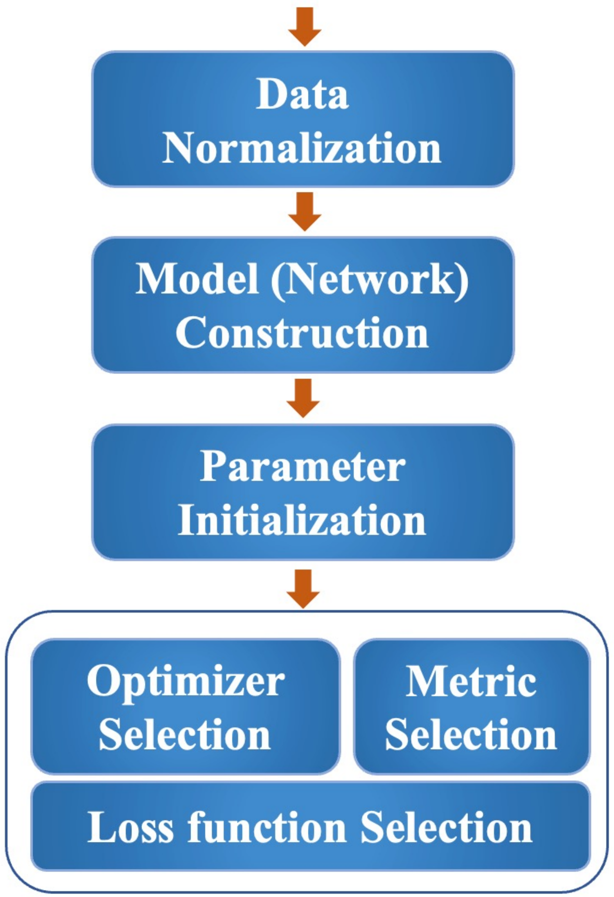
AI VIET NAM
@aivietnam.edu.vn

Insight into Multi-layer Perceptron

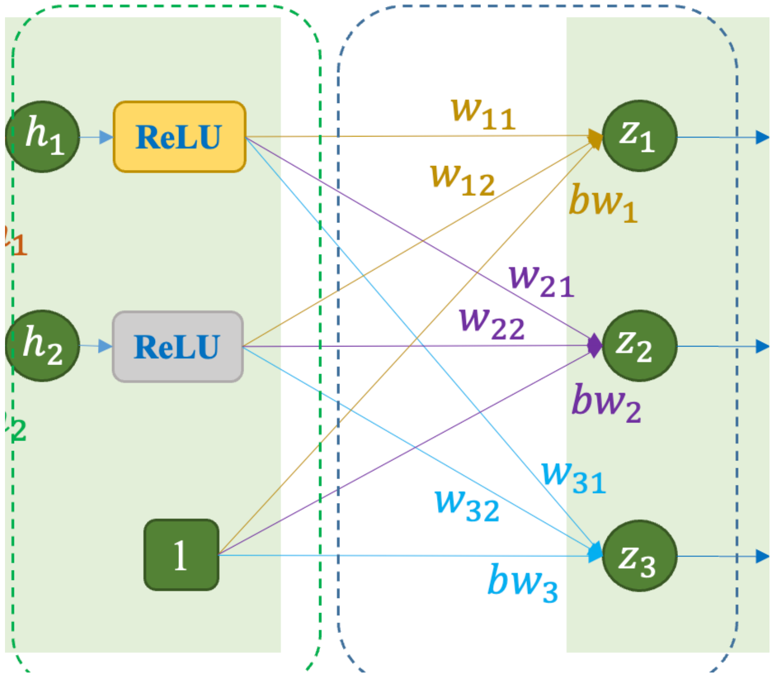
Quang-Vinh Dinh
Ph.D. in Computer Science

Objectives

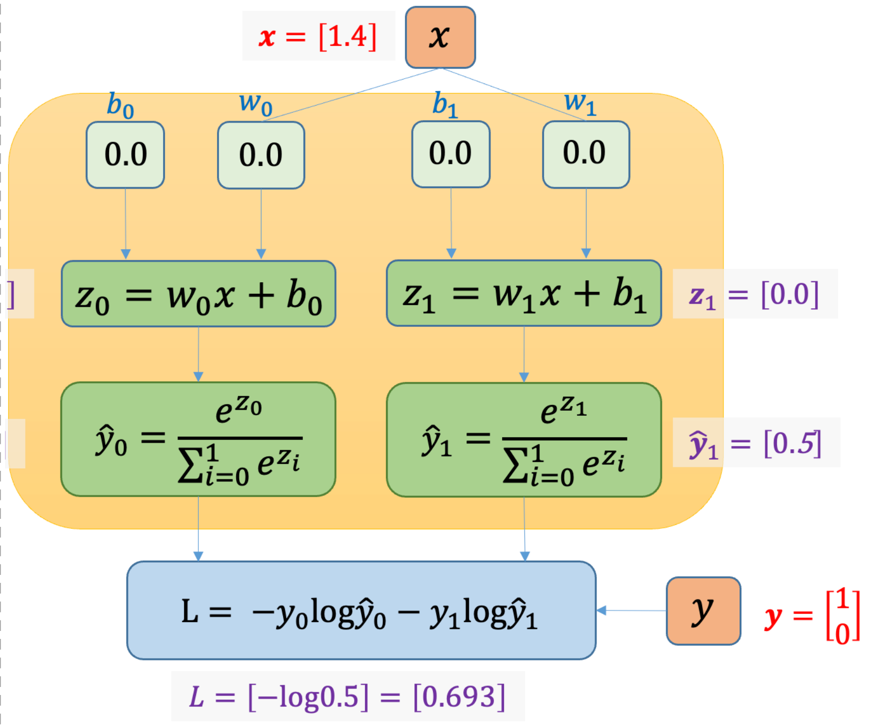
MLP Insight



MLP Examples



Init. Examples



Outline

SECTION 1

MLP Insight

SECTION 2

MLP Examples

SECTION 3

Initialization Examples

Data Normalization

Model (Network) Construction

Parameter Initialization

Optimizer Selection

Metric Selection

Loss function Selection

To-do List for Training

- Data Preparation
- Data Normalization
- Model (Network) Construction
- Parameter Initialization
- Optimizer Selection
- Metric Selection
- Loss function Selection

Data Preparation

Training Data

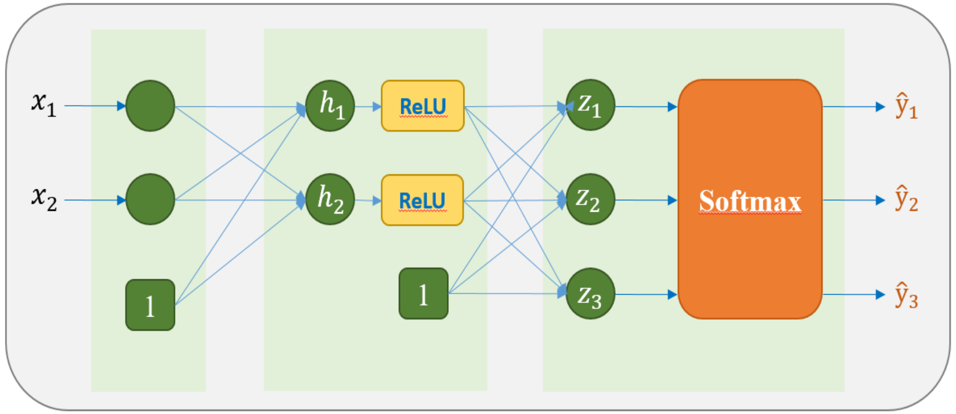
Used to train model
(Teach the model by examples)

≠

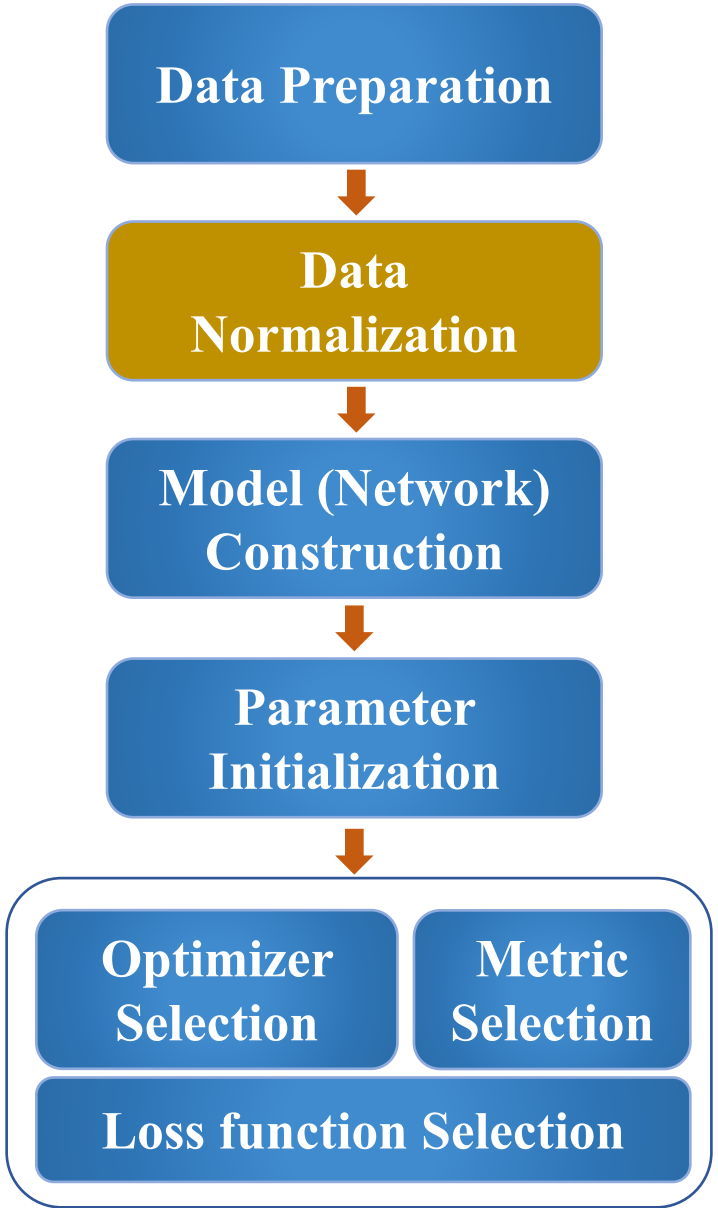
Testing Data

Used to validate model
(Check how good the model is)

Model



Data Normalization



In Theory

$$X \in [0, 255]$$

Convert to the range [0,1]

$$\text{Image} = \frac{\text{Image}}{255}$$

Convert to the range [-1,1]

$$\text{Image} = \frac{\text{Image}}{127.5} - 1$$

Z-score normalization

$$\text{Image} = \frac{\text{Image} - \mu}{\sigma}$$

In Pytorch

$$X \in [0, 1]$$

Normalize(*mean*, std)

$$\text{Image} = \frac{\text{Image} - mean}{\text{std}}$$

[0,1]	mean = 0 ; std = 1
[-1,1]	mean = 0.5; std = 0.5

Compute mean and std from data

```python
transform = transforms.Compose([transforms.ToTensor(), transforms.Normalize((0.5,), (0.5,))])

trainset = torchvision.datasets.FashionMNIST(root='data', train=True, download=True, transform=transform)
trainloader = torch.utils.data.DataLoader(trainset, batch_size=1024, num_workers=10, shuffle=True)

testset = torchvision.datasets.FashionMNIST(root='data', train=False, download=True, transform=transform)
testloader = torch.utils.data.DataLoader(testset, batch_size=1024, num_workers=10, shuffle=False)
```

1

```python
transform = transforms.Compose([transforms.ToTensor(), transforms.Normalize((0,), (1.0,))])

trainset = torchvision.datasets.FashionMNIST(root='data', train=True, download=True, transform=transform)
trainloader = torch.utils.data.DataLoader(trainset, batch_size=1024, num_workers=10, shuffle=True)

testset = torchvision.datasets.FashionMNIST(root='data', train=False, download=True, transform=transform)
testloader = torch.utils.data.DataLoader(testset, batch_size=1024, num_workers=10, shuffle=False)
```

2

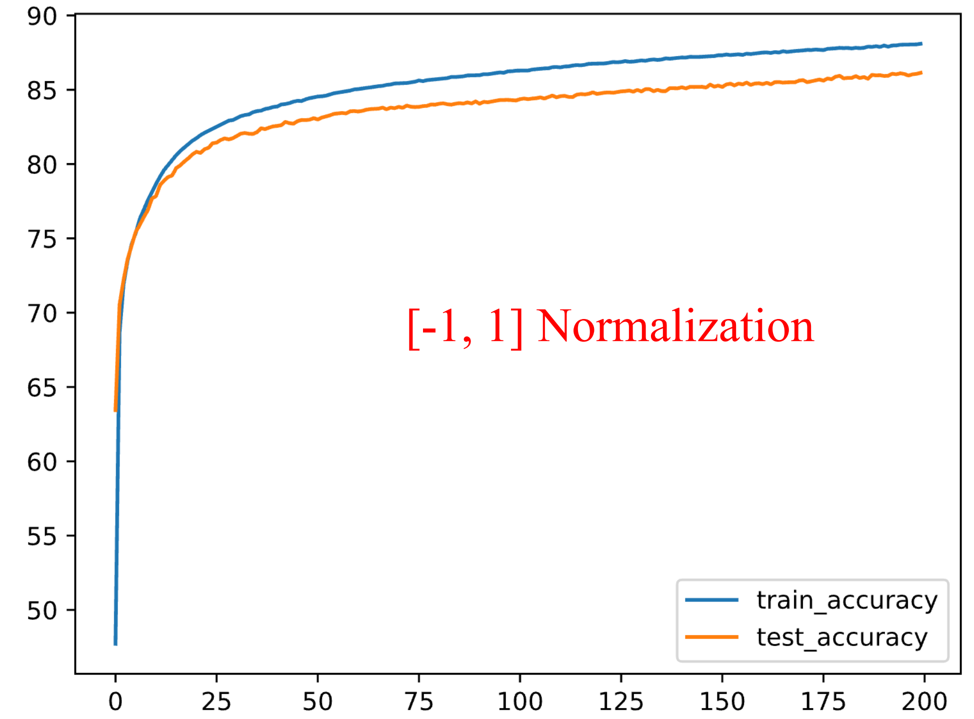
```python
# computed mean and std in advance
transform = transforms.Compose([transforms.ToTensor(), transforms.Normalize((mean,), (std,))])

trainset = torchvision.datasets.FashionMNIST(root='data', train=True, download=True, transform=transform)
trainloader = torch.utils.data.DataLoader(trainset, batch_size=1024, num_workers=10, shuffle=True)

testset = torchvision.datasets.FashionMNIST(root='data', train=False, download=True, transform=transform)
testloader = torch.utils.data.DataLoader(testset, batch_size=1024, num_workers=10, shuffle=False)
```

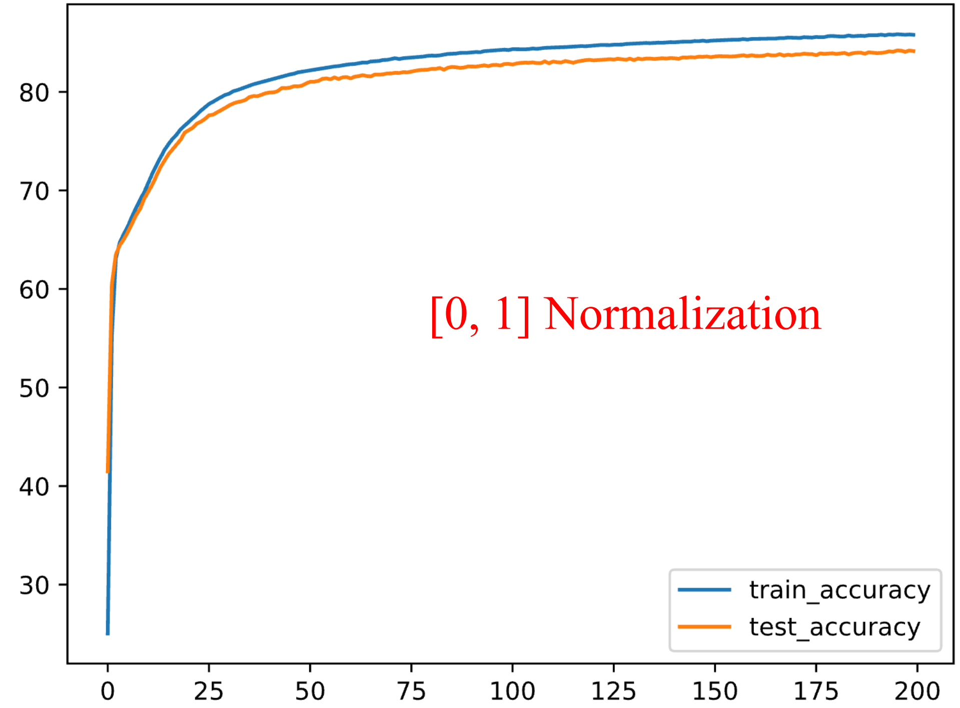
3

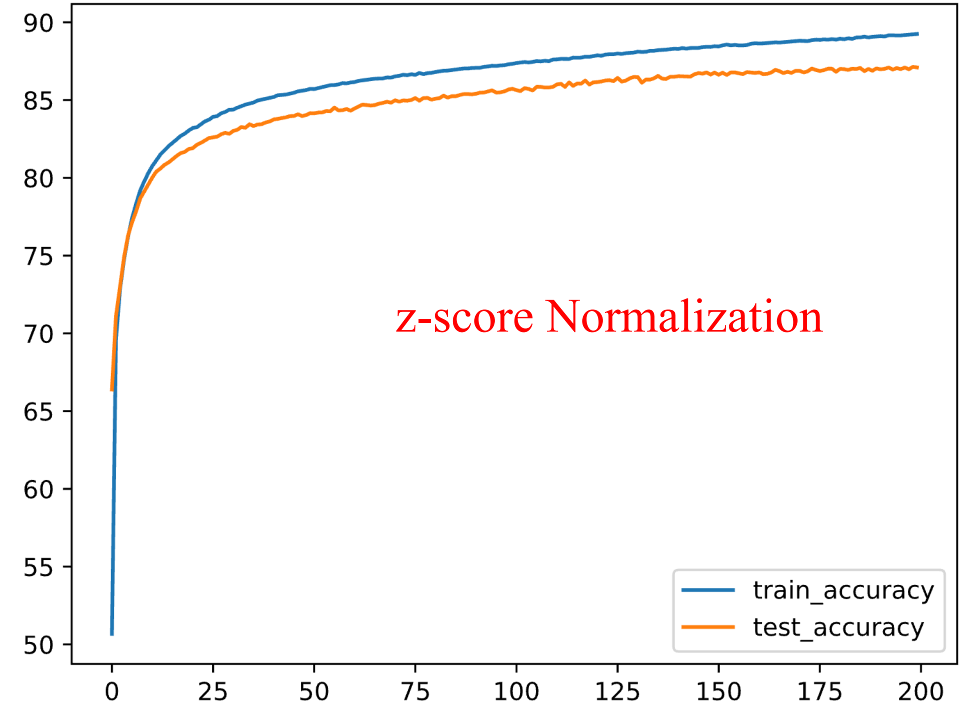
(a) [0, 1] Normalization (b) [-1, 1] Normalization (c) z-score Normalization

Data Normalization

```
model = nn.Sequential(
    nn.Flatten(), nn.Linear(784, 256),
    nn.ReLU(), nn.Linear(256, 10)
)
criterion = nn.CrossEntropyLoss()
optimizer = optim.SGD(model.parameters(),
                      lr=0.01)
```







Training Pipeline

Multi-layer Perceptron

1) #Hidden Layers?

2) #Nodes in a Hidden Layers?

3) Which activation functions?

4) Which Initializers?

Data Preparation

↓

Data Normalization

↓

Model (Network) Construction

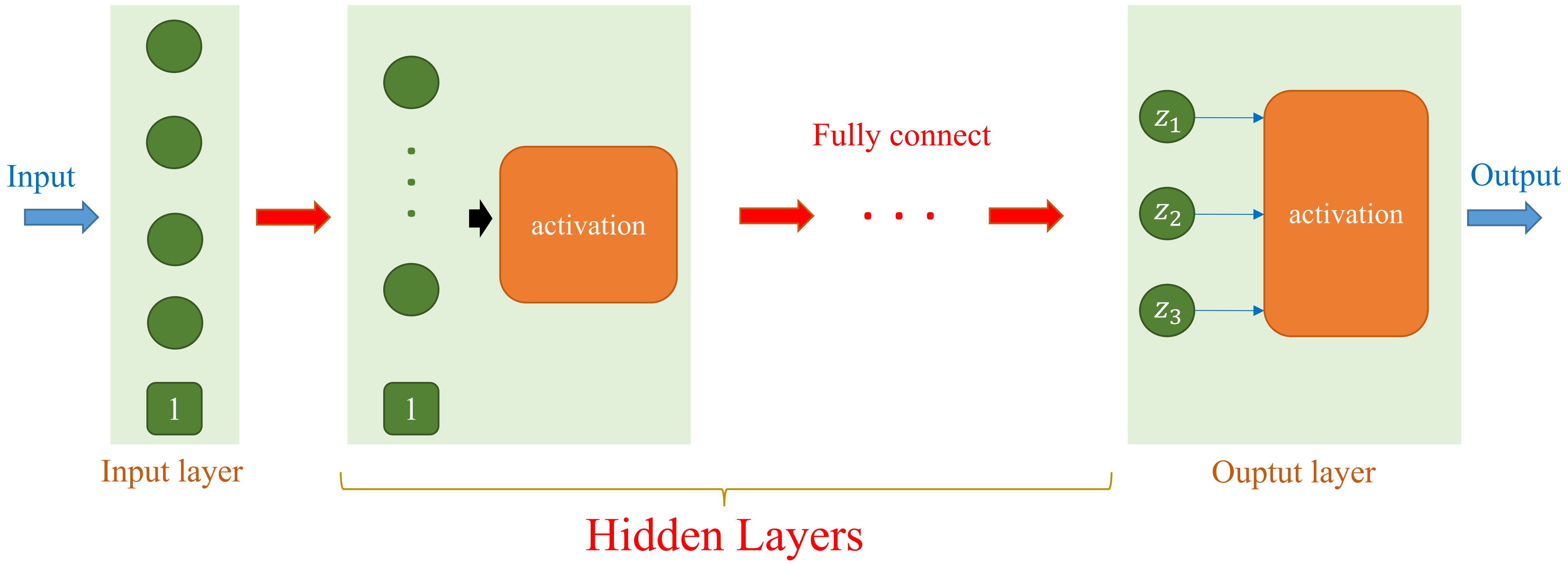
↓

Parameter Initialization →

- **Optimizer Selection**
- **Loss function Selection**
- **Metric Selection**

Training Pipeline

❖ Model (Network) Construction

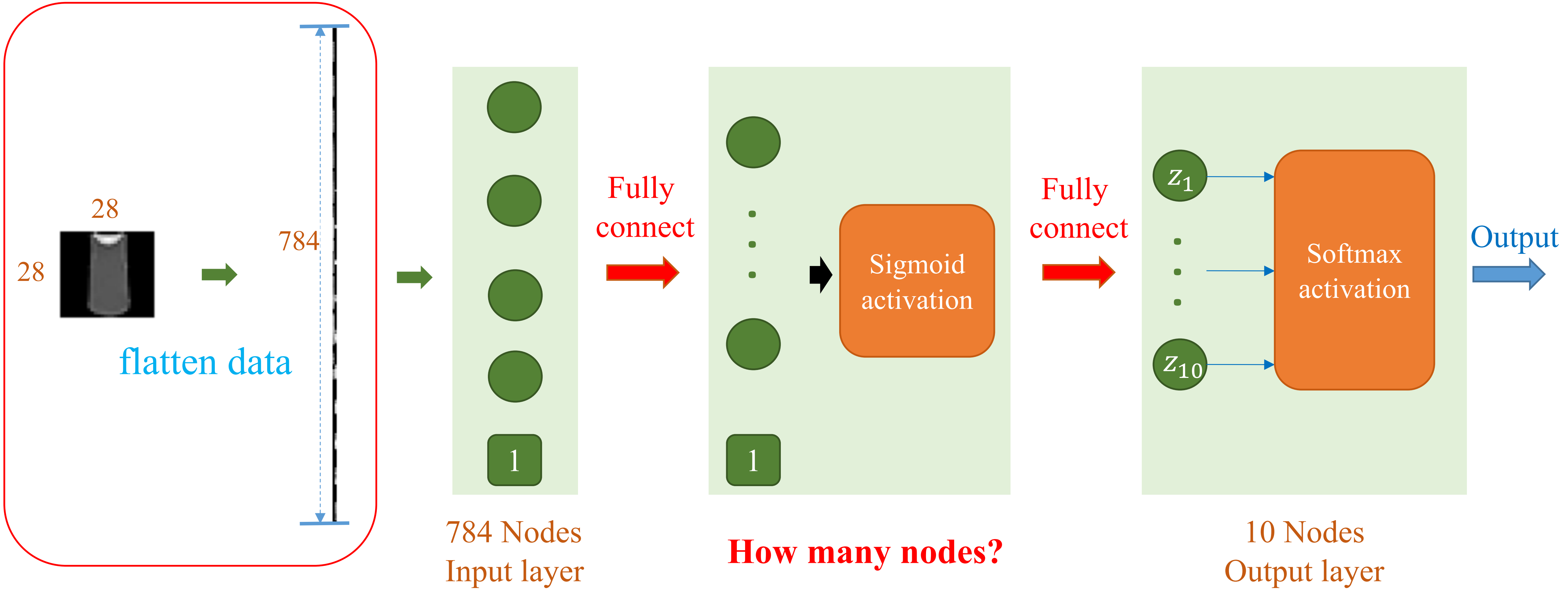


How many hidden layers?
How many nodes in a hidden layer?

Which activation function?
Which network components?

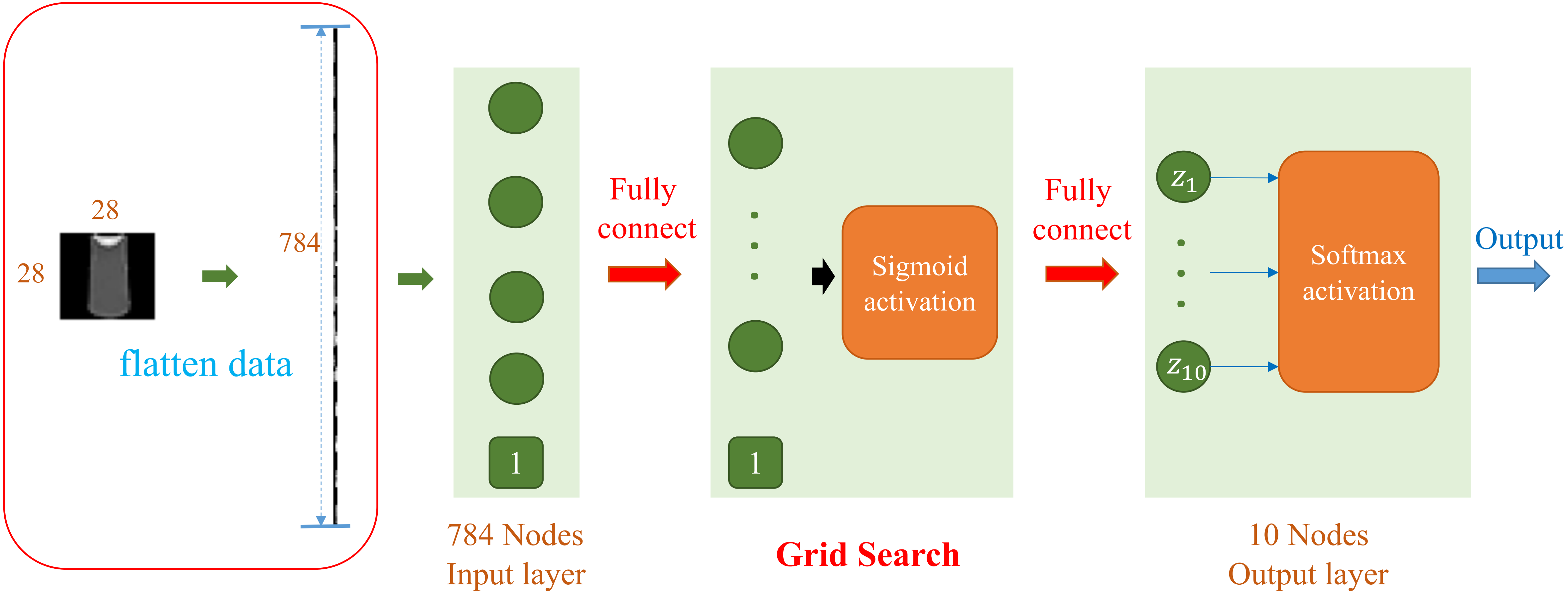
How many nodes?

❖ Model (Network) Construction



How many nodes?

❖ Model (Network) Construction



Grid Search

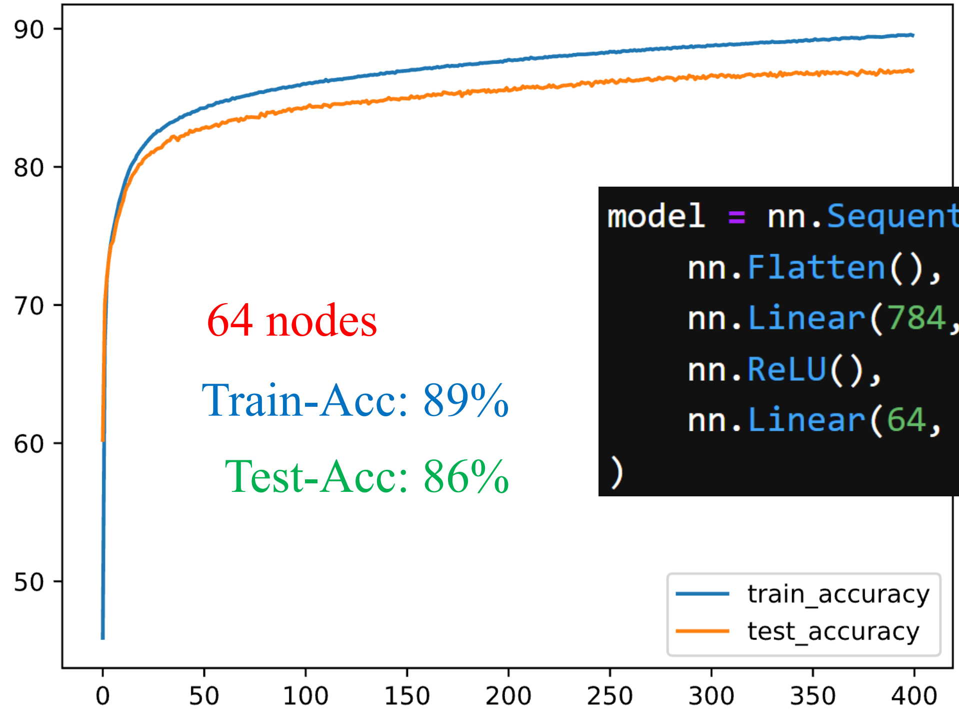
$$[N_1 \quad N_2 \quad \ldots \quad N_k]$$

How many nodes?

