Binary Classifier Forward-Pass Optimization

Samuel Lerman

May 20, 2021

Goal: Learn a binary classifier $\mathbb{P}(Y|\mathbf{x})$ for input $\mathbf{x} \in \mathbb{R}^d$ and binary random variable $Y \in \{0, 1\}$. Without loss of generality, assume $\mathbf{x}$, y_{label} are input-label pairs sampled from a dataset.

Represent $\mathbb{P}(\cdot)$ with:

$$\mathbb{P}(Y = 1|\mathbf{x}) = y_{pred} = \text{sigmoid}(\mathcal{F}_{\mathbf{w}}(\mathbf{x})), \tag{1}$$

where $\mathcal{F}_{\mathbf{w}}$ is a ReLU-activated multi-layer perceptron (MLP) parameterized by weights $\mathbf{w}$ and no biases, and sigmoid is the sigmoid function:

$$\text{relu}(a) = \begin{cases} 0, & \text{if } a \leq 0 \\ a, & \text{otherwise} \end{cases} \tag{2}$$

$$\text{sigmoid}(a) = \frac{1}{1 + e^{-a}}. \tag{3}$$

Let's formally define our L-layer MLP $\mathcal{F}_{\mathbf{w}}(\mathbf{x})$ at each layer $\ell \leq L$:

$$\mathcal{F}_{\mathbf{w}}(\mathbf{x}) = \boldsymbol{\alpha}_j^L, \tag{4}$$

where

$$\begin{aligned} \boldsymbol{\alpha}_j^0 &= \mathbf{x}_j \\ \boldsymbol{\alpha}_j^\ell &= \sigma_\ell(\mathbf{z}_j^\ell) \end{aligned} \tag{5}$$

and

$$\mathbf{z}_j^\ell = \sum_i \mathbf{w}_{ij}^\ell \boldsymbol{\alpha}_i^{\ell-1}, \tag{6}$$

with weights $\mathbf{w}$ and activations σ:

$$\sigma_\ell = \begin{cases} \text{I}, & \text{if } \ell = L \\ \text{relu}, & \text{otherwise} \end{cases}. \tag{7}$$

I is the identity function.

To simplify notation, indices i, j are reused to symbolize the variable number of dimensions (neurons) per layer.

The goal is thus to learn values for parameters $\mathbf{w}$ of MLP $\mathcal{F}_{\mathbf{w}}$ that minimize the error $\mathrm{E}_{\mathbf{w}}(\cdot)$ (say, cross entropy) between label y_{label} and prediction y_{pred}.

Furthermore, we will say a neuron *activated* if and only if its corresponding $\boldsymbol{\alpha}_j^\ell \neq 0$. The hypothetical 0th neural layer of the input $\mathbf{x}$ is always considered activated.

Definition 0.1 (Activated)**.** A "neuron" $\mathbf{n}_{\ell,j}$ is said to be "activated" if and only if $\boldsymbol{\alpha}_j^\ell \neq 0$. $\mathbf{n}_{0,j}$ is always considered activated.

> In biological terms, $\mathbf{z}_j^\ell$ is analogous to the membrane potential in a leaky-integrate-and-fire neuron, with action potential 0 due to the thresholding effect of ReLU, and instantaneous leaking of the neuron's full voltage after every forward pass.
>
> MLPs do not store persistent internal states like recurrent or spiking neural networks, so the full voltage is said to leak after every time step.
>
> Weights may be loosely thought of as dendritic receptors.
>
> Deviating from the biological analogy, our neural outputs $\boldsymbol{\alpha}_i^{\ell-1}$ are not consistent neurotransmitters with fairly reliable excitatory or inhibitory post-synaptic effects, but rather input-adaptive continuous values in $\mathbb{R}^+$ with mixed inhibitory or excitatory effects depending on $\mathbf{w}_{ij}^\ell$.

