Erasmus+ course on machine learning, lecture November 27, 2023

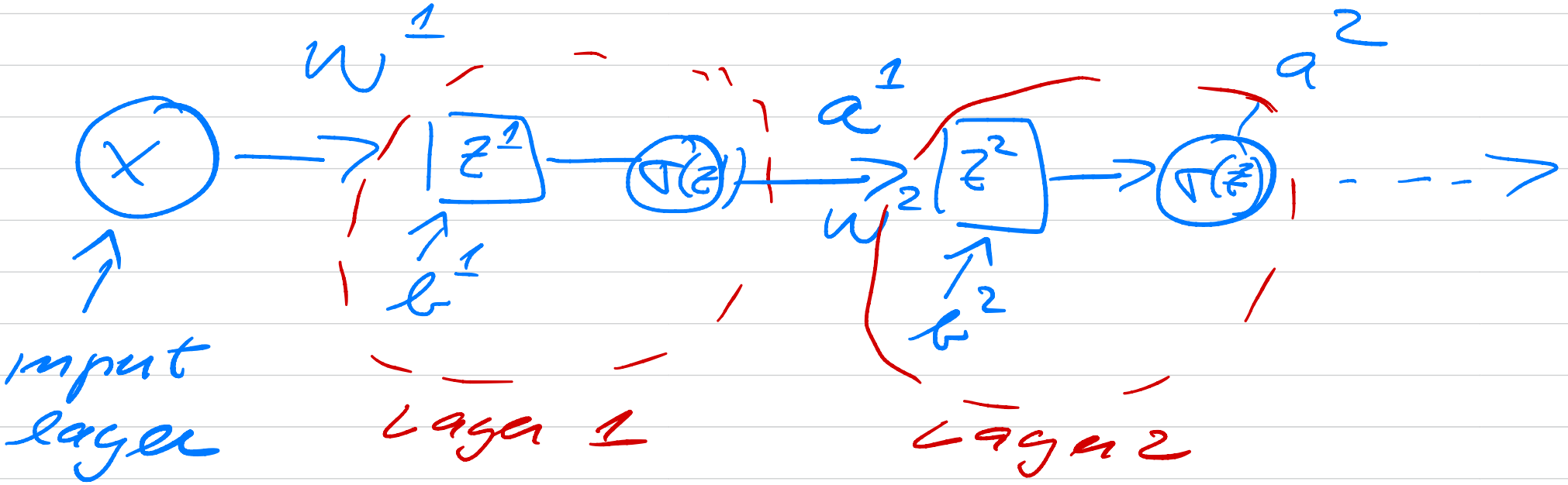
Basic elements of NNs

Define model (architecture)
- # of hidden layers and nodes
- activation functions

Define hyperparameters (shrinkage)
$\lambda \|w\|_2^2$

Gradient descent (GD) algorithms
- Stochastic GD
  - momentum
  - update learning { ADAM, RMSprop, Adagrad

Define cost function $C(\Theta)$
$\Theta = \{ \underbrace{(w^1, b^1)}_{\Theta^1}, (w^2, b^2) \ldots (w^L, b^L) \}$



output

a^{L-1} → w^L → z^L → σ^L → $a^L = y$ → $C(y, t)$ ← t

$C(y, t; \theta)$ cost/cost/error

$$a^L = y = \sigma^L(\sigma^{L-1}(\cdots \sigma^1(z^1)\cdots))$$

$$z^1 = w^1 x + b^1 \qquad \sigma^1(z^1) = a^1$$

$$a^L(x; \Theta)$$

Algorithm : Define network

Feed Forward through all

$l = 1, 2, \ldots L$ layers

$a^1, a^2, \ldots a^L$

initialize w and b

Have $C(a^L, t; \Theta)$

- Backpropagation part
update of w_{jk}^{ℓ} and b_j^{ℓ}

$\ell = L, L-1, L-2, \dots 1$

one Feed Forward pass +
one Backprop part = 1 iteration

$$\frac{\partial C}{\partial \theta^{\ell}} = ?$$

no explicit dependence on θ

$$C(\theta) = \frac{1}{2} \| t - a^{L}(x;\theta) \|_2^2 + \lambda \| w \|_2^2$$

$$\frac{\partial C}{\partial w^l} \quad , \quad \frac{\partial C}{\partial b^l}$$

$$C(a^l; t; \theta) = \frac{1}{2} \| t - a^l \|_2^2$$

$$\frac{\partial C}{\partial w^l} = \frac{\partial C}{\partial a^l} \frac{\partial a^l}{\partial z^l} \frac{\partial z^l}{\partial w^l}$$

$$z^l = w^l \cdot a^{l-1} + b^l$$

$$a^l = \sigma^l(z^l)$$

$$\delta^l = (\sigma^l(z^l))' \frac{\partial C}{\partial a^l} \quad \Rightarrow$$

$$\frac{\partial C}{\partial w^l} = \delta^l a^{l-1}$$

$$\frac{\partial C}{\partial b^L} = \delta^L$$

Backpropagation algo gives

$$\delta^\ell \Rightarrow \delta_j^\ell \qquad \ell = L, L-1, L-2, \ldots, 2, 1$$

