

Assignment - 10

Task 1

- (a) 80 people decided to wait
20 people decided not to wait

Using entropy formula:

$$\begin{aligned} H(A) &= H\left(\frac{80}{100}, \frac{20}{100}\right) = -\frac{80}{100} \log_2\left(\frac{80}{100}\right) - \frac{20}{100} \log_2\left(\frac{20}{100}\right) \\ &= -0.8 \times \log_2(0.8) - (0.2) \times \log_2(0.2) \\ &= -0.8 \times (-0.32192) - (0.2) \times (-2.3219) \\ &= 0.2575424 + 0.46438 \\ &= 0.721928 \end{aligned}$$

- (b) Information Gain is given by

$$\begin{aligned} &= H(A) - \frac{35}{100} \times H\left(\frac{20}{35}, \frac{15}{35}\right) - \frac{65}{100} \left(H\left(\frac{5}{65}, \frac{60}{65}\right) \right) \\ &= 0.72198 - \frac{35}{100} \left(-\frac{20}{35} \log\left(\frac{20}{35}\right) - \frac{15}{35} \log\left(\frac{15}{35}\right) \right) \\ &\quad - \frac{65}{100} \times \left(-\frac{60}{65} \log\left(\frac{60}{65}\right) - \frac{5}{65} \log\left(\frac{5}{65}\right) \right) \end{aligned}$$

$$= 0.721928 - (0.35 \times (-0.571 \times -0.80735) - 0.4285 \times -1.2239) \\ - 0.65 (-0.923 \times (-0.115477) - 0.0769 \times (-3.70087))$$

$$= 0.721928 - 0.3446 - 0.254332$$

$$= 0.721928 - 0.59900$$

$$= 0.1229 //$$

(c) The information gain at node E of using the weekend test ~~is~~ 0, since it is repeated test which is already used at node A.

(d) The test case hungry patron who came in on a rainy day i.e. on tuesday.

The path followed is node A - node C → node F.

The leaf node it ends up in is node F.

The decision tree output for that case is patron will wait.

(e) The test not hungry who came on sunny saturday:

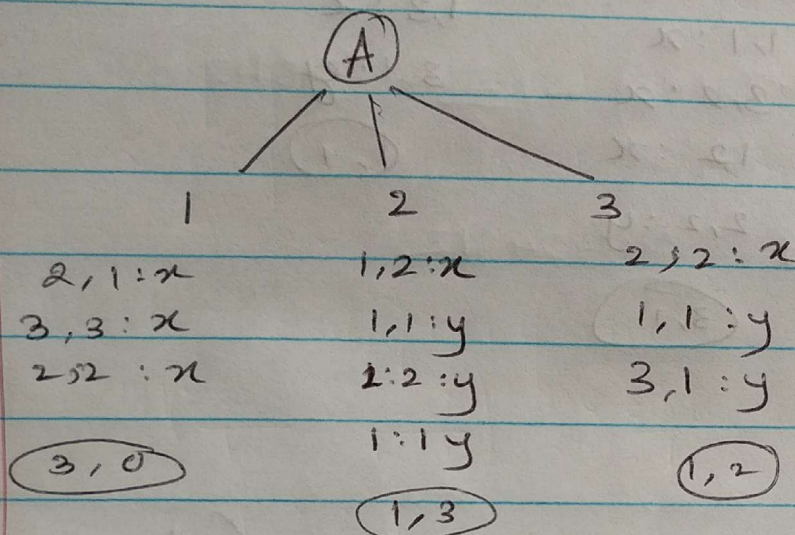
The path followed is node A - node B - node H

The leaf node it ends up in is node H

The decision table output for that case is will not wait

Task 2:

If we choose A as the root node:



$$H(E) = 1, \quad H(E_1) = 0, \quad H(E_2) = -\frac{1}{4} \log_2\left(\frac{1}{4}\right) - \frac{3}{4} \log_2\left(\frac{3}{4}\right)$$

$$= 0.5 + 0.31125 = 0.81125$$

$$H(E_3) = \left(-\frac{1}{3}\right) \log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right) \log_2\left(\frac{2}{3}\right)$$

$$= -0.333 (-1.585) - (0.6666) (-0.58510)$$

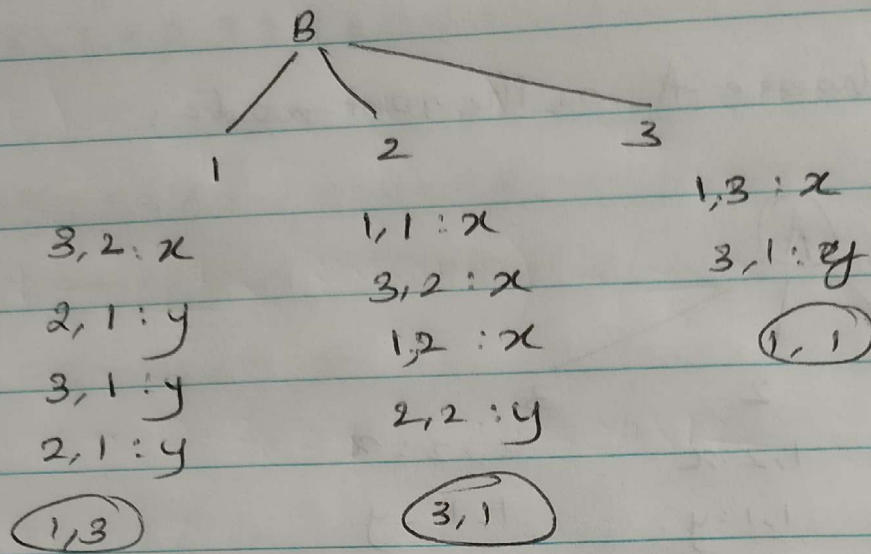
$$= 0.9183483$$

Information gain (A)

