# Back Propagation algorithm proof.

1. Notation

$L$ : total number of layers

$S_l$ = number of units (not including bias unit) in layer $l$.

$K$ = number of output units / classes, in the $L$ Layer

$\theta_{j,i}^{(l)}$ : weight from the ith neuron in the $l^{th}$ layer to the $j^{th}$ neuron in the $(l+1)^{th}$ layer

$a_j^{(l)}$ = activation of the $j^{th}$ neuron in the $l^{th}$ layer

$\delta_j^{(l)}$ : error of the $j^{th}$ neuron in layer $l$.

$\therefore \quad a_j^{(l)} = \sigma\left(\sum_{i=1}^{S_{l-1}+1} \theta_{j,i}^{(l-1)} \cdot a_i^{(l-1)}\right)$ [including the bias unit]

Then vectorized it:

$$a^{(l)} = \sigma\left(\theta^{(l-1)} \cdot a^{(l-1)}\right)$$

To compute $a^{(l)}$, we introduce the intermediate quantity:

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

Then $a^{(l)} = \sigma(z^{(l)})$

and $z_j^{(l)} = \sum_{i=1}^{S_{l}+1} \theta_{j,i}^{(l-1)} \cdot a_i^{(l-1)}$

Our purpose is to calculate the partial derivative of $J(\theta)$; $\frac{\partial}{\partial \theta_{j,i}^{(l)}} J(\theta)$.

The $J(\theta)$ can generialized as:

$J(\theta) = \frac{1}{m} \sum_{x} C_x$. $C_x$ is an individual training example.

$$\underline{J(\theta) = f(a^{(L)})}$$

2. Four fundametal equations behind BP

We need another intermediate quantity. $\delta_j^{(l)}$

$$\underline{\underline{\delta_j^{(l)} = \frac{\partial J}{\partial z_j^{l}}}}$$

$$\delta_j^{(L)} = \frac{\partial J}{\partial a_j^{(L)}} \sigma'(z_j^{(L)}) \quad (BP1)$$

$$\delta^{(l)} = \underline{\left[(\Theta^l)^T \cdot \delta^{(l+1)}\right)} \cdot \sigma'(z^{(l)}) \quad (BP2)$$

$$\frac{\partial}{\partial \Theta_{j \cdot i}^{(l)}} J = a_i^{(l)} \delta_j^{(l+1)}$$

$$(BP3)$$

$$\frac{\partial J}{\partial \Theta_{ji}^{(l)}} = a_{in} \delta_{out}.$$

proof.

$$\delta_j^{(L)} = \frac{\partial J}{\partial z_j^{(L)}} = \frac{\partial J}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}$$

$$\because a_j^{(L)} = \sigma(z_j^{(L)})$$

$$\therefore \delta_j^{(L)} = \frac{\partial J}{\partial a_j^{(L)}} \cdot \sigma'(z_j^{(L)}) \quad (BP1)$$

$\partial z_k^{(l+1)}$

$$\delta_j^{(l)} = \frac{\partial J}{\partial z_j^{(l)}} = \sum_k \frac{\partial J}{\partial z_k^{(l+1)}} \cdot \frac{\partial z_k}{\partial z_j^{(l)}}$$

$$= \sum_k \frac{\partial z_k^{(l+1)}}{\partial z_j^{(l)}} \cdot \delta_k^{(l+1)}$$

$$z_k^{(l+1)} = \sum_{i=1}^{s_{l+1}} \theta_{j,i}^{(l)} \cdot a_i^{(l)}$$

$$= \sum_{i=1}^{s_{l+1}} \theta_{j,i}^{(l)} \cdot \sigma(z_i^{(l)})$$

$$\therefore \frac{\partial z_k^{(l+1)}}{\partial z_j^{(l)}} = \theta_{k,j}^{(l)} \cdot \sigma'(z_j^{(l)})$$

$$\therefore \delta_j^{(l)} = \sum_k \theta_{k,j}^{(l)} \cdot \delta_k^{(l+1)} \cdot \sigma'(z_j^{(l)})$$

$$\frac{\partial J}{\partial \theta^{(l-1)}} = \left|\frac{\partial J}{\partial a^{(l)}}\right| \left|\frac{\partial a^{(l)}}{\partial z^{(l)}}\right| \cdot \frac{\partial z^{(l)}}{\partial \theta^{(l-1)}}$$

$$= \delta^{(l)} \cdot \frac{\partial z}{\partial(\theta^{(l-1)})}$$

$$= \delta^{(l)} \cdot a^{(l-1)}$$

proof complete!

$$\therefore \frac{\partial J}{\partial \theta_{ji}^{(l)}} = \frac{1}{m} \cdot \sum_{t=1}^{m} a_i^{(t)(l)} \cdot \delta_j^{(t)(l+1)}$$