Assignment 5

Durbasmriti Saha
EE708: Fundamentals of Data Science and Machine Intelligence

April 6, 2025

1 a) Total weights between input layer and hidden layer = NxH
Total Weights between hidden layer and output layer = HxC
Hence total weights = NH + HC
b) Total weights between input layer and hidden layer = NxH
Total Weights between hidden layer and output layer = HxC
Total Weights between input layer and output layer = NxC
Hence total weights = NH + HC + NC

2) a) The advantage of network A over network B : Since it directly connects input layer to the output layer with connecting directly to 100 units of output layer, it represents a full-rank linear transformation.
b)The advantage of network B over the network A : This network has fewer parameters hence better generalization. Network A has 100x100 = 10,000 weights whereas Network B has 100x10 + 100x10 = 2,000 weights. Network A may overfit in case of limited data especially. But network B generalizes better hence it has less chance of overfitting.

3) Given,
inputs : x1 = -1, x2 = 1
weights : w1 = 0.1, w2 = 0.5
Activation function : sigmoid with slope parameter ($\alpha = 2$
$\phi(v) = \frac{1}{1+e^{-av}} = \frac{1}{1+e^{-2v}}$
Given $\phi(v1) = 0.73$
Now, v1 = w1x1 + w2x2 + b1 = (0.1)(-1) + (0.5)(1) +b1 = -0.1 + 0.5 + b1 = 0.4 + b1
$\phi(v) = \frac{1}{1+e^{-2(0.4+b1)}} = 0.73$
let z = 0.4+b1. Then ,
$1 + e^{-2z} = \frac{1}{0.73} \Rightarrow e^{-2z} = \frac{1}{0.73} - 1 = \frac{0.27}{0.73} \Rightarrow -2z = \ln\left(\frac{0.27}{0.73}\right)$
$\ln\left(\frac{0.73}{0.27}\right) = \ln(0.36986) \approx -0.995 \Rightarrow -2z = -0.995 \Rightarrow z = 0.4975$
Since z = 0.4 + b1,
$b_1 = z - 0.4 = 0.4975 - 0.4 = 0.0975$

4) a) NOT:

Input x	Output (NOT x)
0	1
1	0

Means,

$$\text{Output} = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

Take w = -2, b = 1
So, when x=0: -2x0 + 1 = 1 $>$ 0 , hence output = 1
when x=1: -2x1 + 1 = $-1 < 0$, hence output = 0

b) NAND :

x_1	x_2	NAND Output
0	0	1
0	1	1
1	0	1
1	1	0

$$\text{Output} = \begin{cases} 1 & \text{if } w_1x_1 + w_2x_2 + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

Take w1 = -2, w2 = -2, b = 3
For various inputs:
$x_1 = 0, \quad x_2 = 0: \quad -2(0) + (-2)(0) + 3 = 3 > 0 \Rightarrow \text{Output} = 1$
$x_1 = 0, \quad x_2 = 1: \quad -2(0) + (-2)(1) + 3 = 1 > 0 \Rightarrow \text{Output} = 1$
$x_1 = 1, \quad x_2 = 0: \quad -2(1) + (-2)(0) + 3 = 1 > 0 \Rightarrow \text{Output} = 1$
$x_1 = 1, \quad x_2 = 1: \quad -2(1) + (-2)(1) + 3 = -1 < 0 \Rightarrow \text{Output} = 0$
Hence,

$$\text{Output} = \begin{cases} 1 & \text{if } -2x_1 - 2x_2 + 3 > 0 \\ 0 & \text{otherwise} \end{cases}$$

5) The parity function over 3 binary inputs outputs:

$$\begin{cases} 1 & \text{if the number of 1s is odd (odd parity)} \\ 0 & \text{if the number of 1s is even (even parity)} \end{cases}$$

x_1	x_2	x_3	Parity (odd number of 1s)
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

Now, $\text{Parity}(x_1, x_2, x_3) = (x_1 \oplus x_2) \oplus x_3$
So, two XORs are needed to be chained together

For 3-input parity:
Input layers:
(x1,x2,x3)
Hidden layers:
1. First layer: Compute $(x_1 \oplus x_2)$
2. Second layer: compute $(x_1 \oplus x_2) \oplus x_3$

Output layer :
Parity = 1 or 0

6)
Input layer (size N): $x \in \mathbb{R}^N$
Hidden layer (size H):
Weights: $W^{(1)} \in \mathbb{R}^{H \times N}$
Bias: $b^{(1)} \in \mathbb{R}^H$
Pre-activation: $z^{(1)} = W^{(1)}x + b^{(1)}$
Activation: $h = \text{ReLU}(z^{(1)}) \in \mathbb{R}^H$
Output layer (1 unit):
Weights: $w^{(2)} \in \mathbb{R}^H$
Bias: $b^{(2)} \in \mathbb{R}$
Output: $\hat{y} = w^{(2)} \cdot h + b^{(2)}$
True label: y

