Hw 10

T1

(a)

Entropy at A before Test.

80 waited, 20 did not wait

So Entropy

$$H = \frac{-80}{100} \log_2 \frac{80}{100} - \frac{20}{100} \log_2 \frac{20}{100}$$

$$= -0.8 \times (-0.3219) - 0.2 \times -2.3$$

$$= 0.7219$$

(b) After split.

Subset 1 [is weekend]: 20 waited.
15 did not wait.

Subset 2 [not weekend]: 60 waited.
5 did not wait.

$$I = H - \frac{35}{100}\left[-\frac{20}{35}\log_2\frac{20}{35} - \frac{15}{35}\log_2\frac{15}{35}\right]$$

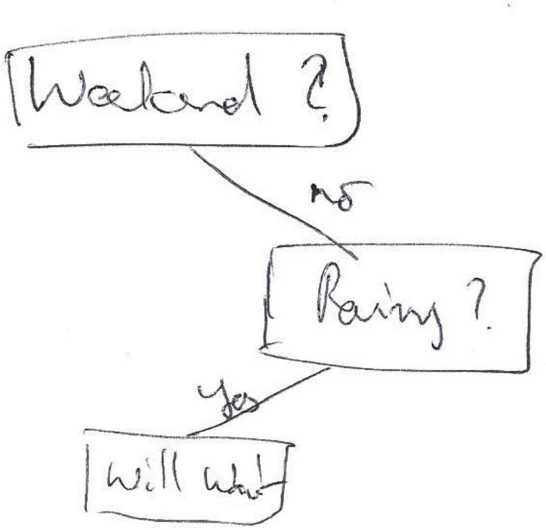
$$- \frac{65}{100}\left[-\frac{60}{65}\log_2\frac{60}{65} - \frac{5}{65}\log_2\frac{5}{65}\right].$$

$$= H - 0.35\,[0.9852] - 0.65\,[0.3912]$$

$$= 0.1228$$

(c) All the samples at E will have Weekend = YES. So using weekend as the test again will just result in information gain of 0

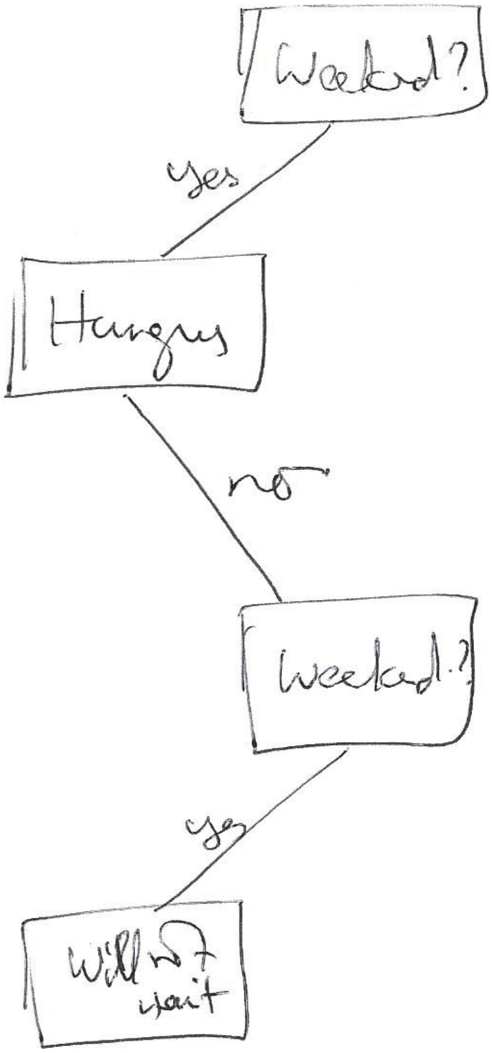
(d)



End up in node F

'Will wait'

(e)



Node H

'Will not wait'

T2

Initial entropy $= -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10}$

$= 1$

$A_1 = 1$

3 x , 0 y.

$$E_{A1=1} = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$$

$$= 0$$

$A_1 = 2$

1 x, 3 y.

