反向传播

单个训练样本的损失函数:

$$\mathrm{C}(\mathbf{w},\mathbf{b})=\frac{1}{2}\|\mathbf{y}-\mathbf{o}\|^2=\frac{1}{2}\sum_{i=0}^{1}(\mathbf{y}_i-\mathbf{o}_i)^2 \tag{33}$$

损失函数 MSE:

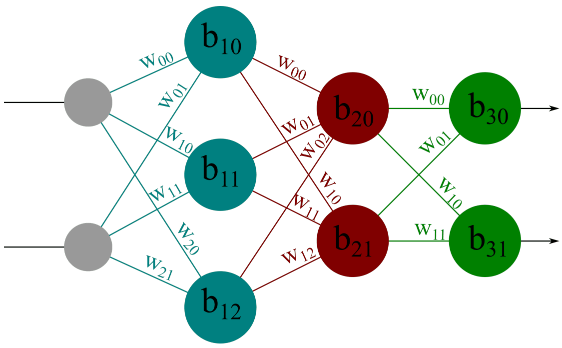
$$\mathrm{L}(\mathbf{w},\mathbf{b})=\sum_{n=1}^{N}\mathrm{C}^n(\mathbf{w},\mathbf{b}) \tag{34}$$

对每一个训练样本得到的梯度分量求和，即可得到总的梯度分量:

$$\begin{aligned}\frac{\partial \mathrm{L}(\mathbf{w},\mathbf{b})}{\partial w}&=\sum_{n=1}^{N}\frac{\partial \mathrm{C}^n(\mathbf{w},\mathbf{b})}{\partial w}\\ \frac{\partial \mathrm{L}(\mathbf{w},\mathbf{b})}{\partial b}&=\sum_{n=1}^{N}\frac{\partial \mathrm{C}^n(\mathbf{w},\mathbf{b})}{\partial b}\end{aligned} \tag{35}$$

损失函数对第 $i=1$层的其中一个权重w_{10}的偏导数为:

$$\frac{\partial C}{\partial w_{00}}=\frac{\partial C}{\partial z_{10}}\frac{\partial z_{10}}{\partial w_{00}} \tag{36}$$



前向传播计算 $\frac{\partial z}{\partial w}$

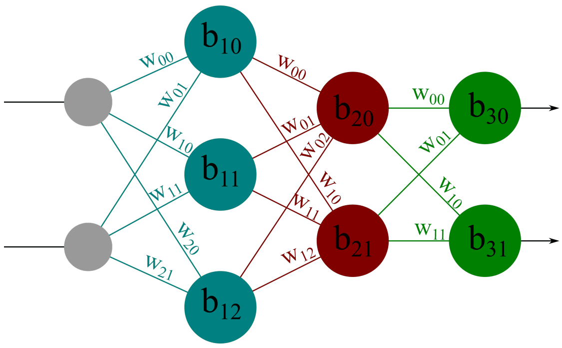
由于激活$a_{10}=\sigma(z_{10})$ 的中间变量 $z_{10}=w_{00}a_{00}+w_{01}a_{01}+b_{10}$ ， 因此有 $\frac{\partial z}{\partial w_{00}}=a_{00}$; 于是可以得到这一层中连接到该神经元的权重$w$对应的$\frac{\partial z}{\partial w}$ 分别为a_{00},a_{01} ; 那么整个第i层所有权重

$$\begin{bmatrix} w_{00} & w_{01}\\ w_{10} & w_{11}\\ w_{20} & w_{21}\end{bmatrix} \tag{37}$$

对应的 $\frac{\partial z}{\partial w}$ 为, 每一组(行)对应的值是上一层的激活，每一行都是一样的， 这些值通过前向传播直接就算出来了:

$$\begin{bmatrix} a_{00} & a_{01} \\ a_{00} & a_{01} \\ a_{00} & a_{01} \end{bmatrix} \tag{38}$$

反向传播计算 $\frac{\partial C}{\partial z}$



$$\frac{\partial C}{\partial z_{10}} = \frac{\partial C}{\partial a_{10}} \frac{\partial a_{10}}{\partial z_{10}} \tag{39}$$

