PS #3.

1. (a) First Split

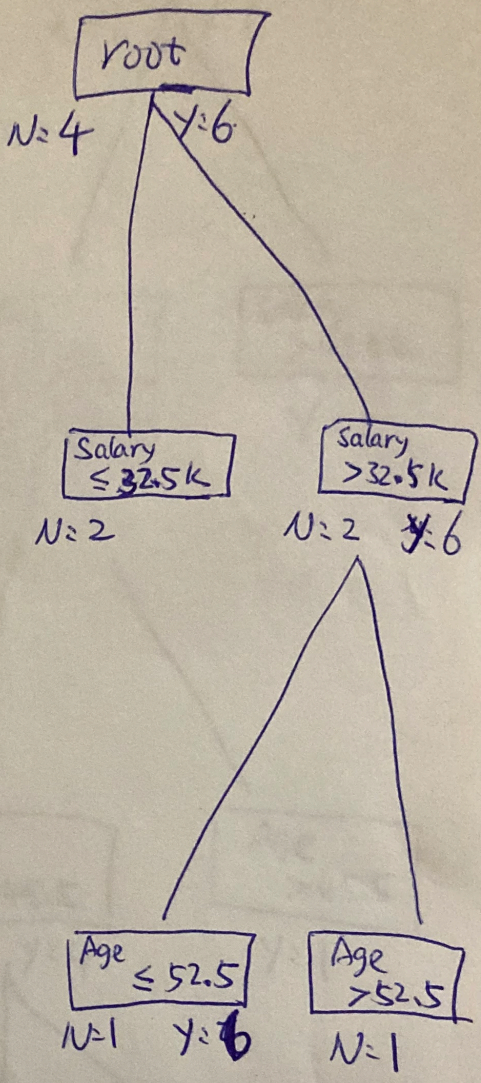
Age:	22	23	24	25	32	43	48	52	52	53
Degree:	Y	N	Y	Y	Y	N	Y	Y	N	N
Error:	$\frac{4}{9}$	$\frac{1}{2}+\frac{3}{8}$	$\frac{1}{3}+\frac{3}{7}$	$\frac{1}{4}+\frac{2}{6}$	$\frac{1}{5}+\frac{2}{5}$	$\frac{2}{6}+\frac{2}{4}$	$\frac{2}{7}+\frac{1}{3}$	[illegible]		$\frac{3}{9}$
	=0.44	=0.875	=0.76	0.75	=0.6	=0.83	=0.62	[illegible]		=0.33

Salary:	25	27	38	40	44	48	52	65	77	110
Degree:	N	N	Y	Y	N	Y	N	Y	Y	Y
Error	$\frac{3}{9}$	$\frac{2}{8}$	$\frac{1}{3}+\frac{2}{7}$	$\frac{3}{4}+\frac{2}{6}$	$\frac{2}{5}+\frac{1}{5}$	$\frac{3}{6}+\frac{1}{4}$	$\frac{3}{7}$	$\frac{4}{8}$	$\frac{4}{9}$	
	=0.33	=0.25	=0.62	=0.83	=0.6	=0.75	0.43	0.5	0.44	

Second split

Age:	22	24	25	32	43	48	52	53
Degree:	Y	Y	Y	Y	N	Y	Y	N
Error:	$\frac{2}{7}$	$\frac{2}{6}$	$\frac{2}{5}$	$\frac{2}{4}$	$\frac{1}{5}+\frac{1}{3}$	$\frac{1}{6}+\frac{1}{2}$	$\frac{1}{1}$	
=	0.29	0.33	0.4	0.5	0.53	0.67	0.14	

Salary	38	40	44	48	52	65	77	110
Degree:	Y	Y	N	Y	N	Y	Y	Y
Error	$\frac{2}{7}$	$\frac{2}{6}$	$\frac{1}{3}+\frac{1}{5}$	$\frac{1}{4}+\frac{1}{4}$	$\frac{2}{5}$	$\frac{2}{6}$	$\frac{2}{7}$	
=	0.29	0.33	0.53	0.5	0.4	0.33	0.29	



1.(a) cont.

third split.

