HW-2 CSC 2515

ANEEK DAS

1. Robust Regression.

a) Plots are attached.

• Reason why Huber Loss is better for handling outliers:

when using squared error loss $L_{SE}(y,t) = \frac{1}{2}(y-t)^2$, we see that when the predicted value is far from the actual value, the loss is very large as the difference is squared.

So, when our model comes across an outlier, the loss is very large. So, the outlier gets more importance as fitting / predicting values ~~close~~ that are close to the outlier is the only way of minimizing loss.

Since, the Huber Loss uses MAE $L_{SE}(y,t) = \text{abs}(y-t)$ for ~~errors~~ difference greater than threshold (δ), it is more robust as the loss scales linearly. This is why Huber loss is more robust to outliers.

1.b) $y = w^T x + b$.

To find $\frac{\partial L_\delta}{\partial w}$, $\frac{\partial L_\delta}{\partial b}$ for Huber loss.

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

$$H_\delta(a) = \begin{cases} \frac{1}{2}a^2 & ; |a| \leq \delta \\ \delta\left(|a| - \frac{1}{2}\delta\right) & ; |a| > \delta. \end{cases}$$

where $a = y - t$.

$y \rightarrow$ predicted value.

$t \rightarrow$ true value.

First we find $\frac{\partial L_\delta}{\partial w}$

case 1 $\quad |a| \leq \delta$

$$\frac{\partial L_\delta}{\partial w} = \frac{\partial H(a)}{\partial w} = \frac{\partial H(a)}{\partial a} \times \frac{\partial a}{\partial y} \times \frac{\partial y}{\partial w}.$$

$$\frac{\partial H(a)}{\partial a} = \frac{\partial}{\partial a}\left(\frac{1}{2}a^2\right) = \frac{1}{2}\frac{\partial}{\partial a}(a^2) = \frac{1}{\cancel{2}} \times \cancel{2}a = \underline{\underline{a}}.$$

$$\frac{\partial a}{\partial y} = \frac{\partial (y-t)}{\partial y} = \underline{\underline{1.}}$$ and $$\frac{\partial y}{\partial w} = \frac{\partial}{\partial w}(wx+b) = \underline{\underline{x}}$$

So, $$\frac{\partial L}{\partial w} = \frac{\partial H(a)}{\partial a} \times \frac{\partial a}{\partial y} \times \frac{\partial y}{\partial w}$$

$$= a \cdot 1 \cdot x = \underline{\underline{ax}}$$ So, $$\frac{\partial L_\delta}{\partial w} = ax \; ; \; |a| \le \delta.$$

Now,

<u>Case 2</u> $|a| > \delta$

$$\Rightarrow -\delta > a > \delta$$

$$\frac{\partial L}{\partial w} = \frac{\partial H(y-t)}{\partial w} = \frac{\partial H(a)}{\partial a} \times \frac{\partial a}{\partial y} \times \frac{\partial y}{\partial w}.$$

$$\frac{\partial H(a)}{\partial a} = \frac{\partial}{\partial a}\left(\delta\left(|a| - \frac{1}{2}\delta\right)\right)$$

$$= \delta \cdot \frac{\partial}{\partial a}|a|$$

when $a > \delta$; $|a| = a$
So, $\delta \frac{\partial |a|}{\partial a} = \delta \frac{\partial a}{\partial a} = \underline{\underline{\delta}}$

when $-\delta > a$; $|a| = -a$.
So, $\delta \frac{\partial |a|}{\partial a} = \underline{\underline{-\delta}}$

$$\frac{\partial a}{\partial y} = \underline{\underline{1}}$$

$$\frac{\partial y}{\partial w} = \frac{\partial}{\partial w}(wx+b) = \underline{\underline{x}}$$

So,

$$\frac{\partial L_\delta}{\partial w} = \begin{cases} \delta \cdot 1 \cdot x = \underline{\underline{\delta x}} \; ; & a > \delta \\ -\delta \cdot 1 \cdot x = -\underline{\underline{\delta x}} \; ; & -\delta > a. \end{cases}$$

