1.

$$f_{0/1}(x) = \underset{y \in \{-1,1\}}{\operatorname{argmax}} P(y|x) = \operatorname{sign}\left(P(1|x) - \tfrac{1}{2}\right)$$

$f_{0/1}$ \ g	+1	-1
+1	no err	1
-1	1	no err

For CIA case,

f_{CIA} \ g	+1	-1
+1	no err	1
-1	1000	no err

$$\begin{cases} 1, & y_n \neq g(x_n) \wedge y_n = +1 \\ 1000, & y_n \neq g(x_n) \wedge y_n = -1 \end{cases}$$

Let $f_{CIA} = \operatorname{sign}(P(1|x) - \alpha)$

Expected cost classifying x as positive: $1000\,P(-1|x)$

Expected cost classifying x as negative: $1 \cdot P(1|x)$

We only classify x as positive when the cost that we classify x as negative is less than classifying as positive.

$\therefore\ 1000\,P(-1|x) < 1 \cdot P(1|x)$

By $1 - P(1|x) = P(-1|x)$

$$1000(1 - P(1|x)) < P(1|x)$$

$$P(1|x) > \frac{1000}{1001} \Rightarrow \alpha = \frac{1000}{1001}$$

#

$$\therefore f_{CIA} = \operatorname{sign}\left(P(1|x) - \frac{1000}{1001}\right)$$

#

2. $P(y=+f(x)|x)=1-\varepsilon$

$P(y=-f(x)|x)=\varepsilon$

Given $E_{out}(g)=\mathcal{E}_{x\sim P(x)}[[g(x)\neq f(x)]]$.

Find $E_{(x,y)\sim P(x,y)}[[g(x)\neq y]]$

<u>case1.</u> when $g(x)=f(x)$ but $y\neq f(x)$ (due to noise)

$$E_1=(1-E_{out}(g))\cdot(\varepsilon) \text{ —— } ①$$

<u>case2.</u> when $g(x)\neq f(x)$, but $y(x)=f(x)$

$$E_2=E_{out}(g)(1-\varepsilon)$$

		$[[y(x)=f(x)]]$	
		+1	−1
$[[g(x)=f(x)]] \rightarrow$	+1	X	err
	−1	err	X

$$\therefore E_{(x,y)\sim P(x,y)}[[g(x)\neq y]]=(1-E_{out}(g))\varepsilon+E_{out}(g)(1-\varepsilon)$$

$$=\varepsilon-2E_{out}(g)+E_{out}(g) \quad \#$$

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

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

To find $\min_{w} E_{in}(w) \Rightarrow$ we take $\nabla E_{in}(w)=0$

$$\therefore \nabla E_{in}(w)=\frac{\partial E}{\partial w}=\frac{1}{N}\sum_{n=1}^{N}(2x_n^2w-2x_ny_n)=0$$

$$\Rightarrow \sum_{n=1}^{N}2x_n^2w-\sum_{n=1}^{N}2x_ny_n=0 \Rightarrow w_{lin}=\frac{\sum_{n=1}^{N}x_ny_n}{\sum_{n=1}^{N}x_n^2} \quad \#$$

4.

$f(x) = ax^2+b$, $h(x) = w_0 + w_1 x$, x sampled from $[0,1]$

$$\therefore E_{sqr}(w_0,w_1) = \int_0^1 (h(x)-f(x))^2 dx$$

$$= \int_0^1 \left((w_0+w_1x)-(ax^2+b)\right)^2 dx = \int_0^1 \left(-ax^2+w_1x+w_0-b\right)^2 dx$$

To Find $\min_{w} E_{sqr}(w_0,w_1) \Rightarrow \nabla E(w_0,w_1)=0$

$$\frac{\partial E}{\partial w_0}=0 \Rightarrow 2\int_0^1 (-ax^2+w_1x+w_0-b)\cdot 1\, dx = 0$$

$$\Rightarrow 2\left(\frac{-a}{3}x^3+\frac{w_1}{2}x^2+(w_0-b)x\right)\Big|_0^1 = 0$$

$$\Rightarrow \frac{-a}{3}+\frac{w_1}{2}+(w_0-b)=0 \Rightarrow w_1+2w_0=\frac{2}{3}a+2b \quad \text{—①}$$