Age	22	24	25	32	43	48	52
Degree:	Y	Y	Y	Y	N	Y	Y
Error	1/6	1/5	1/4	1/3	1/5	1/6	
=	0.17	0.2	0.25	0.33	0.2	0.17	

Salary	38	40	44	48	65	77	110
Degree	Y	Y	N	Y	Y	Y	Y
Error	1/6	1/5	1/3	1/4	1/5	1/6	
=	0.17	0.2	0.33	0.25	0.2	0.17	

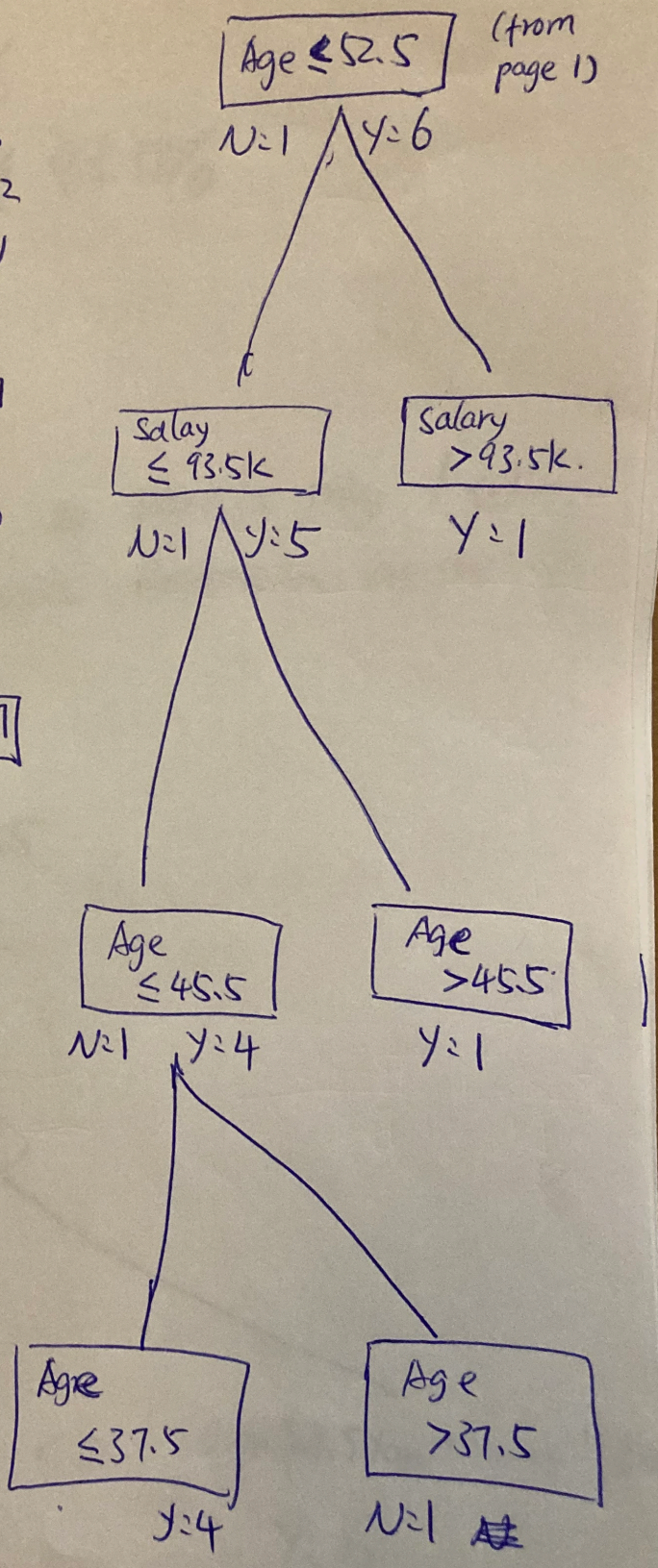
Forth split.

Age	22	24	25	32	43	48
Degree:	Y	Y	Y	Y	N	Y
Error	1/5	1/4	1/3	1/2	1/5	
=	0.2	0.25	0.33	0.5	0.2	

salary	38	40	44	48	65	77
Degree:	Y	Y	N	Y	Y	Y
Error:	1/5	1/4	1/3	1/4	1/5	
=	0.2	0.25	0.33	0.25	0.2	

Fifth split.

Age:	22	24	25	32	43
Degree:	Y	Y	Y	Y	N
Error	1/4	1/3	1/2	0	



1. (b)

accuray: for college : 90%

" " Iris : ~~93.33%~~ 93.33%

(c). data plot

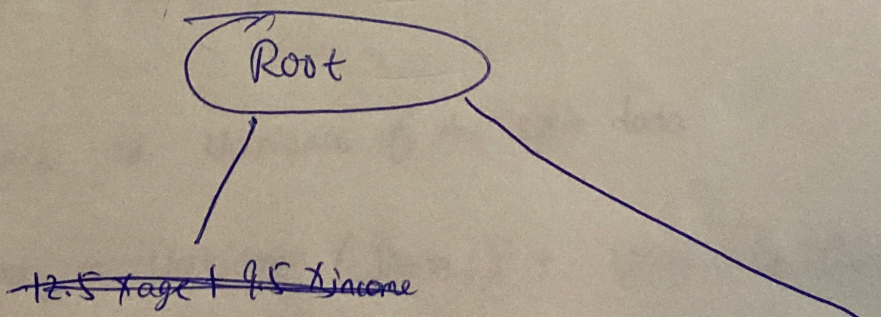
it's a linear separatable data, so there is only 1 split.

