Machine learning HW3

B11901174 傅啟恩

y_n \ $\hat{y}$	+1	-1
+1	0	1
-1	1000	0

1. avg error

$$\hat{y} = \begin{cases} +1 & p(-1|x) \times 1 \\ -1 & p(+1|x) \times 1000 \end{cases}$$

$p(-1|x) + p(+1|x) = 1$

$p(-1|x) = 1 - p(+1|x)$

If $(1 - p(+1|x)) \times 1 < p(+1|x) \times 1000$

$\Leftrightarrow 1001\, p(+1|x) > 1$

$\Rightarrow$ ideal mini-target $p(+1|x) > \alpha$, $f_{\text{ideal}}(x) = +1$

$f_{\text{ideal}}(x) = \text{sign}(p(+1|x) - \alpha)$.

otherwise $f_{\text{ideal}}(x) = -1$

$\alpha = \frac{1}{1001}$ #

2. $E_{(x,y)\sim p(x,y)}(g(x) \neq y) = E_{x\sim p(x)}(g(x) = f(x)) \cdot p(y = -f(x)|x)$

$E_{out}(g)$ ← $+ E_{x\sim p(x)}(g(x) \neq f(x)) \cdot p(y = f(x)|x)$

$= \underline{(1 - E_{out}(g)) \cdot \epsilon + E_{out}(g)(1 - \epsilon)}$ #

3. $E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}(h(x_n) - y_n)^2 = \frac{1}{N}\sum_{n=1}^{N}(wx_n - y_n)^2$

$= \frac{1}{N}\sum_{n=1}^{N}(w^2x_n^2 - 2wx_ny_n + y_n^2)$

$= \frac{1}{N}\left(w^2\sum_{n=1}^{N}x_n^2 - 2w\sum_{n=1}^{N}x_ny_n + \sum_{n=1}^{N}y_n^2\right)$

$\nabla E_{in}(w) = \frac{1}{N}\left(2w\sum_{n=1}^{N}x_n^2 - 2\sum_{n=1}^{N}x_ny_n\right) \overset{\text{set}}{=} 0$

optimal w $\quad w_{LIN} = \dfrac{\sum_{n=1}^{N}x_ny_n}{\sum_{n=1}^{N}x_n^2}$ ※

4. $E_{in}(w) = \frac{1}{N}\int_0^1 ((w_0 + w_1x) - (ax^2 + b))^2\, dx$

$(N=1)$ $= \frac{1}{N}\int_0^1 (-ax^2 + w_1x + (w_0 - b))^2\, dx$

$= \frac{1}{N}\int_0^1 (a^2x^4 - 2aw_1x^3 - 2a(w_0 - b)x^2 + w_1^2x^2 + 2w_1(w_0 - b)x + (w_0 - b)^2)\, dx$

$$= a^2\left(\frac{1}{N}\int_0^1 x^4\,dx\right) - 2aw_1\left(\frac{1}{N}\int_0^1 x^3\,dx\right) + \left(-2a(w_0-b)+w_1^2\right)\left(\frac{1}{N}\int_0^1 x^2\,dx\right)$$

$$+ 2w_1(w_0-b)\left(\frac{1}{N}\int_0^1 x\,dx\right) + (w_0-b)^2 \qquad x \sim \text{UNIF}(0,1)$$

$$= \frac{1}{5}a^2 - \frac{1}{2}aw_1 + \left(-2a(w_0-b)+w_1^2\right)\cdot\frac{1}{3} + w_1(w_0-b) + (w_0-b)^2$$

$$\frac{\partial E_{in}}{\partial w_0} = w_1 - \frac{2}{3}a + 2(w_0-b) = 0 \qquad 2w_1 + 4w_0 = \frac{4}{3}a + 4b$$

$$\frac{\partial E_{in}}{\partial w_1} = \frac{2}{3}w_1 - \frac{1}{2}a + w_0 = 0 \qquad 2w_1 + 3w_0 = \frac{3}{2}a$$

$$\begin{cases} w_0^* = \frac{-1}{6}a + 4b \\ w_1^* = \frac{1}{3}a - 6b \end{cases} \#$$

