1. We know setting weight of false positive to 1000 is the same as copying -1 examples 1000 times (from lecture), while using E_{in}^{w}.
For any x where $P(+1|x) = \frac{a}{a+b}$ ($a = P(+1|x)$, $b = P(-1|x)$)

		h(x)	
		+1	-1
y	+1	0	1
	-1	1000	0

becomes $P'(+1|x) = \frac{a}{a+1000b}$ after copy -1 examples 1000 times
then $f_{LIN}(x) = \text{sign}(P'(+1|x) - \frac{1}{2})$ since we're using E_{in}^{w}
now we know
if $\frac{a}{a+1000b} \geq \frac{1}{2}$, f_{LIN} recognize it as +1

$\Rightarrow 2a \geq a+1000b \Rightarrow a \geq 1000b \Rightarrow a+b \geq 1001b \Rightarrow \frac{1}{1001} \geq \frac{b}{a+b} \Rightarrow \frac{a}{a+b} \geq \frac{1000}{1001}$

which is to say if $\frac{a}{a+b} \geq \frac{1000}{1001}$, recognize it as +1.

$f_{LIN}(x) = \text{sign}(P(+1|x) - \frac{1000}{1001})$, $\alpha = \frac{1000}{1001}$ ※

2.

		g	
		+1	-1
f	+1	a	b
	-1	c	d

Suppose count of TP = a, TN = d, FP = c, FN = b.
we get $E_{out}(g) = \frac{a+d}{a+b+c+d}$

		g	
		+1	-1
y	+1	$a(1-t)+ct$	$b(1-t)+dt$
	-1	$at+c(1-t)$	$bt+d(1-t)$

Since we know for Y, there's t percent that it flips.
For a, there will be ta data become FP, (1-t) stays TP. The whole data will become the right chart.

$$E_{(x,y)\sim P(x,y)}[|g(x) \neq y|] = \frac{(1-t)(a+d)+(b+c)t}{a+b+c+d}$$

$$= (1-t)E_{out}(g) + (1-E_{out}(g))t$$

$$= E_{out}(g) - 2tE_{out}(g) + t$$

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

$$\frac{dE_{in}(w)}{dw} = \frac{1}{N}\sum_{n=1}^{N} 2(wx_n - y_n)x_n = 0$$

$$\Rightarrow \sum_{n=1}^{N} wx_n^2 - y_n x_n = 0$$

$$\Rightarrow w\sum_{n=1}^{N} x_n^2 = \sum_{n=1}^{N} y_n x_n$$

$$\Rightarrow w = \frac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} x_n^2}$$

4.

$$E(h) = \int_0^1 (h(x) - f(x))^2 dx \qquad h(x) = w_0 + w_1 x \;,\; f(x) = ax^2 + b$$

$$= \int_0^1 h(x)^2 - 2h(x)f(x) + f(x)^2 \, dx$$

$$= \int_0^1 \left(w_0^2 + 2w_0 w_1 x + w_1^2 x^2\right) - 2(w_0 + w_1 x)(ax^2 + b) + \left(a^2x^4 + 2abx^2 + b^2\right) dx$$

$$= \int_0^1 w_0^2 + 2w_0 w_1 x + w_1^2 x^2 - 2w_0 a x^2 - 2w_1 a x^3 - 2w_0 b - 2w_1 b x + a^2 x^4 + 2abx^2 + b^2 \, dx$$

$$= w_0^2 x + w_0 w_1 x^2 + \frac{1}{3}w_1^2 x^3 - \frac{2}{3}w_0 a x^3 - \frac{1}{2}w_1 a x^4 - 2w_0 b x - w_1 b x^2 + \frac{1}{5}a^2 x^5 + \frac{2}{3}abx^3 + b^2 x \Big|_0^1$$

$$= w_0^2 + w_0 w_1 + \frac{1}{3}w_1^2 - \frac{2}{3}w_0 a - \frac{1}{2}w_1 a - 2w_0 b - w_1 b + \frac{1}{5}a^2 + \frac{2}{3}ab + b^2$$

$$\frac{dE(h)}{dw_0} = 2w_0 + w_1 - \frac{2}{3}a - 2b = 0 \Rightarrow 4w_0 + 2w_1 = \frac{4}{3}a + 4b \quad \cdots ①$$

$$\frac{dE(h)}{dw_1} = w_0 + \frac{2}{3}w_1 - \frac{1}{2}a - b = 0 \Rightarrow 3w_0 + 2w_1 = \frac{3}{2}a + 3b \quad \cdots ②$$

