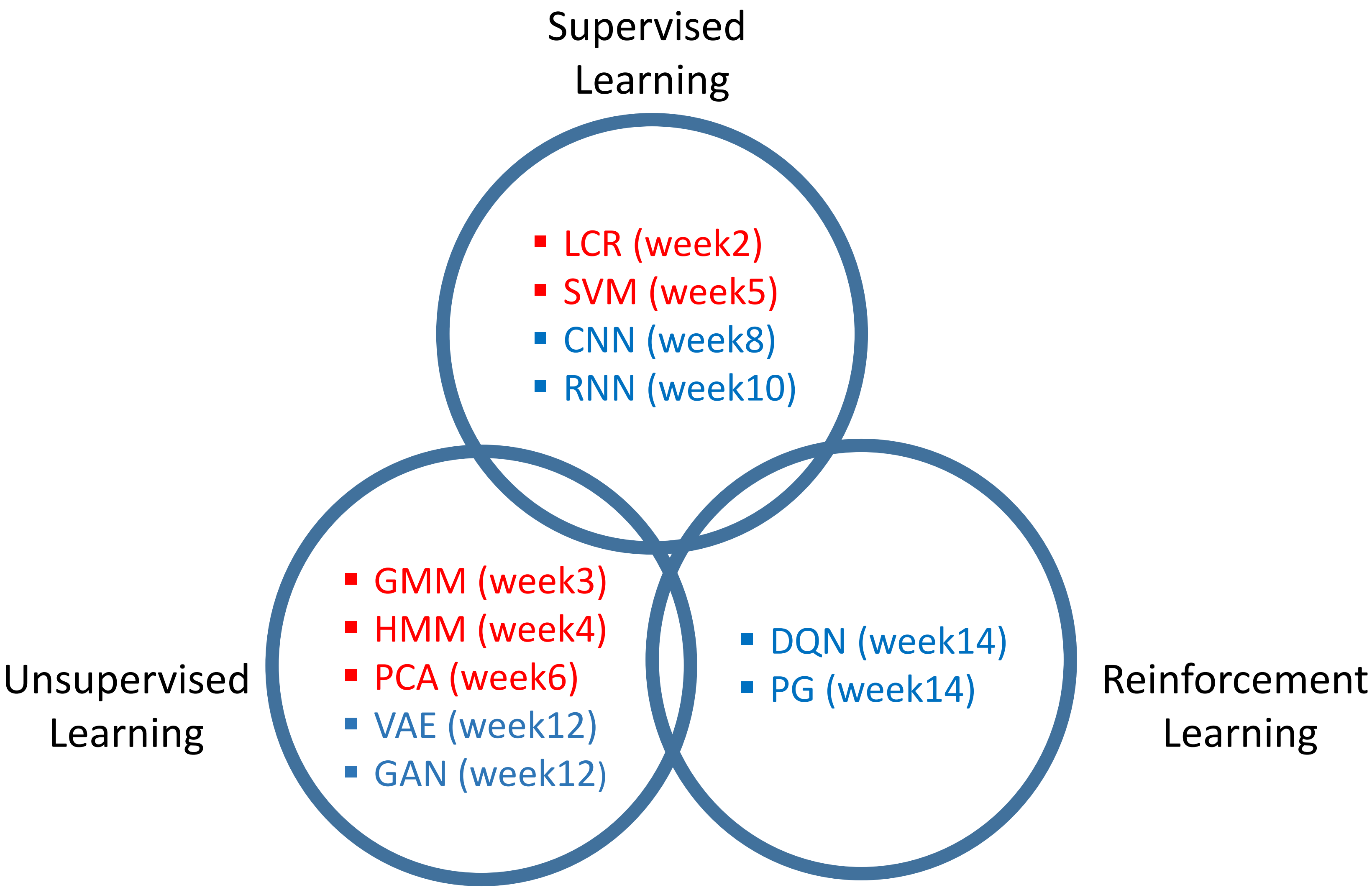
Practical Machine Learning

Lecture 7
Neural Networks

Dr. Suyong Eum

OSAKA UNIVERSITY

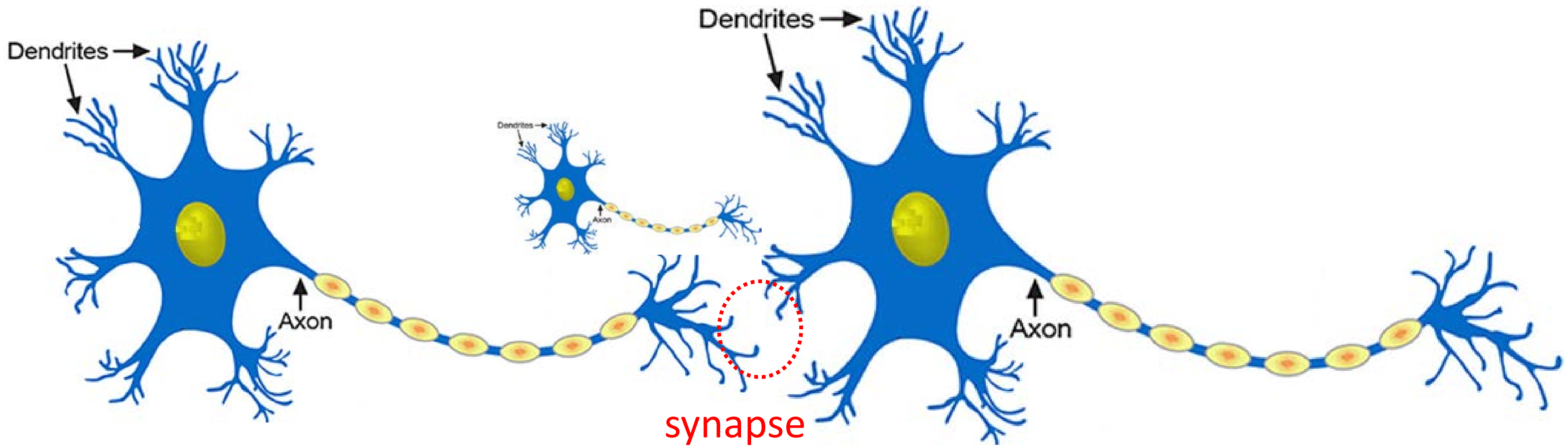
Where we are

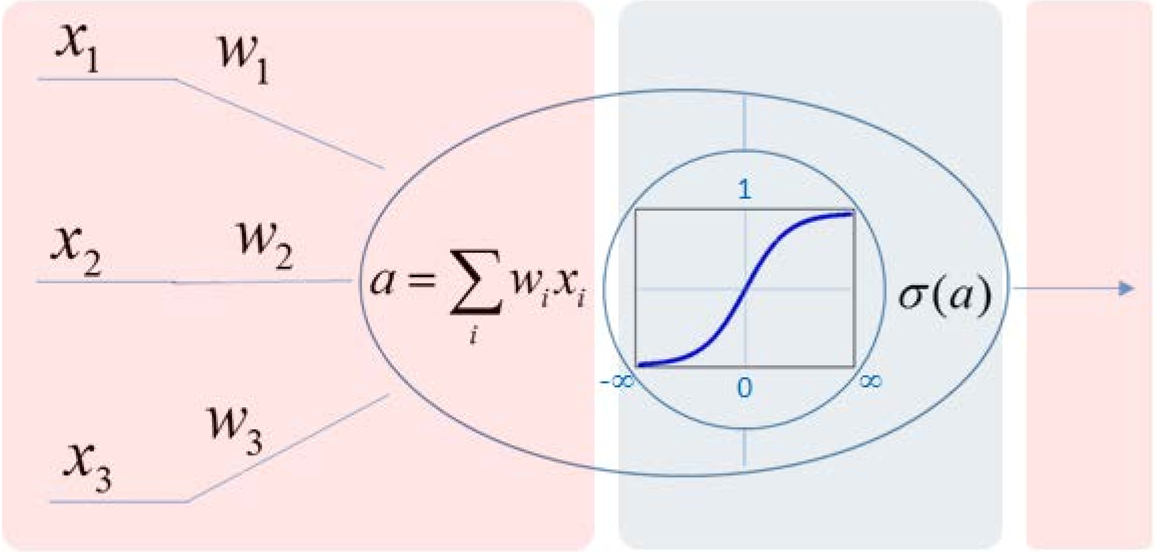

Supervised Learning
LCR (week2)
SVM (week5)
CNN (week8)
RNN (week10)
Unsupervised Learning
GMM (week3)
HMM (week4)
PCA (week6)
VAE (week12)
GAN (week12)
Reinforcement Learning
DQN (week14)
PG (week14)

You are going to learn

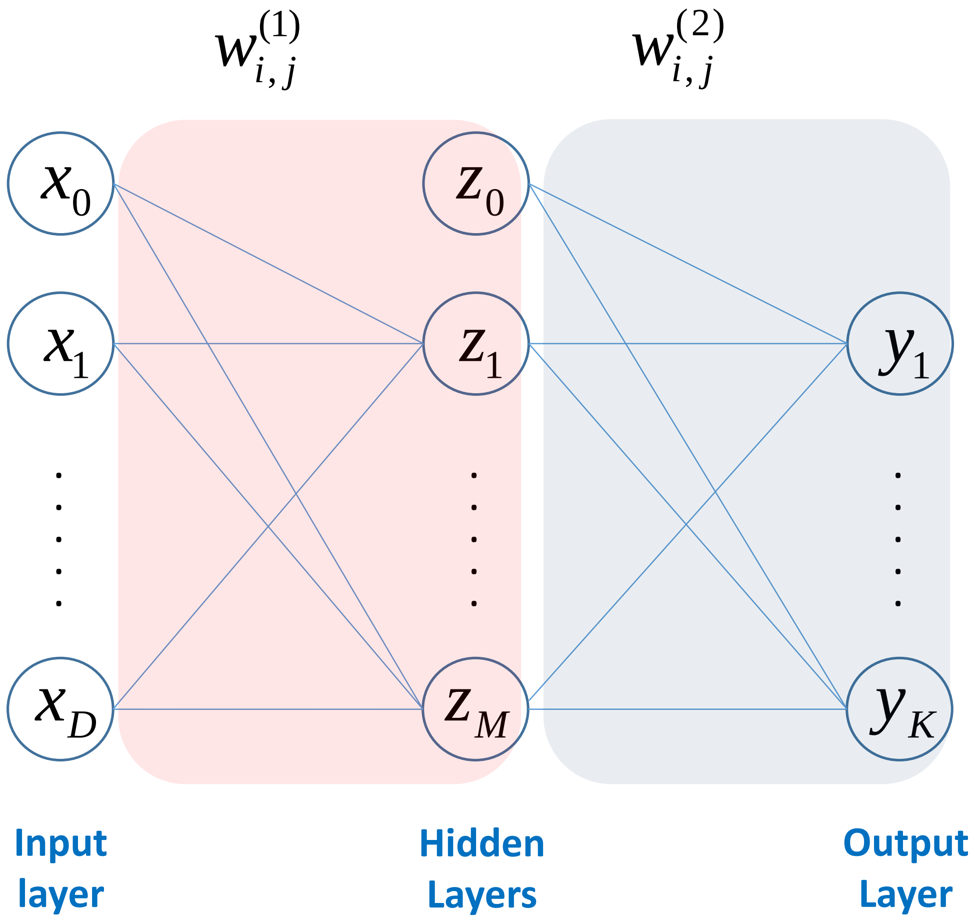
- Terminology for neural networks
- Its basic operation with a multi-layer perceptron model (MLP)
- How to train a neural network
  - Backpropagation
  - An example of backpropagation

A bio-inspired approach





Terminology in neural networks

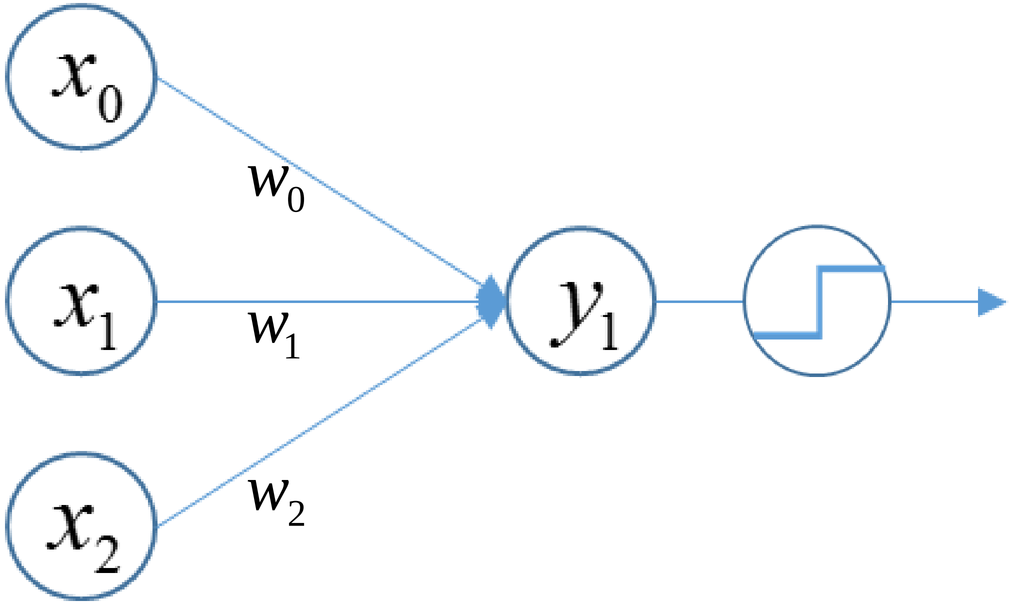


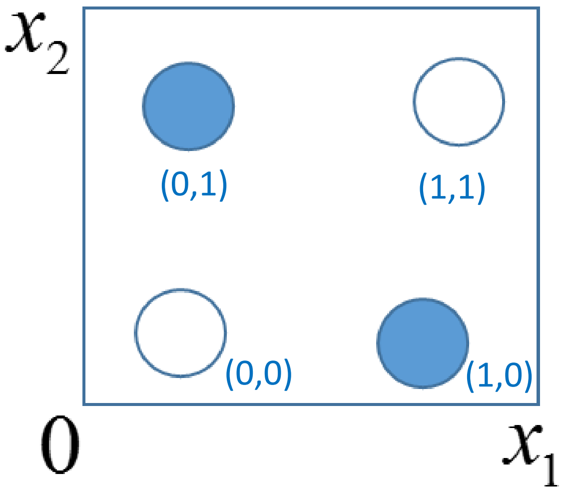
$w_{i,j}^{(\ell)}$: weight on a link at layer (ℓ) between node i and j

- Two types of layers:
  - Nodes: Input/Hidden/Output layer
  - Links: Hidden/Output layer
- In general, a standard L-layer neural network consists of
  - an input layer,
  - (L-1) hidden layers,
  - an output layer.

