Basic Neural Network Math

Jeffrey Zhou

1 Introduction

This document goes over how I implemented the neural network and the math behind it. It shows the proofs for the forward and backward propagation equations.

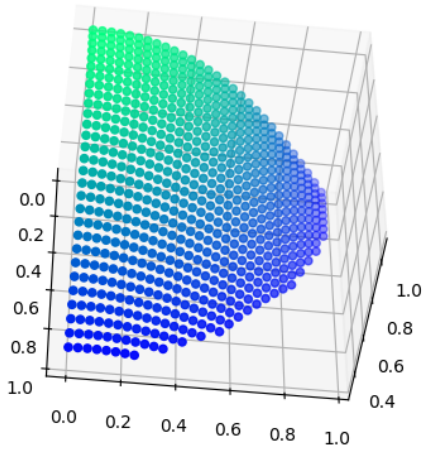

Figure 1: Approximation points on a sphere

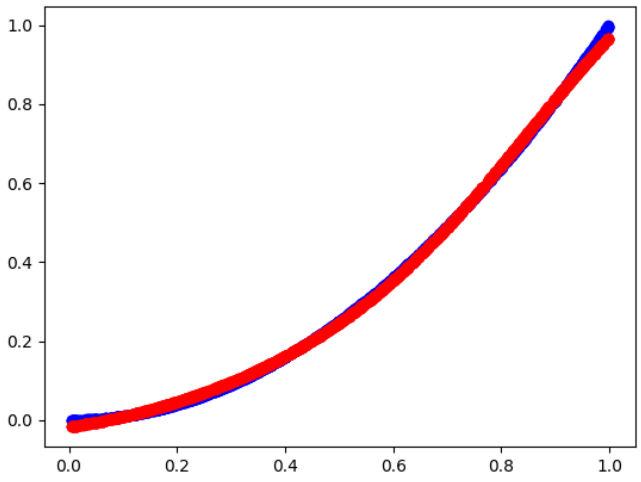

Figure 2: Approximating $f(x) = x^2$

1.1 Notation

There is some notation I use for some proofs:

Let a matrix X be defined as such

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

Then, $F(X)$ means the following:

$$F(X) = \begin{bmatrix} F(x_{11}) & F(x_{12}) & \cdots & F(x_{1n}) \\ F(x_{21}) & F(x_{22}) & \cdots & F(x_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ F(x_{m1}) & F(x_{m2}) & \cdots & F(x_{mn}) \end{bmatrix}$$

2 Forward Propagation

Let z_i be the value of ith output of a layer and F(x) be the activation function.

$$z_i = F(y)$$

where y =

$$(\sum w_{ai} x_i) + b_a$$

w_{ai} represents weight i of node a, and b_a represents the bias

In order to speed this up, matricies can be used. It uses Single Instruction Multiple Data (SIMD) to optimize the run time.

representing all the variables with matricies:

$$Z = \begin{bmatrix} z_1 & z_2 & z_3 & \cdots & z_m \end{bmatrix}^T$$

$$Y = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_m \end{bmatrix}^T$$

$$X = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & y_n \end{bmatrix}^T$$

Therefore, Z can be expressed like this:

$$Z = \begin{bmatrix} F(y_1) & F(y_2) & F(y_3) & \cdots & F(y_m) \end{bmatrix}^T = F(Y^T) \tag{1}$$

This can be implemented by vectorizing F(x) in numpy, which uses SIMD to apply F(x) to all elements of the array

Similarly, the weights can also be represented by a matrix W:

$$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} & \cdots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & w_{m3} & \cdots & w_{mn} \end{bmatrix}$$

where w_{ij} represent the weight for the jth input for the ith node in the current layer

$$Y = W \cdot X \tag{2}$$

(1) and (2) express the forward propagation, which gets fed as input to the next layer. X is (n,1) size, Y is (m,1) size

$$Y = F(W \cdot X) \tag{3}$$

This is the output of the current layer that gets fed to the next layer. (3) is the equation for forward propagation.

3 Back Propagation

The purpose of back propagation is to find the partial derivative of each weight and bias with respect to the error function in order perform gradient descent. This can be done by iterating backwards through the layers. At each layer, there are three things that need to be done. The partial derivative of E with respect to a matrix/vector is the

1. Calculate $\partial E/\partial W$
2. Calculate $\partial E/\partial B$
3. Calculate $\partial E/\partial X$

The first two steps are for updating the current layer, and the last step is for helping the previous layer update its data.

3.1 Calculate $\partial E/\partial W$

For a particular weight, say w_{11}

$$\frac{\partial E}{\partial w_{11}} = \frac{\partial Y_1}{\partial w_{11}} \times \frac{\partial F(Y_1)}{\partial Y_1} \times \frac{\partial E}{\partial F(Y_1)}$$

Evaluating the first component

$$\frac{\partial Y_1}{\partial w_{11}} = \frac{\partial(w_{11}x_1 + w_{12}x_2 + ... + b_1)}{\partial w_{11}} = x_1$$

Since the second part is just the activation function differentiated,

$$\frac{\partial E}{\partial w_{11}} = x_1 \times F'(Y) \times \frac{\partial E}{\partial Z_1}$$

