NAME: ISACK T. NICHOLAUS 20195032 MID-TERM MACHINE LEARNING.

1 ⓐ Given:

$$E_k = \frac{1}{2}\left(y_{dk} - \hat{y}_k\right)^2 + \frac{\lambda}{n}\sum_{k=1}^{l} w_{jk}^2 \qquad \text{Required: } \Delta w_{jk}$$

Solution:

$$\Delta w_{jk} = \alpha \cdot \frac{\partial E_k}{\partial w} \cancel{x_k} = \alpha \cdot \delta_k \cdot x_k$$ where x_k = Input (Assumed to be Independent variable of the model) One is assumed for simplicity

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

$$\approx y_k = \sum_k x_k w_k \approx x_k w_{jk}$$

$$\frac{\partial E_k}{\partial w} = \frac{1}{2}\cdot 2\left(y_{dk} - \hat{y}_k\right)\cdot x_k + \lambda w_{jk}$$

$$\frac{\partial E_k}{\partial w} = \left(y_{dk} - \hat{y}_k\right)\cdot x_k + \lambda w_{jk}$$

$$\therefore \Delta w_{jk} = \alpha \cdot \left(\left(y_{dk} - \hat{y}_k\right)\cdot x_k + \lambda w_{jk}\right) \text{ for Ridge } (L^2)$$

1 ⓑ Required: Δw_{jk} Given: $E_k = \frac{1}{2}\left(y_{d,k} - \hat{y}_k\right)^2 + \frac{\lambda}{n}\sum_{k=1}^{l}|w_{jk}|$

Solution:

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

$\Rightarrow \hat{y}_k \approx x_k w_{jk}$ (Assumed for simplicity)

$$\partial E_k / \partial w = \left(y_{dk} - \hat{y}_k\right) x_k + \lambda$$

$$\therefore \Delta w_{jk} = \alpha\left(\left(y_{d,k} - \hat{y}_k\right)\cdot x_k + \lambda\right) \text{ for LASSO constraint.}$$

1. ⓒ Required! $\Delta w_{jk} = ?$ Given! $E_k = \frac{1}{2}(y_{d,k} - y_k)^2 + \frac{\lambda_1}{n}\sum_{k=1}^{l}|w_{jk}| + \frac{\lambda_2}{n}\sum_{k=1}^{l} w_{jk}^2$

$$\frac{\partial E_k}{\partial w_{jk}} = \frac{\partial\left(\frac{1}{2}(y_{d,k} - \hat{y}_k)^2 + \lambda_1 w_{jk} + \lambda_2 w_{jk}^2\right)}{\partial w_{jk}}$$

Assumption for simplicity;
$\hat{y}_k = X_k w_{jk}$

$$\frac{\partial E_k}{\partial w_{jk}} = (y_{d,k} - y_k)X_k + \lambda_1 + 2\lambda_2 w_{jk}$$

$$\therefore \Delta w_{jk} = \alpha \cdot \left((y_{d,k} - \hat{y}_k)X_k + \lambda_1 + 2\lambda_2 w_{jk}\right)$$

1ⓓ Required! $\Delta w_{jk} = ?$ Given! $E_k = \frac{1}{2}(y_{d,k} - \hat{y}_k)^2 + \frac{\lambda_1}{n}\sum_{k=1}^{l}|w_{jk}|^p$

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

$$\frac{\partial E_k}{\partial w_{jk}} = (y_{d,k} - \hat{y}_k)X_k + \lambda \cdot p \cdot w_{jk}$$

$$\Delta w_{jk} = \alpha \cdot \left((y_{d,k} - \hat{y}_k) \cdot x_k + p \cdot \lambda \cdot w_{jk}\right)$$

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

Solution!

$f'(net) \Rightarrow$ from; $\frac{\partial(u/v)}{\partial x} = \frac{u' \cdot v - v' \cdot u}{v^2}$ u and v are functions of x.