$$\frac{\partial E}{\partial w_1}=0 \Rightarrow 2\int_0^1 (-ax^2+w_1x+w_0-b)\cdot(x)\, dx = 0$$

$$\Rightarrow 2\int_0^1 (-ax^3+w_1x^2+(w_0-b)x)\, dx = 0$$

$$\Rightarrow 2\left(\frac{-a}{4}x^4+\frac{w_1}{3}x^3+\frac{w_0-b}{2}x^2\right)\Big|_0^1 = 0$$

$$\Rightarrow \frac{-a}{4}+\frac{w_1}{3}+\frac{w_0-b}{2}=0 \Rightarrow 4w_1+6(w_0-b)-3a=0$$

$$\Rightarrow 4w_1+6w_0=3a+6b \quad \text{—②}$$

By ①②,

$$\therefore w_1^* = a,\ w_0^* = \frac{-a}{6}+b$$

$\therefore (w_0^*, w_1^*) = \left(\frac{-a}{6}+b, a\right)$ # We have $\min_{w} E_{in}(w_0, w_1)$

5.

$$W_{Lin} = (X^TX)^{-1}X^Ty$$

$$E(W_{Lin}') = \frac{1}{N}\left\|XW_{Lin}' - \left(ay + \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix}\right)\right\|^2$$

$$= \frac{1}{N}\left(W_{Lin}'^T X^T X W_{Lin}' - 2\left(XW_{Lin}'\left(ay + \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix}\right)\right) + \left(ay + \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix}\right)^2 \right)$$

To find $\min_{W} E_{in}(W_{Lin}')$

$$\nabla E(W_{Lin}') = 0$$

$$\Rightarrow 2X^TXW_{Lin}' - 2X^T\left(ay + \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix}\right) = 0$$

Let $W_{Lin}' = W_1 + W_2$ (By linear combination)

For W_1, $X^TXW_1 - X^Tay = 0 \Rightarrow W_1 = a(X^TX)^{-1}X^Ty$

$$= aW_{Lin}$$

For W_2, $X^TXW_2 - X^T\begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix} = X^T\left(XW_2 - \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix}\right) = 0$

$$\because X = \begin{bmatrix} 1 & - & x_1^T & - \\ 1 & - & x_2^T & - \\ \vdots & & & \\ 1 & - & x_N^T & - \end{bmatrix} \quad \therefore W_2 = \begin{bmatrix} b \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\therefore W_{Lin}' = aW_{Lin} + \begin{bmatrix} b \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

($\begin{bmatrix} b \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ is $(d+1)\times 1$ column vector)

#

6. $E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} \ln(1+\exp(-y_n w^T x_n))$, Given $h_t = \frac{1}{1+\exp(w_t^T x)}$

$$\nabla E_{in}(w) = \frac{\partial E}{\partial w} = \frac{1}{N}\sum_{n=1}^{N}(-y_n x_n)\left(\frac{\exp(-y_n w^T x_n)}{1+\exp(-y_n w^T x_n)}\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}(-y_n x_n)\, h_t(y_n w_t^T x_n)$$

($\because h(S) = 1-h(S)$ by definition of logistic function)

For $\nabla^2 E_{in}(w) = \frac{\partial}{\partial w_j}\frac{\partial}{\partial w_i}(E_{in}(w))$, We subsitite $k = \exp(-y_n w^T x_n)$

$$= \left(\frac{1}{N}\sum_{n=1}^{N}(-y_{ni} x_{ni})\frac{(-y_{nj}x_{nj})k\cdot(1+k) - k(-y_{nj}x_{nj})k}{(1+k)^2}\right)$$

$$= \frac{1}{N}\sum_{\substack{n=1\\(i,j)}}^{N}\left(x_{ni}\, x_{nj}\, y_{ni}\, y_{nj}\, h(y_n w_t^T x_n)(1-h_t(y_n w_t^T x_n)\right)$$

The single term denote Hessian Matrix

$\Rightarrow$ Express the sum in matrix form

Let X as matrix is a data point x_n, D_{nn} is a diagnal matrix where n^{th} diagonal entry :

$$\because A_E(w_t)\Big|_{E=E_{in}} = X^T D X$$

$$\therefore D_{nn} = y_n^2 h_t(y_n w_t^T x_n)\left(1-h_t(y_n w_t^T x_n)\right)$$

