Decision tree

Page No.
Date

Q1) A company wants to predict whether a customer will subscribe to a premium membership based on their demographic and browsing behaviour data. Use the decision tree to predict the given example

Age	Gender	Income	Browsing time	subscription
20-30	Male	High	10am - 12pm	Yes
22-30	Female	medium	2pm - 4pm	Yes
30-40	Male	low	8am - 10am	No
30-40	Female	high	4pm - 6pm	Yes
>40	Male	medium	6pm - 8pm	Yes
>40	Female	medium	8am - 10am	No
>40	Male	high	12pm - 2pm	Yes
20-30	Female	low	10am - 12pm	No
20-30	Male	medium	2 pm - 4pm	Yes
30-40	Female	high	8 am - 10 am	Yes

$\Rightarrow S = [7+, 3-]$

$$\text{Entropy}(S) = -\frac{\text{No of yes}}{\text{Total}} \log_2 \frac{\text{No of yes}}{\text{total}} - \frac{\text{No of no}}{\text{total}} \log_2 \frac{\text{No of no}}{\text{total}}$$

$$= -\frac{7}{10} \log_2 \frac{7}{10} - \frac{3}{10} \log_2 \frac{3}{10} = 0.881$$

Attribute : Gender

$S_{\text{male}} = [4+, 1-]$

$$\text{Entropy}(S_{\text{male}}) = -\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} = 0.721$$

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

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

$= 0.971$

$\text{Gain}(S, \text{Gender}) = \text{Entropy}(S) - \sum \frac{|S_v|}{|S|}\text{Entropy}(S_v)$

$= 0.0881 - \frac{5}{10}\text{Entropy}(S_{male}) - \frac{5}{10}\text{Entropy}(S_{female})$

$= 0.881 - \frac{5}{10} \times 0.721 - \frac{5}{10} \times 0.971 = 0.035$

Attribute : age

$S_{20-30} = [3+, 1-] \quad S_{30-40} = [2+, 1-]$

$\text{Entropy}(S_{20-30}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.811$

$\text{Entropy}(S_{30-40}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918$

$S_{>40} = [1+, 1-]$

$\text{Entropy}(S_{>40}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$\text{Gain} = \text{Entropy} - \sum \frac{|S_v|}{|S|}\text{Entropy}(S_v)$

$= 0.0881 - \frac{4}{10}\text{Entropy}(S_{20-30}) - \frac{3}{10}\text{Entropy}(S_{30-40}) - \frac{2}{10}\text{Entropy}(S_{>40})$

$= 0.881 - \frac{4}{10} \times 0.811 - \frac{3}{10} \times 0.918 - \frac{2}{10} \times 1 = 0.0812$

