反向传播算法学习笔记

钰见 · 梵星

2025 年 2 月 8 日

目录

1 前言

本文在理解梯度下降的前提下，采用图文并茂的方式记录学习反向传播算法后的总结笔记，公式推导力求详细不省步骤。

如有谬误，请批评指正。

记号说明：

$X = (x_1, x_2, ..., x_{n_l})^T$ 表示单个样本的输入

n_l 表示第 l 层神经元的个数

$w_{ji}^{(l)}$ 表示第 l 层第 i 个神经元连接到第 $l-1$ 层第 j 个神经元的权重

$b_i^{(l)}$ 表示第 l 层到第 $l+1$ 层第 i 个神经元的偏置（图中未画出）也可用 $w_{bi}^{(l)}$ 表示

$z^{(l)} = (z_1^{(l)}, z_2^{(l)}, ..., z_{n_l}^{(L)})$ 表示第 l 层神经元的加权输入

$a^{(l)} = (a_1^{(l)}, a_2^{(l)}, ..., a_{n_l}^{(l)})$ 表示第 l 层神经元的加权输出（激活值）

E 表示单个样本的误差

下面以图1的神经网络为例说明前向传播和反向传播算法的原理

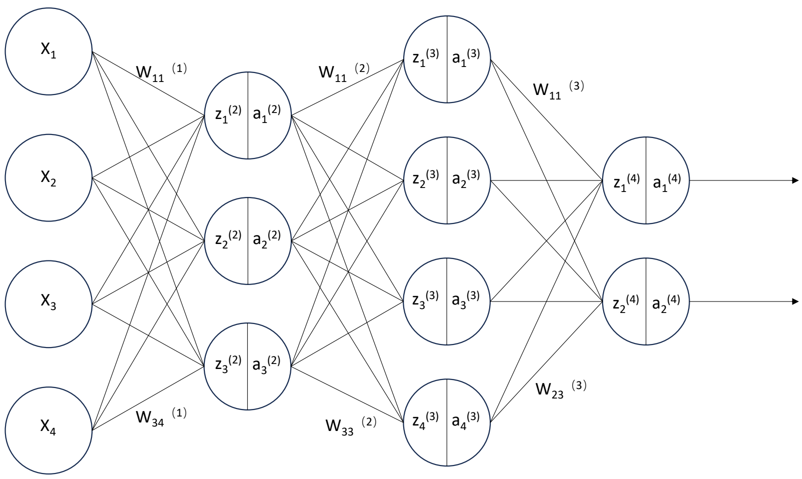

图 1: 神经网络示意图

2 前向传播

如图1，信息的前向传播过程如下：

$$X = a^{(1)} \rightarrow z^{(2)} \rightarrow a^{(2)} \rightarrow z^{(3)} \rightarrow a^{(3)} \rightarrow z^{(4)} \rightarrow a^{(4)} \rightarrow y$$

2.1 第 l 层到第 $l+1$ 层的计算

例

$$z_1^{(2)} = w_{(11)}^{(1)} X_1 + w_{(12)}^{(1)} X_2 + b_1^{(1)}$$
$$z_2^{(3)} = w_{(11)}^{(2)} a_1^{(2)} + w_{(12)}^{(2)} a_2^{(2)} + b_2^{(2)}$$

即

$$z_i^{(l)} = w_{i1}^{(l-1)} a_1^{(l-1)} + w_{i2}^{(l-1)} a_2^{(l-1)} + \ldots + w_{ij}^{(l-1)} a_j^{(l-1)} + b_i^{(l)}$$

写成矩阵形式就是：

$$\begin{bmatrix} w_{11}^{(n_l)} & w_{12}^{(n_l)} & \cdots & w_{1n_l}^{(n_l)} \\ w_{21}^{(n_l)} & w_{22}^{(n_l)} & \cdots & w_{2n_l}^{(n_l)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n_{(l+1)}1}^{(n_l)} & w_{n_{(l+1)}2}^{(n_l)} & \cdots & w_{n_{(l+1)}n_l}^{(n_l)} \end{bmatrix} \begin{bmatrix} a_1^{(n_l)} \\ a_2^{(n_l)} \\ \cdots \\ a_{n_l}^{(n_l)} \end{bmatrix} + \begin{bmatrix} b_1^{(n_l)} \\ b_2^{(n_l)} \\ \cdots \\ b_{n_{(l+1)}}^{(n_l)} \end{bmatrix} = \begin{bmatrix} z_1^{(n_l)} \\ z_2^{(n_l)} \\ \cdots \\ z_{n_{(l+1)}}^{(n_l)} \end{bmatrix}$$