For regression, MSE(Mean Squared error) is loss function
$L = \frac{1}{2}(\hat{y} - y)^2$

Gradients: $\frac{\partial L}{\partial \hat{y}} = \hat{y} - y$
$\frac{\partial \hat{y}}{\partial w^{(2)}} = h \quad \Rightarrow \quad \frac{\partial L}{\partial w^{(2)}} = (\hat{y} - y)h$
$\frac{\partial L}{\partial b^{(2)}} = \hat{y} - y$

Hidden layer Gradient:
$\frac{\partial L}{\partial h} = (\hat{y} - y)w^{(2)}$
ReLU derivative: $\frac{\partial h}{\partial z^{(1)}} = \begin{cases} 1 & \text{if } z_i^{(1)} > 0 \\ 0 & \text{otherwise} \end{cases} = \mathbb{I}_{z^{(1)}>0}$
$\Rightarrow \frac{\partial L}{\partial z^{(1)}} = \left(\frac{\partial L}{\partial h}\right) \circ \mathbb{I}_{z^{(1)}>0} = (\hat{y} - y)w^{(2)} \circ \mathbb{I}_{z^{(1)}>0}$

Now,
$\frac{\partial L}{\partial W^{(1)}} = \left[\frac{\partial L}{\partial z^{(1)}}\right] \cdot x^\top = (\hat{y} - y)\left(w^{(2)} \circ \mathbb{I}_{z^{(1)}>0}\right) \cdot x^\top$
$\frac{\partial L}{\partial b^{(1)}} = (\hat{y} - y)\left(w^{(2)} \circ \mathbb{I}_{z^{(1)}>0}\right)$

Final update equations:
For learning rate η:
$w^{(2)} := w^{(2)} - \eta \cdot (\hat{y} - y)h$
$b^{(2)} := b^{(2)} - \eta \cdot (\hat{y} - y)$

$W^{(1)} := W^{(1)} - \eta \cdot (\hat{y} - y) \left(w^{(2)} \circ \mathbb{I}_{z^{(1)}>0}\right) x^\top$
$b^{(1)} := b^{(1)} - \eta \cdot (\hat{y} - y) \left(w^{(2)} \circ \mathbb{I}_{z^{(1)}>0}\right)$

7)

