Assignment - 5

ANGAD MANJUNATHA
1001718335

Task - 1

a. No of people decided no to wait = 20
No of people decided to wait = 80

$$\text{Entropy} = -(P * \log_2 P + (1-P) \log_2 (1-P))$$

Probability of deciding to wait $= \frac{80}{100} = 0.8$

Probability of deciding not to wait $= 1 - 0.8 = 0.2$

$$\text{Entropy } H(S) = -(0.8 \log_2 0.8 + 0.2 \log_2 0.2)$$
$$= -(-0.258 - 0.464)$$
$$\simeq 0.722$$

b. At node B

$K_1 = 20$, $K_2 = 15$

$K = 20 + 15$
$= 35$

$$H_B = -\left(\frac{20}{35} \log_2\left(\frac{20}{35}\right) + \frac{15}{35} \log_2\left(\frac{15}{35}\right)\right)$$

$= 0.9852$

At node C

$K_1 = 60$, $K_2 = 5$

$K = 60 + 5 = 65$

$$H_C = -\left[\left(\frac{60}{65}\log_2\frac{60}{65}\right) + \left(\frac{5}{65}\log_2\frac{5}{65}\right)\right]$$

$$= 0.3913$$

I.G at node $A = 0.722 - \frac{35}{100} \times 0.9852 - \frac{65}{100} \times 0.3913$

$$= 0.1228$$

c. Let K examples be at node E with Entropy H_E

All of these example will end up at node H_I since weekend = yes, for all of them.

So entropy at node $H = H_H = H_E$

So, I.G.

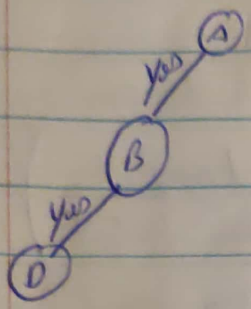
$$I_E = H_E - \left[\frac{K}{K}(H_H) + \frac{0}{K}(H_i)\right]$$

$$= H_E - H_H$$

$$= H_E - H_E$$

$$= 0$$

d. Since hungry patron came of weekend can would end up at node D.



Task-2

We have to determine certain pattern is type x or type x before split, S_x and S_y.

$$H = \frac{-5}{10}\log_2\frac{5}{10} \quad \frac{-5}{10}\log_2\frac{5}{10}$$

$$= 1$$

Use A to split

for $A = 1 \Rightarrow 3x, 0y$

$$H_{A=1} = \frac{-3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$$

$$= 0$$

$H_{A=2} = 1x, 3y$

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

$$= 0.8112$$

$H_A = 3 \Rightarrow 1x, 2y$

$= \frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$= 0.9182$

I.G using A.

$I_A = 1 - \left(\frac{3}{10}(0) + \frac{4}{10}(0.8112) + \frac{3}{10}(0.9182)\right)$

$= 0.4$

Using B to split

for $B=1 \Rightarrow 1x, 3y$

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

$= 0.8112$

for $B=2 \Rightarrow 3x, 1y$

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

$= 0.112$

for $B=3 \Rightarrow 1x, 1y$

$= \frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

$= 1$

$I_B = 1 - \left(\frac{4}{10}(0.8112) - \frac{4}{10}(0.8112) - \frac{2}{10}(1)\right)$

$= 0.15092$

Using C to split

for C=1 ⇒ 1u, 4y

$$= \frac{-1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$$

$$= 0.7921$$

for C=2 ⇒ 3u, 1y

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

$$= 0.8112$$

for C=3 ⇒ 1u, 0y

$$= \frac{-1}{1}\log\frac{1}{1} - \frac{0}{01}\log_2\frac{0}{1}$$

$$= 0$$

$$I_C = 1 - \frac{5}{10}(0.7219) - \frac{4}{5}(0.8112) - \frac{1}{10}(0)$$

$$= 0.3145$$

Therfore A is the bet option to use for root of decision Tree.