因为$a_{10} = \sigma(z_{10})$, 因此

$$\frac{\partial a_{10}}{\partial z_{10}} = \sigma'(z_{10}) \tag{40}$$

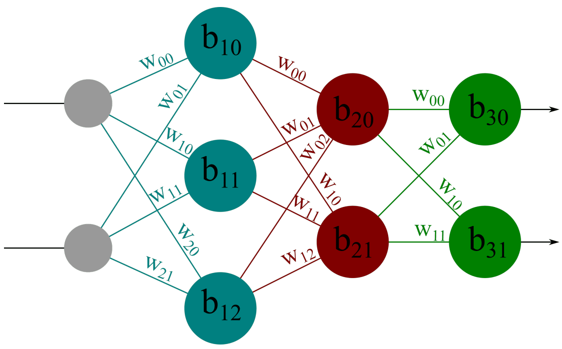
那么这一层所有节点对应的 $\frac{\partial a}{\partial z}$ 构成向量为

$$[\sigma'(z_{10}), \sigma'(z_{11}), \sigma'(z_{12})]^\intercal \tag{41}$$

z 在前向传播的时候可以记录下来，而 $\sigma'(z)$ 也是知道的。

还剩下一项:

$$\frac{\partial C}{\partial a_{10}} = \frac{\partial C}{\partial z_{20}} \frac{\partial z_{20}}{\partial a_{10}} + \frac{\partial C}{\partial z_{21}} \frac{\partial z_{21}}{\partial a_{10}} \tag{42}$$



由于 $z_{20} = w_{00}a_{10} + w_{01}a_{11} + w_{02}a_{12}$, 因此 $\frac{\partial z_{20}}{\partial a_{10}} = w_{00}$, 同理 $\frac{\partial z_{21}}{\partial a_{10}} = w_{10}$, 于是上式变成:

$$\frac{\partial C}{\partial a_{10}} = \frac{\partial C}{\partial z_{20}} w_{00} + \frac{\partial C}{\partial z_{21}} w_{10} \tag{43}$$

同理，这一层其他神经元对应有

$$
\begin{aligned}
\frac{\partial C}{\partial a_{11}} &= \frac{\partial C}{\partial z_{20}} w_{01} + \frac{\partial C}{\partial z_{21}} w_{11} \\
\frac{\partial C}{\partial a_{12}} &= \frac{\partial C}{\partial z_{20}} w_{02} + \frac{\partial C}{\partial z_{21}} w_{12}
\end{aligned} \tag{44}
$$

写成矩阵的形式是这样的

$$
\begin{bmatrix} \frac{\partial C}{\partial a_{10}} \\ \frac{\partial C}{\partial a_{11}} \\ \frac{\partial C}{\partial a_{12}} \end{bmatrix} = \begin{bmatrix} w_{00} & w_{10} \\ w_{01} & w_{11} \\ w_{02} & w_{12} \end{bmatrix} \begin{bmatrix} \frac{\partial C}{\partial z_{20}} \\ \frac{\partial c}{\partial z_{21}} \end{bmatrix} \tag{45}
$$

注意到

$$
\begin{bmatrix} w_{00} & w_{10} \\ w_{01} & w_{11} \\ w_{02} & w_{12} \end{bmatrix} = \begin{bmatrix} w_{00} & w_{01} & w_{02} \\ w_{10} & w_{11} & w_{12} \end{bmatrix}^{\mathsf{T}} \tag{46}
$$

这个关系会给我们带来很大的方便。

回到

$$
\begin{aligned}
\frac{\partial C}{\partial z_{10}} &= \frac{\partial C}{\partial a_{10}} \frac{\partial a_{10}}{\partial z_{10}} \\
&= \sigma'(z_{10}) \frac{\partial C}{\partial a_{10}} \\
&= \sigma'(z_{10}) \left(w_{00} \frac{\partial C}{\partial z_{20}} + w_{10} \frac{\partial C}{\partial z_{21}} \right)
\end{aligned} \tag{47}
$$