$$\delta_j^\ell = \sum_k \delta_k^{\ell+1} w_{kj}^{\ell+1} \sigma'(z_j^\ell)$$

Basic gradient update

$$w_{jk}^\ell \leftarrow w_{jk}^\ell - \eta \, \delta_j^\ell a_k^{\ell-1}$$

$$b_j^\ell \leftarrow b_j^\ell - \eta \, \delta_j^\ell$$

Example

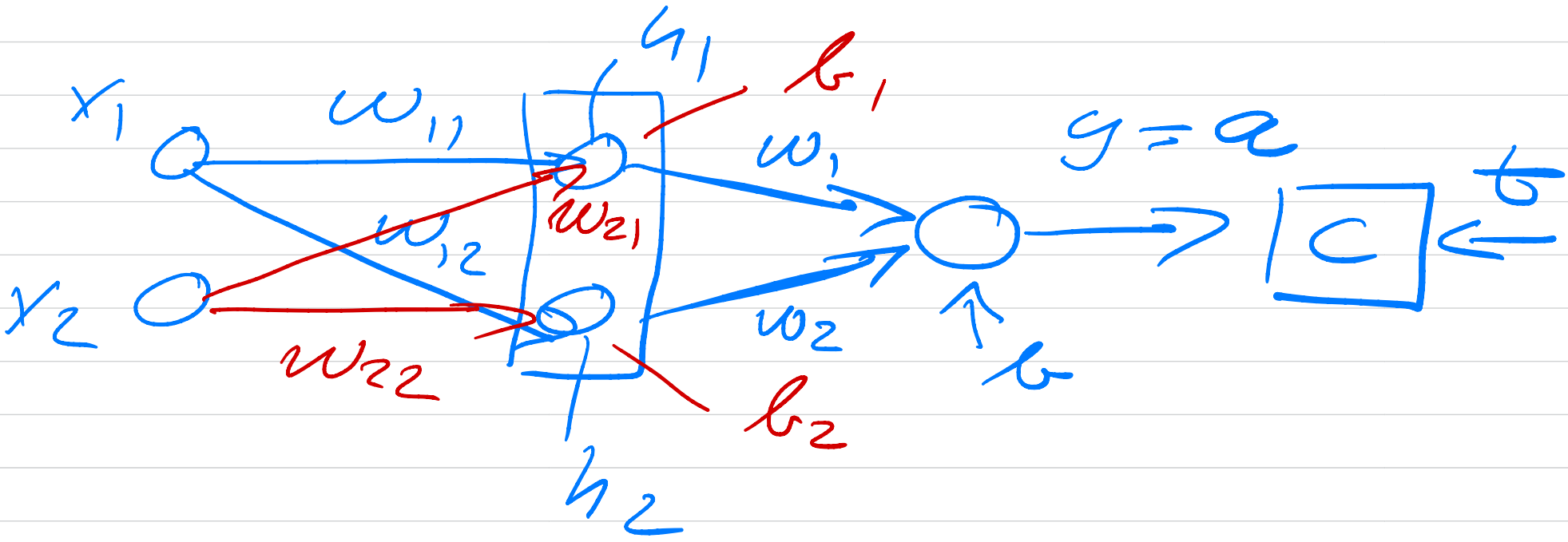
OR-Gate or XOR-gate

Two features (x_1, x_2)

OR-GATE

x_1	x_2	t
0	0	0
0	1	1
1	0	1
1	1	1

$$X = \begin{matrix} x_1 & x_2 \end{matrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$



$$z_1 = \boxed{w_{11}} x_1 + \boxed{w_{21}} x_2 + \boxed{b_1}$$

$$z_2 = \boxed{w_{12}} x_1 + \boxed{w_{22}} x_2 + \boxed{b_2}$$

$$a_1 = \sigma(z_1) \quad a_2 = \sigma(z_2)$$

$$z = \boxed{w_1} a_1 + a_2 \boxed{w_2} + \boxed{b}$$

$$a = \sigma(z)$$

$$W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

$$X = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \in \mathbb{R}^{4\times 2}$$

$$XW \in \mathbb{R}^{4\times 2} \Rightarrow$$

$$a_{1,2} \in \mathbb{R}^{4\times 2}$$

$$w^T = [w_1, w_2] \in \mathbb{R}^{1\times 2}$$

$$w \in R^{2\times 1} \Rightarrow aw \in \mathbb{R}^{4\times 1}$$

Comparison with linear regression

$$\tilde{y}_0 = \beta_0 + \beta_1 x_0 + \beta_2 x_0^2 + \cdots \beta_{p-1} x_0^{p-1}$$

$$\tilde{y}_1 = \beta_0 + \beta_1 x_1 + \cdots$$

$$\tilde{y}_2 = \beta_0 + \beta_1 x_2 - \cdots$$

$$\vdots \qquad \vdots$$

$$\tilde{y}_{n-1}$$

$$\tilde{y} = X\beta$$