Role of hidden layers

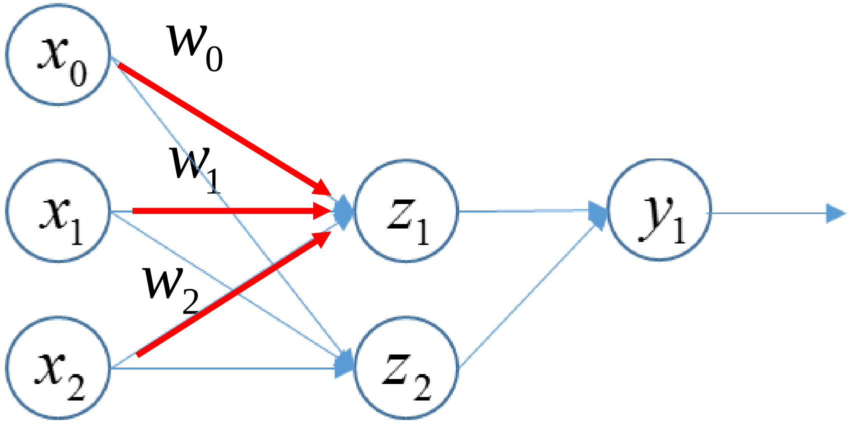
$$y_1 = w_2 x_2 + w_1 x_1 + w_0$$

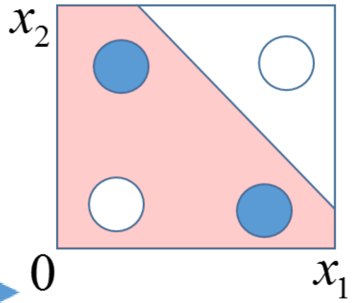




XOR problem

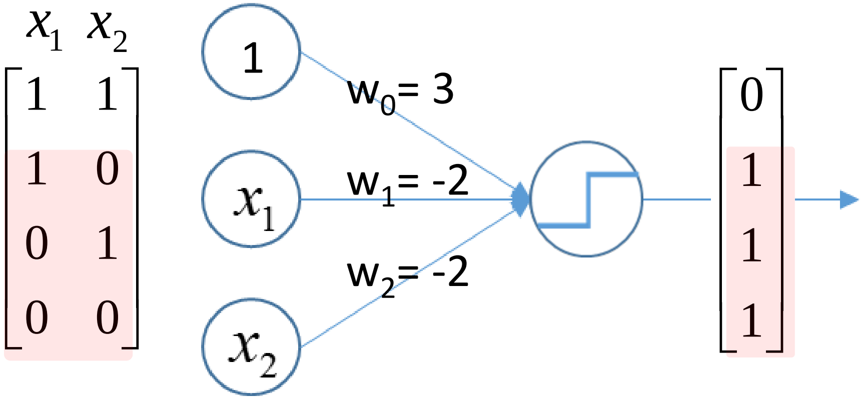
Role of hidden layers: a simple example



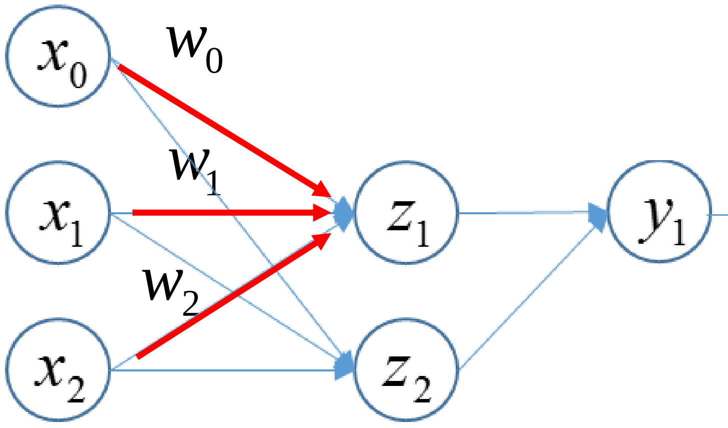


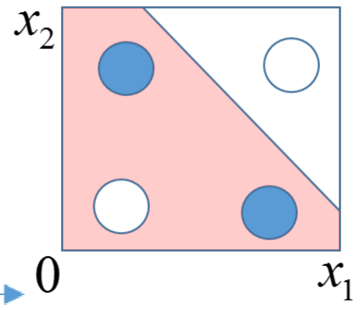
$$z_1 = w_2 x_2 + w_1 x_1 + w_0$$

NAND gate



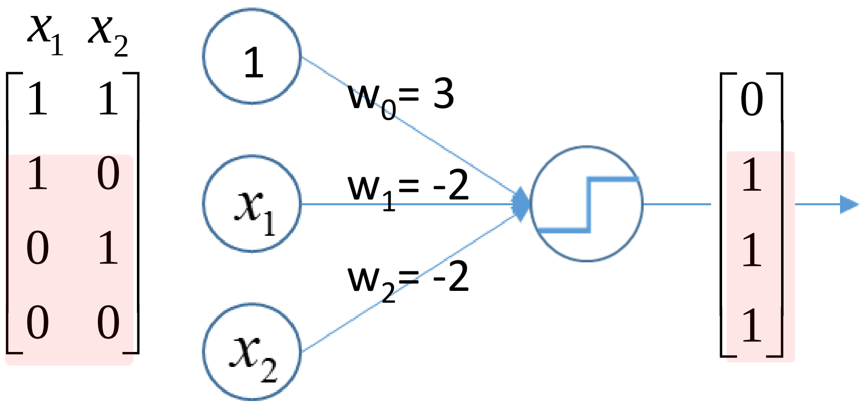
Role of hidden layers: a simple example

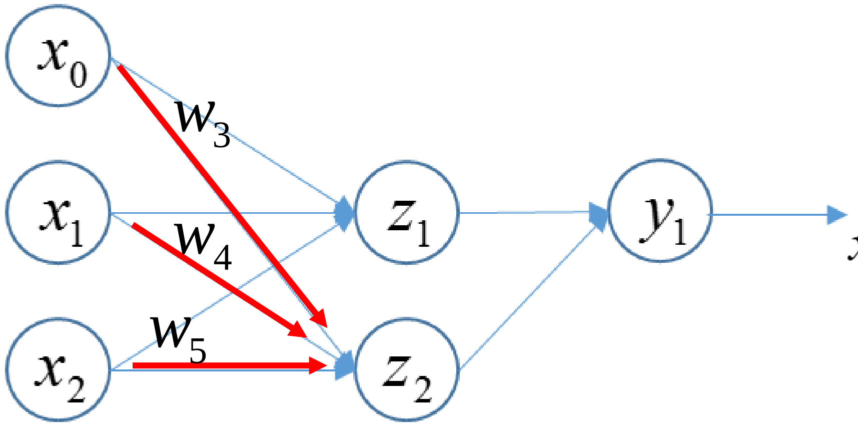


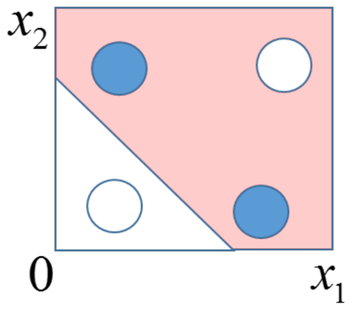


$$z_1 = w_2 x_2 + w_1 x_1 + w_0$$

NAND gate

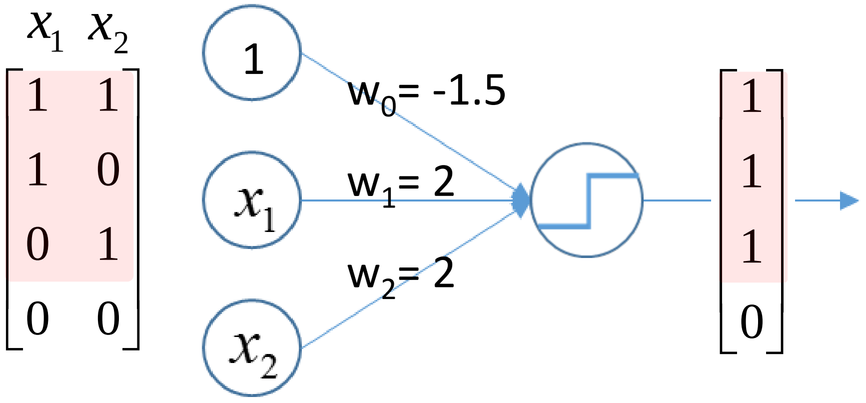




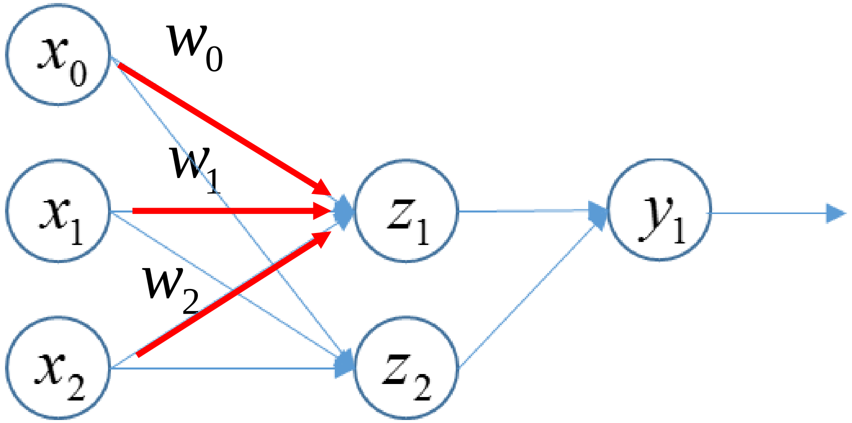


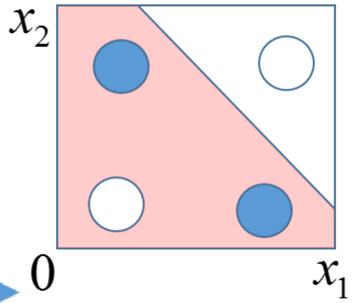
$$z_2 = w_5 x_2 + w_4 x_1 + w_3$$

OR gate

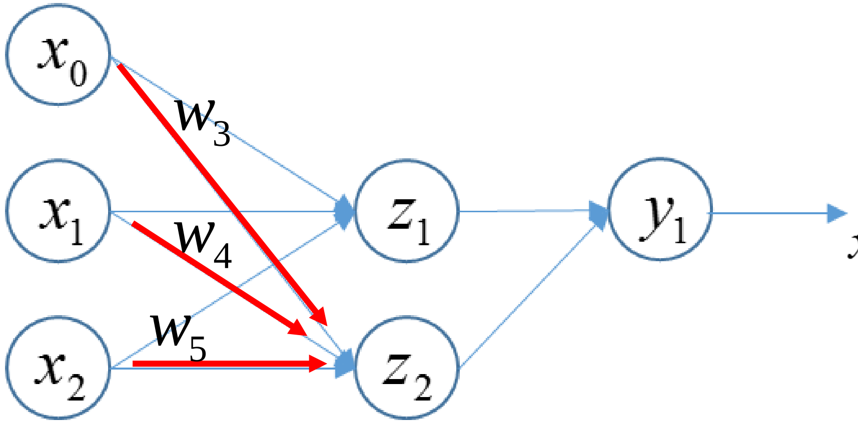


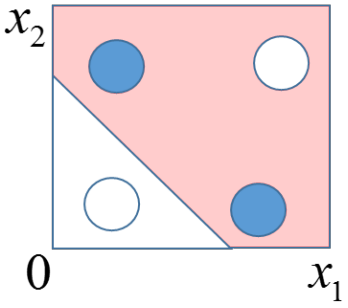
Role of hidden layers: a simple example