①-②

$$w_0 = -\frac{1}{6}a + b$$

$$w_1 = \frac{2}{3}a + 2b - 2w_0 = \frac{2}{3}a + 2b + \frac{1}{3}a - 2b = a$$

XX

5. need to $\min_W E_{in}(w) = \frac{1}{N}\|Xw - Y'\|^2$

$= \frac{1}{N}\|Xw - (aY + B)\|^2$ where B is an $N \times 1$ matrix filled with b

For $E_{in}(w_0) = \frac{1}{N}\|\sum_{n=1}^{N} x_{n,0} \cdot w_0 - (ay_0 + b)\|^2$

$= \frac{1}{N}\left(w_0^2 \underbrace{\left(\sum_{n=1}^{N} x_{n,0}\right)^2}_{=1} - 2(ay_0+b)\underbrace{\left(\sum_{n=1}^{N} x_{n,0}\right)}_{=1} \cdot w_0 + (ay_0+b)^2\right)$

$\frac{dE_{in}(w_0)}{dw_0} = \frac{1}{N}\left(2w_0 - 2(ay_0+b)\right) = 0 \Rightarrow w_0 = ay_0 + b$

Then $E_{in}(W) = \frac{1}{N}\|XW - (aY+B)\|^2$

(insert $w_0 = ay_0 + b$) $= \frac{1}{N}\left\| X \cdot \begin{bmatrix} ay_0+b \\ w_1 \\ \vdots \\ w_d \end{bmatrix} - aY - B \right\|^2$

$= \frac{1}{N}\left\| X \cdot \begin{bmatrix} ay_0 \\ w_1 \\ \vdots \\ w_N \end{bmatrix} + X \cdot \begin{bmatrix} b \\ 0 \\ \vdots \\ 0 \end{bmatrix} - aY - B \right\|^2$

↳ first column of X is 1

$= \frac{1}{N}\left\| X \cdot \begin{bmatrix} ay_0 \\ w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} + B - aY - B \right\|^2$

(set $\begin{bmatrix} ay_0 \\ w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = W_{tmp}$) $= \frac{1}{N}\left\| X \cdot \begin{bmatrix} ay_0 \\ w_1 \\ \vdots \\ w_N \end{bmatrix} - aY \right\|^2$

$= \frac{1}{N}\|X \cdot W_{tmp} - aY\|^2$

$= \frac{1}{N}\|X^T X W_{tmp}^2 - 2X^T aY + a^2 Y^T Y\|$

$\nabla E_{in}(W) = \frac{2}{N}(X^T X W_{tmp} - X^T aY) = 0$

$\Rightarrow X^T X W_{tmp} = X^T aY$

(since $X^T X$ is inverible) $\Rightarrow W_{tmp} = a(X^T X)^{-1} X^T Y = a W_{LIN}$

then $W = W_{tmp} + \begin{bmatrix} b \\ 0 \\ \vdots \\ 0 \end{bmatrix} = a W_{LIN} + b \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, size: $(d+1) \times 1$

$W_{LIN}' = a W_{LIN} + bI$ (I is a $(d+1) \times 1$ matrix, where the first element is 1, rest of it is 0)

b.

$$A_E(\mathbf{w}) = \begin{bmatrix} \frac{\partial^2 E}{\partial w_1^2}(\mathbf{w}) & \frac{\partial^2 E}{\partial w_1 \partial w_2}(\mathbf{w}) & \cdots & \frac{\partial^2 E}{\partial w_1 \partial w_d}(\mathbf{w}) \\ \frac{\partial^2 E}{\partial w_2 \partial w_1}(\mathbf{w}) & \frac{\partial^2 E}{\partial w_2^2}(\mathbf{w}) & \cdots & \frac{\partial^2 E}{\partial w_2 \partial w_d}(\mathbf{w}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 E}{\partial w_d \partial w_1}(\mathbf{w}) & \frac{\partial^2 E}{\partial w_d \partial w_2}(\mathbf{w}) & \cdots & \frac{\partial^2 E}{\partial w_d^2}(\mathbf{w}) \end{bmatrix}_{d \times d}$$

$$\theta(x) = \frac{1}{1+e^{-x}}$$

first, we know $\frac{\partial E_{in}(w)}{\partial w_i} = \frac{1}{N}\sum_{n=1}^{N} \theta(-y_n w^T x_n)(-y_n x_{n,i})$ from lecture notes.

