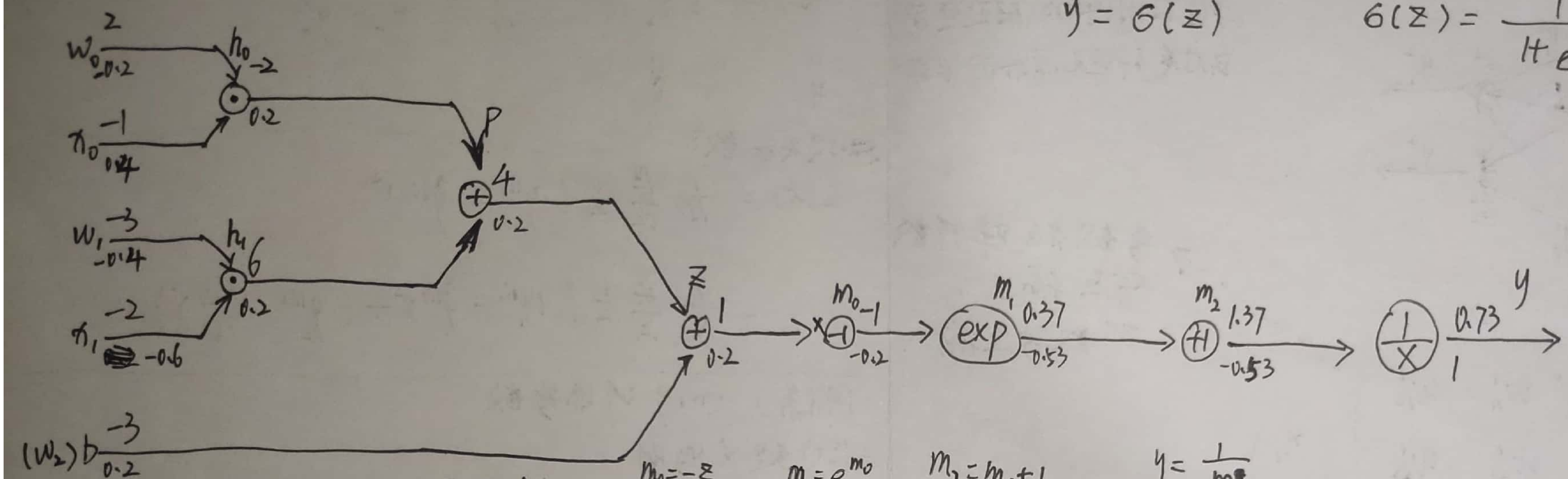
计算图：(带参数) (仅何)