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

2ⓐ $f'(net) = \dfrac{\cancel{ab\,e^{2b\cdot net}} + ab\,e^{b\cdot net} - \cancel{a\cdot b\,e^{2b\cdot net}} + a\cdot b\,e^{b\cdot net}}{(e^{b\cdot net}+1)^2}$

$$f'(net) = \frac{2ab\,e^{b\cdot net}}{(e^{b\cdot net}+1)^2}$$

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

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

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

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

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

2ⓑ $f(-\infty) = a\left[\dfrac{e^{b(-\infty)} - 1}{e^{b(-\infty)}+1}\right] = a\left[\dfrac{0-1}{0+1}\right] = -a$

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

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

2(b) $f'(net) = \dfrac{2abe^{b\cdot net}}{(e^{b\cdot net}+1)^2}$

$f'(-\infty) = \dfrac{2abe^{b(-\infty)}}{(e^{b(\infty)}+1)^2} = 0$, $f'(0) = \dfrac{2abe^{b(0)}}{(e^{b(0)}+1)^2} = \dfrac{2ab}{2^2} = \frac{1}{2}ab$.

$f'(\infty) = \dfrac{2abe^{b(\infty)}}{(e^{b(\infty)}+1)^2} = \dfrac{\infty}{\infty} = 1$

$\therefore f'(\infty) = 0$, $f'(0) = \frac{1}{2}ab$ and $f'(\infty) = 1$.

2(b) $f''(net) = ?$

from; $f'(net) = \dfrac{2abe^{b\cdot net}}{(e^{b\cdot net}+1)^2} = \dfrac{2abe^{b\cdot net}}{e^{2bnet}+2e^{b\cdot net}+1}$

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

$$f''(net) = \frac{2ab^2e^{3bnet} + 4ab^2e^{2b\cdot net} + 2ab^2e^{b\cdot net} - 4ab^2e^{3b\cdot net} - 4ab^2e^{2b\cdot net}}{(e^{b\cdot net}+1)^4}$$

$f''(-\infty) = \dfrac{0}{(0+1)^4} = \dfrac{0}{1} = 0$ $f''(0) = \dfrac{2ab^2+4ab^2+2ab^2-4ab^2-4ab^2}{2^4}$

$f''(0) = 0$

$f''(\infty) = \dfrac{\infty}{(\infty+1)^4} = \dfrac{\infty}{\infty} = 1$

$\therefore f''(-\infty) = 0$, $f''(0) = 0$ and $f''(\infty) = 1$.

3. Given! $E(w) = \frac{1}{2}\sigma^2 - Y_d w + \frac{1}{2} Y_x w^2$

(a) $\frac{\partial E}{\partial w} = ?$

$$\Rightarrow \frac{\partial E}{\partial w} = \frac{\partial\left(\frac{1}{2}\sigma^2 - Y_d w + \frac{1}{2} Y_x w^2\right)}{\partial w}$$

$$\frac{\partial E}{\partial w} = 0 - Y_d + \frac{1}{2}\left(2 Y_x w\right)$$

$$\frac{\partial E}{\partial w} = -Y_d + Y_x w = Y_x w - Y_d$$

$$\therefore \frac{\partial E}{\partial w} = Y_x w - Y_d$$

(b) Update equation for w_{k+1}

Then; $w_{k+1} = w_k + \alpha\left(-\frac{\partial E}{\partial w}\right)$ but $\frac{dE}{\partial w} = Y_x w - Y_d$

$$w_{k+1} = w_k + \alpha\left(-\left(Y_x w - Y_d\right)\right)$$

$$w_{k+1} = w_k + \alpha\left(Y_d - Y_x w\right)$$

$$\therefore w_{k+1} = w_k + \alpha\left(Y_d - Y_x w\right)$$

Where α is the learning rate

(c) (1) Optimum value of w; To get minimal value gradient/slope must be equal to 0 (zeros) in this case

$\frac{\partial E}{\partial w} = 0$; Then; $0 = Y_x w - Y_d$

$$w = Y_d / Y_x$$

$\therefore$ At $w = Y_d / Y_x$ $E(w)$ becomes minimal.

4. Weight update;

(a) ReLU and Leaky ReLU;

Given!