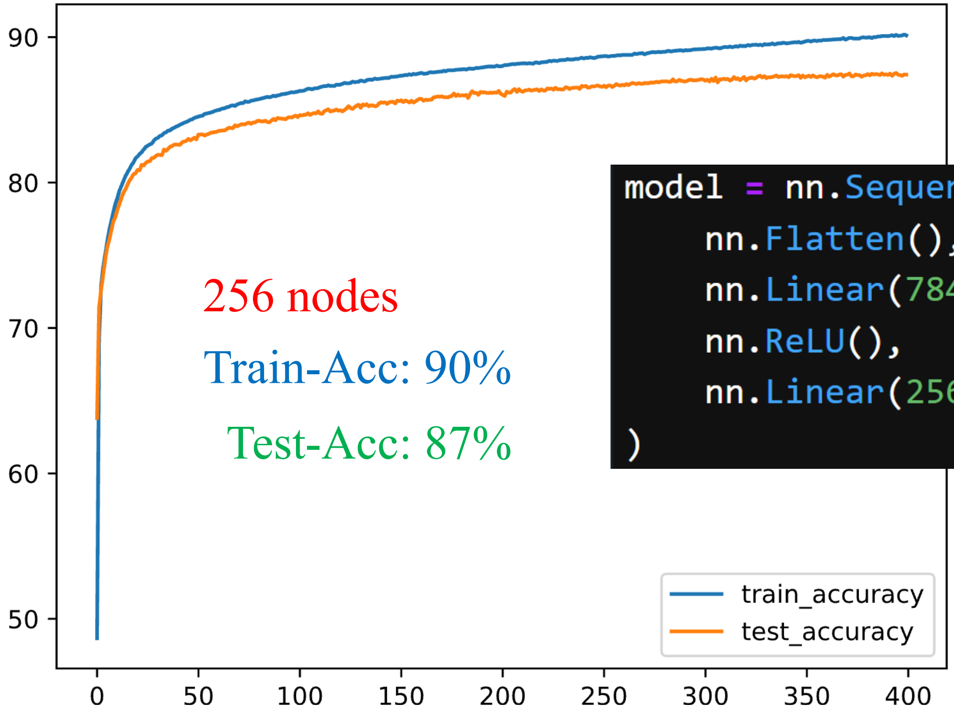
[-1, 1] Normalization

Cross-entropy Loss

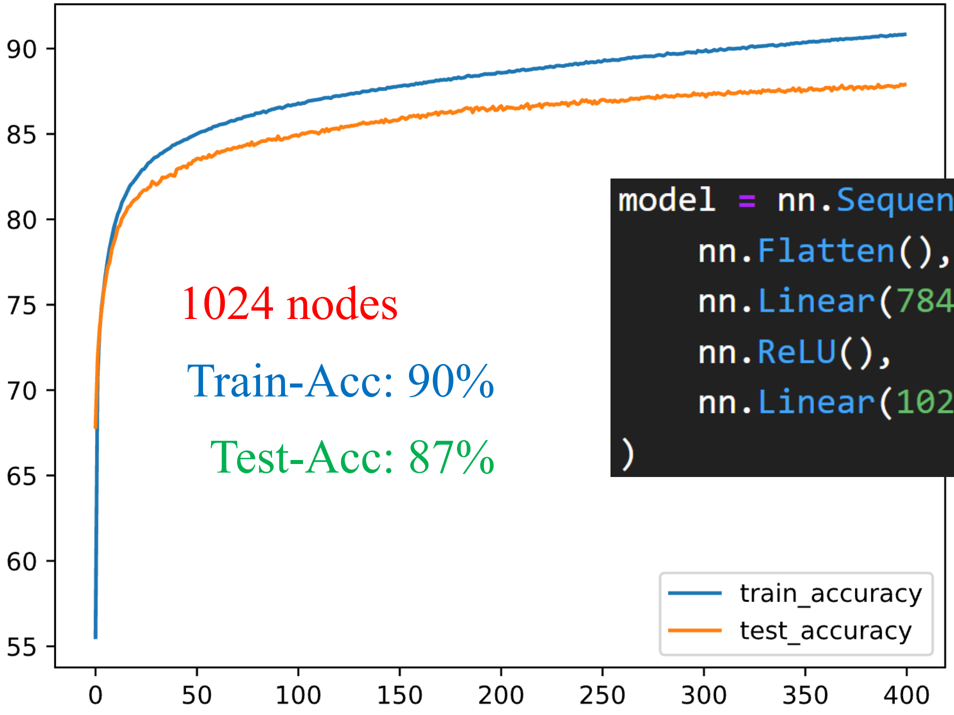
SGD with lr=0.01



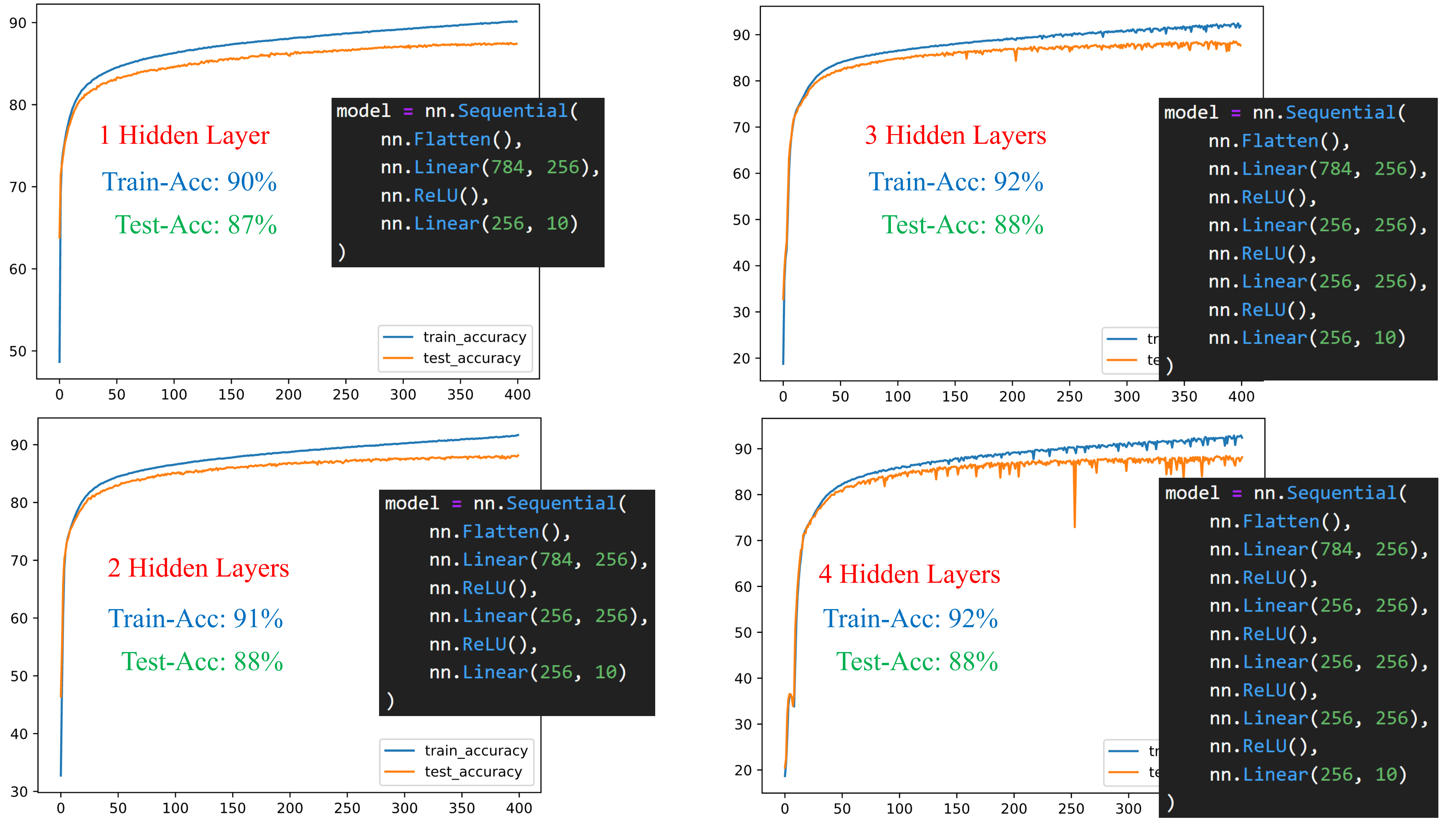

```
model = nn.Sequential(
    nn.Flatten(),
    nn.Linear(784, 64),
    nn.ReLU(),
    nn.Linear(64, 10)
)
```




```
model = nn.Sequential(
    nn.Flatten(),
    nn.Linear(784, 256),
    nn.ReLU(),
    nn.Linear(256, 10)
)
```




```
model = nn.Sequential(
    nn.Flatten(),
    nn.Linear(784, 1024),
    nn.ReLU(),
    nn.Linear(1024, 10)
)
```


1 Hidden Layer
Train-Acc: 90%
Test-Acc: 87%
model = nn.Sequential(
nn.Flatten(),
nn.Linear(784, 256),
nn.ReLU(),
nn.Linear(256, 10)
)
train_accuracy
test_accuracy
3 Hidden Layers
Train-Acc: 92%
Test-Acc: 88%
model = nn.Sequential(
nn.Flatten(),
nn.Linear(784, 256),
nn.ReLU(),
nn.Linear(256, 256),
nn.ReLU(),
nn.Linear(256, 256),
nn.ReLU(),
nn.Linear(256, 10)
)
2 Hidden Layers
Train-Acc: 91%
Test-Acc: 88%
model = nn.Sequential(
nn.Flatten(),
nn.Linear(784, 256),
nn.ReLU(),
nn.Linear(256, 256),
nn.ReLU(),
nn.Linear(256, 10)
)
train_accuracy
test_accuracy
4 Hidden Layers
Train-Acc: 92%
Test-Acc: 88%
model = nn.Sequential(
nn.Flatten(),
nn.Linear(784, 256),
nn.ReLU(),
nn.Linear(256, 256),
nn.ReLU(),
nn.Linear(256, 256),
nn.ReLU(),
nn.Linear(256, 256),
nn.ReLU(),
nn.Linear(256, 10)
)

Activation Functions

❖ Model (Network) Construction

Which activation function?

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$$

2010

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

2017

$$SELU(x) = \begin{cases} \lambda x & \text{if } x \geq 0 \\ \lambda\alpha(e^x - 1) & \text{if } x < 0 \end{cases}$$

$$\lambda \approx 1.0507$$

$$\alpha \approx 1.6733$$

$$\tanh(x) = \frac{2}{1+e^{-2x}} - 1$$

2015

$$\text{ELU}(x) = \begin{cases} \alpha(e^x - 1) & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

2001

$$\text{softplus}(x) = \log(1+e^x)$$

2015

$$\text{PReLU}(x) = \begin{cases} \alpha x & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

2017

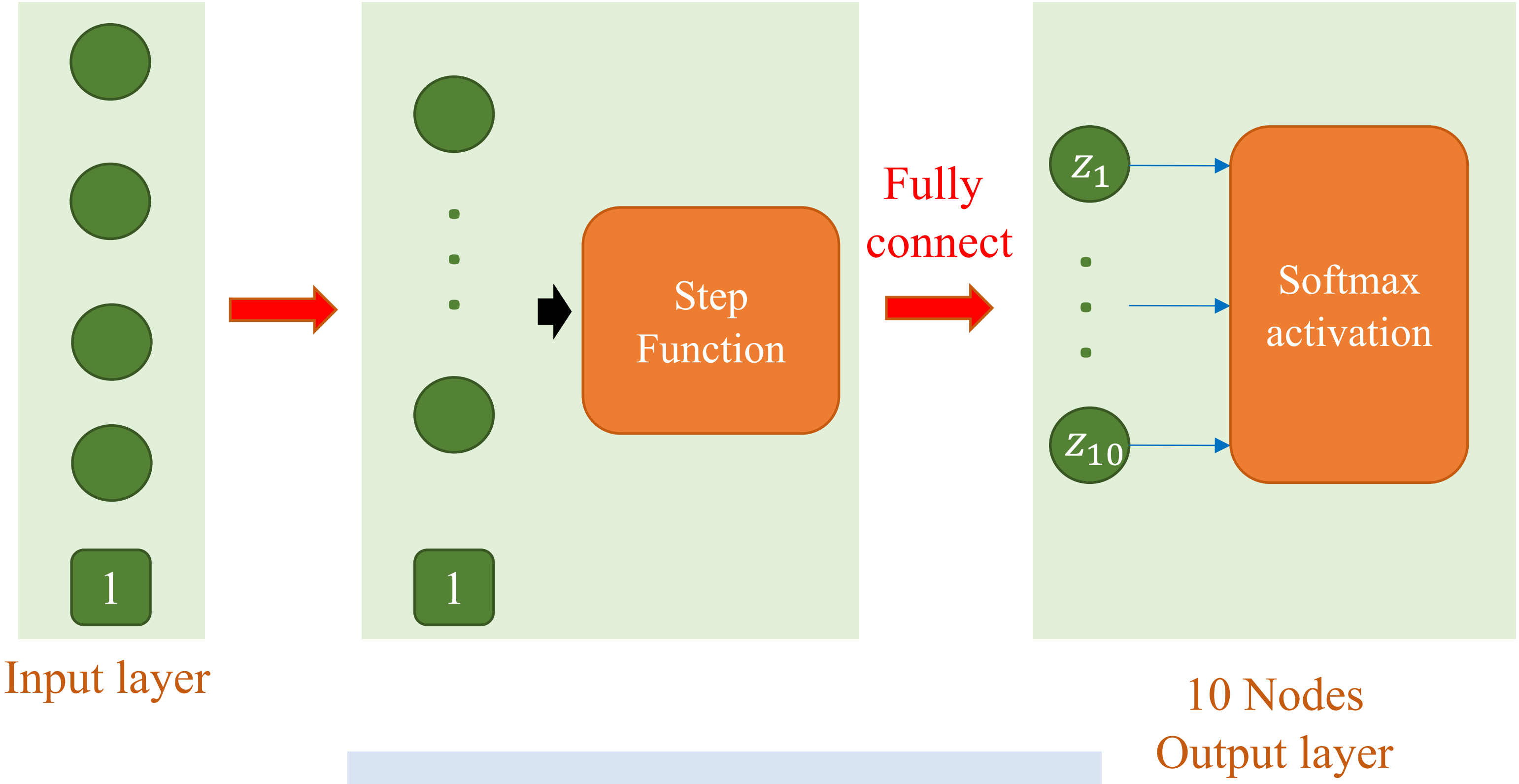
$$swish1(x) = x * \frac{1}{1+e^{-x}}$$

2023

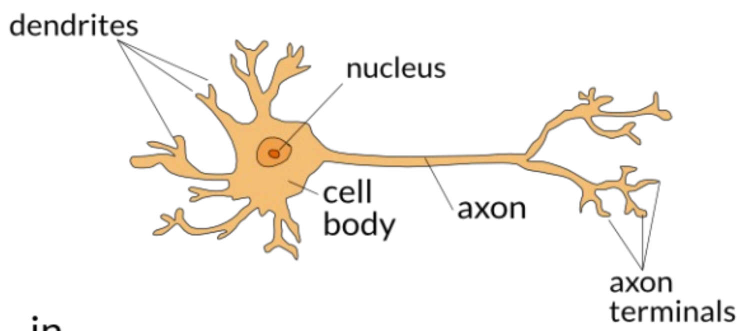
$$\text{GELU}(x) = x\phi(x) \approx x * \text{sigmoid}(1.702x)$$

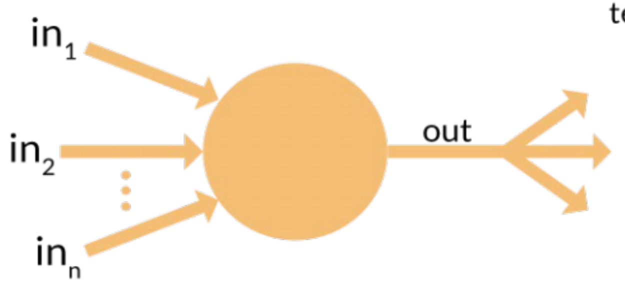
Activation Functions

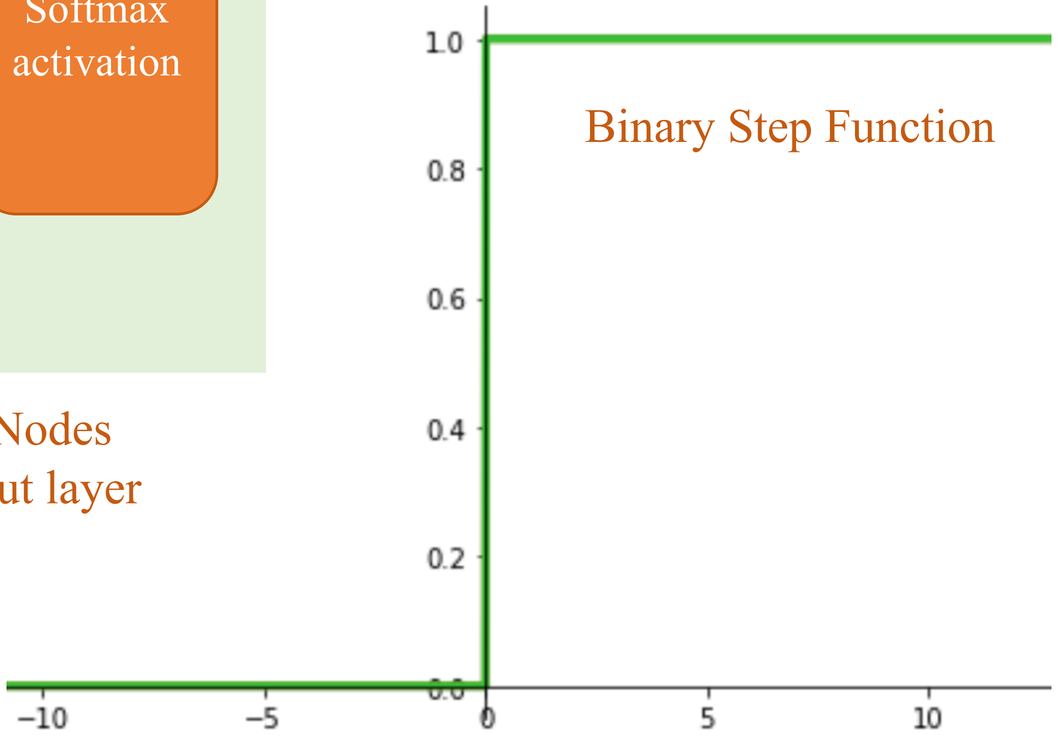
❖ Step function



$$f(x) = \begin{cases} 0 & if \;\; x < 0 \\ 1 & if \;\; x \geq 0 \end{cases}$$







https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/

Activation Functions

❖ Sigmoid function

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

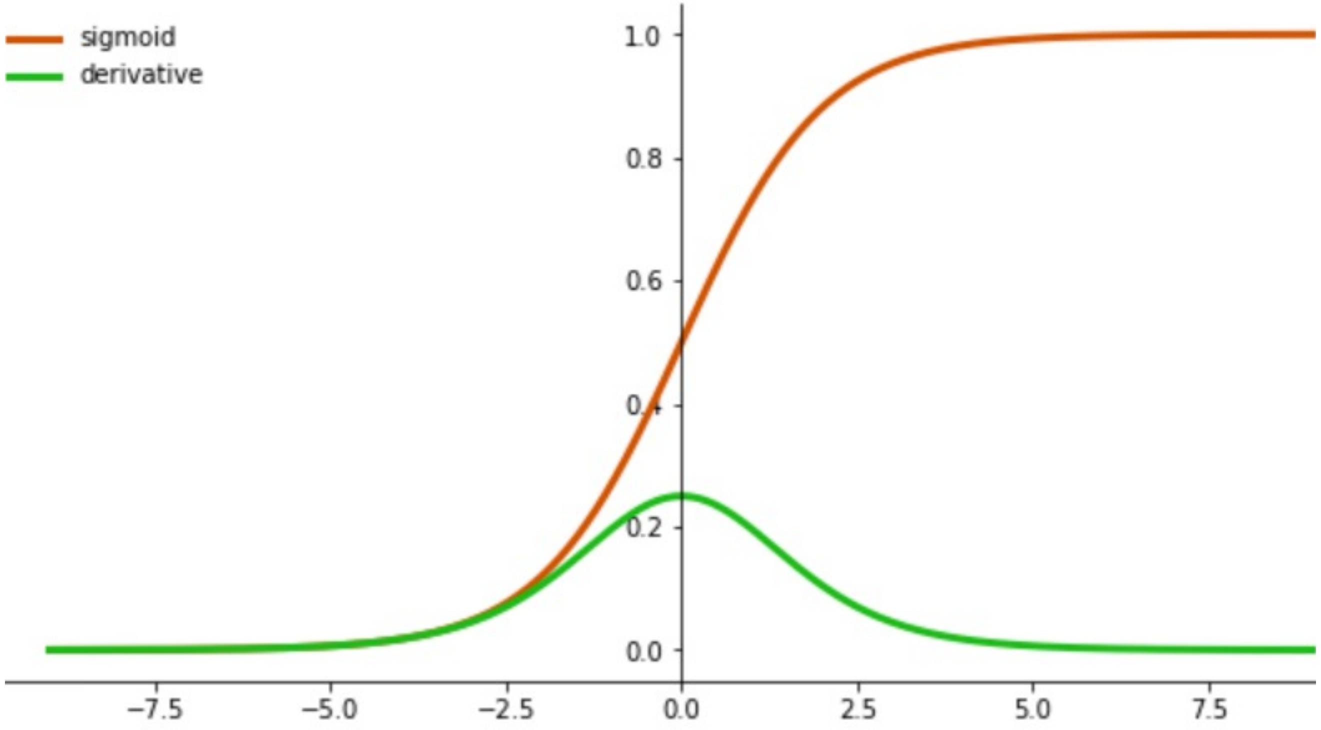
data =

1	5	-4	3	-2

data_a = sigmoid(data)

data_a =

0.731	0.993	0.017	0.95	0.119



$$\text{sigmoid}'(x) = \text{sigmoid}(x)\,(1 - \text{sigmoid}(x))$$

Activation Functions

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$$

$$\text{sigmoid}'(x) = \left(\frac{1}{1+e^{-x}}\right)' = \frac{-1}{(1+e^{-x})^2}(-e^{-x})$$

$$= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}+1-1}{(1+e^{-x})^2}$$

$$= \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2}$$

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

$$= \text{sigmoid}(x)\,(1-\text{sigmoid}(x))$$

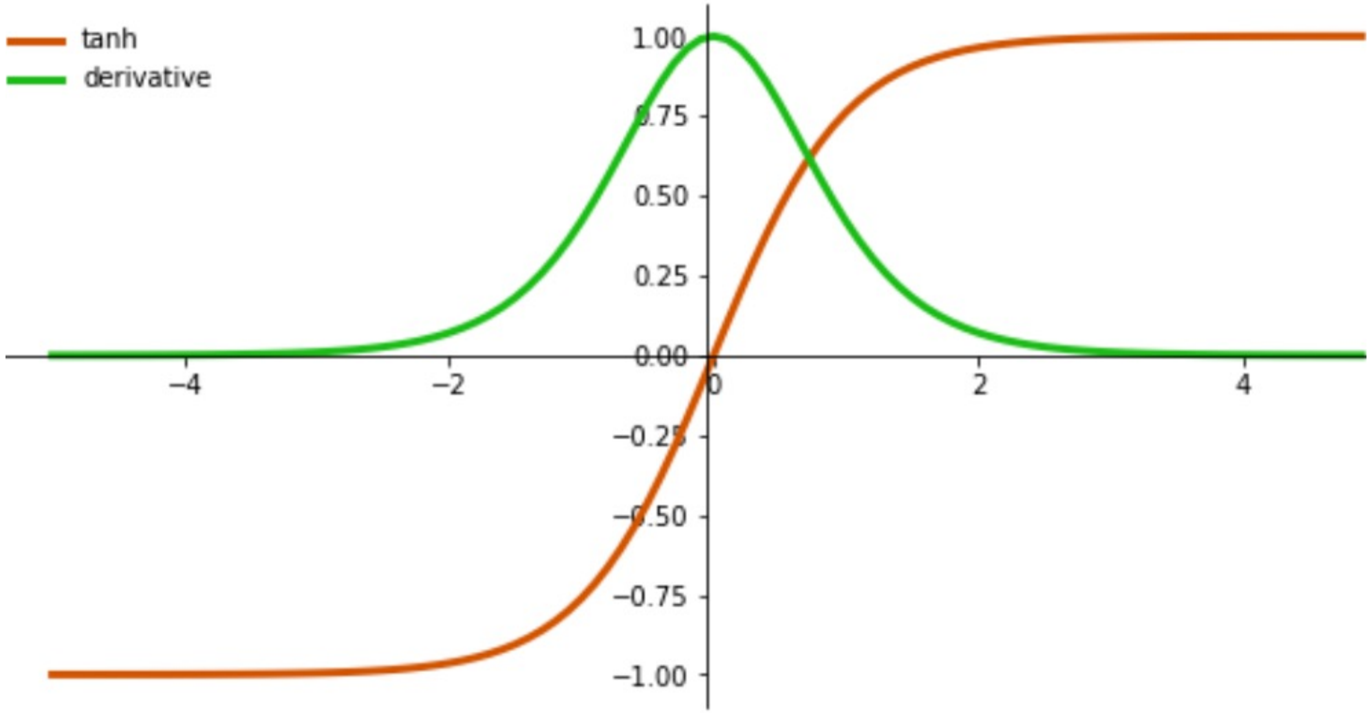
Activation Functions

❖ Tanh function

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$= \frac{2}{1 + e^{-2x}} - 1$$

$$= 1 - \frac{2}{e^{2x} + 1}$$



data =

1	5	-4	3	-2

data_a = tanh(data)

data_a =

0.761	0.999	-0.999	0.995	-0.964

$$tanh'(x) = 1 - tanh^{2}(x)$$

Activation Functions

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 1 - \frac{2}{e^{2x} + 1} = \frac{2}{e^{-2x} + 1} - 1$$

$$\tanh'(x) = \left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)' = \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2}$$

$$= \frac{(e^x + e^{-x})^2 - (e^x - e^{-x})^2}{(e^x + e^{-x})^2}$$

$$= 1 - \left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)^2 = 1 - tanh^2(x)$$

Activation Functions

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 1 - \frac{2}{e^{2x} + 1} = \frac{2}{e^{-2x} + 1} - 1$$

$$tanh'(x) = \left(\frac{2}{e^{-2x} + 1} - 1\right)' = \frac{4e^{-2x}}{(e^{-2x} + 1)^2} = 4\left(\frac{e^{-2x} + 1 - 1}{(e^{-2x} + 1)^2}\right)$$

$$= 4\left(\frac{1}{e^{-2x} + 1} - \frac{1}{(e^{-2x} + 1)^2}\right) = -\left(\frac{4}{(e^{-2x} + 1)^2} - \frac{4}{e^{-2x} + 1}\right)$$

$$= -\left(\frac{4}{(e^{-2x} + 1)^2} - \frac{4}{e^{-2x} + 1} + 1 - 1\right) = 1 - \left(\frac{2}{e^{-2x} + 1} - 1\right)^2 = 1 - tanh^2(x)$$

Activation Functions

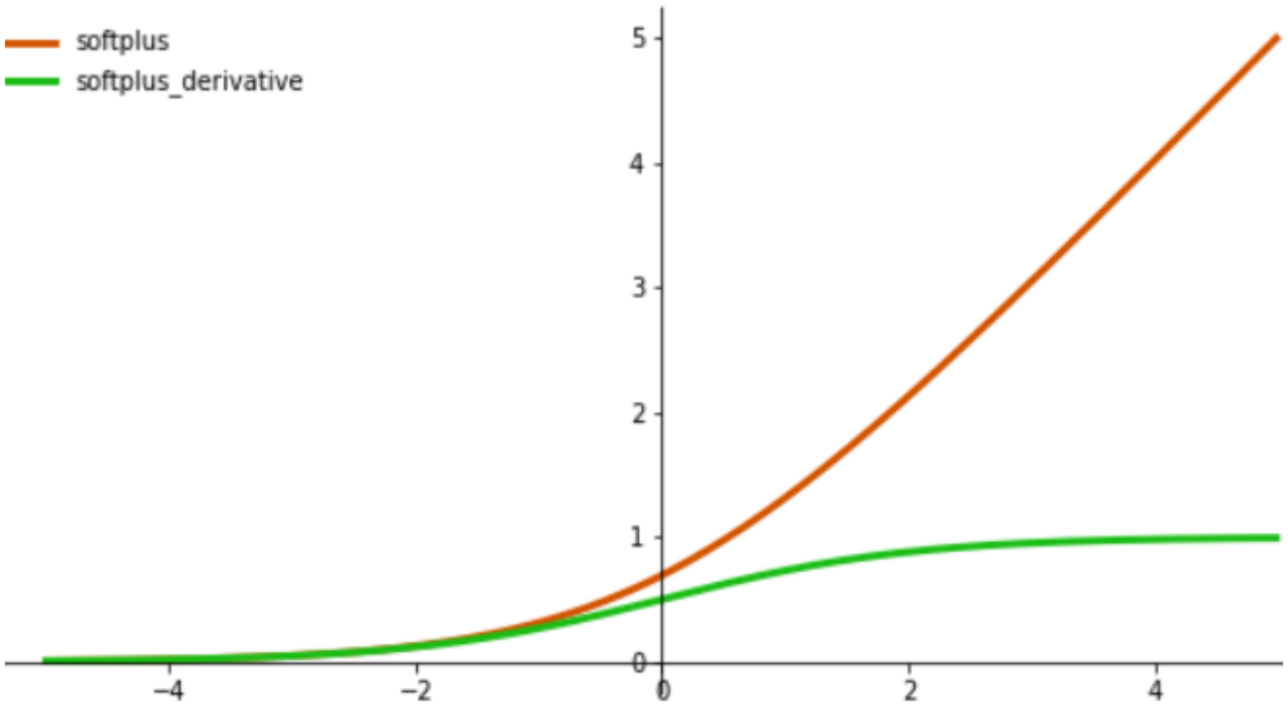
❖ Softplus function

$$\text{softplus}(x) = \log(1 + e^x)$$

data =	1	5	-4	3	-2

data_a = softplus(data)

data_a =	1.313	5.006	0.018	3.048	0.126



$$\text{softplus}'(x) = \frac{1}{1 + e^{-x}}$$

Activation Functions

❖ ReLU function

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

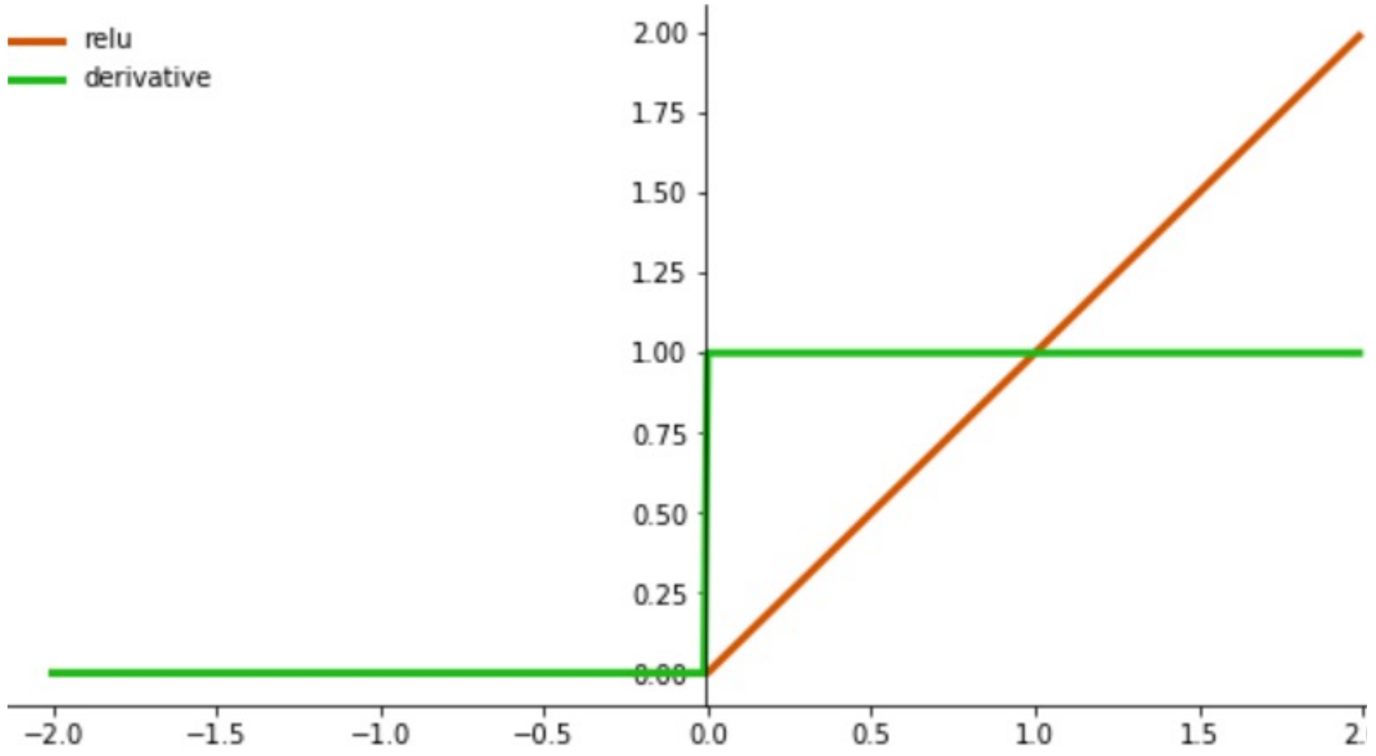
data =

1	5	-4	3	-2

data_a = ReLU(data)

data_a =

1	5	0	3	0



$$\text{ReLU}'(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

AI VIET NAM
@aivietnam.edu.vn

Activation Functions

❖ LeakyReLU function

$$\text{LeakyReLU}(x) = \begin{cases} 0.01x & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