Now, we find $\frac{\partial L_\delta}{\partial b}$

<u>Case 1</u> : $|a| \le \delta$

$$\frac{\partial L(y,t)}{\partial b} = \frac{\partial H(y-t)}{\partial b} = \frac{\partial H(a)}{\partial b} = \frac{\partial H(a)}{\partial a} \times \frac{\partial a}{\partial y} \times \frac{\partial y}{\partial b}.$$

$$= \underbrace{\frac{\partial H(a)}{\partial a}}_{} \times \underbrace{\frac{\partial a}{\partial y}}_{} \times \frac{\partial y}{\partial b} \longrightarrow \frac{\partial(wx+b)}{\partial b} = \underline{\underline{1}}$$

$$\frac{\partial a}{\partial y} \longrightarrow \frac{\partial(y-t)}{\partial y} = \underline{\underline{1}}$$

$$\frac{\partial}{\partial a}\left(\frac{1}{2}a^2\right) = \frac{1}{2}\frac{\partial}{\partial a}(a^2) = \underline{\underline{a}}$$

So, $\frac{\partial L_\delta}{\partial b} = a \; ; \; |a| \leq \delta.$

<u>case 2</u>

$|a| > \delta$

$\Rightarrow \; -\delta > a > \delta$

$$\frac{\partial L}{\partial b} = \frac{\partial H(a)}{\partial a} \times \frac{\partial a}{\partial y} \times \frac{\partial y}{\partial b}.$$

As calculated earlier.

$$\frac{\partial H(a)}{\partial a} = \begin{cases} \delta \; ; & a > \delta \\ -\delta \; ; & -\delta > a. \end{cases}$$

and $\frac{\partial a}{\partial y} = \underline{1}.$

$$\frac{\partial y}{\partial b} = \frac{\partial(wx+b)}{\partial b} = \underline{1}$$

so,

$$\frac{\partial L}{\partial b} = \begin{cases} \delta \; ; & a > \delta \\ -\delta \; ; & -\delta > a. \end{cases}$$

1.c) Code provided as q1.py.

Q2> Locally weighted Regression

Given: $\{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}$ $\rightarrow$ observation pairs.
where
y is actual value.
x is input value.

$a^{(1)}, \dots, a^{(N)}$ $\rightarrow$ +ve weights.

So, each weight multiplies the corresponding ~~input~~ pair.

$$A = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{NN} \end{bmatrix}$$

diagonal matrix of all the weights.

and
$$w^* = \arg\min \underbrace{\frac{1}{2}\sum_{i=1}^{N} a^{(i)} (y^{(i)} - w^T x^{(i)})^2}_{\text{weighted square loss}} + \underbrace{\frac{\lambda}{2}\|w\|^2}_{\text{Regularization term}}$$

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

Let us take $r = y - Xw$

So, the loss can be written as.

$$= \frac{1}{2}\sum_{i=1}^{N} r_i^2 a^{(i)} + \frac{\lambda}{2}\|w\|^2$$

Now, the i^{th} element of Ar is $a^{(i)} \cdot r_i$. So,

$$r \cdot Ar = r^T \cdot A \cdot r$$

So, loss =

$$\frac{1}{2} r^T A \cdot r + \frac{\lambda}{2}\|w\|^2$$

$$= \frac{1}{2}(y - Xw)^T \cdot A (y - Xw) + \frac{\lambda}{2}\|w\|^2$$; putting $r = y - Xw$.

$$L = \frac{1}{2}(y^T A y - 2 w^T X^T A y + w^T X^T A X w + \lambda w^T w)$$

Now, w is of the form $(w_1, w_2, \dots, w_n)$

So, $w^T \cdot w = w_1^2 + w_2^2 + \dots + w_n^2$

$$= \sum_{i=1}^{n} w_i^2$$

So, $\frac{d}{dw} \sum_{i=1}^{n} w_i^2 = \frac{d}{dw}(w^2) = \underline{\underline{2w}}$

So, $\frac{dL}{dw} = \frac{1}{2}\frac{d}{dw}\left(y^TAy - 2w^TX^TAy + w^TX^TAXw + \lambda w^Tw\right)$

$$= \frac{1}{2}\left\{\frac{d}{dw}\left(\cancelto{0}{y^TAy}\right) - 2\frac{d}{dw}\left(w^TX^TAy\right) + \frac{d}{dw}\left(w^TX^TAXw\right) + \lambda\frac{d}{dw}\left(w^Tw\right)\right.$$

$$= \frac{1}{2}\left\{-2X^TAy + 2.X^TAX.w + 2\lambda w\right\}$$

$$\frac{dL}{dw} = -X^TAy + X^TAXw + \lambda w.$$

for $\frac{dL}{dw} = 0. \Rightarrow w = w^*$

$$0 = -X^TAy + X^TAXw^* + \lambda w^*$$

$$\Rightarrow X^TAy = X^TAXw^* + \lambda w^*$$

$$\Rightarrow X^TAy = \left(X^TAX + \lambda I\right)w^*$$

So, $\underline{\underline{(X^TAX + \lambda I)^{-1}. X^TAy = w^*}} \leftarrow$ Ans.