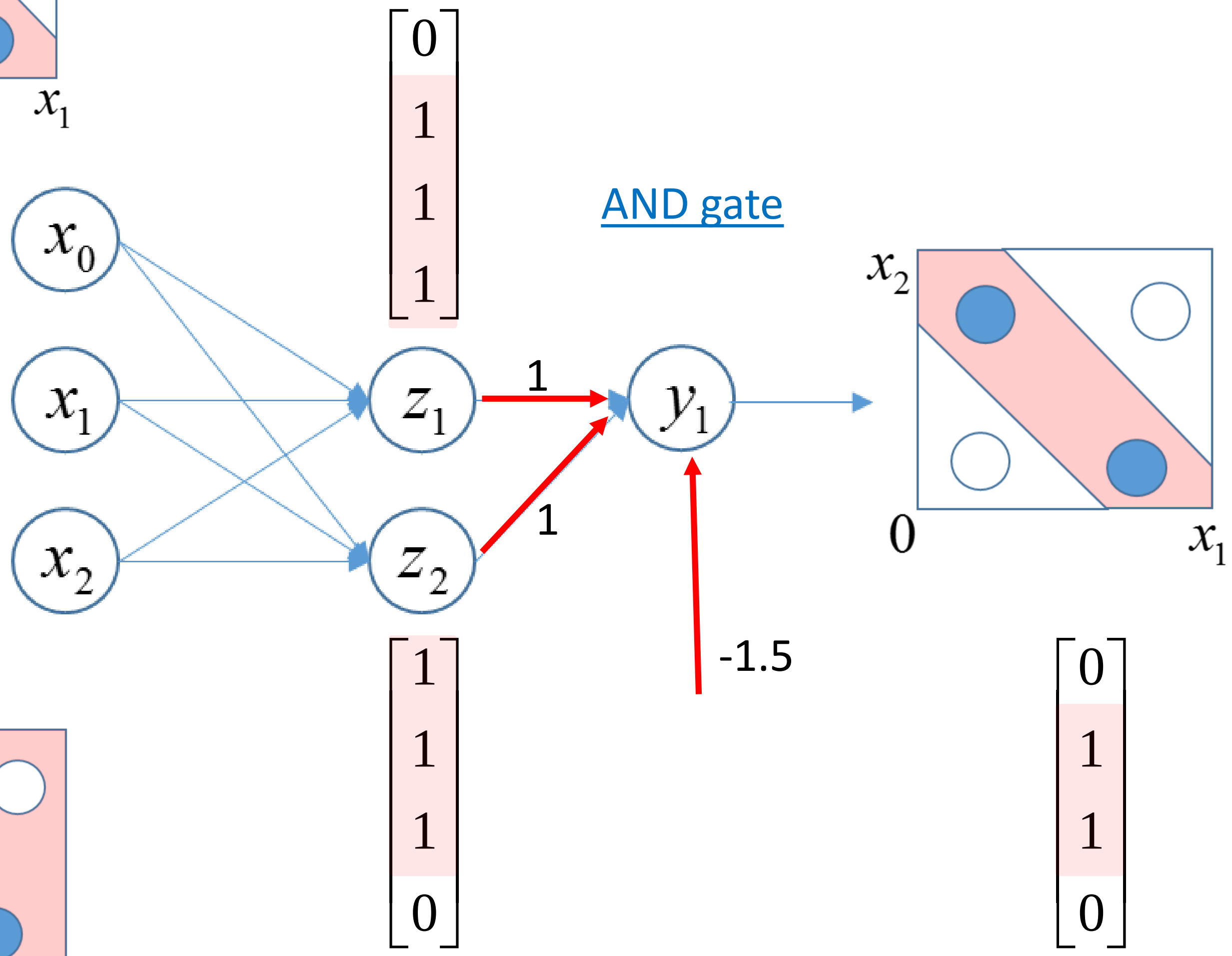


$$z_1 = w_2 x_2 + w_1 x_1 + w_0$$

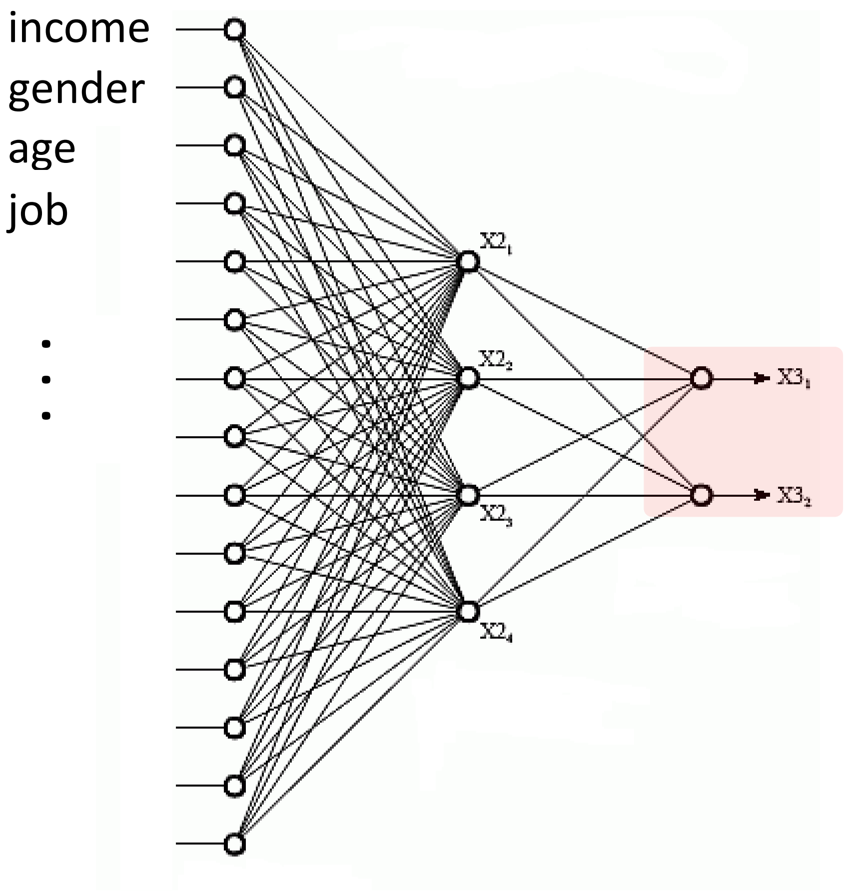




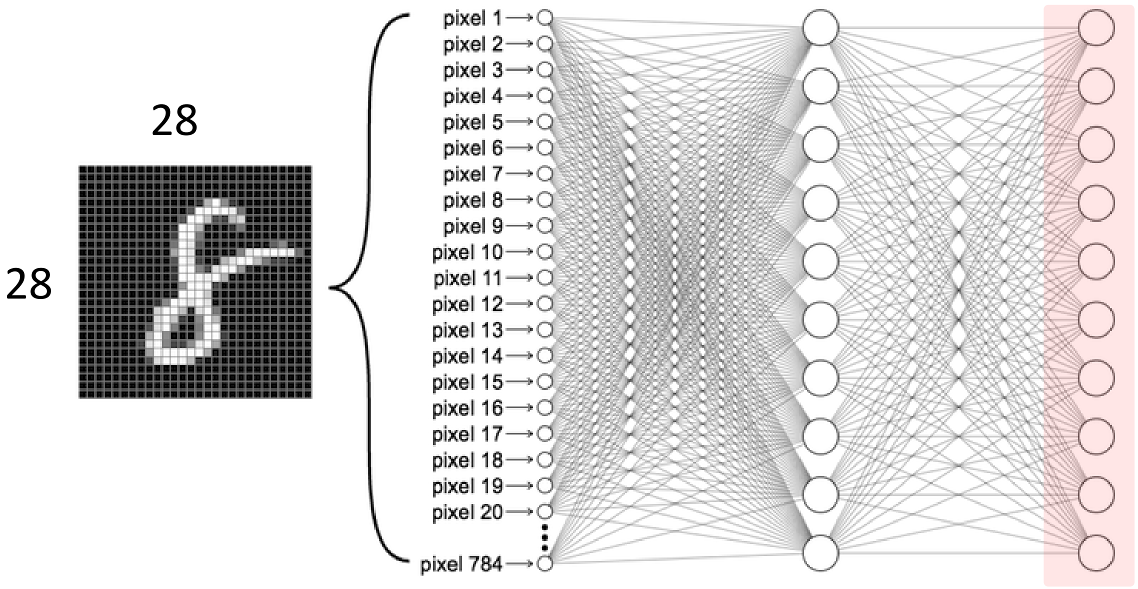
$$z_2 = w_5 x_2 + w_4 x_1 + w_3$$



Example structures of neural networks

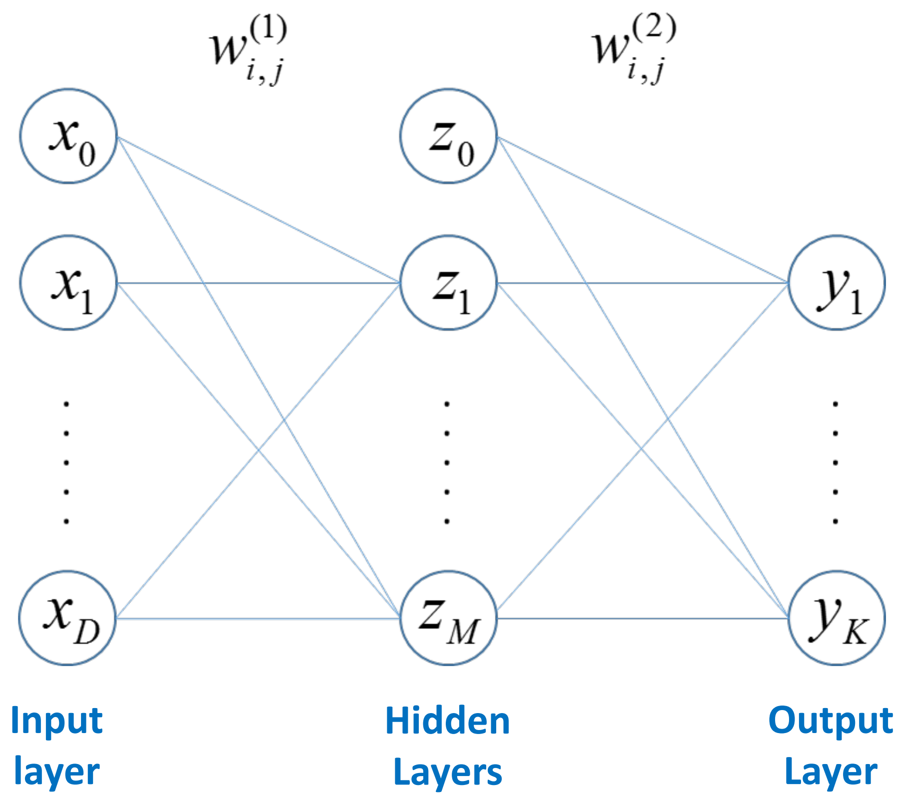


Credit card approval

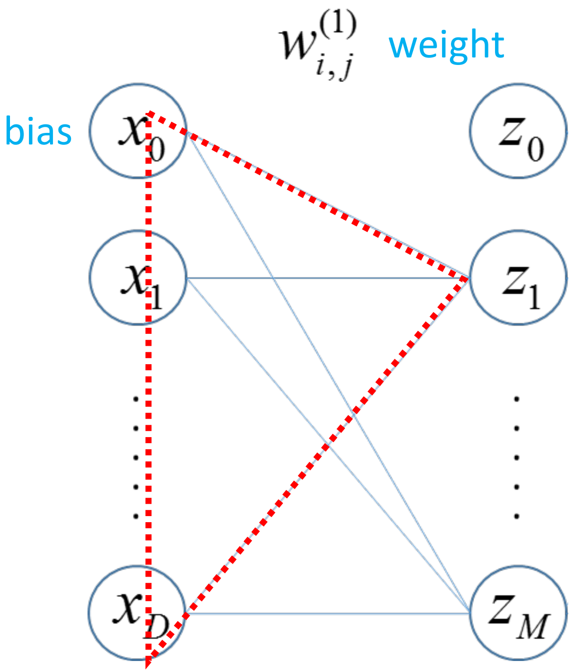


Digit recognition

A neural network model



A neural network model: bias and weight



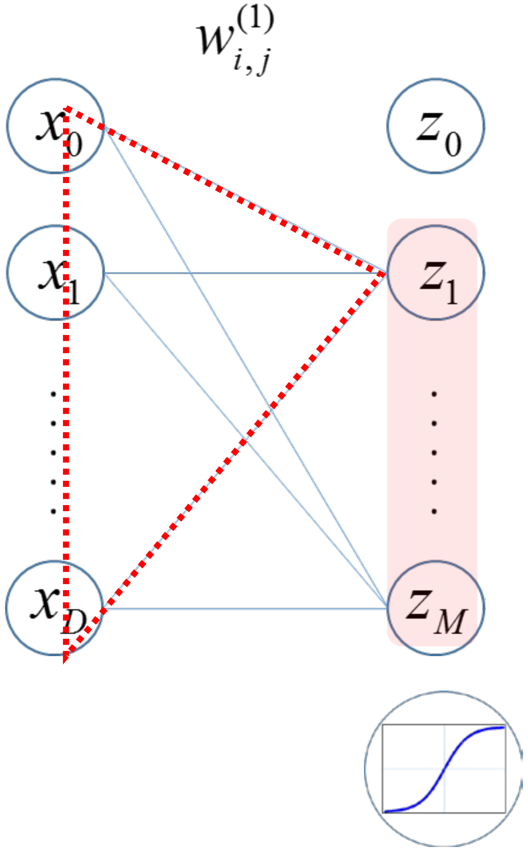
$$\boxed{w_{1,1}^{(1)} x_1 + w_{2,1}^{(1)} x_2 + w_{0,1}^{(1)} x_0}$$

bias

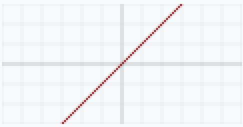
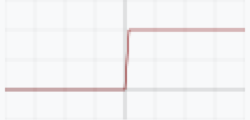
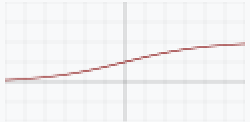
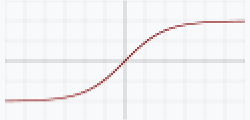
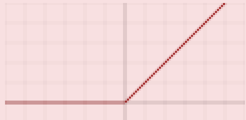
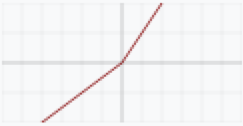
$x_0 = 1$

- ❑ Fully connected vs Partially connected
- ❑ Bias and weight initialization
  - Xarvier (2010) / He (2015)

