) = 0$
$E(group2) = -(1/3 \log_2 1/3 + 2/3 \log_2 2/3)$
$= 0.528 + 0.389$
$= 0.917$
$E(Split) = 0 + 1 \times 0.917$
$= 0.917$
$I = 0.083$

Information gain A_3

group1 = $\{x_3, x_5\}$ group2 = $\{x_4\}$

$E(\text{group1}) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right)$

$= 1$

$E(\text{group2}) = -(1\log_2 1) = 0$

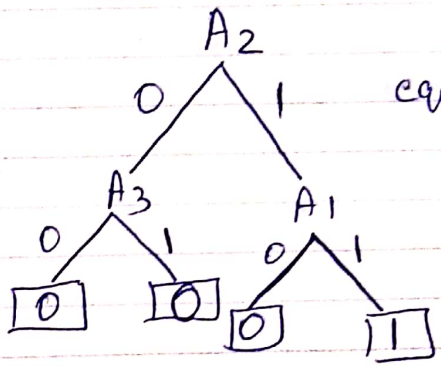
$E(\text{Split}) = \frac{2}{3} \times 1 + \frac{1}{3} \times 0$

$= 0.6667$

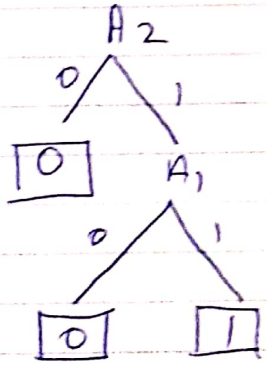
$I = 0.34$

Best split for right node = A_1

Tree



equivalent to



2b) $x_4 = (A_1 = 1, A_2 = ?, A_3 = 1)$

Attribute A_1

group 1 = x_3

group 2 = x_1, x_4, x_2, x_5

$E(\text{group 1}) = 0$

$E(\text{group 2}) = \frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}$

$= 1/2 + 1/2$

$= 1$

$E(\text{Split}) = \frac{1}{5} \times 0 + \frac{4}{5} \times 1$

$= 4/5$

$I = 0.97 - 4/5 = 0.17$

Attribute A_2.

Case 1

$x_4 = 1\ 0\ 1$

group 1 = x_1, x_2, x_4

group 2 = x_3, x_5

$E(\text{group 1}) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.917$

$E(\text{group 2}) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$

$E(\text{Split}) = \frac{3}{5} \times 0.917 + \frac{2}{5} \times 1$

$= 0.9502$

$I = 0.0198$

Case 2

$x_4 = 1\ 1\ 1$

group 1 - x_1, x_2
group 2 - x_3, x_4, x_5

$E(\text{group 1}) = 0$

$E(\text{group 2}) = -\left(\frac{1}{3}\log_2 \frac{1}{3} + \frac{2}{3}\log_2 \frac{2}{3}\right)$
$= 0.917$

$E(\text{Split}) = \frac{2}{5} \times 0 + \frac{3}{5} \times 0.917$

$= 0.5502 \quad I = 0.97 - 0.5502 = 0.4198$

Attribute A_3

group 1 - x_1, x_3, x_5
group 2 - x_2, x_4

$E(\text{group 1}) = -\left(\frac{2}{3}\log_2 \frac{2}{3} + \frac{1}{3}\log_2 \frac{1}{3}\right)$
$= 0.917$

$E(\text{group 2}) = -\left(\frac{1}{2}\log_2 \frac{1}{2} + \frac{1}{2}\log_2 \frac{1}{2}\right)$
$= 1$

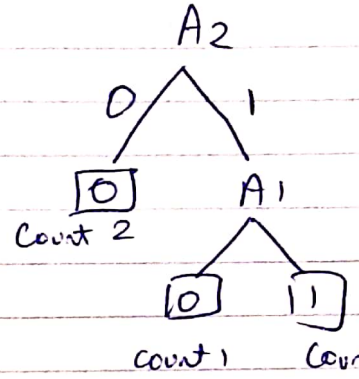
$E(\text{Split}) = \frac{3}{5} \times 0.917 + \frac{2}{5} \times 1$

$= 0.9502$

$I = 0.97 - 0.9502$
$= 0.019$

Since the information gain of A_2 (weighted) is 0.2198 is the highest among all 3 A_2 can be considered as the root.

2a Decision Tree



$X_6 = (A_1=1, A_2=?, A_3=1)$

$X_6 = 101$ goes into the left subtree with output label 0

$X_6 = 111$ goes into the right subtree with output label 1

The weighted avg is $w_0 \times p_0 + w_1 p_1$

$w_0 = \frac{2}{4}$ $w_1 = \frac{2}{4}$ $p_0 = 0$ $p_1 = 1$

$= \frac{1}{2} \times 0 + \frac{1}{2} \times 1 = 0.5$

$$A_2 = \begin{cases} 0 & \text{weighted avg} < 0.5 \\ 1 & \text{weighted avg} \geq 0.5 \end{cases}$$

Since weighted avg $= 0.5$ $A_2 = 1$

$X_6 = A_1 = 1, A_2 = 1, A_3 = 1$

The prediction will be 1.