Q2b) Code in q2.py.

Q2c) As 'τ' increases, the bandwidth of points that should be considered to make a prediction increases.

If 'τ' is very small, you consider only the points which are in the close vicinity of our test point. So, we end up fitting only those selected set of points. This leads to overfitting.

If 'τ' is large, you consider a large set of points when making a prediction. So, you end up with a general curve that fits all of the data but fails to capture irregularities. So, we end up underfitting.

Our training loss should increase as we are adding more points as 'τ' increases. Our validation loss would decrease ~~until~~ as the model considers nearby points that follow a similar trend to the test point. Then, it would increase when other points are considered as 'τ' increases.

Code for q2 is in q2.py.

Q3) **AdaBoost**

target labels $\longrightarrow \{-1, +1\}$ | $h_t \longleftarrow \underset{h \in H}{\text{argmin}} \sum_{i=1}^{N} w_i \, \mathbb{I}\{h(x^{(i)}) \neq t^{(i)}\}$

where $t^{(i)}$ is the true label and $h(x^{(i)})$ is the predicted value.

The Error is given by (I will be using '$\mathcal{E}_t$' instead of 'err_t' to make it easier to read)

$$\mathcal{E}_t = \frac{\sum_{i=1}^{N} w_i \, \mathbb{I}\{h_t(x^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i}$$

$$= \frac{\text{sum of weights of all wrong samples}}{\text{sum of weights of all samples}}.$$

Given $\rightarrow$ $\alpha_t = \frac{1}{2} \log\left(\frac{1-\mathcal{E}_t}{\mathcal{E}_t}\right)$. — (1)

and.

$$w_i' \longleftarrow w_i \cdot e^{-\alpha_t \cdot h_t(x^{(i)}) \cdot t^{(i)}}. \quad \text{— (2)}.$$

Now, $\begin{cases} h_t(x^{(i)}) \cdot t^{(i)} = 1 \text{ ; when } h_t(x^{(i)}) = t^{(i)} \text{ (correct prediction)} \\ h_t(x^{(i)}) \cdot t^{(i)} = -1 \text{ ; when } h_t(x^{(i)}) \neq t^{(i)} \text{ (wrong prediction)}. \end{cases}$

Putting these in (2);

$$w_i' = \begin{cases} w_i \, e^{-\alpha_t} \text{ ; correct prediction} \\ w_i \, e^{\alpha_t} \text{ ; wrong prediction}. \end{cases} \quad \text{— (3)}$$

Substituting the value of α in ③.

$$w_i' = \begin{cases} w_i \cdot e^{-\frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)} & ;\ \text{correct prediction.} \\ w_i \cdot e^{\frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)} & ;\ \text{wrong prediction} \end{cases}$$

Simplifying, we get,

$$\Rightarrow w_i' = \begin{cases} w_i\sqrt{\dfrac{\varepsilon_t}{1-\varepsilon_t}}. & ;\ \text{correct pred.} \\ w_i\sqrt{\dfrac{1-\varepsilon_t}{\varepsilon_t}} & ;\ \text{wrong pred / incorrect pred.} \end{cases} \quad — ⑥$$

We know, $\dfrac{\sum_{\text{Incorrect}} w_i}{\sum_{\text{all}} w_i} = \varepsilon_t. \Rightarrow \dfrac{\sum_{\text{correct}} w_i}{\sum_{\text{all}} w_i} = 1-\varepsilon_t.$

So, $$\sum_{i\in\text{incorrect}} w_i = \varepsilon_t \cdot \sum_{i\in\text{all}} w_i \quad — ④$$

$$\sum_{i\in\text{correct}} w_i = (1-\varepsilon_t)\cdot \sum_{i\in\text{all}} w_i \quad — ⑤$$

~~[illegible]~~.

