CS 189 Introduction to Machine Learning
Summer 2018 DIS7

1 Backpropagation

In this discussion, 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

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

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

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

where $\mathbf{W}^{(1)} \in \mathbb{R}^{d \times d}$, $\mathbf{W}^{(2)} \in \mathbb{R}^{d \times d}$, and $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathbb{R}^d$ 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 $\frac{\partial \ell}{\partial \mathbf{b}}$.

(e) Compare the computation complexity of computing the $\frac{\partial \ell}{\partial \mathbf{W}}$ for Equation (1) using the analytic derivatives and numerical derivatives.

(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).

2 Derivatives of simple functions

Compute the derivatives of the following simple functions used as non-linearities in neural networks.

(a) $\sigma(x) = \frac{1}{1+e^{-x}}$

(b) $\mathsf{ReLu}(x) = \max(x, 0)$

(c) $\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

(d) Leaky ReLu: $f(x) = \max(x, -0.1x)$