$$\Rightarrow \nabla E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} \theta(-y_n w^T x_n)(-y_n x_n)$$

$$\frac{\partial^2 E_{in}(w)}{\partial w_i \partial w_j} = \frac{1}{N}\sum_{n=1}^{N} (-y_n x_{n,i})\left(-\theta(y_n w^T x_n)\theta(-y_n w^T x_n)\, y_n x_{n,j}\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N} \left(x_{n,i}\; x_{n,j}\right)\left(\theta(y_n w^T x_n)\theta(-y_n w^T x_n)\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N} \left(x_{n,i}\, x_{n,j}\right)\left(h(y_n x_n)\, h(-y_n x_n)\right)$$

$$\frac{\partial \theta(-y_n w^T x_n)}{\partial w_j} = \frac{\partial}{\partial w_j}\frac{1}{1+e^{y_n w^T x_n}}$$

$$= -(1+e^{y_n w^T x_n})^{-2} \times e^{y_n x_{n,j}} \times y_n x_{n,j}$$

$$= -\frac{1}{(1+e^{y_n w^T x_n})}\frac{e^{y_n w^T x_n}}{(1+e^{y_n w^T x_n})} \cdot y_n x_{n,j}$$

$$= -\theta(y_n w^T x_n)\theta(-y_n w^T x_n) y_n x_{n,j}$$

then Hessian matrix become $X^T D X$ where D is a diagnal matrix where

$$D = \begin{pmatrix} \frac{1}{N}h(y_1 x_1)h(-y_1 x_1) & & 0 \\ & \ddots & \\ 0 & & \frac{1}{N}h(y_N x_N)h(-y_N x_N) \end{pmatrix}$$

then $D_{nn} = \frac{1}{N} h(y_n x_n) h(-y_n x_n)$ #

7. For SGD.

For $1-y_s > 0$

$err(s, y) = (1 - y w^T x)^2$

$\nabla err(w_t, x_n, y_n) = 2(1 - y_n w_t x_n)(-y_n x_n)$

$w_{t+1} \leftarrow w_t + \eta(-\nabla_w) = w_t + 2\eta(1 - y_n w_t x_n)(y_n x_n)$

For PLA, if $1 - y_s > 0$

$w_{t+1} \leftarrow w_t + y_n x_n$

We can see both algorithms updates when $1 - y_s > 0$, which is to say when $y_n \neq \text{sign}(s)$, the algorithm will update the weights. The weights for both algorithms will be updated according to $y_n x_n$, but with different value multiplied in front.

The difference between two algorithm is they update w_{n+1} with different weights in front of $y_n x_n$.

Moreover, we can see SGD is the softer version of PLA since $1 > 1 - y_s > 0 \Rightarrow \eta > 2\eta(1 - y_s) > 0$, when $\eta < \frac{1}{2}$, SGD will be soft version of PLA.

8.

$$\frac{1}{N} err(W, x, y) = \frac{-1}{N} \ln\left(e^{W_y^T x} \Big/ \sum_{i=1}^{K} e^{W_i^T x} \right) = \frac{1}{N}\left(\ln\left(\sum_{i=1}^{K} e^{W_i^T x} \right) - W_y^T x \right)$$

$$\frac{\partial err(W, x, y)}{\partial W_y} = \frac{1}{N}\left(\frac{x e^{W_y^T x}}{\sum_{i=1}^{K} e^{W_i^T x}} - x \right) = \frac{1}{N}(h(x) - 1)\, x$$

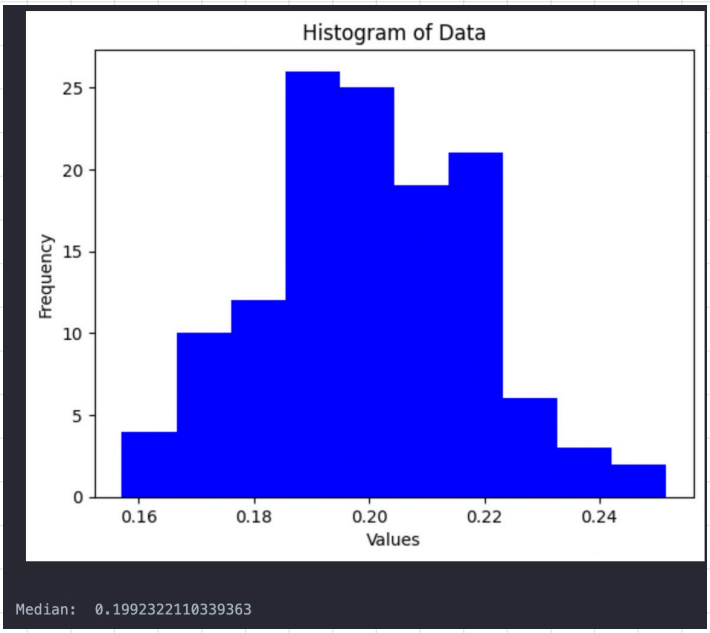
$$\frac{\partial err(W, x, y)}{\partial W_t} = \frac{1}{N}\left(\frac{x e^{W_t^T x}}{\sum_{i=1}^{K} e^{W_i^T x}} \right) = \frac{1}{N} h(x)\, x$$