2.2 第 l 层输入值到激活值的计算

$$a_i^{(l)} = f(z_i^{(l)})$$

常用 sigmoid 函数，即

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

sigmoid 函数的导数为

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

3 反向传播

3.1 误差反向传播

$$E = \frac{1}{N}\sum_{i=1}^{N} E_i$$

其中 E_i 表示第 i 个样本的误差，N 表示样本总数，例如 $E = \frac{1}{N}\sum_{i=1}^{N}(y_i - a_i^{(l)})^2$。
为了通过调整权重和偏置来减小误差，需要计算误差对权重和偏置的偏导数，即 $w_{new} = w_{old} - \eta\nabla E$

$$\begin{aligned}
w'^{(l)}_{ji} &= w^{(l)}_{ji} - \eta\nabla E \\
&= w^{(l)}_{ji} - \eta\frac{\partial E}{\partial w^{(l)}_{ji}} \\
&= w^{(l)}_{ji} - \eta\frac{\partial E}{\partial z^{(l+1)}_i}\frac{\partial z^{(l+1)}_i}{\partial w^{(l)}_{ji}} \\
b'^{(l)}_i &= b^{(l)}_i - \eta\nabla E \\
&= b^{(l)}_i - \eta\frac{\partial E}{\partial b^{(l)}_i} \\
&= b^{(l)}_i - \eta\frac{\partial E}{\partial z^{(l+1)}_i}\frac{\partial z^{(l+1)}_i}{\partial b^{(l)}_i}
\end{aligned}$$

其中 η 为学习率,∇E 表示误差对权重或偏置的偏导数 (此处 w' 表示 w_{new}，不是导数)

3.2 输出层

3.2.1 输出层误差

我们引入记号 δ 使得 $\delta_j^{(l)} = \frac{\partial E}{\partial z_j^{(l)}}$ 表示第 l 层第 j 个神经元的误差。由于 $a_i^{(l)} = f(z_i^{(l)})$，所以 $\frac{\partial a_i^{(l)}}{\partial z_i^{(l)}} = f'(z_i^{(l)})$, 则输出层的误差如图2

$$\delta_1^{(4)} = \frac{\partial E}{\partial z_1^{(4)}} = \frac{\partial E}{\partial a_1^{(4)}}\frac{\partial a_1^{(4)}}{\partial z_1^{(4)}} = \frac{\partial E}{\partial a_1^{(4)}} f'(z_1^{(4)})$$

$$\delta_2^{(4)} = \frac{\partial E}{\partial z_2^{(4)}} = \frac{\partial E}{\partial a_2^{(4)}}\frac{\partial a_2^{(4)}}{\partial z_2^{(4)}} = \frac{\partial E}{\partial a_2^{(4)}} f'(z_2^{(4)})$$