$\therefore$ For diagnal Matrix D,

$$D = \mathrm{diag}\Big[y_1^2 h_t(y_1 w_t^T x_1)(1-h_t(y_1 w_t^T x_1)),\ y_2^2 h_t(y_2 w_t^T x_2)(1-h_t(y_2 w_t^T x_2)),\ \ldots\ldots,\ y_n^2 h_t(y_n w_t^T x_n)(1-h_t(y_n w_t^T x_n))\Big]$$

#

7. Given $err(s,y) = \left(\max(0, 1-ys)\right)^2$, $s = w^T x$

By SGD: $w_{t+1} \leftarrow w_t + \eta\left(-\nabla err(w, x_n, y_n)\right)$ (η is fixed learning rate)

<u>case 1.</u> if $1-ys \geq 0 \Rightarrow err(s,y) = (1-ys)^2$

$$\therefore \frac{\partial err(s,y)}{\partial w} = \frac{\partial err(s,y)}{\partial s} \cdot \frac{\partial s}{\partial w}$$

$$= -2(1-ys)y \cdot x = -2yx(1-ys)$$

<u>case 2.</u> if $1-ys \leq 0 \Rightarrow err(s,y) = 0 \Rightarrow$ The $\nabla err(s,y) = 0$

$\therefore$ SGD update: $w_{t+1} \leftarrow w_t + \eta\left(2yx(1-ys)\right)$ #
(Only when $1-ys \geq 0$)

Comparing to origuinal PLA, Explanations:

① Loss Function: original PLA is 0/1 loss, update only when misclassification, while the new approach uses the truncated squared loss which smoothen the loss surface

② Update rule: In origuinal PLA $w_t \leftarrow w_t + \eta yx$ for misclassified points. The truncated squared loss SGD updates with weighted factor scaled by $2(1-ys)$
(Furthermore, only updates when $1-ys \geq 0$)

③ Convergence: The original PLA might not converge for non-linearly separable data. The truncated squared SGD might converge to a solution even if the data is not linearly-separable due to nature of truncated squared loss #

8.

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

$$E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} err(w, x_n, y_n)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} [[y=k]](-\ln h_k(x))$$

For single point (x, y) (i.e. $y = k$)

$$err(w, x, y) = -\ln \frac{\exp(w_y^T x)}{\sum_{i=1}^{k} \exp(w_i^T x)} = -w_y^T x + \ln\left(\sum_{i=1}^{k} \exp(w_i^T x)\right)$$

① For $y = k$,

$$\nabla_{y=k} (err(w, x, y)) = \frac{\partial(-w_y^T x)}{\partial w_k} + \frac{\partial\left(\ln \sum_{i=1}^{k} \exp(w_i^T x)\right)}{\partial w_k}$$

$$= -x + \left.\frac{\exp(w_k^T x) \cdot x}{\sum_{i=1}^{k} \exp(w_i^T x)}\right|_{y=k}$$

$$= -x + h_k(x) \cdot x$$

② For $y \neq k$

$$\nabla_{y \neq k} (err(w, x, y)) = \frac{\partial\left(\ln \sum_{i=1}^{k} \exp(w_i^T x)\right)}{\partial w_k}$$

$$= \frac{\exp(w_i^T x) \cdot x}{\sum_{i=1}^{k} \exp(w_i^T x)} = h_k(x) \cdot x$$

$$\therefore \nabla err(w, x, y) = \begin{cases} -x + h_k(x) \cdot x, & \text{if } y = k \\ h_k(x) \cdot x, & \text{if } y \neq k. \end{cases}$$

$$\therefore \nabla E_{in} = \frac{1}{N}\sum_{n=1}^{N} \nabla err(w, x_n, y_n)$$

$$\therefore \nabla E_{in} = \begin{cases} \frac{1}{N}\sum_{n=1}^{N}(-x + h_k(x)\cdot x), & \text{if } y = k \\ \frac{1}{N}\sum_{n=1}^{N}(h_k(x)\cdot x), & \text{if } y \neq k. \end{cases}$$

#

(Note each $\nabla err(w, x, y)$ is a matrix of size $(d+1) \times K$, each column is the gradient of corresponding W_k as derived above.)

