Formulae for a Neural Network with 2-hidden layers

Abstract

Neural Networks (NNs) are useful for function approximation based on a given set of relevant data. Shallow and deep NNs are being used extensively. We carry through the algebra for a NN with two hidden layers.

Let $x \in R^D$, be the vector of input variables, i.e. $x = \{x_i | i = 1, 2, \cdots, D\}$. Let the nodes be represented by Cartesian pairs: (L, K), where $L = 1, 2$ specifies the level of the hidden layer and $K = 1, 2, \cdots$, the related neuron. Namely, (L, K) corresponds to the K^{th} node residing on the L^{th} hidden layer. The NN we consider here has a single output $N(x, a, w, c, v, b)$. For the output node, the identity activation is assumed.

The Network

$$N(x, \theta) = \sum_{j=1}^{M_2} a_j O_{2j} \tag{1a}$$

$$O_{2j} = f(A_{2j}), \text{ with } A_{2j} = \sum_{k=1}^{M_1} w_{jk} O_{1k} + c_j, \ \forall j = 1, \cdots, M_2 \tag{1b}$$

$$O_{1k} = f(A_{1k}), \text{ with } A_{1k} = \sum_{l=1}^{D} v_{kl} x_l + b_k, \ \forall k = 1, \cdots, M_1 \tag{1c}$$

O_{LK} denotes the output of neuron (L, K).
v_{kl}, b_k denote here the weights and biases entering the first hidden layer, while w_{jk}, c_j the weights and biases entering the second hidden layer.
$f(.)$ stands for the node activation, and θ stands collectively for a, w, c, v and b.
The first hidden layer is assumed to contain M_1 neurons. Correspondingly, the second layer is assumed to contain M_2 neurons.

The Network's Gradient: $\nabla_\theta N(x, \theta)$

The gradient of the NN wrt its weights and biases

$$\frac{\partial N(x,\theta)}{\partial a_j} = f(A_{2j}), \qquad \forall j = 1, \cdots, M_2 \tag{2a}$$

$$\frac{\partial N(x,\theta)}{\partial w_{jk}} = a_j f'(A_{2j}) f(A_{1k}), \qquad \forall j = 1, \cdots, M_2,\ \forall k = 1, \cdots, M_1 \tag{2b}$$

$$\frac{\partial N(x,\theta)}{\partial c_j} = a_j f'(A_{2j}) \qquad \forall j = 1, \cdots, M_2 \tag{2c}$$

$$\frac{\partial N(x,\theta)}{\partial v_{kl}} = f'(A_{1k}) x_l \sum_{j=1}^{M_2} a_j f'(A_{2j}) w_{jk} \qquad \forall k = 1, \cdots, M_1,\ \forall l = 1, \cdots, D \tag{2d}$$

$$\frac{\partial N(x,\theta)}{\partial b_k} = f'(A_{1k}) \sum_{j=1}^{M_2} a_j f'(A_{2j}) w_{jk} \qquad \forall k = 1, \cdots, M_1 \tag{2e}$$

The Network's Gradient: $\nabla_x N(x, \theta)$

The gradient of the NN wrt its input x

$\forall\ l = 1, \cdots, D$:

$$\frac{\partial N(x,\theta)}{\partial x_l} = \sum_{j=1}^{M_2} a_j \frac{\partial O_{2j}}{\partial x_l} \tag{3a}$$

$$\frac{\partial O_{2j}}{\partial x_l} = f'(A_{2j}) \frac{\partial A_{2j}}{\partial x_l},\ \text{with}\ \frac{\partial A_{2j}}{\partial x_l} = \sum_{k=1}^{M_1} w_{jk} \frac{\partial O_{1k}}{\partial x_l},\ \forall j = 1, \cdots, M_2 \tag{3b}$$

$$\frac{\partial O_{1k}}{\partial x_l} = f'(A_{1k}) \frac{\partial A_{1k}}{\partial x_l},\ \text{with}\ \frac{\partial A_{1k}}{\partial x_l} = v_{kl},\ \forall k = 1, \cdots, M_1 \tag{3c}$$

$$\text{yielding:}\ \frac{\partial N(x,\theta)}{\partial x_l} = \sum_{j=1}^{M_2} a_j f'(A_{2j}) \sum_{k=1}^{M_1} w_{jk} f'(A_{1k}) v_{kl} \tag{3d}$$