$w_{13} = 0.5,\ w_{14} = 0.9,\ w_{24} = 1.0,\ w_{23} = 0.4,\ w_{35} = 1.2,\ w_{45} = 1.1,\ \theta_3 = 0.8,\ \theta_4 = -0.1$
and $\theta_5 = 0.3$

from; $g(x) = \begin{cases} x & \text{if } x \geq 1 \\ 0 & \text{otherwise.} \end{cases}$

$$y_3 = ReLU(x_1 w_{13} + x_2 w_{23} - \theta_3) = g(1 \times 0.5 + 1 \times 0.4 - 0.8) = 0.1$$

$$y_4 = ReLU(x_1 w_{14} + x_2 w_{24} - \theta_4) = g(1 \times 0.9 + 1 \times 1 + 0.1) = 2.0$$

$$y_5 = ReLU(y_3 w_{35} + y_4 w_{45} - \theta_5 = g(0.1 \times 1.2 + 2.0 \times 1.1 - 0.3) = 2.02$$

Then; error;

$$e = y_{d,5} - y_5 = 0 - 2.02 = -2.02$$

Then, error gradient for neuron 5 in output layer;

first; $g'(x) = \dfrac{\partial(g(x))}{\partial x} = \begin{cases} 1 & \text{if } x \geq 1 \\ 0 & \text{otherwise.} \end{cases}$

$\delta_5 = g'(x) \cdot e$ But $x = y_5 = 2.02$.

$\delta_5 = g'(2.02) \cdot -2.02$

for $g'(2.02) = 1$

Hence $\delta_5 = -2.02$.

* Weight corrections

$\Delta W_{35} = \alpha \cdot y_3 \cdot \delta_5 = 0.1 \times 2.02 \times 0.1 = \underline{0.0202}$

$\Delta W_{45} = \alpha \cdot y_4 \cdot \delta_5 = 0.1 \times 2 \times 2.02 = \cancel{0.08}\ \underline{0.404}$

$\Delta \theta_5 = \alpha(-1) \cdot \delta_5 = 0.1 \times (-1) \times 2.02 = \underline{-0.202}$

* Error gradient for neurons 3 and 4 in the hidden layers

$\delta_3 = g'(y_3) \cdot \delta_5 \cdot W_{35}$

for $g'(y_3) = g'(0.1) = 0$

hence; $\delta_3 = 0$

$\delta_4 = g'(y_4) \cdot \delta_5 \cdot W_{45}$

$\delta_4 = g'(2.0) \cdot \delta_5 \cdot W_{45}$

$\delta_4 = 1 \times 2.02 \times 1.1$

$\delta_4 = 2.42$

* Then; Weight corrections;

$\Delta W_{13} = \alpha \cdot X_1 \cdot \delta_3 = 0.1 \times 1 \times 0 = 0$

$\Delta W_{23} = \alpha \cdot X_2 \cdot \delta_3 = 0.1 \times 1 \times 0 = 0$

$\Delta \theta_3 = \alpha \cdot (-1) \cdot \delta_3 = 0.1 \times -1 \times 0 = 0$

$\Delta W_{14} = \alpha \cdot X_1 \cdot \delta_4 = 0.1 \times 1 \times 2.42 = 0.242$

$\Delta W_{24} = \alpha \cdot X_2 \cdot \delta_4 = 0.1 \times 1 \times 2.42 = 0.242$

$\Delta \theta_4 = \alpha \cdot (-1) \cdot \delta_4 = 0.1 \times (-1) \times 2.42 = -0.242$

* Update Weights and biases (Threshold).

$W_{13} = 0.5\ W_{13} + \Delta W_{13} = 0.5 + 0 = 0.5$

$W_{14} = W_{14} + \Delta W_{14} = 0.9 + 0.242 = 1.142$

$W_{23} = W_{23} + \Delta W_{23} = 0.4 + 0 = 0.4$

$W_{24} = W_{24} + \Delta W_{24} = 1.0 + 0.242 = 1.242$

$W_{35} = W_{35} + \Delta W_{35} = -1.2 + 0.0202 = -1.1808$

$W_{45} = W_{45} + \Delta W_{45} = 1.1 + 0.404 = 1.504$

$\theta_3 = \theta_3 + \Delta\theta_3 = 0.8 + 0 = 0.8$

$\theta_4 = \theta_4 + \Delta\theta_4 = -0.1 + 0.242 = 0.142$

$\theta_5 = \theta_5 + \Delta\theta_5 = 0.3 + (-0.202) = 0.008$

4 (b) PReLU

Given! $w_{13} = 0.5$, $w_{14} = 0.9$, $w_{24} = 1.0$, $w_{23} = 0.4$, $W_{35} = 1.2$, $w_{45} = 1.1$, $\theta_3 = 0.8$, $\theta_4 = -0.1$ and $\theta_5 = 0.3$

