CSC 411 Fall 2018

Machine Learning and Data Mining

Homework 3

Family name: Bhatia

Given name: Pooja

solution 1

(a)

$$L_\delta(y,t) = H_\delta(y-t)$$

$$H_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \le \delta \\ \delta\left(|a| - \frac{1}{2}\delta\right) & \text{for } |a| > \delta \end{cases}$$

Now,

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

Now taking $t = 0$

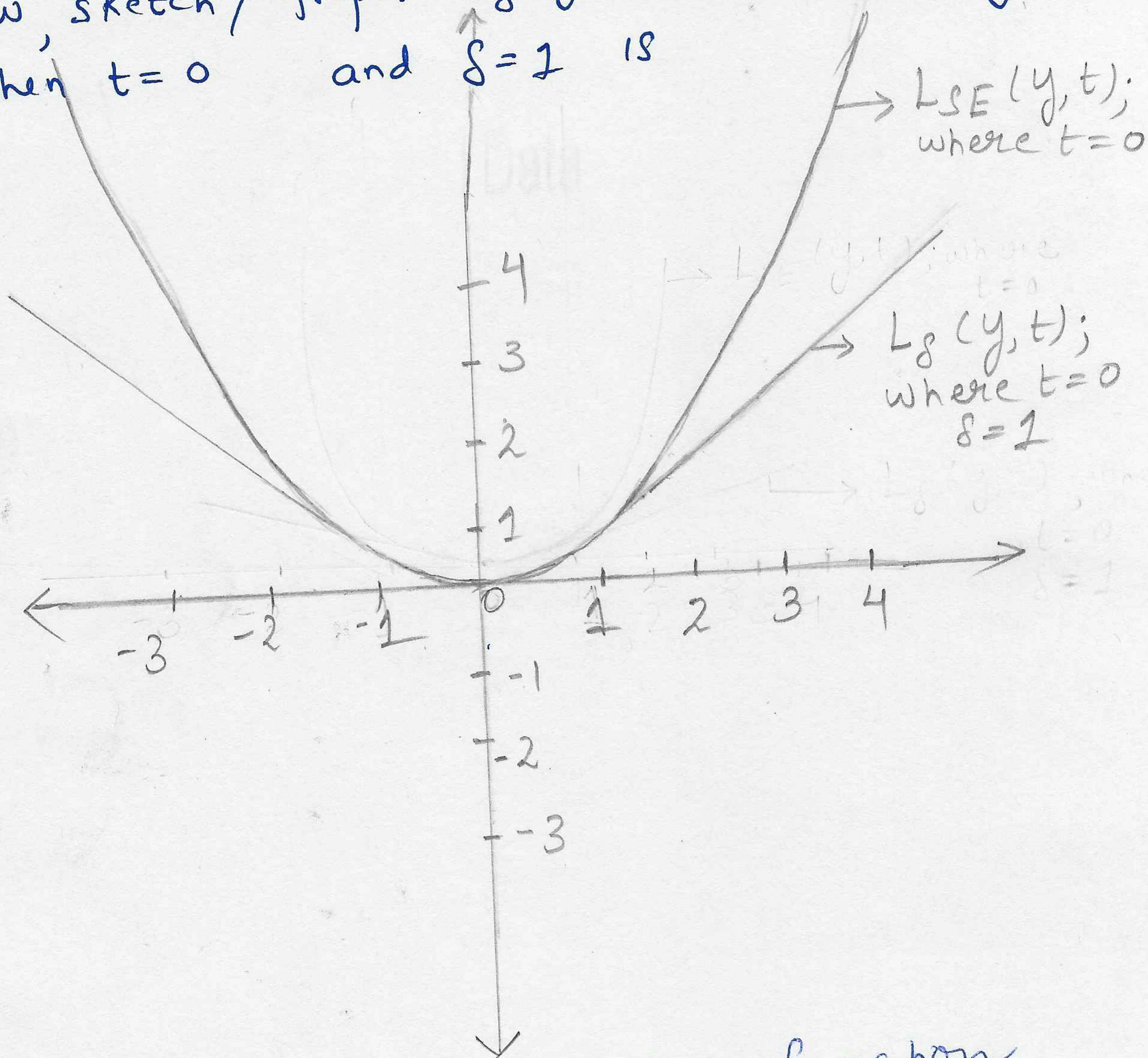
$$L_\delta(y) = \begin{cases} \frac{1}{2}y^2 & \text{if } |y| \le \delta \\ \delta(y) - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$

L_{SE} i.e the squared Loss function

$L_{SE} = \frac{1}{2}(y-t)^2$, Now taking $t = 0$

$$L_{SE} = \frac{1}{2}y^2$$

Now, sketch / graph $L_\delta(y,t)$ and $L_{SE}(y,t)$ when $t=0$ and $\delta=1$ is



Based on the sketch Huber function is strongly convex function in a uniform neighborhood of its minimum $a=0$

i.e $L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a|<\delta \\ \delta\{|a| - \frac{1}{2}\delta\} & \text{otherwise} \end{cases}$

Therefore at the boundaries of the uniform neighborhood the Huber loss function has a differentiable extension to an Affine function at $a=-\delta$ and $a=+\delta$. making it less sensitive/more Robust to outliners than squared Loss error.