$Z = W_0 x_0 + W_1 x_1 + W_2$

$y = \sigma(z) \qquad \sigma(z) = \dfrac{1}{1+e^{-z}}$



$z = p + w_2$

$\dfrac{dy}{dw_2} = \dfrac{dy}{dz}\dfrac{dz}{dw_2}$

$= 0.2 \times 1$

$= 0.2$

$z = p + w_2$

$\dfrac{dy}{dp}$

$= \dfrac{dy}{dz}\dfrac{dz}{dp}$

$= 0.2 \times 1$

$= 0.2$

$m_0 = -z$

$\dfrac{dy}{dz}$

$= \dfrac{dy}{dm_0}\dfrac{dm_0}{dz}$

$= -0.2 \times (-1)$

$= 0.2$

$m_1 = e^{m_0}$

$\dfrac{dy}{dm_0}$

$= \dfrac{dy}{dm_1}\dfrac{dm_1}{dm_0}$

$= -0.53 \times e^{m_0}$

$= -0.53 \times \dfrac{1}{e}$

$= -0.53 \times 0.37$

$= -0.2$

$m_2 = m_1 + 1$

$\dfrac{dy}{dm_1} = \dfrac{dy}{dm_2} \cdot \dfrac{dm_2}{dm_1}$

$= (-0.53) \times 1$

$= -0.53$

$y = \dfrac{1}{m_2}$

$\dfrac{dy}{dm_2} = -\dfrac{1}{m_2^2} = -\dfrac{1}{1.37^2} = -0.53$

$p = h_0 + h_1$

$\dfrac{dy}{dh_0} = \dfrac{dy}{dp} \cdot \dfrac{dp}{dh_0}$

$= 0.2 \times 1$

$= 0.2$

$p = h_0 + h_1$

$\dfrac{dy}{dh_1}$

$= \dfrac{dy}{dp}\dfrac{dp}{dh_1}$

$= 0.2 \times 1$

$= 0.2$

$h_0 = w_0 \cdot x_0$

$\dfrac{dy}{dw_0}$

$= \dfrac{dy}{dh_0} \cdot \dfrac{dh_0}{dw_0}$

$= 0.2 \times x_0$

$= -0.2$

$h_0 = W_0 \cdot x_0$

$\dfrac{dy}{dx_0}$

$= \dfrac{dy}{dh_0}\dfrac{dh_0}{dx_0}$

$= 0.2 \times W_0$

$= 0.2 \times 2 = 0.4$

$h_1 = w_1 \cdot x_1$

$\dfrac{dy}{dw_1}$

$= \dfrac{dy}{dh_1}\dfrac{dh_1}{dw_1}$

$= 0.2 \times x_1$

$= 0.2 \times (-2)$

$= -0.4$

$h_1 = w_1 \cdot x_1$

$\dfrac{dy}{dx_1}$

$= \dfrac{dy}{dh_1}\dfrac{dh_1}{dx_1}$

$= 0.2 \times W_1$

$= 0.2 \times (-3)$

$= -0.6$

简答

计算 神经元正向

推导 神经元反向

计算题：投硬币 问题1：最大似然法

正面：1

反面：0

结果：正正正反反

11100

设硬币正面的概率为 p，设 硬币为正面时，$x=1$

反面 $1-p$ 随机变量 x 反 ，$x=0$

概率密度函数：

$$f(x)=p^x(1-p)^{(1-x)}$$

$$f(x=1)=p$$

$$f(x=0)=1-p$$

最大似然 ~~函数~~ 法）→ 事件发生概率

（函数）

$$L(p)=ppp(1-p)(1-p) \rightarrow L(p)=p^{N_正}(1-p)^{N_反}$$

似然函数 $=p^3(1-p)^2$

← { 取 ln，乘变加

$$\ln L(p)=\ln(p^3(1-p)^2)$$

$$=3\ln p+2\ln(1-p)$$

5 × 20′ = 100′

求导：（找极大值点）

$$\frac{d\ln L(p)}{dp}=\frac{3}{p}+\frac{2\cdot(-1)}{1-p}=0$$

$$\frac{3}{p}=\frac{2}{1-p}$$

$$3(1-p)=2p$$

$$3-3p=2p$$

$$5p=3$$

$$p=\frac{3}{5}$$

正面概率 $=p=\frac{3}{5}$

反面 $=1-p=\frac{2}{5}$

$$\left(\begin{aligned}正&=\frac{N_正}{N_总}=\frac{3}{5}\\反&=\frac{N_反}{N_总}=\frac{2}{5}\end{aligned}\right)$$

问题2：贝叶斯方法，已知：Beta(a,b)

最大似然概率法，eg. $a=20$，$b=20$

设正面概率为p，设$x=1$，正面

反 $1-p$ $x=0$，反面

$$f(x)=p^{x}(1-p)^{(1-x)}$$

设样本集为X

贝叶斯公式 $$P(p|X)=\frac{P(X|p)\,P(p)}{P(X)}$$

$$P(X|p)=L(p)=p^{3}(1-p)^{2}$$

$$L'(p)=P(X|p)\cdot P(p)=p^{3}(1-p)^{2}\,\mathrm{Beta}(a,b)$$

比较：$\mathrm{Beta}(a,b)=\frac{1}{B(a,b)}p^{(a-1)}(1-p)^{(b-1)}$ a和b为 $\longrightarrow$ 伪计数

