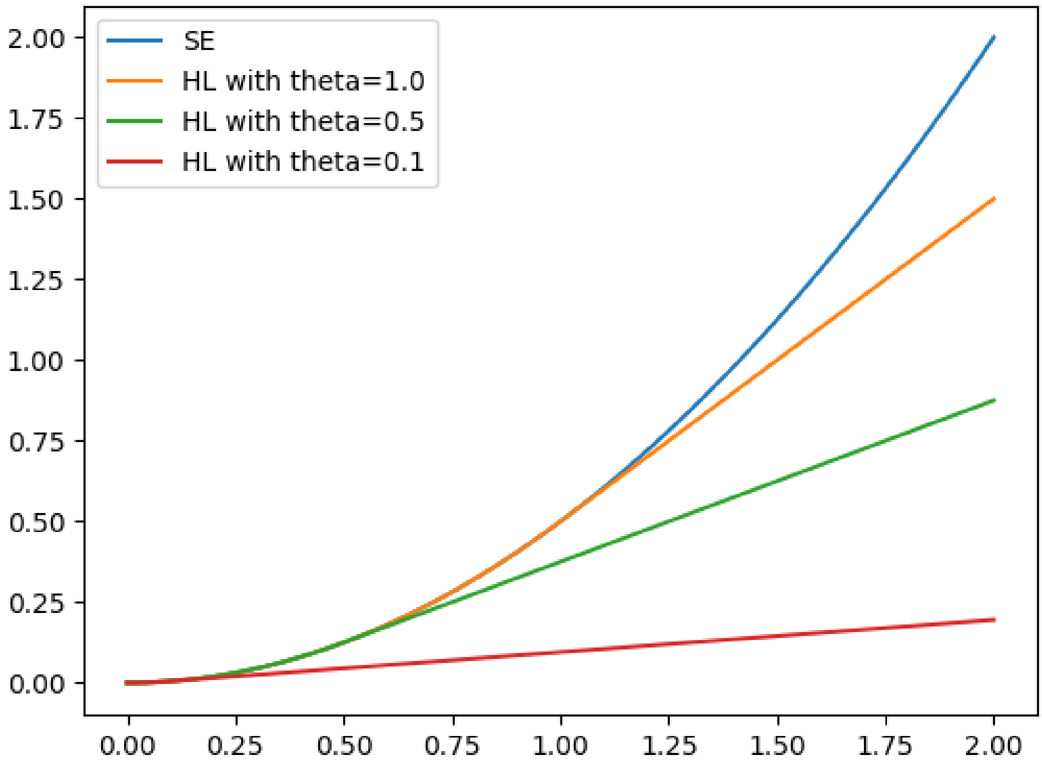
Homework 2

Zihan Zhao

1001103708

1

(a)



Compared with squared error loss, when the residue (y-t) increases, Huber loss is the same as squared error loss. But when it reaches over the threshold δ, i.e. the loss is at outliers, Huber loss becomes linearly increasing by the slope of δ. It becomes less sensitive to outliers than squared error loss. So the optimal weights can be determined more quickly by using Huber loss gradient descent. Therefore it is robust regression.

(b)

Now determines $\frac{\mathrm{d}L_\delta}{\mathrm{d}w}$:

$$\frac{\mathrm{d}L_\delta}{\mathrm{d}w} = \frac{\mathrm{d}H_\delta(a)}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}w}$$

When $|y - t| \leqq \delta$:

$$\begin{aligned}
&= \frac{\mathrm{d}\frac{1}{2}a^2}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}w} \\
&= a * 1 * x = ax = (y-t)x \\
&= (w^\mathsf{T} x + b - t)x
\end{aligned}$$

When $|y - t| > \delta$:

$$\begin{aligned}
&= \frac{\mathrm{d}\delta(|a| - \frac{1}{2}\delta)}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}w} \\
&= \begin{cases} \delta x, & y - t > \delta \\ -\delta x, & y - t < -\delta \end{cases}
\end{aligned}$$