We will follow suit with the traditional paradigm of mini-batch training (*e.g.*, SGD, ADAM), except we will override the gradients for each weight, $\Delta\mathbf{w}_{ij}^\ell = \frac{\partial \mathrm{E}_{\mathbf{w}}(y_{label}, y_{pred})}{\partial \mathbf{w}_{ij}^\ell}$, according to the following rules:

1. If neuron $\mathbf{n}_{\ell,j}$ activated and neuron $\mathbf{n}_{(\ell-1),i}$ activated, then $\Delta\mathbf{w}_{ij}^\ell = \lambda(1 - 2y_{label})|\mathbf{w}_{ij}|$.
2. If neuron $\mathbf{n}_{\ell,j}$ did not activate and neuron $\mathbf{n}_{(\ell-1),i}$ activated, then $\Delta\mathbf{w}_{ij}^\ell = \lambda(2y_{label} - 1)|\mathbf{w}_{ij}|$.
3. If neuron $\mathbf{n}_{(\ell-1),i}$ did not activate, then $\Delta\mathbf{w}_{ij}^\ell = 0$.

To account for negative outputs $\mathbf{z}_j^L$ in our update, we flip the final layer's y_{label} when $\mathbf{z}_j^L < 0$:

$$y_{label} \leftarrow 1 - y_{label}, \text{ if } \ell = L \text{ and } \mathbf{z}_j^L < 0. \tag{8}$$

$\lambda \in (0, 1]$ is a scaling factor analogous to the learning rate.

λ can be substituted with a more adaptive λ' as follows:

$$\lambda' = \frac{\lambda}{(1 - y_{label})\lambda + 1} \tag{9}$$

λ can also adapt w.r.t. error (assuming $0 \leq \mathrm{E}(\cdot) \leq \lambda^{-1}$):

$$\lambda \leftarrow \lambda \mathrm{E}_{\mathbf{w}}(y_{label}, y_{pred}). \tag{10}$$

Rule 2 can also be more selectively applied:

2.1. $\Delta\mathbf{w}_{ij}^\ell = \lambda(2y_{label} - 1)\mathrm{relu}(\mathbf{w}_{ij})$.

Or not applied at all:

2.2. $\Delta \mathbf{w}_{ij}^{\ell} = 0$.

This scheme is akin to Hebbian learning in biologically plausible neural models.

To avoid exploding activations and rounding errors, we substitute $\sigma(\cdot)$ as follows:

$$\sigma_\ell(a) = \begin{cases} a - 0.001, & \text{if } \ell = L \\ \text{clip}_{10}(\text{relu}(a)), & \text{otherwise} \end{cases}, \tag{11}$$

where

$$\text{clip}_{10}(a) = \begin{cases} a, & \text{if } a < 10 \\ 10, & \text{otherwise} \end{cases}. \tag{12}$$

To compute a measure analogous to the natural gradient, we further adapt λ as follows:

$$\lambda' = \frac{\lambda}{\mathbf{z}_j^L}. \tag{13}$$

Here, λ' controls the size of $\mathcal{F}$'s trust region τ to be approximately $(1 - \lambda)^L \leq \tau \leq (1 + \lambda)^L$.

We also permit weights to change signs according to a minimal threshold $\eta \in \mathbb{R}^+$:

$$\mathbf{w}_{ij} \leftarrow -\text{sgn}(\mathbf{w}_{ij})\eta, \text{ if } |\mathbf{w}_{ij}| < \eta, \tag{14}$$

where sgn is the sign function.

> This is a bit like the synthesis of an excitatory neuron in place of an inhibitory neuron, or vice versa, where $n_{\ell-1,i}$ was either inhibitory or excitatory on $n_{\ell,j}$, but underwent apoptosis, followed by the neurogenesis of a new $n_{\ell-1,i}$ that synthesizes a neurotransmitter with the opposite effect on $n_{\ell,j}$ (*e.g.*, GABA to glutamate).
>
> However, the excitatory/inhibitory effect of $n_{\ell-1,i}$ remains the same on other neurons, so perhaps it would be more analogous to say instead that $n_{\ell,j}$'s dendritic receptors begin to modulate the signal from $n_{\ell-1,i}$ in the opposite way.
>
> Or, as a simpler alternative conceptualization, $n_{\ell-1,i}$ can be thought of as synthesizing multiple kinds of neurotransmitters (both inhibitory and excitatory) and transmitting the opposite kind to $n_{\ell,j}$.