13. $$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} \ln(1 + \exp(-y_n w_t^T x_n))$$

$$v = -(X^T D X)^{-1} \nabla E_{in}(w_t)$$

$$D_{nn} = h(x_n)(1 - h(x_n))$$

The analogy between logistic regression and linear regression is that

$$W_{Lin} = \underbrace{(X^T X)^{-1}}_{\hookrightarrow (X^T D X)^{-1}} \underbrace{X^T y}_{\hookrightarrow (-\nabla E_{in}(w_t))} = (\tilde{X}^T \tilde{X})^{-1}(\tilde{X}^T \tilde{y})$$

To find analogy,

$$-\nabla E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} h_t(y_n W^T x_n)(+y_n x_n) = \frac{1}{N}\sum_{n=1}^{N} \frac{y_n x_n}{1 + \exp(y_n w^T x_n)}$$

(rewrite the sum) $\stackrel{\downarrow}{=} \frac{1}{N} X^T \frac{y_n}{1 + e^{y_n w^T x_n}}$

$\therefore$ We define $\tilde{y} = \frac{1}{N\sqrt{D}}\left(\frac{y_n}{1+\exp(y^n w^T x_n)}\right)$

$$\tilde{x} = x\sqrt{D}$$

#

Verify: $\tilde{x}^T\tilde{x} = x^T\sqrt{D}\cdot\sqrt{D}x = x^T D x$ (correct)

$$\tilde{x}^T\tilde{y} = x^T\sqrt{D}\cdot\frac{1}{N\sqrt{D}}\left(\frac{y_n}{1+\exp(y^n w^T x_n)}\right)$$

$$= \frac{1}{N}\left(x^T\cdot\left(\frac{y_n}{1+\exp(y^n w^T x_n)}\right)\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{1+\exp(y^n w^T x_n)}\right)(y_n x_n)$$

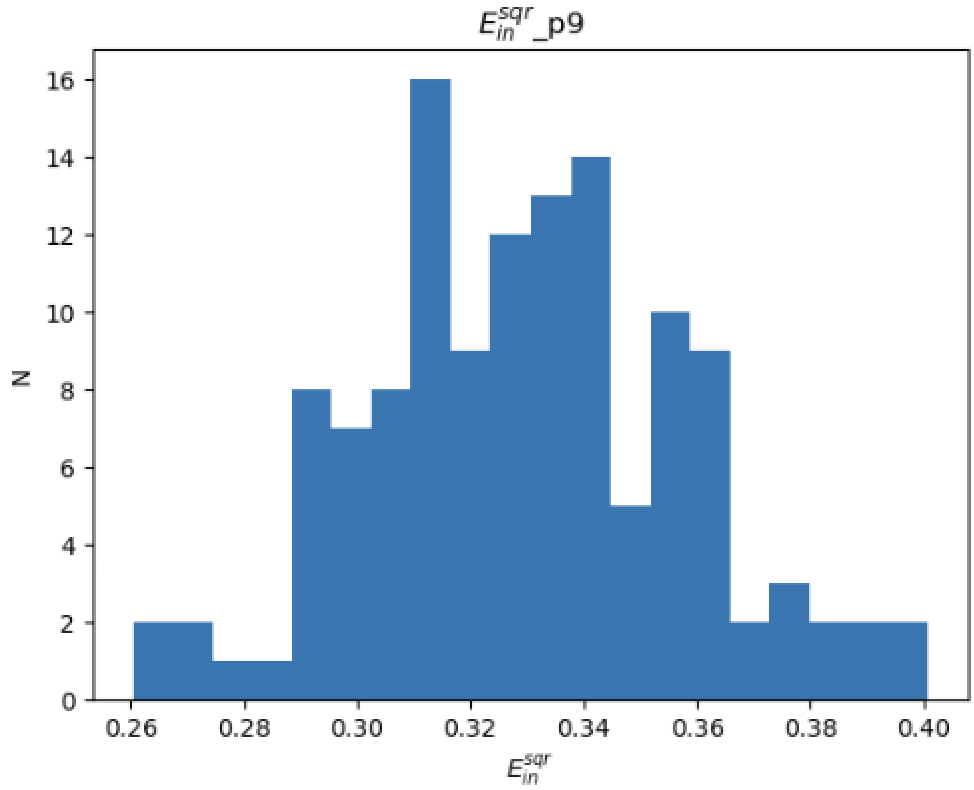
$$= -\nabla E_{in}(w) \quad \text{(correct)}$$

