Name: Yuhan Shao
Student Number: 1002281327
UTORid: shaoyuha

CSC411 Assignment 3

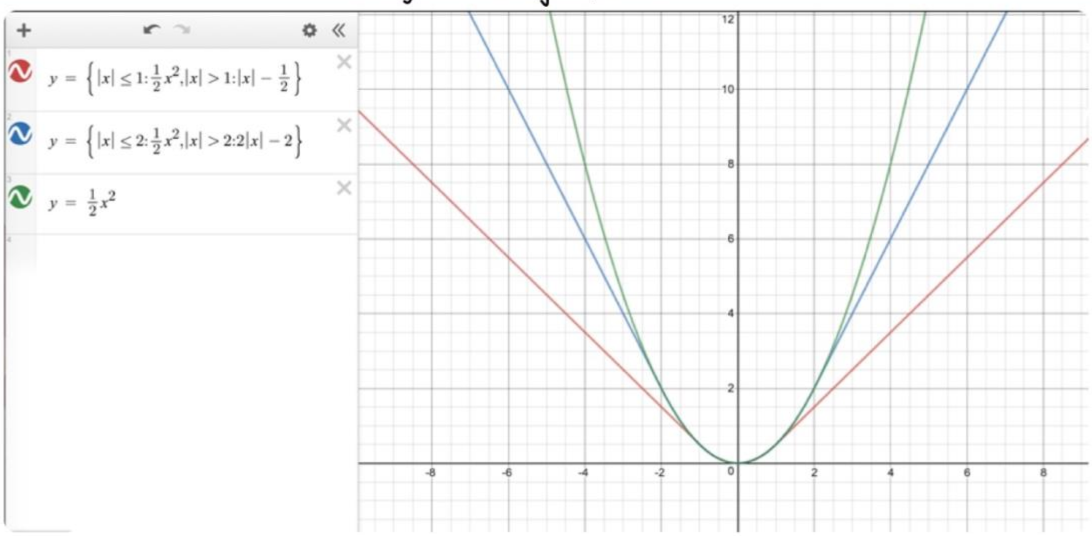
1. (a) When $t=0$,

$$L_{SE}(y,t) = \frac{1}{2}(y-t)^2 = \frac{1}{2}y^2$$

$$L_\delta(y,t) = H_\delta(y-t) = H_\delta(y) = \begin{cases} \frac{1}{2}y^2 & , \text{ if } |y| \le \delta \\ \delta(|y| - \frac{1}{2}\delta) & , \text{ if } |y| > \delta \end{cases}$$

• When $\delta = 1$, $L_\delta(y,t) = \begin{cases} \frac{1}{2}y^2 & , \text{ if } |y| \le 1 \\ |y| - \frac{1}{2} & , \text{ if } |y| > 1 \end{cases}$

• When $\delta = 2$, $L_\delta(y,t) = \begin{cases} \frac{1}{2}y^2 & , \text{ if } |y| \le 2 \\ 2|y| - 2 & , \text{ if } |y| > 2. \end{cases}$



• Huber loss is more robust to outliers, since:
As $|y-t|$ i.e: absolute value of residual grows larger,
[correspond to the x-axis in the graph]
Huber loss grows linearly, while squared error loss grows quadratically.
Huber loss grows much slower than squared error loss,
therefore, Huber loss is less sensitive to the outliers.

(b) $H'_\delta(a) = \begin{cases} a & , \text{ if } |a| \le \delta \\ \delta \frac{a}{|a|} & , \text{ if } |a| > \delta \end{cases}$

$$\xrightarrow[a=y-t]{\text{substitute}} H'_\delta(y-t) = \begin{cases} y-t & , \text{ if } |y-t| \le \delta \\ \delta \frac{y-t}{|y-t|} & , \text{ if } |y-t| > \delta \end{cases}$$

where $y = w^T x + b$

$$\frac{\partial L_\delta}{\partial w} = \frac{\partial L_\delta}{\partial a} \cdot \frac{\partial a}{\partial w} = \left(H'_\delta(y-t)\right) \cdot (x) = x \cdot H'_\delta(y-t)$$

$$\frac{\partial L_\delta}{\partial b} = \frac{\partial L_\delta}{\partial a} \cdot \frac{\partial a}{\partial b} = \left(H'_\delta(y-t)\right) \cdot (1) = H'_\delta(y-t)$$

2. (a) According to Section 3.1 of CSC321:

Let $\mathcal{E} = \frac{1}{2}\left(\sum_{i=1}^{N} a^{(i)}(w^T x^{(i)} - y^{(i)})^2\right) + \frac{\lambda}{2}\|w\|^2$.

$$\frac{\partial \mathcal{E}}{\partial w_j} = \frac{1}{2}\left(\sum_{i=1}^{N} a^{(i)} \cdot 2(w^T x^{(i)} - y^{(i)})x_j^{(i)}\right) + \frac{\lambda}{2} \cdot 2w_j$$

$$= \left(\sum_{i=1}^{N} a^{(i)}(w^T x^{(i)} - y^{(i)})\, x_j^{(i)}\right) + \lambda w_j$$

Therefore, $\frac{\partial \mathcal{E}}{\partial w} = X^T A(Xw - y) + \lambda w$ # A is NxN diagonal

- In order to get the argmin w^*, set $\frac{\partial \mathcal{E}}{\partial w} = 0$.

