Mathematical Analysis of Feedforward Neural Networks

Calvin Osborne

We will denote the weight from node N_j^L to node N_k^{L+1} by $w_{j,k}^L$ and the bias of node N_k^{L+1} by b_k^L. The activation function used will be denoted by σ. From these definitions the forward propagation process can be denoted by

$$y_k^{L+1} = \sigma(z_k^{L+1}) = \sigma(b_k^L + \sum_j w_{j,k}^L y_j^L).$$

An entire layer of the network can then be represented by

$$\begin{bmatrix} y_1^{L+1} \\ y_2^{L+1} \\ \vdots \\ y_k^{L+1} \end{bmatrix} = \begin{bmatrix} \sigma(z_1^{L+1}) \\ \sigma(z_2^{L+1}) \\ \vdots \\ \sigma(z_k^{L+1}) \end{bmatrix} = \begin{bmatrix} \sigma(b_1^L + \sum_j w_{j,1}^L y_j^L) \\ \sigma(b_2^L + \sum_j w_{j,2}^L y_j^L) \\ \vdots \\ \sigma(b_k^L + \sum_j w_{j,k}^L y_j^L) \end{bmatrix}.$$

We will now spend the bulk of our time with the calculus behind back-propagation. We will consider an arbitrary cost function $C(y_1^n, y_2^n, \ldots, y_k^n)$ for a network of n layers. We are interesting in evaluating the three values $\partial C/\partial w$, $\partial C/\partial b$, and $\partial C/\partial y$. We will start by considering only layer $n-1$, but then we will extend this method to any layer's weights or biases.

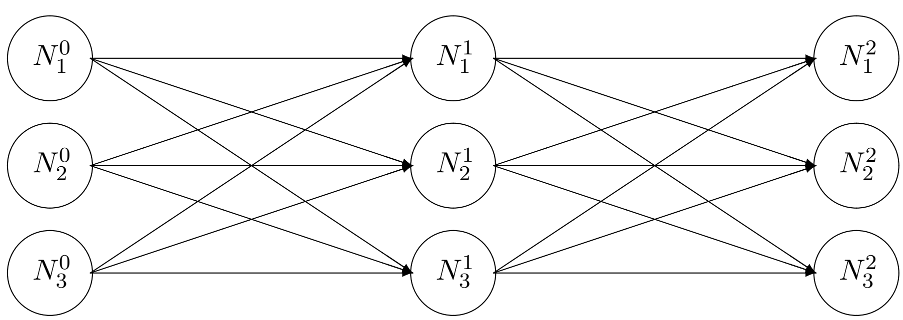

Figure 1: Structure of a basic Feed-forward Neural Network

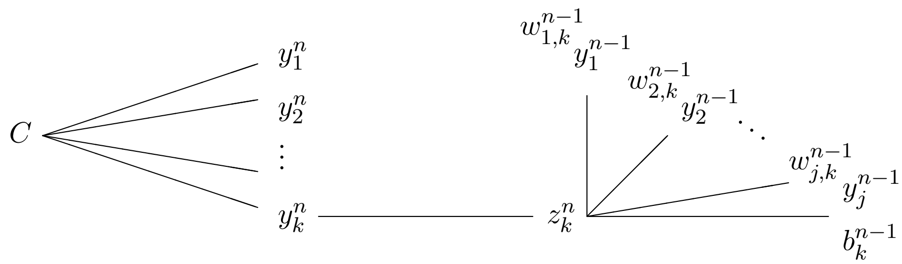

Figure 2: Chain Rule Diagram representing the last layer of the Network

To begin, note that

$$\begin{aligned}\frac{\partial C}{\partial w_{j,k}^{n-1}} &= \frac{\partial C}{\partial y_k^n} \cdot \frac{\partial y_k^n}{\partial z_k^n} \cdot \frac{\partial z_k^n}{\partial w_{j,k}^{n-1}} \\ &= \frac{\partial C}{\partial y_k^n} \cdot \sigma'(z_k^n) \cdot y_j^{n-1}. \qquad (1)\end{aligned}$$

