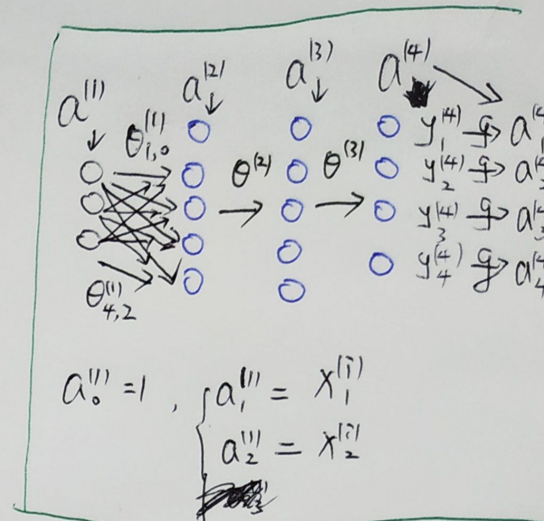
1. $$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log h_\theta(x^{(i)})_k + (1-y_k^{(i)})\log(1-h_\theta(x^{(i)})_k)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(\theta_{j,i}^{(l)}\right)^2$$



(1). 取 $m=1$，求 $\frac{\partial J(\theta)}{\partial y_j^{(4)}}$ （神经网络结构：

↳计算时使用 k（不是 j）.

化简 $J(\theta)$

$$= -\sum_{k=1}^{K}\left(y_k \log \cancel{h_\theta(x_k)}^{\,g(y_k^{(4)})} + (1-y_k)\log(1-\cancel{h_\theta(x_k)}^{\,g(y_k^{(4)})})\right)$$

$$+ \cancel{\text{正则项（与 } y_j^{(4)} \text{ 无关）}}^{\,0}$$

$$= -\sum_{k=1}^{K}\left[y_k \log\left(\frac{1}{1+e^{-y_k^{(4)}}}\right) + (1-y_k)\log\left(1-\frac{1}{1+e^{-y_k^{(4)}}}\right)\right]$$

$$= \log\left(\frac{e^{-y_k^{(4)}}}{1+e^{-y_k^{(4)}}}\right) = \log\left(\frac{1}{1+e^{y_k^{(4)}}}\right)$$

$$= -\sum_{k=1}^{K}\left[-y_k \log(1+e^{-y_k^{(4)}}) + (1-y_k)(-)\log(1+e^{y_k^{(4)}})\right]$$

$$\Rightarrow \frac{\partial J(\theta)}{y_k^{(4)}} = \frac{\partial}{\partial y_k^{(4)}}\left[y_k \cdot \log(1+e^{-y_k^{(4)}}) + (1-y_k)\log(1+e^{y_k^{(4)}})\right]$$

$$= y_k \cdot \frac{1}{1+e^{-y_k^{(4)}}} \cdot (-1)e^{-y_k^{(4)}} + (1-y_k)\cdot\frac{1}{1+e^{y_k^{(4)}}}\cdot e^{y_k^{(4)}}$$

$$= -y_k \cdot \frac{1}{1+e^{y_k^{(4)}}} + \frac{(1-y_k)\cdot e^{y_k^{(4)}}}{1+e^{y_k^{(4)}}}$$

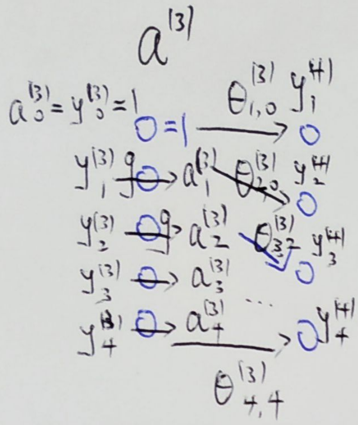
$$= \frac{-y_k(1+e^{y_k^{(4)}}) + e^{y_k^{(4)}}}{1+e^{y_k^{(4)}}} = -y_k + \frac{e^{y_k^{(4)}}}{1+e^{y_k^{(4)}}} = -y_k + g(y_k^{(4)})$$

$$= a_k^{(4)} - y_k$$

因此，我们得到了 $\frac{\partial J(\theta)}{\partial y_j^{(4)}} = a_j^{(4)} - y_j$ 或者

$(L=4)$：$\frac{\partial J(\theta)}{\partial y^{(L)}} = a^{(L)} - y$，（一个样本，最后一层为 sigmoid）.

