U of U IS Deep Learning Study Group

Backpropagation Formula Sheet

Author: Brian Sheng

This is an excerpt from the longer set of notes on computational graphs and backpropagation. The formulas here outline the example case of a fully connnected neural net with one hidden layer, one input layer, one output layer, and non-linear activation functions applied to the hidden and output layers.

$\boldsymbol{x} =$ Inputs

$z_1 = \boldsymbol{W_1 x} + \boldsymbol{b_1} =$ Input values to "hidden" activation function

$\boldsymbol{a} = \boldsymbol{f}(z_1) =$ Hidden Activations

$z_2 = \boldsymbol{W_2 a} + \boldsymbol{b_2} =$ Inputs to "output" activation function

$\hat{y} = \boldsymbol{f}(z_2) =$Output Activations

$$\frac{\partial J}{\partial \boldsymbol{b_2}} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_2} \frac{\partial z_2}{\partial \boldsymbol{b_2}}$$

$$\frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_2} = \frac{\partial J}{\partial \hat{y}} \odot \frac{\partial \hat{y}}{\partial z_2} = \boldsymbol{\delta_2}$$

Where $\odot$ is **element-wise multiplication**, a.k.a the Hadamard Product: https://en.wikipedia.org/wiki/Hadamard_product_(matrices (https://en.wikipedia.org/wiki/Hadamard_product_(matrices)), and $\frac{\partial \hat{y}}{\partial z_2}$ is the derivative of the output activation function w.r.t its inputs.

$$\frac{\partial z_2}{\partial \boldsymbol{b_2}} = \frac{\partial}{\partial \boldsymbol{b_2}}(\boldsymbol{w_2 a} + \boldsymbol{b_2}) = \mathbf{1}$$

Where $\mathbf{1}$ is the ones vector. (i.e. $\mathbf{1} = \begin{bmatrix} 1 \\ \dots \\ 1 \end{bmatrix}$ for some arbitrary length) In the case of $\frac{\partial z_2}{\partial \boldsymbol{b_2}}$, the length of the vector is the same as the length of $\boldsymbol{b_2}$.

$$\frac{\partial J}{\partial \boldsymbol{b_2}} = \boldsymbol{\delta_2} \odot \mathbf{1} = \boldsymbol{\delta_2}$$

$$\frac{\partial J}{\partial \boldsymbol{w_2}} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_2} \frac{\partial z_2}{\partial \boldsymbol{w_2}} = \boldsymbol{\delta_2} \frac{\partial z_2}{\partial \boldsymbol{w_2}}$$

$$\frac{\partial z_2}{\partial \boldsymbol{w_2}} = \frac{\partial}{\partial \boldsymbol{w_2}}(\boldsymbol{w_2 a} + \boldsymbol{b_2}) = \boldsymbol{a}$$

$$\frac{\partial J}{\partial \boldsymbol{w_2}} = \boldsymbol{\delta_2} \otimes \boldsymbol{a} = \boldsymbol{\delta_2 a}^T$$

Where $\otimes$ is the **outer (or tensor) product**: https://en.wikipedia.org/wiki/Outer_product (https://en.wikipedia.org/wiki/Outer_product)

$$\frac{\partial J}{\partial \boldsymbol{b_1}} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_2} \frac{\partial z_2}{\partial \boldsymbol{a}} \frac{\partial \boldsymbol{a}}{\partial z_1} \frac{\partial z_1}{\partial \boldsymbol{b_1}} = \boldsymbol{\delta_2} \frac{\partial z_2}{\partial \boldsymbol{a}} \frac{\partial \boldsymbol{a}}{\partial z_1} \frac{\partial z_1}{\partial \boldsymbol{b_1}}$$

Where $\frac{\partial a}{\partial z_1}$ is the derivative of the hidden activation function w.r.t. its inputs.

$$\frac{\partial z_2}{\partial \boldsymbol{a}} = \frac{\partial}{\partial \boldsymbol{a}}(\boldsymbol{w_2 a} + \boldsymbol{b_2}) = \boldsymbol{w_2}$$

$$\frac{\partial z_1}{\partial \boldsymbol{b_1}} = \frac{\partial}{\partial \boldsymbol{b_1}}(\boldsymbol{w_1 x} + \boldsymbol{b_1}) = \mathbf{1}$$

$$\boldsymbol{\delta_2} \frac{\partial z_2}{\partial \boldsymbol{a}} = \langle \boldsymbol{w_2}, \boldsymbol{\delta_2} \rangle = \boldsymbol{w_2} \cdot \boldsymbol{\delta_2} = \boldsymbol{w_2^T \delta_2}$$

Where $\langle \boldsymbol{w_2}, \boldsymbol{\delta_2} \rangle$ is the **inner product** (in this case, called the **dot product**) of the matrices $\boldsymbol{w_2}$ and $\boldsymbol{\delta_2}$:
https://en.wikipedia.org/wiki/Inner_product_space (https://en.wikipedia.org/wiki/Inner_product_space)
https://en.wikipedia.org/wiki/Dot_product (https://en.wikipedia.org/wiki/Dot_product)

$$\frac{\partial J}{\partial \boldsymbol{b_1}} = \boldsymbol{\delta_2} \frac{\partial z_2}{\partial \boldsymbol{a}} \frac{\partial \boldsymbol{a}}{\partial z_1} \frac{\partial z_1}{\partial \boldsymbol{b_1}} = \langle \boldsymbol{w_2}, \boldsymbol{\delta_2} \rangle \odot \frac{\partial \boldsymbol{a}}{\partial z_1} \odot \mathbf{1} = \boldsymbol{\delta_1} \odot \mathbf{1} = \boldsymbol{\delta_1}$$

$$\frac{\partial J}{\partial \boldsymbol{w_1}} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_2} \frac{\partial z_2}{\partial \boldsymbol{a}} \frac{\partial \boldsymbol{a}}{\partial z_1} \frac{\partial z_1}{\partial \boldsymbol{w_1}} = \boldsymbol{\delta_1} \frac{\partial z_1}{\partial \boldsymbol{w_1}}$$

$$\frac{\partial z_1}{\partial \boldsymbol{w_1}} = \frac{\partial}{\partial \boldsymbol{w_1}}(\boldsymbol{w_1 x} + \boldsymbol{b_1}) = \boldsymbol{x}$$

$$\frac{\partial J}{\partial \boldsymbol{w_1}} = \boldsymbol{\delta_1} \otimes \boldsymbol{x} = \boldsymbol{\delta_1}\boldsymbol{x}^T$$