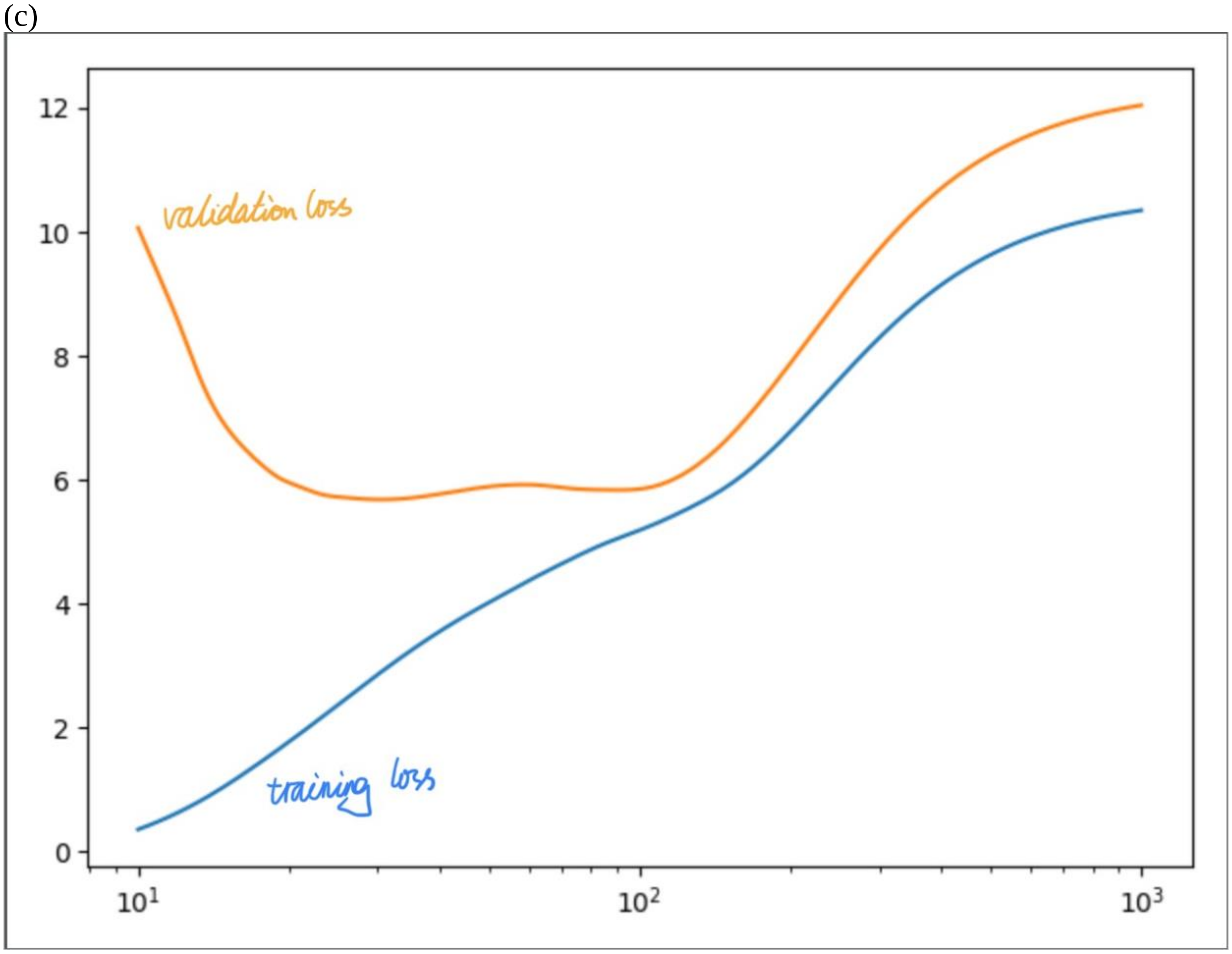
$$X^T A(Xw^* - y) + \lambda w^* = 0.$$

$$(X^T A X + \lambda I)w^* = X^T A y.$$

$$w^* = (X^T A X + \lambda I)^{-1}(X^T A y)$$

- Therefore, $w^* = (X^T A X + \lambda I)^T X^T A y$ as desired.

(c)



(d)

(d) • As $\tau \to \infty$, the weight $w \to$ a constant, i.e: a week predition
and may cause underfit.
⇒ both training loss & validation loss might be high
• As $\tau \to 0$, the weight w decrease suddenly and senstive to distance change
and may cause overfit
⇒ training loss tend to 0 & validation loss is high
• This is matched with the plotted graph.
• To discuss the behavior of $a^{(i)}$,
Consider the equation on left ⇒

$$a^{(i)} = \frac{\exp(-||\mathbf{x}-\mathbf{x}^{(i)}||^2/2\tau^2)}{\sum_j \exp(-||\mathbf{x}-\mathbf{x}^{(j)}||^2/2\tau^2)}$$

$$\Rightarrow a^{(i)} = \frac{\left(e^{\frac{-||x-x^{(i)}||^2}{2}}\right)^{\frac{1}{\tau^2}}}{\left(e^{\frac{-||x-x^{(1)}||^2}{2}}\right)^{\frac{1}{\tau^2}} + \cdots + \left(e^{\frac{-||x-x^{(n)}||^2}{2}}\right)^{\frac{1}{\tau^2}}}$$

• Then, as $\tau \to 0$, $\frac{1}{\tau^2} \to \infty$, $a^{(i)} \to 0$
as $\tau \to \infty$, $\frac{1}{\tau^2} \to 0$, $a^{(i)} \to \infty$