Finally,

$$\sum_{i\in\text{all}} w_i' = \sum_{i\in\text{incorrect}} w_i' + \sum_{i\in\text{correct}} w_i'$$

putting ⑥ in above equation.

$$\cancel{= w_i\, \mathbb{I}\{h(x^{(i)}) \neq t^{(i)}\}\sqrt{1-\varepsilon_t}}$$

$$= \sum_{i\in\text{incorrect}.} w_i \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} + \sum_{i\in\text{correct}.} w_i \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}.$$

putting ④ & ⑤ in above equation.

$$\sum_{i \in all} w_i' = \epsilon_t \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \cdot \sum_{i \in all} w_i + (1-\epsilon_t) \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} \cdot \sum_{i \in all} w_i .$$

$$\sum_{i \in all} w_i' = \sqrt{\epsilon_t \cdot (1-\epsilon_t)} \sum_{i \in all} w_i + \sqrt{\epsilon_t (1-\epsilon_t)} \cdot \sum_{i \in all} w_i$$

$$\boxed{\sum_{i \in all} w_i' = 2\sqrt{\epsilon_t (1-\epsilon_t)} \cdot \sum_{i \in all} w_i .}$$

$$\sum_{i \in incorrect} w_i' = \sum_{i \in incorrect} w_i \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$$

$$= \epsilon_t \cdot \sum_{i \in all} w_i \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} = \sqrt{\epsilon_t (1-\epsilon_t)} \cdot \sum_{i \in all} w_i$$

so, $$\boxed{\sum_{i \in incorrect} w_i' = \sqrt{\epsilon_t (1-\epsilon_t)} \sum_{i \in all} w_i}$$

---

finally, $$\epsilon_t' = \frac{\sum_{i \in incorrect} w_i'}{\sum_{i \in all} w_i'}$$

plugging all the values as found above:

$$\epsilon_t' = \frac{\sqrt{\epsilon_t (1-\epsilon_t)} \cdot \sum_{i \in all} w_i}{2\sqrt{\epsilon_t (1-\epsilon_t)} \cdot \sum_{i \in all} w_i} = \frac{1}{2}$$

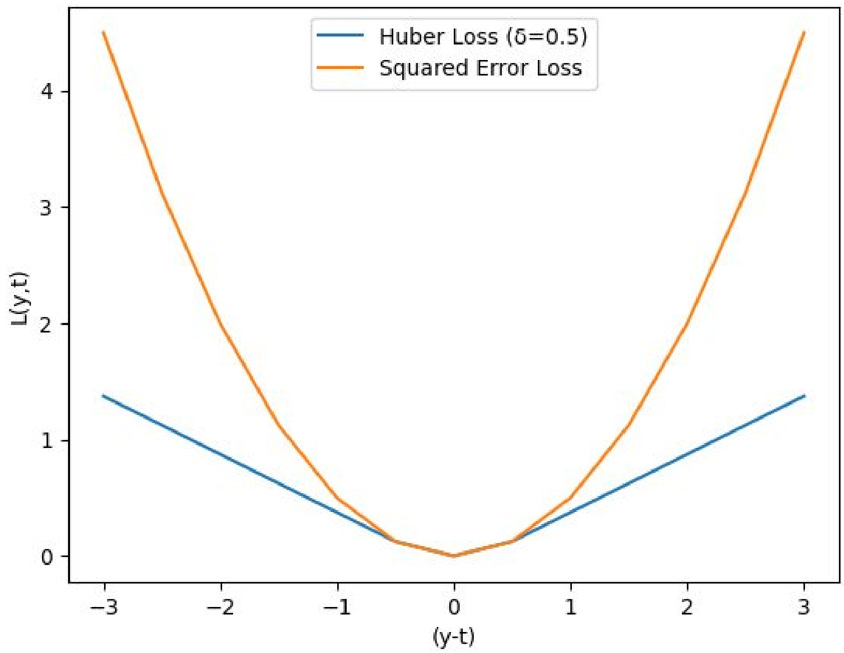
So, $\boxed{\epsilon_t' = \frac{1}{2}}$ (hence proved).

