Q1.

Attribute1: $H(Y) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$

According to $H(Y|A) = \sum p(x) H(Y|A=x)$, we get

$H(Y|A_1) = \frac{4}{5}(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}) + \frac{1}{5}(-1\log_2 1) = 0.8$

$H(Y|A_2) = \frac{3}{5}(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}) + \frac{2}{5}(-1\log_2 1) = 0.551$

$H(Y|A_3) = \frac{2}{5}(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}) + \frac{3}{5}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) = 0.951$

So $IG(Y|A_1) = H(Y) - H(Y|A_1) = 0.171$

$IG(Y|A_2) = H(Y) - H(Y|A_2) = 0.42$

$IG(Y|A_3) = H(Y) - H(Y|A_3) = 0.02$

So we should choose A_2 as the first split attribute

Attribute2:

1) $A_2 = 0$

Example	A_1	A_3	Output y
x_1	1	0	0
x_2	1	1	0

Since x_1 and x_2 have the same output values 0, then we return 0.

2) $A_2 = 1$

Example	A_1	A_3	Output y
x_3	0	0	0
x_4	1	1	1
x_5	1	0	1

As shown above, x_3. x_4. x_5 do not have the same output values.

$H(Y) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$

$H(Y|A_1) = \frac{1}{3}(-1\log_2 1) + \frac{2}{3}(-1\log_2 1) = 0$

$H(Y|A_3) = \frac{1}{3}(-1\log_2 1) + \frac{2}{3}(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}) = 0.667$

So $IG(Y|A_1) = H(Y) - H(Y|A_1) = 0.918$. $IG(Y|A_3) = H(Y) - H(Y|A_3) = 0.251$

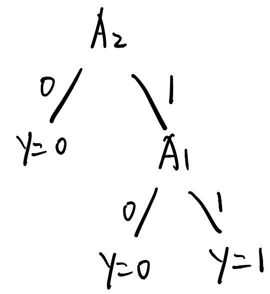
So we chose A_1 as the second split attribute for this subtree.

Attribute 3 :

1) $A_2=1$. $A_1=0$ Since X_3 is the only input and has output 0, then we return 0.

2) $A_2=1$. $A_1=1$ Since X_4 and X_5 have the same output values 1, then we return 1.

So the decision tree is:



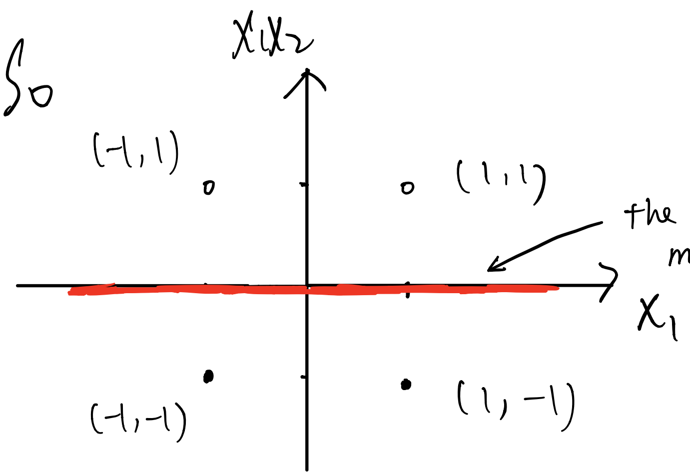
Q2.

(a)

It's easy to get,

$[X_1, X_2]$	X_1	X_1X_2	Xor output
[-1, -1]	-1	1	0
[-1, 1]	-1	-1	1
[1, -1]	1	-1	1
[1, 1]	1	1	0

So



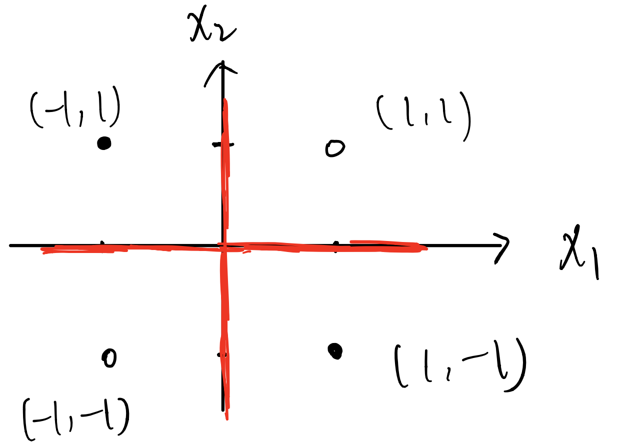
and we use " • " for output 1

" o " for output -1

the maximal margin seperator is: $X_1X_2=0$ which is emphasized by red.

The margin is 2 or 1 for only half of the distance.

(b)



From $x_1x_2=0$, we get $x_1=0$ or $x_2=0$

The sperating line is drawn by the red line.

Q3.

a. We use w^1, w^2 to represent the weights for the hidden layer and Output layer respectively.

For the one hidden layer:

the i-th input: $\sum_k W^1_{ik} X_k$

the i-th output $h_i = g(\sum_k W^1_{ik} X_k) = c\sum_k W^1_{ik} X_k + b$.

For the output layer:

the j-th input: $\sum_n W^2_{jn} \cdot h_n$

the j-th output $O_j = g(\sum_n W^2_{jn} \cdot h_n) = c\sum_n W^2_{jn}(c\sum_k W^1_{nk} X_k + b) + b$

$= c^2 \sum_n \sum_k W^2_{jn} W^1_{nk} X_k + b(c\sum_n W^2_{jn} + 1)$

So there is a network with no hidden units that computes the same function. The parameters are shown as below.

the j-th input: $\sum_n \sum_k W^2_{jn} W^1_{nk} X_k$

$C_{new} = c^2$. $b_{new} = b(c\sum_n W^2_{jn} + 1)$.

$$g(x)_{new} = c^2 x + b\left(c \prod_{n} W_{j}^{2} n + 1\right)$$

b. For a network with n hidden layers, we can use conclusion from part a. to turn it into a network with $(n-1)$ hidden layers. In this way, by induction, we can turn this n-hidden-layer network into a network with no hidden layers.

And the new parameters will be:

$$g_{new}(x) = C_{new}\, x + b_{new} \quad \text{where}$$

$$C_{new} = c^n. \quad b_{new} = b\left(\sum_{i=1}^{n-1} c^i \cdot \prod_{j=1}^{i} W^{n+1-j} + 1\right).$$

c.

For one hidden layer: $h \times n + n \times h = 2hn$

For no hidden layer: $n \times n = n^2$

Since $h \ll n$, the total number of weights becomes much larger, thus after transformation, the performance of the network becomes greatly worse.