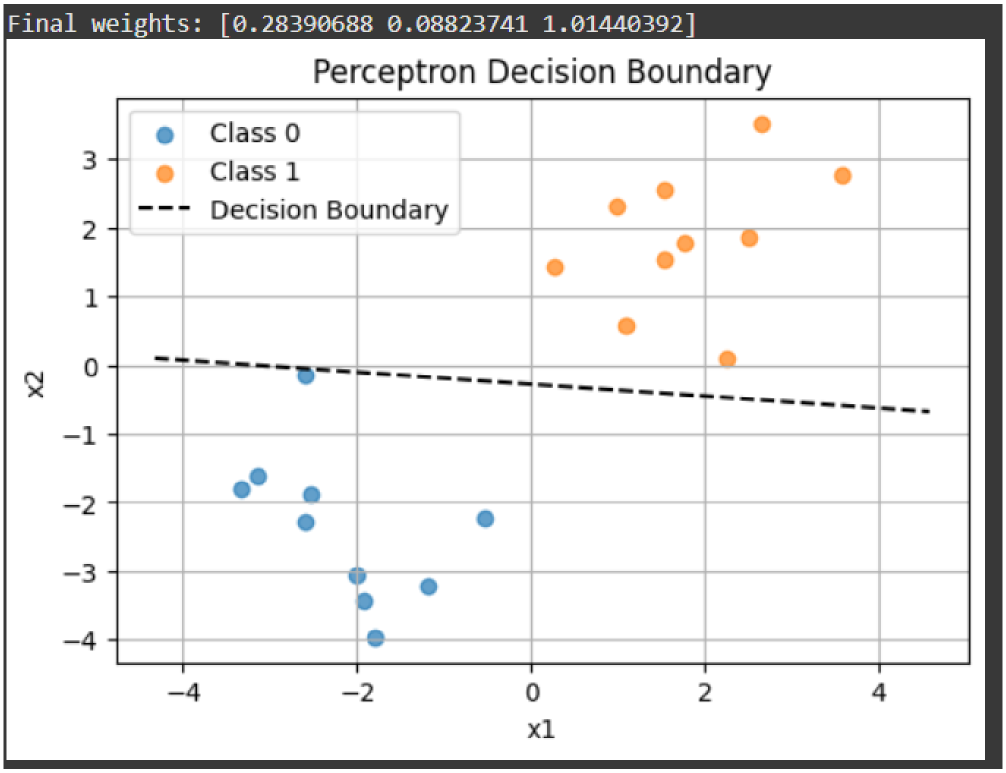

Figure 1: Question 7

8)

```
Epoch 0 | Train Loss: 0.1722 | Test Loss: 0.1606 | Train Acc: 0.5917 | Test Acc: 0.6333
Epoch 100 | Train Loss: 0.0100 | Test Loss: 0.0036 | Train Acc: 0.9833 | Test Acc: 1.0000
Epoch 200 | Train Loss: 0.0093 | Test Loss: 0.0022 | Train Acc: 0.9833 | Test Acc: 1.0000
Epoch 300 | Train Loss: 0.0090 | Test Loss: 0.0018 | Train Acc: 0.9833 | Test Acc: 1.0000
Epoch 400 | Train Loss: 0.0086 | Test Loss: 0.0014 | Train Acc: 0.9833 | Test Acc: 1.0000
Epoch 500 | Train Loss: 0.0082 | Test Loss: 0.0011 | Train Acc: 0.9833 | Test Acc: 1.0000
Epoch 600 | Train Loss: 0.0078 | Test Loss: 0.0009 | Train Acc: 0.9833 | Test Acc: 1.0000
Epoch 700 | Train Loss: 0.0074 | Test Loss: 0.0008 | Train Acc: 0.9833 | Test Acc: 1.0000
Epoch 800 | Train Loss: 0.0065 | Test Loss: 0.0016 | Train Acc: 0.9917 | Test Acc: 1.0000
Epoch 900 | Train Loss: 0.0052 | Test Loss: 0.0058 | Train Acc: 0.9917 | Test Acc: 1.0000
```

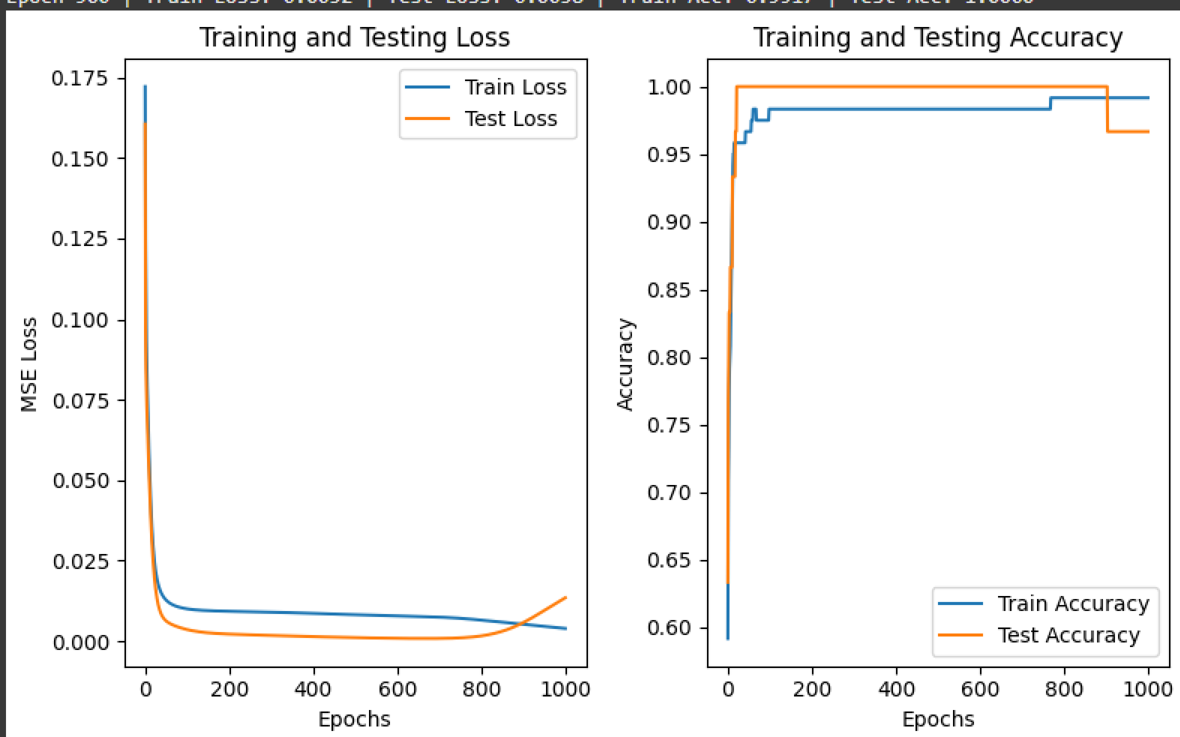

Figure 2: Question 8