Similarly, we can derive that

$$\begin{aligned}\frac{\partial C}{\partial b_k^{n-1}} &= \frac{\partial C}{\partial y_k^n} \cdot \frac{\partial y_k^n}{\partial z_k^n} \cdot \frac{\partial z_k^n}{\partial b_k^{n-1}} \\ &= \frac{\partial C}{\partial y_k^n} \cdot \sigma'(z_k^n). \qquad (2)\end{aligned}$$

Finally, to evaluate $\partial C / \partial y_j^{n-1}$, note that this value is used in calculating every z^n since each node between layers is connected. Hence,

$$\begin{aligned}\frac{\partial C}{\partial y_j^{n-1}} &= \frac{\partial C}{\partial y_1^n} \cdot \frac{\partial y_1^n}{\partial z_1^n} \cdot \frac{\partial z_1^n}{\partial y_j^{n-1}} + \frac{\partial C}{\partial y_2^n} \cdot \frac{\partial y_2^n}{\partial z_2^n} \cdot \frac{\partial z_2^n}{\partial y_j^{n-1}} + \cdots + \frac{\partial C}{\partial y_k^n} \cdot \frac{\partial y_k^n}{\partial z_k^n} \cdot \frac{\partial z_k^n}{\partial y_j^{n-1}} \\ &= \sum_k \frac{\partial C}{\partial y_k^n} \cdot \frac{\partial y_k^n}{\partial z_k^n} \cdot \frac{\partial z_k^n}{\partial y_j^{n-1}} \\ &= \sum_k \frac{\partial C}{\partial y_k^n} \cdot \sigma'(z_k^n) \cdot w_{j,k}^{n-1}. \qquad (3)\end{aligned}$$

It isn't too hard to extend these formulas to any layer in the network by noting that

$$\frac{\partial C}{\partial w_{j,k}^l} = \frac{\partial C}{\partial y_k^{l+1}} \cdot \frac{\partial y_k^{l+1}}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial w_{j,k}^l} = \frac{\partial C}{\partial y_k^{l+1}} \cdot \sigma'(z_k^{l+1}) \cdot y_j^l. \qquad (4)$$

Similarly it follows that

$$\frac{\partial C}{\partial b_k^l} = \frac{\partial C}{\partial y_k^{l+1}} \cdot \sigma'(z_k^{l+1}) \tag{5}$$

and

$$\frac{\partial C}{\partial y_j^l} = \sum_k \frac{\partial C}{\partial y_k^{l+1}} \cdot \sigma'(z_k^{l+1}) \cdot w_{j,k}^l. \tag{6}$$

Representing these equations in vector form, we can derive that

$$\begin{bmatrix} \partial C/\partial b_1^l \\ \partial C/\partial b_2^l \\ \vdots \\ \partial C/\partial b_k^l \end{bmatrix} = \begin{bmatrix} \partial C/\partial y_1^{l+1} \\ \partial C/\partial y_2^{l+1} \\ \vdots \\ \partial C/\partial y_k^{l+1} \end{bmatrix} \circ \begin{bmatrix} \sigma'(z_1^{l+1}) \\ \sigma'(z_2^{l+1}) \\ \vdots \\ \sigma'(z_k^{l+1}) \end{bmatrix}$$

or

$$\nabla_{b^l} \boldsymbol{C} = \nabla_{y^{l+1}} \boldsymbol{C} \circ \sigma'(\boldsymbol{z}^{l+1}). \tag{7}$$