Generalizing to all weights,

$$\frac{\partial E}{\partial W} = \begin{bmatrix} x_{11}F'(y_1)\frac{\partial E}{\partial Z_1} & x_{12}F'(y_1)\frac{\partial E}{\partial Z_1} & F'(y_1)\frac{\partial E}{\partial Z_1}x_{13} & \cdots & x_{1n}F'(y_1)\frac{\partial E}{\partial Z_1} \\ x_{21}F'(y_2)\frac{\partial E}{\partial Z_2} & x_{22}F'(y_2)\frac{\partial E}{\partial Z_2} & x_{23}F'(y_2)\frac{\partial E}{\partial Z_2} & \cdots & x_{2n}F'(y_2)\frac{\partial E}{\partial Z_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1}F'(y_m)\frac{\partial E}{\partial Z_m} & x_{m2}F'(y_m)\frac{\partial E}{\partial Z_m} & x_{m3}F'(y_m)\frac{\partial E}{\partial Z_m} & \cdots & x_{mn}F'(y_m)\frac{\partial E}{\partial Z_m} \end{bmatrix}$$

This can be represented as a multiplication, taking advantage of SIMD

$$\frac{\partial E}{\partial W} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} F'(y_1)\frac{\partial E}{\partial Z_1} & F'(y_2)\frac{\partial E}{\partial Z_2} & \cdots & F'(y_m)\frac{\partial E}{\partial Z_m} \end{bmatrix}$$

Using element wise multiplication to separate the right matrix, the simplified result is

$$\frac{\partial E}{\partial W} = X(F'(Y)) \odot \frac{\partial E}{\partial Z} \tag{4}$$

Where $F'(Y)$ represents applying the function to each element of Y.

3.2 Calculate $\partial E/\partial B$

For a particular bias, say b_1

$$\frac{\partial E}{\partial b_1} = \frac{\partial Y_1}{\partial b_1} \times \frac{\partial F(Y_1)}{\partial Y_1} \times \frac{\partial E}{\partial F(Y_1)}$$

To calculate the first partial:

$$\frac{\partial Y_1}{\partial b_1} = \frac{\partial (w_{11}x_1 + w_{12}x_2 + ... + b_1)}{\partial b_1} = 1$$

The expression then evaluates to:

$$\frac{\partial E}{\partial b_1} = \frac{\partial F(Y_1)}{\partial Y_1} \times \frac{\partial E}{\partial F(Y_1)}$$

Similar to the calculations for the weights, it is possible to generalize this to solve for B with matrix operations.

$$\frac{\partial E}{\partial B} = \begin{bmatrix} F'(y_1)\frac{\partial E}{\partial Z_1} \\ F'(y_2)\frac{\partial E}{\partial Z_2} \\ \vdots \\ F'(y_m)\frac{\partial E}{\partial Z_m} \end{bmatrix}$$

Like before, this evaluates to:

$$\frac{\partial E}{\partial B} = F'(Y) \odot \frac{\partial E}{\partial Z} \tag{5}$$

Where F'(Y) represents applying the function to each element of Y

3.3 Calculate $\partial E/\partial X$

For a particular input, say x_1

$$\frac{\partial E}{\partial x_1} = \frac{\partial Y_1}{\partial x_1} \times \frac{\partial F(Y_1)}{\partial Y_1} \times \frac{\partial E}{\partial F(Y_1)}$$

This is different from the weights example, because x_1 is used in many variables.

$$Y_1, Y_2 ... Y_m$$

all use x_1 as a part of their input

Therefore, x_1 needs to be calculated using chain rule. More specifically,

$$\frac{\partial E}{\partial x_1} = \frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial E}{\partial y_2}\frac{\partial y_2}{\partial x_1} + \cdots + \frac{\partial E}{\partial y_m}\frac{\partial y_m}{\partial x_1}$$

For y_1:

$$\frac{\partial y_1}{\partial x_1} = \frac{\partial(w_{11}x_1 + w_{12}x_2 + ... + b_1)}{\partial x_1} = w_{11}$$

Going back

$$\frac{\partial E}{\partial x_1} = \frac{\partial E}{\partial y_1}w_{11} + \frac{\partial E}{\partial y_2}w_{21} + \cdots + \frac{\partial E}{\partial y_m}w_{m1} \tag{6}$$

From the next layer, we are given the partial derivative of the error with respect to its inputs, which is the same as the partial derivative of the error with respect to this layer's outputs.

$$\frac{\partial E}{\partial Z} = \begin{bmatrix} \frac{\partial E}{\partial z_1} \\ \frac{\partial E}{\partial z_2} \\ \vdots \\ \frac{\partial E}{\partial z_m} \end{bmatrix}$$

Since the the relation between z_i and y_i is $z_i = F(y_i)$

$$\frac{\partial E}{\partial y_1} = \frac{\partial E}{\partial z_1}\frac{\partial z_1}{\partial y_1} = \frac{\partial E}{\partial z_1}F'(y_1)$$

Combining this with equation (6) gives:

$$\frac{\partial E}{\partial x_1} = \frac{\partial E}{\partial z_1}F'(y_1)w_{11} + \frac{\partial E}{\partial z_2}F'(y_2)w_{21} + \cdots + \frac{\partial E}{\partial y_m}F'(y_m)w_{m1} \tag{7}$$

This result can be generalized to X. Expressed in matrix form

$$\frac{\partial E}{\partial X} = \begin{bmatrix} w_{11} & w_{21} & w_{31} & \cdots & w_{m1} \\ w_{12} & w_{22} & w_{32} & \cdots & w_{m2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{1n} & w_{2n} & w_{3n} & \cdots & w_{mn} \end{bmatrix} \begin{bmatrix} \frac{\partial E}{\partial z_1} \\ \frac{\partial E}{\partial z_2} \\ \vdots \\ \frac{\partial E}{\partial z_m} \end{bmatrix} \odot \begin{bmatrix} F'(y_1) \\ F'(y_2) \\ \vdots \\ F'(y_m) \end{bmatrix}$$

Finally, this can be re-written in terms already defined.

$$\frac{\partial E}{\partial X} = W^T(Z \odot F'(Y)) \tag{8}$$