(1) Architha Harinath
1001657433

1) a) Intial entropy $= H\left(\frac{80}{100}, \frac{20}{100}\right)$

$$= \frac{-80}{100}\log_2\left(\frac{80}{100}\right) - \frac{20}{100}\log_2\left(\frac{20}{100}\right)$$

$$= -0.8\log(0.8) - 0.2\log(0.2)$$

$$= -0.8\times(-0.322) - 0.2(-2.322)$$

$$= 0.2575 + 0.4643$$

$$H(A) = 0.7218$$

b) Info gain $= H(A) - \frac{35}{100}\times H\left(\frac{20}{35}, \frac{15}{35}\right) - \frac{65}{100}\times H\left(\frac{5}{65}, \frac{60}{65}\right)$

$$= 0.7218 + \left[\frac{35}{100}\times\left(\frac{-20}{35}\log_2\left(\frac{20}{35}\right) - \frac{15}{35}\log_2\left(\frac{15}{35}\right)\right)\right]$$

$$+ \left[\frac{-65}{100}\times\left(\frac{-5}{65}\log_2\left(\frac{5}{65}\right) - \frac{60}{65}\log_2\left(\frac{60}{65}\right)\right)\right]$$

$$= 0.7218 - (0.35\times(-0.571\times -0.80735) - (0.428\times -1.2239))$$

$$+ (-0.65(-0.923\times(-0.115) - 0.0769\times(-3.700)))$$

$$= 0.7218 - 0.344 - 0.2543$$

$$= 0.7218 - 0.5990$$

$$= 0.1229$$

c) Info gain is 0 because it is repeated & same so no changes are observed.

(2)

d) leaf node : F $\quad A \rightarrow C \rightarrow F$
ouput : will wait

e) leaf node : H $\quad A \rightarrow B \rightarrow E \rightarrow H$
output: will not wait

3)

a) Highest entropy if equally distributed

$$1000 \rightarrow 250(A), 250(B), 250(C), 250(D)$$

$$H \text{ entropy} = H(EA) + H(EB) + H(EC) + H(ED)$$

$$= 4\left(\frac{-250}{1000} \times \log_2\left(\frac{250}{1000}\right)\right)$$

$$= 4\left(-0.25 \times \log_2(0.25)\right)$$

$$= 4(-0.25 \times -2)$$

$$= 2$$

lowest entropy when all the values are in single class

$$\text{Entropy} = H(EA)$$

$$= \frac{-1000}{1000} \times \log_2(1)$$

$$= 0$$

3

b) Highest info gain is ~~2~~. when training data has only 2 classes and they are evenly split.

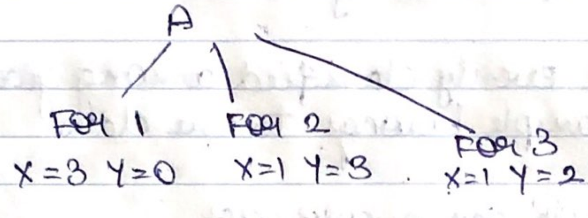
lowest possible info gain is 0 when data is evenly split among root as well as leaf nodes.

4) Classifier can be improved based on true cases in the data set. The data set should be vibrant and diverse

We cannot guarantee about achieving better than 60%. It completely depends upon data set and training data.

(4)

2) FOR A



$H(E) = 1$

$H(E1) = 0$

$H(E2) = [(-1/4) * \log_2(1/4)] - [(-3/4) * \log_2(3/4)]$

$= [-0.25 * 2] - [-0.75 * -0.41504]$

$= 0.5 + 0.31128$

$= 0.81128$

$H(E3) = [(-1/3) * \log_2(1/3)] - [(2/3) * (\log_2(2/3))]$

$= [-0.33 * -1.584] - [0.667 * -0.577]$

$= 0.52832 + 0.38518$

$= 0.9135$

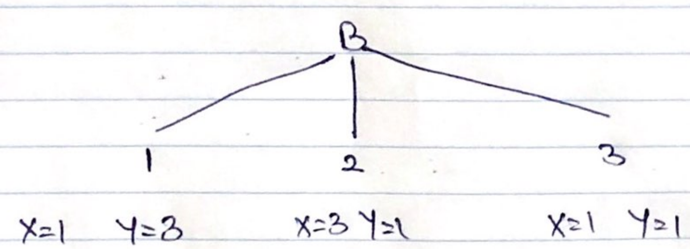
$H(E, A) = H(E) - [3/10 * H(E1)] - [4/10 * H(E2)] - [3/10 * H(E3)]$

$= 1 - [0.3 * 0] - [0.4 * 0.811] - [0.3 * 0.9135]$

$= 1 - 0.32451 - 0.27405$

$= 0.40144$

FOR B



(5)