A neural network model: activation function



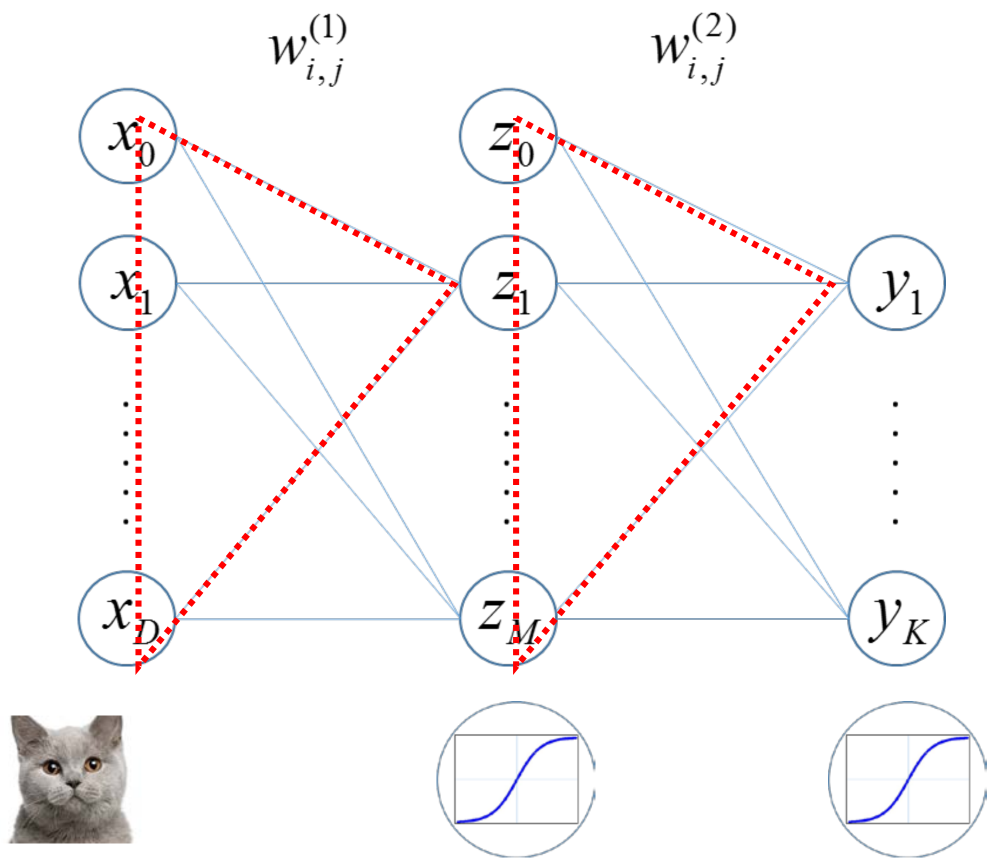
$$\boxed{w_{1,1}^{(1)}x_1 + w_{2,1}^{(1)}x_2 + w_{0,1}^{(1)}x_0}$$

Identity		$f(x) = x$	$f'(x) = 1$
Binary step		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x \neq 0 \\ ? & \text{for } x = 0 \end{cases}$
Logistic (a.k.a. Sigmoid or Soft step)		$f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$ [1]	$f'(x) = f(x)(1 - f(x))$
TanH		$f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$	$f'(x) = 1 - f(x)^2$
Rectified linear unit (ReLU)[11]		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Leaky rectified linear unit (Leaky ReLU)[12]		$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0.01 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$

- ❑ Vanishing gradient problem
- ❑ Convergence speed (6 times fater) [http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf]

https://en.wikipedia.org/wiki/Activation_function

A neural network model: cross entropy with softmax



$w_{1,1}^{(1)} x_1 + w_{2,1}^{(1)} x_2 + w_{0,1}^{(1)} x_0$

$w_{1,1}^{(2)} z_1 + w_{2,1}^{(2)} z_2 + w_{0,1}^{(2)} z_0$

y

output		Sigmoid	Normalization	Softmax
-5	frog	0.00669	0.00334	0.00004
-1	bird	0.26894	0.13447	0.00243
1	dog	0.73106	0.36553	0.01794
5	cat	0.99331	0.49666	0.97959

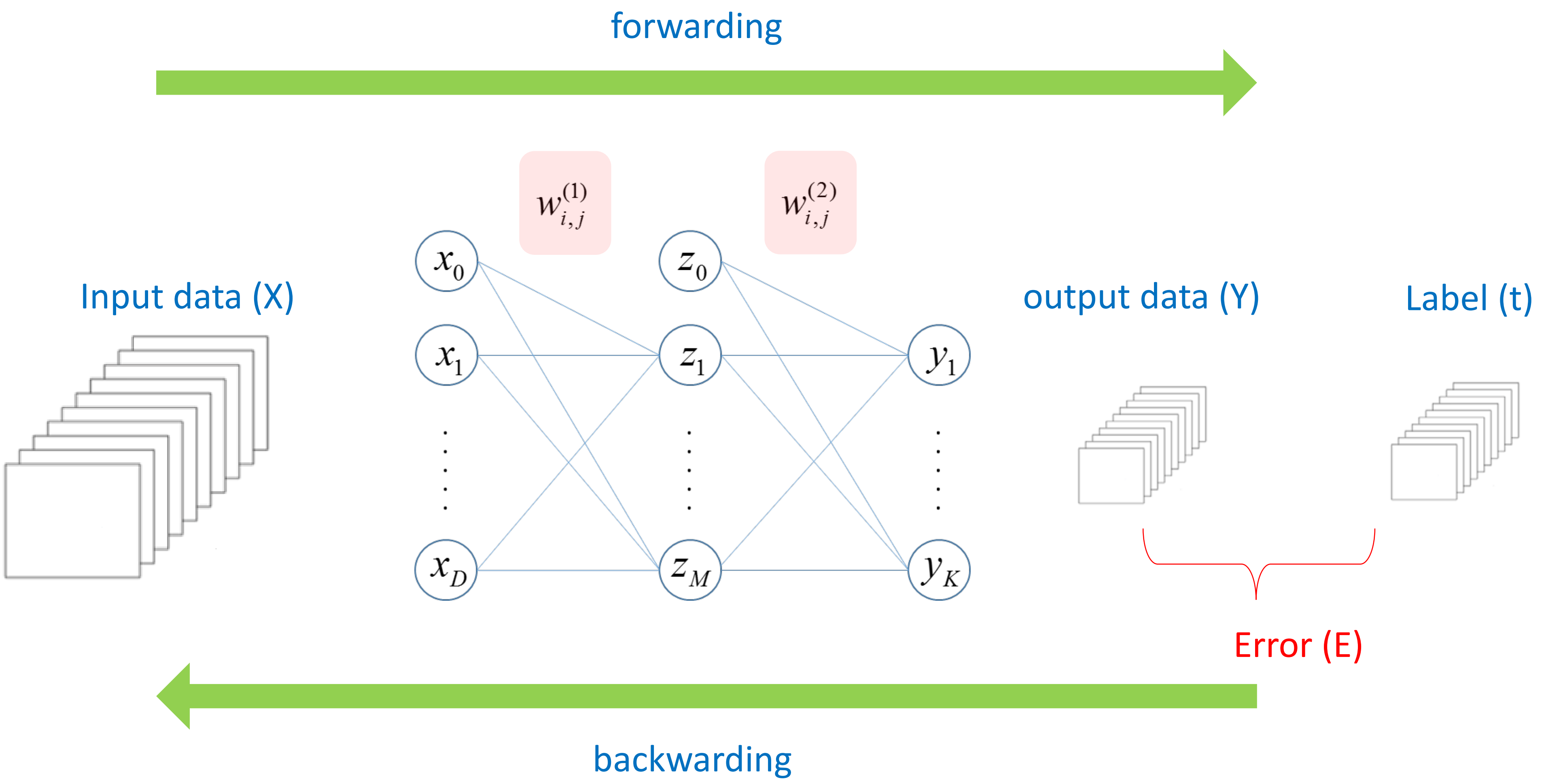
t

Label
0
0
0
1

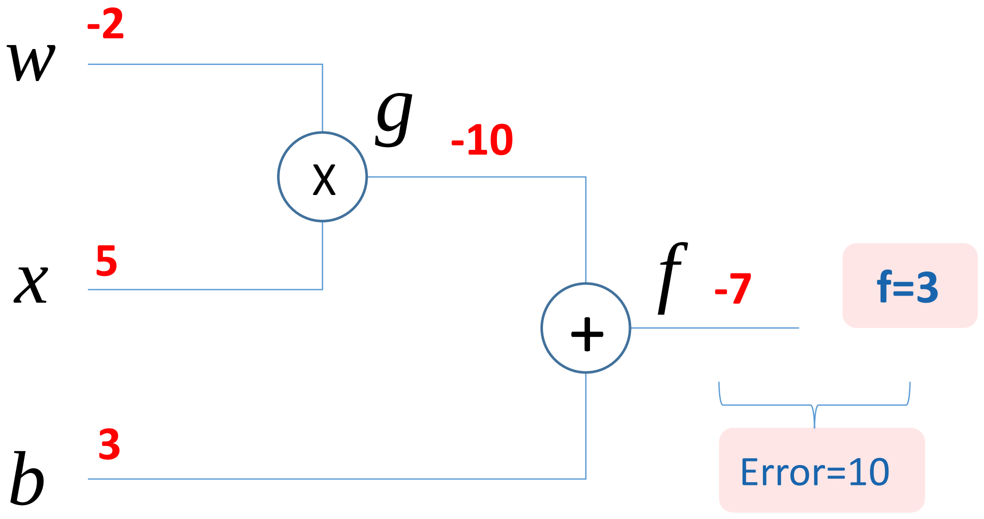
$$H(\mathrm{y}) = -\sum_{\mathrm{i}} \mathrm{t_i} \log(y_i) = 0.020621$$

Operation

Overview of the operation



Backpropagation: a toy example



$$\frac{\partial f}{\partial w} = \frac{\partial f}{\partial g}\frac{\partial g}{\partial w} = x = 5$$

$$\frac{\partial f}{\partial b} = 1$$

$$f = g + b$$

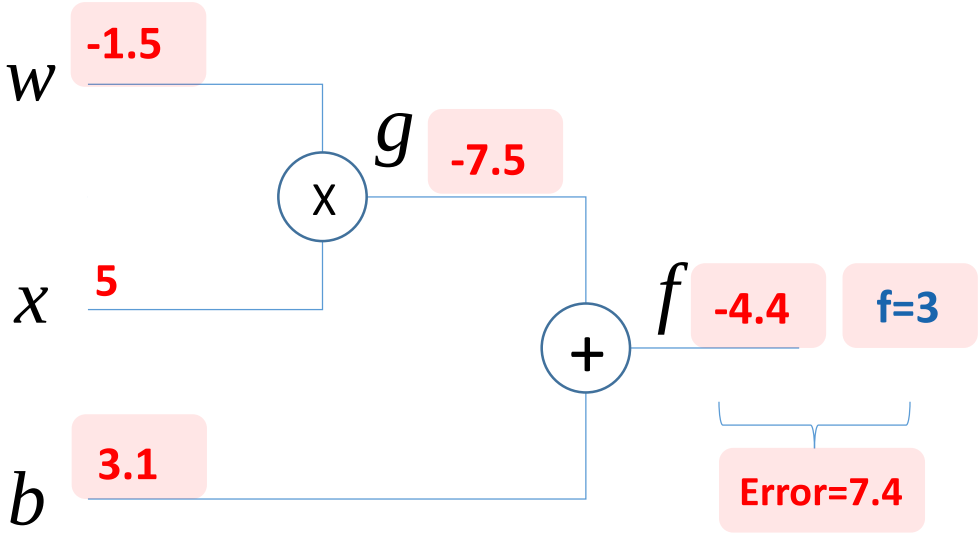
$$g = wx$$

- ❑ Assuming that the value of f should be “3”.
- ❑ How to update variables which you are interested?

$$W_{new} = W_{old} + \eta \frac{\partial f}{\partial w_i} \qquad b_{new} = b_{old} + \eta \frac{\partial f}{\partial b}$$

$$W_{new} = -2 + 0.1 \times 5 = -1.5 \qquad b_{new} = 3 + 0.1 \times 1 = 3.1$$

Backpropagation: a toy example



$$\frac{\partial f}{\partial w} = \frac{\partial f}{\partial g}\frac{\partial g}{\partial w} = x = 5$$

$$\frac{\partial f}{\partial b} = 1$$

$$f = g + b$$

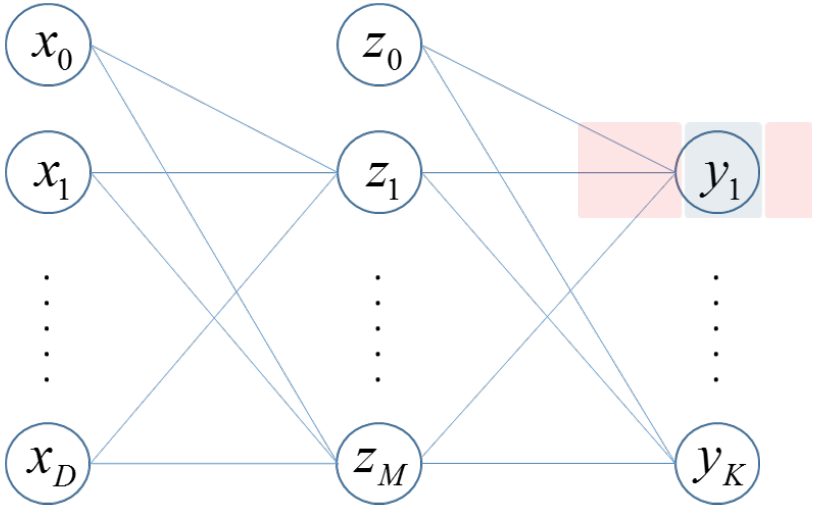
$$g = wx$$

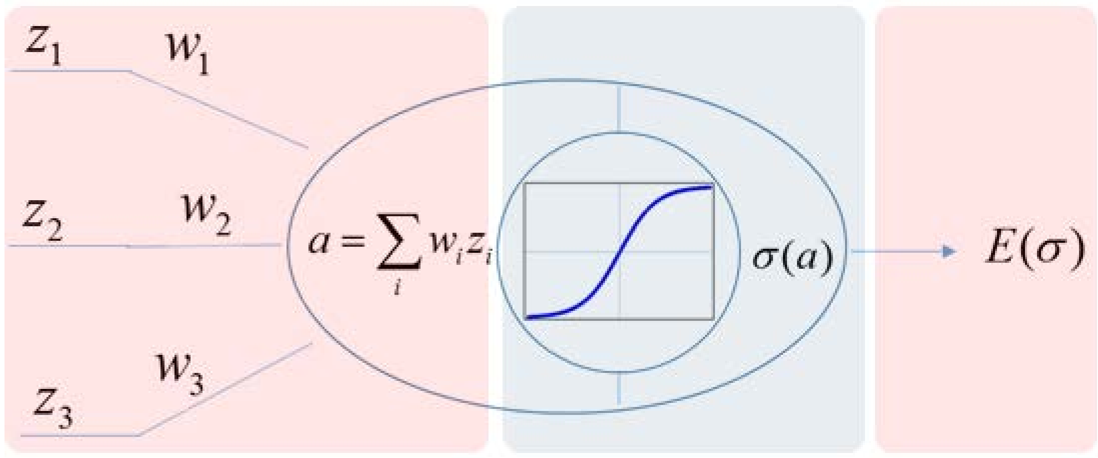
- ❑ Assuming that the value of f should be "3".
- ❑ How to update variables which you are interested?