由此可得

$$\delta_i^{(l)} = \frac{\partial E}{\partial z_i^{(l)}} f'(z_i^{(l)})$$

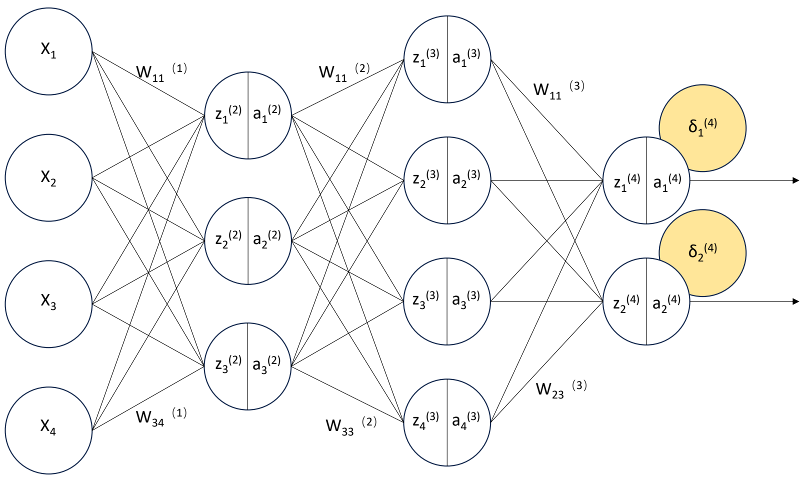

图 2: 输出层误差

以损失函数为 $E=\frac{1}{2}\sum_{i=1}^{2}(y_i-a_i^{(4)})^2$，激活函数为 sigmoid 函数为例，则

$$\frac{\partial E}{\partial a_1^{(4)}}=\frac{\partial}{\partial a_1^{(4)}}\frac{1}{2}\left((y_1-a_1^{(4)})^2+(y_2-a_2^{(4)})^2\right)=-(y_1-a_1^{(4)})$$

$$\frac{\partial a_1^{(4)}}{\partial z_1^{(4)}}=\frac{\partial}{\partial z_1^{(4)}}\sigma(z_1^{(4)})=\sigma'(z_1^{(4)})=\sigma(z_1^{(4)})(1-\sigma(z_1^{(4)}))=a_1^{(4)}(1-a_1^{(4)})$$

代入以上两式，得

$$\frac{\partial E}{\partial z_1^{(4)}}=\frac{\partial E}{\partial a_1^{(4)}}\frac{\partial a_1^{(4)}}{\partial z_1^{(4)}}=-(y_1-a_1^{(4)})a_1^{(4)}(1-a_1^{(4)})$$

同理可得

$$\frac{\partial E}{\partial z_i^{(l)}}=\frac{\partial E}{\partial a_i^{(l)}}\frac{\partial a_i^{(l)}}{\partial z_i^{(l)}}=-(y_i-a_i^{(l)})a_i^{(l)}(1-a_i^{(l)})$$

3.2.2 输出层参数更新

由 $w'^{(3)}_{11}=w^{(3)}_{11}-\eta\frac{\partial E}{\partial w^{(3)}_{11}}$, $\frac{\partial E}{\partial w^{(3)}_{11}}=\frac{\partial E}{\partial z_1^{(4)}}\frac{\partial z_1^{(4)}}{\partial w^{(3)}_{11}}=\delta_1^{(4)}\frac{\partial z_1^{(4)}}{\partial w^{(3)}_{11}}$，我们现在已经求得 $\delta_1^{(4)}$，则只需求出 $\frac{\partial z_1^{(4)}}{\partial w^{(3)}_{11}}$，即可得到更新后的 w'。（这里引入 δ 的作用还不是很明显，后续在隐藏层中将会看到引入 δ 可以大大简化表达形式并且利用计算结果）

由矩阵

$$\begin{bmatrix} w_{11}^{(3)} & w_{12}^{(3)} & w_{13}^{(3)} & w_{14}^{(3)} \\ w_{21}^{(3)} & w_{22}^{(3)} & w_{23}^{(3)} & w_{24}^{(3)} \\ w_{31}^{(3)} & w_{32}^{(3)} & w_{33}^{(3)} & w_{34}^{(3)} \end{bmatrix}\begin{bmatrix} a_1^{(3)} \\ a_2^{(3)} \\ a_3^{(3)} \\ a_4^{(3)} \end{bmatrix}+\begin{bmatrix} b_1^{(3)} \\ b_2^{(3)} \\ b_3^{(3)} \end{bmatrix}=\begin{bmatrix} z_1^{(4)} \\ z_2^{(4)} \\ z_3^{(4)} \end{bmatrix}$$

可知 $z_1^{(4)}=w_{11}^{(3)}a_1^{(3)}+w_{12}^{(3)}a_2^{(3)}+w_{13}^{(3)}a_3^{(3)}+w_{14}^{(3)}a_4^{(3)}+b_1^{(3)}$，所以 $\frac{\partial z_1^{(4)}}{\partial w_{11}^{(3)}}=a_1^{(3)}, \frac{\partial z_1^{(4)}}{\partial b_1^{(3)}}=1$

