10 marks

1. In this part you are asked to design a multi-layer perceptron that implements the binary decision region depicted in Fig. 1. The two regions shaded in grey should map to +1 and the rest to −1.

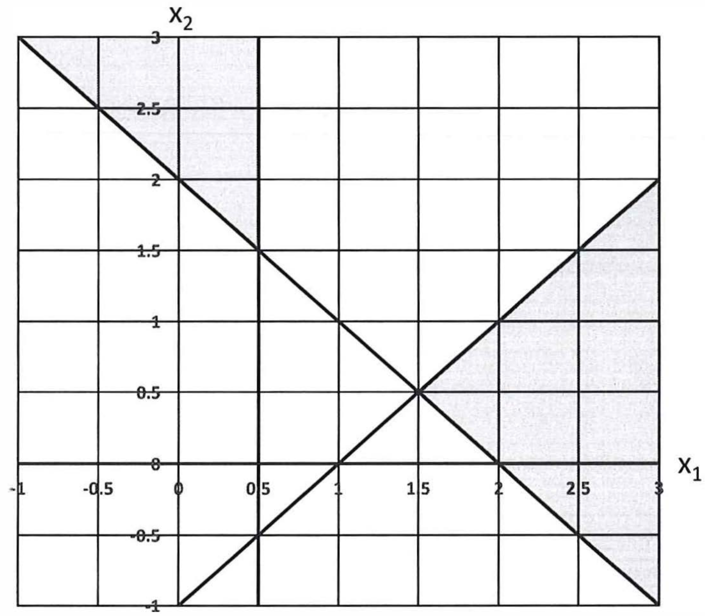


Figure 1: Decision region of a multi-layer perceptron.

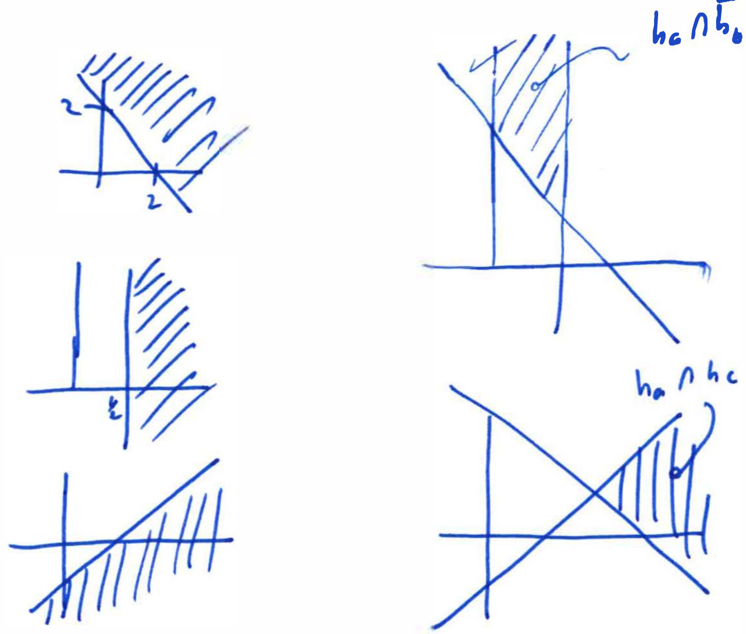
First, make a clear drawing of the architecture of your multilayer-perceptron, clearly indicating the weights associated with each edge. Second, list your edge weights in standard form in a weight matrix $W^{(l)}$ for each layer of weights. Finally, describe (clearly & concisely!) what is going on at each layer to describe the functioning of your architecture. (Additional space for your work and solution is provided on the next page.)