The Network's Hessian: $\nabla_x^2 N(x,\theta)$

The Hessian of the NN wrt its input x

$$\frac{\partial^2 N(x,\theta)}{\partial x_i \partial x_l} =$$
$$\sum_{j=1}^{M_2}\sum_{k=1}^{M_1} a_j w_{jk} v_{kl}\left(f'(A_{2j})f''(A_{1k})v_{ki} + f''(A_{2j})f'(A_{1k})\sum_{m=1}^{M_1} w_{jm}f'(A_{1m})v_{mi}\right) \tag{4}$$

Mixed derivatives: $\nabla_{x\theta}^2 N(x,\theta)$

Derivatives wrt x and wrt θ

We use the Kronecker symbol $\delta_{li} = 1$,if $l = i$ and zero otherwise.

$$\frac{\partial^2 N(x,\theta)}{\partial x_i \partial a_j} = f'(A_{2j})\sum_{k=1}^{M_1} w_{jk}f'(A_{1k})v_{ki}, \forall j = 1,\cdots,M_2 \tag{5a}$$

$$\frac{\partial^2 N(x,\theta)}{\partial x_i \partial w_{jk}} = a_j\left(f''(A_{2j})f(A_{1k})\sum_{m=1}^{M_1} w_{jm}f'(A_{1m})v_{mi} + f'(A_{2j})f'(A_{1k})v_{ki}\right) \tag{5b}$$

$$\frac{\partial^2 N(x,\theta)}{\partial x_i \partial c_j} = a_j f''(A_{2j})\sum_{k=1}^{M_1} w_{jk}f'(A_{1k})v_{ki} \tag{5c}$$

$$\begin{aligned}\frac{\partial^2 N(x,\theta)}{\partial x_i \partial v_{kl}} &= \left(f''(A_{1k})v_{ki}x_l + f'(A_{1k})\delta_{li}\right)\sum_{j=1}^{M_2} a_j w_{jk}f'(A_{2j}) \\ &+ f'(A_{1k})x_l\sum_{j=1}^{M_2} a_j w_{jk}f''(A_{2j})\sum_{m=1}^{M_1} w_{jm}f'(A_{1m})v_{mi}\end{aligned} \tag{5d}$$

$$\begin{aligned}\frac{\partial^2 N(x,\theta)}{\partial x_i \partial b_k} &= f''(A_{1k})v_{ki}\sum_{j=1}^{M_2} a_j w_{jk}f'(A_{2j}) \\ &+ f'(A_{1k})\sum_{m=1}^{M_1} a_j w_{jm}f''(A_{2j})\sum_{m=1}^{M_1} w_{jm}v_{mi}f'(A_{1m})\end{aligned} \tag{5e}$$

Mixed third order derivatives $\nabla_\theta \nabla_x^2$

Second order derivatives wrt x and first order wrt θ

In order to simplify the expressions we will use the following shorthands:

$$B_{ji} \equiv \sum_{m=1}^{M_1} w_{jm} v_{mi} f'(A_{1m}),\ C_{jil} \equiv \frac{\partial B_{ji}}{\partial x_l} = \sum_{m=1}^{M_1} w_{jm} v_{mi} v_{ml} f''(A_{1m})$$

$$\frac{\partial^3 N(x,\theta)}{\partial x_i \partial x_l \partial a_j} = f''(A_{2j}) B_{ji} B_{jl} + f'(A_{2j}) C_{jil} \tag{6a}$$

