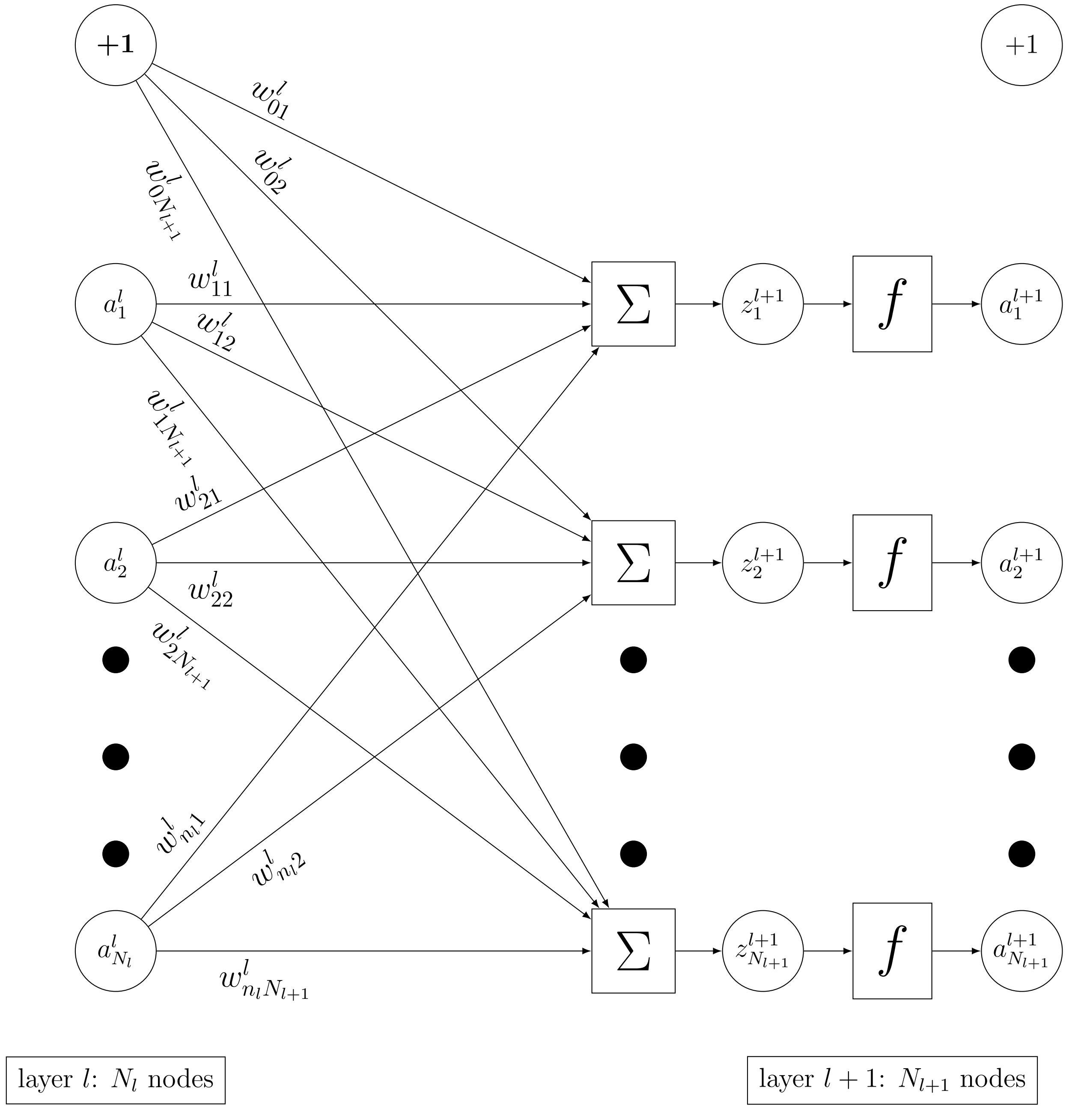
Backpropagation

June 21, 2017

1 Derivation of backpropagation equations in the general case



+1
+1
w^l_{01}
w^l_{02}
$w^l_{0N_{l+1}}$
a^l_1
w^l_{11}
w^l_{12}
$w^l_{1N_{l+1}}$
w^l_{21}
a^l_2
w^l_{22}
$w^l_{2N_{l+1}}$
$w^l_{n_l1}$
$w^l_{n_l2}$
$a^l_{N_l}$
$w^l_{n_lN_{l+1}}$
Σ
z^{l+1}_1
f
a^{l+1}_1
z^{l+1}_2
a^{l+1}_2
$z^{l+1}_{N_{l+1}}$
$a^{l+1}_{N_{l+1}}$
layer l: N_l nodes
layer $l+1$: N_{l+1} nodes

The image above illustrates forward propagation, in other words, how the values related to a layer $l+1$ of a network are calculated using the values of the previous layer. Let us define the following:

L: number of layers
N_l: number of nodes in layer l
N_L: number of nodes in the last layer, also the number of classes of the problem
P: number of data points/examples
Q: number of features of each example/coordinates for each data point
$h^p_{n_L}$: output of node n_L of the last layer when the model is applied to data point $\mathbf{x}^p = (x_1, ..., x_q, ..., x_Q)$
$y^p_{n_L}$: true output corresponding to $h^p_{n_L}$
$z^{p,l}_{n_l}$: value in layer l corresponding to data point p before activation
f: activation function
$a^{p,l}_{n_l} = f(z^{p,l}_{n_l})$: value in layer l corresponding to data point p after activation
$w^l_{n_l n_{l+1}}$: weight in layer l multiplying $a^{p,l}_{n_l}$