则

$$w'^{(3)}_{11}=w^{(3)}_{11}-\eta\frac{\partial z_1^{(4)}}{\partial w^{(3)}_{11}}\delta_1^{(4)}=w^{(3)}_{11}-\eta a_1^{(3)}\delta_1^{(4)}$$

$$b'^{(3)}_1=b^{(3)}_1-\eta\frac{\partial z_1^{(4)}}{\partial b^{(3)}_1}\delta_1^{(4)}=b^{(3)}_1-\eta\delta_1^{(4)}$$

一般地，对于输出层参数更新，有

$$w'^{(l)}_{ij} = w^{(l)}_{ij} - \eta a^{(l)}_j \delta^{(l+1)}_i$$

$$b'^{(l)}_i = b^{(l)}_i - \eta \delta^{(l+1)}_i$$

3.3 隐藏层

3.3.1 隐藏层误差

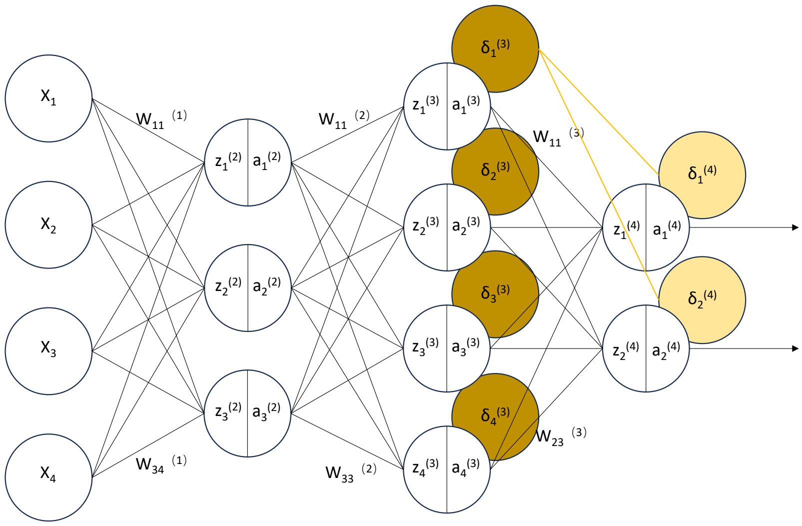

图 3: 隐藏层误差（1）

为什么要引入 δ，以及为什么要求隐藏层误差？

因为我们的目标是要根据损失函数 E 对各个参数的偏导来求出更新后的参数值，那就要先求出 E 对 w 和 b 的偏导，例如 $\frac{\partial E}{\partial w^{(2)}_{11}}$，根据链式法则可以写成 $\frac{\partial E}{\partial z^{(3)}_1}\frac{\partial z^{(3)}_1}{\partial w^{(2)}_{11}}$，其中 $\frac{\partial z^{(3)}_1}{\partial w^{(2)}_{11}}$ 容易得出，$\frac{\partial E}{\partial z^{(3)}_1}$ 就是 $\delta^{(3)}_1$，而 $\delta^{(3)}_1$ 又可根据之前算出的其他 δ 值计算，这样就可以方便且简洁地求出我们想要的偏导数，具体过程如下

如图3，分析变量，将 $z^{(3)}_1$ 看做自变量，$z^{(4)}_1$，$a^{(4)}_1$ 和 $z^{(4)}_2$，$a^{(4)}_2$ 看做中间变量，E 看做因变量，即

$$z^{(3)}_1 \to a^{(3)}_1 \to \begin{cases} z^{(4)}_1 \to a^{(4)}_1 \\ z^{(4)}_2 \to a^{(4)}_2 \end{cases} \to E$$

