Back Propagation

- L 表示神经网络的总层数
- S_l表示第l层神经网络unit个数，不包括偏差单元`bias unit`
- k表示第几个输出单元
- $\Theta_{i,j}^{(l)}$ 第l层到第$l+1$层的权值矩阵的第i行第j列的分量
- $Z_i^{(j)}$ 第j层第i个神经元的输入值
- $a_i^{(j)}$ 第j层第i个神经元的输出值
- $a^{(j)} = g(Z^{(j)})$

1. **chain Rule**

$y = g(x) \quad z = h(y)$

$\Delta x \rightarrow \Delta y \rightarrow \Delta z \quad \frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$

$x = g(s) \quad y = h(s) \quad z = k(x, y)$

$\Delta s \rightarrow \Delta x \rightarrow \Delta z \quad \Delta s \rightarrow \Delta y \rightarrow \Delta z$

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

2. **loss function**

$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(h_\Theta(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - (h_\Theta(x^{(i)}))_k)\right]$$

我们知道代价函数loss function后，下一步就是按照梯度下降法来计算θ求解loss function的最优 解。使用梯度下降法首先要求出梯度，即偏导项$\frac{\partial}{\partial \Theta_{ij}^{(l)}} J(\Theta)$，计算偏导项的过程我们称为back propagation。

根据上面的feed forward computation 我们已经计算得到了 $a^{(1)}$ ，$a^{(2)}$， $a^{(3)}$ ，$Z^{(2)}$，$Z^{(3)}$。

3. **output layer to hidden layer**

$$\frac{\partial}{\partial \Theta_{i,j}^{(L-1)}} J(\Theta) = \frac{\partial J(\Theta)}{\partial h_\theta(x)_i}\frac{\partial h_\theta(x)_i}{\partial z_i^{(L)}}\frac{\partial z_i^{(L)}}{\partial \Theta_{i,j}^{(L-1)}} = \frac{\partial J(\Theta)}{\partial a_i^{(L)}}\frac{\partial a_i^{(L)}}{\partial z_i^{(L)}}\frac{\partial z_i^{(L)}}{\partial \Theta_{i,j}^{(L-1)}}$$

$$h_\Theta(x) = a^{(L)} = g(z^{(L)})$$

$$z^{(l)} = \Theta^{(l-1)} a^{(l-1)}$$

$$loss(\Theta) = -y^{(i)} \log(h_\Theta(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_\Theta(x^{(i)}))$$

$$\frac{\partial J(\Theta)}{\partial a_i^{(L)}} = \frac{a_i^{(L)} - y_i}{(1 - a_i^{(L)})a_i^{(L)}}$$

仅仅从激活函数的求导就可以推出

$$\begin{aligned}
\frac{\partial g(z)}{\partial z} &= -\left(\frac{1}{1+e^{-z}}\right)^2 \frac{\partial}{\partial z}(1+e^{-z}) \\
&= -\left(\frac{1}{1+e^{-z}}\right)^2 e^{-z}\,(-1) \\
&= \left(\frac{1}{1+e^{-z}}\right)\left(\frac{1}{1+e^{-z}}\right)(e^{-z}) \\
&= \left(\frac{1}{1+e^{-z}}\right)\left(\frac{e^{-z}}{1+e^{-z}}\right) \\
&= \left(\frac{1}{1+e^{-z}}\right)\left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}}\right) \\
&= g(z)\,(1-g(z))
\end{aligned}$$

$$\frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} = \frac{\partial g(z_i^{(L)})}{\partial z_i^{(L)}} = g(z_i^{(L)})(1 - g(z_i^{(L)})) = a_i^{(L)}(1 - a_i^{(L)})$$

$$\frac{\partial z_i^{(L)}}{\partial \Theta_{i,j}^{(L-1)}} = a_{i,j}^{(L-1)}$$

综上

$$\begin{aligned}
\frac{\partial}{\partial \Theta_{i,j}^{(L-1)}} J(\Theta) &= \frac{\partial J(\Theta)}{\partial a_i^{(L)}} \frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} \frac{\partial z_i^{(L)}}{\partial \Theta_{i,j}^{(L-1)}} \\
&= \frac{a_i^{(L)} - y_i}{(1 - a_i^{(L)})a_i^{(L)}} a_i^{(L)}(1 - a_i^{(L)}) a_j^{(L-1)} \\
&= (a_i^{(L)} - y_i) a_j^{(L-1)}
\end{aligned}$$

4. **hidden layer to hidden layer**

$$\frac{\partial}{\partial \Theta_{i,j}^{(l-1)}} J(\Theta) = \frac{\partial J(\Theta)}{\partial a_i^{(l)}} \frac{\partial a_i^{(l)}}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial \Theta_{i,j}^{(l-1)}} (l = 2, 3, 4, \ldots, L-1)$$