$$W_{new} = W_{old} + \eta \frac{\partial f}{\partial w_i} \qquad b_{new} = b_{old} + \eta \frac{\partial f}{\partial b}$$

$$W_{new} = -2 + 0.1 \times 5 = -1.5 \qquad b_{new} = 3 + 0.1 \times 1 = 3.1$$

Backpropagation in a neuron





Linear summation function

$$a = w_1 z_1 + w_2 z_2 + w_3 z_3$$

Activation function

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

Error function

$$E(\sigma(a)) = \frac{1}{2} \sum_{i=1} (t_i - \sigma(a))^2$$

$$\frac{\partial E(\sigma(a))}{\partial \mathrm{w}} = \frac{\partial E(\sigma(a))}{\partial \sigma} \times \frac{\partial \sigma}{\partial a} \times \frac{\partial a}{\partial \mathrm{w}}$$

Backpropagation in a neuron

Linear summation function

$$a = w_1 z_1 + w_2 z_2 + w_3 z_3$$

$$\partial a / \partial w_1 = z_1$$

$$\partial a / \partial w_2 = z_2$$

$$\partial a / \partial w_3 = z_3$$

Activation function

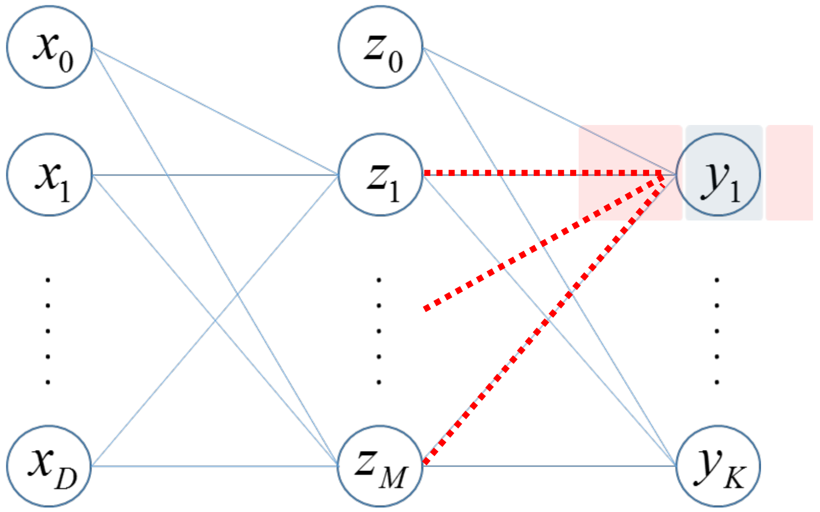
$$\sigma(a) = \frac{1}{1+e^{-a}}$$

$$\frac{\partial \sigma}{\partial a} = \sigma(a)(1-\sigma(a))$$

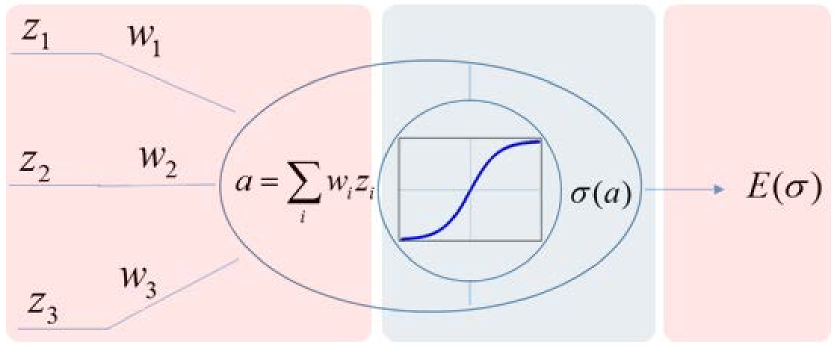
Error function

$$E(\sigma(a)) = \frac{1}{2}\sum_{i=1}(t_i - \sigma(a))^2$$

$$\frac{\partial E(\sigma(a))}{\partial \sigma} = -(t_i - \sigma(a))$$



Linear summation function **Activation function** **Error function**



$$\frac{\partial E(w)}{\partial w_1} = \frac{\partial E(\sigma(a))}{\partial \sigma} \times \frac{\partial \sigma}{\partial a} \times \frac{\partial a}{\partial w_1}$$

$$\frac{\partial E(w)}{\partial w_2} = \frac{\partial E(\sigma(a))}{\partial \sigma} \times \frac{\partial \sigma}{\partial a} \times \frac{\partial a}{\partial w_2}$$

$$\frac{\partial E(w)}{\partial w_3} = \frac{\partial E(\sigma(a))}{\partial \sigma} \times \frac{\partial \sigma}{\partial a} \times \frac{\partial a}{\partial w_3}$$

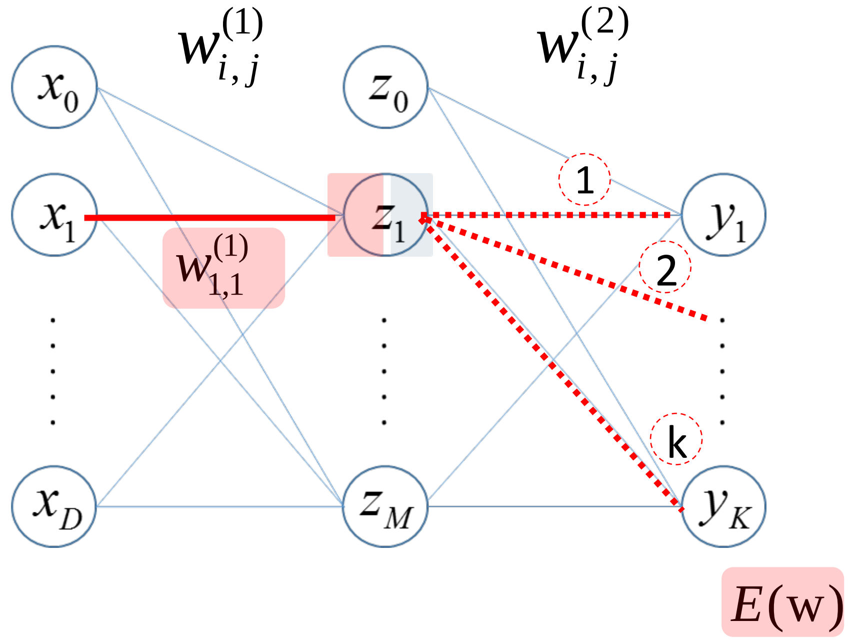
Backpropagation in the next layer

$$\frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}} = \frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}} \times \frac{\partial z_{1(ouput)}}{\partial z_{1(input)}} \times \frac{\partial z_{1(input)}}{\partial w_{1,1}^{(1)}}$$

$$\frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}} = \overset{1}{\frac{\partial E(\mathrm{w})_{y_1}}{\partial z_{1(ouput)}}} + \overset{2}{\frac{\partial E(\mathrm{w})_{y_2}}{\partial z_{1(ouput)}}} + \cdots + \overset{k}{\frac{\partial E(\mathrm{w})_{y_K}}{\partial z_{1(ouput)}}}$$

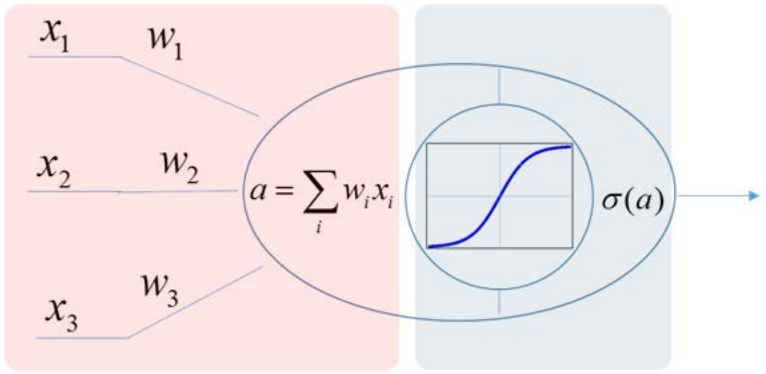
$$\frac{\partial E(\mathrm{w})_{y_1}}{\partial z_{1(ouput)}} = \underbrace{\frac{\partial E(\mathrm{w})_{y_1}}{\partial y_{1(output)}} \times \frac{\partial y_{1(output)}}{\partial y_{1(input)}}}_{\text{Previously obtained}} \times \underbrace{\frac{\partial y_{1(input)}}{\partial z_{1(ouput)}}}_{w_{1,1}^{(2)}}$$