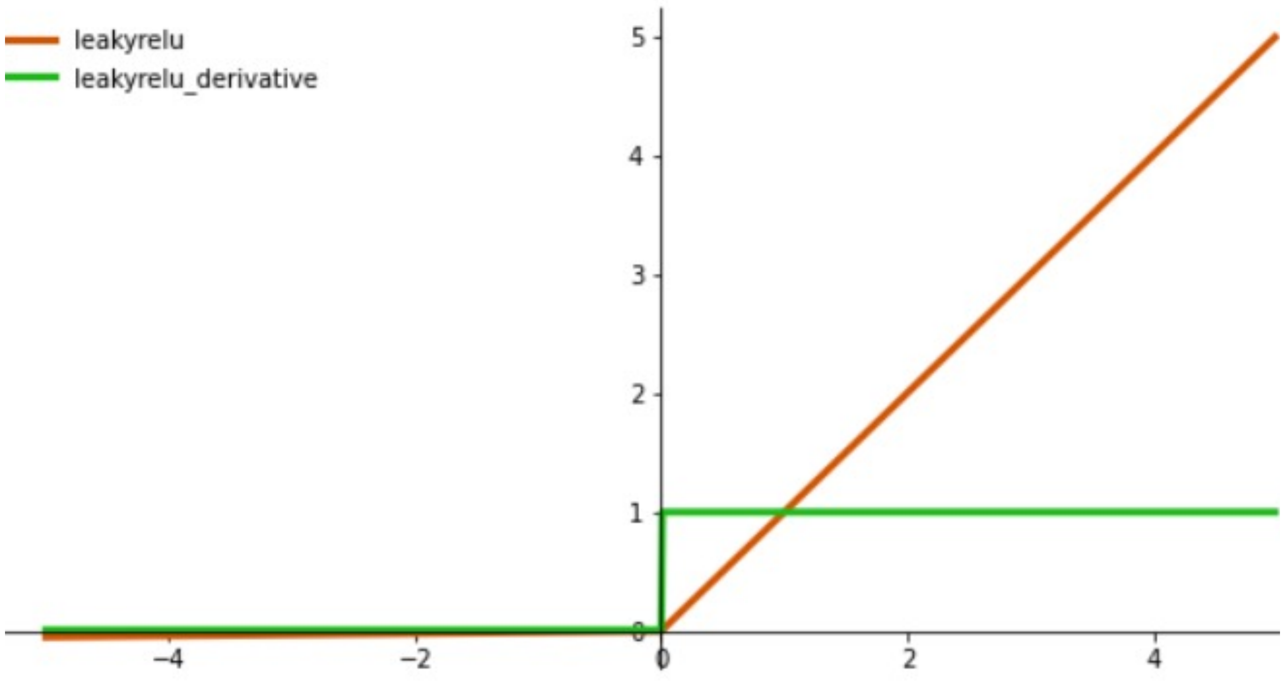
data =

1	5	-4	3	-2

data_a = leakyrelu(data)

data_a =

1	5	-0.04	3	-0.02



$$\text{LeakyReLU}'(x) = \begin{cases} 0.01 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Activation Functions

❖ ELU function

$$\text{ELU}(x) = \begin{cases} \alpha(e^x - 1) & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

$$\alpha = 0.1$$

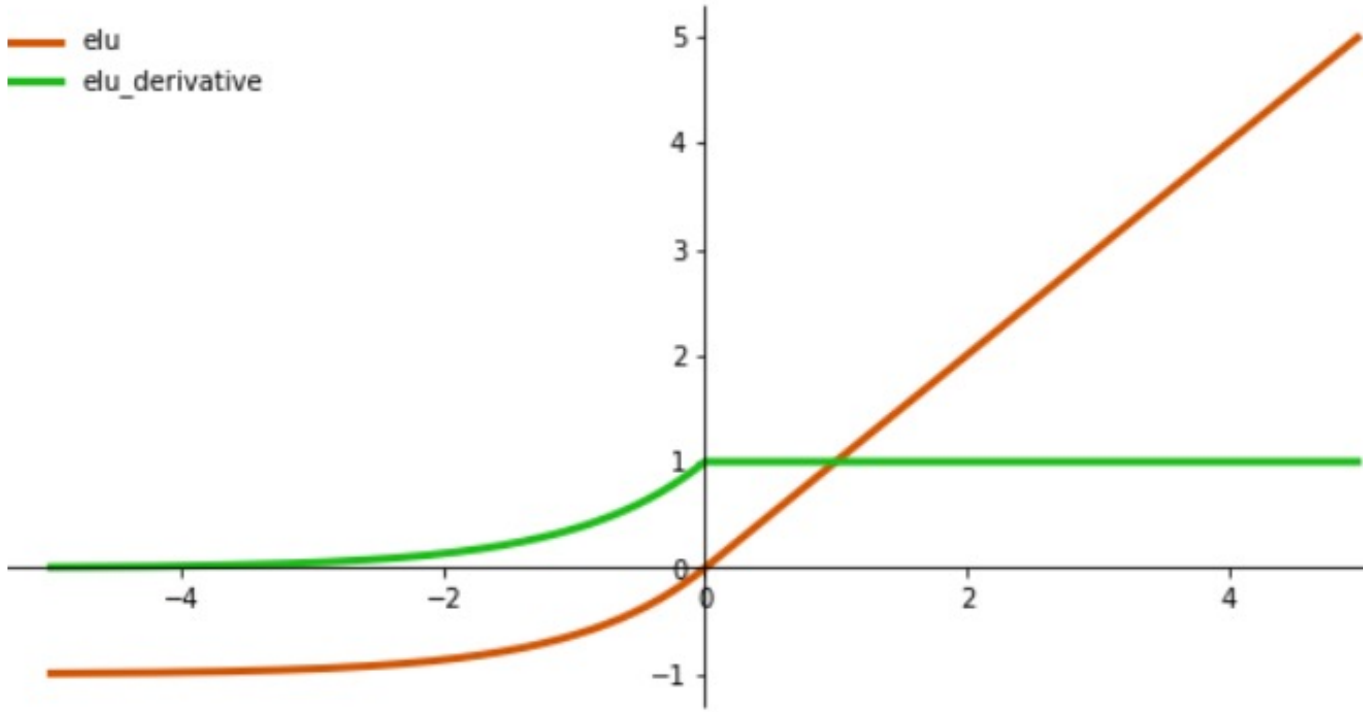
data =

1	5	-4	3	-2

data_a = ELU(data)

data_a =

1	5	-0.098	3	-0.086



$$\text{ELU}'(x) = \begin{cases} \alpha e^x & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Activation Functions

❖ PReLU function

$$\text{PReLU}(x) = \begin{cases} \alpha x & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

$\alpha = 0.1$

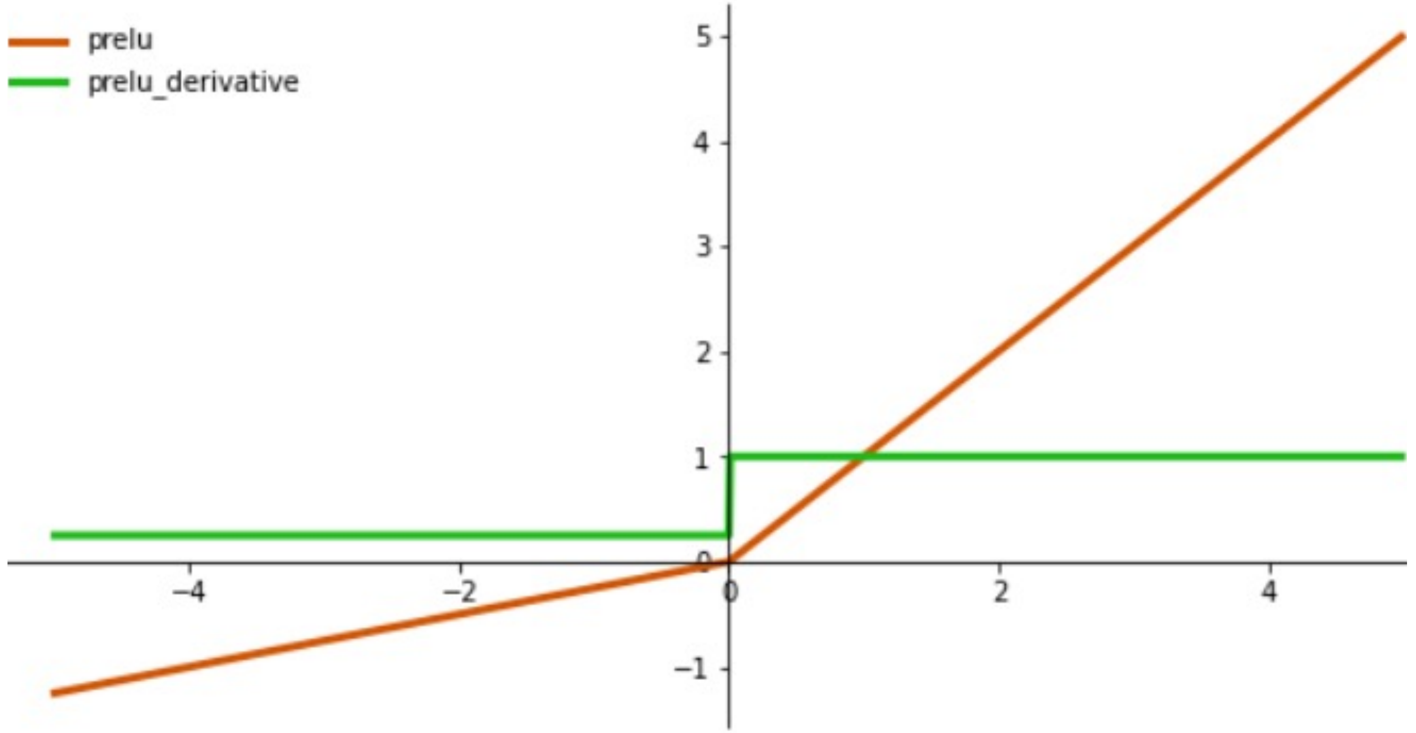
data =

1	5	-4	3	-2

data_a = PRELU(data)

data_a =

1	5	-0.4	3	-0.2



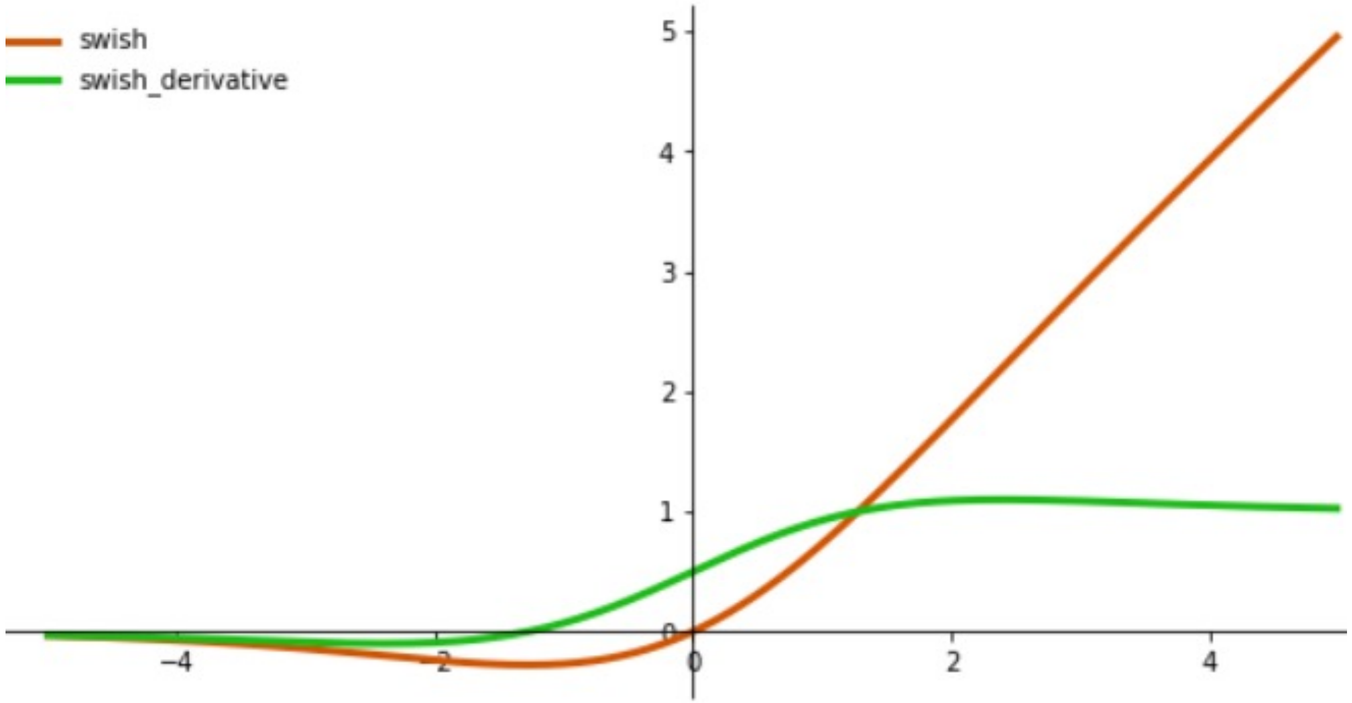
$$\text{PReLU}'(x) = \begin{cases} \alpha & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Activation Functions

❖ Swish function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$swish(x) = \frac{x}{1 + e^{-x}} = x\,\sigma(x)$$



data =

1	5	-4	3	-2

data_a = swish(data)

data_a =

0.731	4.966	-0.071	2.857	-0.238

$$swish'(x) = swish(x) + \sigma(x)\,(1 - swish(x))$$

Activation Functions

❖ Swish function

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$swish(x) = \frac{x}{1+e^{-x}} = x\,\sigma(x)$$

$$\begin{aligned}
swish'(x) &= (x\,\sigma(x))' = (x)'\,\sigma(x) + x(\sigma(x))' \\
&= \sigma(x) + x\,\sigma(x)\,(1-\sigma(x)) \\
&= \sigma(x) + x\,\sigma(x) - x\,\sigma(x)^2 \\
&= x\,\sigma(x) + \sigma(x)(1 - x\,\sigma(x)) \\
&= swish(x) + \sigma(x)\,(1 - swish(x))
\end{aligned}$$

Outline

SECTION 1

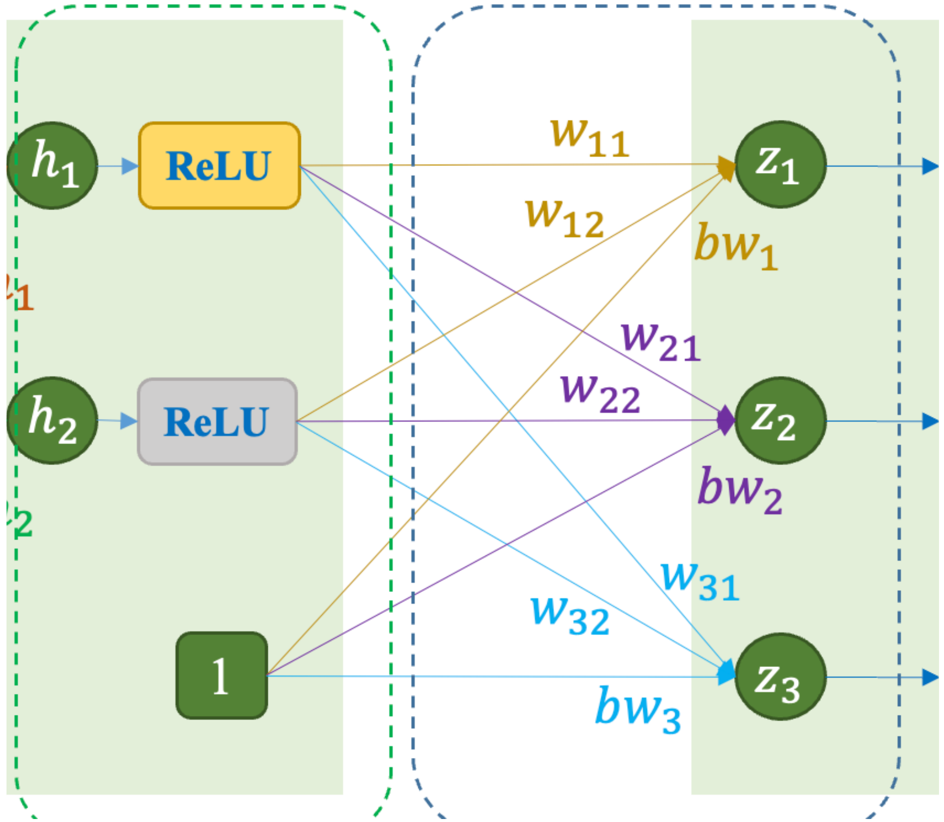
MLP Insight

SECTION 2

MLP Examples

SECTION 3

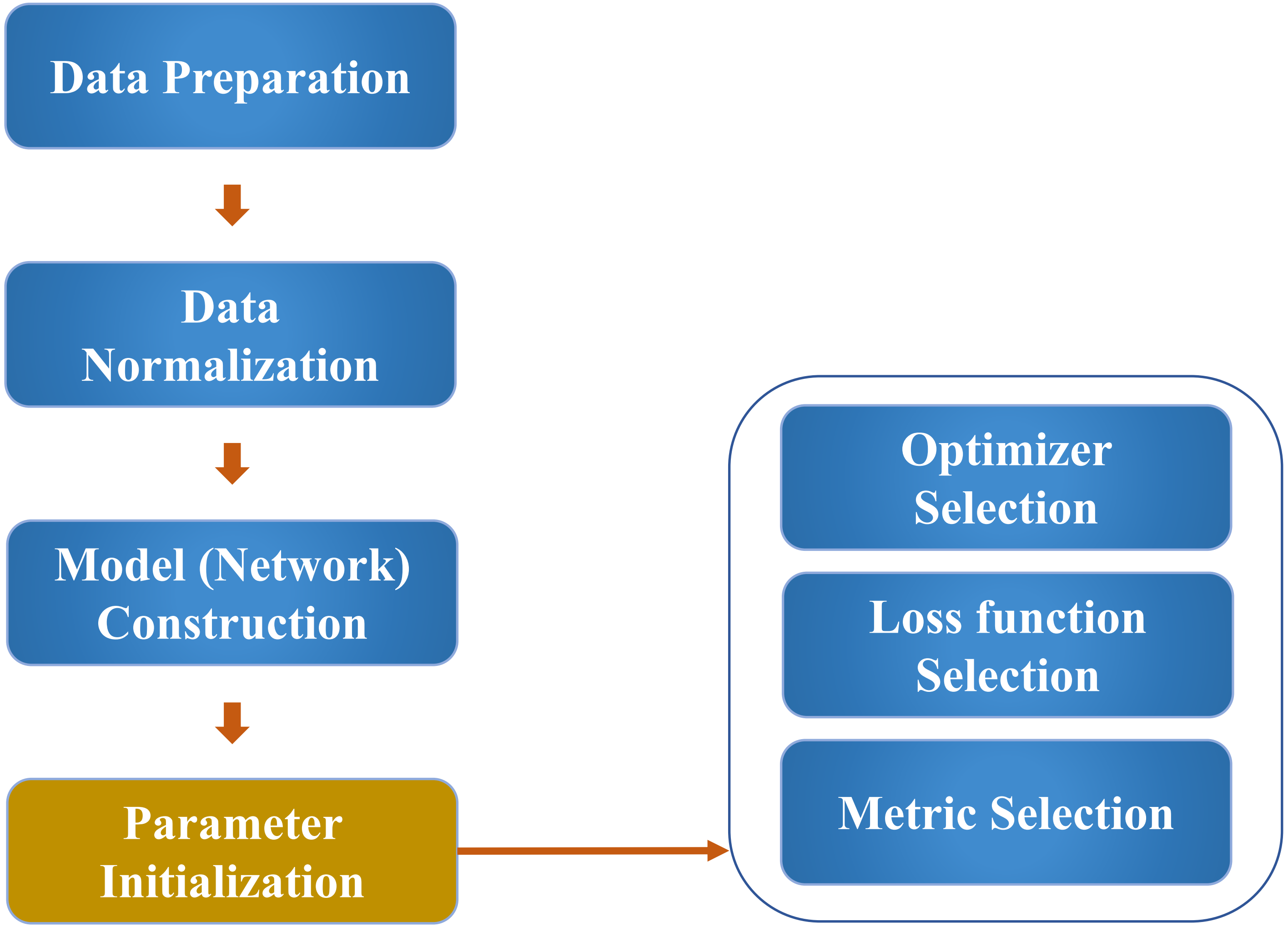
Initialization Examples



AI VIET NAM
@aivietnam.edu.vn

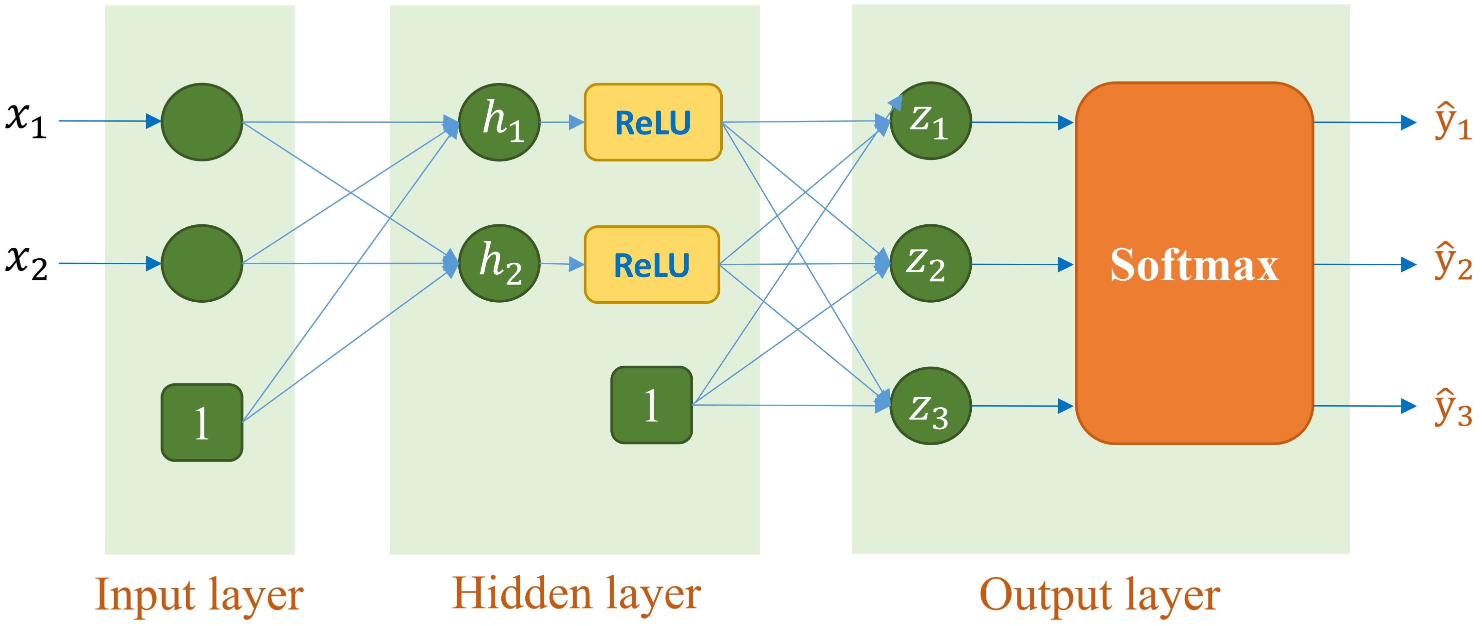
To-do List for Training

❖ Train a model



MLP Example 1

Feature		Label
Petal_Length	Petal_Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2



$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \boldsymbol{x}^{(3)} \end{bmatrix} = \begin{bmatrix} 1.5 & 0.2 \\ 4.7 & 1.6 \\ 5.6 & 2.2 \end{bmatrix}$$

$$\boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

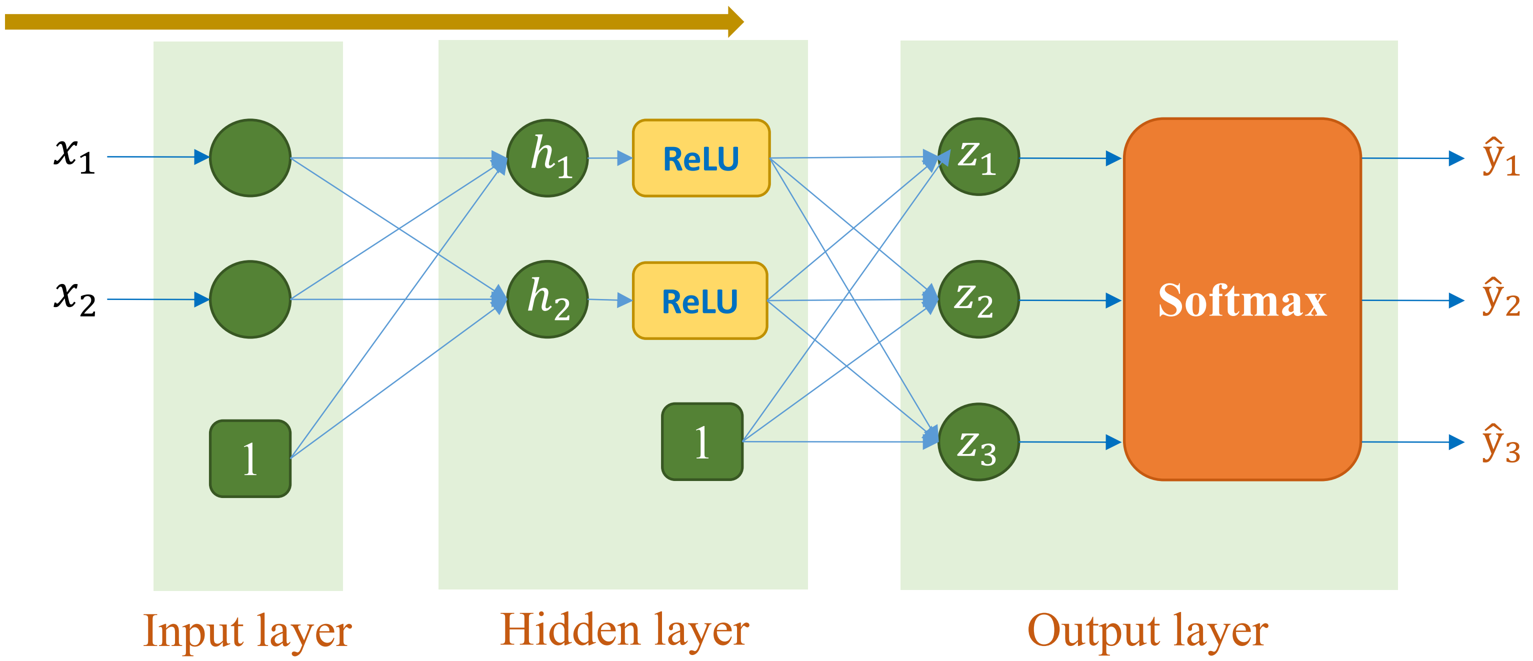
$$\boldsymbol{W}_h = [\boldsymbol{W}_{h1} \quad \boldsymbol{W}_{h2}] = \begin{bmatrix} 0.0 & 0.0 \\ 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$\boldsymbol{W}_z = [\boldsymbol{W}_{z1} \quad \boldsymbol{W}_{z2} \quad \boldsymbol{W}_{z3}] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$\boldsymbol{h} = \boldsymbol{x}\boldsymbol{W}_h = \begin{bmatrix} 1 & 1.5 & 0.2 \\ 1 & 4.7 & 1.6 \\ 1 & 5.6 & 2.2 \end{bmatrix} \begin{bmatrix} 0.0 & 0.0 \\ 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix} = \begin{bmatrix} 1.373 & -1.696 \\ 4.708 & -5.951 \\ 5.731 & -7.281 \end{bmatrix}$$

$$\mathrm{ReLU}(\boldsymbol{h}) = \begin{bmatrix} 1.373 & 0 \\ 4.708 & 0 \\ 5.731 & 0 \end{bmatrix}$$

Feature		Label
Petal_Length	Petal_Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2



$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \boldsymbol{x}^{(3)} \end{bmatrix} = \begin{bmatrix} 1 & 1.5 & 0.2 \\ 1 & 4.7 & 1.6 \\ 1 & 5.6 & 2.2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

$$\boldsymbol{W}_h = [\boldsymbol{W}_{h1} \quad \boldsymbol{W}_{h2}] = \begin{bmatrix} 0.0 & 0.0 \\ 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$\boldsymbol{W}_z = [\boldsymbol{W}_{z1} \quad \boldsymbol{W}_{z2} \quad \boldsymbol{W}_{z3}] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

