EECS 189 Introduction to Machine Learning
Fall 2020

DIS7

This discussion was released **Friday, October 16**.

This discussion serves as an introduction to neural networks. You will through a simple example of backpropagation and build intuition for the ReLU nonlinearity and gradient descent training process through visualizations in the Jupyter **[notebook]**. This notebook takes a long time to do the initial network training! You should begin the training process then return to the theory part of this discussion while you wait.

1 ReLU SGD Visualization

Work through the **[notebook]** to explore how a simple network with ReLU non-linearities adapts to model a function with SGD updates. Training the networks takes 5-10 minutes depending on whether you run locally or on datahub and the server load, so you should start the training process (run through the train all layers cell) then return to the theory part of the discussion while training occurs.

As you walk through the notebook, pay attention to how the slopes and elbows of the ReLU functions change during training and how they impact the shape of the final learned function.

2 Backpropagation

In this problem, we will explore the chain rule of differentiation, and provide some algorithmic motivation for the backpropagation algorithm. Those of you who have taken CS170 may recognize a particular style of algorithmic thinking that underlies the computation of gradients.

Let us begin by working with simple functions of two variables.

(a) Define the functions $f(x) = x^2$ and $g(x) = x$, and $h(x_1, x_2) = x_1^2 + x_2^2$. Compute the derivative of $\ell(x) = h(f(x), g(x))$ with respect to x.

(b) Chain rule of multiple variables: Assume that you have a function given by $f(x_1, x_2, \ldots, x_n)$, and that $g_i(w) = x_i$ for a scalar variable w. How would you compute $\frac{\mathrm{d}}{\mathrm{d}w} f(g_1(w), g_2(w), \ldots, g_n(w))$? What is its computation graph?

(c) Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n \in \mathbb{R}^d$, and we refer to these variables together as $\mathbf{W} \in \mathbb{R}^{n \times d}$. We also have $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Consider the function

$$f(\mathbf{W}, \mathbf{x}, y) = \left(y - \sum_{i=1}^{n} \phi(\mathbf{w}_i^\top \mathbf{x} + \mathbf{b}_i) \right)^2 .$$

Write out the function computation graph (also sometimes referred to as a pictorial representation of the network). This is a directed graph of decomposed function computations, with the function at one end (which we will call the sink), and the variables $\mathbf{W}, \mathbf{x}, y$ at the other end (which we will call the sources).

(d) Define the cost function

$$\ell(\mathbf{x}) = \frac{1}{2}\|\mathbf{W}^{(2)}\Phi\left(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}\right) - \mathbf{y}\|_2^2, \tag{1}$$

where $\mathbf{W}^{(1)} \in \mathbb{R}^{d\times d}$, $\mathbf{W}^{(2)} \in \mathbb{R}^{d\times d}$, and $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ is some nonlinear transformation. **Compute the partial derivatives $\frac{\partial \ell}{\partial \mathbf{x}}$, $\frac{\partial \ell}{\partial \mathbf{W}^{(1)}}$, $\frac{\partial \ell}{\partial \mathbf{W}^{(2)}}$, and $\frac{\partial \ell}{\partial \mathbf{b}}$. Track the dimensions of each element of the derivatives as you go and make sure they make sense.**

In order to keep track of the partial derivatives, it is incredibly helpful to define intermediate variables. We suggest using the ones below, but you are free to define your own.

$$\begin{aligned}
\mathbf{x}^{(1)} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b} \\
\mathbf{x}^{(2)} &= \Phi(\mathbf{x}^{(1)}) \\
\mathbf{x}^{(3)} &= \mathbf{W}^{(2)}\mathbf{x}^{(2)} \\
\mathbf{x}^{(4)} &= \mathbf{x}^{(3)} - \mathbf{y} \\
\ell &= \frac{1}{2}\|\mathbf{x}^{(4)}\|_2^2.
\end{aligned}$$

Remember that the superscripts represent the index rather than the power operator.

(e) Compare the computation complexity of computing the $\frac{\partial \ell}{\partial \mathbf{W}}$ for Equation (1) using the analytic derivatives and numerical derivatives. Remember that if we want to compute the derivative of some function $f(x)$ at $x = 3$, we can use

$$\frac{d}{dx}f(x)|_{x=3} = \lim_{\epsilon\to 0}\frac{f(3+\epsilon) - f(3)}{\epsilon}$$

(f) What is the intuitive interpretation of taking a partial derivative of the output with respect to a particular node of this function graph?

(g) Discuss how gradient descent would work on the function $f(\mathbf{W}, \mathbf{x}, y)$ if we use backpropagation as a subroutine to compute gradients with respect to the parameters $\mathbf{W}$ (with $\mathbf{x}$ and y given).

Contributors:

- Anant Sahai
- Ashwin Pananjady
- Josh Sanz
- Yichao Zhou