Math 156 Term Project 3

Deadline: Monday May 18th 11:59pm

Problem 1

ReLu function is a common activation function for neural networks. It is given by

$$\text{ReLu}(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Note that it is differentiable except at input $x = 0$. Show that $\text{ReLu} : \mathbb{R} \to \mathbb{R}$ is a convex function. (10 points) Show that the perceptron criterion for binary classification problem

$$E_P(\mathbf{w}) = \sum_{i=1}^{N} \text{ReLu}(-\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}^{(i)}) t^{(i)})$$

is convex in $\mathbf{w}$. Here the data set is given by $\mathcal{D} = \{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^{N}$ where $t^{(i)} = 1$ if $i \in C_{[1]}$ and $t^{(i)} = -1$ if $i \in C_{[2]}$. (20 points)

Problem 2

In convex optimization, we can use *subderivatives* to find descending directions when the objective function is not differentiable. The *subdifferential* of a convex function $f : \mathbb{R}^K \to \mathbb{R}$ at point $\mathbf{x} \in \mathbb{R}^K$ is defined to be the set

$$\partial f(\mathbf{x}) = \{\mathbf{p} \in \mathbb{R}^K | \text{ for all } \mathbf{y} \in \mathbb{R}^K, f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{p}^T(\mathbf{y} - \mathbf{x})\}.$$

Elements of this set are called *subderivatives* (or *subgradient*). Show that any nonzero subderivative is an ascending direction, that is, for any vector $\mathbf{p} \in \partial f(\mathbf{x}), \mathbf{p} \neq \mathbf{0}$ and any positive scalar $\eta > 0$,

$$f(\mathbf{x} + \eta \mathbf{p}) > f(\mathbf{x}). \text{ (20 points)}$$

Problem 3

In textbook §4.1.7, we introduced a stochastic gradient descent method to minimize the perceptron criterion. Alternatively, we can use the subdifferential of ReLu (as defined in Problem 1) combined with backpropagation to find the optimal weights $\mathbf{W}^{(l)}$ and biases $\mathbf{v}^{(l)}$. Show that

$$\begin{aligned} x > 0 &\Rightarrow \partial\text{ReLu}(x) = \{1\} \\ x = 0 &\Rightarrow \partial\text{ReLu}(0) = \{p \in \mathbb{R} \mid 0 \leq p \leq 1\} \\ x < 0 &\Rightarrow \partial\text{ReLu}(x) = \{0\}. \end{aligned}$$

In particular, we see that the subgradient $\partial f(\mathbf{x})$ at $\mathbf{x}$ is a singleton set containing $\nabla f(\mathbf{x})$ when the function f is differentiable at $\mathbf{x}$. (20 points)

Problem 4

In this problem, you need the notion of *big-O notation*. You can seek help from Wikipedia to familiarize yourself with it. Consider the M-layer perceptron network,

$$\mathbf{y}(\mathbf{x}, \mathbf{W}^{(1)}, \cdots, \mathbf{W}^{(M)}, \mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(M)}) = \sigma \circ \psi^{(M)} \circ h \circ \psi^{(M-1)} \circ \cdots \circ h \circ \psi^{(1)}(\mathbf{x}),$$

where $h = \tanh$ is the activation function used for each layer, and the affine map for layer l is given by $\psi^{(l)}(\mathbf{z}) = \mathbf{v}^{(l)} + \mathbf{W}^{(l)}\mathbf{z}$. Denote the (hidden) units by

$$\begin{aligned}
\mathbf{z}^{(0)} &= \mathbf{x} \in \mathbb{R}^{K_0} \\
\mathbf{a}^{(1)} &= \mathbf{v}^{(1)} + \mathbf{W}^{(1)}\mathbf{z}^{(0)} \in \mathbb{R}^{K_1} \\
\mathbf{z}^{(1)} &= h(\mathbf{a}^{(1)}) \in \mathbb{R}^{K_1} \\
\mathbf{a}^{(2)} &= \mathbf{v}^{(2)} + \mathbf{W}^{(2)}\mathbf{z}^{(1)} \in \mathbb{R}^{K_2} \\
\mathbf{z}^{(2)} &= h(\mathbf{a}^{(2)}) \in \mathbb{R}^{K_2} \\
&\vdots \\
\mathbf{a}^{(M)} &= \mathbf{v}^{(M)} + \mathbf{W}^{(M)}\mathbf{z}^{(M-1)} \in \mathbb{R}^{K_M} \\
\mathbf{z}^{(M)} &= \mathbf{y}(\mathbf{x}, \mathbf{W}^{(1)}, \cdots, \mathbf{W}^{(M)}, \mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(M)}) = \sigma(\mathbf{a}^{(M)}).
\end{aligned}$$

Suppose for simplicity that $K_1 = K_2 = \cdots = K_M = K$, express the cost of backpropagation in terms of the K and M. (30 points)