由多元函数微分知识，有

$$\begin{aligned}
\delta^{(3)}_1 &= \frac{\partial E}{\partial z^{(3)}_1} \\
&= \frac{\partial E}{\partial z^{(4)}_1}\frac{\partial z^{(4)}_1}{\partial z^{(3)}_1} + \frac{\partial E}{\partial z^{(4)}_2}\frac{\partial z^{(4)}_2}{\partial z^{(3)}_1} \\
&= \frac{\partial E}{\partial a^{(4)}_1}\frac{\partial a^{(4)}_1}{\partial z^{(4)}_1}\frac{\partial z^{(4)}_1}{\partial a^{(3)}_1}\frac{\partial a^{(3)}_1}{\partial z^{(3)}_1} + \frac{\partial E}{\partial a^{(4)}_2}\frac{\partial a^{(4)}_2}{\partial z^{(4)}_2}\frac{\partial z^{(4)}_2}{\partial a^{(3)}_1}\frac{\partial a^{(3)}_1}{\partial z^{(3)}_1} \\
&= \delta^{(4)}_1 w^{(3)}_{11} f'(z^{(3)}_1) + \delta^{(4)}_2 w^{(3)}_{21} f'(z^{(3)}_1) \\
&= \left(\sum_{i=1}^{2} \delta^{(4)}_i w^{(3)}_{i1}\right) f'(z^{(3)}_1)
\end{aligned}$$

同理可得

$$
\begin{aligned}
\delta_2^{(3)} &= \frac{\partial E}{\partial z_2^{(3)}} \\
&= \frac{\partial E}{\partial z_1^{(4)}} \frac{\partial z_1^{(4)}}{\partial z_2^{(3)}} + \frac{\partial E}{\partial z_2^{(4)}} \frac{\partial z_2^{(4)}}{\partial z_2^{(3)}} \\
&= \frac{\partial E}{\partial a_1^{(4)}} \frac{\partial a_1^{(4)}}{\partial z_1^{(4)}} \frac{\partial z_1^{(4)}}{\partial a_2^{(3)}} \frac{\partial a_2^{(3)}}{\partial z_2^{(3)}} + \frac{\partial E}{\partial a_2^{(4)}} \frac{\partial a_2^{(4)}}{\partial z_2^{(4)}} \frac{\partial z_2^{(4)}}{\partial a_2^{(3)}} \frac{\partial a_2^{(3)}}{\partial z_2^{(3)}} \\
&= \delta_1^{(4)} w_{12}^{(3)} f'(z_2^{(3)}) + \delta_2^{(4)} w_{22}^{(3)} f'(z_2^{(3)}) \\
&= \left(\sum_{i=1}^{2} \delta_i^{(4)} w_{i2}^{(3)} \right) f'(z_2^{(3)})
\end{aligned}
$$

这样我们就得到了第三层神经元的误差

$$
\delta_i^{(3)} = \left(\sum_{j=1}^{2} \delta_j^{(4)} w_{ji}^{(3)} \right) f'(z_i^{(3)})
$$