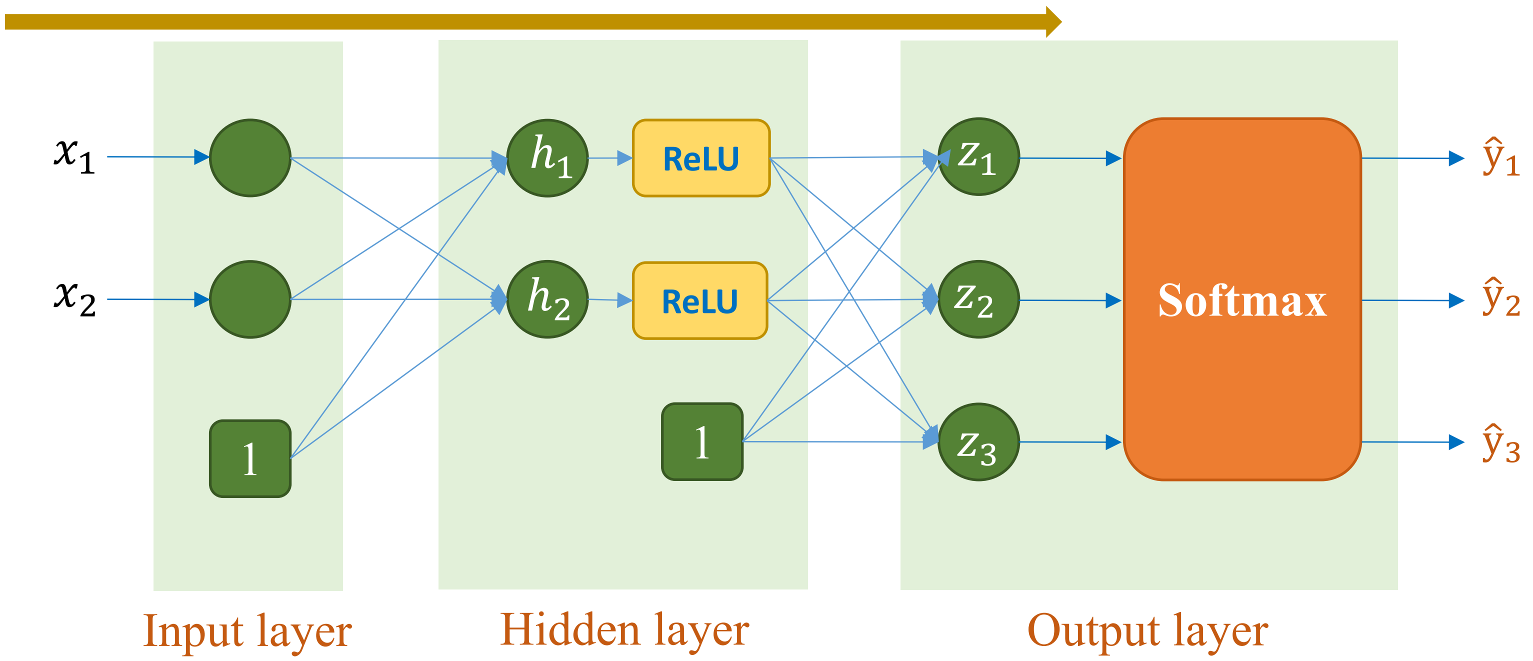
$$\text{ReLU}(\boldsymbol{h}) = \begin{bmatrix} 1.373 & 0 \\ 4.708 & 0 \\ 5.731 & 0 \end{bmatrix}$$

$$[\mathbf{1} \quad \text{ReLU}(\boldsymbol{h})] = \begin{bmatrix} 1 & 1.373 & 0 \\ 1 & 4.708 & 0 \\ 1 & 5.731 & 0 \end{bmatrix}$$

$$\boldsymbol{z} = [\mathbf{1} \quad \text{ReLU}(\boldsymbol{h})]\boldsymbol{W}_z = \begin{bmatrix} 1 & 1.373 & 0 \\ 1 & 4.708 & 0 \\ 1 & 5.731 & 0 \end{bmatrix} \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$= \begin{bmatrix} 0.439 & 0.356 & 0.195 \\ 1.507 & 1.220 & 0.670 \\ 1.835 & 1.485 & 0.816 \end{bmatrix}$$

Feature		Label
Petal_Length	Petal_Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2



$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \boldsymbol{x}^{(3)} \end{bmatrix} = \begin{bmatrix} 1 & 1.5 & 0.2 \\ 1 & 4.7 & 1.6 \\ 1 & 5.6 & 2.2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

$$\boldsymbol{W}_h = [\boldsymbol{W}_{h1} \quad \boldsymbol{W}_{h2}] = \begin{bmatrix} 0.0 & 0.0 \\ 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

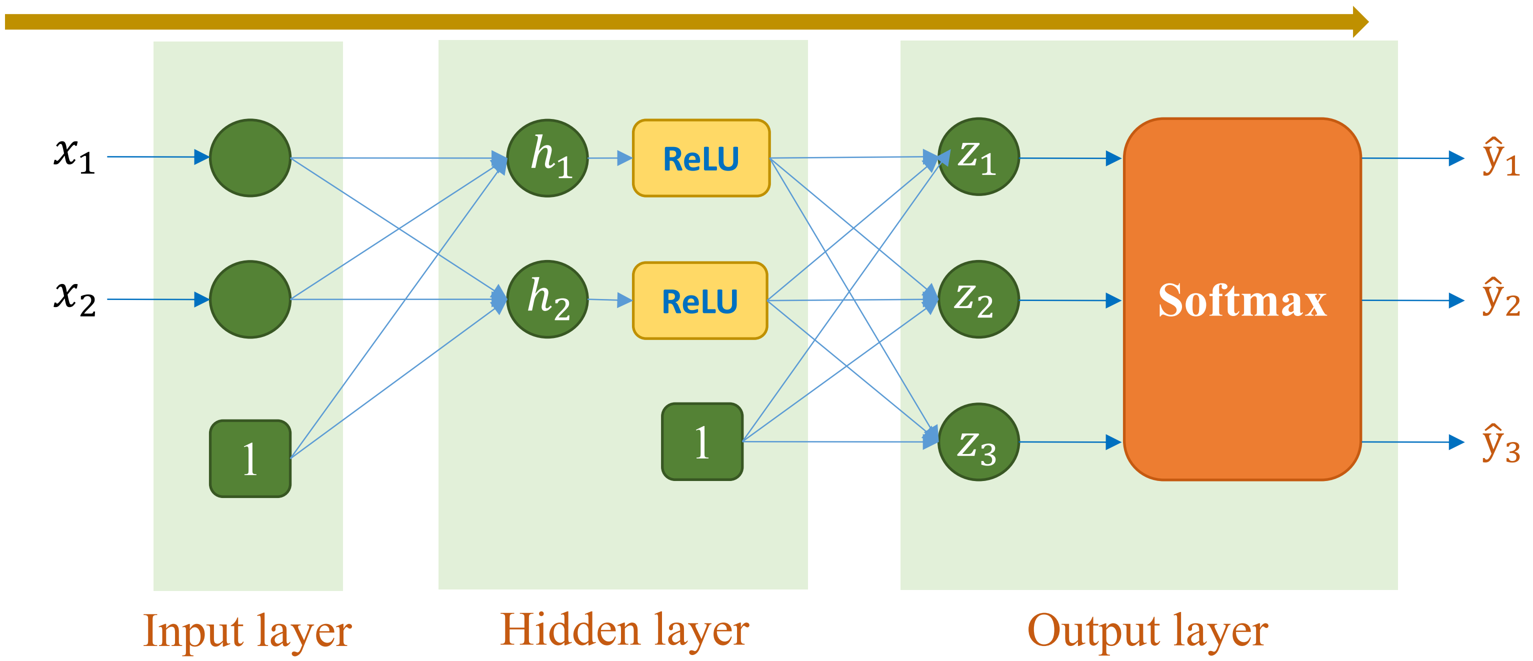
$$\boldsymbol{W}_z = [\boldsymbol{W}_{z1} \quad \boldsymbol{W}_{z2} \quad \boldsymbol{W}_{z3}] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$\boldsymbol{z} = \begin{bmatrix} 0.439 & 0.356 & 0.195 \\ 1.507 & 1.220 & 0.670 \\ 1.835 & 1.485 & 0.816 \end{bmatrix}$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{z}) = \begin{bmatrix} \hat{\boldsymbol{y}}^{(1)} \\ \hat{\boldsymbol{y}}^{(2)} \\ \hat{\boldsymbol{y}}^{(3)} \end{bmatrix} = \begin{bmatrix} 0.369 & 0.340 & 0.289 \\ 0.458 & 0.343 & 0.198 \\ 0.484 & 0.341 & 0.174 \end{bmatrix}$$

loss = 1.269

Feature		Label
Petal_Length	**Petal_Width**	**Label**
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2



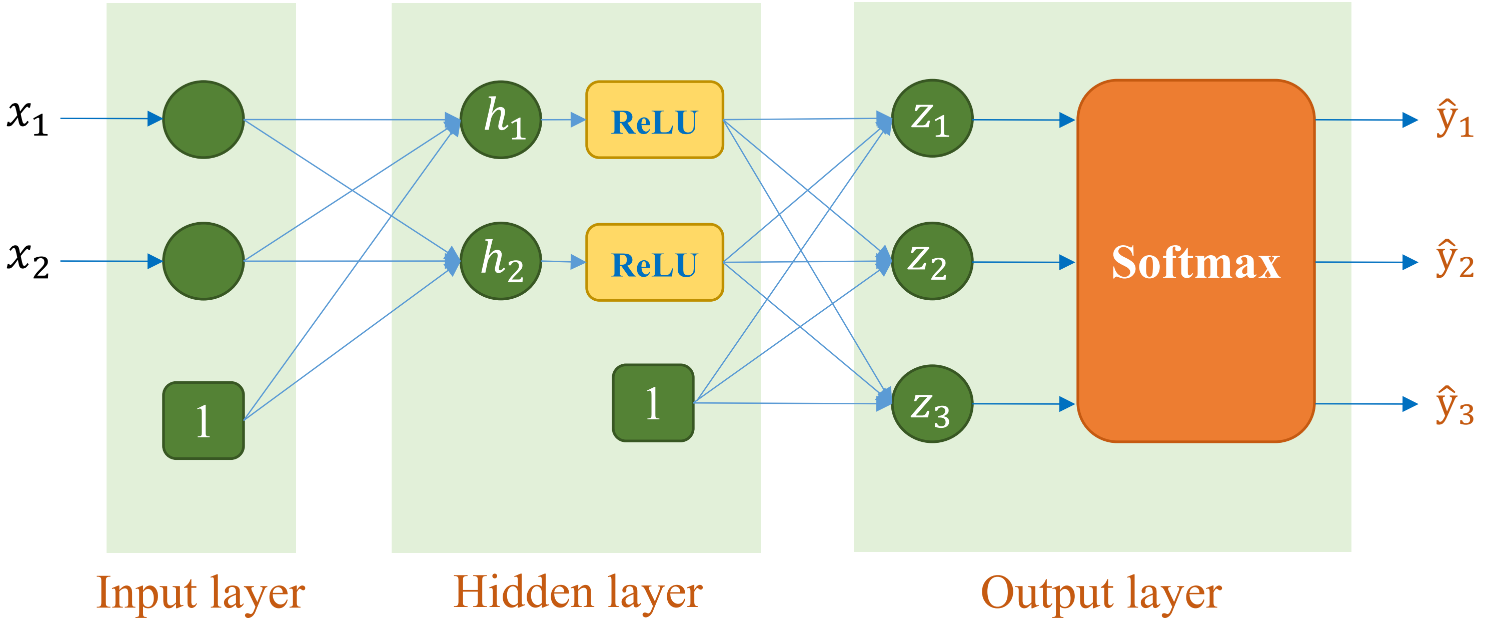
$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \boldsymbol{x}^{(3)} \end{bmatrix} = \begin{bmatrix} 1 & 1.5 & 0.2 \\ 1 & 4.7 & 1.6 \\ 1 & 5.6 & 2.2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

$$\boldsymbol{W}_h = [\boldsymbol{W}_{h1} \quad \boldsymbol{W}_{h2}] = \begin{bmatrix} 0.0 & 0.0 \\ 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$\boldsymbol{W}_z = [\boldsymbol{W}_{z1} \quad \boldsymbol{W}_{z2} \quad \boldsymbol{W}_{z3}] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

Example 2 - Dying ReLU

Petal_Length	Petal_Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2

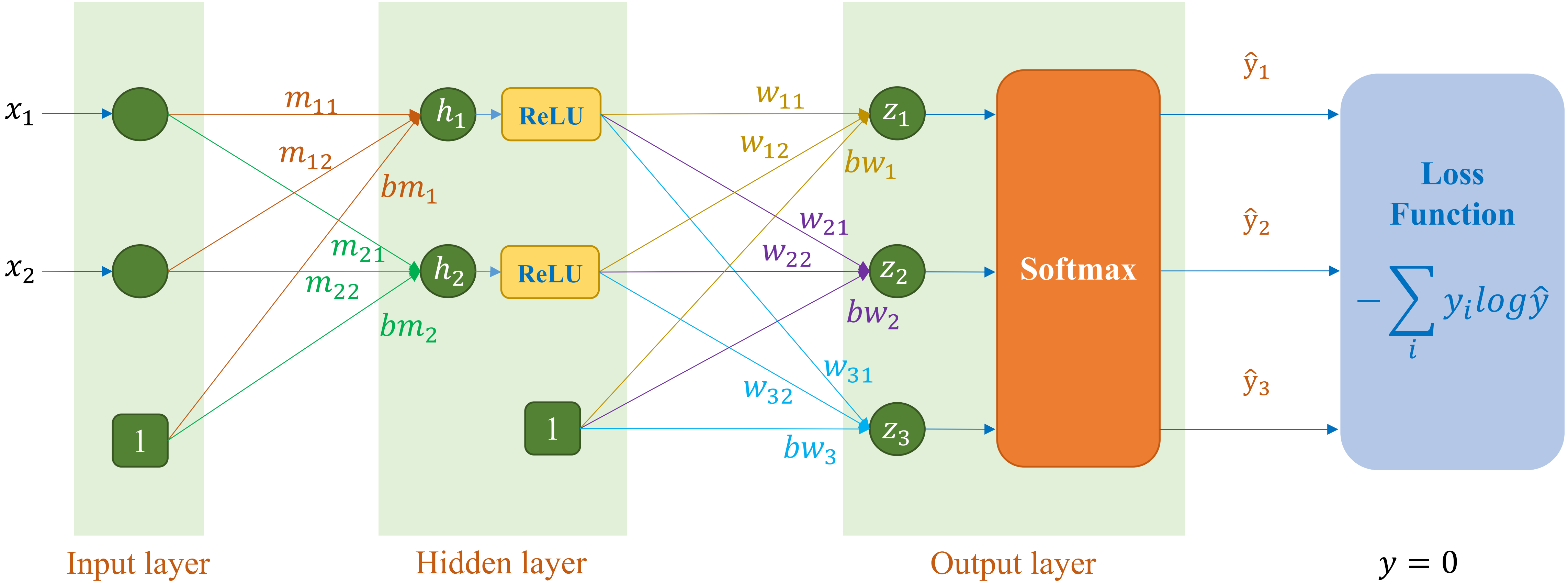


$$\boldsymbol{x} = \begin{bmatrix} 1.5 \\ 0.2 \end{bmatrix} \qquad y = 0$$

$$\boldsymbol{m} = [\boldsymbol{m}_1 \quad \boldsymbol{m}_2] = \begin{bmatrix} 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$\boldsymbol{w} = [\boldsymbol{w}_1 \quad \boldsymbol{w}_2 \quad \boldsymbol{w}_3] = \begin{bmatrix} 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$\boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix} \qquad \boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$



$$\boldsymbol{x} = \begin{bmatrix} 1.5 \\ 0.2 \end{bmatrix}$$

$$\boldsymbol{m} = [\boldsymbol{m_1} \quad \boldsymbol{m}_2] = \begin{bmatrix} 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix}$$

$$\boldsymbol{w} = [\boldsymbol{w}_1 \quad \boldsymbol{w}_2 \quad \boldsymbol{w}_3] = \begin{bmatrix} 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$$

$$\rightarrow \boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix} \qquad \boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

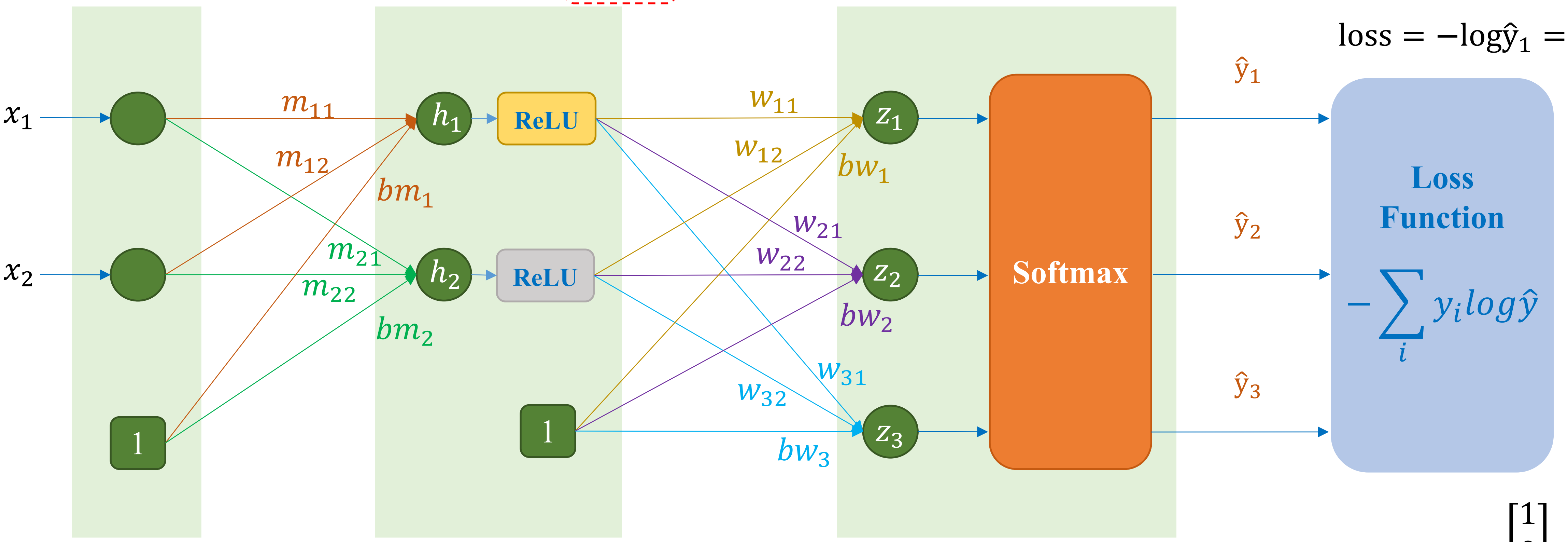
Forward pass

zero value

$$\boldsymbol{h} = \begin{bmatrix} 1.372 \\ -1.68 \end{bmatrix}$$

$$\boldsymbol{x} = \begin{bmatrix} 1.5 \\ 0.2 \end{bmatrix} \qquad \textbf{ReLU} = \begin{bmatrix} 1.372 \\ 0.0 \end{bmatrix} \qquad \boldsymbol{z} = \begin{bmatrix} 0.439 \\ 0.343 \\ 0.192 \end{bmatrix} \qquad \hat{\boldsymbol{y}} = \begin{bmatrix} 0.372 \\ 0.338 \\ 0.290 \end{bmatrix}$$

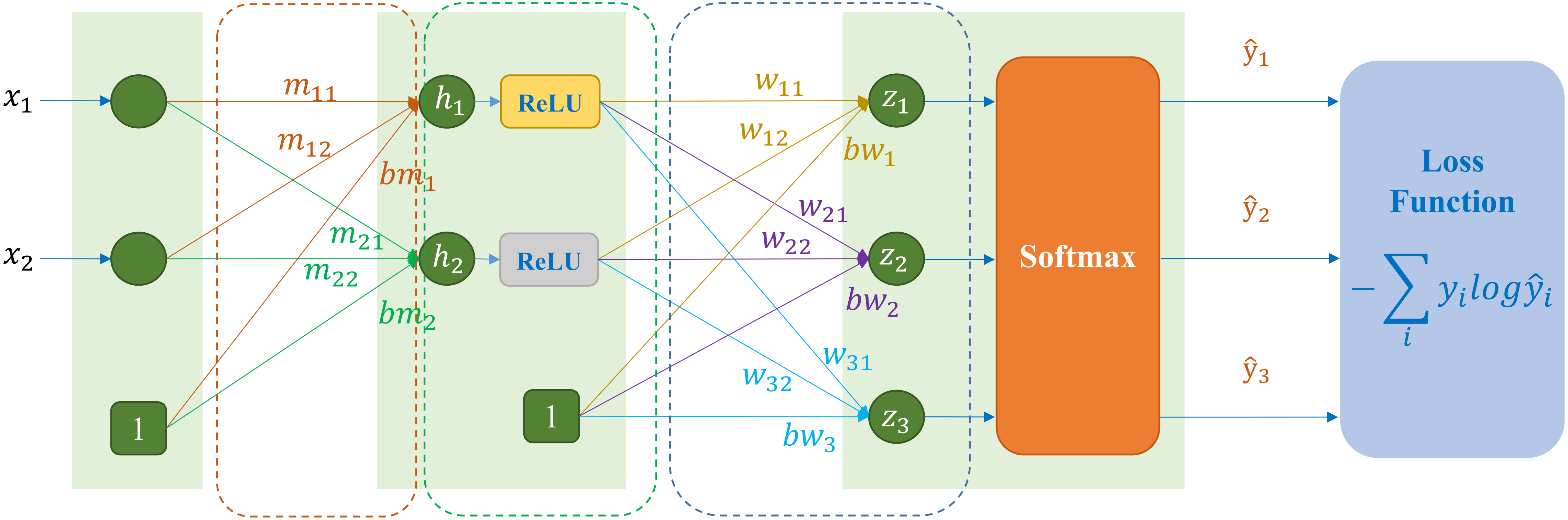
$$\text{loss} = -\log\hat{y}_1 = 0.989$$



$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{m} = [\boldsymbol{m_1} \quad \boldsymbol{m_2}] = \begin{bmatrix} 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix} \qquad \boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{w} = [\boldsymbol{w_1} \quad \boldsymbol{w_2} \quad \boldsymbol{w_3}] = \begin{bmatrix} 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix} \qquad \boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$



$$\frac{\partial L}{\partial m_{jk}} = x_k \frac{\partial L}{\partial h_j}$$

$$\frac{\partial L}{\partial bm_j} = \frac{\partial L}{\partial h_j}$$

$$\frac{\partial L}{\partial relu_j} = \sum_i w_{ij} \frac{\partial L}{\partial z_i}$$

$$\text{ReLU}'(h_j) = \begin{cases} 0 & \text{if } h_j \leq 0 \\ 1 & \text{if } h_j > 0 \end{cases}$$

$$\frac{\partial L}{\partial h_j} = \begin{cases} 0 & \text{if } h_j \leq 0 \\ \frac{\partial L}{\partial relu_j} & \text{if } h_j > 0 \end{cases}$$

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

$$\frac{\partial L}{\partial w_{ij}} = \text{ReLU}_j \frac{\partial L}{\partial z_i}$$

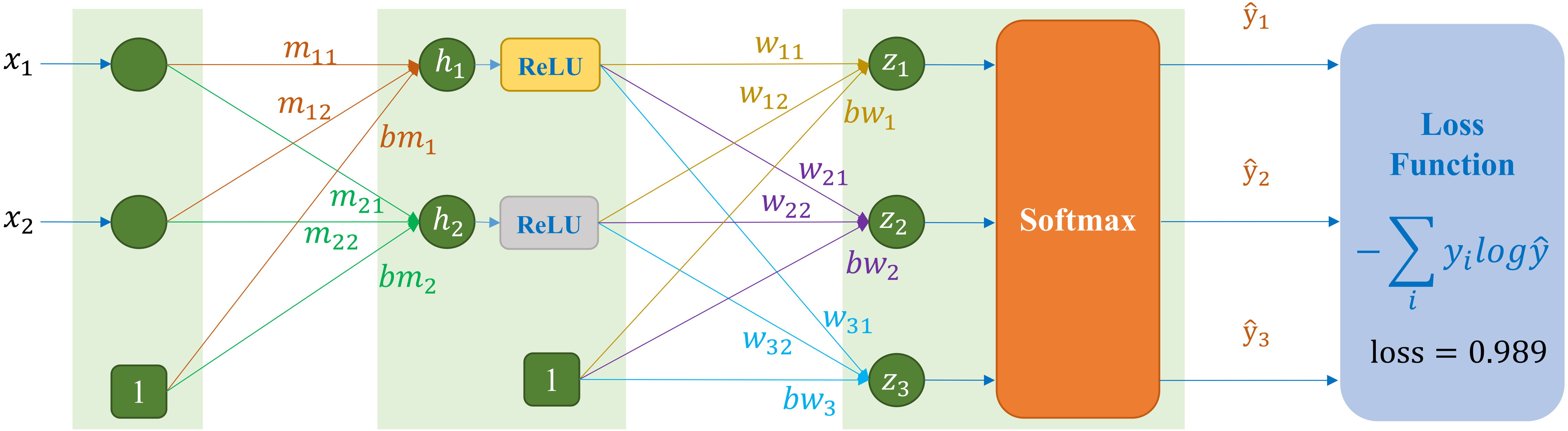
$$\frac{\partial L}{\partial bw_i} = \frac{\partial L}{\partial z_i}$$

Backward pass

Backward pass

$$\boldsymbol{x} = \begin{bmatrix}1.5\\0.2\end{bmatrix} \quad \boldsymbol{h} = \begin{bmatrix}1.372\\-1.68\end{bmatrix} \quad \textbf{ReLU} = \begin{bmatrix}1.372\\0.0\end{bmatrix} \quad \boldsymbol{z} = \begin{bmatrix}0.439\\0.343\\0.192\end{bmatrix} \quad \hat{\boldsymbol{y}} = \begin{bmatrix}0.372\\0.338\\0.290\end{bmatrix}$$

$$\boldsymbol{m} = \begin{bmatrix}0.86 & -1.04\\0.41 & -0.65\end{bmatrix} \quad \boldsymbol{bm} = \begin{bmatrix}0.0\\0.0\end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix}0.32 & 0.25 & 0.14\\-0.47 & -1.06 & 0.063\end{bmatrix} \quad \boldsymbol{bw} = \begin{bmatrix}0.0\\0.0\\0.0\end{bmatrix}$$



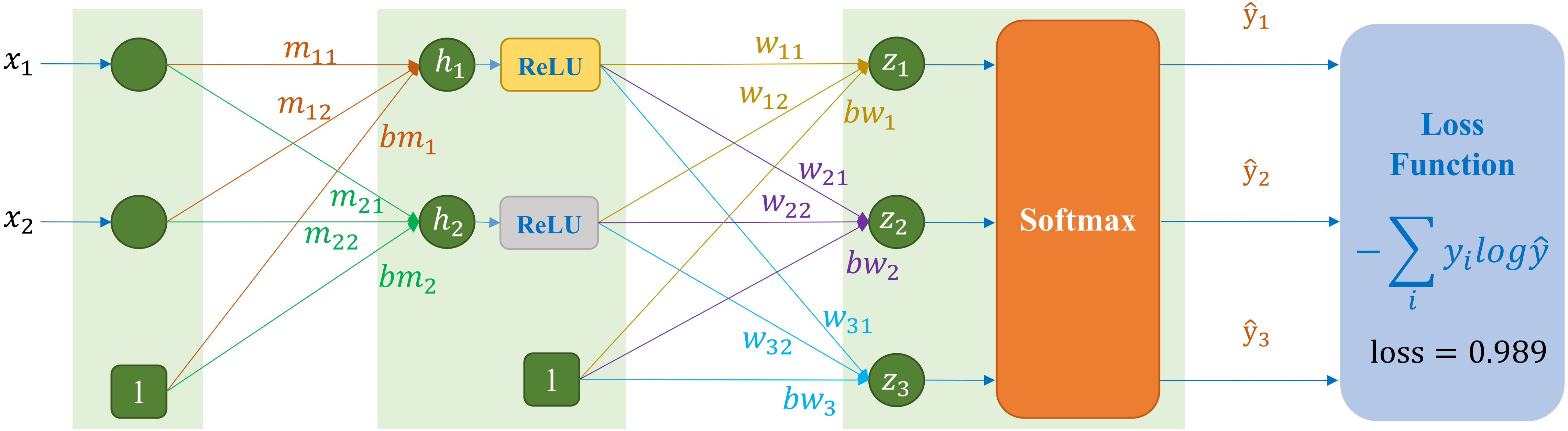
$$\frac{\partial L}{\partial relu_j} = \sum_i w_{ij}\frac{\partial L}{\partial z_i} \qquad \frac{\partial L}{\partial w_{ij}} = \text{ReLU}_j\frac{\partial L}{\partial z_i} \qquad \frac{\partial L}{\partial bw_i} = \frac{\partial L}{\partial z_i} \qquad \frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

$$\boldsymbol{\nabla}_{\textbf{ReLU}}L = \begin{bmatrix}-0.0759\\-0.0445\end{bmatrix} \quad \boldsymbol{\nabla_w}L = \begin{bmatrix}-0.861 & 0.463 & 0.398\\0.0 & 0.0 & 0.0\end{bmatrix} \quad \boldsymbol{\nabla_{bw}}L = \begin{bmatrix}-0.628\\0.338\\0.290\end{bmatrix} \quad \boldsymbol{\nabla_z}L = \begin{bmatrix}-0.628\\0.338\\0.290\end{bmatrix}$$

Backward pass

$$\boldsymbol{x} = \begin{bmatrix}1.5\\0.2\end{bmatrix} \quad \boldsymbol{h} = \begin{bmatrix}1.372\\-1.68\end{bmatrix} \quad \textbf{ReLU} = \begin{bmatrix}1.372\\0.0\end{bmatrix} \quad \boldsymbol{z} = \begin{bmatrix}0.439\\0.343\\0.192\end{bmatrix} \quad \hat{\boldsymbol{y}} = \begin{bmatrix}0.372\\0.338\\0.290\end{bmatrix}$$

$$\boldsymbol{m} = \begin{bmatrix}0.86 & -1.04\\0.41 & -0.65\end{bmatrix} \quad \boldsymbol{bm} = \begin{bmatrix}0.0\\0.0\end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix}0.32 & 0.25 & 0.14\\-0.47 & -1.06 & 0.063\end{bmatrix} \quad \boldsymbol{bw} = \begin{bmatrix}0.0\\0.0\\0.0\end{bmatrix}$$



$$\frac{\partial L}{\partial m_{jk}} = x_k \frac{\partial L}{\partial h_j}$$

$$\boldsymbol{\nabla}_{\boldsymbol{m}} L = \begin{bmatrix}-0.114 & 0.0\\-0.015 & 0.0\end{bmatrix}$$

$$\frac{\partial L}{\partial bm_j} = \frac{\partial L}{\partial h_j}$$

$$\boldsymbol{\nabla}_{\boldsymbol{bm}} L = \begin{bmatrix}-0.0759\\0.0\end{bmatrix}$$

$$\frac{\partial L}{\partial h_j} = \begin{cases} 0 & \text{if } \; h_j \leq 0 \\ \dfrac{\partial L}{\partial relu_j} & \text{if } \; h_j > 0 \end{cases}$$

