Machine Learning Mid-Term Exam (Homework #2)

Indra Imanuel Gunawan / 20195118

1.) a.) $E_k = \frac{1}{2}(y_{d,k} - \hat{y}_k)^2 + \frac{\lambda}{n}\sum_{k=1}^{l} w_{jk}^2$

$\lambda$ = regularization parameter

Derive $\Delta W_{jk}$ for this $E_k$

$$\frac{\partial E_k}{\partial W_{jk}} = \frac{\partial\left(\frac{1}{2}(y_{d,k} - \hat{y}_k)^2 + \frac{\lambda}{n}\sum_{k=1}^{l} W_{jk}^2\right)}{\partial W_{jk}}$$

$$= 2 \cdot \frac{\lambda}{n} \sum_{k=1}^{l} W_{jk}$$

$$\Delta W_{jk} = \alpha \cdot \frac{\partial E_k}{\partial W_{jk}} \cdot x_k$$

$$= \alpha \cdot 2 \frac{\lambda}{n} W_{jk} \cdot x_k$$

b.) Derive $\Delta W_{jk}$ for $E_k = \frac{1}{2}(y_{d,k} - \hat{y}_k)^2 + \frac{\lambda}{n}\sum_{k=1}^{l} |W_{jk}|$

$$\frac{\partial E_k}{\partial W_{jk}} = \frac{\partial\left(\frac{1}{2}(y_{d,k} - \hat{y}_k)^2 + \frac{\lambda}{n}\sum_{k=1}^{l} |W_{jk}|\right)}{\partial W_{jk}}$$

$$= \frac{\lambda}{n}$$

$$\Delta W_{jk} = \alpha \cdot \frac{\partial E_k}{\partial W_{jk}} \cdot x_k$$

$$= \alpha \cdot \frac{\lambda}{n} \cdot x_k$$

c.) Derive $\Delta W_{jk}$ when $E_k = \frac{1}{2}\left(y_{d,k} - \hat{y}_k\right)^2 + \frac{\lambda_1}{n}\sum_{k=1}^{l} |w_{jk}| + \frac{\lambda_2}{n}\sum_{k=1}^{l} W_{jk}^2$

$$\frac{\partial E_k}{\partial w_{jk}} = \frac{\partial\left(\frac{1}{2}(y_{dk} - \hat{y}_k)^2 + \frac{\lambda_1}{n}\sum_{k=1}^{l} |W_{jk}| + \frac{\lambda_2}{n}\sum_{k=1}^{l} W_{jk}^2\right)}{\partial W_{jk}}$$

$$= \frac{\lambda_1}{n} \cancel{W_{jk}} + 2 \cdot \frac{\lambda_2}{n} W_{jk}$$

$$\Delta W_{jk} = \alpha \cdot \frac{\partial E_k}{\partial w_{jk}} \cdot x_k$$

$$= \alpha \cdot \left(\frac{\lambda_1}{n} + 2 \cdot \frac{\lambda_2}{n} w_{jk}\right) \cdot x_k$$

d.) Derive $\Delta W_{jk}$ when $E_k = \frac{1}{2}\left(y_{dk} - \hat{y}_k\right)^2 + \frac{\lambda_1}{n}\sum_{k=1}^{l} |W_{jk}|^p$

$$\frac{\partial E_k}{\partial w_{jk}} = \frac{\partial\left(\frac{1}{2}(y_{d,k} - \hat{y}_k)^2 + \frac{\lambda_1}{n}\sum_{k=1}^{l} |W_{jk}|^p\right)}{\partial w_{jk}}$$

$$= p \cdot \frac{\lambda_1}{n} \cdot W_{jk}^{p-1}$$

$$\Delta W_{jk} = \alpha \cdot \frac{\partial E_k}{\partial w_{jk}} \cdot x_k$$

$$= \alpha \cdot \left(p \cdot \frac{\lambda_1}{n} \cdot W_{jk}^{p-1}\right) \cdot x_k$$