Attribute : Income

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

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

$S_{low} = [0+, 2-]$

Entropy $(S_{High}) = 0$

Entropy $(S_{medium}) = 0.811$

Entropy $(S_{low}) = 0$

Gain $= 0.881 - 0.811 \times \frac{3}{10} = 0.6377$

Attribute : Browsing time

$S_{10am-12pm} = [1+, 1-]$

$S_{2pm-4pm} = [2+, 0-]$

$S_{8am-10am} = [1+, 1-]$

$S_{4pm-6pm} = [1+, 0-]$

$S_{6pm-8pm} = [1+, 0-]$

$S_{12pm-2pm} = [1+, 0-]$

Entropy $(S_{10am-12pm}) = 0$

Entropy $(S_{2pm-4pm}) = 0$

Entropy $(S_{8am-10am}) = 0.918$

Entropy $(S_{4pm-6pm}) = 0$

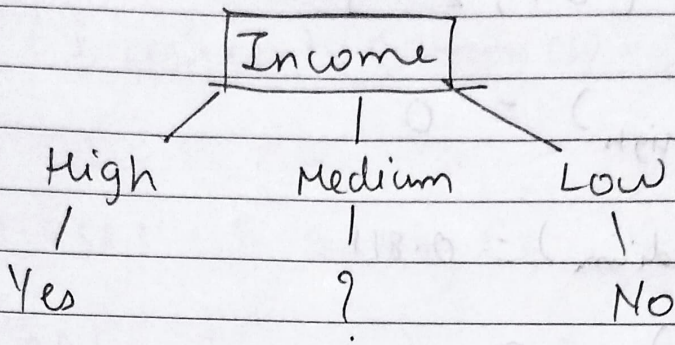
Entropy $(S_{6pm-8pm}) = 0$

Entropy $(S_{12pm-2pm}) = 0$

Page No.
Date

Gain $= 0.881 - \frac{3}{10} \times 0.918 = 0.6056$

We will consider income as root node



Age	Gender	Browsing time	Subscription
20-30	Female	2pm - 4pm	Yes
>40	Male	6pm - 8pm	Yes
>40	Female	8am - 10am	No
20-30	Male	2pm - 4pm	Yes

$S = [3+, 1-]$

$S = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.811$

Attribute : Age

$S_{20-30} = [2+, 0-]$

$S_{>40} = [1+, 1-]$

Entropy $(S_{20-30}) = 0$

Entropy $(S_{>40}) = 1$

Gain $= 0.811 - 1 \times \frac{2}{5} = 0.411$

Page No.
Date

Attribute : Gender

$S_{male} = [2+, 0-]$

$S_{female} = [1+, 1-]$

Entropy $(S_{male}) = 0$

Entropy $(S_{female}) = 1$

Gain $= 0.811 - 1 \times \frac{2}{5} = 0.411$

Attribute : Browsing time

$S_{2pm-4pm} = [2+, 0-]$

$S_{6pm-8pm} = [1+, 0-]$

$S_{8am-10am} = [0+, 1-]$

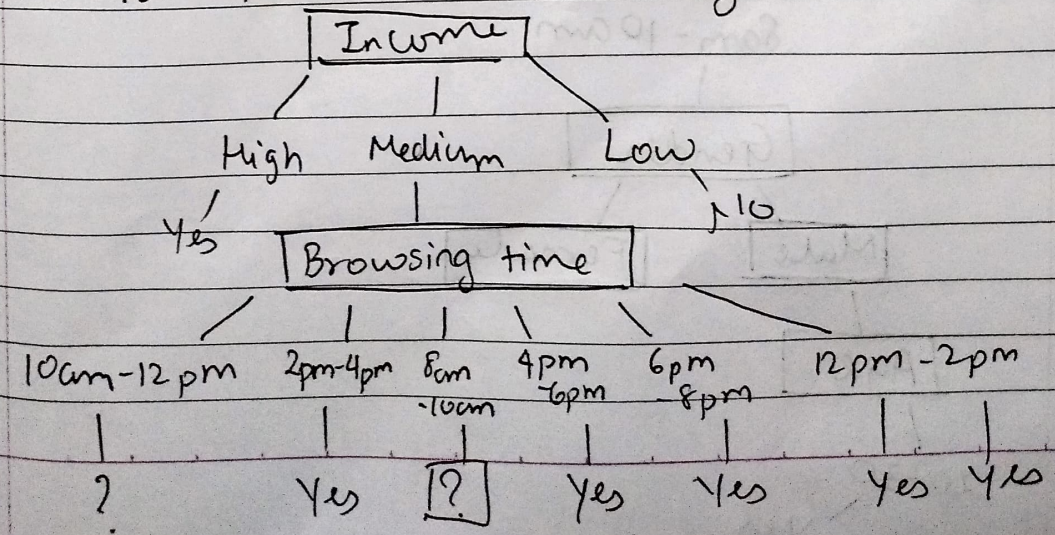
Entropy $(S_{2pm-4pm}) = 0$

Entropy $(S_{6pm-8pm}) = 0$

Entropy $(S_{8am-10am}) = 0$

Gain $= 0.811 - 0 = 0.811$

The root node becomes browsing time



For 10am - 12pm

	Age	Gender	Subscription
20-30		Male	Yes
20-30		Female	No

Attribute : Age

$E(S_{20-30}) = 1$

Gain = 1 − 1 × 1

Gain = 0

Attribute : Gender

$E(S_{Male}) = 0$

$E(S_{Female}) = 0$

Gain = 1

Considering gender as root node