The squared Loss Error is more sensitive to outliers as while taking summation the sample mean gets influenced by a large value of y in this case. i.e

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

$$L_{SE}(y) = \frac{1}{2}(y)^2 \text{; when } t=0$$

Ans

P. T. O

solution 1

(b)

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

$\cancel{L_\delta(y,t) = H_\delta(y-t)}$

$$L_\delta(y,t) = H_\delta(y-t)$$

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

Now, $y = w^T x + b$

$$H_\delta(y-t) = \begin{cases} \frac{1}{2}(w^T x + b - t)^2 & \text{if } |w^T x + b - t| \le \delta \\ \delta(|w^T x + b - t| - \frac{1}{2}\delta) & \text{if } |w^T x + b - t| > \delta \end{cases}$$

$\to$ 1

$$\frac{\partial L_\delta}{\partial w} = \frac{\partial(H_\delta(y-t))}{\partial w}$$

$$= \begin{cases} \frac{1}{2}\frac{\partial}{\partial w}(w^T x + b - t)^2, & \text{if } |w^T x + b - t| \le \delta \\ \delta \frac{\partial}{\partial w}\left(|w^T x + b - t| - \frac{1}{2}\delta\right) & \text{if } |w^T x + b - t| > \delta \end{cases}$$

$$\boxed{\frac{\partial L_\delta}{\partial w} = \begin{cases} (w^T x + b - t)(x) & \text{if } |w^T x + b - t| \le \delta \\ \delta(x) & \text{if } |w^T x + b - t| > \delta \end{cases}}$$

ans.

$\frac{\partial L_\delta}{\partial w}$ in terms of y

$$\boxed{\frac{\partial L_\delta}{\partial w} = \begin{cases} (y-t)\cdot(x) & \text{if } |y-t| \le \delta \\ (\delta)\cdot(x) & \text{if } |y-t| > \delta. \end{cases}}$$

Ans

Similarly from equation (1)

$$\frac{\partial L_\delta}{\partial b} = \frac{\partial (H_\delta (y-t))}{\partial b}$$

$$= \begin{cases} \frac{1}{2} \frac{\partial}{\partial b} (w^T x + b - t)^2 & \text{if } |w^T x + b - t| \le \delta \\ \delta \frac{\partial}{\partial b} (|w^T x + b - t| - \frac{1}{2} \delta) & \text{if } |w^T x + b - t| > \delta \end{cases}$$

$$\boxed{\frac{\partial L_\delta}{\partial b} = \begin{cases} (w^T x + b - t) & \text{if } |w^T x + b - t| \le \delta \\ \delta & \text{if } |w^T x + b - t| > 0 \end{cases}}$$

Ans.

$\frac{\partial L_\delta}{\partial b}$ in terms of y

$$\boxed{\frac{\partial L_\delta}{\partial b} = \begin{cases} (y-t) & \text{if } |y-t| \le \delta \\ \delta & \text{if } |y| > \delta \end{cases}}$$

Ans

Solution 2 (a)

Given Training data $\{(x^{(1)}, y^{(1)}) \ldots \ldots (x^{(N)}, y^{(N)})\}$ and positive weights $a^{(1)} \ldots \ldots a^{(N)}$

Now, writing/representing them in Matrix Notation

$$y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix}_{N \times 1} \qquad X = \begin{bmatrix} 1 & \cdots & \cdots & x^{(1)(D)} \\ \vdots & \vdots & & \vdots \\ 1 & x^{(N)(1)} & \cdots & x^{(N)(D)} \end{bmatrix}_{N \times (D+1)}$$

$$A = \begin{bmatrix} a^{(1)} & \cdots 0 & 0 \\ \vdots & a^{(2)} & \vdots \\ 0 & \cdots & a^{(N)} \end{bmatrix}_{N \times N} \qquad w = \begin{bmatrix} b \\ w^{(1)} \\ \vdots \\ w^{(D)} \end{bmatrix}_{(D+1) \times 1}$$

$y = w^T x + b$ (in case of linear regression)

Now, The objective in this case would be

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

Now, representing the same in Matrix Notation

$$E(w) = \frac{A}{2}(Y - XW)^T(Y - XW) + \frac{\lambda}{2} w^T w \quad \to (1)$$

Now, we need to minimize the square loss function along with regulizer term in above equation

Now, taking derivative w.r.t w of equation (1) and equating it to zero

$$\nabla E(w) = -X^T A(Y - XW) + \lambda W = 0$$

$$= -X^T AY + X^T AXW + \lambda W = 0$$

$$\Rightarrow X^T AY = X^T AXW + \lambda W$$

$$X^T AY = (X^T AX + \lambda I) w$$

(where I is the identity matrix)

Taking inverse on both sides of equation

$$X^T A Y = (X^T A X + \lambda I) w$$

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

$\therefore$

Solution to the weighted least square problem

$$w^* = \arg\min \frac{1}{2} \sum_{i=1}^{N} a^{(i)} (y^{(i)} - w^T x^{(i)})^2 + \frac{\lambda}{2} \|w\|^2$$

is

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

Ans

Solution Q2
(d)

When $\tau \to 0$
Then both the training loss
and test loss are
approaching the same value.
In the graph you will witness
the plots train_losses
test_losses is merging.

When $\tau \to \infty$
Then gap between the training loss
and test loss is increasing.
They are approaching different
values.
In the graph you will witness
the plots train_losses and
test_losses is diverging