5. $y_n' = ay_n + b \qquad X^T y = w_{LIN}$

$$w_{LIN}' = a w_{LIN} + \underbrace{(X^TX)^{-1}X^T\hat{b}} \qquad \text{we know that } \hat{b} = \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix}$$

and $X = \begin{bmatrix} 1 & -x_1^T- \\ 1 & -x_2^T- \\ \vdots & \\ 1 & -x_N^T- \end{bmatrix}$

$\Leftrightarrow X^TX w_b - X^Tb = 0$,

$\Leftrightarrow X^T(Xw_b - \hat{b}) = 0$, $w_b = (X^TX)^{-1}X^Tb$

$$= \begin{bmatrix} b \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \begin{matrix} w_b[0] = b & i=0 \\ w_b[i] = 0 & i \neq 0 \end{matrix}$$

$$w_{LIN}' = a w_{LIN} + \left(w_b = \begin{bmatrix} b \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right) \#$$

6. $\nabla E_{in}(w_t) = \frac{1}{N}\sum_n \frac{1}{1+\exp(y_n w_t^T x_n)}(-y_n x_n)$

$$A_{i,i} = \frac{\partial(\nabla E_{in}(w_t)\,i)}{\partial w_i} = \frac{1}{N}\sum_n \left(\frac{+1}{1+\exp(y_n w_t^T x_n)}\right)^2 y_n^2 (x_{n,i})^2 \exp(y_n w_t^T x_n)$$

$$= \frac{1}{N}\sum_n (h_t(y_n x_n))^2 \exp(y_n w_t^T x_n)\cdot y_n^2 (x_{n,i})(x_{n,i})$$

$$A_{i,j} = \frac{1}{N}\sum_{n=1}^{N}[h_t(y_n x_n)]^2 \exp(y_n \hat{w}_t^T \hat{x}_n)\cdot y_n^2 (x_{ni})(x_{nj})$$

$$= \sum_{n=1}^{N}(X_{:,n})^T (x_{n,j}) D_{nn}, \qquad D_{nn} = \begin{bmatrix} D_{11} & \cdots & 0 \\ \vdots & D_{22} & \vdots \\ \vdots & & \ddots \\ 0 & \cdots & D_{NN} \end{bmatrix}$$

$$A = X^T D X$$

$$D_{i,j} = \begin{cases} \frac{1}{N}\left[h_t(y_n x_n)\right]^2 \exp\left(y_n \hat{w}_t^T \hat{x}_n\right) y_n^2 & i = j \\ 0 & i \neq j \end{cases} \quad \#$$

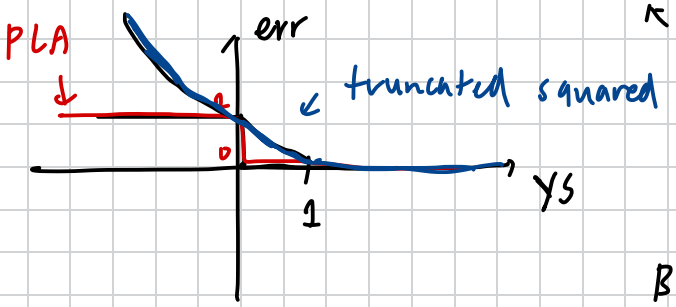
7. $err(s, y) = \begin{cases} (1-sy)^2 & \text{if } sy \leq 1 \\ 0 & \text{otherwise} \end{cases}$

Applying SGD ← truncated squared loss

$$w_{t+1} \leftarrow w_t + \eta\left(1(ys \leq 1)\, 2(y_n - w_t^T x_n) x_n\right)$$

$$w_{t+1} \leftarrow w_t + \eta\left(1(ys \leq 0)\, y_n x_n\right)$$

↖ PLA



Compared to PLA, the error is the same $=0$ if $ys = yw^T x \geq 1$

But if $ys = yw^T x \leq 1$, the truncated squared loss would be the upper bound of PLA method #

