Mathematics for Neural Network implementation

Martin Lazo

April 2022

1 Notation

$\odot$	Hadamard (element-wise) product
$\mathbf{v}$	Vector, vectors are denoted in lower case and bold
$\mathbf{M}$	Matrix, matrices are denoted in upper case and bold
$\mathbf{v}_i$	ith element of a vector $\mathbf{v}$
$\mathbf{M}_{i,j}$	Element in the ith row and jth column of a matrix $\mathbf{M}$
$\mathbf{M}_{i,}$	ith row of a matrix $\mathbf{M}$
$\mathbf{M}_{,i}$	ith column of a matrix $\mathbf{M}$
$\mathbf{o}_i$	Vector which preserves only the ith element of a vector

$$\mathbf{o}_3\mathbf{v} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = c$$

$\mathbf{O}_i$	Square matrix which preserves only the ith element of a vector

$$\mathbf{O}_3\mathbf{v} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix}$$

Hidden layers are indexed from 0 to n with n being the output layer and 0 being the input layer.

$\mathbf{X}$	Features
$\mathbf{Y}$	Labels
$\mathbf{x}$	Features data point
$\mathbf{y}$	Labels data point
$\mathbf{b}^m$	Bias vector of the mth layer
$\mathbf{W}^m$	Weights matrix of the mth layer
ϕ^m	Activation function of the mth layer, applied element-wise
$\phi^{m\prime}$	Derivative of the activation function of the mth layer

2 Overview

Neural network output for input $\mathbf{x}$:

$$\hat{\mathbf{y}} = \phi^n(\mathbf{W}^n\phi^{n-1}(..\phi^0(\mathbf{W}^0\mathbf{x} + \mathbf{b}^0)..) + \mathbf{b}^n)$$

Partial derivatives w.r.t. weights:

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{W}^m_{i,j}} = \phi^{n\prime}(..) \odot .. \odot (\mathbf{W}^{m+2}(\phi^{m+1\prime}(..) \odot (\mathbf{W}^{m+1}(\mathbf{O}_i\phi^{m\prime}(..)\mathbf{o}_j\phi^{m-1}(..)))))$$

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{W}^0_{i,j}} = \phi^{n\prime}(..) \odot .. \odot (\mathbf{W}^2(\phi^{1\prime}(..) \odot (\mathbf{W}^1(\mathbf{O}_i\phi^{0\prime}(..)\mathbf{x}_j))))$$

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{W}^n_{i,j}} = \mathbf{O}_i\phi^{n\prime}(..)\mathbf{o}_j\phi^{n-1}(..)$$

Partial derivatives w.r.t. biases:

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{b}^m_i} = \phi^{n\prime}(..) \odot .. \odot (\mathbf{W}^{m+2}(\phi^{m+1\prime}(..) \odot (\mathbf{W}^{m+1}(\mathbf{O}_i\phi^{m\prime}(..)))))$$

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{b}^0_i} = \phi^{n\prime}(..) \odot .. \odot (\mathbf{W}^2(\phi^{1\prime}(..) \odot (\mathbf{W}^1(\mathbf{O}_i\phi^{0\prime}(..)))))$$

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{b}^n_i} = \mathbf{O}_i\phi^{n\prime}(..)$$