by using hw2 Logistic Regression on the

~~(d)~~ dataset. it learned:

$\alpha = -1.5$

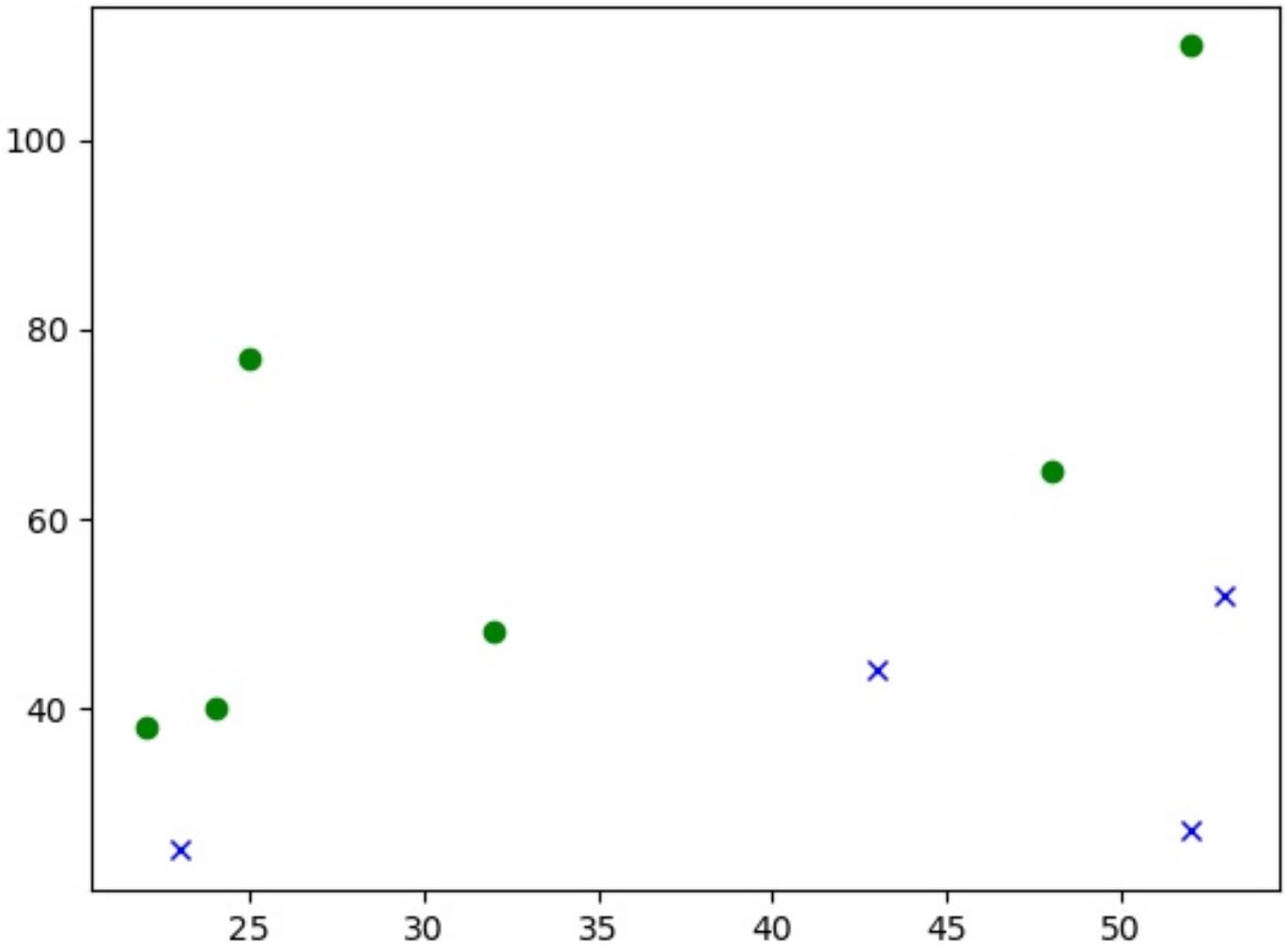
$\beta = 2.$



Sign $\{-1.5 X_{age} + 2X_{income} - 1\} < 0$ N: 4

Sign $\{-1.5 X_{age} + 2X_{income} - 1\} \geq 0$ Y: 6

Classification Error: 0.


100
80
60
40
25
30
35
40
45
50

d), (d)

Pro: ① tree has less depth. less complicated.

② can cover more decision boundry. eg:
univariate tree can only have decision boundry parallel to the feature axis. multivariate tree can cover other linear boundry.