PReLU activation function: $g(x) = \begin{cases} x & \text{if } x > 0 \\ a_i x & \text{if otherwise} \end{cases}$

But in PReLU activation function

a_i = Coefficient controlling the slope

for a_i is learnable parameter.

hence, $\dfrac{\partial(g(x_i))}{\partial a_i} = \begin{cases} 0, & \text{if } x_i > 0 \\ x_i, & \text{if } x_i \leq 0 \end{cases}$

So;

$$y_k = g\left(\sum_{j=1}^{m} x_{jk}(p) \cdot w_{jk}(p) - \theta_k\right)$$

$$y_j = g\left(\sum_{i=1}^{n} x_i(p) \cdot w_{ij}(p) - \theta_j\right)$$

Weight Training/Update;

$\delta_k(p) = g'(y_k) \cdot e_k(p)$ Where $e_k(p) = y_{d,k}(p) - y_k(p)$

Weight Correction!

$$\Delta w_{jk}(p) = \alpha \cdot y_j(p) \cdot \delta_k(p).$$

$$\therefore w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p).$$

$$\delta_j(p) = g'(y_j) \cdot \sum_{k=1}^{l} \delta_k(p) w_{jk}(p)$$

$$\Delta w_{ij}(p) = \alpha \cdot x_i(p) \cdot \delta_j(p)$$

$$\therefore w_{ij}(p) = w_{ij}(p) + \Delta w_{ij}(p).$$

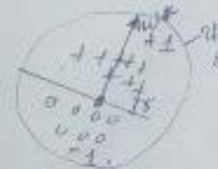
5. What we always what;
for each data point

$\{x : w^T x + b = 0\}$

For generalisation;

$w^T x = 0$.

Example: Given;



w^* Exist hyper-plane defined by (w^*) with $\|w\|^* = 1$.

γ - Distance of hyperplane to the closest datapoint.

So; $y = +1 \Rightarrow w^T x > 0$
$y = -1 \Rightarrow w^T x < 0$ $\Big\}\ y w^T x > 0$

Hence; Update rule;

$y = +1 \Rightarrow w \leftarrow w + x$
$y = -1 \Rightarrow w \leftarrow w - x$ $\Big\}\ w \leftarrow w + yx$

~~o~~ If the above holds then $1/\gamma^2$ mistakes are ~~made~~ made at most.

Proof needed to show $y w^T x > 0$ will be true, for every data point by separating hyper-plane and classifies all data point correctly.

$\exists w$ such that $\forall (x, y) \leftarrow D$

$\alpha\, y w^* x > 0$

Scalar value

$w^* = \alpha w^*$

$\|w^*\| = 1$ Mean all datapoints will be ~~in the~~ within the radius of 1 unit.

$\forall_i \|x_i\| \le 1$.

So, $x_i = \dfrac{x_i}{\max_j |x_j|}$ so that the maximum one should have exactly norm 1.

Then; On update of ~~xx~~ $w^T w^*$

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

This means; $y(x^T w^*) = |x^T w^*| \geq \gamma$

On single update step; $w^T w^*$ grows by $\approx \gamma$ (at least)

Then; Update on $w^T w^*$

$$(w+yx)^T (w+yx) = w^T w^* + 2y(w^T w) + y^2(x^T x) \leq w^T w + 1$$

$2y(w^T \cancel{x}) < 0$ means x is misclassified, update is required.

$0 \leq y^2(x^T x) \leq 1$ as $y^2 = 1$ and $x^T x \leq 1$ because $||x|| \leq 1$.

Means $w^T w$ grows at most 1. for each update.

Therefore; after K updates

$w^T w \leq K$ and $w^T w^* \geq K(\gamma)$ these must be true

$K\gamma \leq w^T w^*$

$= ||w|| \cos(\theta)$ Inner product, θ is angle between w and w^*

$\leq ||w||$ Cos (θ) must ~~have~~ have $\cos\theta \leq 1$.

$= \sqrt{w^T w}$ by definition of $||w||$

$\leq \sqrt{K}$ because $w^T w \leq K$

Then; $K\gamma \leq \sqrt{K}$

$K^2 \gamma^2 \leq K$

$\therefore K \leq 1/\gamma^2$ K is bounded ~~from~~ by a constant γ