$$\begin{aligned}\frac{\partial^3 N(x,\theta)}{\partial x_i \partial x_l \partial w_{jk}} = a_j \big[& f'''(A_{2j}) f(A_{1k}) B_{ji} B_{jl} + f''(A_{2j}) f'(A_{1k}) B_{ji} (v_{ki} + v_{kl}) \\ & + f''(A_{2j}) f(A_{1k}) C_{jil} + f'(A_{2j}) f''(A_{1k}) v_{ki} v_{kl} \big]\end{aligned} \tag{6b}$$

$$\frac{\partial^3 N(x,\theta)}{\partial x_i \partial x_l \partial c_j} = a_j \left[f'''(A_{2j}) B_{ji} B_{jl} + f''(A_{2j}) C_{jil} \right] \tag{6c}$$

$$\begin{aligned}\frac{\partial^3 N(x,\theta)}{\partial x_i \partial x_l \partial v_{kn}} = & \left[f'''(A_{1k}) v_{kl} v_{ki} x_n + f''(A_{1k}) \left(v_{ki} \delta_{nl} + v_{kl} \delta_{ni} \right) \right] \sum_{j=1}^{M_2} a_j w_{jk} f'(A_{2j}) \\ & + \left[f''(A_{1k}) v_{ki} x_n + f'(A_{1k}) \delta_{ni} \right] \sum_{j=1}^{M_2} a_j w_{jk} f''(A_{2j}) B_{jl} \\ & + \left[f''(A_{1k}) v_{kl} x_n + f'(A_{1k}) \delta_{nl} \right] \sum_{j=1}^{M_2} a_j w_{jk} f''(A_{2j}) B_{ji} \\ & + f'(A_{1k}) x_n \sum_{j=1}^{M_2} a_j w_{jk} \left[f'''(A_{2j}) B_{jl} B_{ji} + f''(A_{2j}) C_{jil} \right]\end{aligned} \tag{6d}$$

$$\begin{aligned}\frac{\partial^3 N(x,\theta)}{\partial x_i \partial x_l \partial b_k} = & f'''(A_{1k}) v_{kl} v_{ki} \sum_{j=1}^{M_2} a_j w_{jk} f'(A_{2j}) \\ & + f''(A_{1k}) \sum_{j=1}^{M_2} a_j w_{jk} f''(A_{2j}) (v_{ki} B_{jl} + v_{kl} B_{ji}) \\ & + f'(A_{1k}) \sum_{j=1}^{M_2} a_j w_{jk} \left[f'''(A_{2j}) B_{jl} B_{ji} + f''(A_{2j}) C_{jil} \right]\end{aligned} \tag{6e}$$

Relating $(\boldsymbol{a}, \boldsymbol{w}, \boldsymbol{c}, \boldsymbol{v}, \boldsymbol{b})$ to $\boldsymbol{\theta}$

We would like to use, for implementation reasons mostly, a single index parameter, actually θ_m, to represent all the weights. Hence, a correspondence of the used notation $(a_j, w_{jk}, c_j, v_{kl}, b_k)$ to θ_m is necessary.

$$a_j = \theta_j \qquad j = 1, \cdots, M_2 \tag{7a}$$

$$w_{jk} = \theta_{M_2+M_1(j-1)+k} \qquad j = 1, \cdots, M_2, \qquad k = 1, \cdots, M_1 \tag{7b}$$

$$c_j = \theta_{M_2+M_1M_2+j} \qquad j = 1, \cdots, M_2 \tag{7c}$$

$$v_{kl} = \theta_{2M_2+M_1M_2+D(k-1)+l} \qquad k = 1, \cdots, M_1 \qquad l = 1, \cdots, D \tag{7d}$$

$$b_k = \theta_{2M_2+M_1M_2+M_1D+k} \qquad k = 1, \cdots, M_1 \tag{7e}$$

Note that the total number of parameters (weights) equals to:

$$N_T = M_1(1 + D + M_2) + 2M_2$$