此时就可以看出引入 δ 确实可以大大简化表达形式并且利用之前的计算结果

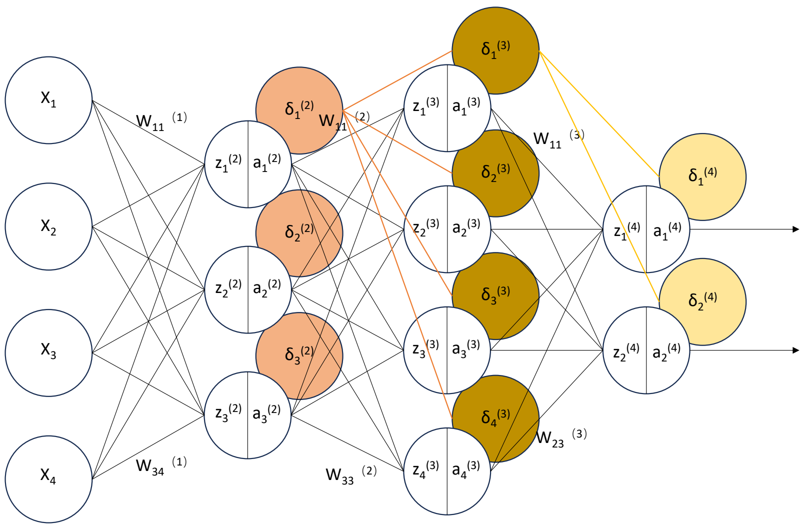

图 4: 隐藏层误差（2）

如图4，对于第二层神经网络的误差，与第三层类似，此处只写出 $\delta_1^{(2)}$ 的推导

$$
\begin{aligned}
\delta_1^{(2)} &= \frac{\partial E}{\partial z_1^{(2)}} \\
&= \frac{\partial E}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial z_1^{(2)}} + \frac{\partial E}{\partial z_2^{(3)}} \frac{\partial z_2^{(3)}}{\partial z_1^{(2)}} + \frac{\partial E}{\partial z_3^{(3)}} \frac{\partial z_3^{(3)}}{\partial z_1^{(2)}} + \frac{\partial E}{\partial z_4^{(3)}} \frac{\partial z_4^{(3)}}{\partial z_1^{(2)}} \\
&= \frac{\partial E}{\partial a_1^{(3)}} \frac{\partial a_1^{(3)}}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial z_1^{(2)}} + \frac{\partial E}{\partial a_2^{(3)}} \frac{\partial a_2^{(3)}}{\partial z_2^{(3)}} \frac{\partial z_2^{(3)}}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial z_1^{(2)}} + \frac{\partial E}{\partial a_3^{(3)}} \frac{\partial a_3^{(3)}}{\partial z_3^{(3)}} \frac{\partial z_3^{(3)}}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial z_1^{(2)}} + \frac{\partial E}{\partial a_4^{(3)}} \frac{\partial a_4^{(3)}}{\partial z_4^{(3)}} \frac{\partial z_4^{(3)}}{\partial a_4^{(2)}} \frac{\partial a_1^{(2)}}{\partial z_1^{(2)}} \\
&= \delta_1^{(3)} w_{11}^{(2)} f'(z_1^{(2)}) + \delta_2^{(3)} w_{21}^{(2)} f'(z_1^{(2)}) + \delta_3^{(3)} w_{31}^{(2)} f'(z_1^{(2)}) + \delta_4^{(3)} w_{41}^{(2)} f'(z_1^{(2)}) \\
&= \left(\sum_{i=1}^{4} \delta_i^{(3)} w_{i1}^{(2)} \right) f'(z_1^{(2)})
\end{aligned}
$$

综上所述，对于隐藏层的误差，可以总结为

$$\delta_i^{(l)} = \left(\sum_{j=1}^{n_{l+1}} \delta_j^{(l+1)} w_{ji}^{(l)}\right) f'(z_i^{(l)})$$

3.3.2 隐藏层参数更新

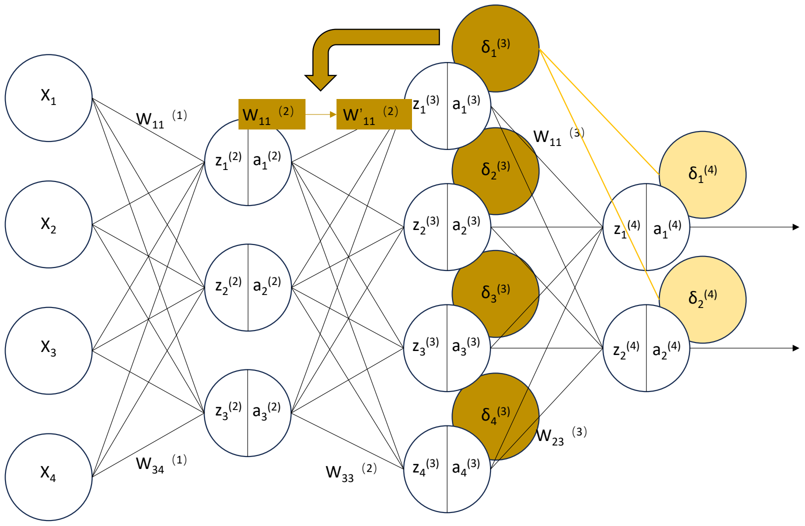

图 5: 隐藏层参数更新（1）

如图5，由 $z_1^{(3)} = w_{11}^{(2)} a_1^{(2)} + w_{12}^{(2)} a_2^{(2)} + w_{13}^{(2)} a_3^{(2)} + b_1^{(2)}$，可以得到 $\frac{\partial z_1^{(2)}}{\partial w_{11}^{(2)}} = a_1^{(2)}$ 和 $\frac{\partial z_1(2)}{\partial b_1^{(2)}} = 1$，因此

$$\begin{aligned} w'^{(2)}_{11} &= w_{11}^{(2)} - \eta \frac{\partial E}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial w_{11}^{(2)}} \\ &= w_{11}^{(2)} - \eta \delta_1^{(3)} a_1^{(2)} \\ b'^{(2)}_1 &= b_1^{(2)} - \eta \frac{\partial E}{\partial z_1^{(3)}} \frac{\partial z_1^{(3)}}{\partial b_1^{(2)}} \\ &= b_1^{(2)} - \eta \delta_1^{(3)} \end{aligned}$$