$X=5 \quad Y=5$

$H(E)=1$

$H(E1)= [(-1/4) * \log_2(1/4)] - [3/4 * \log_2(3/4)]$

$= [-0.25 * -2] - [0.75 * -0.41504]$

$= 0.5 + 0.31128$

$= 0.81128$

$H(E2) = [(-3/4) * \log_2(3/4)] - [(1/4) * \log_2(1/4)]$

$= [-0.75 * -0.41504] - [0.25 * -2]$

$= 0.81128$

$H(E3) = 1$

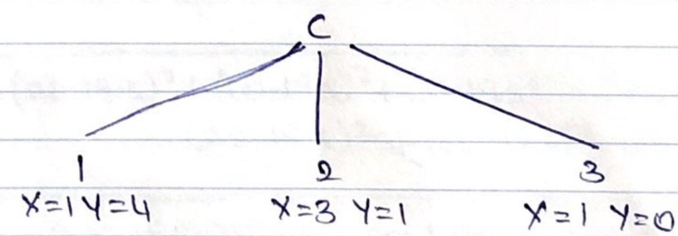
$I(E,B) = H(E) - [4/10 * (H(E1))] - [4/10 * H(E2)] - [2/10 * H(E3)]$

$= 1 - [0.4 * 0.81128] - [0.4 * 0.81128] - [0.2 * 1]$

$= 1 - 0.32451 - 0.32451 - 0.2$

$= 0.15098$

For C



$X=5 \quad Y=5$

$H(E)=1$

$H(E1) = [(-1/5) * \log_2(1/5)] + [(-4/5) * \log_2(4/5)]$

$= [-0.2 * -2.3193] + [-0.8 * -0.3219]$

$= 0.46439 + 0.25754$

$= 0.72193$

(6)

$H(E2) = [-3/4 * \log(3/4)] + [-1/4 * \log(1/4)]$
$= [-0.75 * -0.41504] - [0.25 * -2]$
$= 0.31128 + 0.5$
$= 0.81128$

$H(E3) = 0$

$I(E, C) = H(E) - [5/10 * H(E1)] - [4/10 * H(E2)] - [1/10 * H(E3)]$
$= 1 - [0.5 * 0.72193] - [0.4 * 0.81128] - [0.1 * 0]$
$= 0.31453$

A Has highest information gain.

(7)

5) $\mu A1 \rightarrow$ Mean of attribute1 belonging to class A
$\mu A2$ = Mean of attri 2 belonging to class A
$\mu B1 \rightarrow$ Mean of attribute 1 belonging to class B
$\mu B2 \Rightarrow$ Mean of attribute 2 belonging to class B,

$\mu A1 = 15+18/2 = 16.5$
$\mu A2 = 28+32/2 = 30$

$\mu B1 = 20+32+25/3 = 25.66$
$\mu B2 = 10+15+15/3 = 13.33$

$\sigma A1, \sigma A2, \sigma B1, \sigma B2$ be SD

$\sigma A1 = \sqrt{1(15-16.5)^2 + (18-16.5)^2} = \sqrt{2.25+6.25} = \sqrt{8.25} = 2.91$

$\sigma A2 = \sqrt{1(28-30)^2 + (32-30)^2} = \sqrt{4+4} = \sqrt{8} = 2.82$

$\sigma B1 = \sqrt{\frac{1}{2}(25.66-20)^2 + (32-25.66)^2 + (25-25.66)^2}$

$= \sqrt{\frac{1}{2}(5.66)^2 + (6.34)^2 + (0.66)^2}$

$= \sqrt{\frac{1}{2}(72.665)}$

$= 6.027$

$\sigma B2 = \sqrt{\frac{1}{2}(10-13.3)^2 + (15-13.3)^2 + (15-13.3)^2}$

$= \sqrt{\frac{1}{2}(3.3)^2 + (2.3)^2 + (2.3)^2}$

$= \sqrt{\frac{1}{2}(10.89+5.29+5.29)}$

8

$$\sigma_{B2} = \sqrt{\frac{21.47}{2}}$$

$$= \sqrt{10.735} = 3.276.$$

$$P(X=x_1|C=A) = \frac{1}{2.91\sqrt{2\pi}} e^{-\frac{(x-16.5)^2}{2\times 8.5}}$$

$$P(X=x_2|C=A) = \frac{1}{\cancel{2.88}\,2.82\sqrt{2\pi}} e^{-\frac{(x-30)^2}{2\times 8}}$$

$$P(X=x_1|C=B) = \frac{1}{6.027\sqrt{2\pi}} e^{-\frac{(x-25.66)^2}{2\times 36.3325}}$$

$$P(X=x_2|C=B) = \frac{1}{3.276\sqrt{2\pi}} e^{-\frac{(x-13.3)^2}{2\times 10.735}}$$

(8)

b) $3x-4y=15 = (3x-15 >= 4y)$ AND $(3x-15 <= 4y)$

$3x-15 >= 4y = 3x-4y-15 >= 0 \Rightarrow N1$

$3x-15 <= 4y = 3x-4y-15 <= 0 \Rightarrow N2$