$$E_{A1=2} = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$

$$= 0.8113$$

$A_1 = 3$

1 x, 2 y.

$$E_{A1=3} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

$$= 0.9183$$

Inf gain of using A_1.

$$I_{A_1} = 1 - \frac{3}{10}(0) - \frac{4}{10}(0.8113) - \frac{3}{10}(0.9183)$$

$$= 0.4.$$

Inf of $A_2 = 1$

1 X, 3 Y.

$$E_{A_2=1} = -\frac{1}{4}\log_2 \frac{1}{4} - \frac{3}{4}\log_2 \frac{3}{4} = 0.8113$$

$A_2 = 2$

3 X, 1 Y.

$$E_{A_2=2} = -\frac{3}{4}\log_2 \frac{3}{4} - \frac{1}{4}\log_2 \frac{1}{2} = 0.8113$$

$A_2 = 3$ 1 X, 1 Y

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

Inf gain of using A_2

$$I_{A_2} = 1 - \frac{4}{10}(0.8113) - \frac{4}{10}(0.8113) - \frac{2}{10}(1)$$

$$= 0.1510$$

$A_3 = 1$

1x, 4y.

$$E_{A_3=1} = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.7219$$

$A_3 = 2$ 3x, 1y.

$$E_{A_3=2} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$$

$A_3 = 3$ 1x, 0y.

$$E_{A_3=3} = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 1$$

Inf gain of A_3.

$$I_{A_3} = 1 - \frac{5}{10} \times 0.7219 - \frac{4}{10} 0.8113 - \frac{1}{10}(1)$$

$$= 0.3145$$

Highest Inf gain is for A_1

So we use A_1 as the best cut root node.

T3

(a) Highest Entropy possible $\log_2 4 = 2$

Lowest Entropy possible $= 0$

(b) Highest happens if Entropy before split is 2 and after split is 0

So 2. Lowest is 0.

T_4

Let $c(x)$ is the classifier given
Consider the following classifier.

```
f(x)
{
    if c(x) is 'will win'
        return 'will not win'
    if c(x) is 'will not win'
        return 'will win'
}.
```

Any pattern $c(x)$ would have classified incorrectly this will classify correctly any that $c(x)$ would classify correctly this will classify incorrectly

Error rate of $f(x)$ = 1 − Error rate of $c(x)$

$= 1 - 0.28$

$= 0.72$

$f(x)$ is 72% accurate which meets our requirement.

T_5

Patterns for A:

15	28	A
18	32	A

Patterns for B

20	10	B
32	15	B
25	15	B

$$\hat{\mu} = \frac{15+18}{2} = 16.5 \qquad \hat{\sigma} = \sqrt{\frac{(15-16.5)^2 + (8-16.5)^2}{1}} = 2.1213$$

$$P\left(\text{Attr 1} / \text{class} = A\right) = N_{A1}(16.5,\ 2.1213)$$

where N is normal pdf.

$$\hat{\mu} = \frac{28+32}{2} = 30 \qquad \hat{\sigma} = \sqrt{\frac{(28-30)^2 + (32-30)^2}{2}}$$

$$= 2.8284.$$

$$P\left(\text{Attr 2} / \text{class} = A\right) = N_{A2}(30,\ 2.8284)$$

$$\boxed{P\left(X / \text{class} = A\right) = N_{A1}(16.5,\ 2.1213) \times N_{A2}(30, 2.8284)}$$

$\hat{\mu} = 25.6667 \qquad \hat{\sigma} = 6.0277$

$$P\left(\text{Attr 1} / \text{class} = B\right) = N_{A1}(25.6667, 6.0277)$$

$\hat{\mu} = 13.3333 \qquad \hat{\sigma} = 2.8868$

$$P\left(\text{Attr 2} / \text{class} = B\right) = N_{A2}(13.3333, 2.8868)$$

$$\boxed{P\left(X / \text{class} = B\right) = N_{A1}(25.6667, 6.0277) \times N_{A2}(13.3333, 2.8868)}$$

$$\boxed{\begin{aligned} P(\text{Class} = A) &= 2/5 \\ P(\text{Class} = B) &= 3/5 \end{aligned}}$$

