Prem A. Jethwa

1001861810

Assignment-4-

* Task 1:

Part a:

$$P(Color = \neg Green \mid Vehicle = Truck)$$

$$= P(Color = Red \mid Vehicle = Truck) + P(Color = Blue \mid Veh = Truck)$$

$$= \frac{P(Color = Red \wedge Vehicle = Truck)}{P(Vehicle = Truck)} + \frac{P(Color = Blue \wedge Vehicle = Truck)}{P(Vehicle = Truck)}$$

$$= \frac{0.1554}{0.1554 + 0.1680 + 0.0966} + \frac{0.0966}{0.1554 + 0.1680 + 0.0966}$$

$$= 0.60$$

Part b:

Proove 'Vehicle' & 'Color' are independent of each other.

To do so, we need to prooke,

$$P(Vehicle \mid Color) = P(Vehicle)$$

or

$$P(Color \mid Vehicle) = P(Color)$$

For Color = ¬Green and Vehicle = Car:

$P(Color = \neg Green \mid Vehicle = TruCK = 0.60$ — (Part a)

$P(Color = \neg Green) = 1 - P(Color = Green)$

$= 1 - (0.1280 + 0.0480 + 0.1680 + 0.0560)$

$= 0.60$

We can see that,

$P(Color = \neg Green \mid Vehicle = TruCK) = P(Color = Green)$

As both the values are same, ∴ variable color & vehicle are independent of each other.

* Task 2:

Part a:

Given:

A → 7 values & B_n → 8 values each where n = 1 to 10.

∴ No. of values to be stored for Joint P.D. (theoritically)

$= 7 \times 8^{10} = 7.51 \times 10^{9}$ (approx)

And practically = $\underline{(7 \times 8^{10}) - 1}$

Part b:

It is given that B_i is conditionally independent of B_j (where $j \neq i$) given A.

$\therefore P(B_1, B_2, B_3, \ldots, B_{10} | A) = \prod_{j=1}^{10} P(B_j | A)$ — ①

∴ Using product rule we can say that,

$P(A, B_1, B_2, B_3, \ldots B_{10}) = P(B_1, B_2, B_3, \ldots B_{10} | A) \cdot P(A)$

$= \left(\prod_{j=1}^{10} P(B_j | A) \right) \cdot P(A)$

— from ①

$$\left(\prod_{j=1}^{10} P(B_j|A)\right) \cdot P(A)$$

Each B_i variable can take 8 values & A can take 7 values.

∴ No. of values to store for each $P(B_j|A) = 8 \times 7$

And as there are 10 B_i variables.

∴ Total no. of values to be stored for $P(B_j|A)$

$= 8 \times 7 \times 10$ — ②

$= 7 \times 7 \times 10$ — ④

A variable can take 7 values

∴ No. of values to store for $P(A) = 7$ — ③

$= 6$ — ⑤

∴ Total no. of values to be stored $= (8 \times 7 \times 10) + 7$ — from ② & ③

$= \underline{570}$ (theoritically).

And practically $= (7 \times 7 \times 10) + 6$ — from ④ & ⑤

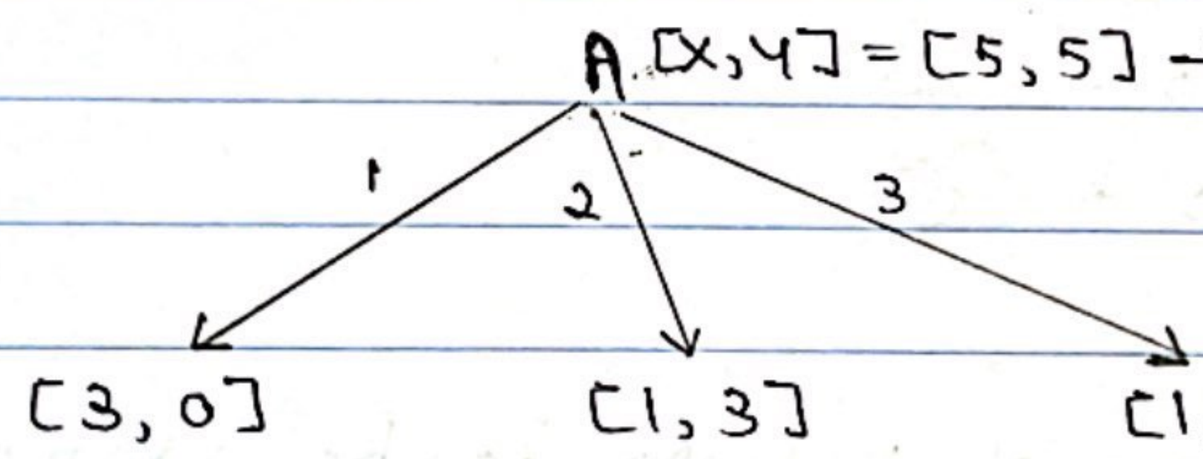
$= \underline{496}$ (Practically).