Implement the 3 lines

$h_a(\underline{x}) = \text{sign}(x_1 + x_2 - 2)$

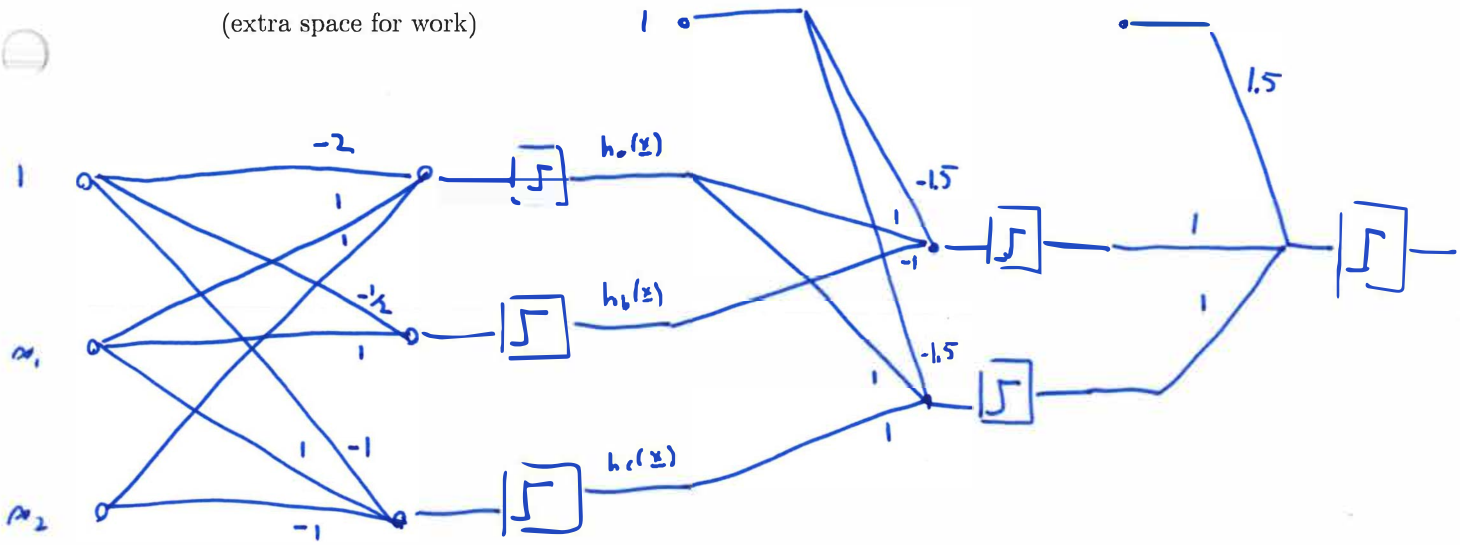
$h_b(\underline{x}) = \text{sign}(x_1 - \frac{1}{2})$

$h_c(\underline{x}) = \text{sign}(x_1 - x_2 - 1)$



total/–

(extra space for work)



$$W^{(1)} = \begin{bmatrix} -2 & -\frac{1}{2} & -1 \\ 1 & 1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \qquad W^{(2)} = \begin{bmatrix} -1.5 & -1.5 \\ 1 & 1 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \qquad W^{(3)} = \begin{bmatrix} 1.5 \\ 1 \\ 1 \end{bmatrix}$$

→ First layer implements h_a, h_b, h_c
Each column of $W^{(1)}$ corresponds to one hypothesis

→ 2^{nd} layer corresponds to two AND gates
$h_a \cap \bar{h}_b$, $h_a \cap h_c$ note $\bar{h}_b$ is negation (via weight -1)

→ 3^{rd} layer implements an OR gate

total/10

25 marks

2. In this problem we consider a neural network with 2 layers as shown in Fig. 2. Note that $w_{i,j}^{(l)}$ denotes the weight on the edge between node i in layer $l-1$ and node j in layer l. The input symbols are denoted by x_1 and x_2 and the output symbol is denoted by $x_1^{(2)}$. The symbol Σ denotes summation. For example, $s_1^{(1)} = w_{0,1}^{(1)} x_0 + w_{1,1}^{(1)} x_1 + w_{2,1}^{(1)} x_2$. The activation functions at the hidden layer are "rectified linear units" (ReLUs) and the activation function at the output layer is the "logistic function". Recall that the ReLU nonlinearity implements