8. Find $\nabla E_{in}(w)$

$$W = \begin{bmatrix} | & | & & | \\ w_1 & w_2 & \cdots & w_K \\ | & | & & | \end{bmatrix}_{(d+1)\times K} \quad \begin{matrix} \leftarrow w_{y_n} \\ \uparrow w_{i,j} \end{matrix}$$

$$h_y(x) = \frac{\exp(w_y^T x)}{\sum_{i=1}^{K} \exp(w_i^T x)}$$

and $E_{in}(\underset{\text{matrix}}{w}) = \frac{1}{N}\sum_n err(w, x_n, y_n) = \frac{1}{N}\sum_n -\ln h_{y_n}(x_n)$

$$= \frac{-1}{N}\left(\ln h_{y_1}(x_1) + \ln h_{y_2}(x_2) + \cdots + \ln h_{y_N}(x_N)\right)$$

$$\ln h_{y_n}(x_n) = \left(w_{y_n}^T x_n\right) - \ln\left(\sum_{i=1}^{K} \exp(w_i^T x_n)\right)$$

$\frac{\exp(w_{ij} x_{n,i})}{\exp(w_1^T x_n) + \cdots + \exp(w_K^T x_n)}$

$w_{11} x_n + \cdots + w_{1K} x_n$

First solve $\frac{\partial E_{in}(w)}{\partial w_{i,j}}$ ← element / entry

$$\frac{\partial \left(\ln h_{y_n}(x_n)\right)}{\partial w_{i,j}} = \begin{cases} x_{n,i} - \frac{1}{\sum_{i=1}^{K} \exp(w_i^T x_n)} \exp(w_{ij} x_{n,j}) \cdot x_{n,j} & \text{if } j = y_n \end{cases}$$

$$\frac{\partial E_{in}(w)}{\partial w_{ij}} = \begin{cases} -\frac{1}{\sum_{i=1}^{K} \exp(w_i^T x_n)} \exp(w_{ij} x_{n,i})\, x_{n,i} & \text{if } j \neq y_n \\ \text{as above} & \text{if } j = y_n \end{cases}$$

$$\frac{\partial E_{in}(w)}{\partial w} = \begin{bmatrix} | & | & | & | \\ \frac{\partial E_{in}(w)}{\partial w_{0,1}} & \frac{\partial E_{in}(w)}{\partial w_{0,2}} & \cdots & \frac{\partial E_{in}(w)}{\partial w_{0,K}} \\ | & | & | & | \\ \frac{\partial E_{in}(w)}{\partial w_{1,1}} & \frac{\partial E_{in}(w)}{\partial w_{1,2}} & \cdots & \frac{\partial E_{in}(w)}{\partial w_{1,K}} \\ \vdots & \vdots & \ddots & \vdots \\ | & | & | & | \\ \frac{\partial E_{in}(w)}{\partial w_{d,1}} & \frac{\partial E_{in}(w)}{\partial w_{d,2}} & \cdots & \frac{\partial E_{in}(w)}{\partial w_{d,K}} \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ w_1 & w_2 & \cdots & w_K \\ | & | & | & | \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \ddots & \\ 0 & & & \end{bmatrix}$$

each element ↑ is as above

$i = [0, d+1)$, $j = [0, K)$, y_n

$(d+1) \times K$ #

13. By problem 6

$$A = X^T D X = \tilde{X}^T \tilde{X} = \sum_{n=1}^{N} (\tilde{x}_{i,n})^T (\tilde{x}_{n,j})$$

$$A_{i,j} = \sum_{n=1}^{N} (x_{i,n})^T (x_{n,j})\, D_{nn}$$

⇒ compare $\tilde{X}^T \tilde{X}$ and $X^T D X$, we can know that

$$(\tilde{x}_{i,n}) = (x_{i,n}) \sqrt{D_{nn}}, \quad (\tilde{x}_{n,j}) = (x_{n,j}) \sqrt{D_{nn}},$$

where $\sqrt{D_{nn}} = \begin{bmatrix} \sqrt{D_{11}} & \cdots & 0 \\ \vdots & \sqrt{D_{22}} & \vdots \\ 0 & \cdots & \sqrt{D_{NN}} \end{bmatrix}$

$$\boxed{\tilde{X} = X \cdot \sqrt{D}}$$

$\sqrt{\frac{1}{N}} \cdot \sqrt{\exp} \cdot ht \cdot y_n$

$$\sqrt{D_{i,j}} = \begin{cases} \sqrt{\frac{1}{N} [ht(y_n x_n)]^2 \exp(y_n \hat{w}_t^T \hat{x}_n)\, y_n^2} & i = j \\ 0 & i \neq j \end{cases} \quad *$$