(2). $\theta^{(3)}$ 如下图所示



$$\begin{pmatrix} y^{(4)}_1 \\ y^{(4)}_2 \\ y^{(4)}_3 \\ y^{(4)}_4 \end{pmatrix}_{4\times 1} = \begin{pmatrix} \theta^{(3)}_{1,0} & \theta^{(3)}_{1,1} & \theta^{(3)}_{1,2} & \theta^{(3)}_{1,3} & \theta^{(3)}_{1,4} \\ \theta^{(3)}_{2,0} & \theta^{(3)}_{2,1} & \theta^{(3)}_{2,2} & \theta^{(3)}_{2,3} & \theta^{(3)}_{2,4} \\ \theta^{(3)}_{3,0} & \theta^{(3)}_{3,1} & \theta^{(3)}_{3,2} & \theta^{(3)}_{3,3} & \theta^{(3)}_{3,4} \\ \theta^{(3)}_{4,0} & \theta^{(3)}_{4,1} & \theta^{(3)}_{4,2} & \theta^{(3)}_{4,3} & \theta^{(3)}_{4,4} \end{pmatrix}_{4\times 5} \begin{pmatrix} a^{(3)}_0 = 1 \\ a^{(3)}_1 \\ a^{(3)}_2 \\ a^{(3)}_3 \\ a^{(3)}_4 \end{pmatrix}_{5\times 1} \cdots (*)$$

$\theta^{(3)}_b$ $\qquad$ $\theta^{(3)}_w$

b: bias
w: weights

或者 $\cancel{y^{(4)} = \theta^{(3)}_w y^{3}}$

$$= \theta^{(3)} \begin{pmatrix} a^{(3)}_0 = y^{(3)}_0 = 1 \\ g(y^{(3)}_1) \\ g(y^{(3)}_2) \\ g(y^{(3)}_3) \\ g(y^{(3)}_3) \end{pmatrix}$$

我们希望求解 $\delta^{(3)} = \dfrac{\partial J(\theta)}{\partial y^{(3)}} \overset{y^{(3)}_0=1,\text{我}}{\underset{\text{们不计算偏导.}}{=\!=\!=\!=}} \begin{pmatrix} \dfrac{\partial J}{\partial y^{(3)}_1} \\ \partial J/\partial y^{(3)}_2 \\ \partial J/\partial y^{(3)}_3 \\ \partial J/\partial y^{(3)}_4 \end{pmatrix}$

由(*)知 每个 $y^{(3)}_i$, $i=1,2,3,4$ 都对 每个 $y^{(4)}_j$, $j=1,2,3,4$ 有贡献，根据偏导数法则，有

$$\frac{\partial J}{\partial y^{(3)}_i} = \sum_{j=1}^{4} \frac{\partial J}{\partial y^{(4)}_j} \cdot \frac{\partial y^{(4)}_j}{\partial y^{(3)}_i},$$

如 $\dfrac{\partial J}{\partial y^{(3)}_1} = \dfrac{\partial J}{\partial y^{(4)}_1} \cdot \dfrac{\partial y^{(4)}_1}{\partial y^{(3)}_1} + \dfrac{\partial J}{\partial y^{(4)}_2} \cdot \dfrac{\partial y^{(4)}_2}{\partial y^{(3)}_1} + \dfrac{\partial J}{\partial y^{(4)}_3} \cdot \dfrac{\partial y^{(4)}_3}{\partial y^{(3)}_1} + \dfrac{\partial J}{\partial y^{(4)}_4} \cdot \dfrac{\partial y^{(4)}_4}{\partial y^{(3)}_1}$

$$= \sum_{j=1}^{4} \delta^{(4)}_j \cdot \frac{\partial y^{(4)}_j}{\partial y^{(3)}_i}$$

