CS 189 Introduction to Machine Learning
Fall 2018 DIS8

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

2 Backpropagation

In this discussion, we will explore the chain rule of differentiation, and provide some algorithmic motivation for the backpropagation algorithm.

(a) 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?

(b) 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

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

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, and the variables $\mathbf{W}, \mathbf{b}, \mathbf{x}, y$ at the other end, where $\mathbf{b} = [b_1, \cdots, b_n]$.

(c) Suppose $\boldsymbol{\phi}(x) = \sigma(x)$ (from problem 1a). Compute the partial derivatives $\frac{\partial f}{\partial \mathbf{w}_i}$ and $\frac{\partial f}{\partial b_i}$. Use the computational graph you drew in the previous part to guide you.

(d) Write down a single gradient descent update for $\mathbf{w}_i^{(t+1)}$ and $b_i^{(t+1)}$, assuming step size η. You answer should be in terms of $\mathbf{w}_i^{(t)}$, $b_i^{(t)}$, $\mathbf{x}$, and y.

(e) (optional) 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 $\mathbf{\Phi} : \mathbb{R}^d \to \mathbb{R}^d$ 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 $\frac{\partial \ell}{\partial \mathbf{b}}$.

(f) (optional) Suppose $\mathbf{\Phi}$ is the identity map. Write down a single gradient descent update for $\mathbf{W}^{(1)}_{t+1}$ and $\mathbf{W}^{(2)}_{t+1}$ assuming step size η. Your answer should be in terms of $\mathbf{W}^{(1)}_t, \mathbf{W}^{(2)}_t, \mathbf{b}_t$ and $\mathbf{x}, \mathbf{y}$.