Mathematics for Neural Networks

styxofsyntax

July 9, 2024

1 Introduction

This document explains the mathematical concepts used in the implementation of a neural network from scratch.

2 Equations

2.1 Calculating net input

$m =$ number of examples
$n =$ number of inputs
$k =$ number of neurons

$$\boldsymbol{Z} = \boldsymbol{X} \cdot \boldsymbol{W} + \boldsymbol{j} \otimes \boldsymbol{b}$$

- $\boldsymbol{X}$ is the input matrix with dimensions $m \times n$.
- $\boldsymbol{W}$ is the weight matrix with dimensions $n \times k$
- $\boldsymbol{j}$ is all ones with dimensions $m \times 1$
- $\boldsymbol{b}$ is the bias matrix with dimensions $1 \times k$

The result $\boldsymbol{Z}$ will have dimensions $m \times k$.

$$\boldsymbol{X} = \begin{bmatrix} x_1^1 & x_2^1 & x_3^1 & \cdots & x_n^1 \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ x_1^3 & x_2^3 & x_3^3 & \cdots & x_n^3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^m & x_2^m & x_3^m & \cdots & x_n^m \end{bmatrix}_{m \times n} \qquad \boldsymbol{W} = \begin{bmatrix} w_1^1 & w_1^2 & w_1^3 & \cdots & w_1^k \\ w_2^1 & w_2^2 & w_2^3 & \cdots & w_2^k \\ w_3^1 & w_3^2 & w_3^3 & \cdots & w_3^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_n^1 & w_n^2 & w_n^3 & \cdots & w_n^k \end{bmatrix}_{n \times k}$$

$$\boldsymbol{J} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{m \times 1} \qquad \boldsymbol{b} = \begin{bmatrix} b_1 & b_2 & b_3 & \cdots & b_k \end{bmatrix}_{1 \times k}$$

Note: The kronecker product is not required to be explicitly implemented in code, as B will be added to all rows of $\boldsymbol{W}$ regardless.

2.2 Gradients

- **Gradient of the Error with respect to the activation of the last layer**

  The error functions is given by:

$$E = \frac{1}{m} \sum_{i=0}^{m} \left(\boldsymbol{a}_i^{[L]} - \boldsymbol{y}_i \right)^2$$

  Where $\boldsymbol{a_i^{[L]}}$ is the activation of the last layer and $\boldsymbol{y}_i$ is the true output, over one example.

  The gradient of the error with respect to the activations of the last layer is:

$$\begin{aligned} \frac{\partial E}{\partial \boldsymbol{a}_j^{[L]}} &= \frac{1}{m} \cdot 2 \sum_{i=0}^{m} \left(\boldsymbol{a}_i^{[L]} - \boldsymbol{y}_i \right) \cdot 1 \\ &= \frac{2}{m} \left(\boldsymbol{a}_j^{[L]} - \boldsymbol{y}_j \right) \end{aligned}$$

  Where $\boldsymbol{a_j^{[L]}}$ is one particular activation of the last layer over all examples.

  Matrix notation of gradient of the error with respect to activations of the last layer is:

$$\frac{\partial E}{\partial \boldsymbol{A}^{[L]}} = \frac{2}{m} \left(\boldsymbol{A}^{[L]} - \boldsymbol{Y} \right) \qquad \text{(dimensions: } m \times k\text{)}$$

  Where $\boldsymbol{A}^{[L]}$ is the activation of the last layer and $\boldsymbol{Y}$ are true outputs, over all examples.

- **Gradient of the activations of the last layer with respect to the net input of the last layer**

$$\frac{\partial \boldsymbol{A}^{[L]}}{\partial \boldsymbol{Z}^{[L]}} = h'(\boldsymbol{Z}^{[L]}) \qquad \text{(dimensions: } m \times k\text{)}$$

  where $\boldsymbol{Z}^{[L]}$ are the net inputs over all examples.