$\uparrow$

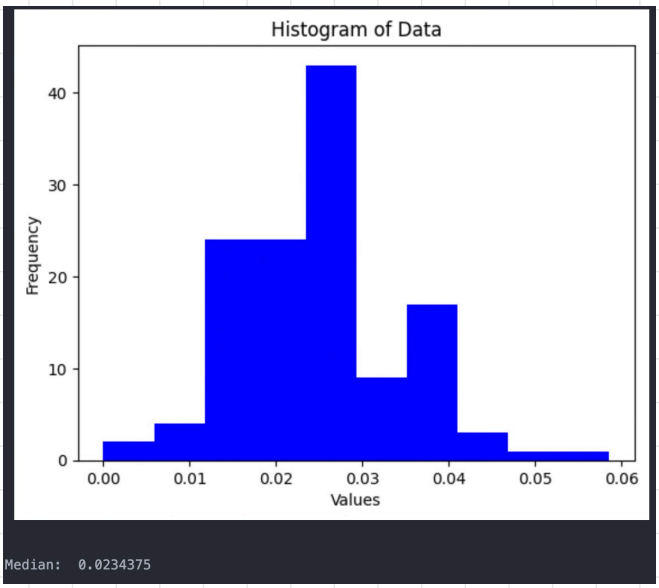
$t \neq y$

Then $\nabla E_{in}(W) = \sum_{n=1}^{N}$ (a $(d+1) \times K$ matrix whose y_n-th column is $\frac{1}{N}(h_{y_n}(x_n) - 1)\vec{x}_n$ and whose other columns are $\frac{1}{N} h_{y_n}(x_n)\vec{x}_n$)

9

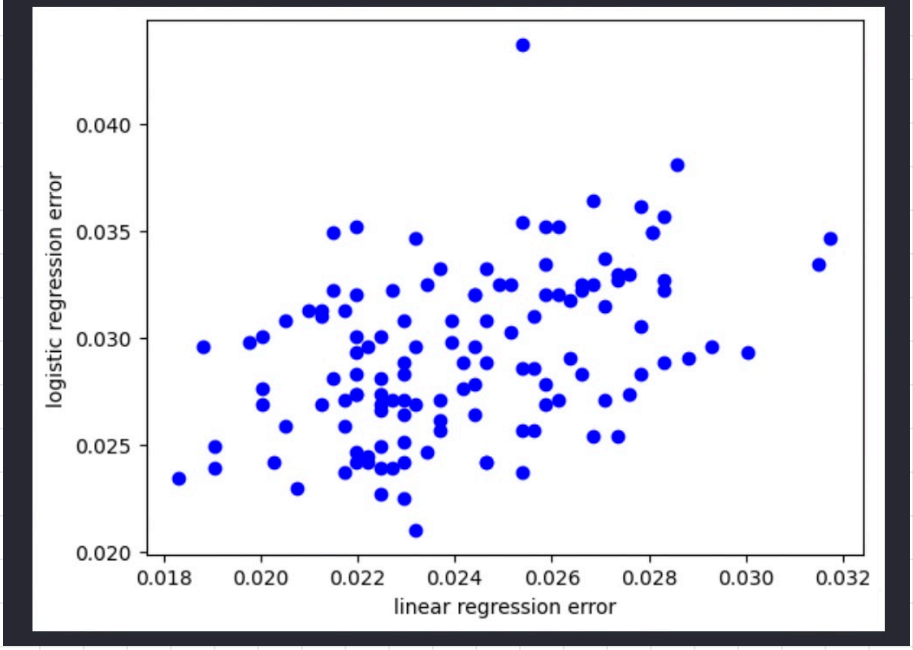

Histogram of Data
Frequency
Values
0
5
10
15
20
25
0.16
0.18
0.20
0.22
0.24
Median: 0.1992322110339363