$$\boxed{y_1 = w_{1,1}^{(2)} z_1 + w_{2,1}^{(2)} z_2 + w_{0,1}^{(2)} z_0}$$



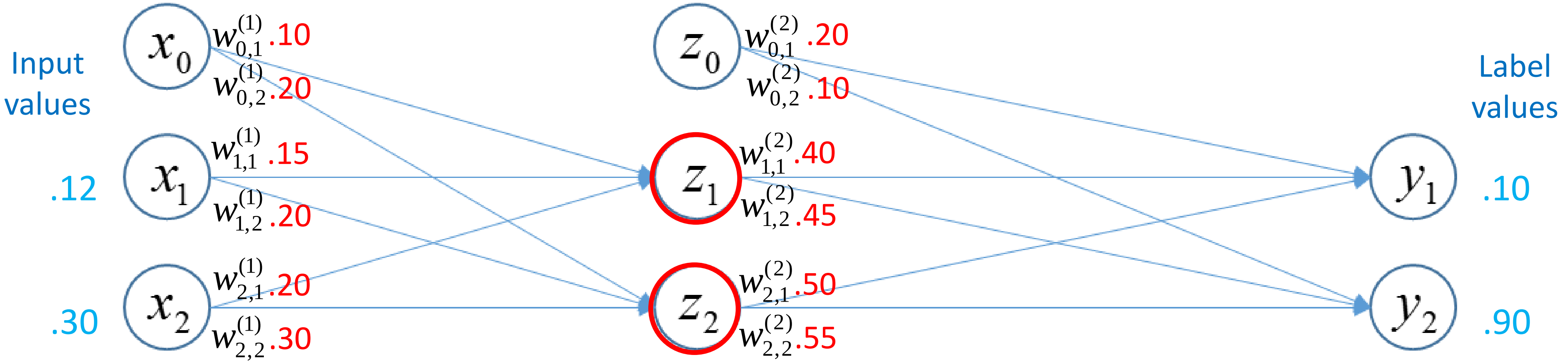
Linear summation function

Activation function

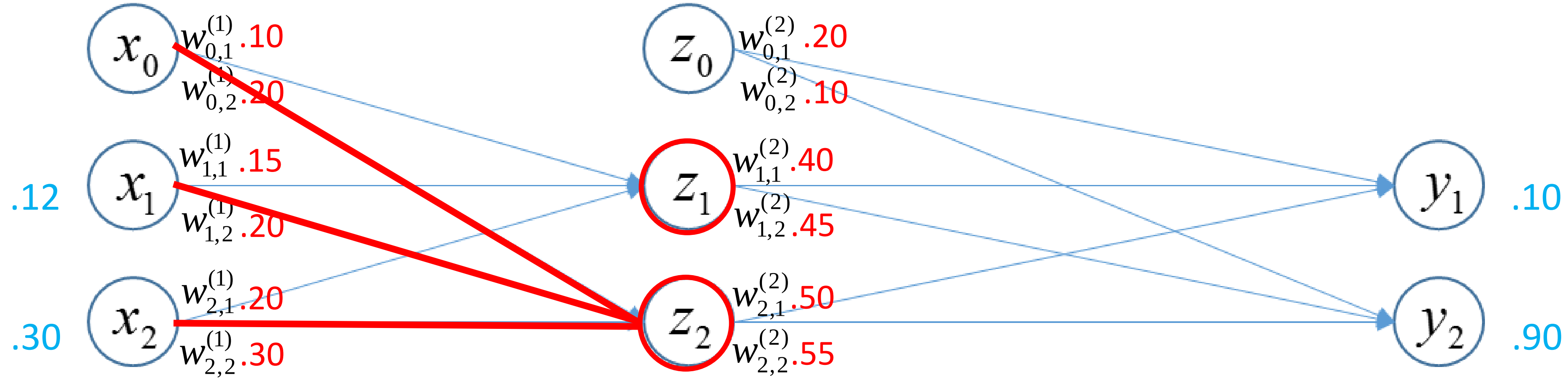


Backpropagation example

Backpropagation algorithm - Forwarding



Backpropagation algorithm - Forwarding



$$a_1 = w^{(1)}_{1,1}x_1 + w^{(1)}_{2,1}x_2 + w^{(1)}_{0,1}$$

$$a_1 = 0.15 \times 0.12 + 0.2 \times 0.3 + 0.1$$

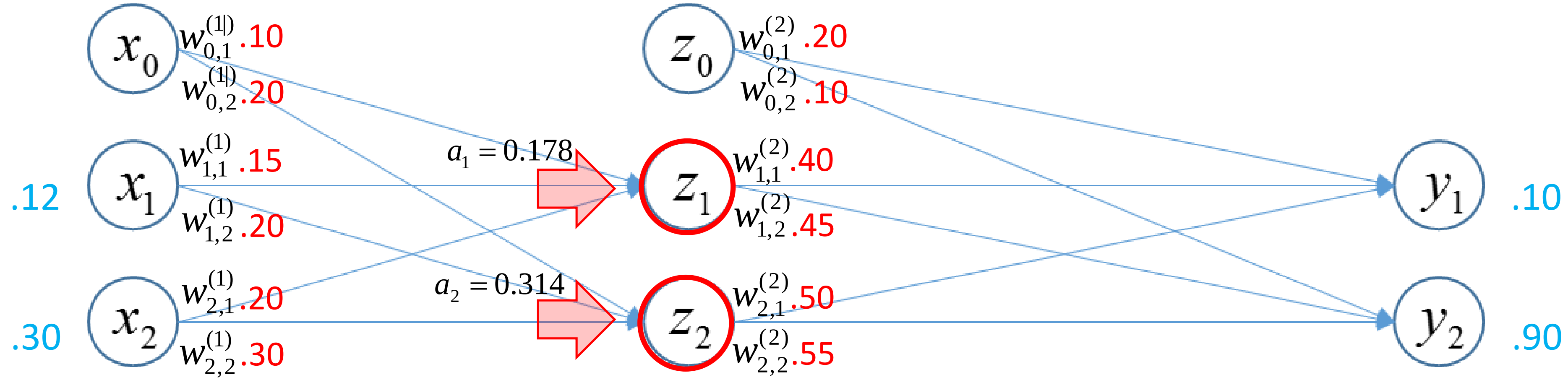
$$a_1 = 0.178$$

$$a_2 = w^{(1)}_{1,2}x_1 + w^{(1)}_{2,2}x_2 + w^{(1)}_{0,2}$$

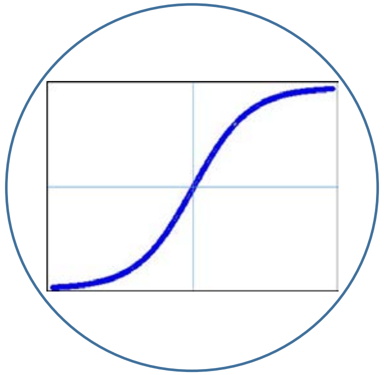
$$a_2 = 0.2 \times 0.12 + 0.3 \times 0.3 + 0.2$$

$$a_1 = 0.314$$

Backpropagation algorithm - Forwarding



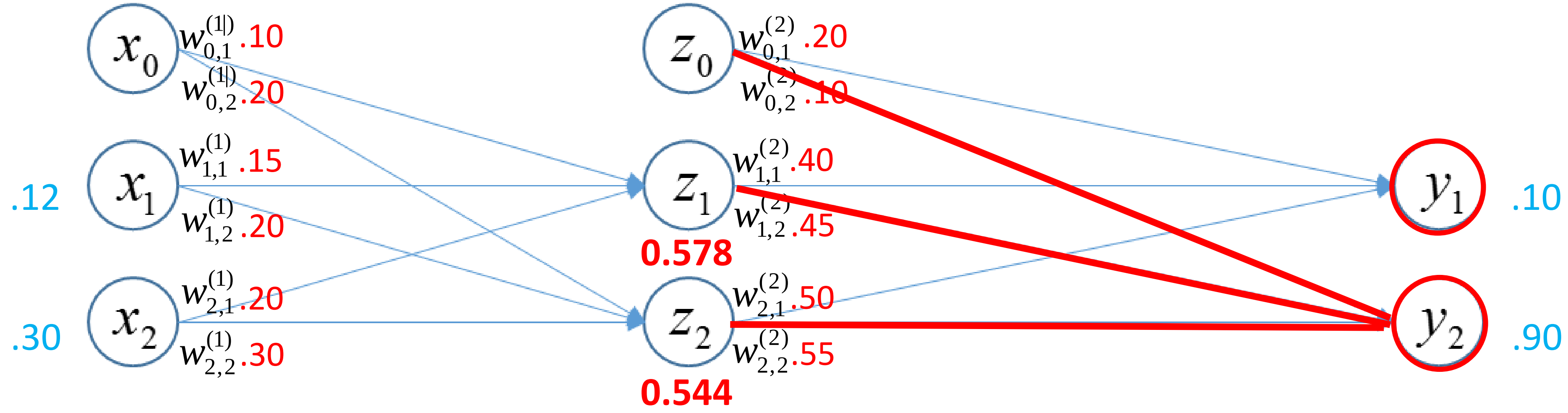
$$a_1 = 0.178$$
$$a_2 = 0.314$$



$$\sigma(a_1) = 0.544$$
$$\sigma(a_2) = 0.578$$

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

Backpropagation algorithm - Forwarding



$$a_1 = w^{(2)}_{1,1} z_1 + w^{(2)}_{2,1} z_2 + w^{(2)}_{0,1}$$

$$a_1 = 0.40 \times 0.578 + 0.5 \times 0.544 + 0.2$$

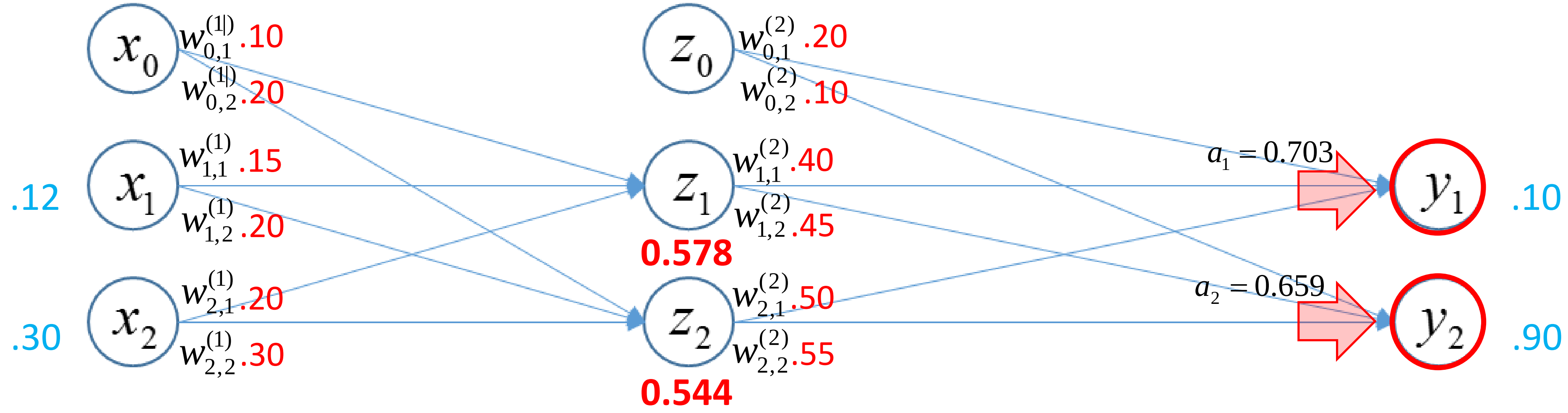
$$a_1 = 0.703$$

$$a_2 = w^{(2)}_{1,2} z_1 + w^{(2)}_{2,2} z_2 + w^{(2)}_{0,2}$$

$$a_2 = 0.45 \times 0.578 + 0.55 \times 0.544 + 0.1$$

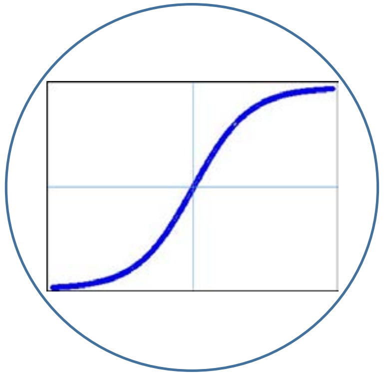
$$a_2 = 0.659$$

Backpropagation algorithm - Forwarding



$a_1 = 0.703$

$a_2 = 0.659$

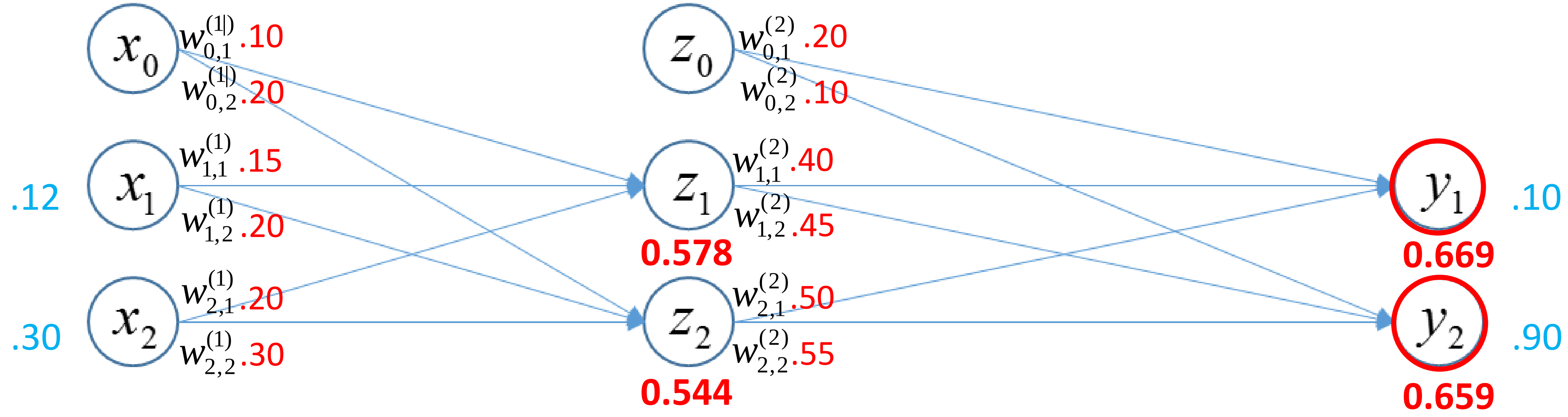


$\sigma(a_1) = 0.669$

$\sigma(\alpha_2) = 0.659$

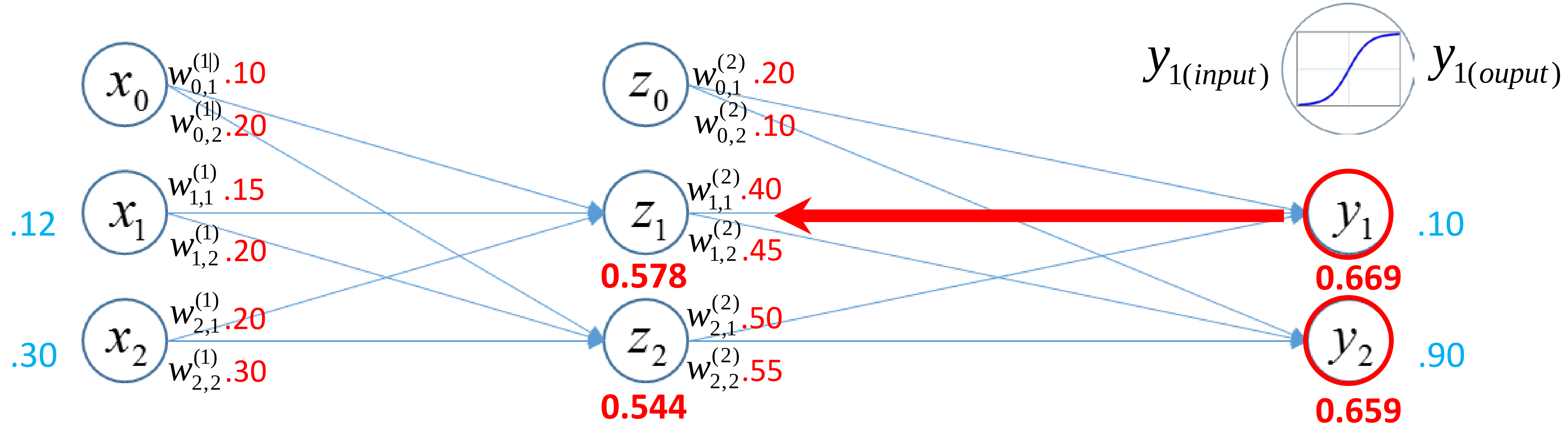
$$\sigma(a) = \frac{1}{1+e^{-a}}$$

Backpropagation algorithm - Forwarding



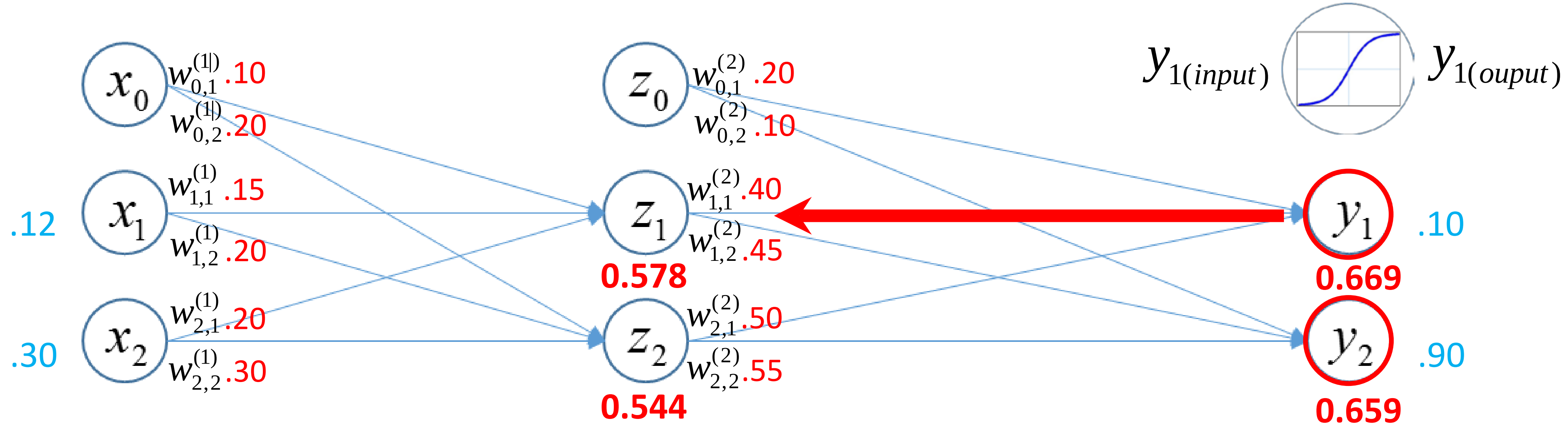
$$E(\mathrm{w}) = \frac{1}{2}\left((0.1 - 0.669)^2 + (0.9 - 0.659)^2\right) = 0.191$$