$$\max\{0, s\}.$$

Recall that the logistic function is

$$\frac{e^s}{1+e^s}.$$

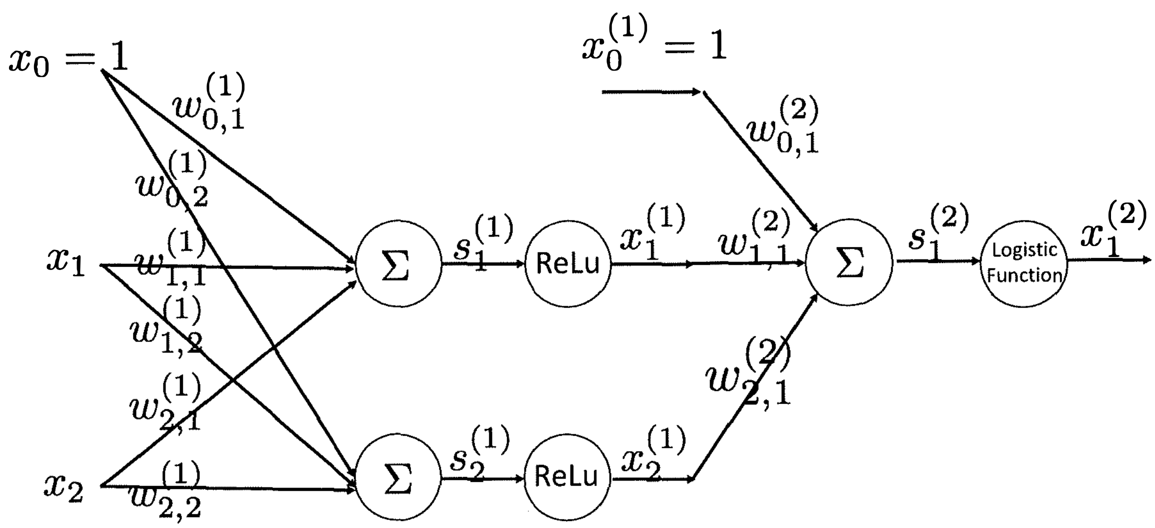


Figure 2: A neural network architecture that uses rectified linear (ReLu) activation functions at layer-1 and the logistic function at the output layer.

The weights are

$$\begin{aligned} w_{0,1}^{(1)} &= -1, & w_{1,1}^{(1)} &= -1, & w_{2,1}^{(1)} &= 1, \\ w_{0,2}^{(1)} &= 4, & w_{1,2}^{(1)} &= -2, & w_{2,2}^{(1)} &= -1, \\ w_{0,1}^{(2)} &= 1.5, & w_{1,1}^{(2)} &= 1, & w_{2,1}^{(2)} &= -1. \end{aligned}$$

Given a training example $\mathbf{x} = (x_1, x_2)$ with label y and network output $x_1^{(2)}$, let the loss function $e(x_1^{(2)}, y)$ be logistic loss. I.e.,

$$e(x_1^{(2)}, y) = -\mathbb{I}(y=+1)\log_e[x_1^{(2)}] - \mathbb{I}(y=-1)\log_e[1-x_1^{(2)}],$$

where $\mathbb{I}(\cdot)$ is the indicator function taking on value +1 if the argument is true and zero otherwise.

There are four parts to this problem – **(a)**, **(b)**, **(c)** and **(d)** – on the following pages.

total/–

10 marks

(a) Compute the initialization of the backpropogation for the above architecture. In other words, find the general form for $\delta_1^{(2)} = \frac{\partial}{\partial s_1^{(2)}} e(x_1^{(2)}, y)$ in terms of $s_1^{(2)}$, $x_1^{(2)}$ and y.

Method I: $\delta_1^{(2)} = \theta'(s) \frac{d}{dx} e(x, y)$

$$\theta'(s) = \frac{d}{ds} \frac{e^s}{1+e^s} = \frac{e^s}{1+e^s} - \frac{(e^s)^2}{(1+e^s)^2} = \frac{e^s}{(1+e^s)^2}$$

$$\frac{d}{dx} e(x,y) = -\mathbb{I}(y=1)\frac{1}{x} - \mathbb{I}(y=-1)\left(\frac{-1}{1-x}\right)$$

$$\boxed{\delta_1^{(2)} = \frac{e^{s_1^{(2)}}}{(1+e^{s_1^{(2)}})^2}\left[-\mathbb{I}(y=1)\frac{1}{x_1^{(2)}} + \mathbb{I}(y=-1)\frac{1}{1-x_1^{(2)}}\right]}$$

Method II: Get rid of dependence on x

$$\tilde{e}(s, y) = -\mathbb{I}(y=1)\log\left(\frac{e^s}{1+e^s}\right) - \mathbb{I}(y=-1)\log\left(\frac{1}{1+e^s}\right)$$

$$= -\log\left(\frac{e^{ys}}{1+e^{ys}}\right) = \log\left(\frac{1+e^{ys}}{e^{ys}}\right) = \log\left(1+e^{-ys}\right)$$

$$\frac{d}{ds}\tilde{e}(s,y) = \frac{1}{1+e^{-ys}} e^{-ys}(-y)$$

$$\boxed{\delta_1^{(2)} = \frac{-y\, e^{-y s_1^{(2)}}}{1+e^{-y s_1^{(2)}}}}$$

total/10

5 marks

(b) Verify your answer to part (a) by showing that for $(s_1^{(2)}, y) = (0.7, 1)$ the initialization of the backpropogation is $\delta_1^{(2)} = -1/3$. (To verify, approximate $e^{0.7}$ as 2 in your calculations.)