#

HTML hw3 solution

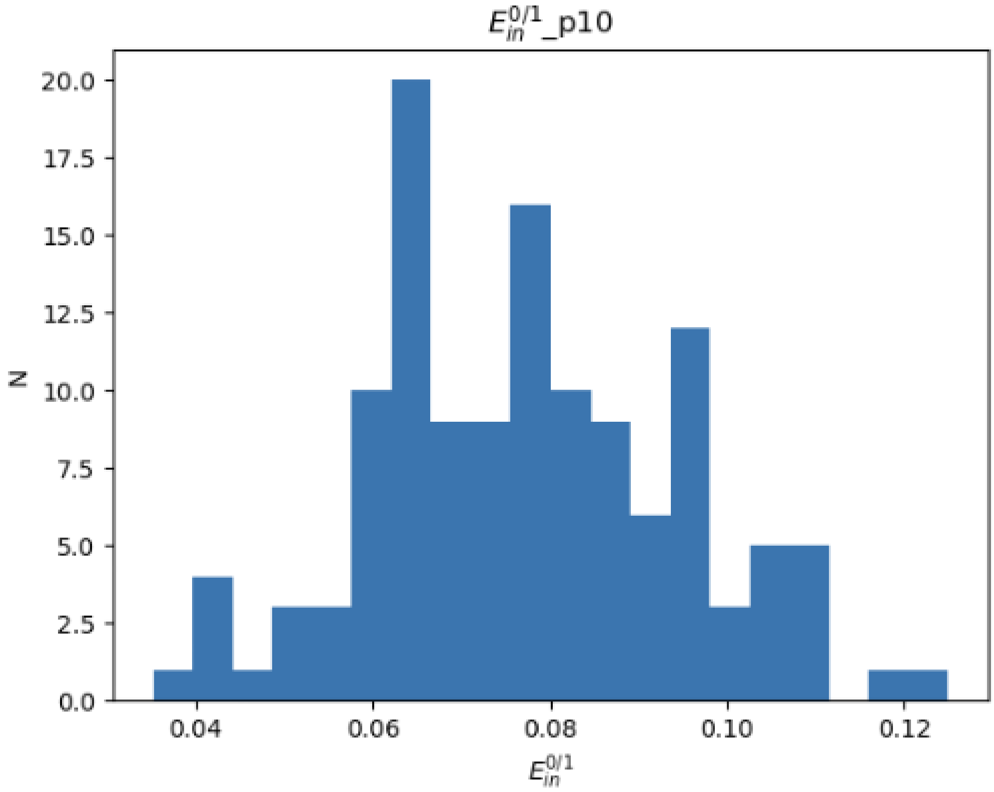
姓名: 謝銘倫, 系級:電機三, 學號:B10502166

9.



