HW 2

Haoyi Shi

①. $L(y_i | x_i, w) = y_i \log(\sigma(w^T x_i)) + (1-y)\log(\sigma(-w^T x_i))$

Derive the gradient of $L(y_i | x_i w)$ respect to w_j

According to the derivative of sigmoid function

$$\sigma'(x) = \sigma(x)(1-\sigma(x))$$

∵ For first assume $z = \sigma(w^T x_i)$

∴ $$\frac{\partial}{\partial w_j} \log(\sigma(w^T x_i)) = \frac{\partial}{\partial z}\frac{\partial z}{\partial w_j}$$ Chain rule

$$= \frac{1}{z} \cdot \sigma(w^T x_i) \cdot (1-\sigma(w^T x_i)) \cdot \frac{\sigma w^T x_i}{\sigma w_j}$$

$$= (1-\sigma(w^T x_i)) \cdot x_{ij}$$

For second part using the same way

$$(1-y)\log \sigma(-w^T x_i) \frac{\partial}{\partial w_j}$$

$$= (1-y) \frac{1}{\sigma(-w^T x_i)} \sigma(-w^T x_i) \cdot (1-\sigma(-w^T x_i)) \cdot -x_i$$

$$= (1-y) \cdot (1-\sigma(-w^T x_i)) \cdot -x_{ij}$$

$$\therefore \frac{\partial L(y_i | x_i w)}{\partial w_j} = y(1-\sigma(w^T x) \cdot x_{ij} + (1-y) \cdot (1-\sigma(-w^T x_i)) \cdot -x_{ij}$$

$$= y(1-\sigma(w^T x) \cdot x_{ij} + (1-y)\, \sigma(w^T x_i) \cdot -x_{ij}$$

$$= x_{ij}\,(y - y\cancel{\sigma(w^T x)} - \sigma(w^T x_i) + y\cancel{\sigma(w^T x)})$$

$$= x_{ij}\,(y - \sigma(w^T x_i))$$

(13) $f(w) = \frac{1}{2}\|W\|_2^2 + C\sum_{i=1}^{n} \max(0, 1-y_i(w^T x_i + b))$

find $\frac{\partial}{\partial w_j} f(w)$

for gradient of hinge Loss

if $L = 0$, then the gradient is 0

if $L > 0$,

$$1-y_i(w^T x_i + b)\frac{\partial}{\partial w}$$

$$= -x_{ij} y_i$$

$$\therefore \frac{\partial f(w)}{\partial w} = W_j - C\sum y_i x_{ij}$$ if hinge loss greater than 0

$$\frac{\partial f w}{\partial w} = W_j$$ if hinge loss is equal to 0

HW2

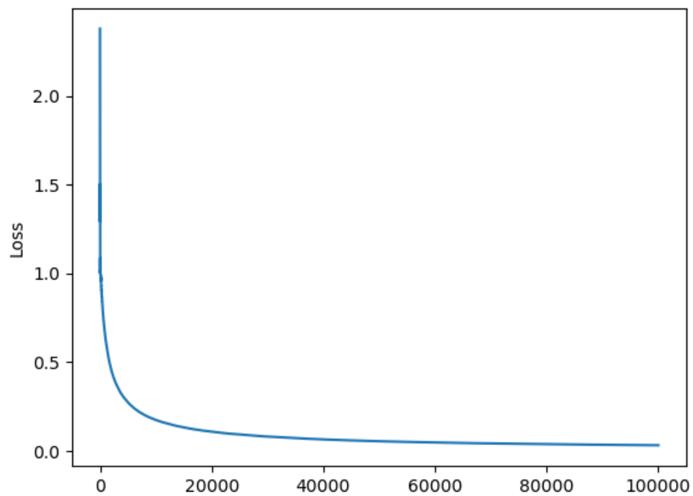
Question2

- Logistic Regression

eta_value	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Mean	Std
0.001	0.0.0125	0.00625	0.0125	0.0125	0.01875	0.0125	0.00625	0	0.00625	0.03125	0.0118	0.008125
0.01	0.01875	0.0125	0.0125	0.0125	0.01875	0.0125	0.0125	0	0.01252	0.03125	0.01437	0.007421
0.1	0.025	0.00625	0.0125	0.01875	0.01875	0.01875	0	0.0125	0.0375	0.01875	0.016875	0.0097

Which value of η is optimal?

- Best η for **logistic regression** with eta 0.001: MSE is 0.005



According to training average loss plot, the loss keep going lower until 5000 iteration. If I using Gradient Decent, the loss will be more smoother.

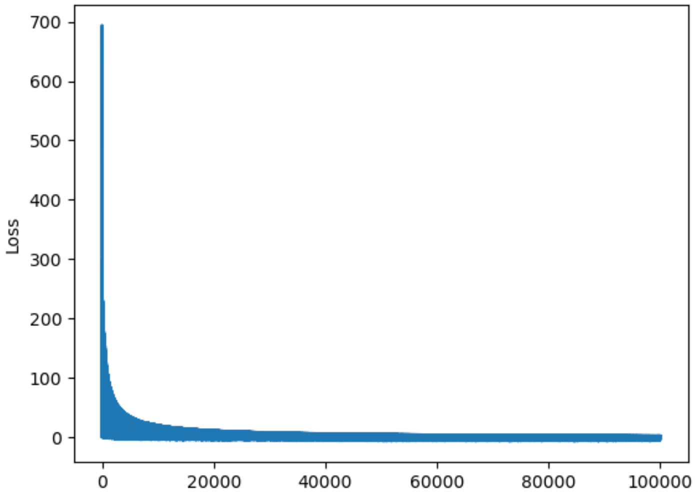
Question4

- SVM

C	Learning Rate	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Mean	Std
1	1e-05	0.075	0.025	0.05	0.05	0.075	0.075	0.075	0	0.075	0.15	0.065	0.03741657
10	1e-05	0.075	0.05	0.05	0.05	0.075	0.075	0.075	0	0.025	0.125	0.06	0.03201562
100	1e-05	0.05	0.025	0.05	0.025	0.075	0.05	0	0	0.05	0.125	0.0575	0.03172144
1	0.0001	0.075	0.025	0.05	0.05	0.075	0.075	0.075	0	0.05	0.15	0.0625	0.0375
10	0.0001	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0	0.05	0.125	0.0525	0.02839454
100	0.0001	0.05	0.075	0.05	0.05	0.05	0.05	0.075	0	0.05	0.125	0.0575	0.02968586
1	0.001	0.05	0.025	0.025	0.05	0.1	0.075	0.075	0.05	0.075	0.15	0.0675	0.03544362
10	0.001	0.075	0.025	0.05	0.05	0.075	0.075	0.05	0	0.025	0.125	0.055	0.03316625
100	0.001	0.1	0.075	0.005	0.025	0.05	0.075	0.075	0.025	0.075	0.175	0.0725	0.041

Which value of η and C are optimal?

- Best η for **logistic regression** with eta 0.00001 and C 100: MSE is 0.02 and zero_one loss is 0.005



According to training average loss plot, the loss keep going lower until 2000 ~ 4000 iteration. Also, the training loss is very noisy. If I using Gradient Decent, the loss will be more smoother.