~~2.)~~

2.) $f(net) = a.\tanh(b \cdot net) = a\left[\frac{e^{b\cdot net}-1}{e^{b\cdot net}+1}\right] = \frac{2a}{1+e^{-b\cdot net}} - a$

a.) $f'(net) = 2a \cdot \left(\frac{1}{e^{-b\cdot net}+1}\right)' + (-a)'$

$$= -2 \cdot \frac{(e^{-b\cdot net}+1)'}{(e^{-b\cdot net}+1)^2} \cdot a + 0$$

$$= -\frac{2\left((e^{-bnet})' + (1)'\right) \cdot a}{(e^{-bnet}+1)^2}$$

$$= -\frac{2 \cdot e^{-bnet} \cdot a\left(-b \cdot (x)'\right)}{(e^{-bnet}+1)^2}$$

$$= \frac{2 \cdot abe^{-bnet}}{(e^{-bnet}+1)^2} = \frac{2 \cdot a \cdot b \cdot e^{bnet}}{(e^{b\cdot net}+1)^2} \to 2ab\left(\frac{e^{bnet}}{(e^{bnet}+1)^2}\right)$$

$$= 2ab\left(\frac{\cancel{e^{bnet}+1}}{(e^{bnet}+1)^{\cancel{2}}} - \frac{1}{(e^{bnet}+1)^2}\right)$$

$$= 2ab\left(\frac{1}{e^{bnet}+1} - \frac{1}{(e^{bnet}+1)^2}\right)$$

$$f''(net) = \left(\frac{2 \cdot a \cdot b \cdot e^{bnet}}{(e^{b\cdot net}+1)^2}\right)'$$

$$= 2 \cdot \frac{(e^{bnet})' \cdot (e^{bnet}+1)^2 - e^{bnet}\left((e^{bnet}+1)^2\right)'}{\left((e^{bnet}+1)^2\right)^2}$$

ab

$$= \frac{2ab\left(e^{bnet} b(e^{bnet}+1)^2 - 2(e^{bnet}+1) \cdot e^{2bnet} \cdot b\right)}{(e^{bnet}+1)^4}$$

b.) $net = -\infty$

$$f(-\infty) = a\left[\frac{e^{b\cdot(-\infty)}-1}{e^{b\cdot(-\infty)}+1}\right] = a \cdot \left[\frac{0-1}{0+1}\right] = -a$$

$$f'(-\infty) = \frac{2 \cdot a \cdot b \cdot e^{b \cdot -\infty}}{(e^{b \cdot -\infty}+1)^2} = 0$$

$$f''(-\infty) = \frac{2 \cdot a \cdot b \cdot \left(e^{b(-\infty)} \cdot b(e^{-\infty}+1)^2 - 2(e^{b(-\infty)}+1) \cdot e^{2b(-\infty)} \cdot b\right)}{(e^{b(-\infty)}+1)^4} = \frac{0}{(0+1)^4} = 0$$

$net = 0$

$$f(0) = a\left[\frac{e^{b(0)}-1}{e^{b\cdot 0}+1}\right] = a\left[\frac{1-1}{1+1}\right] = 0$$

$$f'(0) = \frac{2 \cdot a \cdot b \cdot e^{b(0)}}{(e^{b\cdot 0}+1)^2} = \frac{\cancel{2} \cdot a \cdot b}{\cancel{2} \cdot 2} = \frac{a \cdot b}{2}$$

$$f''(0) = \frac{\cancel{2ab^2 + 4ab^2 + 2ab^2 - 4ab^2 - 4ab^2}}{2^4} = 0$$

2.) net $= \infty$

$$f(\infty) = a\left[\frac{e^{b(\infty)}-1}{e^{b.\infty}+1}\right] = a\left[\frac{\infty-1}{\infty+1}\right] = \underline{\underline{a}} \quad c.)$$

$$f'(\infty) = \frac{2.a.b.e^{b(\infty)}}{(e^{b.\infty}+1)^2} = \frac{\infty}{\infty} = \underline{\underline{1}}$$

$$f''(\infty) = \cancel{\ldots}$$

$$= \frac{\infty}{\infty} = \underline{\underline{1}}$$

3.) $E(w) = \frac{1}{2}\sigma^2 - Y_{dw} + \frac{1}{2}Y_x w^2$

a) $$\frac{\partial E}{\partial w} = \frac{\partial(\frac{1}{2}\sigma^2 - Y_{dw} + \frac{1}{2}Y_x W^2)}{\partial w}$$