The expression to calculate $z^{p,l+1}_{n_{l+1}}$ using values related to the previous layer is:

$$z^{p,l+1}_{n_{l+1}} = w^l_{n_0 n_{l+1}} + \sum_{n_l=1}^{N_l} w^l_{n_l n_{l+1}} a^{p,l}_{n_l} = w_{n_0 n_{l+1}} + \sum_{n_l=1}^{N_l} w^l_{n_l n_{l+1}} f(z^{p,l}_{n_l}) \tag{1}$$

The cost function can be generally defined as:

$$C = \sum_{p=1}^{P} \sum_{n_L=1}^{N_L} c(y^p_{n_L}, h^p_{n_L}) = \sum_{p=1}^{P} \sum_{n_L=1}^{N_L} c^p_{n_L} \tag{2}$$

Taking the derivative of C with respect to a weight $w^l_{n_l n_{l+1}}$:

$$\frac{\partial C}{\partial w^l_{n_l n_{l+1}}} = \sum_{p=1}^{P} \sum_{n_L=1}^{N_L} \frac{\partial c^p_{n_L}}{\partial w^l_{n_l n_{l+1}}} \tag{3}$$

using the derivative chain rule:

$$\frac{\partial c^p_{n_L}}{\partial w^l_{n_l n_{l+1}}} = \frac{\partial c^p_{n_L}}{\partial z^{p,l+1}_{n_{l+1}}} \frac{\partial z^{p,l+1}_{n_{l+1}}}{\partial w^l_{n_l n_{l+1}}} \tag{4}$$

$$\frac{\partial z^{p,l+1}_{n_{l+1}}}{\partial w^l_{0 n_{l+1}}} = 1 \tag{5}$$

$$\frac{\partial z^{p,l+1}_{n_{l+1}}}{\partial w^l_{n_l n_{l+1}}} = a^{p,l}_{n_l} \tag{6}$$

$$\frac{\partial c^p_{n_L}}{\partial z^{p,l+1}_{n_{l+1}}} = \sum_{n_{L+2}=1}^{N_{l+2}} \frac{\partial c^p_{n_L}}{\partial z^{p,l+2}_{n_{l+2}}} \frac{\partial z^{p,l+2}_{n_{l+2}}}{\partial z^{p,l+1}_{n_{l+1}}} \tag{7}$$