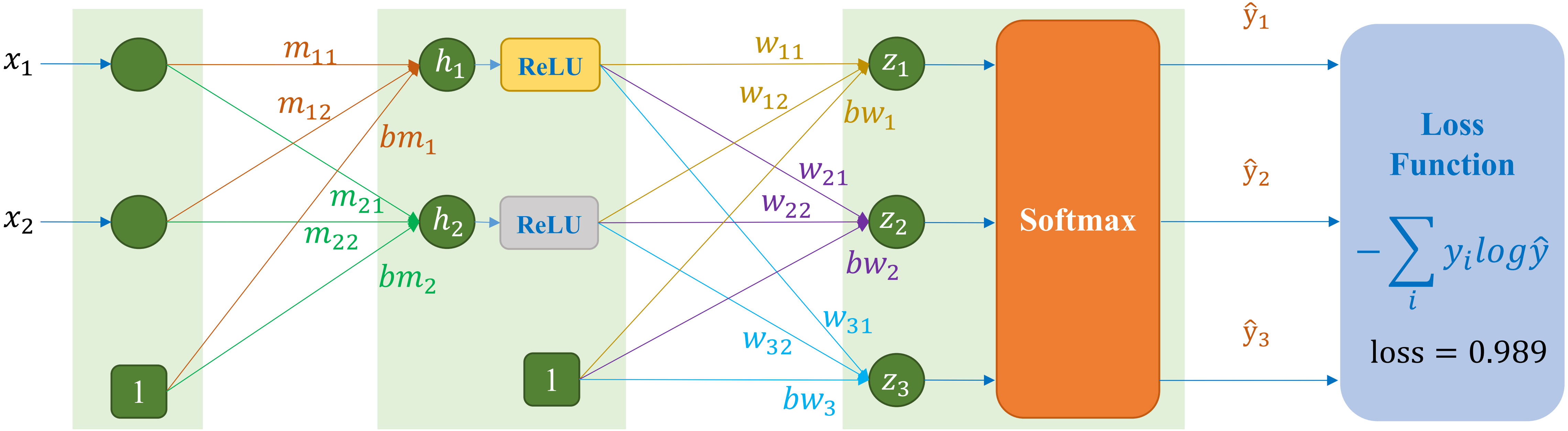
$$\boldsymbol{\nabla}_{\mathbf{h}} L = \begin{bmatrix}-0.0759\\0.0\end{bmatrix}$$

$$\frac{\partial L}{\partial relu_j} = \sum_i w_{ij} \frac{\partial L}{\partial z_i}$$

$$\boldsymbol{\nabla}_{\textbf{ReLU}} L = \begin{bmatrix}-0.0759\\-0.0445\end{bmatrix}$$

$$\boldsymbol{m} = \begin{bmatrix} 0.86 & -1.04 \\ 0.41 & -0.65 \end{bmatrix} \quad \boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} 0.32 & 0.25 & 0.14 \\ -0.47 & -1.06 & 0.063 \end{bmatrix} \quad \boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

$$\nabla_{\boldsymbol{m}} L = \begin{bmatrix} -0.114 & 0.0 \\ -0.015 & 0.0 \end{bmatrix} \quad \nabla_{\boldsymbol{bm}} L = \begin{bmatrix} -0.0759 \\ 0.0 \end{bmatrix} \quad \nabla_{\boldsymbol{w}} L = \begin{bmatrix} -0.628 & 0.338 & 0.29 \\ 0.0 & 0.0 & 0.0 \end{bmatrix} \quad \nabla_{\boldsymbol{bw}} L = \begin{bmatrix} -0.628 \\ 0.338 \\ 0.290 \end{bmatrix}$$



Update the parameters with $\eta = 0.01$

$$\boldsymbol{m} = \begin{bmatrix} 0.861 & -1.04 \\ 0.4105 & -0.65 \end{bmatrix} \quad \boldsymbol{bm} = \begin{bmatrix} 0.000759 \\ 0.0 \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} 0.328 & 0.245 & 0.136 \\ -0.47 & -1.06 & 0.063 \end{bmatrix} \quad \boldsymbol{bw} = \begin{bmatrix} 0.0062 \\ -0.0033 \\ -0.0029 \end{bmatrix}$$

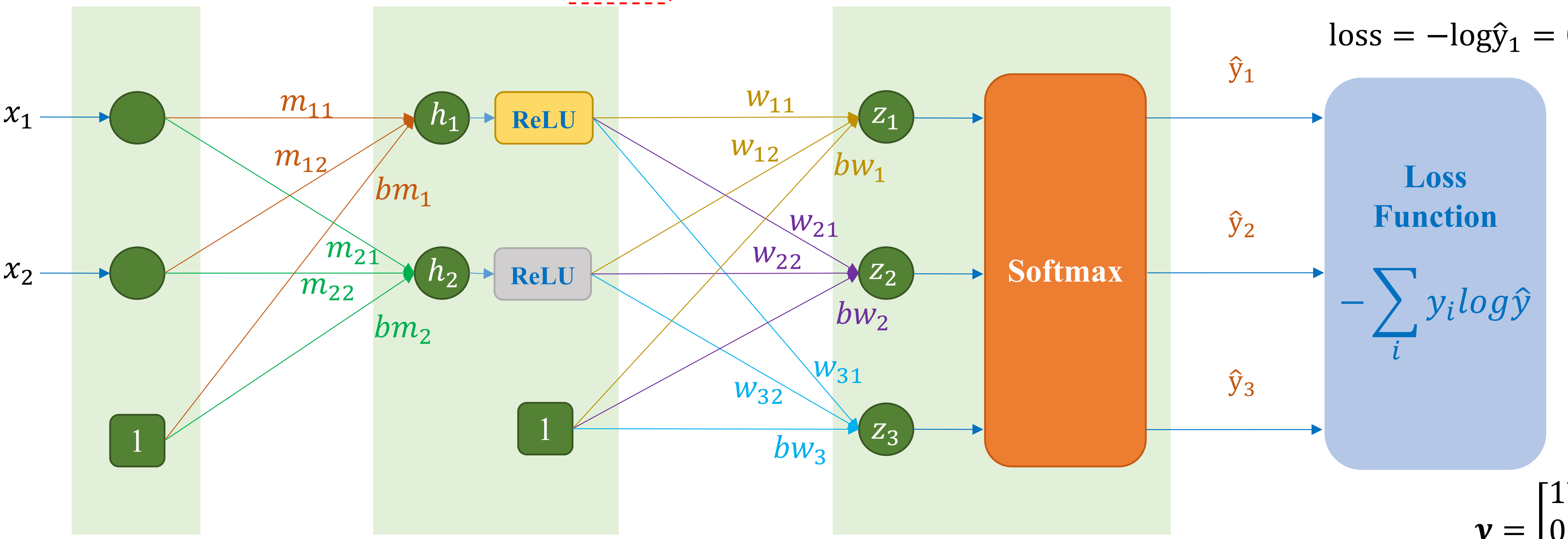
Forward pass again

$\boldsymbol{h} = \begin{bmatrix} 1.374 \\ -1.68 \end{bmatrix}$

still zero value

$\boldsymbol{x} = \begin{bmatrix} 1.5 \\ 0.2 \end{bmatrix}$ $\quad \textbf{ReLU} = \begin{bmatrix} 1.374 \\ \boxed{0.0} \end{bmatrix}$ $\quad \boldsymbol{z} = \begin{bmatrix} 0.458 \\ 0.334 \\ 0.184 \end{bmatrix}$ $\quad \hat{\boldsymbol{y}} = \begin{bmatrix} 0.378 \\ 0.334 \\ 0.287 \end{bmatrix}$

$\text{loss} = -\log \hat{y}_1 = 0.972$



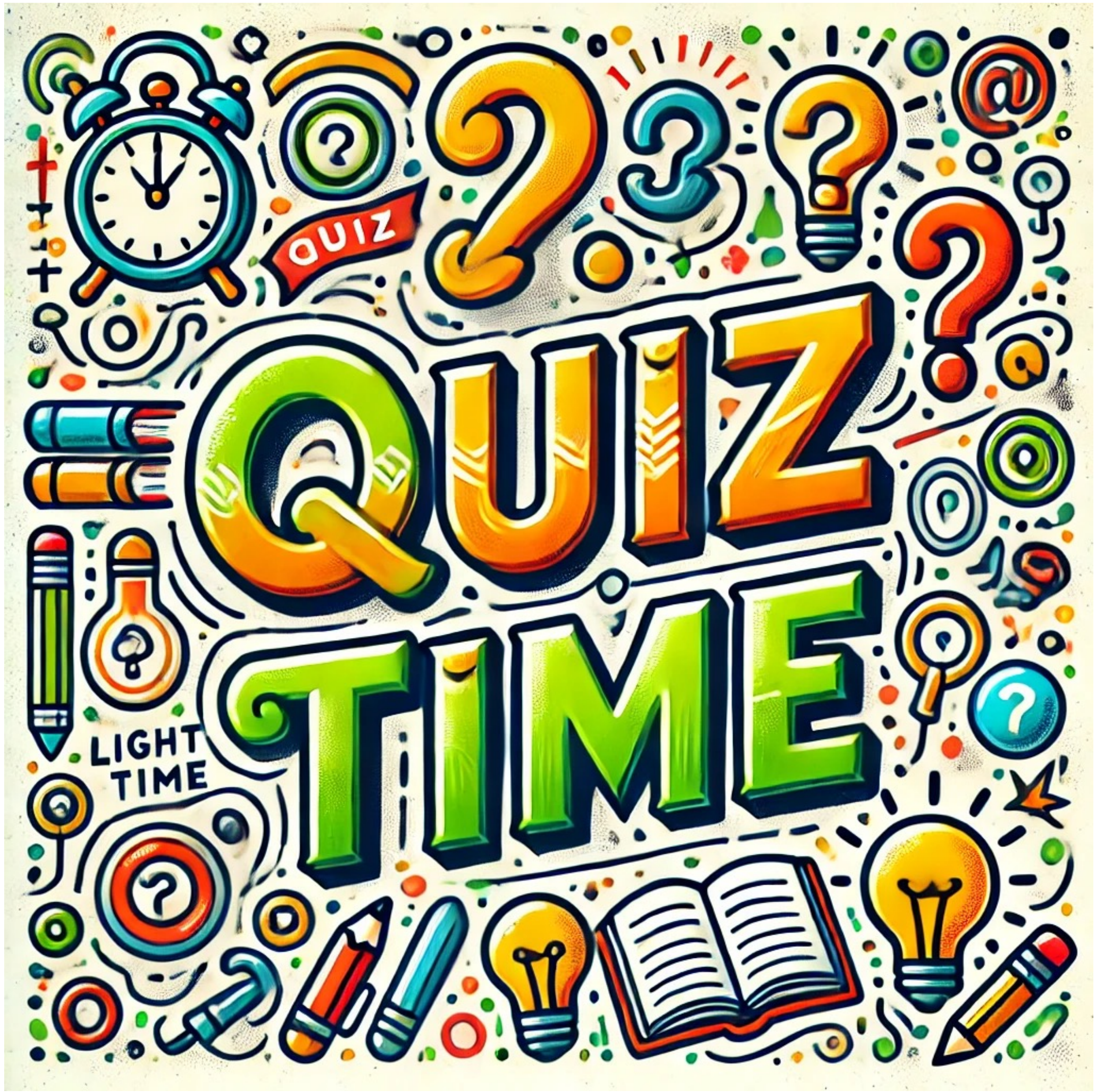
$\boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

$\boldsymbol{m} = [\boldsymbol{m_1} \quad \boldsymbol{m}_2] = \begin{bmatrix} 0.861 & -1.04 \\ 0.4105 & -0.65 \end{bmatrix}$

$\boldsymbol{bm} = \begin{bmatrix} 0.000759 \\ 0.0 \end{bmatrix}$

$\boldsymbol{w} = [\boldsymbol{w}_1 \quad \boldsymbol{w}_2 \quad \boldsymbol{w}_3] = \begin{bmatrix} 0.328 & 0.245 & 0.136 \\ -0.47 & -1.06 & 0.063 \end{bmatrix}$

$\boldsymbol{bw} = \begin{bmatrix} 0.0062 \\ -0.0033 \\ -0.0029 \end{bmatrix}$


QUIZ
QUIZ
TIME
LIGHT
TIME

Question 1

❖ Chuẩn hóa dữ liệu ảnh nào có giá trị trung bình của data bằng 0 (chọn nhiều đáp án)?

a) Sau chuẩn hóa có range là [0, 255]

b) Có range là [0, 1]

c) Có range là [-1, 1]

d) Dạng z-score

Question 2

❖ Code nào chuẩn hóa data và kết quả thuộc đoạn [0, 255]?

1
```
Compose([transforms.ToTensor(), transforms.Normalize((0.5,), (0.5,))])
```

2
```
Compose([transforms.ToTensor(), transforms.Normalize((0,), (1.0,))])
```

3
```
Compose([transforms.ToTensor(), transforms.Normalize((mean,), (std,))])
```

a) Code 1

b) Code 2

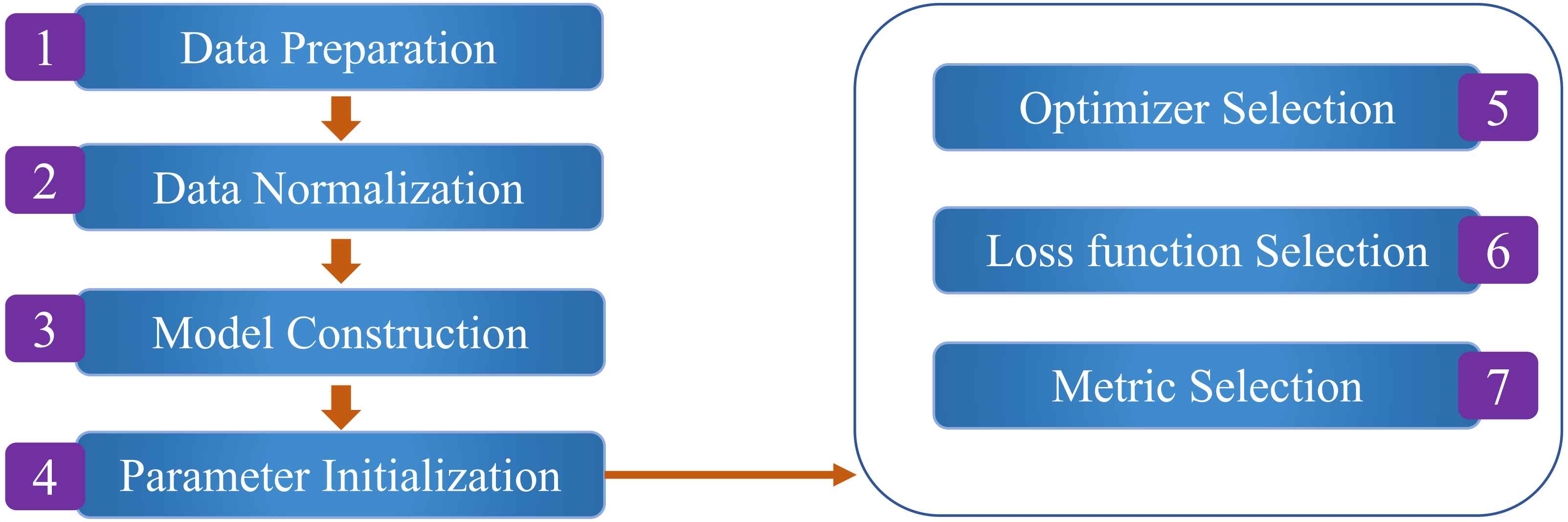
c) Code 3

d) Không code nào ở trên

AI VIET NAM
@aivietnam.edu.vn

Question 3

❖ Chọn 2 thành phần ít quan trọng nhất từ hình pipeline huấn luyện sau?



a) Thành phần (1) hoặc (2)

b) Thành phần (3) hoặc (4)

c) Thành phần (5) hoặc (6)

d) Thành phần (7)

Question 4

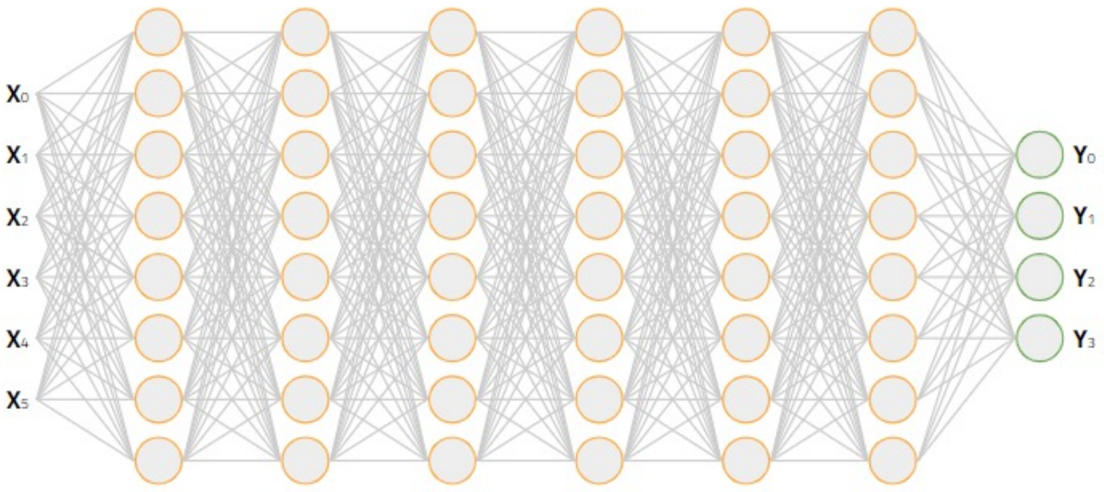
❖ Activation nào không nên dùng cho mô hình sau?

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$$

$$\tanh(x) = \frac{2}{1+e^{-2x}} - 1$$

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$$

$$\text{GELU(x)} = \text{x}\phi(\text{x}) \approx \text{x} * \text{sigmoid}(1.702\text{x})$$



a) Sigmoid(.)

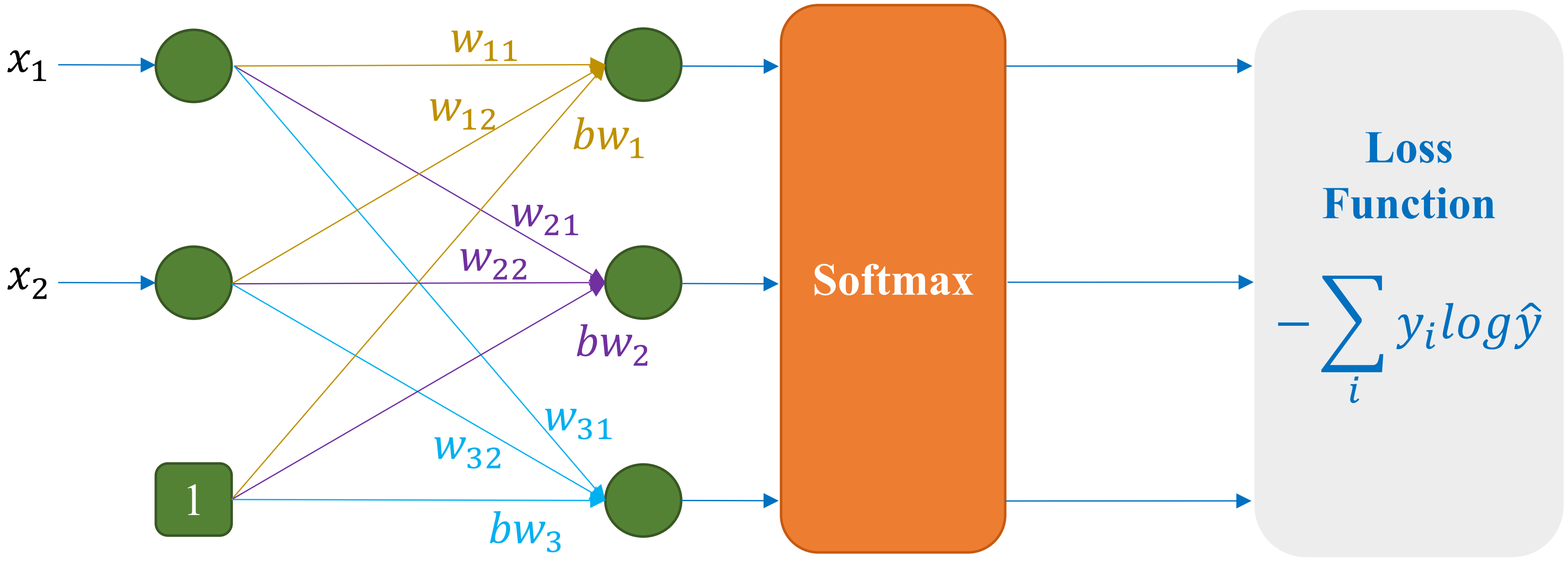
b) Tanh(.)

c) ReLU(.)

d) GELU(.)

Question 5

❖ Khởi tạo tất cả tham số của model sau đều bằng 0. Việc huấn luyện mô hình sẽ như thế nào?



a) Vẫn huấn luyện được

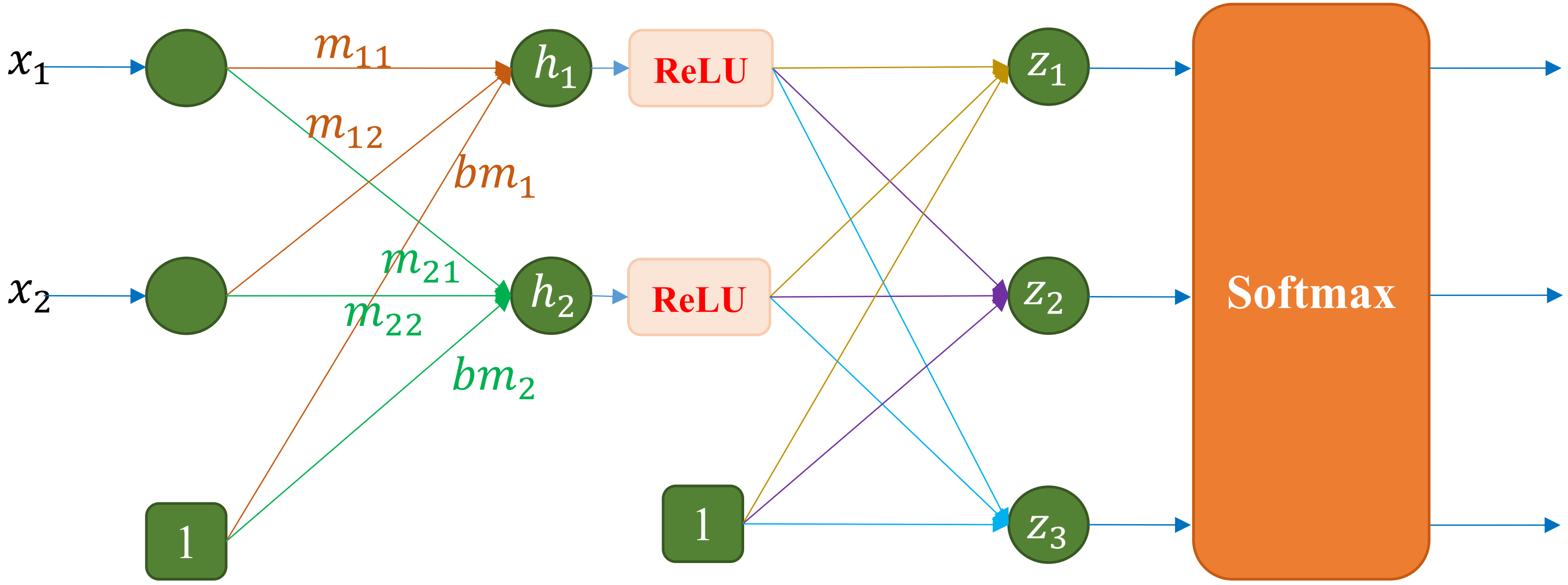
b) Không huấn luyện được

c) Không xác định được

d) Các câu trả lời trên đều sai

Question 6

❖ Khởi tạo tất cả tham số của model sau đều bằng 0. Việc huấn luyện mô hình sẽ như thế nào?