Task - 3

Let's consider max and min threshold in the D.T so, the range of A are

class	Range	A
x	22-25	3
x	26-29	2
y	12-14	2
y	15-18	1
y	19-21	2

Before split : 2x, 3y

$$\text{Entropy} = \frac{-3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5}$$

$$H = 0.970$$

for attribute A

when $A=1$ $x=0$ $y=1$

$A=2$ $x=1$ $y=2$

$A=3$ $x=1$ $y=0$

$$H_{A_1} = \frac{-0}{1} \log_2 \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1} = 0$$

$$H_{A_2} = \frac{-1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.9182$$

$$H_{A_3} = \frac{-1}{1} \log_2 \frac{1}{1} - \frac{0}{1} \log_2 \frac{0}{1} = 0$$

$$I_A = 0.970 - \frac{1}{5}(0) - \frac{3}{5}(0.9182) - \frac{1}{5}(0)$$

$$= 0.4199$$

Range of B

class	Range	B
x	13-17	1
x	18-22	2
y	23-24	1
y	25-29	2
y	30-32	3

when B = 1 x = 1 y = 0

B = 2 x = 2 y = 1

B = 3 x = 0 y = 1

$$H_{B1} = \frac{-0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1} = 0$$

$$H_{B2} = \frac{-2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.9182$$

$$H_{B3} = \frac{-0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1} = 0$$

$$I_B = 0.9708 - \frac{3}{15}(0.9182)$$

$$= 0.4199$$

Range for C

class	Range	C
u	21-24	2
u	25-28	1
u	29-31	2
y	13-16	2
y	17-20	3

when C=1 u=1 y=0

C=2 u=2 y=1

C=3 u=0 y=1

$H_{C_1} = \frac{-1}{1} \log\frac{1}{1} - \frac{0}{1} \log\frac{0}{1}$

$= 0$

$H_{C_2} = \frac{-2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log\frac{1}{3} = 0.9182$

$H_{C_3} = \frac{-0}{1} \log_2 \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1}$

$I_C = 0.9708 - \frac{3}{15}(0.9182)$

$= 0.4199$

$\therefore I_A = I_B = I_C$

All I.G are equal.

Task-4

Number of training samples = 1000

Number of possible class labels = 4

a. Highest entropy value = $\log_2 N$

$= \log_2 4 = 2$

Lowest Entropy value = 0

b. Let Entropy of N be H_N

Let Entropy of attribute K is split it into i subset

then $I_K = H_N - \left[\sum_{i=1}^{J} \frac{c_i}{C} H \right]$

c_i = No of in Subset C - Total Number N

where H_i is Entropy of each subsets

Lowest when all $H_i = A_N$

lowest = 0

Highest when all $H_i = 0$

Highest = H_N (Entropy at node N).

Task - 5

$B(attr1, attr2) = Avg\ Max\ P(c|attr1, attr2)$

$$P(c(attr1, attr2) = \frac{P(attr1, attr2)}{C}\ P(1)$$

$$= \sum_{c} \frac{P(attr1, attr2)}{C}\ P(1)$$

Naive bayes assumption

$$\frac{P(attr1, attr2)}{C} = \frac{P(attr1)}{C} \cdot \frac{P(attr2)}{C}$$

for class A

$P(attr2|A)$

Fit a gaussian into $C(15, 25, A)$

$n = 3$

$\mu = 19$

$$\sigma = \sqrt{\frac{1}{2}(15-19)^2 + (17-19)^2 + (25-19)^2}$$

$$= 5.291$$

$$P(attr1|A) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(A-\mu A)}{2\sigma A^2}}$$

$$= \frac{1}{5.291\sqrt{2\pi}} * \left[e^{\frac{-(15-19)^2}{2\times(5.291)^2}} + e^{\frac{-(17-19)^2}{2\times(5.19)^2}} + e^{\frac{-(25-19)^2}{2\times(5.291)^2}} \right]$$

$$= 0.0754\,[0.7514 + 0.1310 + 0.5256$$

$$= 0.1664$$

$$P(attr_2, CA) = \frac{1}{5.131\sqrt{2\pi}}\left(e^{\frac{-(18.83)}{2\times(5.13)^2}} + e^{-\frac{32.63}{2\times10.13^2}} + e^{\frac{-1.186}{2\times3.13^2}}\right)$$