Method I:

$$\delta_1^{(2)} = \frac{e^{0.7}}{(1+e^{0.7})^2}\left[-\frac{1}{x_1^{(2)}} + 0\right]$$

$$= \frac{\cancel{e^{0.7}}}{(1+e^{0.7})^{\cancel{2}}}(-1)\frac{\cancel{(1+e^{0.7})}}{\cancel{e^{0.7}}} = \frac{-1}{1+e^{0.7}} \simeq -\frac{1}{1+2}$$

$$\boxed{\delta_1^{(2)} \cong \frac{-1}{3}}$$

Method II:

$$\delta_1^{(2)} = \frac{-1\, e^{-1\cdot 0.7}}{1+e^{-1\cdot 0.7}} = \frac{-1\, e^{-0.7}}{1+e^{-0.7}} \simeq \frac{-1\left(\frac{1}{2}\right)}{1+\frac{1}{2}} = \frac{-\frac{1}{2}}{\frac{3}{2}}$$

$$\boxed{\delta_1^{(2)} \simeq -\frac{1}{3}}$$

✓ match

total/5

5 marks

(c) For $(x_1, x_2) = (2, -0.8)$ and $y = 1$ solve for $\delta_1^{(2)}$ and $\frac{\partial}{\partial w_{1,1}^{(1)}} e(x_1^{(2)}, y)$. Explain your steps.

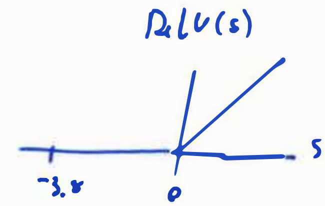
$$s_1^{(1)} = -1 - 1(2) + 1(-0.8) = -1 - 2 - 0.8 = -3.8$$

$$x_1^{(1)} = \max\{0, -3.8\} = 0$$

Now, note on back-prop

$$\delta_1^{(1)} = \theta'(s_1^{(1)}) \quad \delta_1^{(2)} \quad w_{11}^{(2)}$$

↑ this = 0 when $s_1^{(1)} = -3.8$



↑ slope = 0

And since

$$\frac{\partial e}{\partial w_{11}^{(1)}} = x_1 \; \delta_1^{(1)} = x_1 \cdot 0 = 0$$

∴

$$\boxed{\frac{\partial}{\partial w_{11}^{(1)}} e(x_1^{(2)}, y) = 0}$$

$$\boxed{\delta_1^{(2)} = -\frac{1}{3}}$$

← see logic next page

total/5

5 marks

(d) For $(x_1, x_2) = (2, -0.8)$ and $y = 1$ solve for $\delta_1^{(2)}$ and $\frac{\partial}{\partial w_{2,2}^{(1)}} e(x_1^{(2)}, y)$. Explain your steps.

$s_2^{(1)} = 4 - 2(2) - 1(-0.8) = 4 - 4 + 0.8 = 0.8$

$x_2^{(1)} = \text{ReLU}(0.8) = 0.8$

$s_1^{(2)} = 1.5 + \underbrace{1 \cdot 0}_{\text{from part (c)}} - 1(0.8) = 0.7 \sim (s_1^{(2)}, y) = (0.7, 1)$

from part (b) given

$\boxed{\delta_1^{(2)} = -1/3}$

$\delta_2^{(1)} = \theta'(s_2^{(1)}) \left[\delta_1^{(2)} w_{21}^{(2)} \right]$

$= \theta'(0.8) \left[-\frac{1}{3} \cdot (-1) \right]$

$= 1 \left(-\frac{1}{3}\right)(-1) = \frac{1}{3}$

$\frac{d e(x_1^{(2)}, y)}{d w_{22}^{(1)}} = x_2 \delta_2^{(1)} = (-0.8) \cdot \frac{1}{3} = -\frac{8}{10} \cdot \frac{1}{3} = -\frac{8}{30} = -\frac{4}{15}$

$$\boxed{\frac{d e(x_1^{(2)}, y)}{d w_{22}^{(1)}} = -\frac{4}{15}}$$

total/5