取对数：$$\ln L'(p)=\ln\left(p^{(3+a-1)}(1-p)^{(2+b-1)}\right)+\left(\ln\frac{1}{B(a,b)}\right)\text{常数}$$

$$=(3+a-1)\ln p+(2+b-1)\ln(1-p)+\text{常数}$$

$$\frac{d\ln L'(p)}{dp}=\frac{3+a-1}{p}+\frac{2+b-1}{1-p}(-1)=0$$

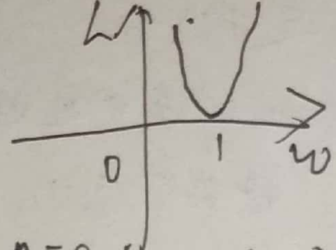
$$\frac{3+a-1}{p}=\frac{2+b-1}{1-p}$$

$$p=\frac{3+a-1}{3+a-1+2+b-1}$$

$$=\frac{a+2}{a+b+3}=\frac{22}{43}$$

$$1-p=\frac{21}{43}$$

梯度下降



$$L(w) = (y - aw - 10)^2$$

$a=1$, $y=11$, $w^0=0$, $\eta=0.1$ → 学习率

$$L(w) = (11 - w - 10)^2$$

$$= (1-w)^2$$

$$= (w-1)^2$$

1′ 化简 $L(w)$

2′ 对 $L(w)$ 求导

$$\frac{dL(w)}{dw} = 2(w-1)\cdot 1 = 2(w-1)$$

$$w^{n+1} = w^n - \eta \frac{dL(w)}{dw}\Big|_{w^n}$$

$$w^1 = w^0 - \eta \frac{dL(w)}{dw}\Big|_{w^0}$$

$$= 0 - 0.1\cdot 2(w^0-1)$$

$$= 0 - 0.1\times 2(0-1)$$

$$= 0.2$$

$$w^2 = w^1 - \eta(2(w_1-1)$$

$$= 0.2 - 0.2(0.2-1)$$

$$= 0.2 + 0.2\times 0.8$$

$$= 0.2 + 0.16 = 0.36$$

$$w^3 = w^2 - \eta(2(w_2-1))$$

$$= 0.36 - 0.1\times 2\times(0.36-1)$$

$$= 0.36 + 0.2\times 0.64$$

$$= 0.36 + 0.128$$

$$= 0.488 \approx 0.49$$

$$w^4 = w^3 - \eta \frac{dL(w)}{dw}\Big|_{w^3}$$

$$= 0.49 - 0.1\times 2(w^3-1)$$

$$= 0.592$$

$$\approx 0.6$$

$$w^5 = w^4 - \eta \frac{dL(w)}{dw}\Big|_{w^4}$$

$$= 0.6 - 0.1\times 2(w^4-1)$$

$$= 0.6 - 0.1\times 2\times(0.6-1) = 0.68$$

$$w^6 = w^5 - \eta \frac{dL(w)}{dw}\Big|_{w^5}$$

$$= 0.68 - 0.1 \times 2(w_5 - 1)$$

0.68.

$$= 0.68 + 0.064$$

过拟合

在训练的过程中，训练的误差随着迭代逐渐减小，验证(或测试)的误差逐渐增大，这种现象是过拟合现象

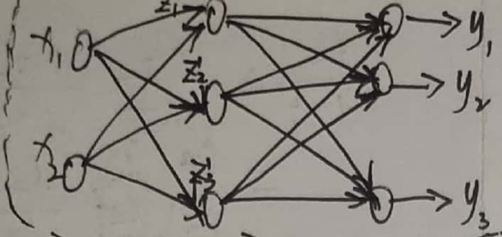
成因：模型过于复杂，不仅学到了数据的知识，还学到了没有用的噪音。

3种方法：

1. 正则化
2. 增大数据集
3. 减小噪音

均方误差：

$wx^{(i)}+b = f(x^{(i)})$



$$L(w) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(\hat{y}^{(i)} - f(x^{(i)})\right)^2 \longrightarrow \frac{dL}{dw} = \frac{1}{N}\sum_{i=1}^{N} -\left(\hat{y}^{(i)} - wx^{(i)} - b\right)$$

