CS 189/289A Introduction to Machine Learning
Fall 2023 Jennifer Listgarten, Jitendra Malik

DIS4

1 Initialization of Weights for Backpropagation

Assume a fully-connected 1-hidden-layer network. Denote the dimensionalities of the input, hidden, and output layers as $d^{(0)}$, $d^{(1)}$, and $d^{(2)}$. That is, the input (which we will denote with a superscript (0)) is a vector of the form $x_1^{(0)}, \ldots, x_{d^{(0)}}^{(0)}$. Let g denote the activation function applied at each layer. We will let $S_j^{(l)} = \sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)}$ be the weighted input to node j in layer l, and let $\delta_j^{(l)} = \frac{\partial \ell}{\partial S_j^{(l)}}$ be the partial derivative of the final loss ℓ with respect to $S_j^{(l)}$.

Recall that backpropagation is an efficient method to compute the gradient of the loss function so we can use it for gradient descent. Gradient descent requires the parameters to be initialized to some value(s).

(a) To better orient yourself with the operations described in this 1-hidden-layer network, draw out a diagram of the layers, including weights, activation functions, and the outputs of each operation during the forward pass. In addition, identify where the partial derivatives $\delta_j^{(l)}$ are calculated during backpropagation.

(b) Imagine that we initialize every element of each weight $w^{(l)}$ to be the same constant scalar value a. After performing the forward pass, what is the value of $x_j^{(1)}$ in terms of the elements of $\{x_i^{(0)} : i = 1, \ldots, d^{(0)}\}$? What is the relationship between each $x_j^{(1)}$?

(c) Following from the previous part, after the backward pass of backpropagation, compute the values for each member of the set $\{\delta_i^{(1)} : i = 1, \ldots, d^{(1)}\}$, assuming we have calculated $\{\delta_j^{(2)} : j = 1, \ldots, d^{(2)}\}$. What is the relationship between each $\delta_i^{(1)}$?

(d) For a reasonable loss function, are all of the $\delta_i^{(2)}$ equal to each other? Why or why not?

(e) In the previous part, you showed that $w_{ij}^{(2)}$ is different for each j, but for a fixed j, it is the same for each i. In fact, no matter how many subsequent iterations of gradient descent you take, this property will continue to be true. Show why this is the case.

(f) To solve this problem, we randomly initialize our weights. This is called symmetry breaking. Note that for logistic regression, we don't run into this issue; that is, gradient descent will find the optimal values of the weights even if we initialize them at 0. Explain why this discrepancy exists between our 1-hidden-layer neural network and logistic regression.

2 Backpropagation Practice

(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. What is its computation graph? Sketch out a diagram of what the computation graph would look like. How would you compute

$$\frac{\mathrm{d}}{\mathrm{d}w} f(g_1(w), g_2(w), \ldots, g_n(w))$$

?

(b) Let $w_1, w_2, \ldots, w_n \in \mathbb{R}^d$, and we refer to these weights together as $W \in \mathbb{R}^{n \times d}$. We also have $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Consider the function

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

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 output of the function at one end, and the input to the function, x at the other end, where b are the bias terms corresponding to each weight vector, i.e. $b = [b_1, \cdots, b_n]$.

(c) Suppose $\phi(x)$ (from the previous part) is the sigmoid function, $\sigma(x)$. Compute the partial derivatives $\frac{\partial f}{\partial w_i}$ 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 $w_i^{(t+1)}$ and $b_i^{(t+1)}$, assuming step size η. You answer should be in terms of $w_i^{(t)}$, $b_i^{(t)}$, x, and y.

(e) Define the cost function

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

where $W^{(1)} \in \mathbb{R}^{d \times d}$, $W^{(2)} \in \mathbb{R}^{d \times d}$, and $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ is some nonlinear transformation. Compute the partial derivatives $\frac{\partial \ell}{\partial x}, \frac{\partial \ell}{\partial W^{(1)}}, \frac{\partial \ell}{\partial W^{(2)}}$, and $\frac{\partial \ell}{\partial b}$.

(f) Suppose Φ is the identity map. Write down a single gradient descent update for $W_{t+1}^{(1)}$ and $W_{t+1}^{(2)}$ assuming step size η. Your answer should be in terms of $W_t^{(1)}, W_t^{(2)}, b_t$ and x, y.