Backpropagation algorithm - Backwarding



$$\frac{\partial y_{1(input)}}{\partial w_{1,1}^{(2)}} \times \frac{\partial y_{1(ouput)}}{\partial y_{1(input)}} \times \frac{\partial E(\mathrm{w})}{\partial y_{1(ouput)}} = \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(2)}}$$

Backpropagation algorithm - Backwarding



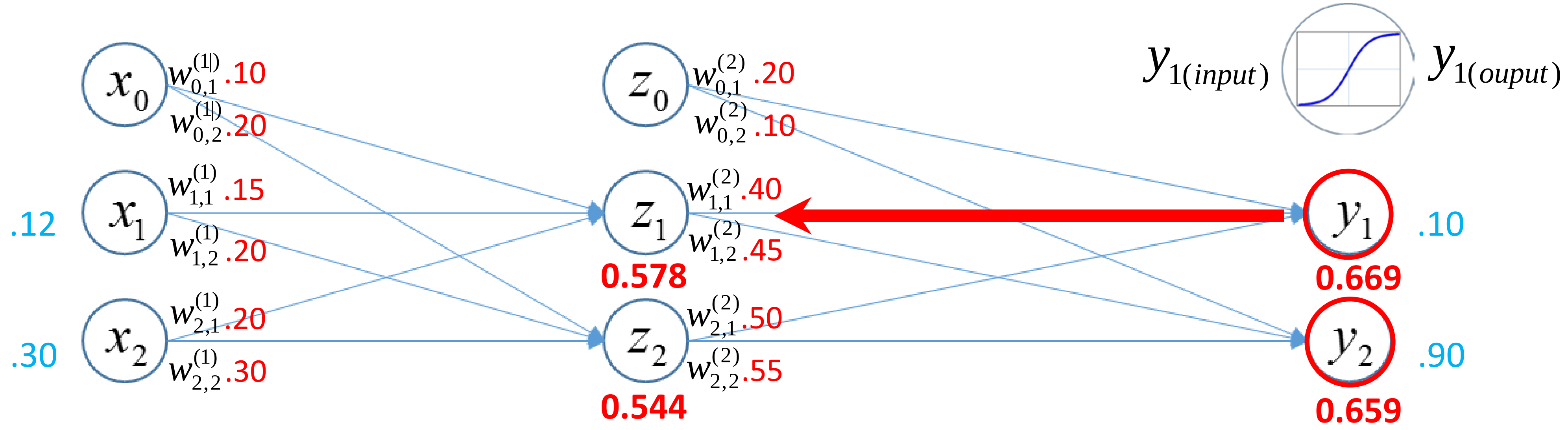
$$\frac{\partial y_{1(input)}}{\partial w_{1,1}^{(2)}} \times \frac{\partial y_{1(ouput)}}{\partial y_{1(input)}} \times \boxed{\frac{\partial E(\mathrm{w})}{\partial y_{1(ouput)}}} = \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(2)}}$$

0.569

$$E(\mathrm{w}) = \frac{1}{2}\left((0.1 - y_{1(output)})^2 + (0.9 - y_{2(output)})^2\right)$$

$$\frac{\partial E(\mathrm{w})}{\partial y_{1(output)}} = -(0.1 - y_{1(output)}) = -(0.1 - 0.669) = 0.569$$

Backpropagation algorithm - Backwarding



$$\frac{\partial y_{1(input)}}{\partial w^{(2)}_{1,1}} \times \frac{\partial y_{1(ouput)}}{\partial y_{1(input)}} \times \frac{\partial E(\mathrm{w})}{\partial y_{1(ouput)}} = \frac{\partial E(\mathrm{w})}{\partial w^{(2)}_{1,1}}$$

0.221 **0.569**

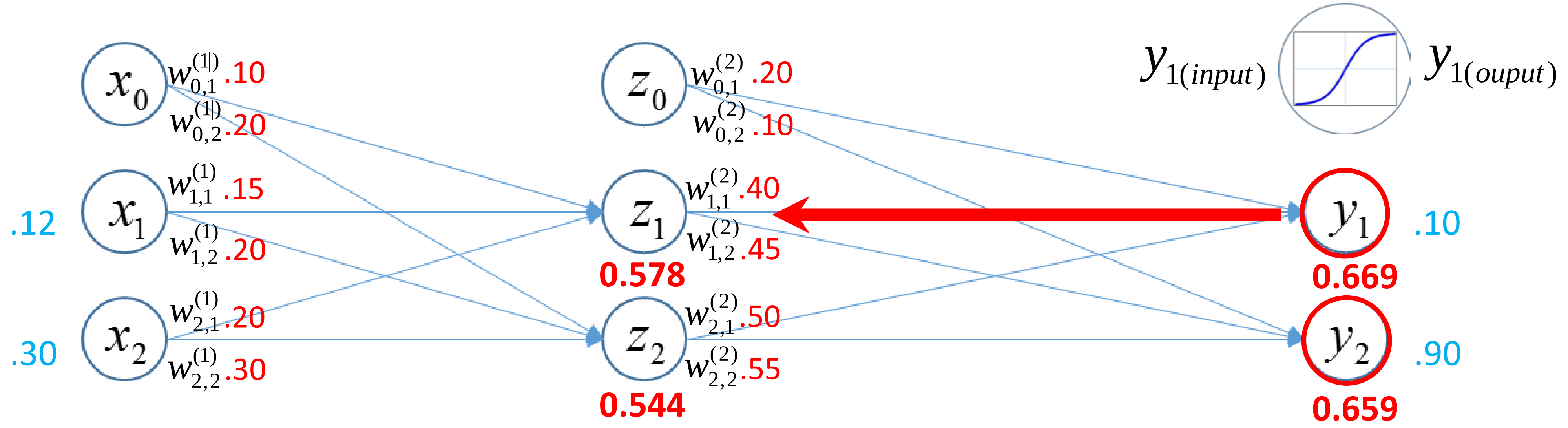
$$y_{1(ouput)} = \frac{1}{1+e^{-y_{1(input)}}}$$

$$\frac{\partial y_{1(output)}}{\partial y_{1(input)}} = \sigma(y_{1(input)})\left(1-\sigma(y_{1(input)})\right)$$

$$= y_{1(output)}(1-y_{1(output)})$$

$$= 0.669\times(1-0.669) = 0.221$$

Backpropagation algorithm - Backwarding



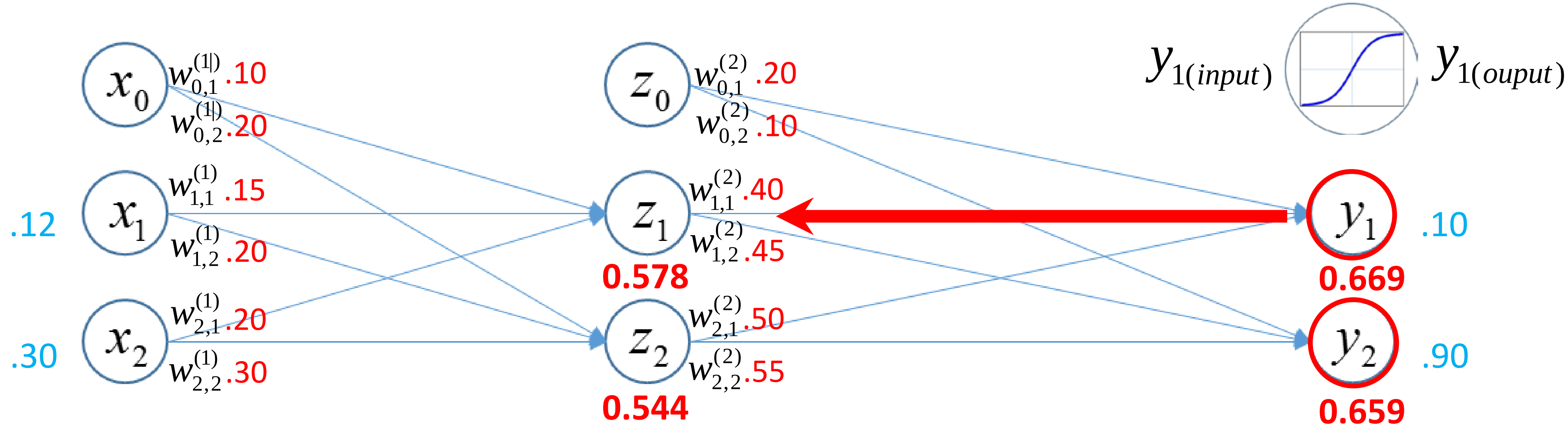
$$\boxed{\frac{\partial y_{1(input)}}{\partial w_{1,1}^{(2)}}} \times \frac{\partial y_{1(ouput)}}{\partial y_{1(input)}} \times \frac{\partial E(\mathrm{w})}{\partial y_{1(ouput)}} = \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(2)}}$$

0.578 **0.221** **0.569**

$$y_{1(input)} = w_{1,1}^{(2)} z_1 + w_{2,1}^{(2)} z_2 + w_{0,1}^{(2)}$$

$$\frac{\partial y_{1(input)}}{\partial w_{1,1}^{(2)}} = z_1 = 0.578$$

Backpropagation algorithm - Backwarding



$$\boxed{w_{1,1}^{(2)}*} = w_{1,1}^{(2)} + \eta \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(2)}}$$

$$= 0.4 + 0.5 \times 0.0727$$

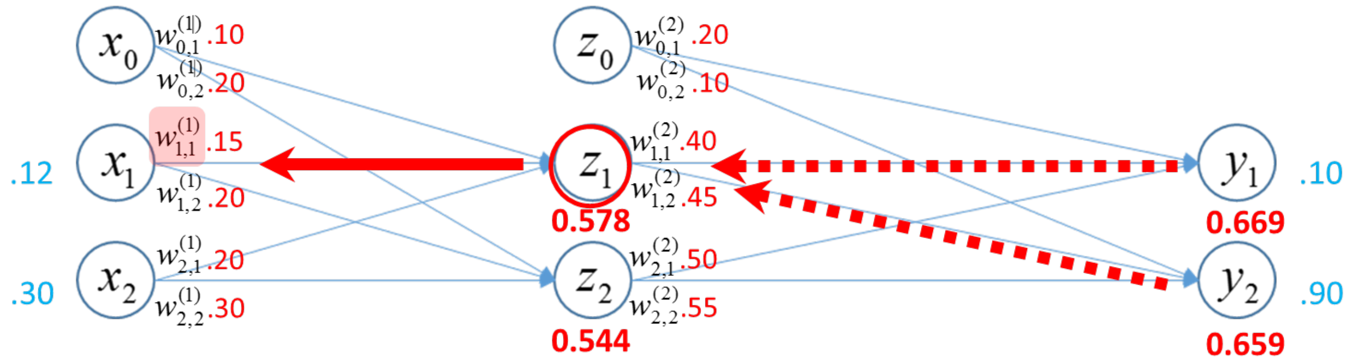
$$= \boxed{\mathbf{0.436}}$$

$$\underset{\mathbf{0.578}}{\frac{\partial y_{1(input)}}{\partial w_{1,1}^{(2)}}} \times \underset{\mathbf{0.221}}{\frac{\partial y_{1(ouput)}}{\partial y_{1(input)}}} \times \underset{\mathbf{0.569}}{\frac{\partial E(\mathrm{w})}{\partial y_{1(ouput)}}} = \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(2)}}$$

$$\frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(2)}} = \mathbf{0.0727}$$

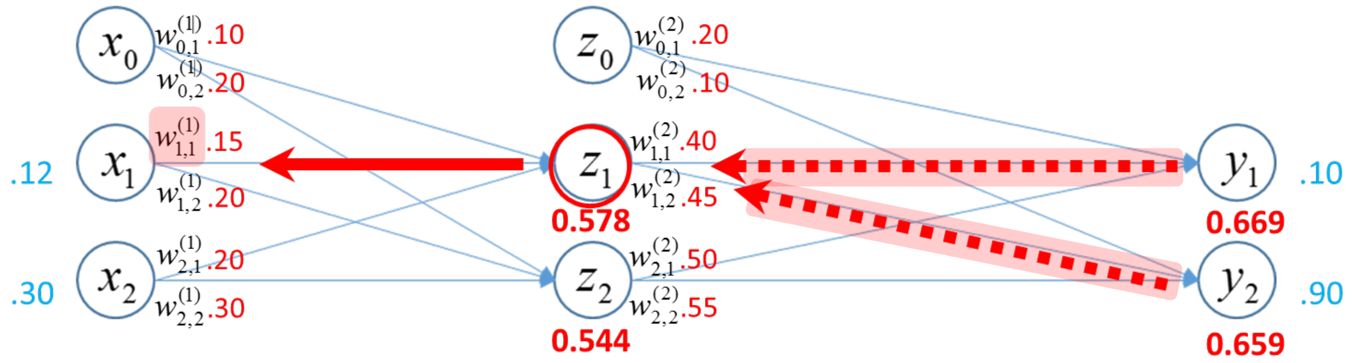
Same procedures are applied for $\boxed{w_{1,2}^{(2)}, w_{2,1}^{(2)}, w_{2,2}^{(2)}}$

Backpropagation algorithm - Backwarding



$$\frac{\partial z_{1(input)}}{\partial w_{1,1}^{(1)}} \times \frac{\partial z_{1(ouput)}}{\partial z_{1(input)}} \times \frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}} = \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}}$$