Similarly it follows that

$$\begin{bmatrix} \partial C/\partial w_{1,1}^l & \partial C/\partial w_{2,1}^l & \dots & \partial C/\partial w_{j,1}^l \\ \partial C/\partial w_{1,2}^l & \partial C/\partial w_{2,2}^l & \dots & \partial C/\partial w_{j,2}^l \\ \vdots & \vdots & \ddots & \vdots \\ \partial C/\partial w_{1,k}^l & \partial C/\partial w_{2,k}^l & \dots & \partial C/\partial w_{j,k}^l \end{bmatrix} = \left(\begin{bmatrix} \partial C/\partial y_1^{l+1} \\ \partial C/\partial y_2^{l+1} \\ \vdots \\ \partial C/\partial y_k^{l+1} \end{bmatrix} \circ \begin{bmatrix} \sigma'(z_1^{l+1}) \\ \sigma'(z_2^{l+1}) \\ \vdots \\ \sigma'(z_k^{l+1}) \end{bmatrix} \right) \begin{bmatrix} y_1^l \\ y_2^l \\ \vdots \\ y_j^l \end{bmatrix}^\top$$

which can be alternatively written as

$$\nabla_{w^l} \boldsymbol{C} = (\nabla_{y^{l+1}} \boldsymbol{C} \circ \sigma'(\boldsymbol{z}^{l+1})) \cdot (\boldsymbol{y}^l)^\top \tag{8}$$

or even

$$\nabla_{w^l} \boldsymbol{C} = \nabla_{b^l} \boldsymbol{C} \cdot (\boldsymbol{y}^l)^\top. \tag{9}$$

Finally, we can express $\nabla_{y^l} \boldsymbol{C}$ as

$$\begin{bmatrix} \partial C/\partial y_1^l \\ \partial C/\partial y_2^l \\ \vdots \\ \partial C/\partial y_j^l \end{bmatrix} = \begin{bmatrix} w_{1,1}^l & w_{2,1}^l & \dots & w_{j,1}^l \\ w_{1,2}^l & w_{2,2}^l & \dots & w_{j,2}^l \\ \vdots & \vdots & \ddots & \vdots \\ w_{1,k}^l & w_{2,k}^l & \dots & w_{j,k}^l \end{bmatrix}^\top \cdot \left(\begin{bmatrix} \sigma'(z_1^{l+1}) \\ \sigma'(z_2^{l+1}) \\ \vdots \\ \sigma'(z_k^{l+1}) \end{bmatrix} \circ \begin{bmatrix} \partial C/\partial y_1^{l+1} \\ \partial C/\partial y_2^{l+1} \\ \vdots \\ \partial C/\partial y_k^{l+1} \end{bmatrix} \right)$$

or

$$\nabla_{y^l} \boldsymbol{C} = (\boldsymbol{w}^l)^\top \cdot (\sigma'(\boldsymbol{z}^{l+1}) \circ \nabla_{y^{l+1}} \boldsymbol{C}). \tag{10}$$

We will use a small shift in notation to make these formulas cleaner to compute in practice. We will let

$$\boldsymbol{\delta}^l = ((\boldsymbol{w}^l)^\top \boldsymbol{\delta}^{l+1}) \circ \sigma'(\boldsymbol{z}^l) \tag{11}$$

for all $l < n$ and

$$\boldsymbol{\delta}^n = \sigma'(\boldsymbol{z}^n) \circ \nabla_{y^n} \boldsymbol{C}. \tag{12}$$

Note that from these definitions $\boldsymbol{\delta}^l = \nabla_{y^l} \boldsymbol{C} \circ \sigma'(\boldsymbol{z}^l)$. The following expressions also hold, being only slight modifications from (7) and (9):

$$\nabla_{b^l} \boldsymbol{C} = \boldsymbol{\delta}^{l+1} \tag{13}$$

and

$$\nabla_{w^l} \boldsymbol{C} = \boldsymbol{\delta}^{l+1} \cdot (\boldsymbol{y}^l)^\top. \tag{14}$$

Thus, computing $\nabla_{b^l} \boldsymbol{C}$ and $\nabla_{w^l} \boldsymbol{C}$ depend only on calculating successive values of $\boldsymbol{\delta}^{l+1}$.