10


Histogram of Data
Frequency
Values
0
10
20
30
40
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Median: 0.0234375

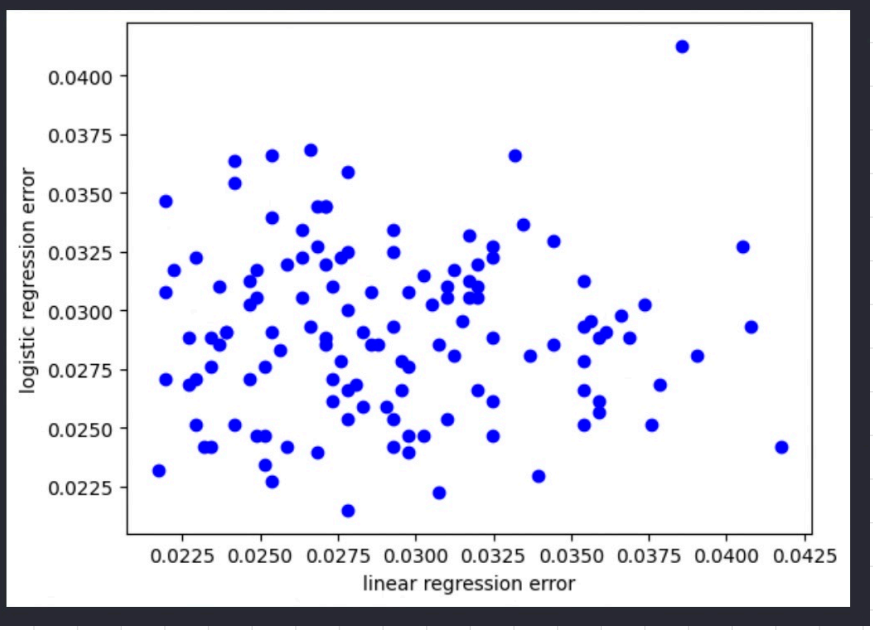
11

```
Median Lin:  0.0238037109375
Median Log:  0.0289306640625
```



12

```
Median Lin:  0.0284423828125
Median Log:  0.0289306640625
```



The median didn't changed much for logistic regression, even the distribution didn't changed much as well.

The top is still bounded at around 0.04, the bottom is around 0.0225.

However, the outcome of linear regression changed siginficantly. The median rises from 0.0238 to 0.0289.

The top becomes 0.0425, the bottom becomes 0.0225. The original top was 0.032, and bottom was 0.018.

Most values for linear regression shifted right, meaning their Err increases, and the distribution becomes more separate compare to the last problem.

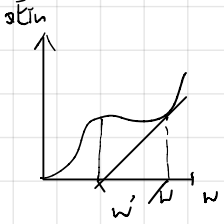
In these two questions we can see, with noiseless data, linear regression has better performance. However, if we added some noise into the data, the performance of two models would be very alike. (Without looking at the outliers, most data points are within (0.0375,0.0375) to (0.0225, 0.0225))

13.

For { logistic regression, we're trying to minimize E_{in},
Linear regression

Suppose $\nabla^2 E_{in}$ is positive definite, we know when ∇E_{in} decreases, E_{in} decreases,

Therefore minimizing ∇E_{in} equals to minimizing E_{in}



The internal linear regression problem is to find w such that minimize $\nabla E_{in}(w)$, which is to find $\nabla E_{in}(w) = 0$.

In order to find $\nabla E_{in}(w) = 0$, we use Newtons method to approach.

each iteration will update w by $-(\nabla^2 E_{in}(w))^{-1}(\nabla E_{in}(w)) = -(X^T D X)^{-1} X^T y$

$w' = (X^T D X)^{-1} X^T y$ is another solution for a linear regression problem.

$\Rightarrow X^T D X w' = X^T y$

$\Rightarrow \tilde{X}^T \tilde{X} w' = \tilde{X}^T \tilde{y}$

$\Rightarrow \tilde{X} w' = \tilde{y}$

$H_{ij} = \frac{1}{N} \sum_{n=1}^{N} (x_{n,i} x_{n,j}) (h(y_n x_n) h(-y_n x_n))$

$(x_{n,i} x_{n,j}) \quad (1 - h(y_n x_n))^2$