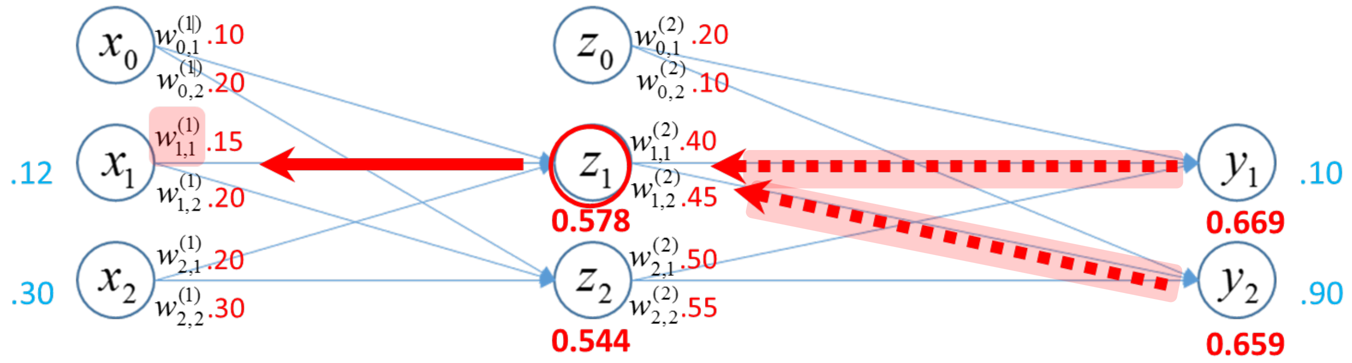
Backpropagation algorithm - Backwarding



$$\frac{\partial z_{1(input)}}{\partial w_{1,1}^{(1)}} \times \frac{\partial z_{1(ouput)}}{\partial z_{1(input)}} \times \frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}} = \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}}$$

$$\frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}} = \frac{\partial E(\mathrm{w})_{y_1}}{\partial z_{1(ouput)}} + \frac{\partial E(\mathrm{w})_{y_2}}{\partial z_{1(ouput)}}$$

Backpropagation algorithm - Backwarding



$$y_{2(input)} = w^{(2)}_{1,2} z_1 + w^{(2)}_{2,2} z_2 + w^{(2)}_{0,2}$$

$$\frac{\partial y_{2(input)}}{\partial z_{1(output)}} = w^{(2)}_{1,2} = 0.45$$

$$\frac{\partial z_{1(input)}}{\partial w^{(1)}_{1,1}} \times \frac{\partial z_{1(ouput)}}{\partial z_{1(input)}} \times \frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}} = \frac{\partial E(\mathrm{w})}{\partial w^{(1)}_{1,1}}$$

0.0259

$$\frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}} = \frac{\partial E(\mathrm{w})_{y_1}}{\partial z_{1(ouput)}} + \frac{\partial E(\mathrm{w})_{y_2}}{\partial z_{1(ouput)}}$$

$$= 0.0503 - 0.0244 = 0.0259$$

Previously calculated

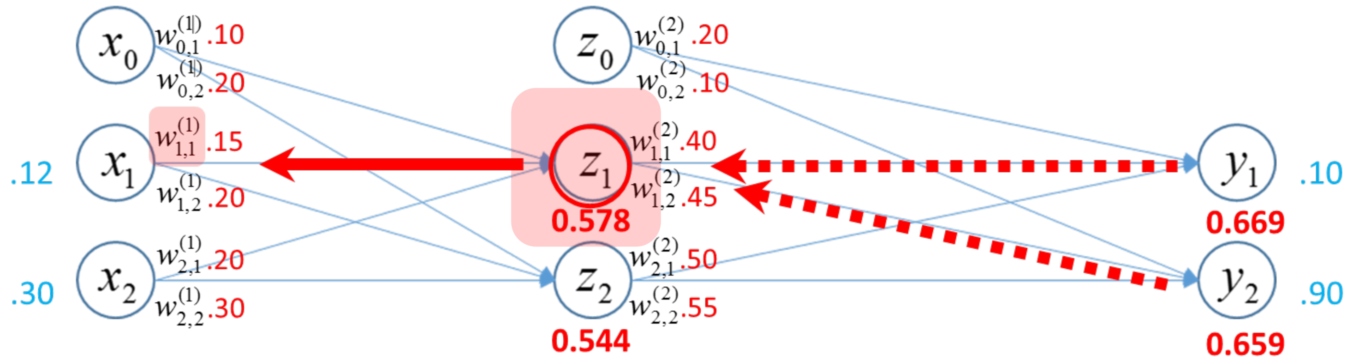
$$\frac{\partial E(\mathrm{w})_{y_1}}{\partial z_{1(ouput)}} = \frac{\partial E(\mathrm{w})_{y_1}}{\partial y_{1(output)}} \times \frac{\partial y_{1(output)}}{\partial y_{1(input)}} \times \frac{\partial y_{1(input)}}{\partial z_{1(ouput)}}$$

0.0503 **0.569** **0.221** **0.4**

$$\frac{\partial E(\mathrm{w})_{y_2}}{\partial z_{1(ouput)}} = \frac{\partial E(\mathrm{w})_{y_2}}{\partial y_{2(output)}} \times \frac{\partial y_{2(output)}}{\partial y_{2(input)}} \times \frac{\partial y_{2(input)}}{\partial z_{2(ouput)}}$$

-0.0244 **-0.241** **0.225** **0.45**

Backpropagation algorithm - Backwarding

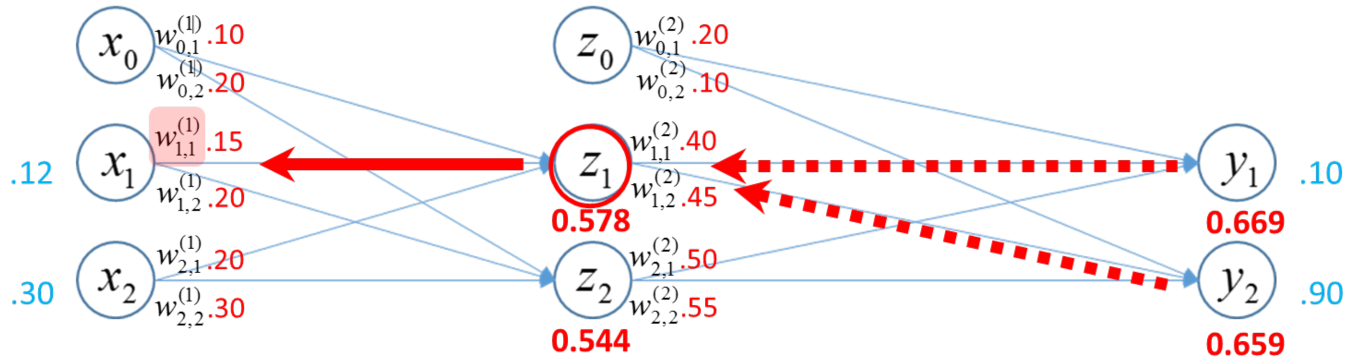


$$\frac{\partial z_{1(input)}}{\partial w_{1,1}^{(1)}} \times \underbrace{\frac{\partial z_{1(ouput)}}{\partial z_{1(input)}}}_{0.2439} \times \underbrace{\frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}}}_{0.0259} = \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}}$$

$$z_{1(ouput)} = \frac{1}{1+e^{-z_{1(input)}}}$$

$$\begin{aligned}\frac{\partial z_{1(output)}}{\partial z_{1(input)}} &= \sigma(z_{1(input)})\left(1-\sigma(z_{1(input)})\right) \\ &= z_{1(output)}\left(1-z_{1(output)}\right) \\ &= 0.578\times\left(1-0.578\right) = 0.2439\end{aligned}$$

Backpropagation algorithm - Backwarding



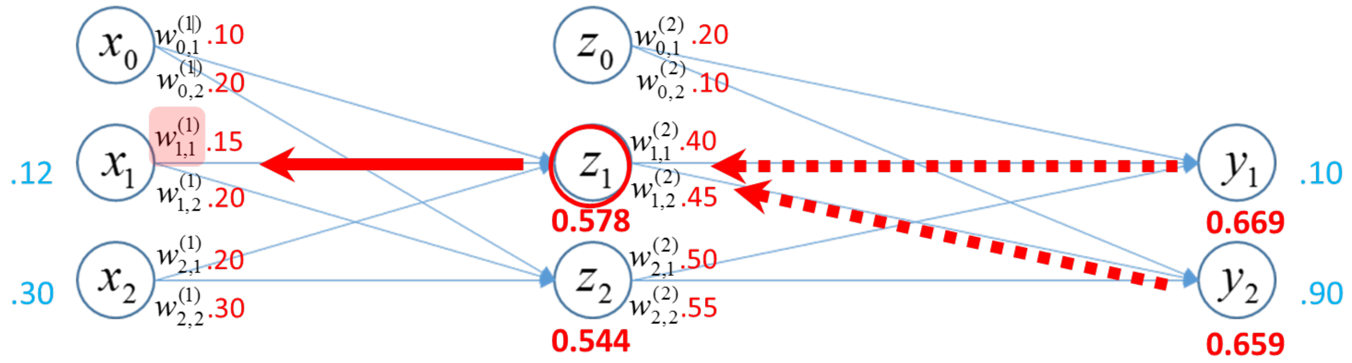
$$\frac{\partial z_{1(input)}}{\partial w_{1,1}^{(1)}} \times \frac{\partial z_{1(ouput)}}{\partial z_{1(input)}} \times \frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}} = \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}}$$

0.12 0.2439 0.0259

$$z_{1(input)} = w_{1,1}^{(1)} x_1 + w_{2,1}^{(1)} x_2 + w_{0,1}^{(1)}$$

$$\frac{\partial z_{1(input)}}{\partial w_{1,1}^{(1)}} = x_1 = 0.12$$

Backpropagation algorithm - Backwarding



$$\frac{\partial z_{1(input)}}{\partial w_{1,1}^{(1)}} \times \frac{\partial z_{1(ouput)}}{\partial z_{1(input)}} \times \frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}} = \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}}$$

0.12 **0.2439** **0.0259**

$$\boxed{\frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}}} = \mathbf{0.00075}$$

$$w_{1,1}^{(1)*} = w_{1,1}^{(1)} + \eta \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}}$$

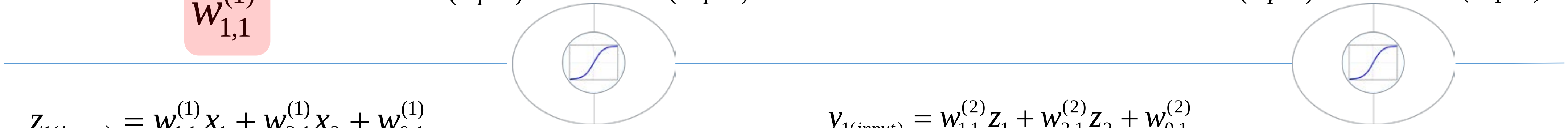
$$= 0.15 + 0.5 \times 0.00075$$

$$= \boxed{\mathbf{0.1504}}$$

Same procedures are applied for $w_{1,2}^{(1)}, w_{2,1}^{(1)}, w_{2,2}^{(1)}$

Backpropagation algorithm - Backwarding

$w_{1,1}^{(1)}$ $\quad z_{1(input)} \quad z_{1(ouput)} \quad y_{1(input)} \quad y_{1(ouput)} \quad E(\mathrm{w})$



$$z_{1(input)} = w_{1,1}^{(1)} x_1 + w_{2,1}^{(1)} x_2 + w_{0,1}^{(1)}$$

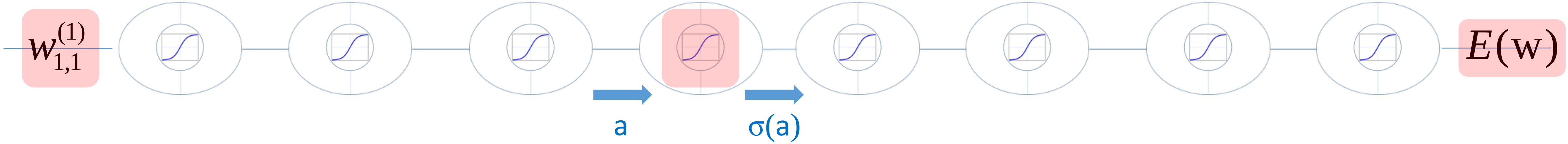
$$z_{1(ouput)} = \frac{1}{1+e^{-z_{1(input)}}}$$

$$y_{1(input)} = w_{1,1}^{(2)} z_1 + w_{2,1}^{(2)} z_2 + w_{0,1}^{(2)}$$

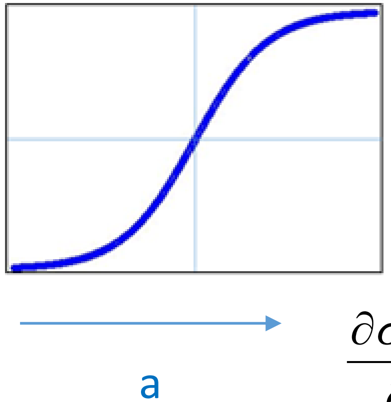
$$y_{1(ouput)} = \frac{1}{1+e^{-y_{1(input)}}}$$

$$\frac{\partial z_{1(input)}}{\partial w_{1,1}^{(1)}} \times \frac{\partial z_{1(ouput)}}{\partial z_{1(input)}} \times \frac{\partial y_{1(output)}}{\partial y_{1(input)}} \times \frac{\partial E(\mathrm{w})}{\partial y_{1(output)}} = \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}}$$

Vanishing gradient problem



$$\frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}} = \frac{\partial E(\mathrm{w})}{\partial z_{1(ouput)}} \times \cdots\cdots\cdots \times \frac{\partial \sigma(a)}{\partial a} \times \cdots\cdots\cdots \times \frac{\partial z_{1(input)}}{\partial w_{1,1}^{(1)}}$$



$$\frac{\partial \sigma(a)}{\partial a} \approx 0$$

$$w_{new} = w_{old} + \eta \frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}}$$

$$\frac{\partial E(\mathrm{w})}{\partial w_{1,1}^{(1)}} \approx 0$$

Gradient vanishing

Backup Slides