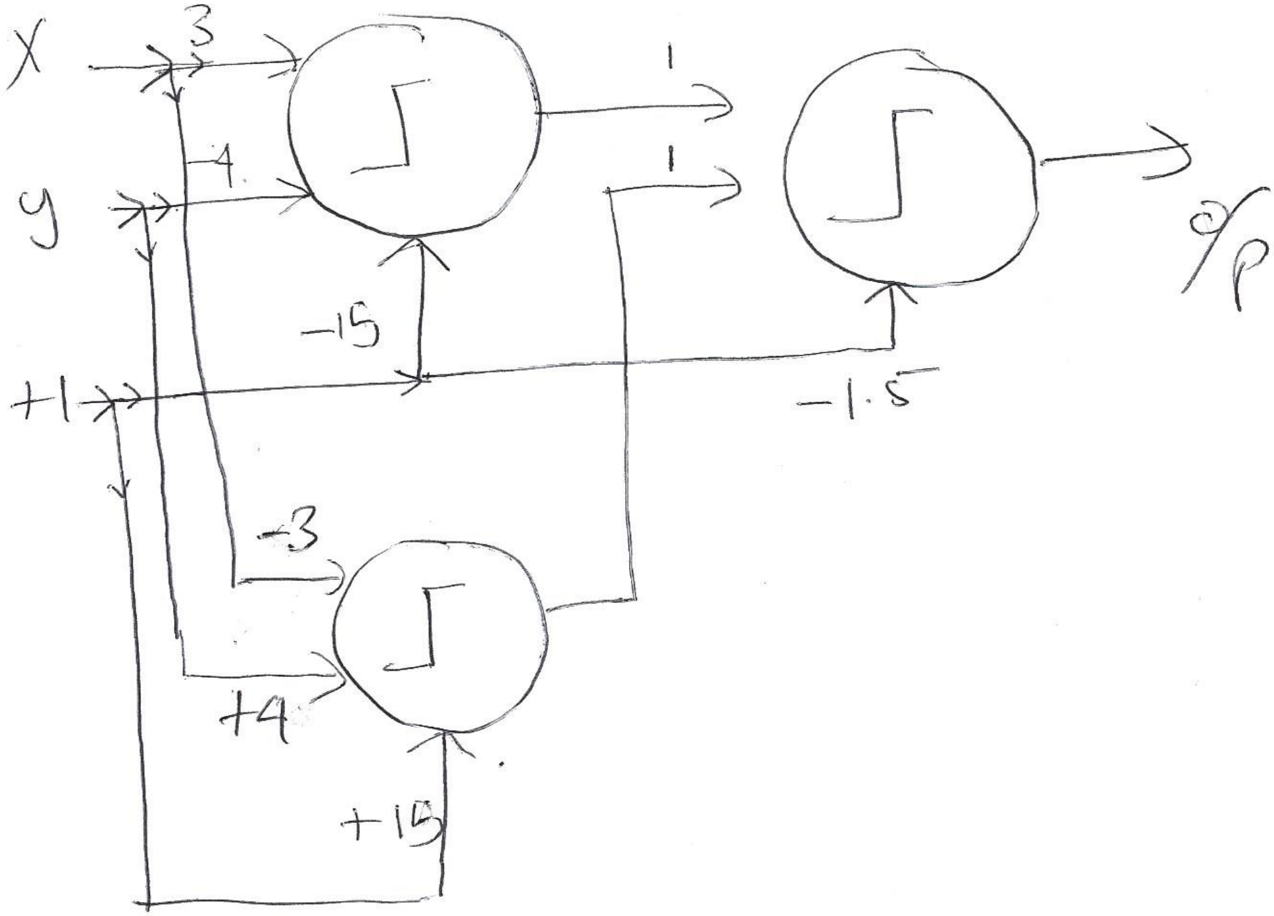
Task 6

This function is not linearly seperable. So we cannot use a single neuron for it.

We can write it as

$(3x - 4y \geq 15)$ AND $(3x - 4y \leq 15)$

Each of those is linearly seperable so we can use a network to represent it.



is the neuron.