$$\frac{\partial z^{p,l+2}_{n_{l+2}}}{z^{p,l+1}_{n_{l+1}}} = w^{l+1}_{n_{l+1} n_{l+2}} f'(z^{p,l+1}_{n_{l+1}}) \tag{8}$$

$$\begin{aligned} \frac{\partial c^p_{n_L}}{\partial z^{p,l+1}_{n_{l+1}}} &= \sum_{n_{l+2}=1}^{N_{l+2}} \frac{\partial c^p_{n_L}}{\partial z^{p,l+2}_{n_{l+2}}} w^{l+1}_{n_{l+1} n_{l+2}} f'(z^{p,l+1}_{n_{l+1}}) \\ &= f'(z^{p,l+1}_{n_{l+1}}) \sum_{n_{l+2}=1}^{N_{l+2}} w^{l+1}_{n_{l+1} n_{l+2}} \frac{\partial c^p_{n_L}}{\partial z^{p,l+2}_{n_{l+2}}} \end{aligned}$$

$$\begin{cases} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l+1}_{n_{l+1}}} & = f'(z^{p,l+1}_{n_{l+1}}) \sum_{n_{l+2}=1}^{N_{l+2}} w^{l+1}_{n_{l+1}n_{l+2}} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l+2}_{n_{l+2}}} \\ \dfrac{\partial C}{\partial w^l_{0n_{l+1}}} & = \sum_{p=1}^{P} \sum_{n_L=1}^{N_L} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l+1}_{n_{l+1}}} \\ \dfrac{\partial C}{\partial w^l_{n_l n_{l+1}}} & = \sum_{p=1}^{P} a^{p,l}_{n_l} \sum_{n_L=1}^{N_L} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l+1}_{n_{l+1}}} \end{cases} \tag{9}$$

All w, z and a values have already been computed during forward propagation, and $f'(z)$ value can easily be computed. Therefore, if the $\dfrac{\partial c^p}{\partial z^{p,l+2}_{n_{l+2}}}$ values are known, $\dfrac{\partial c^p_{n_L}}{\partial z^{p,l+1}_{n_{l+1}}}$ and thus $\dfrac{\partial C}{\partial w^l_{0n_{l+1}}}$ and $\dfrac{\partial C}{\partial w^l_{n_l n_{l+1}}}$ can be computed using the above equalities. Starting from the last layer, we can computer all those value up to the first layer. Since L is the last layer, for clarity we can rewrite the above system as:

$$\begin{cases} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l-1}_{n_{l-1}}} & = f'(z^{p,l-1}_{n_{l-1}}) \sum_{n_l=1}^{N_l} w^{l-1}_{n_{l-1}n_l} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l}_{n_l}} \\ \dfrac{\partial C}{\partial w^{l-2}_{0n_{l-1}}} & = \sum_{p=1}^{P} \sum_{n_L=1}^{N_L} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l-1}_{n_{l-1}}} \\ \dfrac{\partial C}{\partial w^{l-2}_{n_{l-2}n_{l-1}}} & = \sum_{p=1}^{P} a^{p,l-2}_{n_{l-2}} \sum_{n_L=1}^{N_L} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l-1}_{n_{l-1}}} \end{cases} \tag{10}$$

2 One output layer

If the output layer only contains one neuron, $N_l = 1$, so there is no need for the inner summation.

$$\begin{cases} \dfrac{\partial c^p}{\partial z^{p,l-1}_{n_{l-1}}} & = f'(z^{p,l-1}_{n_{l-1}}) w^{l-1}_{n_{l-1}n_l} \dfrac{\partial c^p}{\partial z^{p,l}_{n_l}} \\ \dfrac{\partial C}{\partial w^{l-2}_{0n_{l-1}}} & = \sum_{p=1}^{P} \dfrac{\partial c^p}{\partial z^{p,l-1}_{n_{l-1}}} \\ \dfrac{\partial C}{\partial w^{l-2}_{n_{l-2}n_{l-1}}} & = \sum_{p=1}^{P} a^{p,l-2}_{n_{l-2}} \dfrac{\partial c^p}{\partial z^{p,l-1}_{n_{l-1}}} \end{cases} \tag{11}$$