$X^T (y_n \cdot \quad y \cdot (X w_t)$

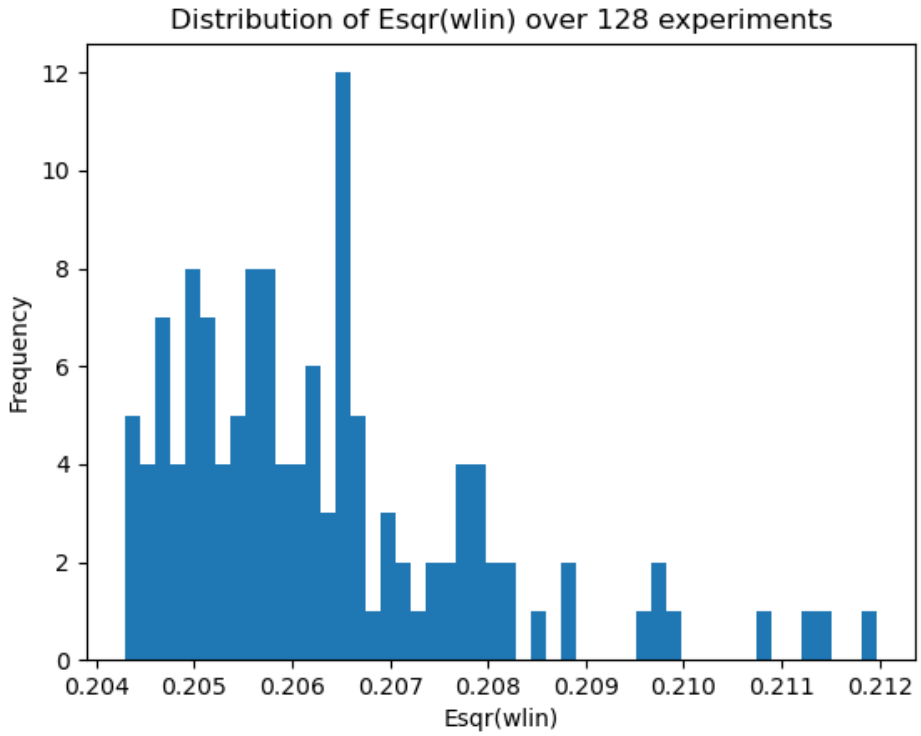
and $\frac{+1}{N} \sum_n \frac{1}{1 + \exp(y_n w_t^T x_n)} (+y_n \hat{x}_n) = -\nabla E_{in}(w_t) = \tilde{X}^T \hat{y}$

$$-\nabla E_{in}(w_t)_i = \sum_n \tilde{x}_{i,n} \tilde{y}_n = \sum_n (x_{i,n}) \sqrt{D_{nn}}\, \tilde{y}_n$$