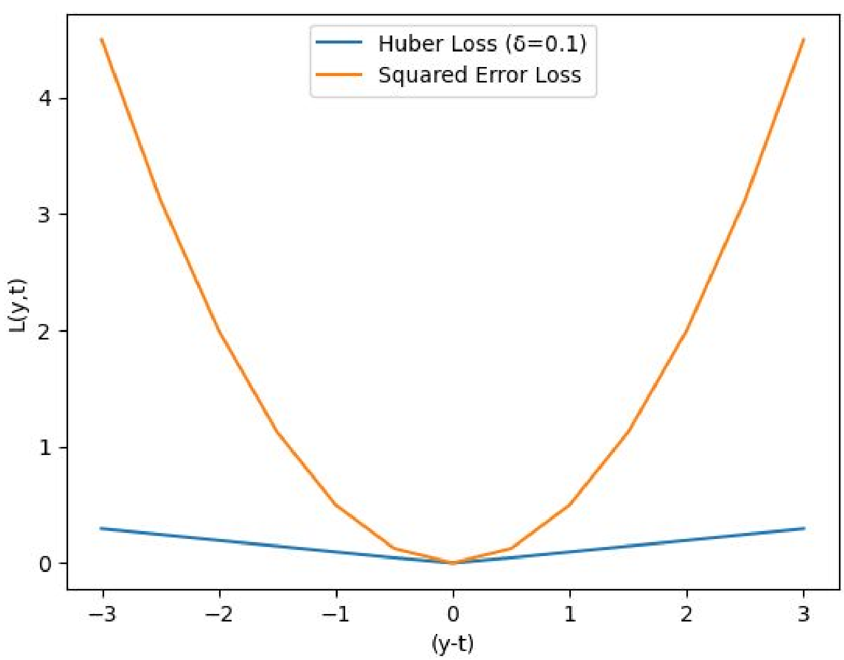
Interpretation

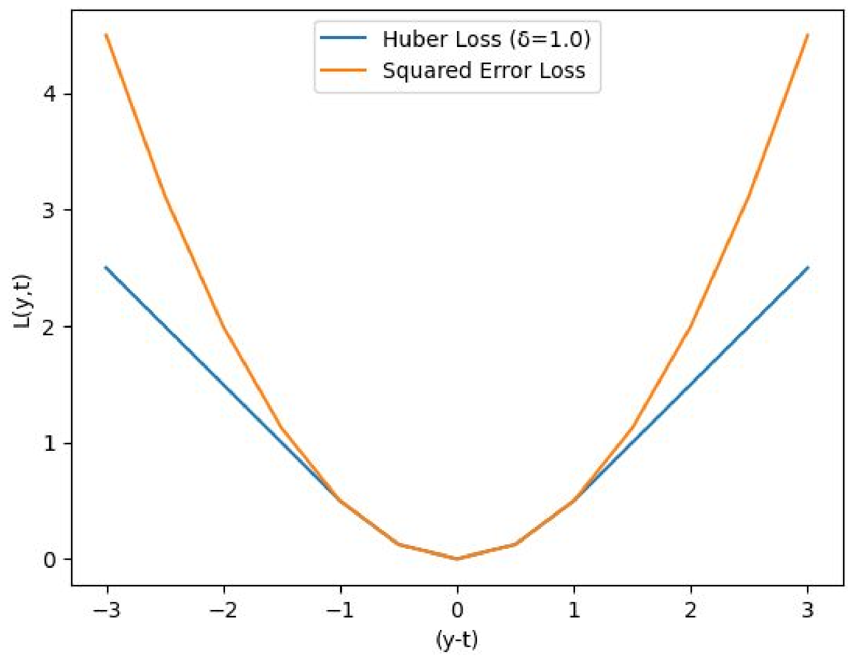
$$\epsilon_t' = \frac{\sum_{i \in \text{incorrect}} w_i'}{\sum_{i \in \text{all}} w_i'} = \frac{1}{2}$$

This means that the sum of all correct sample weights is equal to the sum of all incorrect sample weights. So, if we have weights w_i in the t^{th} time step, we will scale the weights at the $t+1^{th}$ time step, so, that ~~weights~~ sum of all incorrect samples and correct samples are equal.

Q1.a. Plots :


Huber Loss (δ=0.5)
Squared Error Loss
L(y,t)
(y-t)


Huber Loss (δ=0.1)
Squared Error Loss
L(y,t)
(y-t)



Q2c. Plot: You can also generate the plot by running the code in q2.py