a) Vẫn huấn luyện được

b) Không huấn luyện được

c) Không xác định được

d) Các câu trả lời trên đều sai

Outline

SECTION 1

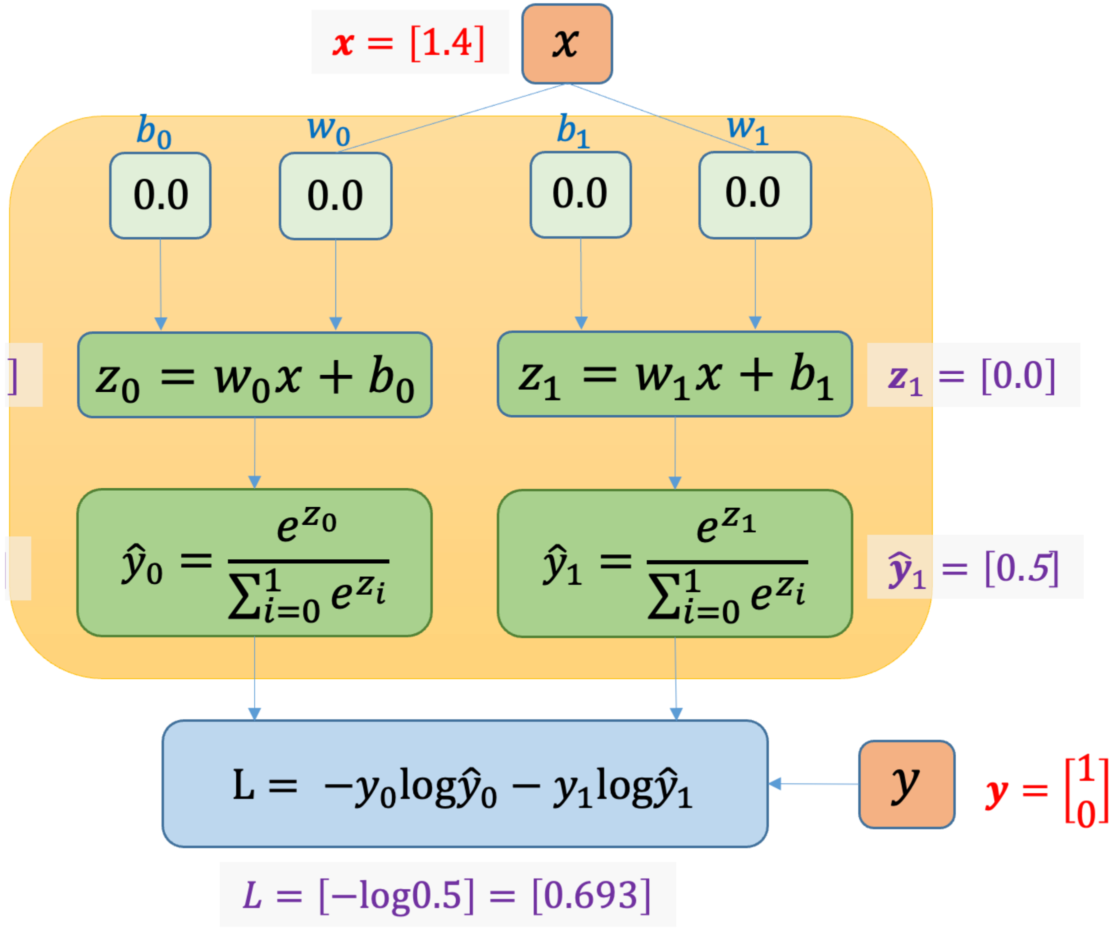
MLP Insight

SECTION 2

MLP Examples

SECTION 3

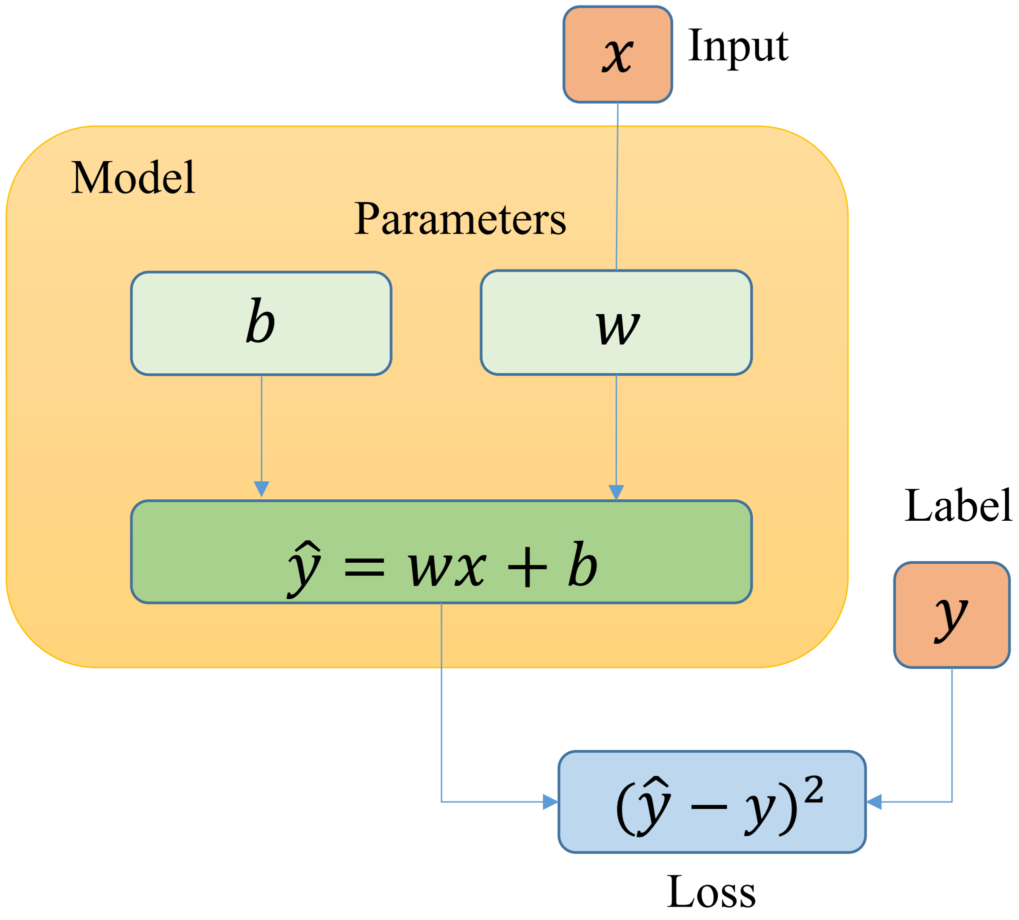
Initialization Examples



Example 3 - Zero Initialization

❖ Linear regression

Diagram



Cheat sheet

Compute the output $\hat{y}$

$$\hat{y} = wx + b$$

Compute the loss

$$L = (\hat{y} - y)^2$$

Compute derivative

$$L'_w = 2x(\hat{y} - y)$$

$$L'_b = 2(\hat{y} - y)$$

Update parameters

$$w = w - \eta L'_w$$

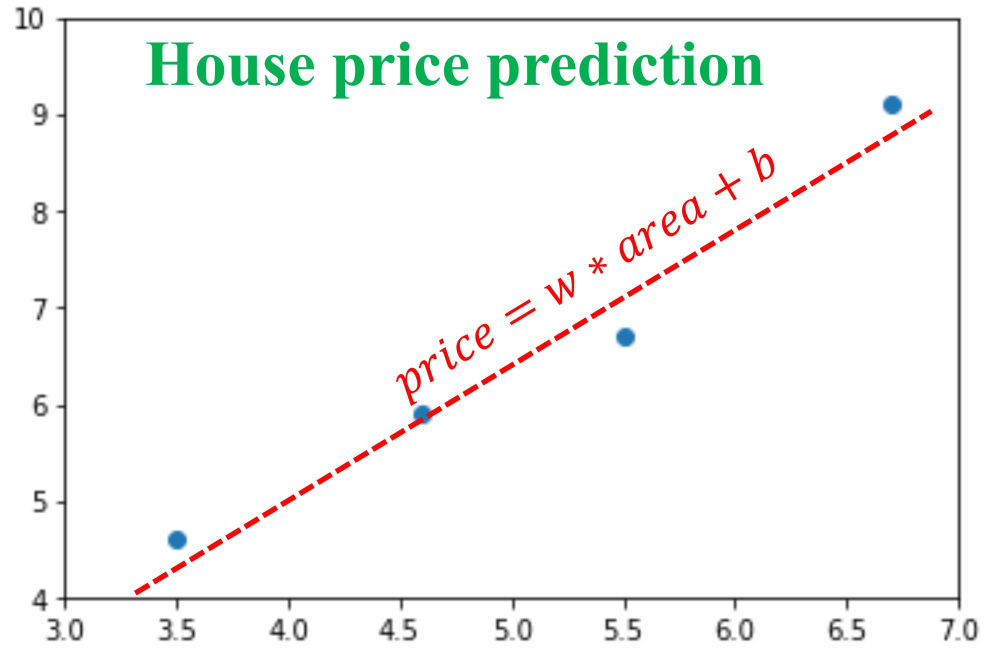
$$b = b - \eta L'_b$$

Example 3 - Zero Initialization

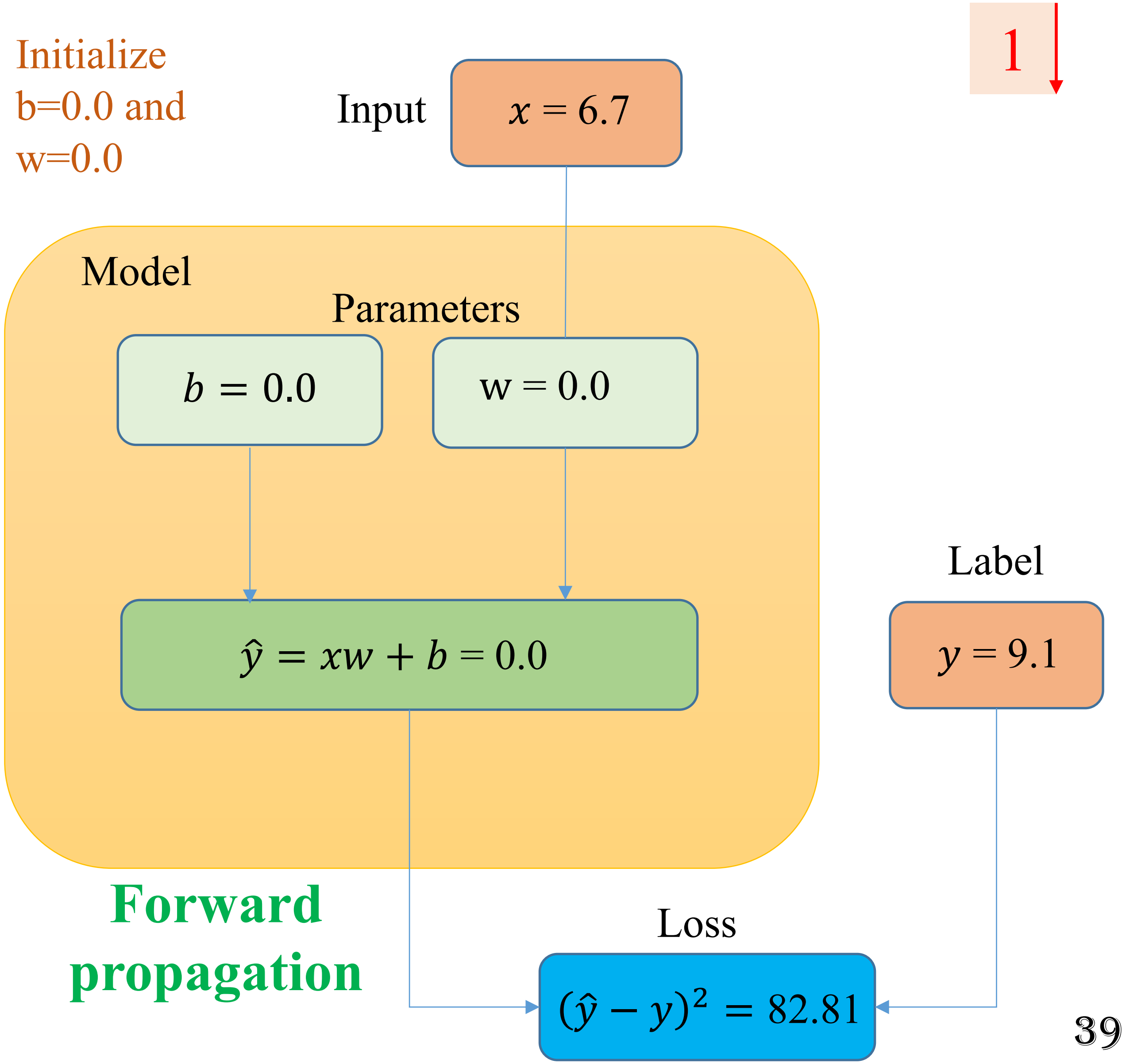
Given sample data

	area	price
	6.7	9.1
	4.6	5.9
	3.5	4.6
	5.5	6.7

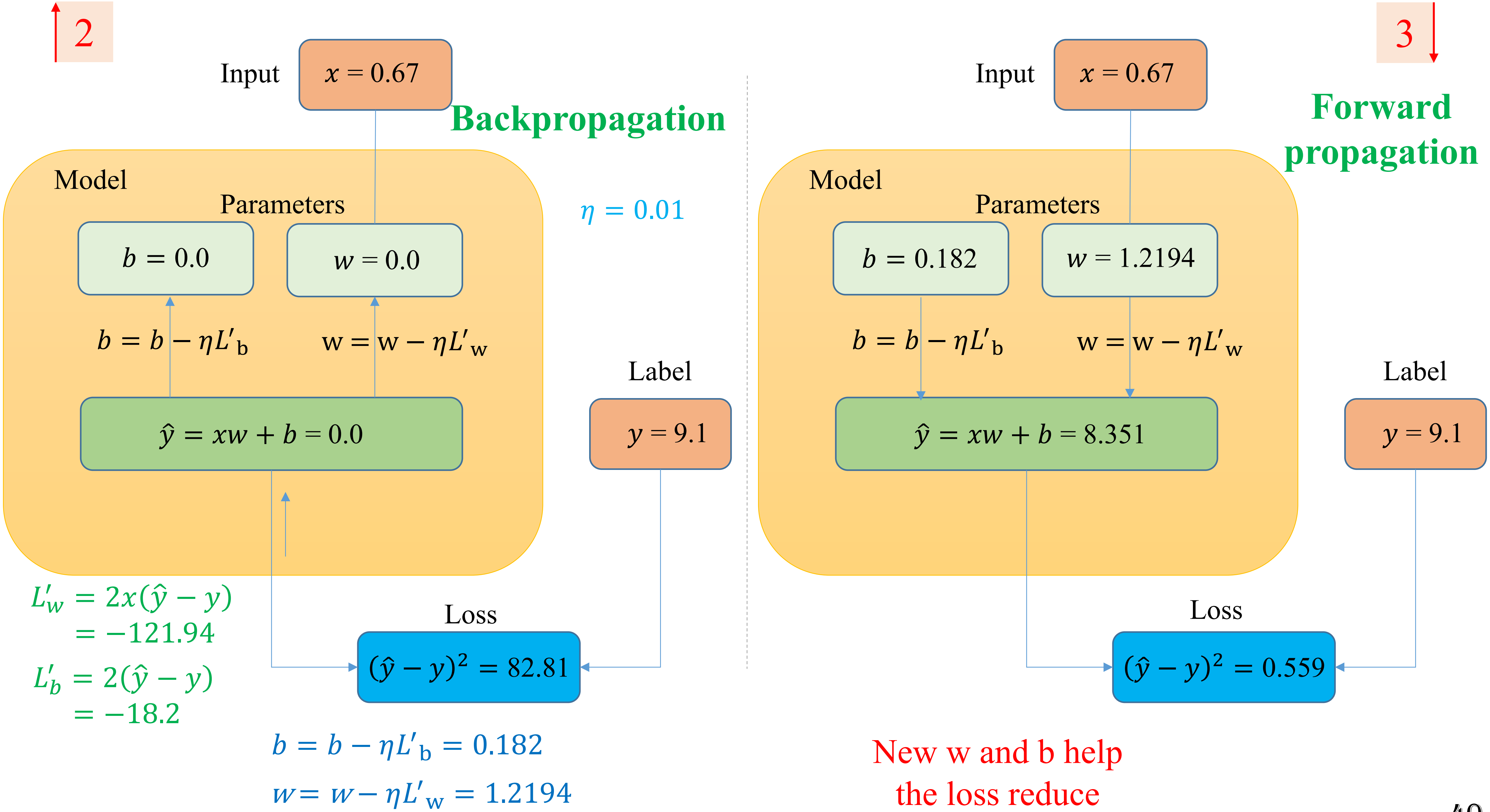
Feature: area; Label: price



Initialize b=0.0 and w=0.0



Forward propagation


2
Input
$x = 0.67$
Backpropagation
Model
Parameters
$b = 0.0$
$w = 0.0$
$\eta = 0.01$
$b = b - \eta L'_b$
$w = w - \eta L'_w$
Label
$\hat{y} = xw + b = 0.0$
$y = 9.1$
$L'_w = 2x(\hat{y} - y)$
$= -121.94$
Loss
$(\hat{y} - y)^2 = 82.81$
$L'_b = 2(\hat{y} - y)$
$= -18.2$
$b = b - \eta L'_b = 0.182$
$w = w - \eta L'_w = 1.2194$
3
Input
$x = 0.67$
Forward propagation
Model
Parameters
$b = 0.182$
$w = 1.2194$
$b = b - \eta L'_b$
$w = w - \eta L'_w$
Label
$\hat{y} = xw + b = 8.351$
$y = 9.1$
Loss
$(\hat{y} - y)^2 = 0.559$
New w and b help
the loss reduce

Example 4 - Zero Initialization

❖ Logistic regression

$$\boldsymbol{\theta}^T = [b \quad w_1 \quad w_2]$$

$$\boldsymbol{x}^T = [1 \quad x_1 \quad x_2]$$

1) Pick a sample (x, y) from training data

2) Compute output $\hat{y}$

$$z = \boldsymbol{\theta}^T \boldsymbol{x}$$

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

3) Compute loss

$$L(\boldsymbol{\theta}) = (-\text{y}\log\hat{\text{y}} - (1-\text{y})\log(1-\hat{\text{y}}\ ))$$

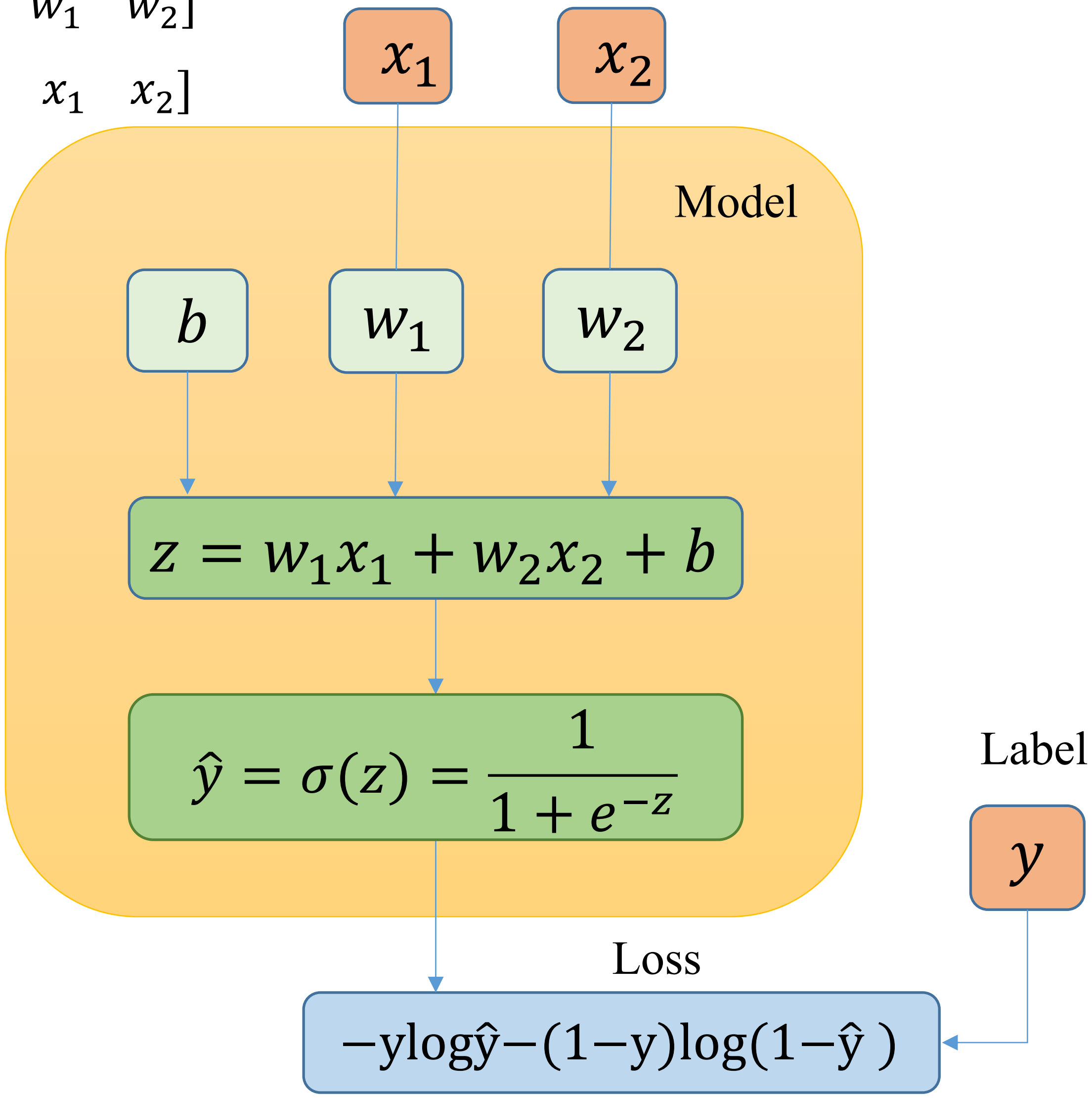
4) Compute derivative

$$\nabla_{\boldsymbol{\theta}} L = \mathbf{x}(\hat{\text{y}} - y)$$

5) Update parameters

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \eta L'_{\boldsymbol{\theta}}$$

η is learning rate

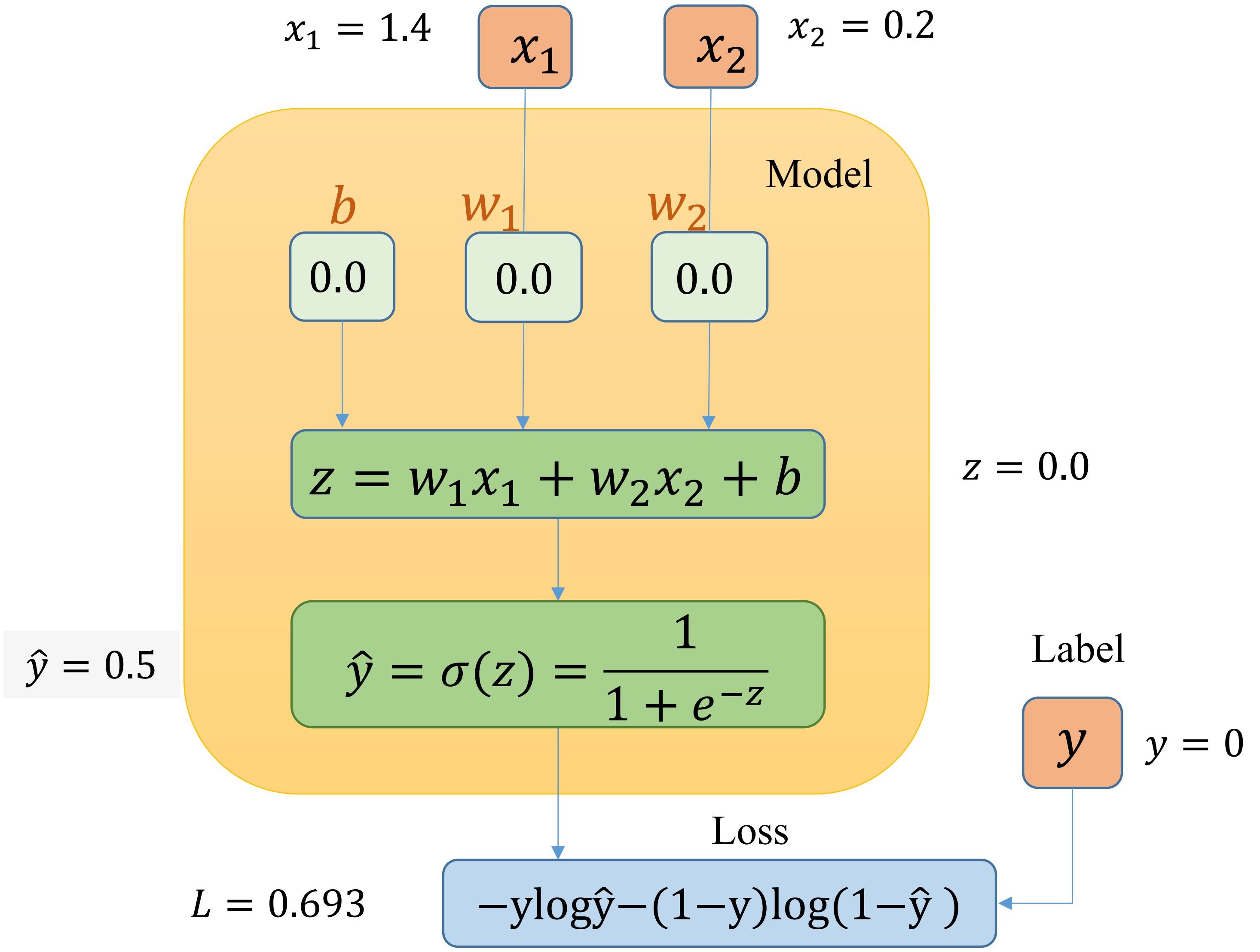


Example 4 - Zero Initialization

Dataset

Petal_Length	Petal_Width	Label
1.4	0.2	0
1.5	0.2	0
3	1.1	1
4.1	1.3	1

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \qquad \boldsymbol{y} = [0]$$



Example 4 - Zero Initialization

Dataset

Petal_Length	Petal_Width	Label
1.4	0.2	0
1.5	0.2	0
3	1.1	1
4.1	1.3	1

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \qquad \boldsymbol{y} = [0]$$

$\eta = 0.01$

$b = -0.005$

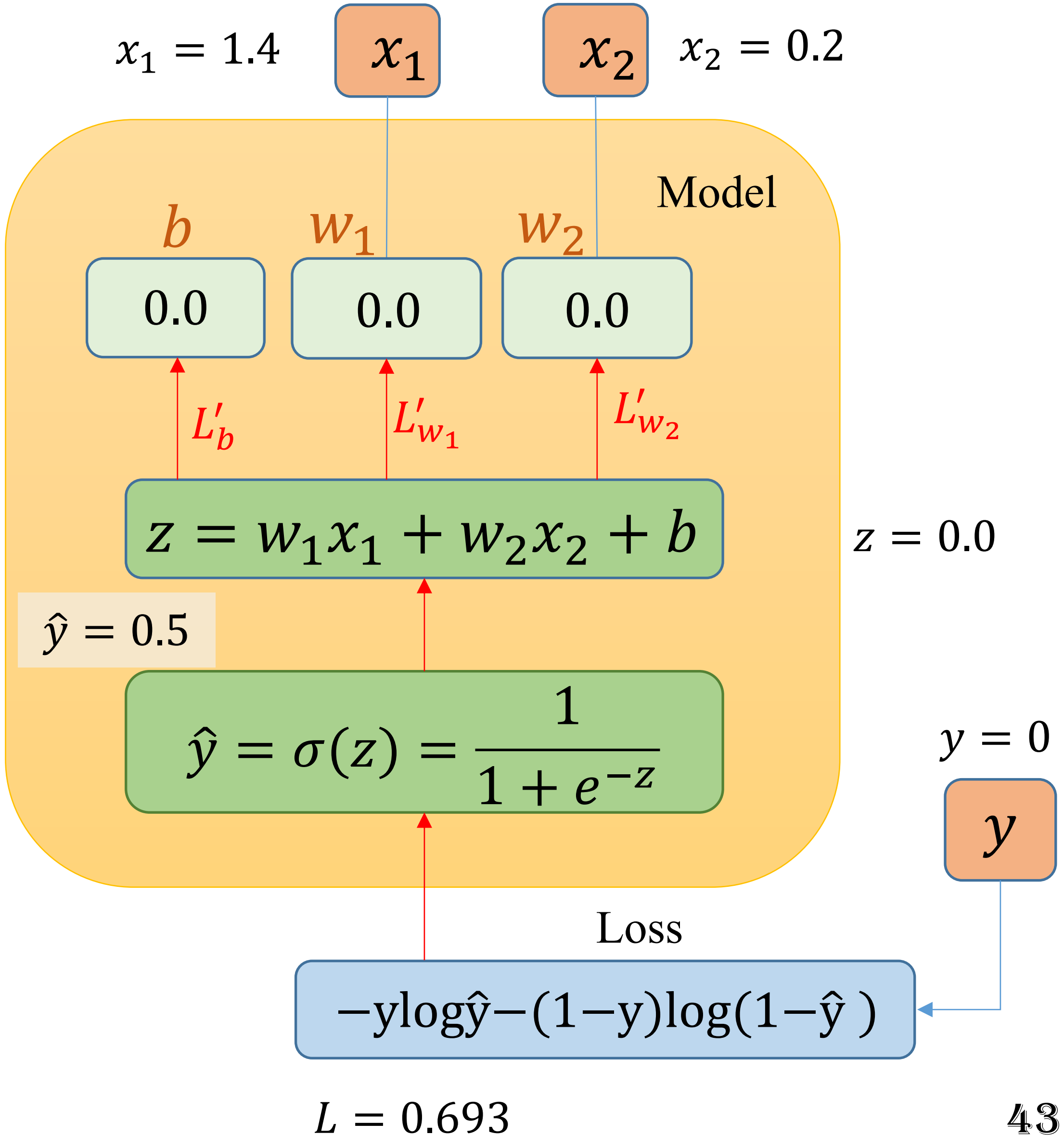
$w_1 = -0.007$

$w_2 = -0.001$

$$L'_{\boldsymbol{\theta}} = \mathbf{x}(\hat{y} - y)$$

$$= \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} [0.5]$$

$$= \begin{bmatrix} 0.5 \\ 0.7 \\ 0.1 \end{bmatrix} = \begin{bmatrix} L'_b \\ L'_{w_1} \\ L'_{w_2} \end{bmatrix}$$



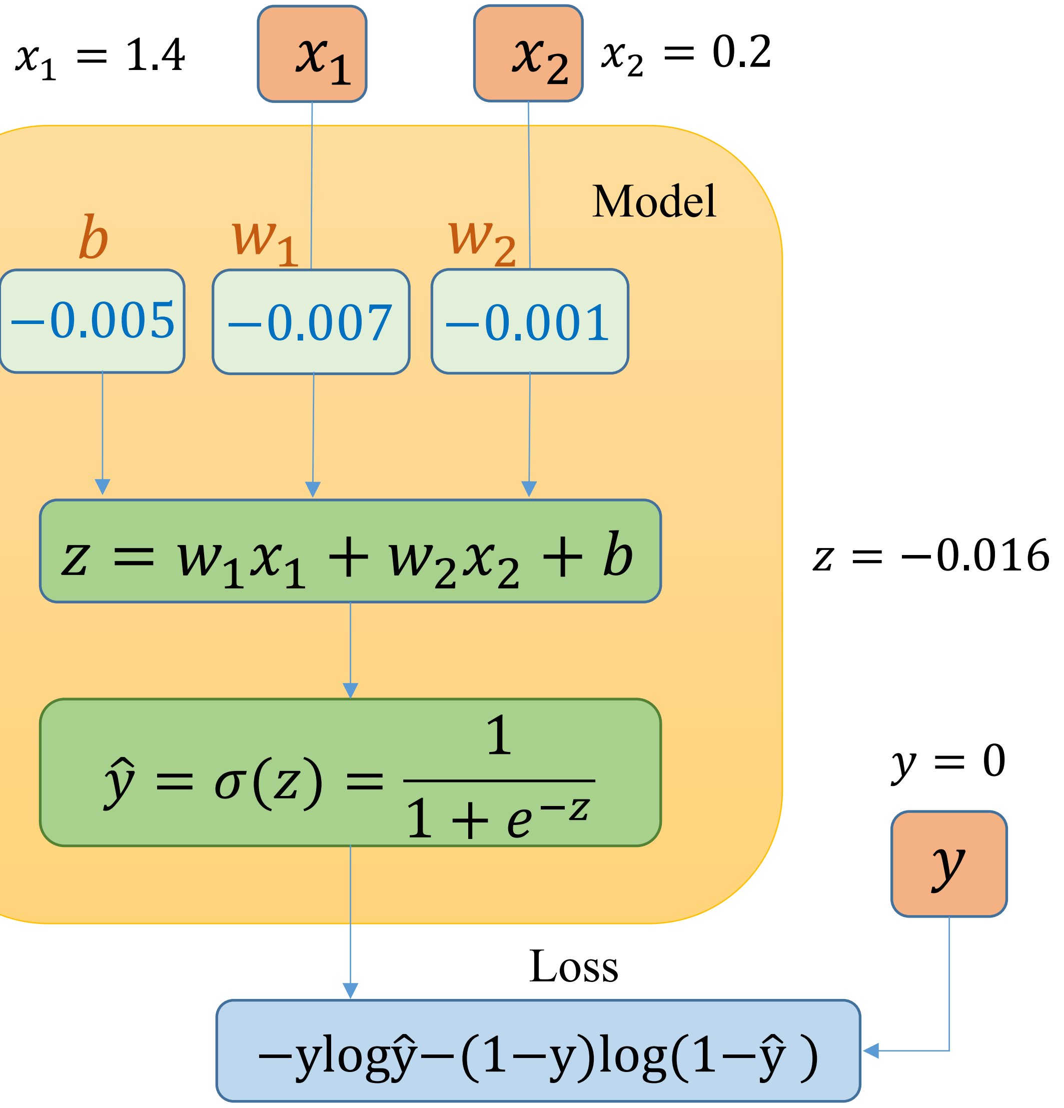
Example 4 - Zero Initialization

Dataset

Petal_Length	Petal_Width	Label
1.4	0.2	0
1.5	0.2	0
3	1.1	1
4.1	1.3	1

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 1.4 \\ 0.2 \end{bmatrix} \qquad \boldsymbol{y} = [0]$$

$\hat{y} = 0.49$



previous $\boldsymbol{L} = 1.1573$

$L = 0.68$

Example 5 - Zero Initialization

❖ Softmax regression

Training data

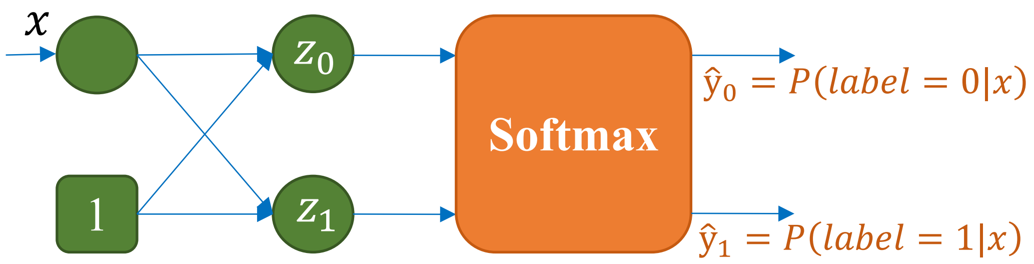
Feature	Label	
Petal_Length	Label	
1.4	0	Category A
1.3	0	Category A
1.5	0	Category A
4.5	1	Category B
4.1	1	Category B
4.6	1	Category B

One-hot encoding for labels

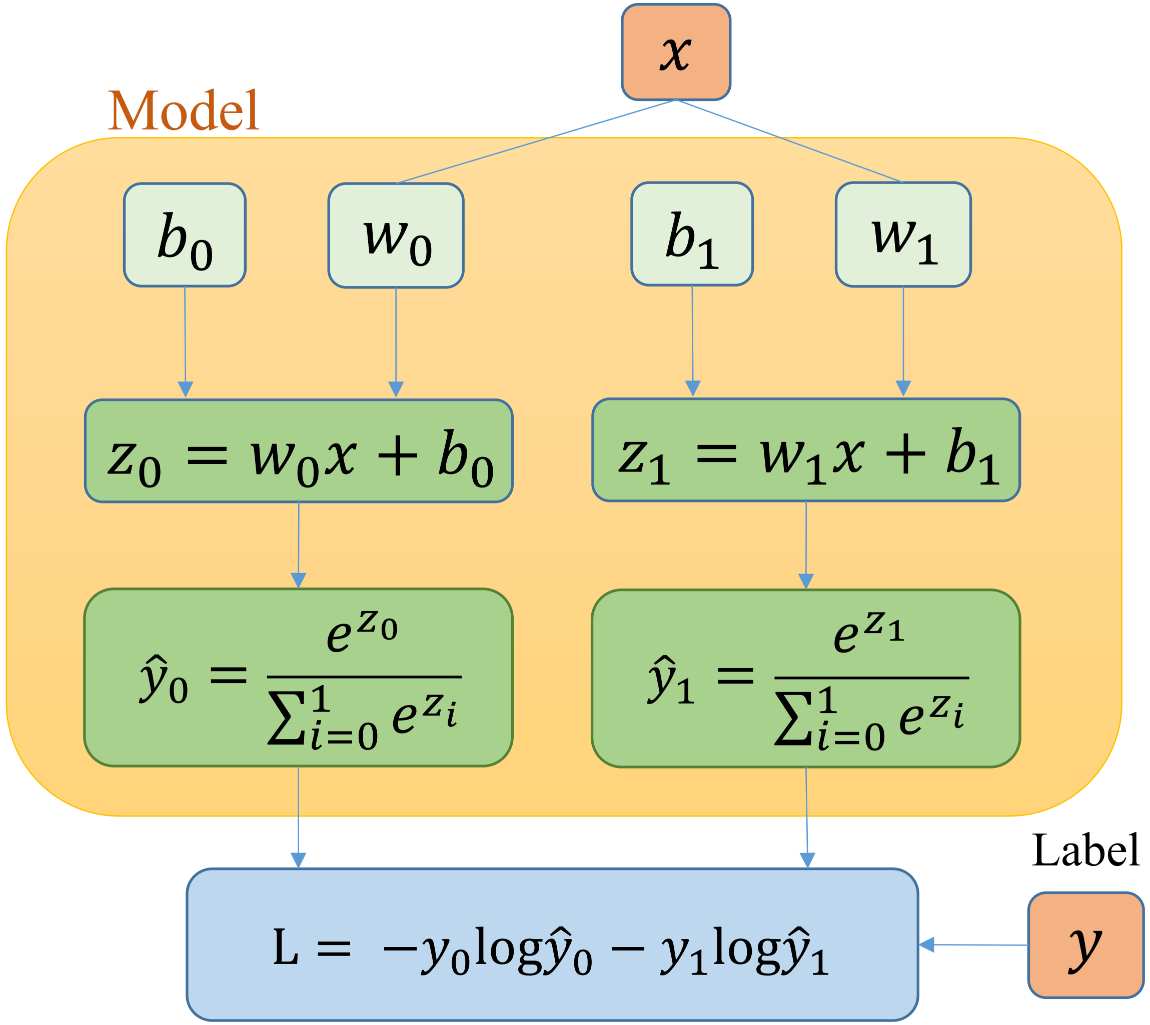
index 0 1

$$y = 0 \rightarrow \boldsymbol{y}^T = [1, 0]$$

$$y = 1 \rightarrow \boldsymbol{y}^T = [0, 1]$$



Model



Example 5 - Zero Initialization

❖ Softmax regression

Feature **Label**

Petal_Length	Label
1.4	0
1.3	0
1.5	0
4.5	1
4.1	1
4.6	1

#class=2

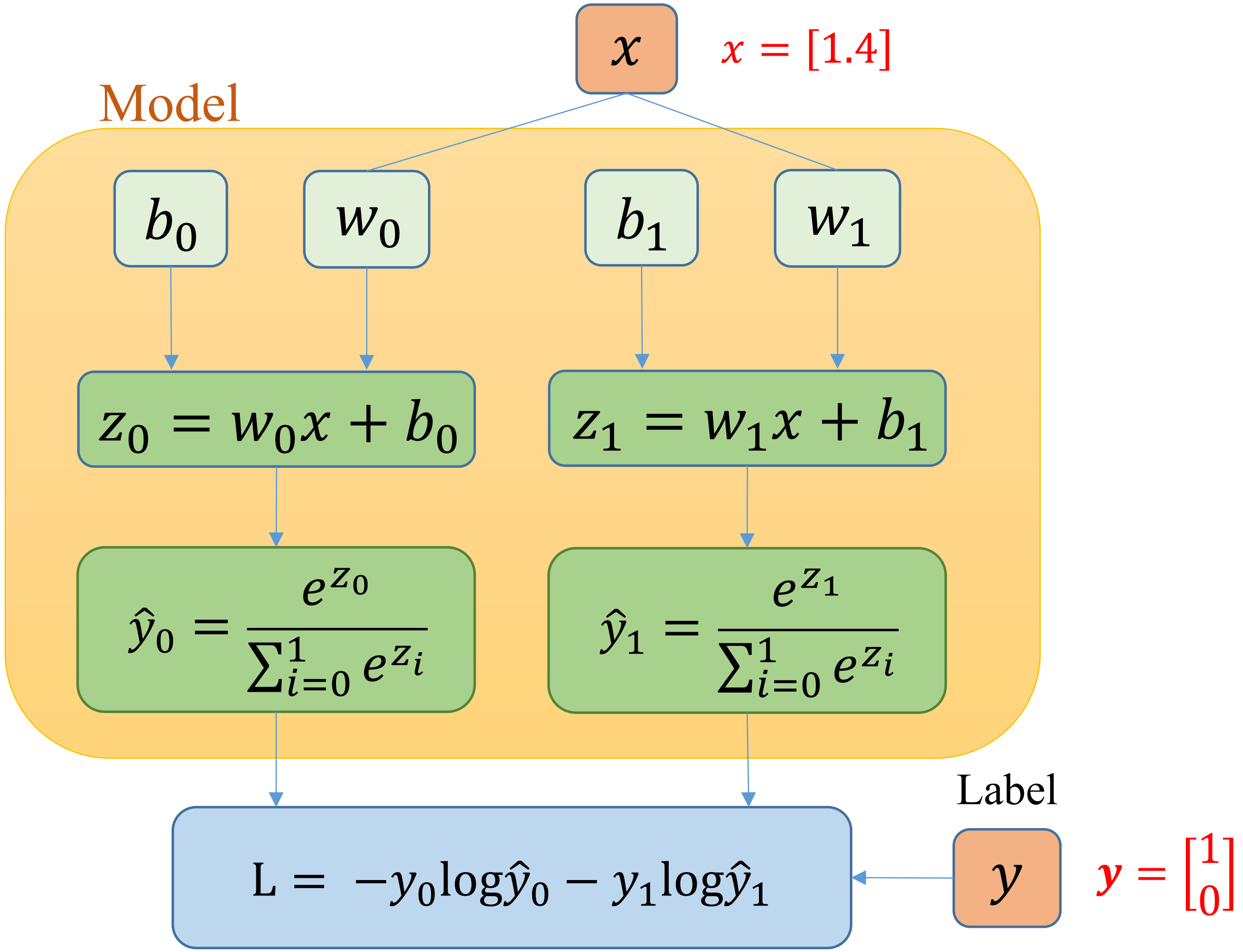
#feature=1

One-hot encoding for label

$$\begin{array}{l} \qquad\qquad\qquad\quad\; y_0 \;\; y_1 \\ y = 0 \rightarrow \boldsymbol{y}^T = [1 \quad 0] \\ y = 1 \rightarrow \boldsymbol{y}^T = [0 \quad 1] \end{array}$$

