CSE 546 Homework 3B

Jake Gonzales

November 23, 2023

Collaborators: n/a

1 Perceptron

B1.

(a) We are minimizing the following:

$$L((x_i, y_i), w) = \frac{1}{n} \sum_{i=1}^{n} l((x_i, y_i), w)$$

Let w_t be the current step, then we have

$$\begin{aligned} w_{t+1} &= w_t - \eta \nabla_w [L((x_i, y_i), w_t)] \\ &= w_t - \eta (\frac{1}{n} \sum_{i=1}^{n} \nabla_w [l((x_i, y_i), w_t)]) \end{aligned}$$

where

$$\nabla_w [l((x_i, y_i), w_t)] = \begin{cases} -y_i x_i & \text{if } -y_i(w_t \cdot x_i) \geq 0 \\ 0 & \text{if } -y_i(w_t \cdot x_i) < 0 \end{cases} \tag{1}$$

therefore we have shown an expression for a single step of gradient descent.

(b) Let's consider a simple SGD approach with a batch size of 1. Our expression that we derivded above for a single step of gradient descent reduces to

$$w_{t+1} = w_t - \eta \nabla_w [l((x_i, y_i), w_t)]$$

because of the nature of the perceptron we have two scenarios: one where we predict correctly and the other incorrectly. So if we have $w_t \cdot x_i < 0$ and we have predicted correctly then according to the piece-wise function above, we have $y_i(w_t \cdot x_i) > 0$. Similarly if we have $w_t \cdot x_i > 0$. Therefore, for a correct prediction the SGD step would be

$$w_t = w_t.$$

So now suppose we predict incorrectly. This means $w_t \cdot x_i$ has the opposite sign of y_i which implies the next SGD step as

$$\begin{aligned} w_{t+1} = w_t &= \eta(-y_i x_i) \\ &= w_t + \eta(y_i x_i). \end{aligned}$$

So for $y_i = 1$

$$w_{t+1} = w_t + \eta(x_i)$$

and for $y_i = -1$

$$w_{t+1} = w_t - \eta(x_i)$$

Therefore, we see that for $\eta = 1$ we have $w_{t+1} \leftarrow w_t + x_i$ for incorrect on positive y_i and $w_{t+1} \leftarrow w_t - x_i$ for negative y_i. This is the same as the Perceptron for SGD applied with $\eta = 1$.

(c) The hinge loss is given by

$$l((x, y), w) = \max\{0, 1 - y(w \cdot x)\}.$$

The difference is the hinge loss can warrant a negative given the "1 -" in the hinge loss. The hinge loss warrants the possibility of a non-zero loss even in the case of a correct prediction, so the data points in the hinge loss will be used to update the weights but not in what we derived above. This can possibly increase the decision boundary and thus be preferred.