SRIHARI THYAGARAJAN
I066
B2

01/04/23 ML LAB 11 MANUAL WORK

(I) ENTROPY:

* We calculate Entropy of entire dataset:

① $S = [10+, 10-]$

$\therefore \text{Entropy}(S) = -\frac{10}{14}\log_2\frac{10}{14} - \frac{10}{14}\log_2\frac{10}{14}$

$= 0.693$

Considering features - Gender, Car Type and Shirt Size:

② Gender:

$S_{Gender='M'} \leftarrow [4+, 6-] \Rightarrow \text{Entropy} = -\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10}$

$= 0.97$

$S_{Gender='F'} \leftarrow [6+, 4-] \Rightarrow \text{Entropy} = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10}$

$= 0.97...$

$\Rightarrow \text{Gain}(S, \text{Gender}) = \text{Entropy}(S) - \sum_{v \in (Male, Female)} \frac{|S_v|}{|S|}\text{Entropy}(S_v)$

$= 0.693 - \frac{10}{14}(0.97) - \frac{10}{14}(0.97)$

$= -0.693$

Similarly:

③ Car Type:

$S_{Family} \leftarrow [3+, 1-] \Rightarrow \text{Entropy} = 0.811$

$S_{sports} \leftarrow [0+, 8-] \Rightarrow \text{Entropy} = 0$

$S_{Luxury} \leftarrow [7+, 1-] \Rightarrow \text{Entropy} = 0.543$

$\Rightarrow \text{Gain}(S, \text{Car Type}) = \text{Entropy}(S) - \sum_{v \in sports, family, luxury} \frac{|S_v|}{|S|}\text{Entropy}(S_v)$

$= 0.693 - \frac{4}{14}(0.811) - 0 - \frac{8}{14}(0.543) = 0.151$

Similarly.

④ Shirt size :

$S_{small} \leftarrow [2+, 3-] \Rightarrow Entropy = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.971$

$S_{medium} \leftarrow [4+, 3-] \Rightarrow$ Entropy $= 0.985$

$S_{Large} \leftarrow [1+, 2-] \Rightarrow$ Entropy $= 0.918$

$S_{Extralarge} \leftarrow [2+, 2-] \Rightarrow$ Entropy $= 1$

$\Rightarrow Gain(S, \text{Shirt size}) = 0.693 - \frac{5}{14}(0.971) - \frac{7}{14}(0.985)$ $\frac{10}{20}$

$- \frac{3}{14}(0.918) - \frac{4}{14}(1) = -0.629$

From ② ③ and ④

Gain(S, Gender) $= -0.693$

Gain(S, Car Type) $= 0.151$

Gain(S, Shirt size) $= -0.629$

Hence Car Type is root node (Max value).

⑪ GINI :

* We calculate GINI of entire dataset :

① $G = [10+ ; 10-]$ I O

$\Rightarrow GINI(S) = 1 - \left[\left(\frac{10}{20}\right)^2 + \left(\frac{10}{20}\right)^2\right]$

$= 1 - \frac{1}{2} = \frac{1}{2}$

② GINI index for Gender :

$$\text{Gini (gender = 0)} = 1 - \left[\left(\frac{4}{10}\right)^2 + \left(\frac{6}{10}\right)^2\right]$$

$$= 0.48$$

$$\text{Gini (Gender = 1)}$$

$$= 1 - \left[\left(\frac{5}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right]$$

$$= 0.48$$

$$\text{Weighted average} = 0.48\left(\frac{10}{20}\right) + 0.48\left(\frac{10}{20}\right)$$

$$= 0.48$$

③ GINI Index Car Type :

$$\text{Gini (Car Type 0)} = 1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right]$$

$$= 0.375$$

$$\text{Gini (Car Type 1)} = 1 - \left[\left(\frac{1}{8}\right)^2 + \left(\frac{7}{8}\right)^2\right]$$

$$= 0.21875$$

$$\text{Gini (Car Type 2)} = 1 - \left(\frac{8}{8}\right)^2$$

$$= 0$$

$$\text{Weighted} = 0.375\left(\frac{4}{20}\right) + 0.21875\left(\frac{8}{20}\right) + 0$$

$$= 0.1625$$

④ GINI Index Shirt size :

$$\text{For Shirt size 0 :} = \left(1 - \left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right)$$

$$= 0.5$$

GINI (Shirt size 1):

$$= 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$$

$$= 0.5$$

GINI (Shirt size = 2):

$$= 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right]$$

$$= 0.489$$

GINI (Shirt size = 3)

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

$$= 0.48$$

$$\text{Weighted} = 0.5\left(\frac{4}{20}\right) + 0.5\left(\frac{4}{20}\right) + 0.489\left(\frac{7}{20}\right) + 0.49\left(\frac{5}{20}\right)$$

$$= 0.4913$$

From ②, ③, ④ :

GINI : 0.48 , 0.1625 , 0.4913

Least = 0.1625 = Car Type

Hence root node.