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

$$\frac{\partial z_i^{(l)}}{\partial \Theta_{i,j}^{(l-1)}} = a_j^{(l-1)}$$

第一项的求解就比较难了，需要用上上面的链式法则

$$\frac{\partial J(\Theta)}{\partial a_i^{(l)}} = \sum_{k=1}^{S_{l-1}} \left[\frac{\partial J(\Theta)}{\partial a_k^{(l+1)}} \frac{\partial a_k^{(l+1)}}{\partial z_k^{(l+1)}} \frac{\partial z_k^{(l+1)}}{\partial a_i^{(l)}} \right]$$

$$\frac{\partial a_k^{(l+1)}}{\partial z_k^{(l+1)}} = a_k^{(l+1)}(1 - a_k^{(l+1)})$$

$$\frac{\partial z_k^{(l+1)}}{\partial a_i^{(l)}} = \Theta_{k,i}^{(l)}$$

$$\begin{aligned} \frac{\partial}{\partial \Theta_{i,j}^{(L-1)}} J(\Theta) &= \sum_{k=1}^{S_{l-1}} \left[\frac{\partial J(\Theta)}{\partial a_k^{(l+1)}} \frac{\partial a_k^{(l+1)}}{\partial z_k^{(l+1)}} \frac{\partial z_k^{(l+1)}}{\partial a_i^{(l)}} \right] \\ &= \sum_{k=1}^{S_{l-1}} \left[\frac{\partial J(\Theta)}{\partial a_k^{(l+1)}} \frac{\partial a_k^{(l+1)}}{\partial z_k^{(l+1)}} \Theta_i^{(l)} \right] \\ &= \sum_{k=1}^{S_{l-1}} \left[\frac{\partial J(\Theta)}{\partial a_k^{(l+1)}} a_k^{(l+1)}(1 - a_k^{(l+1)}) \Theta_{k,i}^{(l)} \right] \end{aligned}$$

定义第l层第i个节点的的误差为

$$\begin{aligned} \delta_i^{(l)} &= \frac{\partial}{\partial z_i^{(l)}} J(\Theta) \\ &= \frac{\partial J(\Theta)}{\partial a_i^{(l)}} \frac{\partial a_i^{(l)}}{\partial z_i^{(l)}} \\ &= \frac{\partial J(\Theta)}{\partial a_i^{(l)}} a_i^{(l)}(1 - a_i^{(l)}) \\ &= \sum_{k=1}^{S_{l-1}} \left[\frac{\partial J(\Theta)}{\partial a_k^{(l+1)}} \frac{\partial a_k^{(l+1)}}{\partial z_k^{(l+1)}} \Theta_i^{(l)} \right] a_i^{(l)}(1 - a_i^{(l)}) \\ &= \sum_{k=1}^{S_{l-1}} \left[\delta_k^{(l+1)} \Theta_{k,i}^{(l)} \right] a_i^{(l)}(1 - a_i^{(l)}) \end{aligned}$$

输出层的误差为

$$
\begin{aligned}
\delta_i^{(L)} &= \frac{\partial J(\Theta)}{\partial z_i^{(L)}} \\
&= \frac{\partial J(\Theta)}{\partial a_i^{(L)}} \frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} \\
&= \frac{a_i^{(L)} - y_i}{(1 - a_i^{(L)}) a_i^{(L)}} a_i^{(L)} (1 - a_i^{(L)}) a_j^{(L-1)} \\
&= a_i^{(L)} - y_i
\end{aligned}
$$

final 代价函数的偏导数为

$$
\begin{aligned}
\frac{\partial}{\partial z_i^{(l-1)}} J(\Theta) &= \frac{\partial J(\Theta)}{\partial a_i^{(l)}} \frac{\partial a_i^{(l)}}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial \Theta_{i,j}^{(l-1)}} \\
&= \frac{\partial J(\Theta)}{\partial z_i^{(l)}} \\
&= \delta_i^{(l)} \frac{\partial z_i^{(l)}}{\partial \Theta_{i,j}^{(l-1)}} \\
&= \delta_i^{(l)} a_i^{(l-1)}
\end{aligned}
$$

总结

> 输 出 层 误 差 δ_i^L
>
> $$\delta_i^L = a_i^{(L)} - y_i$$

> 隐 层 误 差 δ_i^{l-1}
>
> $$\delta_i^{(l)} = \sum_{k=1}^{S_{l-1}} \left[\delta_k^{(l+1)} \Theta_{k,i}^{(l)} \right] a_i^{(l)} (1 - a_i^{(l)})$$

> 代 价 函 数 的 偏 导 项 $\frac{\partial}{\partial z_i^{(l)}} J(\Theta)$
>
> $$\frac{\partial}{\partial z_i^{(l-1)}} J(\Theta) = \delta_i^{(l)} a_i^{(l-1)}$$
>
> 即

$$\frac{\partial}{\partial z_i^{(l)}} J(\Theta) = \delta_i^{(l+1)} a_i^{(l)}$$