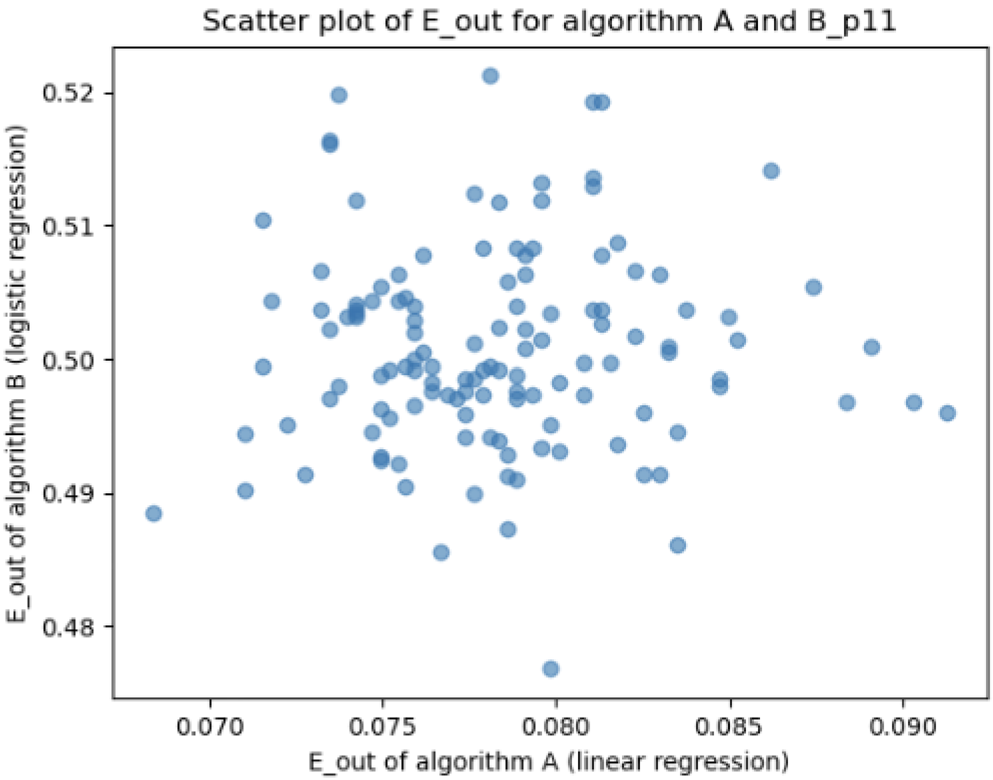
median of E_in: 0.329

10.



median of E_in: 0.078

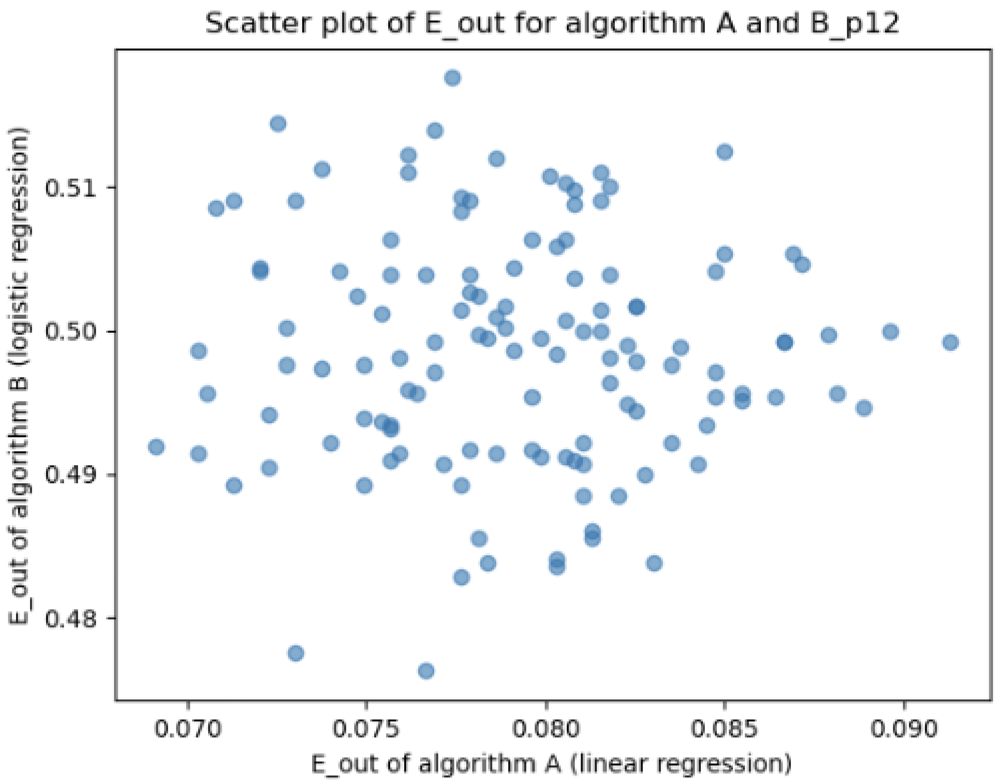
11.



median of $E_{out}(A(D))$ $=$ 0.078

median of $E_{out}(B(D))$ $=$ 0.499

12.



median of $E_{out}(A(D'))$ $=$ 0.079

median of $E_{out}(B(D'))$ $=$ 0.498

Findings:

1. 第11題與第12題的圖結果相似 median of E_out也相似

2. 因為logistic regression是用sigmoid function 這個函數是個monotone function並不容易受到outlier影響 因此結果相似

3. outlier的數量並不多 僅16筆 相對於原本training data的256筆 僅佔少數 因此對整體圖形的影響並不大