现在计算 $\dfrac{\partial y^{(4)}_j}{\partial y^{(3)}_i} \overset{(*)}{=} \dfrac{\partial}{\partial y^{(3)}_i} \left[\theta^{(3)}_{j,0} \cdot 1 + \theta^{(3)}_{j,1} \cdot g(y^{(3)}_1) + \theta^{(3)}_{j,2} \cdot g(y^{(3)}_2) + \theta^{(3)}_{j,3} \cdot g(y^{(3)}_3) + \theta^{(3)}_{j,4} \cdot g(y^{(3)}_4) \right]$

$\underset{=}{\underline{\text{why?}}}$ $\theta^{(3)}_{j,i} \cdot g'(y^{(3)}_i) \Leftarrow \theta^{(3)}_{j,i} \cdot (y^{(3)}_i) \cdot (1 - y^{(3)}_i)$, 如果 g 是 sigmoid.

所以得到 $\delta^{(3)}_i = \dfrac{\partial J}{\partial y^{(3)}_i} = \sum_{j=1}^{4} \delta^{(4)}_j \cdot \theta^{(3)}_{j,i} \cdot g'(y^{(3)}_i)$, 这里 $i \geq 1$, $j \geq 1$.

(2). 续：观察

$$\delta_i^{(3)} = \frac{\partial J}{\partial y_i^{(3)}} = \sum_{j=1}^{4} \delta_j^{(4)} \cdot \theta_{j,i}^{(3)} \cdot g'(y_i^{(3)})$$

可单独提出来，放在求和外面

$$= \left(\sum_{j=1}^{4} \delta_j^{(4)} \cdot \theta_{j,i}^{(3)}\right) \cdot g'(y_i^{(3)})$$

回顾矩阵乘法 $C_{m\times n} = A_{m\times k} B_{k\times n}$，

$$C_{i,j} = \sum_{k=1}^{k} a_{i,k} \cdot b_{k,j}.$$

$$= \left[\left(\theta_{1,i}^{(3)}, \theta_{2,i}^{(3)}, \cdots, \theta_{j,i}^{(3)}, \cdots, \theta_{4,i}^{(3)}\right) \begin{pmatrix} \delta_1^{(4)} \\ \delta_2^{(4)} \\ \vdots \\ \delta_j^{(4)} \\ \vdots \\ \delta_4^{(4)} \end{pmatrix}\right] \cdot g'(y_i^{(3)})$$

$$= \left[\underbrace{\left(\theta_{1,i}^{(3)}, \theta_{2,i}^{(3)}, \cdots, \theta_{j,i}^{(3)}, \cdots \theta_{4,i}^{(3)}\right)}_{\text{行向量}} \underbrace{\delta^{(4)}}_{\text{列向量}}\right] \cdot g'(y_i^{(3)})$$

计算内积

现在我们记

$$\begin{pmatrix} \delta_1^{(3)} \\ \delta_2^{(3)} \\ \delta_3^{(3)} \\ \delta_4^{(3)} \end{pmatrix} = \begin{pmatrix} \left[\left(\theta_{1,1}^{(3)}, \theta_{2,1}^{(3)}, \theta_{3,1}^{(3)}, \theta_{4,1}^{(3)}\right)\delta^{(4)}\right] \cdot g'(y_1^{(3)}) \\ \left[\left(\theta_{1,2}^{(3)}, \theta_{2,2}^{(3)}, \theta_{3,2}^{(3)}, \theta_{4,2}^{(3)}\right)\delta^{(4)}\right] \cdot g'(y_2^{(3)}) \\ \left[\left(\theta_{1,3}^{(3)}, \theta_{2,3}^{(3)}, \theta_{3,3}^{(3)}, \theta_{4,3}^{(3)}\right)\delta^{(4)}\right] \cdot g'(y_3^{(3)}) \\ \left[\left(\theta_{1,4}^{(3)}, \theta_{2,4}^{(3)}, \theta_{3,4}^{(3)}, \theta_{4,4}^{(3)}\right)\delta^{(4)}\right] \cdot g'(y_4^{(3)}) \end{pmatrix}$$