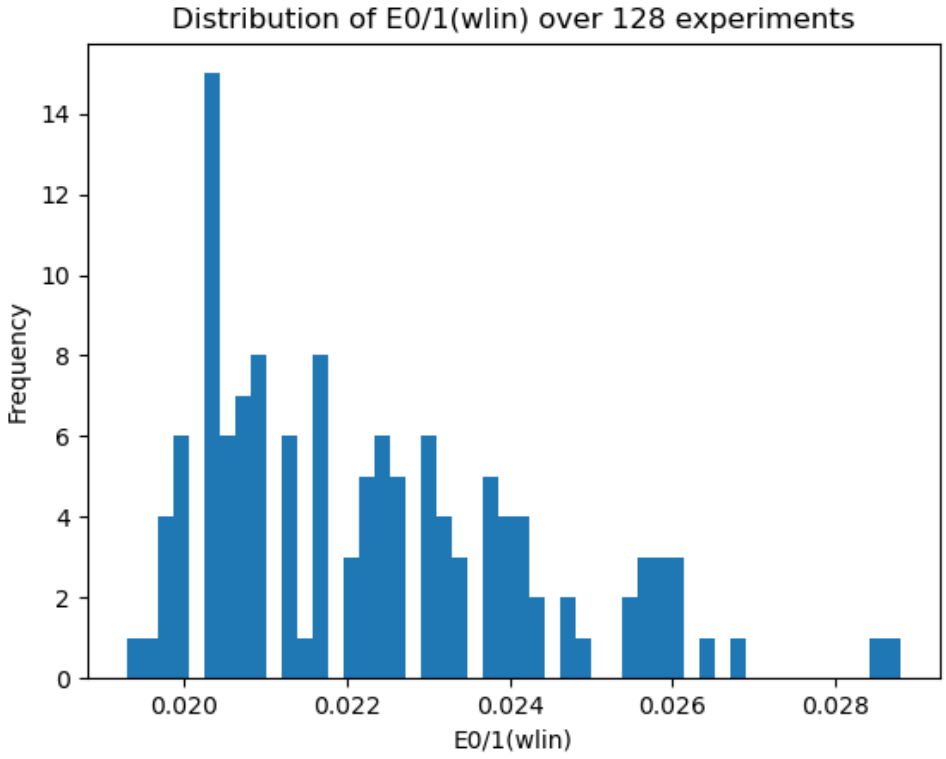
$$\tilde{y}_n = \sqrt{\frac{1}{N}} \frac{1}{\sqrt{\exp(y_n w_t^T x_n)}}, \quad \tilde{y} = \sqrt{\frac{1}{N}} \cdot \frac{1}{\sqrt{\exp(y \cdot X w_t)}} \quad *$$

array matrix

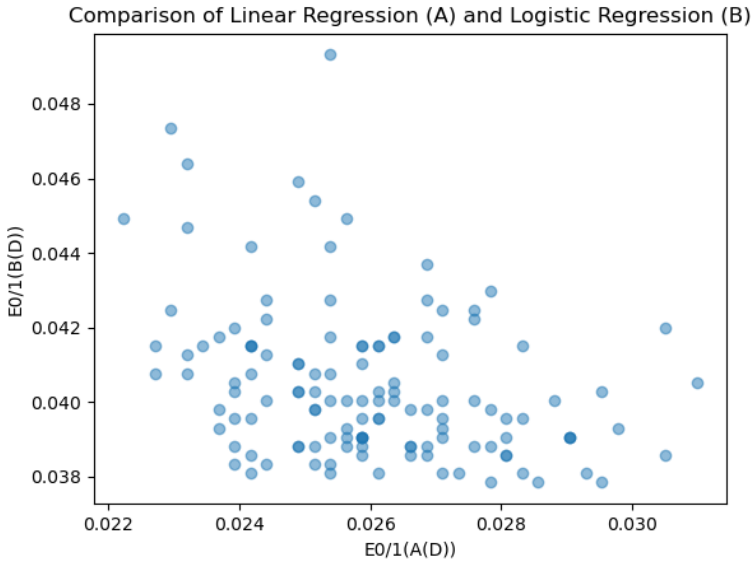
9. Median Esqr over 128 experiments: 0.20595399486341354



10. Median E0/1 over 128 experiments: 0.02197265625



11.
Median E0/1(A(D)) over 128 experiments: 0.022216796875
Median E0/1(B(D)) over 128 experiments: 0.0357666015625



12.
Median E0/1(A(D)) over 128 experiments: 0.0350341796875
Median E0/1(B(D)) over 128 experiments: 0.038818359375
We see that as we add outlier examples, the error rate is getting higher for linear regression and for logistic regression, which is due to new data with different mean and covariance being added. Compared to two types of regressions, we see that linear regression changes more obviously than logistic one.