Training example

$$(x, y) = (1.4, 0)$$



Example 5 - Zero Initialization

❖ Softmax regression

Feature	Label
Petal_Length	Label
1.4	0
1.3	0
1.5	0
4.5	1
4.1	1
4.6	1

#class=2

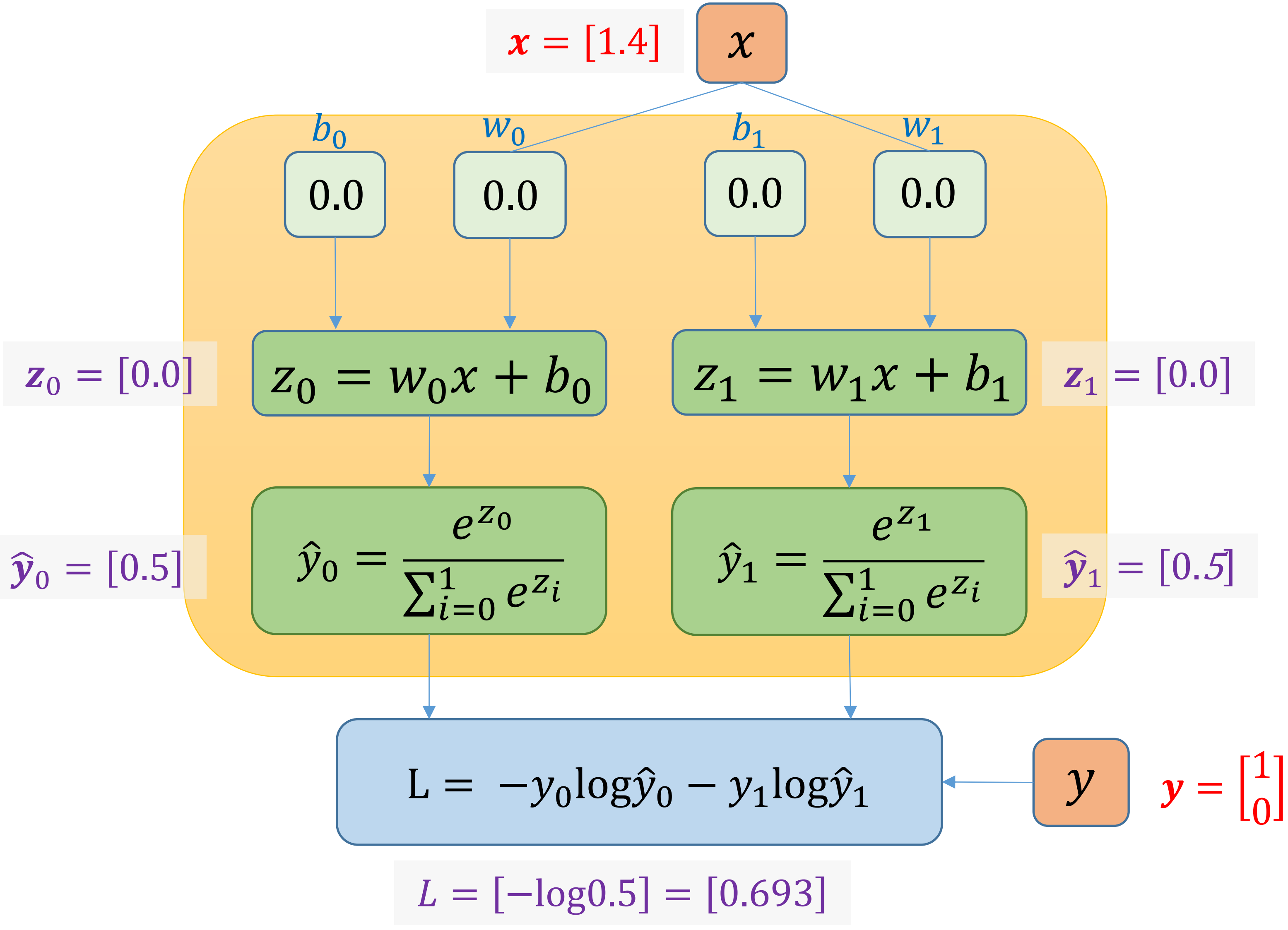
#feature=1

One-hot encoding for label

$$\begin{array}{l} \qquad\qquad\qquad\qquad\quad y_0 \quad y_1 \\ y = 0 \rightarrow \boldsymbol{y}^T = [1 \quad 0] \\ y = 1 \rightarrow \boldsymbol{y}^T = [0 \quad 1] \end{array}$$

Training example

$$(x, y) = (1.4, 0)$$



AI VIET NAM
@aivietnam.edu.vn

Example 5 - Zero Initialization

❖ Softmax regression

Derivative

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

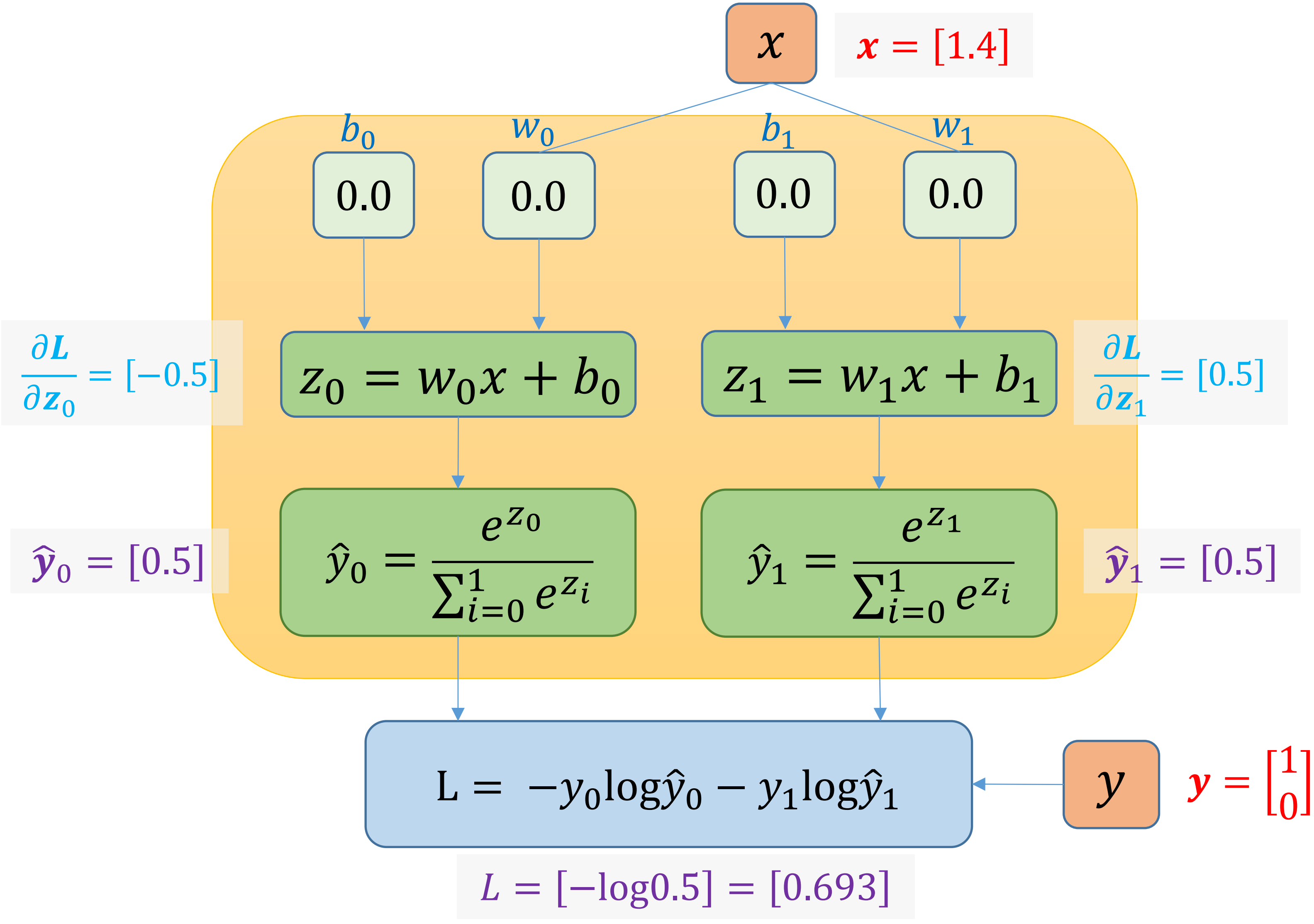
$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - y_i$$

$$\begin{array}{l} \\ y = 0 \rightarrow \boldsymbol{y}^T = \\ y = 1 \rightarrow \boldsymbol{y}^T = \end{array} \begin{array}{cc} y_0 & y_1 \\ [1 & 0] \\ [0 & 1] \end{array}$$

$$\frac{\partial L}{\partial \boldsymbol{z}_0} = \hat{y}_0 - 1$$
$$= 0.5 - 1 = -0.5$$

$$\frac{\partial L}{\partial \boldsymbol{z}_1} = \hat{y}_1 - 0 = 0.5$$



Example 5 - Zero Initialization

❖ Softmax regression

Derivative

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

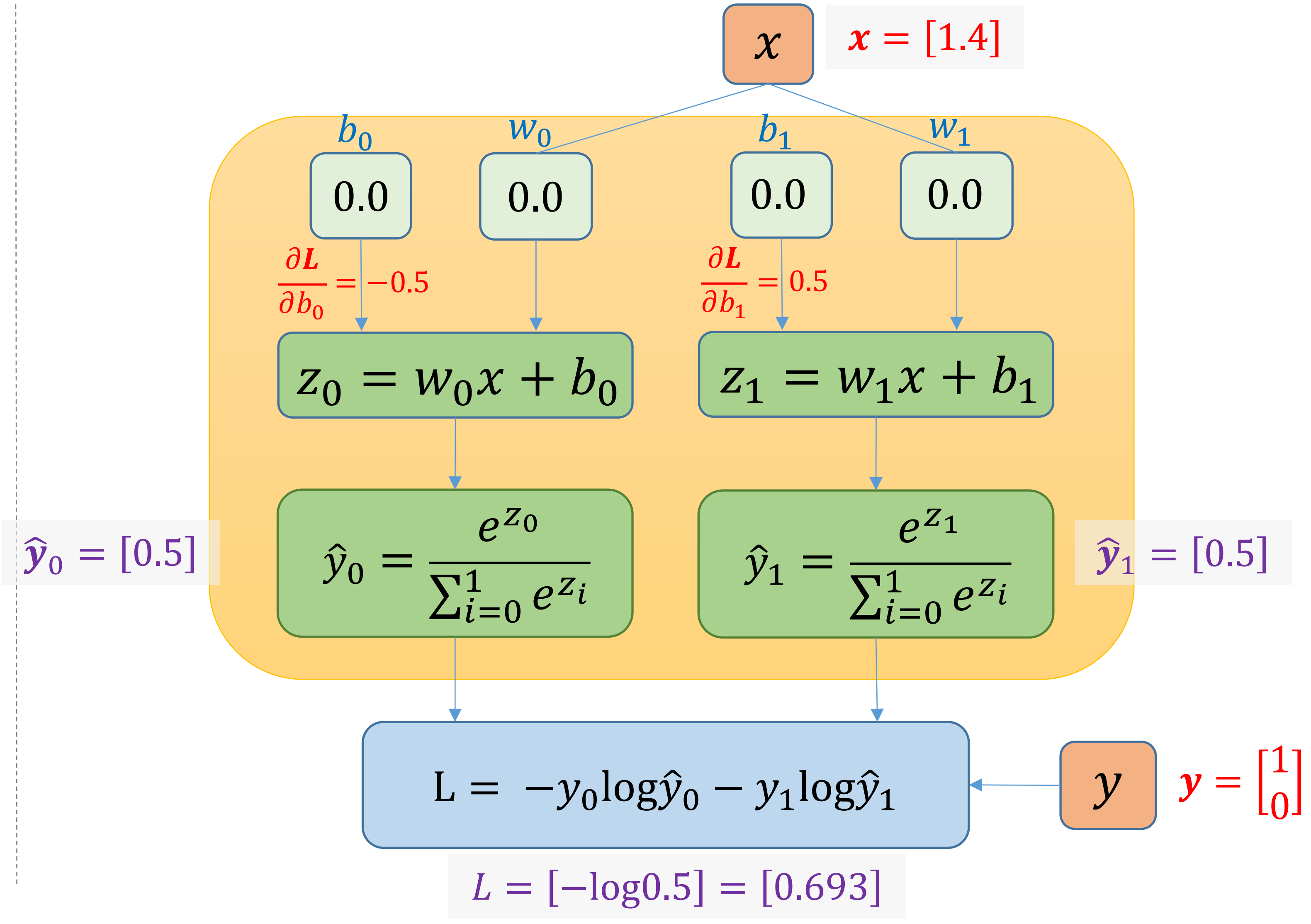
$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - y_i$$

$$\begin{array}{lcc} & y_0 & y_1 \\ y = 0 \rightarrow \boldsymbol{y}^T = & [1 & 0] \\ y = 1 \rightarrow \boldsymbol{y}^T = & [0 & 1] \end{array}$$

$$\frac{\partial \boldsymbol{L}}{\partial b_0} = (\hat{y}_0 - 1\ ) = -0.5$$

$$\frac{\partial \boldsymbol{L}}{\partial b_1} = (\hat{y}_1 - 0\ ) = 0.5$$



Example 5 - Zero Initialization

❖ Softmax regression

$$\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$$

$$\frac{\partial L}{\partial w_i} = x(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - y_i$$

$$y = 0 \rightarrow \boldsymbol{y}^T = \begin{array}{cc} y_0 & y_1 \\ \hline [1 & 0] \end{array}$$

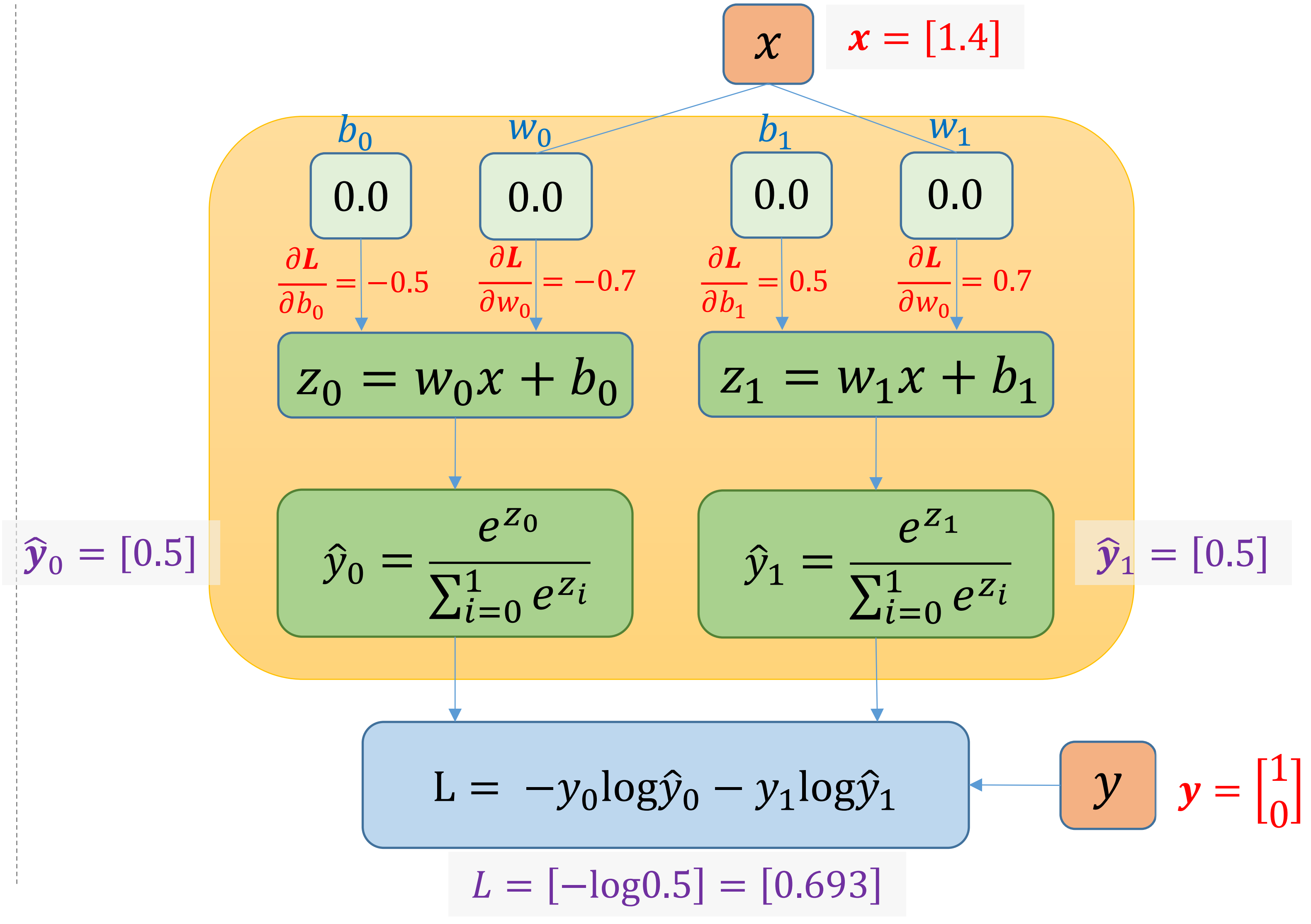
$$y = 1 \rightarrow \boldsymbol{y}^T = [0 \quad 1]$$

$$\frac{\partial \boldsymbol{L}}{\partial w_0} = x(\hat{y}_0 - 1\ )$$

$$= -0.5*1.4 = -0.7$$

$$\frac{\partial \boldsymbol{L}}{\partial w_1} = x(\hat{y}_1 - 0\ )$$

$$= 0.5*1.4 = 0.7$$



Example 5 - Zero Initialization

❖ Softmax regression

Update parameters

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \eta L'_{\boldsymbol{\theta}}$$

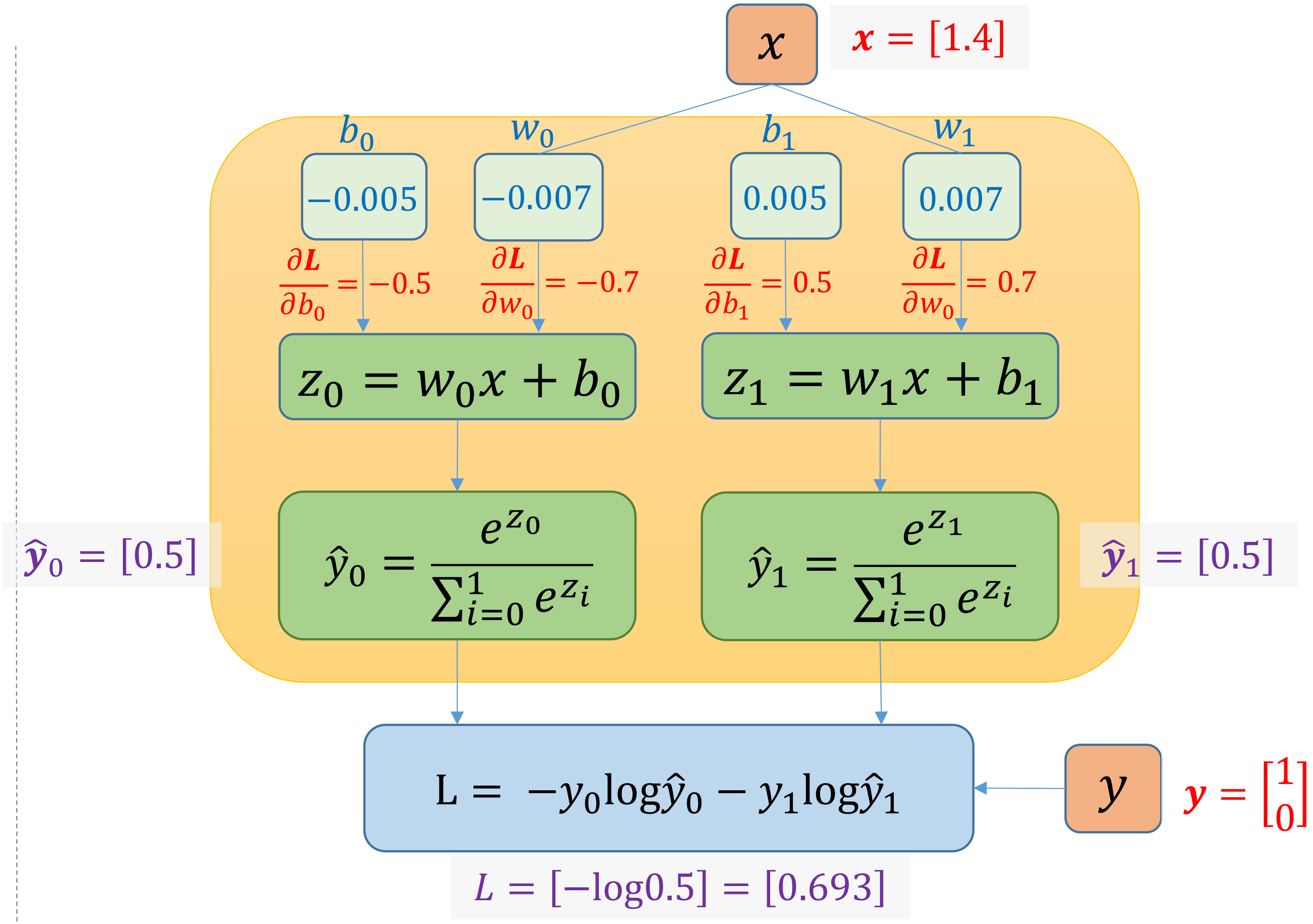
η is learning rate

$$\boldsymbol{\theta} = \begin{bmatrix} b_0 & b_1 \\ w_0 & w_1 \end{bmatrix} \qquad L'_{\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial L}{\partial b_0} & \frac{\partial L}{\partial b_1} \\ \frac{\partial L}{\partial w_0} & \frac{\partial L}{\partial w_1} \end{bmatrix}$$

$$\eta = 0.1$$

$$\boldsymbol{\theta} = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix} - 0.01 \begin{bmatrix} -0.5 & 0.5 \\ -0.7 & 0.7 \end{bmatrix}$$

$$= \begin{bmatrix} -0.005 & 0.005 \\ -0.007 & 0.007 \end{bmatrix}$$



Example 5 - Zero Initialization

❖ Softmax regression

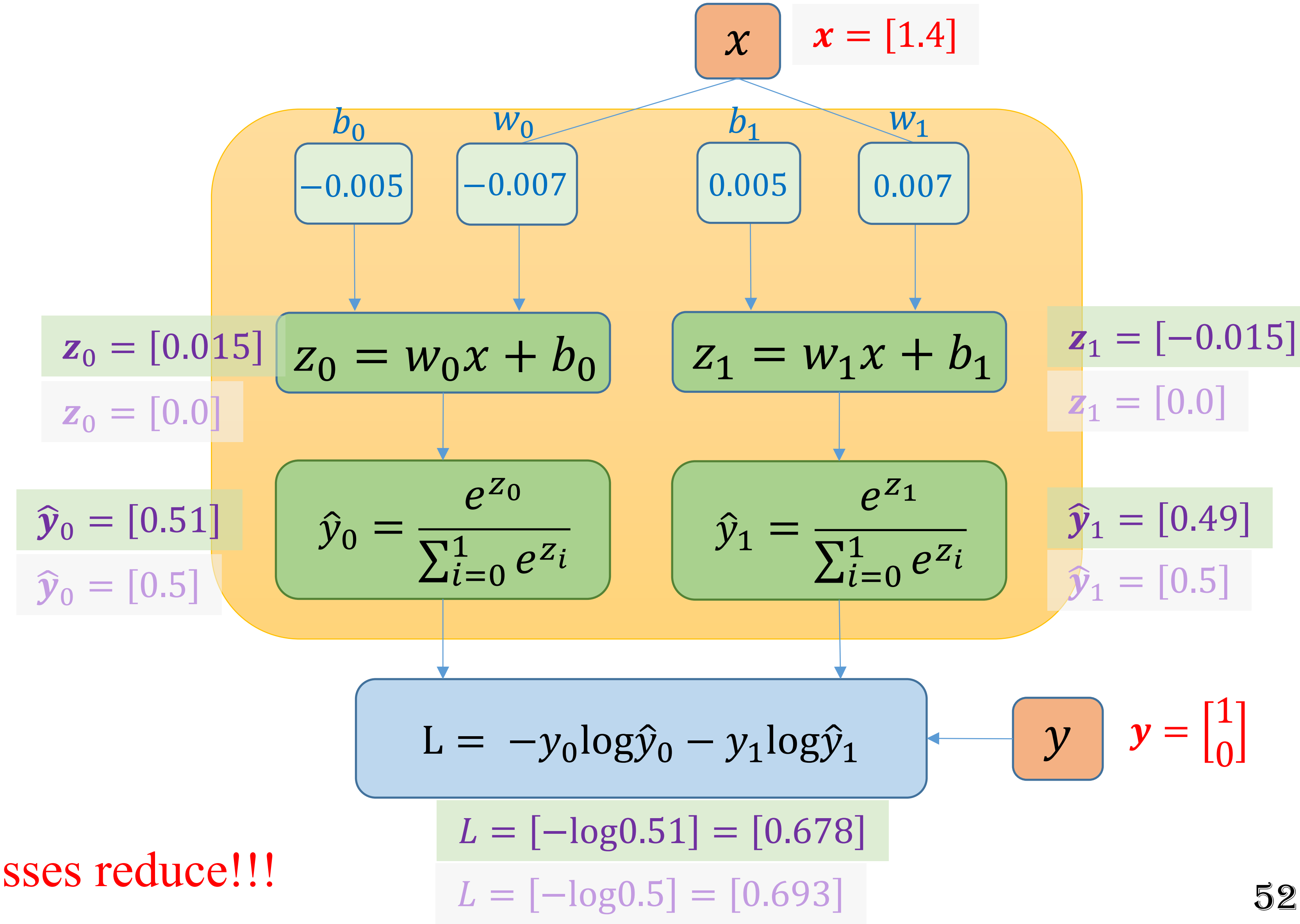
Feature	Label
Petal_Length	Label
1.4	0
1.3	0
1.5	0
4.5	1
4.1	1
4.6	1

One-hot encoding for label

$$\begin{array}{l} \qquad\qquad\qquad\qquad\quad\; y_0 \;\; y_1 \\ y = 0 \rightarrow \boldsymbol{y}^T = [1 \quad 0] \\ y = 1 \rightarrow \boldsymbol{y}^T = [0 \quad 1] \end{array}$$

Training example

$$(x, y) = (1.4, 0)$$



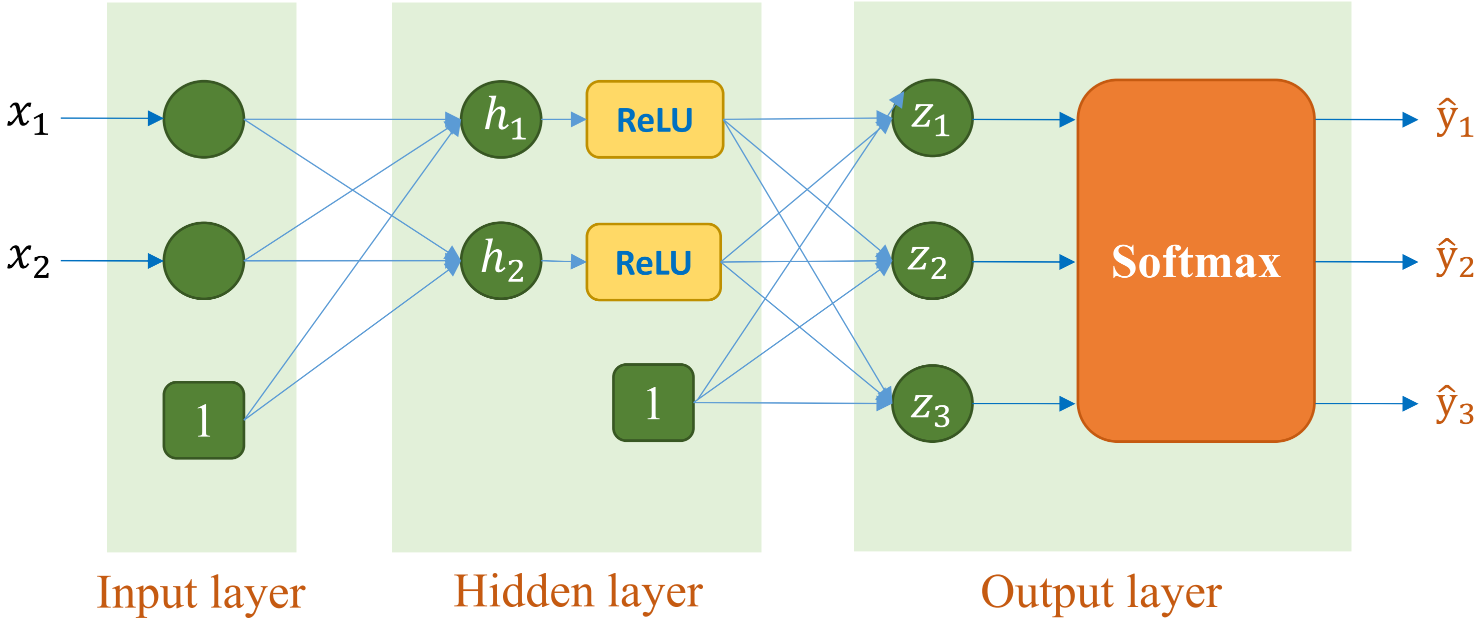
losses reduce!!!

