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

$\delta_i^{(l+1)}$ ：第 l+1 层第 i 个单元的计算误差（delta） $\delta_i^{(l+1)}=\dfrac{\partial J}{\partial z_i^{(l+1)}}$

s_l ：第 l 层的单元数（不包括偏置单元）。

假设当前为第 l+1 层，单元下标为 i。其中 l 层单元下标为 j，l+1 层单元下标为 k。

$$\frac{\partial J}{\partial \Theta_{ji}^{l}}=\frac{\partial J}{\partial z_i^{(l+1)}} * \frac{\partial z_i^{(l+1)}}{\partial \Theta_{ji}}=\delta_i^{(l+1)} * \frac{\partial z_i^{(l+1)}}{\partial \Theta_{ji}^{l}} \qquad \frac{\partial J}{\partial \Theta_{ji}^{l}}=\frac{\partial J}{\partial a_i^{(l+1)}} * \frac{\partial a_i^{(l+1)}}{\partial z_i^{(l+1)}} * \frac{\partial z_i^{(l+1)}}{\partial \Theta_{ji}^{l}}$$

① $\dfrac{\partial J}{\partial a_i^{(l+1)}}$

若 l+1 层为输出层：

$\dfrac{\partial J}{\partial a_i^{(l+1)}}=a_i-d_i$ ， d_i 为第 i 个样本标签

若 l+1 层为隐藏层：

$$\frac{\partial J}{\partial a_i^{(l+1)}}=\sum_{k=1}^{s_{l+2}}\frac{\partial J}{\partial z_k^{(l+2)}} * \frac{\partial z_k^{(l+2)}}{\partial a_i^{(l+1)}}=\sum_{k=1}^{s_{l+2}}\delta_k^{(l+2)} * \frac{\partial z_k^{(l+2)}}{\partial a_i^{(l+1)}}$$

$$\frac{\partial J}{\partial a_i^{(l+1)}}=\sum_{k=1}^{s_{l+2}}\delta_k^{(l+2)} * \frac{\partial\left(\Theta_{0k}^{(l+1)} * a_0^{(l+1)}+\Theta_{1k}^{(l+1)} * a_1^{(l+1)}+...+\Theta_{s_{l+1}k}^{(l+1)} * a_{s_{l+1}}^{(l+1)}\right)}{\partial a_i^{(l+1)}}$$

$$\frac{\partial J}{\partial a_i^{(l+1)}}=\sum_{k=1}^{s_{l+2}}\delta_k^{(l+2)} * \Theta_{ik}^{(l+1)}$$

② $\dfrac{\partial a_i^{(l+1)}}{\partial z_i^{(l+1)}}=\dfrac{\partial g\left(z_i^{(l+1)}\right)}{\partial z_i^{(l+1)}}=g\left(z_i^{(l+1)}\right) * \left(1-g\left(z_i^{(l+1)}\right)\right)=a_i^{(l+1)} * \left(1-a_i^{(l+1)}\right)$

③ $\dfrac{\partial z_i^{(l+1)}}{\partial \Theta_{ji}^{(l)}}=\dfrac{\partial\left(\sum\limits_{j=0}^{s_l}\Theta_{ji}^{(l)}\ *\ a_j^{(l)}\right)}{\partial \Theta_{ji}^{(l)}}=\dfrac{\partial\left(\Theta_{0i}^{(l)}\ *\ a_0^{(l)}+\Theta_{1i}^{(l)}\ *\ a_1^{(l)}+...+\Theta_{s_li}^{(l)}\ *\ a_{s_l}^{(l)}\right)}{\partial \Theta_{ji}^{(l)}}=a_j^{(l)}$

总结：

$\dfrac{\partial J}{\partial \Theta_{ji}^{(l)}}$ = ① * ② * ③

①：传播到隐藏层的误差（errors propagated to the hidden layer）
②：隐藏层梯度（hidden layer gradients）
① * ②：隐藏层计算误差（hidden layer delta）

若 l+1 层为输出层：

$\delta^{(l+1)}=\dfrac{\partial J}{\partial z^{(l+1)}}$ = ① * ② = $\left(a_i-d_i\right)\ *\ a^{(l+1)}\ *\left(1-a^{(l+1)}\right)$

若 l+1 层为隐藏层：

$\delta^{(l+1)}=\dfrac{\partial J}{\partial z^{(l+1)}}$ = ① * ② = $\left(\Theta^{(l+1)}\right)^T\circ\delta^{(l+2)}\ *\ a^{(l+1)}\ *\left(1-a^{(l+1)}\right)$

$\dfrac{\partial J}{\partial \Theta^{(l)}}=\delta^{(l+1)}\circ\left(a^{(l)}\right)^T$

$\Delta \Theta=\eta\ *\ \dfrac{\partial J}{\partial \Theta^{(l)}}$ ， η 是学习率（learning_rate）

参考索引

https://my.oschina.net/findbill/blog/529001

http://blog.csdn.net/qrlhl/article/details/50885527