Part c:

Here we know that the variables $B_1, B_2, \ldots B_{10}$ are independent of each other given A. In a way, we can say that the effects $(B_1, B_2, \ldots B_{10})$ are independent of each other given cause (A). Thus we can say that this scenario follows the Naive Bayes Model.

* Task 4:

Attribute A:



$H_1 =$
$-(3/3)\log_2(3/3)$
$-(0/3)\log_2(0/3)$
$= 0$

$H_2 =$
$-(1/4)\log_2(1/4)$
$-(3/4)\log_2(3/4)$
$= 0.81125$

$H_3 =$
$-(1/3)\log_2(1/3)$
$-(2/3)\log_2(2/3)$
$= 0.918.$

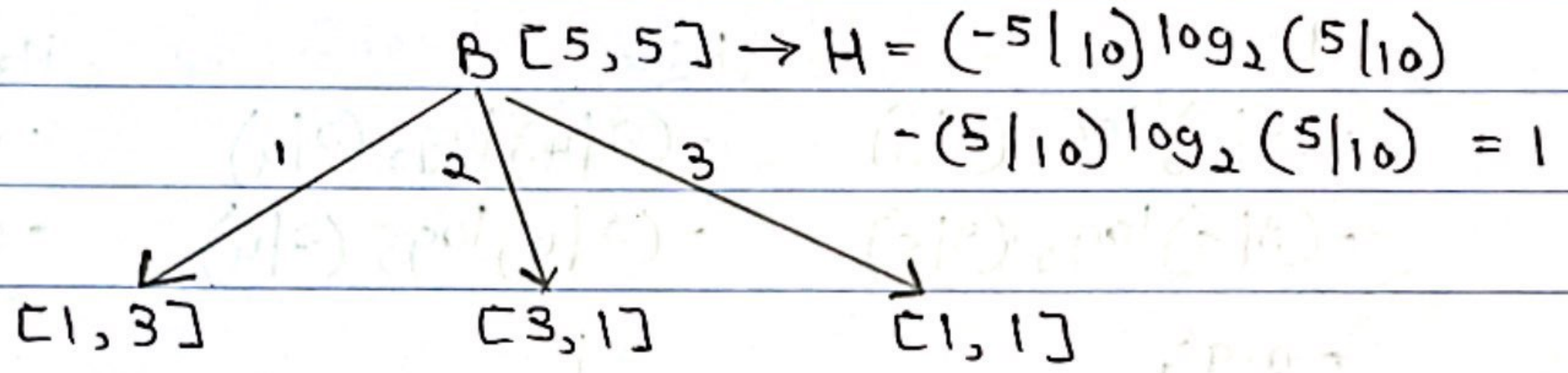
Information gain, for attribute A

$$= H - \sum_{j=1}^{L} (Ri/R) \cdot Hj$$

$$= 1 - \left[\left(\frac{3}{10}\right)(0) + \left(\frac{4}{10}\right)(0.8112) + \left(\frac{3}{10}\right)(0.918) \right]$$

$$= 0.400112 \approx 0.4$$

Attribute B:



$H_1 =$

$-(1/4)\log_2(1/4)$

$-(3/4)\log_2(3/4)$

$= 0.8112$

$H_2 =$

$-(3/4)\log_2(3/4)$

$-(1/4)\log_2(1/4)$

$= 0.8112$

$H_3 =$

$-(1/2)\log_2(1/2)$

$-(1/2)\log_2(1/2)$

$= 1$

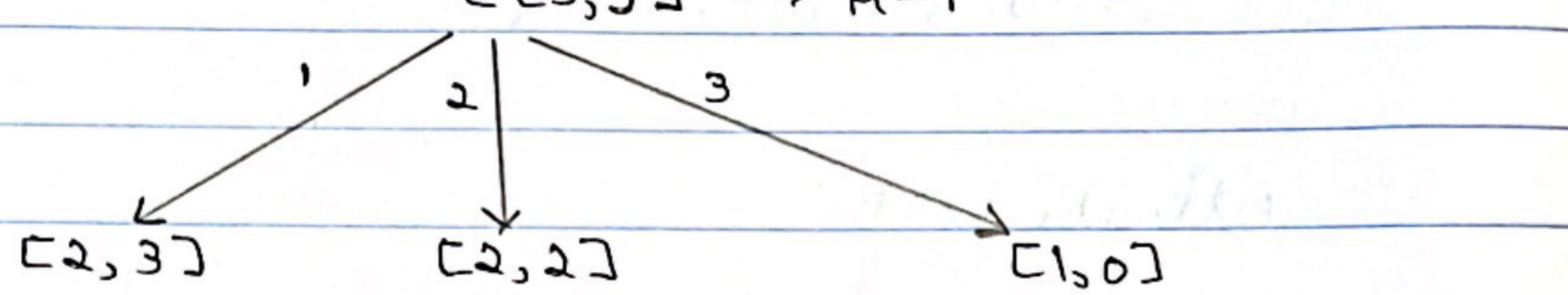
∴ Information gain for attribute B,

$$= H - \sum_{j=1}^{L} (Ki/K) \cdot Hi$$

$$= 1 - \left[\left(\frac{4}{10}\right) \cdot (0.8112) + \left(\frac{4}{10}\right) \cdot (0.8112) + \left(\frac{2}{10}\right)(1) \right]$$

$$= 0.15104$$

Attribute c:



H_1 =

$-(2/5)\log_2(2/5)$

$-(3/5)\log_2(3/5)$

$= 0.97$

H_2 =

$-(2/4)\log_2(2/4)$

$-(2/4)\log_2(2/4)$

$= 1$

H_3 =

$-(1/1)\log_2(1/1)$

$-(0/2)\log_2(0/2)$

$= 0$

∴ Information gain for attribute c

$$= 1 - \left[\left(\frac{5}{10}\right).(0.97) + \left(\frac{4}{10}\right).(1) + \left(\frac{1}{10}\right).(0)\right]$$

$$= 0.115$$

Attribute	Information gain
A	0.4
B	0.15
C	0.11

∴ A has maximum information gain, we can select it at root.

* Task 3:

From the given network.

P(not (Baseball game on TV) | not (George Feeds cat)

$= P\left(\frac{\neg bGTv}{\neg GFc}\right)$

$= \frac{P(\neg bGTv \wedge \neg GFc)}{P(\neg GFc)}$

Inference by Enumeration;

$\alpha < P(\neg bGTv \wedge \neg GFc)$

$P(\neg bGTv \wedge \neg GFc)$

$= P(\neg bGTv \wedge GWTv \wedge DCF \wedge \neg GFc) +$
$P(\neg bGTv \wedge GWTv \wedge \neg DCF \wedge \neg GFc) +$
$P(\neg bGTv \wedge \neg GWTv \wedge DCF \wedge \neg GFc) +$
$P(\neg bGTv \wedge GWTv \wedge \neg DCF \wedge \neg GFc)$

$= 0.695890411 \times 0.118110236 \times 0.169863014 \times 0.958333333$
$+ 0.695890411 \times 0.118110236 \times 0.830136986 \times 0.293577982$

+0.695890411 x 0.881889764 x 0.169863014 x
0.684210054
+0.695890411 x 0.881889764 x 0.830136986 x
0.412377113
= 0.125744303.

Similary, P(BGTV ∧ ¬GFC)

= 0.304109589 x 0.927927928 x 0.169863014 x
0.958333333
+ 0.304109589 x 0.927927928 x 0.830136906 x
0.293577982
+0.304109589 x 0.720720720 x 0.169863014 x
0.684210526
+0.304109589 x 0.720720720 x 0.830136906 x
0.041237113
= 0.118007272

tuple,

α < 0.125744303	0.118007272 >
= < 0.51587073	0.48412927 >

* Task 5 -

Entropy Before Split.

$$H = -\frac{5}{10}\log_2\left(\frac{5}{10}\right) - \frac{5}{10}\log_2\left(\frac{5}{10}\right)$$

$$= 1$$

Using A threshold 15,

$$X = 5, Y = 3$$

$$H_{A>15} = -\frac{5}{8}\log_2\left(\frac{5}{8}\right) - \frac{3}{8}\log_2\left(\frac{5}{8}\right)$$

$$= 0.9544$$

$$H_{A<15} \rightarrow X = 0, Y = 2$$

$$= -\frac{0}{2}\log\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right)$$

$$= 0$$

$$(\text{Infogain A}) = 1 - 0.95 + 4 \times \frac{8}{10} + 0 = 0.7635$$

$$= 1 - 0.7635$$

$$= 0.2365$$

Using B threshold 15,

$B \geq 15 \rightarrow X = 3, Y = 5$

$H_{B \geq 15} = -\frac{3}{8} \log_2 \left(\frac{3}{8}\right) - \frac{5}{8} \log_2 \left(\frac{5}{8}\right)$