$$= \begin{pmatrix} g'(y_1^{(3)}) & 0 & 0 & 0 \\ 0 & g'(y_2^{(3)}) & 0 & 0 \\ 0 & 0 & g'(y_3^{(3)}) & 0 \\ 0 & 0 & 0 & g'(y_4^{(3)}) \end{pmatrix} \begin{pmatrix} \left(\theta_{1,1}^{(3)}, \theta_{2,1}^{(3)}, \theta_{3,1}^{(3)}, \theta_{4,1}^{(3)}\right)\delta^{(4)} \\ \left(\theta_{1,2}^{(3)}, \theta_{2,2}^{(3)}, \theta_{3,2}^{(3)}, \theta_{4,2}^{(3)}\right)\delta^{(4)} \\ \vdots \\ \vdots \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial a_1^{(3)}}{\partial y_1^{(3)}} & \frac{\partial a_1^{(3)}}{\partial y_2^{(3)}} & \frac{\partial a_1^{(3)}}{\partial y_3^{(3)}} & \frac{\partial a_1^{(3)}}{\partial y_4^{(3)}} \\ \frac{\partial a_2^{(3)}}{\partial y_1^{(3)}} & \frac{\partial a_2^{(3)}}{\partial y_2^{(3)}} & \frac{\partial a_2^{(3)}}{\partial y_3^{(3)}} & \frac{\partial a_2^{(3)}}{\partial y_4^{(3)}} \\ \frac{\partial a_3^{(3)}}{\partial y_1^{(3)}} & \frac{\partial a_3^{(3)}}{\partial y_2^{(3)}} & \frac{\partial a_3^{(3)}}{\partial y_3^{(3)}} & \frac{\partial a_3^{(3)}}{\partial y_4^{(3)}} \\ \frac{\partial a_4^{(3)}}{\partial y_1^{(3)}} & \frac{\partial a_4^{(3)}}{\partial y_2^{(3)}} & \frac{\partial a_4^{(3)}}{\partial y_3^{(3)}} & \frac{\partial a_4^{(3)}}{\partial y_4^{(3)}} \end{pmatrix} \left(\theta_w^{(3)}\right)^T \delta^{(4)}$$

$$= \underbrace{g'(y^{(3)})}_{4\times4} \underbrace{\left(\theta_w^{(3)}\right)^T}_{4\times4} \underbrace{\delta^{(4)}}_{4\times1}$$