$$w^{(n+1)} = w^{(n)} - \eta \left.\frac{dL(w)}{dw}\right|_{w^{(n)}}$$

w是线

$$= w^{(n)} - \eta\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}^{(i)} - f(x^{(i)})\right)\left.\frac{df}{dw}\right|_{w^n}\right)$$

交叉熵：

$$f(x^n) = wx + b$$

$$w^{(n+1)} = w^{(n)} - \eta\left(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}^{(i)} - wx^{(i)} - b\right)x^{(i)}\right)$$

$$L(w) = \frac{-1}{N}\sum_{i=1}^{N}\left(\hat{y}^{(i)}\ln y^{(i)} + (1-\hat{y}^{(i)})\ln(1-y^{(i)})\right)$$

$$= \frac{-1}{N}\sum_{i=1}^{N}\left(\hat{y}^{(i)}\ln \underbrace{f(x^{(i)})}_{\sigma(wx^{(i)}+b)} + (1-\hat{y}^{(i)})\ln(1-\underbrace{f(x^{(i)})}_{\sigma(wx^{(i)}+b)})\right)$$

$$f(x^{(i)}) = \sigma(wx^{(i)}+b)$$

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

$$w^{n+1} = w^{n} - \eta\left.\frac{dL(w)}{dw}\right|_{w^n}$$

$$= w^{n} - \eta\left(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}^{(i)} - f(x^{(i)})\right)x^{(i)}\right)$$

神经元网络的损失函数 L：

交叉熵： $(x^{(i)}, \hat{y}^{(i)})$

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{M}\hat{y}_j^{(i)}\ln y_j\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{M}\hat{y}_j^{(i)}\ln f(x^{(i)})\right)$$

均方误差：

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(\hat{y}^{(i)} - y^{(i)}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\sum_{j=1}^{M}\left(\hat{y}_j^{(i)} - f_j(x^{(i)})\right)^2$$

正向传播：

$$\begin{pmatrix} z_1^1 \\ z_2^1 \\ z_3^1 \end{pmatrix} = \begin{pmatrix} w_{11}^1 & w_{12}^1 \\ w_{21}^1 & w_{22}^1 \\ w_{31}^1 & w_{32}^1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \begin{pmatrix} y_1^1 \\ y_2^1 \\ y_3^1 \end{pmatrix} = \begin{pmatrix} \sigma(z_1^1) \\ \sigma(z_2^1) \\ \sigma(z_3^1) \end{pmatrix}$$

$$\longrightarrow \frac{dL}{dw} = \frac{1}{N}\sum_{i=1}^{N} -\left(\hat{y}^{(i)} - \sigma(wx^{(i)}+b)\right)$$

$$x_1^2 = y_1^1$$

$$x_2^2 = y_2^1$$

$$x_3^2 = y_3^1$$

$$\begin{pmatrix} z_1^2 \\ z_2^2 \\ z_3^2 \end{pmatrix} = \begin{pmatrix} w_{11}^2 & w_{12}^2 & w_{13}^2 \\ w_{21}^2 & w_{22}^2 & w_{23}^2 \\ w_{31}^2 & w_{32}^2 & w_{33}^2 \end{pmatrix}\begin{pmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \end{pmatrix} \qquad \begin{pmatrix} y_1^2 \\ y_2^2 \\ y_3^2 \end{pmatrix} = \begin{pmatrix} \sigma(z_1^2) \\ \sigma(z_2^2) \\ \sigma(z_3^2) \end{pmatrix}$$