3 Sigmoid activation function

If sigmoid activation functions are used:

$$c^p_{n_L} = y^p_{n_L} \log(h^p_{n_L}) + (1 - y^p_{n_L}) \log(1 - h^p_{n_L}) \tag{12}$$

Where

$$\sigma(z^{p,l-1}_{n_{l-1}}) = a^{p,l-1}_{n_{l-1}}$$
$$h^p_{n_L} = \sigma(z^{p,L}_{n_L}) = a^{p,L}_{n_L}$$

It can be checked that $\dfrac{\partial c^p_{n_L}}{\partial z^{p,L}_{n_L}} = y^p_{n_L} - h^p_{n_L}$ which will be denoted $d^p_{n_L}$.

$$\begin{cases} \dfrac{\partial c^p_{n_L}}{\partial z^{p,L}_{n_L}} & = y^p_{n_L} - h^p_{n_L} \\ \dfrac{\partial c^p_{n_L}}{\partial z^{p,l-1}_{n_{l-1}}} & = a^{p,l-1}_{n_{l-1}}(1 - a^{p,l-1}_{n_{l-1}}) \sum_{n_l=1}^{N_l} w^{l-1}_{n_{l-1}n_l} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l}_{n_l}} \\ \dfrac{\partial C}{\partial w^{l-2}_{0n_{l-1}}} & = \sum_{p=1}^{P} \sum_{n_L=1}^{N_L} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l-1}_{n_{l-1}}} \\ \dfrac{\partial C}{\partial w^{l-2}_{n_{l-2}n_{l-1}}} & = \sum_{p=1}^{P} a^{p,l-2}_{n_{l-2}} \sum_{n_L=1}^{N_L} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l-1}_{n_{l-1}}} \end{cases} \tag{13}$$

4 Two-layer network

In this case $a^{p,l-2}_{n_{l-2}}$ comes from the input layer which does not have an activation function (or has the identity activation function if you prefer) . Let us write denote $n_0 = q$ so $a^{p,l-2}_{n_{l-2}} = x^p_q$.

$$\begin{cases} \dfrac{\partial c^p_{n_2}}{\partial z^{p,1}_{n_1}} &= f'(z^{p,1}_{n_1}) \sum_{n_l=1}^{N_2} w^1_{n_1 n_l} \dfrac{\partial c^p_{n_2}}{\partial z^{p,2}_{n_2}} \\ \dfrac{\partial C}{\partial w^0_{0n_1}} &= \sum_{p=1}^{P} \sum_{n_2=1}^{N_2} \dfrac{\partial c^p_{n_2}}{\partial z^{p,l-1}_{n_1}} \\ \dfrac{\partial C}{\partial w^0_{qn_1}} &= \sum_{p=1}^{P} x^p_q \sum_{n_L=1}^{N_L} \dfrac{\partial c^p_{n_L}}{\partial z^{p,l-1}_{n_{l-1}}} \end{cases} \tag{14}$$

5 The 2-2-1 architecture with sigmoid activation

Here, by 2-2-1 is meant a 1 hidden layer network, with 2 inputs, 2 neurons in the hidden layer and 1 output.

$$\begin{cases} \dfrac{\partial C}{\partial w^0_{0n_1}} &= w^1_{n_1} \sum_{p=1}^{2} a^{p,1}_{n_1}(1 - a^{p1}_{n_1})(y^p - h^p) \\ \dfrac{\partial C}{\partial w^0_{qn_1}} &= w^1_{n_1} \sum_{p=1}^{2} x^p_q a^{p,1}_{n_1}(1 - a^{p1}_{n_1})(y^p - h^p) \end{cases} \tag{15}$$

If you understand the notations, (which you should!) you will readily notice those equations are the same we had before fo $\mathbf{U}$. For v, which is the coefficient vector for the layer before the output, we can take the derivative directly and there is not backpropagation equation!