$$H(E) - \frac{4}{10} H(E_2) - \frac{3}{10} H(E_3) = 1 - 0.4 \times 0.81125 - 0.3 \times 0.918$$

$$= 1 - 0.324511 - 0.2755044 = 0.39998$$

If we choose B as the root node:



$$H(E) = 1$$

$$H(E_1) = \left(-\frac{1}{4}\right) \log_2\left(\frac{1}{4}\right) - \frac{3}{4} \log_2\left(\frac{3}{4}\right)$$

$$= 0.81125$$

$$H(E_2) = \left(-\frac{3}{4}\right) \log_2\left(\frac{3}{4}\right) - \frac{1}{4} \log_2\left(\frac{1}{4}\right)$$

$$= 0.81125$$

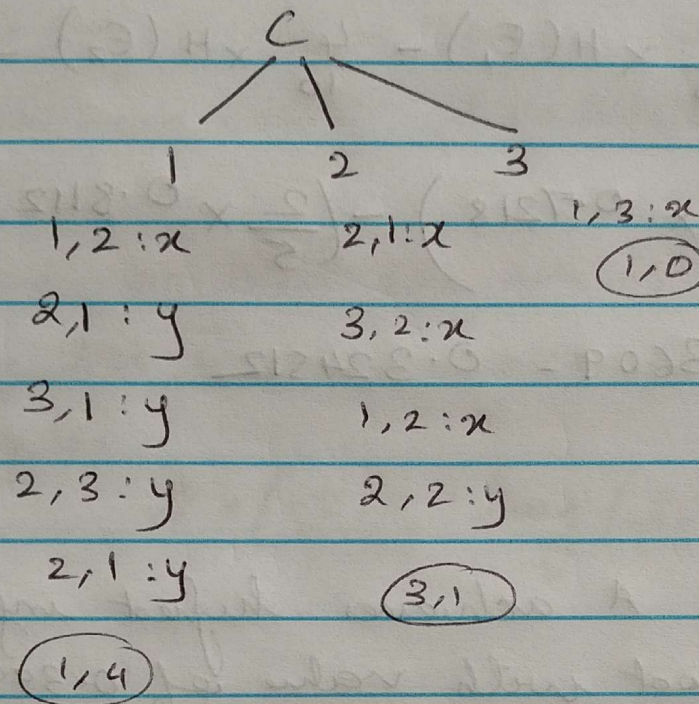
$$H(E_3) = 1$$

Info Gain for node B:

$$= 1 - \left(\frac{4}{10}\right) H(E_1) - \frac{4}{10} H(E_2) - \frac{2}{10} H(E_3)$$

$$= 1 - 0.32448 - 0.3248 - 0.2 = 0.1507$$

If we choose C as the root node.



$$H(E) = 1$$

$$H(E_1) = -\frac{1}{5} \log_2\left(\frac{1}{5}\right) - \frac{4}{5} \log_2\left(\frac{4}{5}\right)$$

$$= -(0.2) \log_2(0.2) - 0.8 \log_2(0.8)$$

$$= 0.4643 + 0.2575 = 0.7218$$

$$H(E_2) = \left(-\frac{3}{4}\right) \log_2\left(\frac{3}{4}\right) - \frac{1}{4} \log_2\left(\frac{1}{4}\right)$$

$$= 0.81128$$

$$H(E_3) = 0$$

Information gain for node C

$$H(E) - \frac{5}{10} \times H(E_1) - \frac{4}{10} \times H(E_2) - \frac{1}{10} H(E_3)$$

$$= 1 - \left(\frac{1}{2} \times 0.7218 \right) - \left(\frac{2}{5} \times 0.81128 \right)$$

$$= 1 - 0.3609 - 0.324512$$

$$= 0.3147$$

The attribute A achieves highest information gain at the root with value of 0.39998

Task 3:

The total number of distinct decision trees with n boolean attributes is equal to the number of distinct truth table with 2^n rows
 $= 2^{2^n}$

\therefore With 5 boolean attributes we have

$$n(\text{boolean } 2^{2^n} = 2^{2^5} = 2^{32} = \cancel{33554432})$$

$$= 4294967296$$

Task 4

- (a) Increase of 2 classes highest entropy is 1 when examples are evenly distributed.

So now when the examples are evenly distributed among the 4 classes, each class will have 250 examples.

$$\therefore 4 \times \left(-\frac{250}{1000} \log \left(\frac{250}{1000} \right) \right) = -4 \times \frac{1}{4} (-2) = 2$$

\therefore Highest entropy could be 2

lowest ~~could~~ entropy is when all the examples are distributed to one single class.

$$\therefore -\frac{1000}{1000} \log \left(\frac{1000}{1000} \right) = 0.$$

- (b) When the entropy ~~is~~ is 2 in the above case, where all the examples are equally distributed among all the classes then we have,

$$2 - \frac{1000}{1000} \left(-4 \times \frac{250}{250} \times \log \left(\frac{250}{250} \right) \right) = 2 - 0 = 2.$$

\therefore highest information gain is 2.

when, all examples belong to one class, then

$$H(E) = 0,$$

$$\therefore 0 = \frac{1000}{1000} \left(-\frac{1000}{1000} \log\left(\frac{1000}{1000}\right) \right) = 0.$$

\therefore Lowest information gain is 0.

Task 5 %

To improve the accuracy there is need of more training data set to be experimented.

So if accuracy is 28%, i.e. in order to increase accuracy to 60%. there is need of more data set - provided that data set are not false. We cannot guarantee better than 60% accuracy because it depends on the data set.