CS189: Introduction to Machine Learning

Homework 4

Problem 1 - Derivations (30 points)

We will use the following notation:

- L, the number of hidden and output layers ($L = 2$ for this homework)
- $d^{(l)}$, the number of units in layer l excluding the bias unit
- $x_k^{(l)}$, the value of the kth unit of layer l ($x_0^{(l)} = 1$ for $0 \leq l \leq L-1$)
- $w_{ij}^{(l)}$, the weight connecting the ith unit of layer $l-1$ to the jth unit of layer l
- $s_j^{(l)} = \sum_{i=0}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)}$
- g, the non-linear activation function
- η, the learning rate
- $\mathbf{y}$ is the label reformatted into $\mathbb{R}^{10}$

The gradient updates should be as follows.

$$w_{ij}^{(l)} = w_{ij}^{(l)} - \eta \frac{\partial J}{\partial w_{ij}^{(l)}}$$

To find the partial derivatives, we utilize the following chain rule.

$$\frac{\partial J}{\partial w_{ij}^{(l)}} = \frac{\partial J}{\partial s_j^{(l)}} \frac{\partial s_j^{(l)}}{\partial w_{ij}^{(l)}}$$

The second term is simply $x_i^{(l-1)}$. We denote the first term as $\delta_j^{(l)}$. We compute the $x_j^{(l)}$ for all l and j as follows.

1. Initialize $x_i^{(0)} = x_i$ and $x_0^{(0)} = 1$.
2. For $1 \leq l \leq L$, $x_j^{(l)} = g(\sum_{i=0}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)})$.

For this homework, g here should be the ReLU for the hidden layers and the softmax function for the output layer ($l = L$). We compute the $\delta_j^{(l)}$ for all l and j as follows.

1. Initialize $\delta_j^{(L)} = \frac{\partial J}{\partial x_j^{(L)}} \frac{\partial x_j^{(L)}}{\partial s_j^{(L)}}$.

2. For $L - 1 \geq l \geq 1$, $\delta_i^{(l-1)} = g'(s_i^{(l-1)}) \sum_{j=1}^{d^{(l)}} w_{ij}^{(l)} \delta_j^{(l)}$.

When using cross-entropy error ($J = -\sum_{j=1}^{n_{out}} y_j \ln x_j^{(L)}$):

$$\delta_j^{(L)} = \sum_{i=1}^{10} \frac{\partial J}{\partial x_i^{(L)}} \frac{\partial x_i^{(L)}}{\partial s_j^{(L)}} = -\sum_{i=1}^{10} \frac{y_i}{x_i^{(L)}} \frac{\partial x_i^{(L)}}{\partial s_j^{(L)}} = -\frac{y_j}{x_j^{(L)}} x_j^{(L)}(1 - x_j^{(L)}) - \sum_{i \neq j}^{10} \frac{y_i}{x_i^{(L)}}(-x_i^{(L)} x_j^{(L)})$$

$$= -y_j + y_j x_j^{(L)} + \sum_{i \neq j}^{10} y_i x_j^{(L)} = -y_j + \sum_{i=1}^{10} y_i x_j^{(L)} = -y_j + x_j^{(L)}(\sum_{i=1}^{10} y_i) = -y_j + x_j^{(L)}(1) = x_j^{(L)} - y_j$$

For this homework, $g'(s_i^{(l-1)}) = \max(0, s_i^{(l-1)})$. Once all values of $x_j^{(l)}$ and $\delta_j^{(l)}$ have been found, the stochastic gradient descent update is as follows.

$$w_{ij}^{(l)} = w_{ij}^{(l)} - \eta x_i^{(l-1)} \delta_j^{(l)}$$

We can also vectorize this notation into the following updates.

$$W^{(l)} = W^{(l)} - \eta x^{(l-1)} (\delta^{(l)})^\top$$