再回到

$$
\begin{aligned}
\frac{\partial C}{\partial w_{00}} &= \frac{\partial C}{\partial z_{10}} \frac{\partial z_{10}}{\partial w_{00}} \\
&= a_{00} \frac{\partial C}{\partial z_{10}} \\
&= a_{00} \sigma'(z_{10}) \left(w_{00} \frac{\partial C}{\partial z_{20}} + w_{10} \frac{\partial C}{\partial z_{21}} \right)
\end{aligned} \tag{48}
$$

$$
\frac{\partial C}{\partial w_{00}^{\mathrm{i}}} = a_{00} \sigma'(z_{10}) \left(w_{00}^{\mathrm{j}} \frac{\partial C}{\partial z_{20}} + w_{10}^{\mathrm{j}} \frac{\partial C}{\partial z_{21}} \right) \tag{49}
$$

汇总一下:

$$
\nabla \mathbf{w}^{i} = \begin{bmatrix} a_{00} & a_{01} \\ a_{00} & a_{01} \\ a_{00} & a_{01} \end{bmatrix} \odot \begin{bmatrix} \frac{\partial C}{\partial z_{10}} \\ \frac{\partial C}{\partial z_{11}} \\ \frac{\partial C}{\partial z_{12}} \end{bmatrix} \tag{50}
$$

其中,

$$
\begin{aligned}
\begin{bmatrix} \frac{\partial C}{\partial z_{10}} \\ \frac{\partial C}{\partial z_{11}} \\ \frac{\partial C}{\partial z_{12}} \end{bmatrix} &= \begin{bmatrix} \sigma'(z_{10}) \\ \sigma'(z_{11}) \\ \sigma'(z_{12}) \end{bmatrix} \odot \begin{bmatrix} \frac{\partial C}{\partial a_{10}} \\ \frac{\partial C}{\partial a_{11}} \\ \frac{\partial C}{\partial a_{12}} \end{bmatrix} \\
&= \begin{bmatrix} \sigma'(z_{10}) \\ \sigma'(z_{11}) \\ \sigma'(z_{12}) \end{bmatrix} \odot \left(\begin{bmatrix} w_{00} & w_{10} \\ w_{01} & w_{11} \\ w_{02} & w_{12} \end{bmatrix} \begin{bmatrix} \frac{\partial C}{\partial z_{20}} \\ \frac{\partial c}{\partial z_{21}} \end{bmatrix} \right) \\
&= \begin{bmatrix} \sigma'(z_{10}) \\ \sigma'(z_{11}) \\ \sigma'(z_{12}) \end{bmatrix} \odot \left(\begin{bmatrix} w_{00} & w_{01} & w_{02} \\ w_{10} & w_{11} & w_{12} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \frac{\partial C}{\partial z_{20}} \\ \frac{\partial c}{\partial z_{21}} \end{bmatrix} \right) \\
&= \begin{bmatrix} \sigma'(z_{10}) \\ \sigma'(z_{11}) \\ \sigma'(z_{12}) \end{bmatrix} \odot \left(\mathbf{w}^{j\,\mathsf{T}} \begin{bmatrix} \frac{\partial C}{\partial z_{20}} \\ \frac{\partial c}{\partial z_{21}} \end{bmatrix} \right)
\end{aligned} \tag{51}
$$

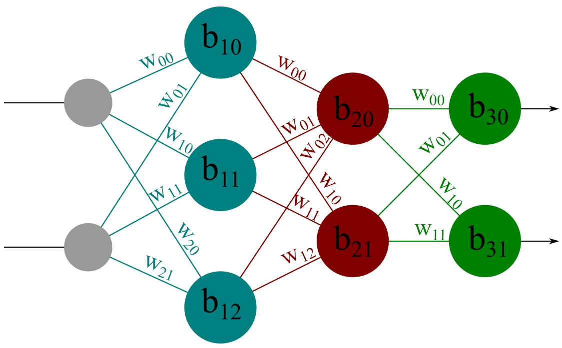
即

$$
\frac{\partial C}{\partial \mathbf{z_1}} = \sigma'(\mathbf{z_1}) \odot \left(\mathbf{w}^{2\,\mathsf{T}} \frac{\partial C}{\partial \mathbf{z_2}} \right) \tag{52}
$$