所以我们证明了

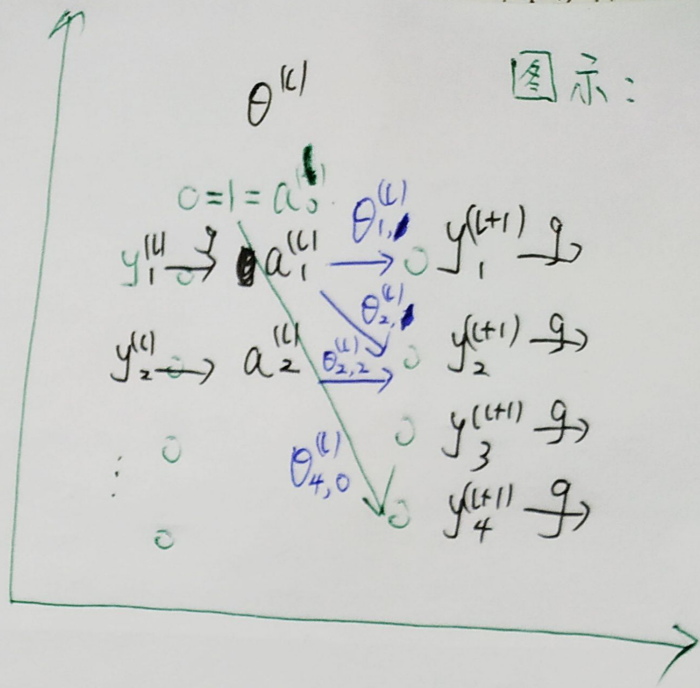
$$\delta^{(3)} = g'(y^{(3)}) \left[\theta_w^{(3)}\right]^T \delta^{(4)}.$$

(3) $$\frac{\partial J(\theta)}{\partial \theta_{j,i}^{(l)}} = \frac{\partial J(\theta)}{\partial y_j^{(l+1)}} \cdot \frac{\partial y_j^{(l+1)}}{\partial \theta_{j,i}^{(l)}}$$

$$y_j^{(l+1)} = \theta_{j,0}^{(l)} a_0^{(l)} + \theta_{j,1}^{(l)} a_1^{(l)} + \cdots + \theta_{j,i}^{(l)} a_i^{(l)} + \cdots$$

$$= \delta_j^{(l+1)} \cdot a_i^{(l)}$$

图示：



即有

$$\frac{\partial J(\theta)}{\partial \theta_{j,i}^{(l)}} = \delta_j^{(l+1)} a_i^{(l)}, \quad i \geq 0, \quad j \geq 1.$$

$$\frac{\partial J(\theta)}{\partial \theta^{(l)}} \overset{\text{记号}}{=\!=} \begin{pmatrix} \frac{\partial J(\theta)}{\partial \theta_{1,0}^{(l)}} & \frac{\partial J(\theta)}{\partial \theta_{1,1}^{(l)}} & \cdots & \frac{\partial J(\theta)}{\partial \theta_{1,i}^{(l)}} & \cdots \\ \vdots & & & & \\ \frac{\partial J(\theta)}{\partial \theta_{j,0}^{(l)}} & \frac{\partial J(\theta)}{\partial \theta_{j,1}^{(l)}} & \cdots & \frac{\partial J(\theta)}{\partial \theta_{j,i}^{(l)}} & \cdots \\ & & \ddots & & \end{pmatrix}$$

$$\theta^{(l)} = \begin{pmatrix} \theta_{1,0}^{(l)}, & \theta_{1,1}^{(l)}, \cdots, & \theta_{1,i}^{(l)}, & \cdots \\ \vdots & & & \\ \theta_{j,0}^{(l)}, & \theta_{j,1}^{(l)}, \cdots & \theta_{j,i}^{(l)}, & \cdots \end{pmatrix}$$

$$= \begin{pmatrix} \delta_1^{(l+1)} a_0^{(l)}, & \delta_1^{(l+1)} a_1^{(l)}, & \cdots & \delta_1^{(l+1)} a_i^{(l)}, & \cdots \\ \vdots & & & & \\ \delta_j^{(l+1)} a_0^{(l)}, & \delta_j^{(l+1)} a_1^{(l)}, & \cdots & \delta_j^{(l+1)} a_i^{(l)}, & \cdots \\ & & \ddots & & \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} \delta_1^{(l+1)} \\ \vdots \\ \delta_j^{(l+1)} \\ \vdots \end{pmatrix}}_{\text{列向量}} \underbrace{\begin{pmatrix} a_0^{(l)}, & a_1^{(l)}, & \cdots & a_i^{(l)}, & \cdots \end{pmatrix}}_{\text{行向量}} = \delta^{(l+1)} \left[a^{(l)}\right]^T$$