Example 6 - Zero Initialization

AI VIET NAM
@aivietnam.edu.vn

❖ MLP

Feature		Label
Petal_Length	Petal_Width	Label
1.5	0.2	0
1.4	0.2	0
1.6	0.2	0
4.7	1.6	1
3.3	1.1	1
4.6	1.3	1
5.6	2.2	2
5.1	1.5	2
5.6	1.4	2



$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \boldsymbol{x}^{(3)} \end{bmatrix} = \begin{bmatrix} 1.5 & 0.2 \\ 4.7 & 1.6 \\ 5.6 & 2.2 \end{bmatrix}$$

$$\boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

$$\boldsymbol{h} = [\boldsymbol{h}_1 \quad \boldsymbol{h}_2] = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{w} = [\boldsymbol{w}_1 \quad \boldsymbol{w}_2 \quad \boldsymbol{w}_3] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{b}_h = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{b}_w = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{h} = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

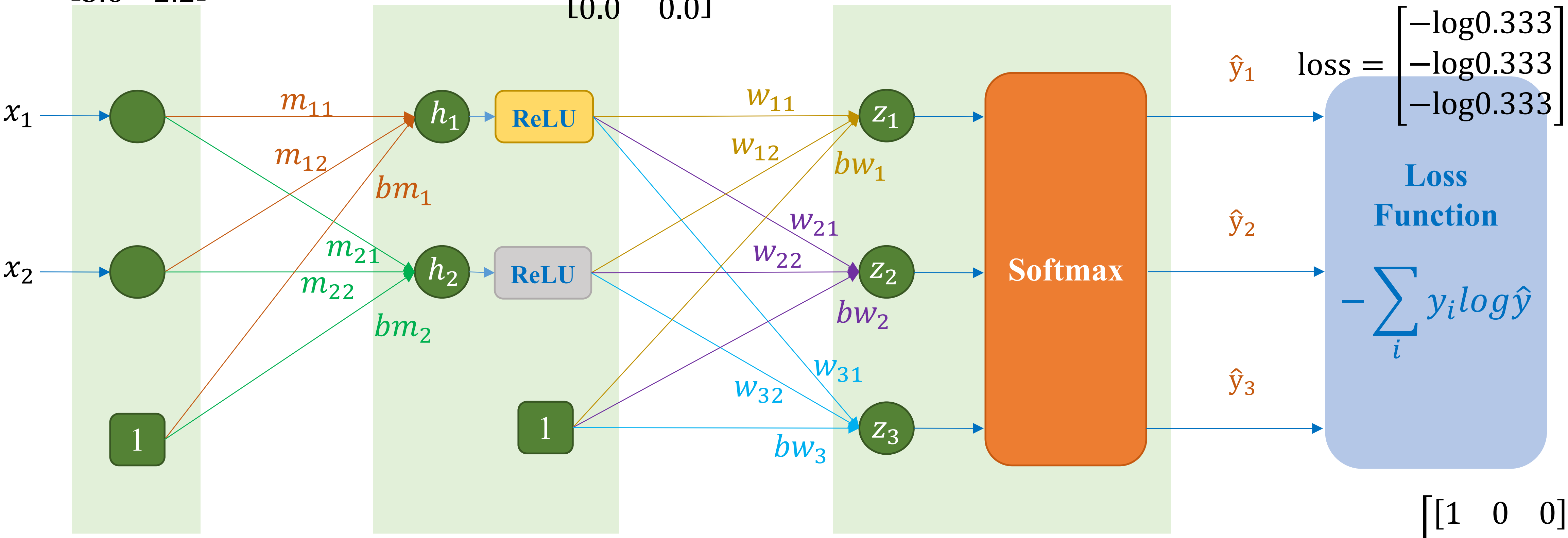
$$\boldsymbol{x} = \begin{bmatrix} 1.5 & 0.2 \\ 4.7 & 1.6 \\ 5.6 & 2.2 \end{bmatrix}$$

$$\textbf{ReLU} = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{z} = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\hat{y}} = \begin{bmatrix} 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \end{bmatrix}$$

$$\text{loss} = \begin{bmatrix} -\log 0.333 \\ -\log 0.333 \\ -\log 0.333 \end{bmatrix}$$



$$\boldsymbol{y} = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \end{bmatrix}$$

$$\boldsymbol{m} = [\boldsymbol{m_1} \quad \boldsymbol{m_2}] = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{w} = [\boldsymbol{w_1} \quad \boldsymbol{w_2} \quad \boldsymbol{w_3}] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix}$$

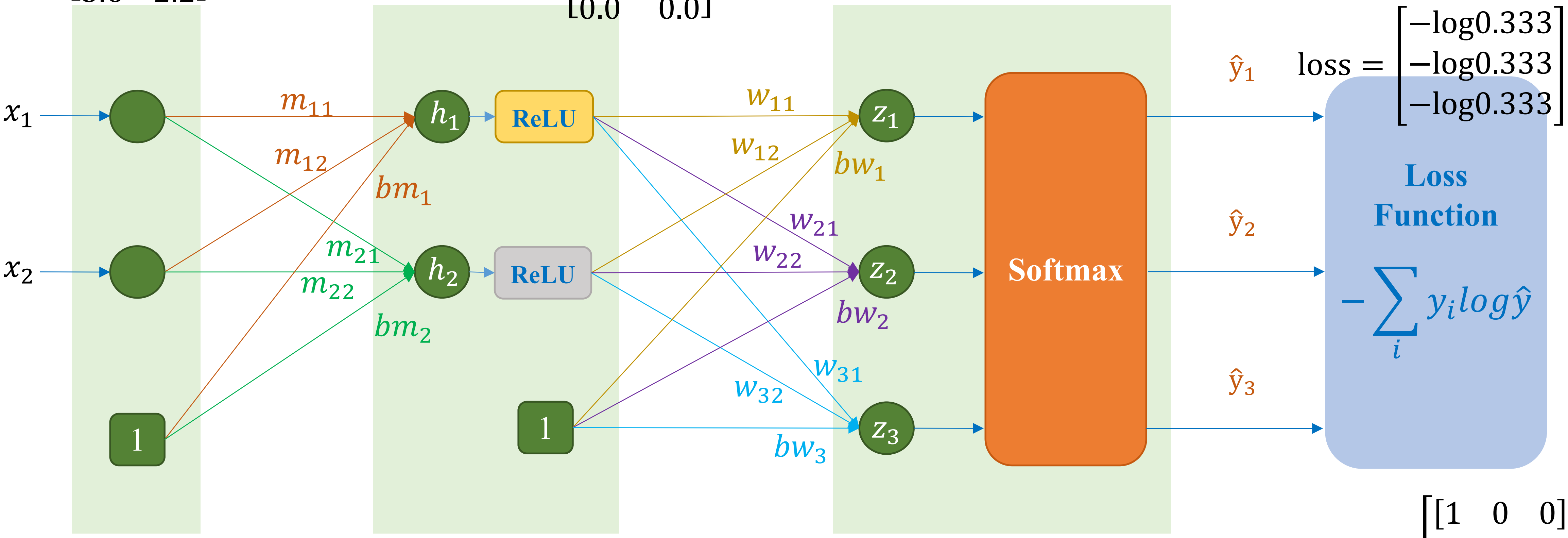
$$\boldsymbol{bw} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{h} = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{x} = \begin{bmatrix} 1.5 & 0.2 \\ 4.7 & 1.6 \\ 5.6 & 2.2 \end{bmatrix} \qquad \textbf{ReLU} = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{z} = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix} \qquad \hat{\boldsymbol{y}} = \begin{bmatrix} 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \\ 0.333 & 0.333 & 0.333 \end{bmatrix}$$

$$\text{loss} = \begin{bmatrix} -\text{log}0.333 \\ -\text{log}0.333 \\ -\text{log}0.333 \end{bmatrix}$$



$$\boldsymbol{y} = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \end{bmatrix}$$

$$\boldsymbol{m} = [\boldsymbol{m}_1 \quad \boldsymbol{m}_2] = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix} \qquad \boldsymbol{bm} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{w} = [\boldsymbol{w}_1 \quad \boldsymbol{w}_2 \quad \boldsymbol{w}_3] = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix} \qquad \boldsymbol{bw} = \begin{bmatrix} v \\ v \\ v \end{bmatrix}$$

Optimizers

❖ Optimizer Selection

Data Preparation

↓

Data Normalization

↓

Model (Network) Construction

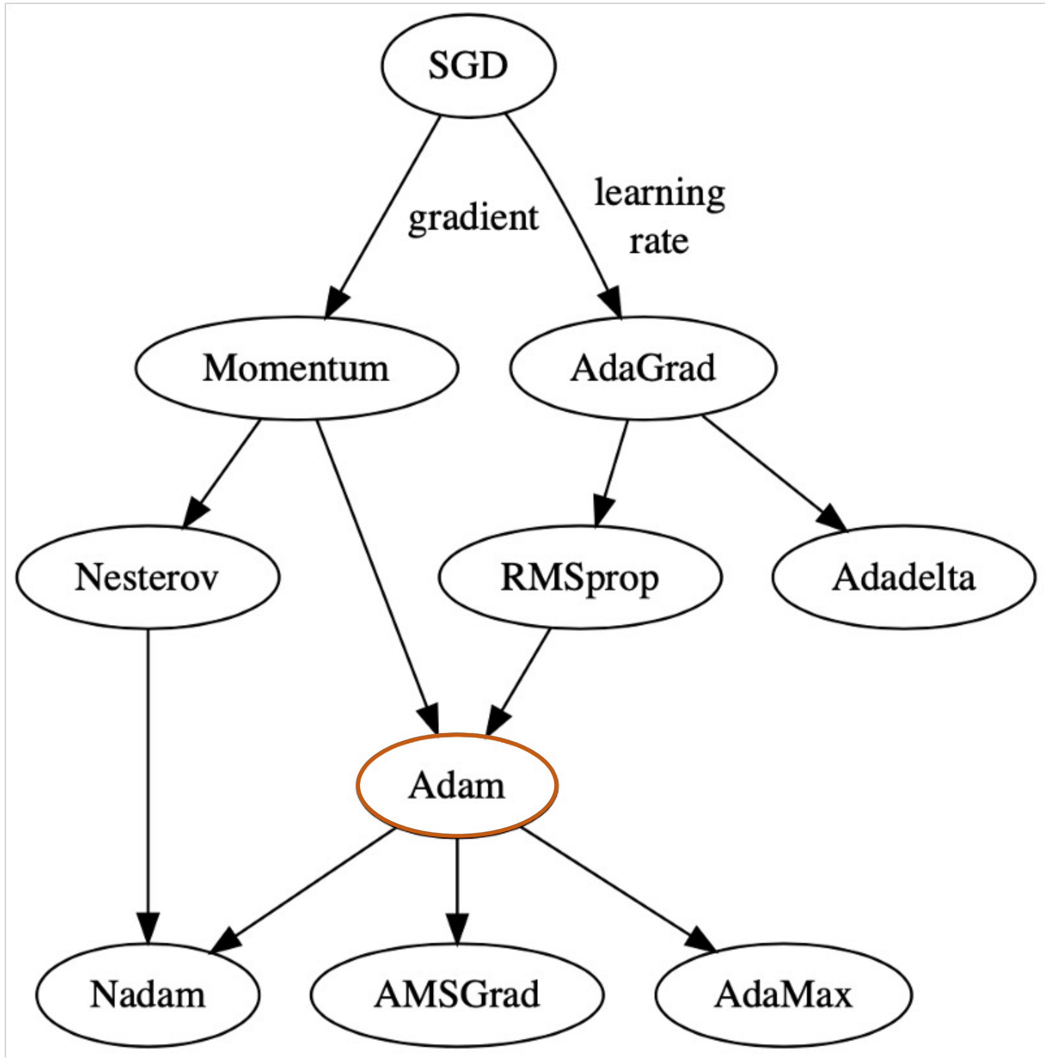
↓

Parameter Initialization

→

Define a way to update parameters

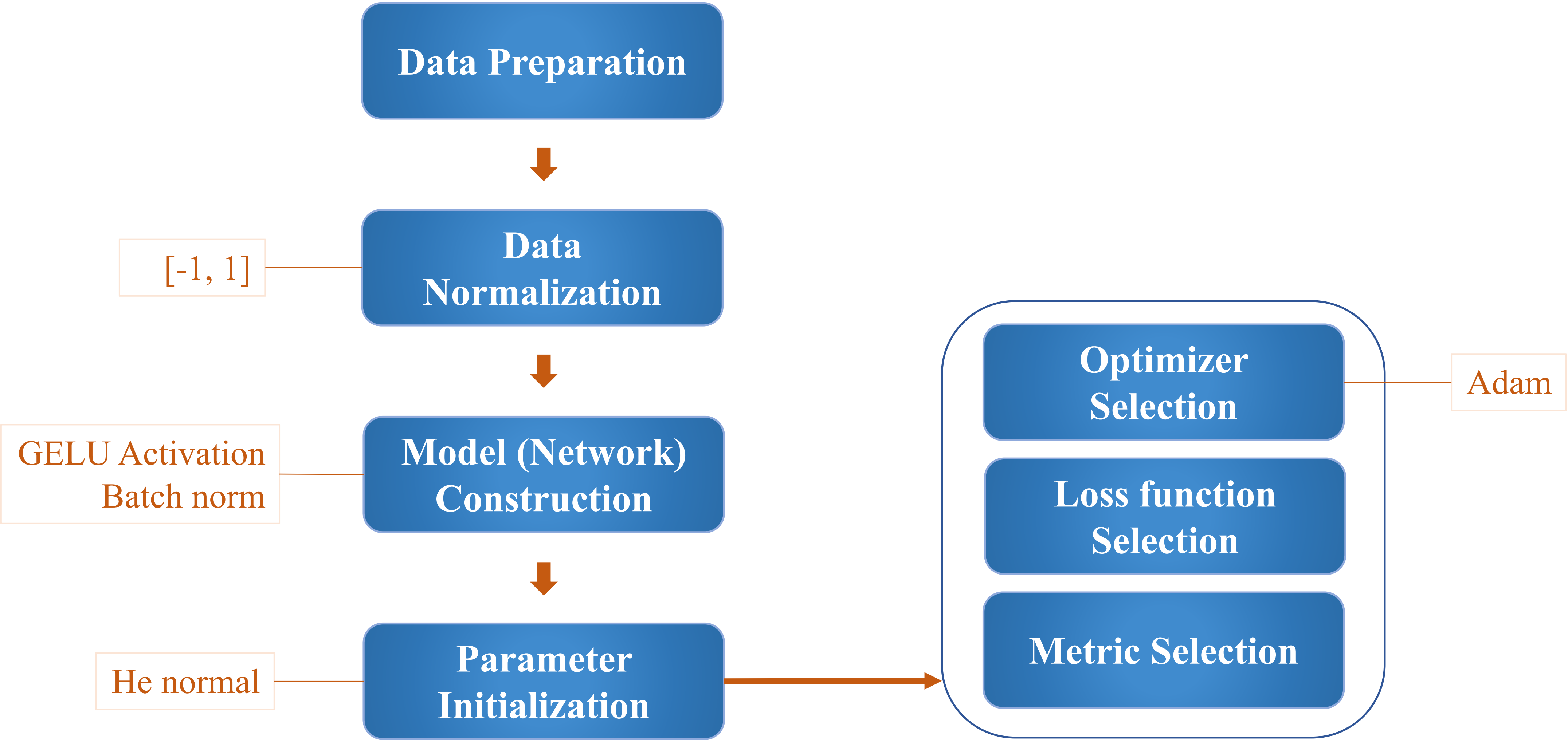
- Optimizer Selection
- Loss function Selection
- Metric Selection



Summary and Discussion

Summary

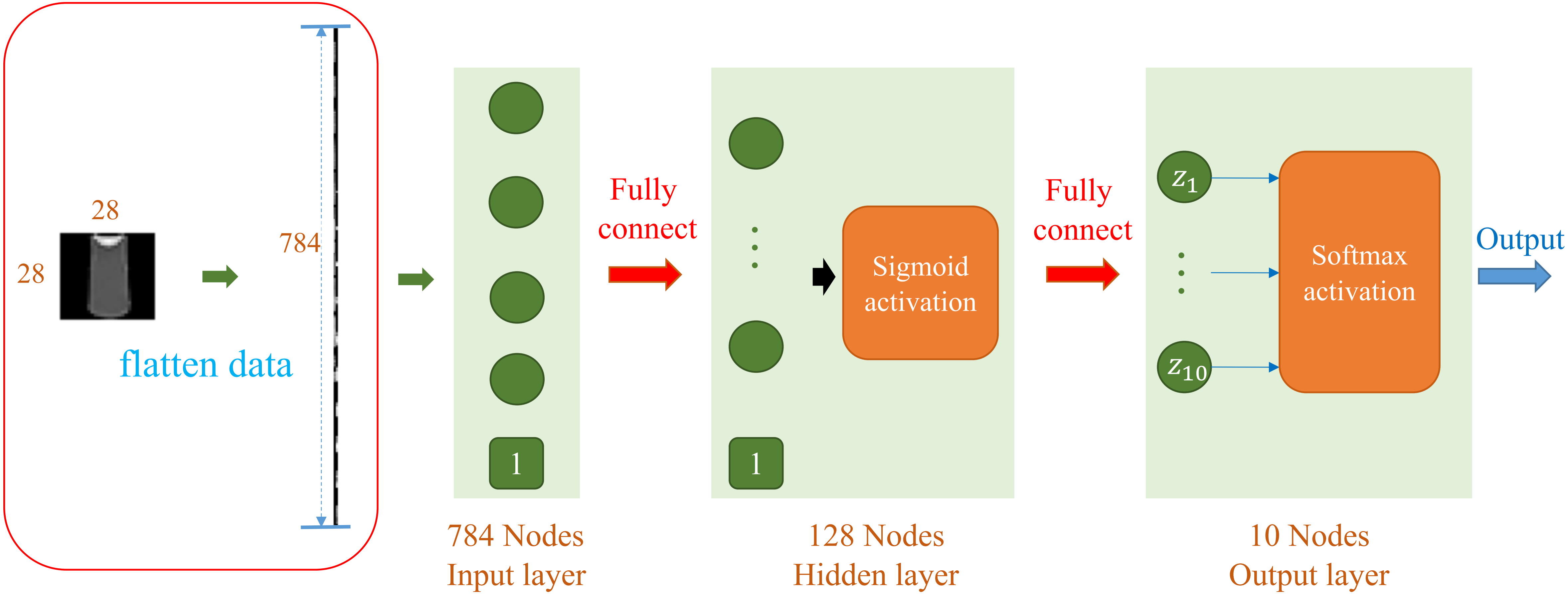
❖ Recommendation



Discussion

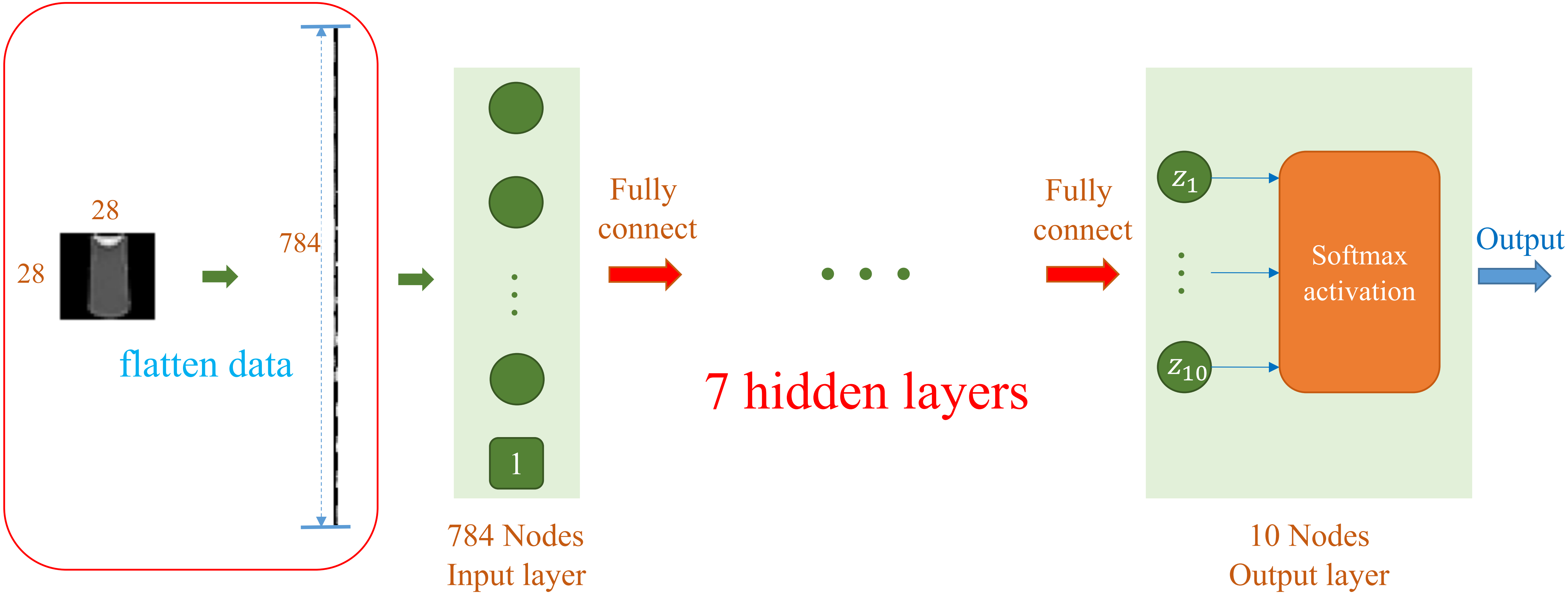
❖ Sigmoid and SGD

❖ W/o using normalization



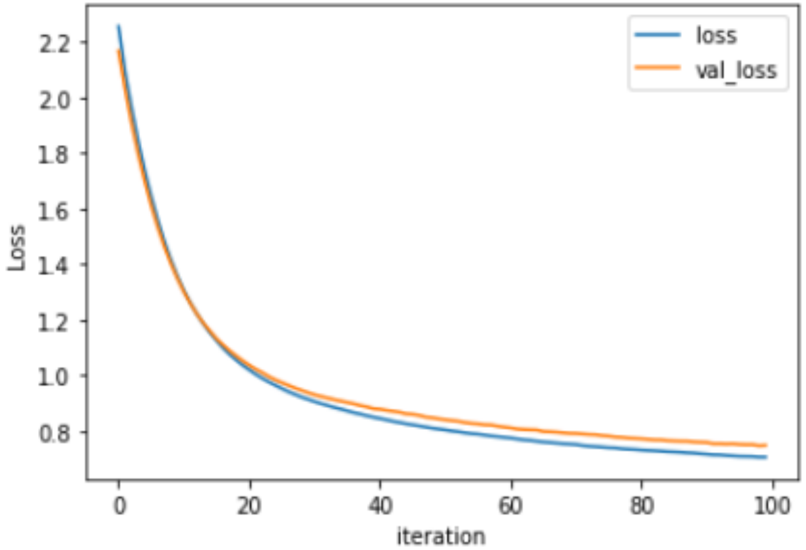
Discussion

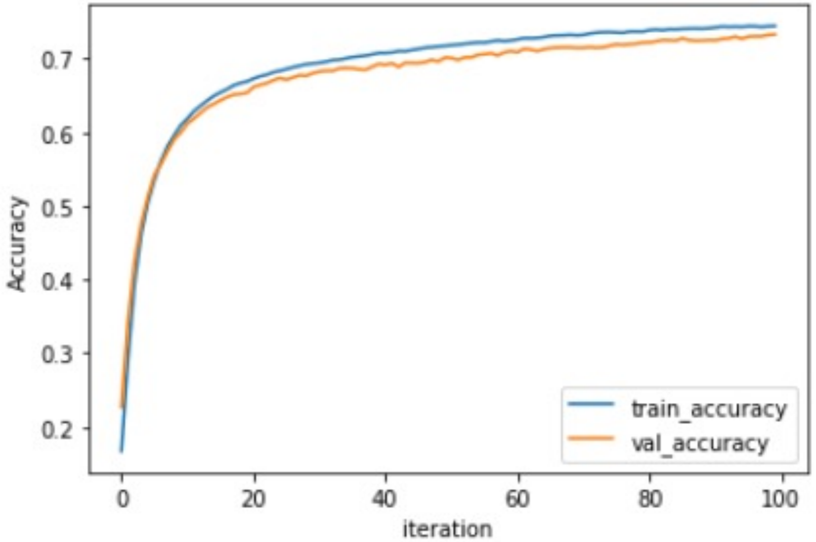
- ❖ **Sigmoid and SGD**
- ❖ **W/o using normalization**



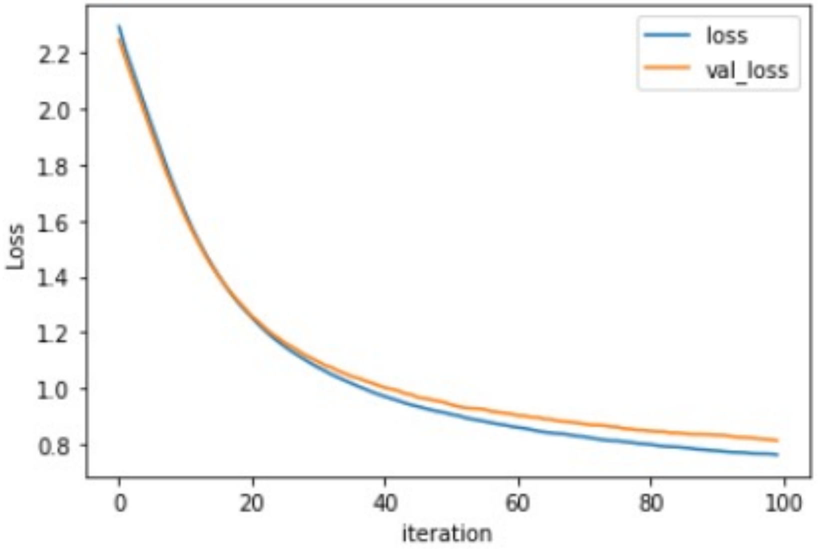
Discussion

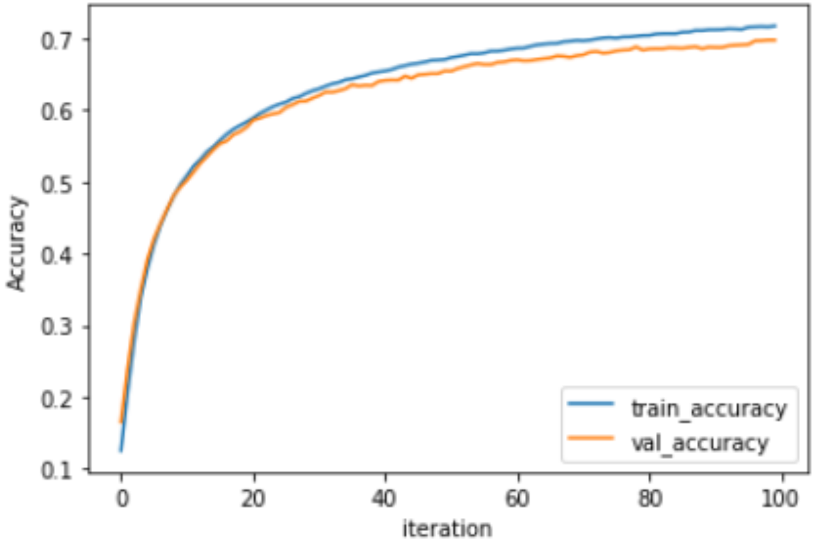
2 hidden layers



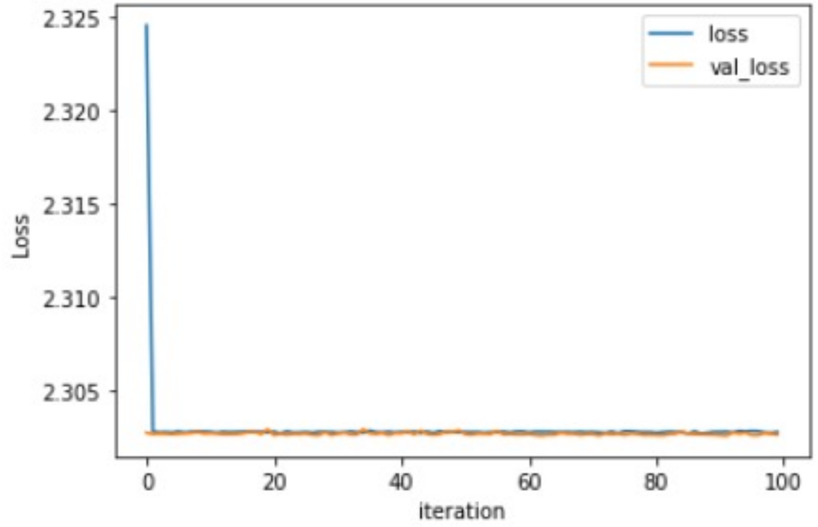


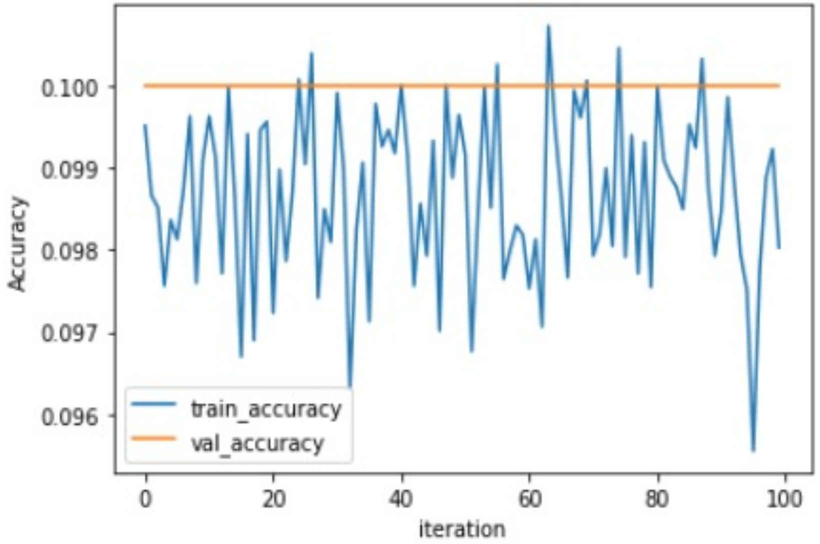
5 hidden layers (!)





7 hidden layers





Further Reading

Dying ReLU

https://towardsdatascience.com/the-dying-relu-problem-clearly-explained-42d0c54e0d24

Initialization

https://www.deeplearning.ai/ai-notes/initialization/index.html