$$y_1 = y_1^2$$

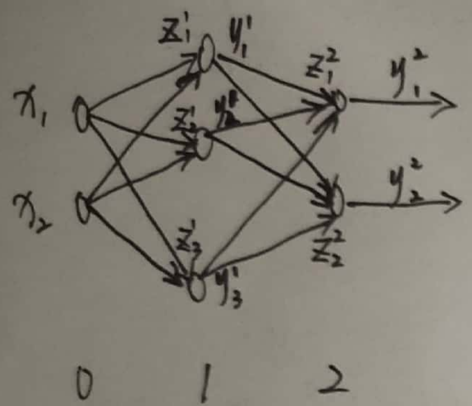
$$y_2 = y_2^2$$

$$y_3 = y_3^2$$

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

$$\sigma'(z) = \sigma(z)(1-\sigma(z))$$

神经元网络前向传播（给输入算输出）（预测）



神经元网络的假设空间
就是神经元网络的结构

参数矩阵行数
对应输出，
列数对应输入

$$\begin{pmatrix} z_1^1 \\ z_2^1 \\ z_3^1 \end{pmatrix} = \begin{pmatrix} w_{11}^1 & w_{12}^1 \\ w_{21}^1 & w_{22}^1 \\ w_{31}^1 & w_{32}^1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\begin{pmatrix} y_1^1 \\ y_2^1 \\ y_3^1 \end{pmatrix} = \begin{pmatrix} \sigma(z_1^1) \\ \sigma(z_2^1) \\ \sigma(z_3^1) \end{pmatrix}$$

$x_1^2 = y_1^1$

$x_2^2 = y_2^1$

$x_3^2 = y_3^1$

$$\begin{pmatrix} z_1^2 \\ z_2^2 \end{pmatrix} = \begin{pmatrix} w_{11}^2 & w_{12}^2 & w_{13}^2 \\ w_{21}^2 & w_{22}^2 & w_{23}^2 \end{pmatrix} \begin{pmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \end{pmatrix}$$

$$\begin{pmatrix} y_1^2 \\ y_2^2 \end{pmatrix} = \begin{pmatrix} \sigma(z_1^2) \\ \sigma(z_2^2) \end{pmatrix} \qquad \sigma(z) = \frac{1}{1+e^{-z}}$$

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_1^2 \\ y_2^2 \end{pmatrix}$$

$$\begin{matrix} \| & & \| \\ y & & \hat{y} \end{matrix}$$

$y_1 = y_1^2$

$y_2 = y_2^2$

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

损失函数：

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left((y_1^{(i)} - \hat{y}_1^{(i)})^2 + (y_2^{(i)} - \hat{y}_2^{(i)})^2\right)$$

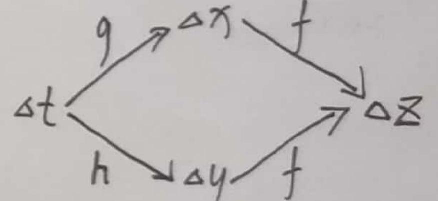
（训练）如何训练参数

方法1）链式法则

$y = f(x) \quad z = h(y) \quad \frac{dz}{dx} = h'(y) f'(x)$

$z = f(x, y) \quad x = g(t) \quad y = h(t)$

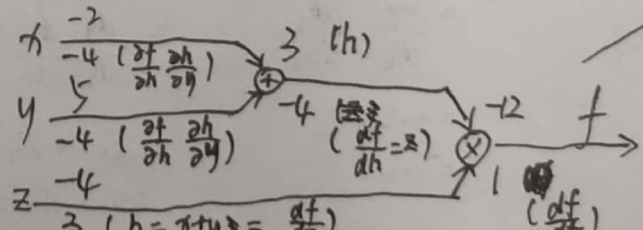
$$\frac{dz}{dt} = \frac{\partial f}{\partial x} g'(t) + \frac{\partial f}{\partial y} h'(t)$$