同理，如图6，可以求得该层参数的更新值为

$$\begin{aligned} w'^{(1)}_{23} &= w_{23}^{(1)} - \eta \frac{\partial E}{\partial z_2^{(2)}} \frac{\partial z_2^{(2)}}{\partial w_{23}^{(1)}} \\ &= w_{23}^{(1)} - \eta \delta_2^{(2)} X_3 \\ b'^{(1)}_3 &= b_3^{(1)} - \eta \frac{\partial E}{\partial z_2^{(2)}} \frac{\partial z_2^{(2)}}{\partial b_2^{(1)}} \\ &= b_2^{(1)} - \eta \delta_2^{(2)} \end{aligned}$$

综上所述，对于隐藏层的参数更新，可以总结为

$$\begin{cases} w'^{(l)}_{ij} = w_{ij}^{(l)} - \eta \delta_i^{(l+1)} a_j^{(l)} \\ b'^{(l)}_i = b_i^{(l)} - \eta \delta_i^{(l+1)} \end{cases}$$

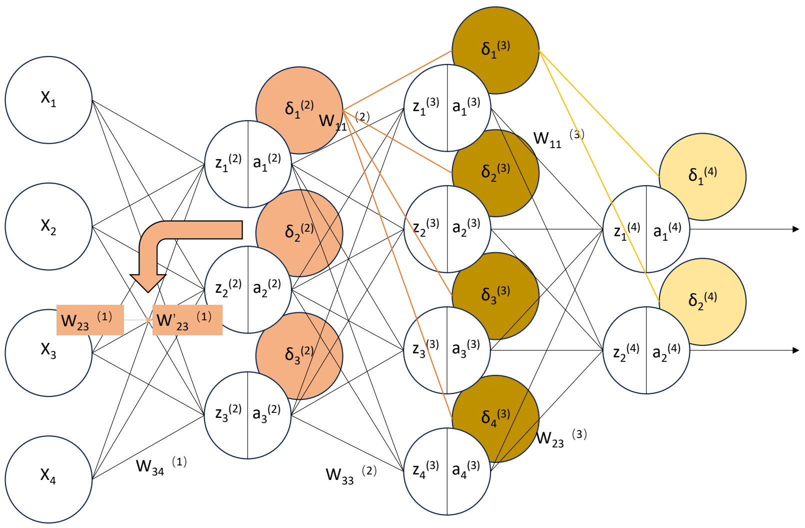

图 6: 隐藏层参数更新（2）

4 总结

反向传播误差：

$$\delta_i^{(l)} = \frac{\partial E}{\partial z_i^{(l)}}$$

输出层误差：

$$\delta_i^{(l)} = \frac{\partial E}{\partial z_i^{(l)}} f'(z_i^{(l)})$$

隐藏层误差：

$$\delta_i^{(l)} = \left(\sum_{j=1}^{n_{l+1}} \delta_j^{(l+1)} w_{ji}^{(l)} \right) f'(z_i^{(l)})$$

参数更新：

$$w'^{(l)}_{ij} = w_{ij}^{(l)} - \eta \delta_i^{(l+1)} a_j^{(l)}$$
$$b'^{(l)}_i = b_i^{(l)} - \eta \delta_i^{(l+1)}$$

5 Reference

[1]3B1B 【官方双语】深度学习之反向传播算法上/下 Part 3 ver 0.9 beta
[2]解读反向传播算法（图与公式结合）
[3]机器学习笔记丨神经网络的反向传播原理及过程（图文并茂 + 浅显易懂）

欢迎关注网易云音乐人**钰见 · 梵星！**

↑

(click here!)