Now determines $\frac{\mathrm{d}L_\delta}{\mathrm{d}b}$:

$$\frac{\mathrm{d}L_\delta}{\mathrm{d}b} = \frac{\mathrm{d}H_\delta(a)}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}b}$$

When $|y - t| \leqq \delta$:

$$\begin{aligned}
&= \frac{\mathrm{d}\frac{1}{2}a^2}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}b} \\
&= a * 1 * 1 = x \\
&= w^\mathsf{T} x + b - t
\end{aligned}$$

When $|y - t| > \delta$:

$$\begin{aligned}
&= \frac{\mathrm{d}\delta(|a| - \frac{1}{2}\delta)}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}w} \\
&= \begin{cases} \delta, & y - t > \delta \\ -\delta, & y - t < -\delta \end{cases}
\end{aligned}$$

(c)

Look at q1.py.

2

(a)

First factor the Loss formula:

$$
\begin{aligned}
L &= \frac{1}{2}\sum_{i=1}^{N} a^{(i)}(y^{(i)} - w^\intercal x^{(i)})^2 + \frac{\lambda}{2}||w||^2 \\
&= \frac{1}{2}A||Y - Xw||^2 + \frac{\lambda}{2}w^\intercal w \text{ (where Y is Nx1, X is Nxd, A is NxN, and w is dx1)} \\
&= \frac{1}{2}(Y - Xw)^\intercal(A(Y - Xw)) + \frac{\lambda}{2}w^\intercal w \\
&= \frac{1}{2}(Y^\intercal AY - Y^\intercal AXw - (Xw)^\intercal AY + w^\intercal X^\intercal AXw) + \frac{\lambda}{2}w^\intercal w \\
&= \frac{1}{2}(Y^\intercal AY - Y^\intercal AXw - (AY)^\intercal Xw + w^\intercal(X^\intercal AX)w) + \frac{\lambda}{2}w^\intercal w \\
&= \frac{1}{2}(Y^\intercal AY - Y^\intercal AXw - Y^\intercal A^\intercal Xw + w^\intercal(X^\intercal AX)w) + \frac{\lambda}{2}w^\intercal w
\end{aligned}
$$

Since $A = A^\intercal$,

$$
\begin{aligned}
&= \frac{1}{2}(Y^\intercal AY - Y^\intercal AXw - Y^\intercal AXw + w^\intercal(X^\intercal AX)w) + \frac{\lambda}{2}w^\intercal w \\
&= \frac{1}{2}Y^\intercal AY - Y^\intercal AXw + \frac{1}{2}w^\intercal(X^\intercal AX)w + \frac{\lambda}{2}w^\intercal w
\end{aligned}
$$

Now take derivative of L by w:

Since $A = A^\intercal$, so $X^\intercal AX$ is symmatric as well, then

$$
\begin{aligned}
\frac{\mathrm{d}L}{\mathrm{d}w} &= 0 - Y^\intercal AX + \frac{1}{2}2(X^\intercal AX)w + \frac{1}{2}2\lambda w \\
&= -Y^\intercal AX + (X^\intercal AX)w + \lambda w
\end{aligned}
$$

Let $\frac{\mathrm{d}L}{\mathrm{d}w} = 0$, we got

$$
\begin{aligned}
-Y^\intercal AX + (X^\intercal AX)w + \lambda w &= 0 \\
(X^\intercal AX + \lambda I)w &= Y^\intercal AX \\
w &= (X^\intercal AX + \lambda I)^{-1}Y^\intercal AX
\end{aligned}
$$

Since $Y^\intercal AX = (AX)^\intercal Y = X^\intercal A^\intercal Y = X^\intercal AY$,

$$
w = (X^\intercal AX + \lambda I)^{-1}X^\intercal AY
$$

Done.

(b)

Look at q2.py.

(c)

Look at q2.py.