$$= \cancel{\ldots} \underline{-Y_d + Y_x W}$$

b.) $$W_{k+1} = W_k + \alpha\left(-\frac{\partial E}{\partial w}\right)$$ $\alpha$ = learning rate

$$= W_k + \alpha\left(-(-Y_d + Y_x W)\right)$$

$$= \underline{W_k + \alpha . (Y_d - Y_k W)}$$

c.) Optimum value for $W$ for which $E(w)$ becomes minimal

$$\frac{\partial E}{\partial W} = 0 \; \cancel{\ldots}$$

$$-Y_d + Y_x W = 0$$

$$Y_x W = Y_d$$

$$\underline{W = \frac{Y_d}{Y_x}}$$

4.) a) ReLU and Leaky ReLU

$$ReLU(x) = \begin{cases} x & \text{if } x \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$Leaky\ ReLU(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0 \end{cases}$$

b.) $$PReLU(y_i) = \begin{cases} y_i, & \text{if } y_i > 0 \\ a_i y_i, & \text{if } y_i \leq 0 \end{cases}$$

c.) $$SELU(x) = \lambda \begin{cases} x, & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \leq 0 \end{cases}$$ or $$SELU(x) = \lambda \begin{cases} x & \text{if } x \geq 0 \\ \alpha(\exp(x) - 1) & \text{if } x < 0 \end{cases}$$

d.) $Swish(x) = x \cdot Sigmoid(\beta x)$

5.) If a data set is linearly separable, the Perceptron will find a separating hyperplane in a finite number of updates.

Suppose $\exists w^*$ such that $y_i(x\, w^*) > 0 \; \forall (x_i, y_i) \in D$

Suppose that we rescale each data point and the $w^*$ such that

$\|w^*\| = 1$ and $\|x_i\| \leq 1 \; \forall x_i \in D$

Here's the proof:

Consider the effect of an update ($w$ becomes $w + yx$) on the two terms $w^T w^*$ and $w^T w$. We'll use 2 facts:

- $y(x^T w) \leq 0 \rightarrow x$ is misclassified by $w$
- $y(x^T w^*) > 0 \rightarrow w^*$ is a separating hyperplane and classifies all points correctly.

1.) Consider the effect of an update on $w^T w^*$

$$(w + yx)^T w^* = w^T w^* + y(x^T w^*) \geq w^T w^* + \gamma$$

The inequality forms from the distance from the hyperplane defined by $w^*$ to $x$ must be at leas $\gamma$. This means, for each update, $w^T w^*$ grows at least by $\gamma$

next page

(2) Consider the effect of an update on $W^TW$:

$$(W+yx)^T(W+yx) = W^TW + 2\underbrace{y(W^Tx)}_{<0} + \underbrace{y^2(x^Tx)}_{0\leq\,\leq 1} \leq W^TW+1$$

the inequality forms beacuse:

- $2y(W^Tx) < 0$ as we had to make an update, meaning $x$ was misclassified
- $0 \leq y^2(x^Tx) \leq 1$ as $y^2 = 1$ and all $x^Tx \leq 1$ (because $\|x\| \leq 1$)

(3) Now we know that after M updates the following two inequalities must hold:

(1) $W^TW^* \geq M\gamma$

(2) $W^TW \leq M$

We can complete the proof:

$M\gamma \leq W^TW^*$    By (1)

$= \|W\|\cos\theta$    by definiton of inner-product, where $\theta$ is the angle between W and W*

$\leq \|W\|$    by definition of cos, we must have $\cos(\theta) \leq 1$

$= \sqrt{W^TW}$    by definition of $\|W\|$

$\leq \sqrt{M}$    By (2)

$\Rightarrow M\gamma \leq \sqrt{M}$

$\Rightarrow M^2\gamma^2 \leq M$

$\Rightarrow M \leq \frac{1}{\gamma^2}$    And hence, the number of updates M is bounded from above by a constant