$$= 0.1718$$

$$\underset{C}{P(attr1, attr2)} = \underset{C}{P(attr1)} \times \underset{C}{P(attr2)}$$

$$= 0.1664 \times 0.1718$$

$$= 0.2858$$

for class B

fit a gaussian into 20, 32, 25

$$n = 3$$

$$\mu = 25.66$$

$$\sigma = 6.02$$

$$P(attr1, B) = 0.066[0.6427 + 0.5793 + 0.9940]$$

$$= 0.1465$$

$P(attr_2|B)$ = fit a gaussian into 10, 15, 15

$$n = 3$$

$$\mu = 13.33$$

$$\sigma = 2.886$$

$$P(attr_2|B) = \frac{1}{2.886(2.506)}(0.5139 + 2(0.8458)$$

$$= 0.30496$$

$P(\underset{C}{attr1, attr2}) = 0.1465 * 0.30496$

$= 0.446$

Now $P(C/attr1, attr2) = \dfrac{P(\underset{C}{attr1, attr2})\ P(C)}{\sum_C P(\underset{C}{attr1, attr2})\ P(C)}$

$\sum_C P(\underset{C}{attr1, attr2})\ P(C) = 0.02858(0.5) + 0.046(0.5)$

$= 0.03659$

for A $P(C/attr1, attr2) = \dfrac{0.02858(0.5)}{0.03659}$

$= 0.3903$

for B $P(C/attr1, attr2) = \dfrac{0.046 * 0.5}{0.03655}$

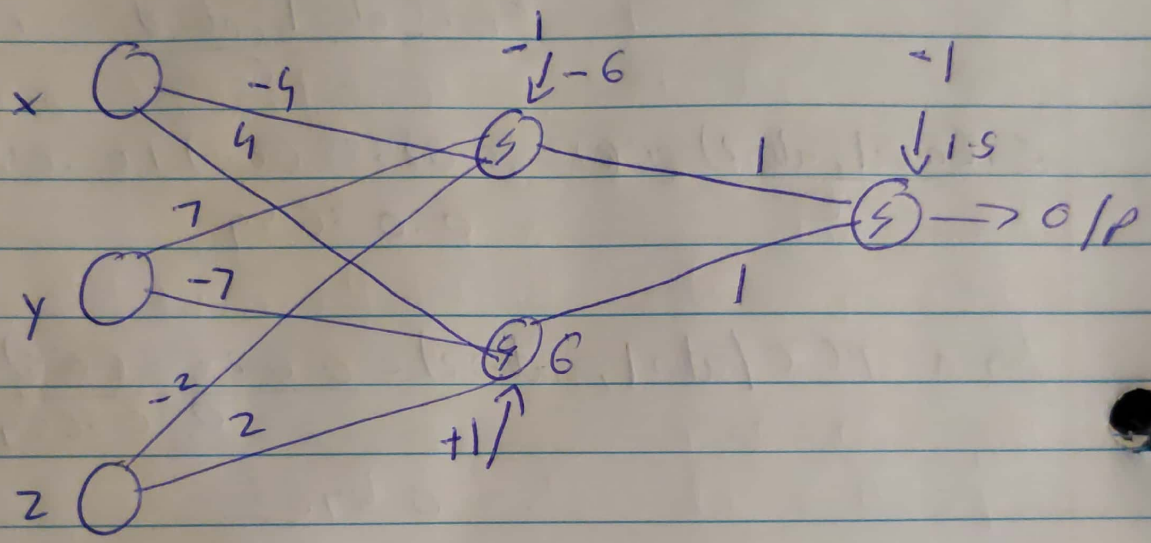
$= 0.6094$

Task-6

given $4x - 7y + 22 - 6 = 0$

we can write the above eq as below

$(-4x + 7y - 22 + 6 \geq 0) \wedge (4x - 7y + 22 - 6 \geq 0)$



Task-7

Since it is a binary classifier, there are only 2 possible outcome.

28% accuracy means it give wrong answer 72% of the time, Hence if we flip the answer of the classifier, we get a classifier with a guaranteed 72% accuracy.