- **Gradient of the net input of the last layer with respect to the activations of the previous layer**

  Given that:

$$\boldsymbol{Z}^{[L]} = \boldsymbol{A}^{[L-1]} \cdot \boldsymbol{W}^{[L]} + \boldsymbol{B}^{[L]}$$

  The partial derivative of $\boldsymbol{Z}^{[L]}$ with respect to $\boldsymbol{A}^{[L-1]}$ is:

$$\frac{\partial \boldsymbol{Z}^{[L]}}{\partial \boldsymbol{A}^{[L-1]}} = \boldsymbol{W}^{[L]} \qquad \text{(dimensions: } n \times k\text{)}$$

- **Gradient of the net input of the last layer with respect to the weights**
  The partial derivative of $\boldsymbol{Z}^{[L]}$ with respect to $\boldsymbol{W}^{[L]}$ is:

$$\frac{\partial \boldsymbol{Z}^{[L]}}{\partial \boldsymbol{W}^{[L]}} = \boldsymbol{A}^{[L-1]} \qquad \text{(dimensions: } m \times n\text{)}$$

  where $\boldsymbol{A}^{[L-1]}$ are the activations of previous layer over all examples.

- **Gradient of the net input of the last layer with respect to the biases**
  The partial derivative of $\boldsymbol{Z}^{[L]}$ with respect to $\boldsymbol{b}^{[L]}$ is:

$$\frac{\partial \boldsymbol{Z}^{[L]}}{\partial \boldsymbol{b}^{[L]}} = 1$$

2.3 Combining Gradients

Combining these gradients using the chain rule:

To generalize the equations over all layers, $\boldsymbol{\partial}(l)$ can be used in place of $\frac{\partial E}{\partial \boldsymbol{A}^{[l]}}$ where l indicates l-th Layer.

The gradient of the last layer $\boldsymbol{\partial}(L)$ is given as,

$$\boldsymbol{\partial}(L) = \frac{2}{m}\left(\boldsymbol{A}^{[l]} - \boldsymbol{Y}\right)$$

- **Gradient with respect to the previous layer's activations**

$$\begin{aligned}\boldsymbol{\partial}(l-1) &= \left(\frac{\partial \boldsymbol{A}^{[l]}}{\partial \boldsymbol{Z}^{[l]}} \times \boldsymbol{\partial}(l)\right) \cdot \left(\frac{\partial \boldsymbol{Z}^{[l]}}{\partial \boldsymbol{A}^{[l-1]}}\right)^T \\ &= \left[h'(\boldsymbol{Z}^{[l]}) \times \boldsymbol{\partial}(l)\right] \cdot (\boldsymbol{W}^{[l]})^T \qquad \text{(dimensions: } m \times n\text{)}\end{aligned}$$

- **Gradients with respect to the weights**

$$\begin{aligned}\frac{\partial E}{\partial w^{[l]}} &= \left(\frac{\partial \boldsymbol{Z}^{[l]}}{\partial w^{[l]}}\right)^T \cdot \left(\frac{\partial \boldsymbol{A}^{[l]}}{\partial \boldsymbol{Z}^{[l]}} \times \boldsymbol{\partial}(l)\right) \\ &= (\boldsymbol{A}^{[l-1]})^T \cdot \left[h'(\boldsymbol{Z}^{[l]}) \times \boldsymbol{\partial}(l)\right] \qquad \text{(dimensions: } n \times k\text{)}\end{aligned}$$

- **Gradients with respect to the biases**

$$\frac{\partial E}{\partial \boldsymbol{B}^{[l]}} = \frac{\partial \boldsymbol{Z}^{[l]}}{\partial \boldsymbol{b}^{[l]}} \cdot \left(\frac{\partial \boldsymbol{A}^{[l]}}{\partial \boldsymbol{Z}^{[l]}} \times \boldsymbol{\partial}(l)\right)$$

since $\frac{\partial \boldsymbol{Z}^{[l]}}{\partial \boldsymbol{b}^{[l]}} = 1$, we have

$$\begin{aligned}&= \frac{\partial \boldsymbol{A}^{[l]}}{\partial \boldsymbol{Z}^{[l]}} \times \boldsymbol{\partial}(l) \\ &= h'(\boldsymbol{Z}^{[l]}) \times \boldsymbol{\partial}(l) \qquad \text{(dimensions: } m \times k\text{)}\end{aligned}$$

2.4 Updating parameters

- **Updating Weights**

$$\boldsymbol{W}'^{[l]} = \boldsymbol{W}^{[l]} - \alpha \frac{\partial E}{\partial \boldsymbol{W}^{[l]}}$$

- **Updating Biases**

  The partial derivatives over all examples are added,

$$\boldsymbol{b}'^{[l]} = \boldsymbol{b}^{[l]} - \alpha \sum_{i=0}^{m} \frac{\partial E}{\partial \boldsymbol{B}_i^{[l]}}$$

  Here α is the learning rate.