于是:

$$
\nabla \mathbf{w}^1 = \begin{bmatrix} a_{00} & a_{01} \\ a_{00} & a_{01} \\ a_{00} & a_{01} \end{bmatrix} \odot \begin{bmatrix} \frac{\partial C}{\partial z_{10}} \\ \frac{\partial C}{\partial z_{11}} \\ \frac{\partial C}{\partial z_{12}} \end{bmatrix} = \begin{bmatrix} a_{00} & a_{01} \\ a_{00} & a_{01} \\ a_{00} & a_{01} \end{bmatrix} \odot \left\{ \begin{bmatrix} \sigma'(z_{10}) \\ \sigma'(z_{11}) \\ \sigma'(z_{12}) \end{bmatrix} \odot \left(\mathbf{w}^{2\,\mathsf{T}} \begin{bmatrix} \frac{\partial C}{\partial z_{20}} \\ \frac{\partial C}{\partial z_{21}} \end{bmatrix} \right) \right\} \tag{53}
$$

即只要算出每个神经元的 $\frac{\partial c}{\partial z}$ 即可算出梯度的改变量， 先把最后一层的 $\frac{\partial c}{\partial z}$ 算出，一层层就可往前算出了。



输出层，

$$
\mathrm{C}(\mathbf{w}, \mathbf{b}) = \frac{1}{2} \|\mathbf{y} - \mathbf{o}\|^2 = \frac{1}{2} \sum_{i=0}^{1} (\mathbf{y}_i - \mathbf{o}_i)^2 = \frac{1}{2} \left(y_0^2 - 2y_0 o_0 + o_0^2 + y_1^2 - 2y_1 o_1 + o_1^2 \right) \tag{54}
$$

$$
o_0 = \sigma(z_{30}) \tag{55}
$$

$$
\frac{\partial C}{\partial z_{30}} = \frac{\partial C}{\partial o_0} \frac{\partial o_0}{\partial z_{30}} = (o_0 - y_0)\sigma'(z_{30}) \tag{56}
$$

同理

$$
\frac{\partial C}{\partial z_{31}} = \frac{\partial C}{\partial o_1} \frac{\partial o_1}{\partial z_{31}} = (o_1 - y_1)\sigma'(z_{31}) \tag{57}
$$

$$\frac{\partial C}{\partial \mathbf{z_3}} = (\mathbf{o} - \mathbf{y})\sigma'(\mathbf{z_3}) \tag{58}$$

总结

相邻两层i, j关键量的计算关系:

$$\frac{\partial C}{\partial \mathbf{z_i}} = \sigma'(\mathbf{z_i}) \odot \left(\mathbf{w^j}^\intercal \frac{\partial C}{\partial \mathbf{z_j}} \right) \tag{59}$$

最后一层若j=o, 即第j层是输出层:

$$\frac{\partial C}{\partial \mathbf{z_o}} = (\mathbf{o} - \mathbf{y})\sigma'(\mathbf{z_o}) \tag{60}$$

第i层w的改变量:

$$\nabla \mathbf{w}^i = [\mathbf{a}_h, \mathbf{a}_h, \mathbf{a}_h]^\intercal \odot \frac{\partial C}{\partial \mathbf{z}_i} \tag{61}$$

对于偏置

仿照类似的方法可以得到:

$$\frac{\partial C}{\partial b_{10}} = \frac{\partial C}{\partial z_{10}} \frac{\partial z_{10}}{\partial b_{10}} \tag{62}$$

前向传播计算 $\frac{\partial z}{\partial b}$

$a_{10} = \sigma(z_{10})$; $z_{10} = w_{00}a_{00} + w_{01}a_{01} + b_{10}$; 因此, $\frac{\partial z_{10}}{\partial b_{10}} = 1$

那么,

$$\frac{\partial \mathbf{z}_1}{\partial \mathbf{b}_1} = [1, 1, 1]^\intercal \tag{63}$$

反向传播计算 $\frac{\partial C}{\partial z}$

$$\nabla \mathbf{b}^i = \frac{\partial C}{\partial \mathbf{z}_i} \tag{64}$$