$= 0.9544$

$B < 15 \rightarrow X = 2, Y = 0$

$H_{B < 15} = -\frac{2}{2} \log_2 \left(\frac{2}{2}\right) - \frac{0}{2} \log_2 \left(\frac{2}{2}\right)$

$= 0$

$(\text{Information Gain})_B = 1 - 0 + 0.9544 \times \frac{8}{10}$

$= 1 - 0.7635$

$= 0.2365$

Using C threshold 15,

$C \geq 15 \rightarrow X = 5, Y = 4$

$H_{C \geq 15} = -\frac{5}{9} \log_2 \left(\frac{5}{9}\right) - \frac{4}{9} \log_2 \left(\frac{4}{9}\right)$

$= 0.9910$

$C_{215} \rightarrow X=0 \;, Y=1$

$H_{C215} = -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right)$

$= 0$

(Information Gain) $C_{15} = 1 - 0 + 0.9910 \times \frac{9}{10} = 0.8919$

$= 1 - 0.8919$

$= 0.1081$

Using A threshold $= 20$;

$A \geq 20 \rightarrow X=5 \;, Y=2$

$H_{A \geq 20} = -\frac{5}{7}\log_2\left(\frac{5}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right)$

$= 0.8631$

$H_{A<20} \rightarrow X=0 \;, Y=3$

$H_{A<20} = \frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log_2\left(\frac{3}{3}\right)$

$= 0$

Using B threshold = 20,

$B \geq 20 \rightarrow X = 3, Y = 5$

$H_{B \geq 20} = -\frac{3}{8} \log_2 \left(\frac{3}{8}\right) - \frac{5}{8} \log_2 \left(\frac{5}{8}\right)$

$= 0.9544$

$B < 20 \rightarrow X = 2, Y = 0$

$H_{B < 20} = -\frac{2}{2} \log_2 \left(\frac{2}{2}\right) - \frac{0}{2} \log_2 \left(\frac{0}{2}\right)$

$= 0$

Info Gain $(B) 20 = 1 - 0.7635$

$= 0.2365$

Using C threshold = 20,

$C \geq 20 \Rightarrow X = 5, Y = 1$

$H_{C \geq 20} = -\frac{5}{6} \log_2 \left(\frac{5}{6}\right) - \frac{1}{6} \log \left(\frac{1}{6}\right)$

$= 0.6500.$

$C < 20 \Rightarrow X = 0, Y = 4$

$H_{C < 20} = -\frac{0}{4} \log_2 \left(\frac{0}{4}\right) - \frac{4}{4} \log_2 \left(\frac{4}{4}\right)$

$= 0$

[Info Gain] C_{20} = 0.4 − 0.6500 × $\left(\frac{6}{10}\right)$

= 1 − 0.3900 = 0.6100

Using A threshold = 25,

A ≥ 25 → X = 3, Y = 0

$H_{A \geq 25} = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)$

= 0

$H_{A<25}$ = X = 2, Y = 5

$= -\frac{2}{7}\log_2\left(\frac{2}{7}\right) - \frac{5}{7}\log_2\left(\frac{5}{7}\right)$

= 0.8631

[Info Gain] A_{25} = 0.8631 × $\frac{7}{10}$

= 1 − [0.6041] = 0.3959

Using B threshold = 25,

B > 25 → X = 0, Y = 5

$H_{B \geq 25} = -\frac{0}{5}\log_2\left(\frac{0}{5}\right) - \frac{5}{5}\log_2\left(\frac{5}{5}\right)$

= 0

$B \geq 25 \rightarrow X = 5 \;, Y = 0$

$H_{B \geq 25} = 0$

(Info gain $B)_{25} = 1 - 0 = 1$

Using c threshold = 25 ,

$C \geq 25 \rightarrow X = 3 \;, Y = 0$

$H_{C \geq 25} = -\frac{3}{3} \log_2 \left(\frac{3}{3}\right) - \log_2 \left(\frac{0}{3}\right)$

$= 0$

$C \leq 25 \rightarrow X = 2 \;, Y = 5$

$H_{C \leq 25} = -\frac{2}{7} \log \left(\frac{2}{7}\right) - \frac{5}{7} \log_2 \left(\frac{5}{7}\right)$

$= 0.8631$

(Info gain) $C_{25} = 1 - (0 + 0.8631 \times 7/10)$

$= 0.3959$

Highest information is obtained when using B with threshold 25.