确定参数（梯度下降）

计算图：

$f(x, y, z) = (x+y)z = hz \qquad h = x+y$



上面是前向传播计算结果

下面是导数

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial h}\frac{\partial h}{\partial x} = -4 \cdot 1 = -4$$

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial h}\frac{\partial h}{\partial y} = -4 \cdot 1 = -4$$

反向传播：

$\boxed{\text{所有 } \Delta \text{ 都是 } \nabla}$

$$\frac{dL}{dy_i^n} = \frac{dL}{dy_i}\frac{dy_i}{dy_i^n} = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{y}_i^{(n)} - y_i\right)$$

$$\frac{dL}{dy_i^2} = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{y}_i^{(n)} - y_i\right) = \Delta L_i$$

$$\frac{dL}{dz_i^2} = \frac{dL}{dy_i^2}\frac{dy_i^2}{dz_i^2} = \Delta L_i\, \sigma'(z)\Big|_{z_i^2} = \sigma(z_i^2)(1-\sigma(z_i^2))\Delta L_i$$

$$\frac{dL}{dw_{ij}^2} = \frac{dL}{dz_i^2}\frac{dz_i^2}{dw_{ij}^2} = \sigma(z_i^2)(1-\sigma(z_i^2))\Delta L_i\, x_j^2$$

$$z_i^2 = w_{i1}^2 x_1^2 + w_{i2}^2 x_2^2 + w_{i3}^2 x_3^2$$

$$\frac{dz_i^2}{dw_{i1}^2} = x_1^2 \qquad \frac{dz_i^2}{dw_{i2}^2} = x_2^2 \qquad \frac{dz_i^2}{dw_{i3}^2} = x_3^2$$

$$\therefore \frac{dz_i^2}{dw_{ij}^2} = x_j^2$$

$$\frac{dL}{dx_j^2} = \cancel{\frac{dL}{dz_i^2}\frac{dz_i^2}{dx_j^2}}$$

$$= \sum_{i=1}^{M_n}\left(\frac{dL}{dz_i^2}\frac{dz_i^2}{dx_j^2}\right)$$

$$= \frac{dL}{dz_1^2}\frac{dz_1^2}{dx_j^2} + \frac{dL}{dz_2^2}\frac{dz_2^2}{dx_j^2} + \frac{dL}{dz_3^2}\frac{dz_3^2}{dx_j^2}$$

$$= \Delta L_1 w_{1j}^2 + \Delta L_2 w_{2j}^2 + \Delta L_3 w_{3j}^2$$

不写：

$z = f(x, y)$

$x = h(t)$

$y = g(t)$

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}$$

不写（条件速查）

$z_1^2 = w_{11}^2 x_1^2 + w_{12}^2 x_2^2 + w_{13}^2 x_3^2$

$z_2^2 = w_{21}^2 x_1^2 + w_{22}^2 x_2^2 + w_{23}^2 x_3^2$

$z_3^2 = w_{31}^2 x_1^2 + w_{32}^2 x_2^2 + w_{33}^2 x_3^2$

$$\frac{dL}{z_i^1} = \frac{dL}{dy_i^1}\frac{dy_i^1}{dz_i^1}$$

$$= \frac{dL}{dx_i^2}\frac{dx_i^2}{dz_i^1}$$

$$= (\Delta L_1 w_{1i}^2 + \Delta L_2 w_{2i}^2 + \Delta L_3 w_{3i}^2)(\sigma(z_i^2)(1-\sigma(z_i^2)))$$

$$\frac{dL}{dw_{ij}^1} = \frac{dL}{dz_i^1}\frac{dz_i^1}{dw_{ij}^1} = (\downarrow \ \downarrow \ \downarrow \ \downarrow)\, x_j^1$$