Con: ① hard to reason about. eg.
$-1.5 \times age + 2 \times income$ doesn't have any meaning.

② ~~doesn't fit the~~
can much easier to overfit.

d), (e) (i) for regression, the appropiate output would be the average of training data points' y value.

Line 3: return Leaf($\frac{1}{n} \sum_{i=1}^{n} y^{(i)}$)

(ii) minimize the variance of the split data.

Line 8: $Error_{i,v}$ = variance ($Data_1$) + variance ($Data_2$)

$$\text{variance} = \frac{\sum_{i=1}^{N} (y^{(i)} - \bar{y})^2}{N}, \quad \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y^{(i)},$$

2. (a)

$$G(R_m) = P_{m_1}(1-P_{m_1}) + P_{m_2}(1-P_{m_2})$$

$P_{m_1} = \frac{m_1}{m_1+m_2}$ = m_1 is the # of class 1, m_2 is the # of class 2.

$P_{m_2} = \frac{m_2}{m_1+m_2} = 1 - P_{m_2}$. $m = m_1+m_2$ is the # of all classes

$$G(R_m) = P_{m_1}P_{m_2} + P_{m_2}\cdot P_{m_1}$$

$$= 2P_{m_1}P_{m_2}$$

$$= 2\frac{m_1 \cdot m_2}{m_1+m_2}$$

$$= 2\frac{m_1(m-m_1)}{m} = 2\frac{m_2(m-m_2)}{m}$$

$$\nabla G(R_m)_{m_1,m_2} = \begin{bmatrix} \frac{2}{m}(m-2m_1) \\ \frac{2}{m}(m-2m_2) \end{bmatrix} = \begin{bmatrix} 2-\frac{4m_1}{m} \\ 2-\frac{4m_2}{m} \end{bmatrix}$$

~~$m_1 \geq 0$ $m_2 \geq 0$~~ and m_1 and m_2 cannot be 0 at the same time

∴ ~~$\nabla G(R_m)_{m_1,m_2}$ is strictly concave.~~

$$H\,G(R_m) = \begin{bmatrix} -\frac{4}{m} & 0 \\ 0 & -\frac{4}{m} \end{bmatrix} = -\frac{4}{m}\cdot I$$

negative definite

∴ $G(R_m)$ is strictly concave.

~~the, let $P_1 = m_1$ $P_2 = m_2$ $t = \frac{m_1}{m_1+m_2}$ $t \in (0,1)$~~

~~$G(tP_1 + (1-t)P_2) > \frac{m_1}{m_1-m_2} G(m_1) = \frac{m_2}{} G(m$~~

2.(a) cont.

~~$G(M) = G(0.5M + 0.5M) > 0.5\,G(0.5M) + 0.5\,G(0.5M)$~~

$G(M) = 2 \cdot G(0.5M)$ by definition.

~~$G(m_1 + m_2)$~~

$G(0.5M)$ $\quad P_1 = m_1,\ P_2 = m_2,\ t = 0.5.$

$= G(0.5m_1 + 0.5m_2) > 0.5\,G(m_1) + 0.5\,G(m_2)$

$2\,G(0.5m_1 + 0.5m_2) > G(m_1) + G(m_2)$

$\hookrightarrow = 2G(0.5M)$

$= G(M)$

$\therefore G(M) > G(m_1) + G(m_2)$

2.(b) ① when $G(m) = 0$: loss is 0, all children are in the same class, so $G(m_1)$ will be 0, $G(m_2)$

② when $P_1 = P_2$, the children region has the exact same propotion of classes examples as the parent.

because we aim to reduce the loss, we always looks for a split that has smaller loss than the parent. Therefore, it will converge

2.(c) let M be the root, N_{m_1} be # of class 1 in m

N_{m_2} " " " " 2 " "

assume M split at some feature and result in X, y, child node

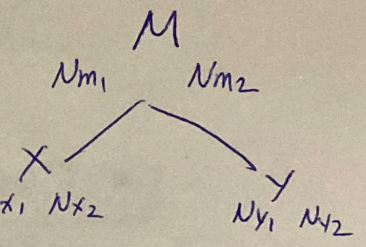
let N_{x_1} be # of class 1 in X

N_{x_2} " " " " 2 " "

N_{y_1} " " " " 1 " Y

N_{y_2} " " " " 2 " y

further, assume $N_{m_1} \leq N_{m_2}$



by definition:

$$L(M) = \frac{N_{m_1}}{N_m}, \quad N_{m_1} + N_{m_2} = N_m$$

$$N_{x_1} + N_{y_1} = N_{m_1}, \quad N_{x_2} + N_{y_2} = N_{m_2}, \quad N_x + N_y = N_M$$

$$L(X) + L(y) = \frac{\min(N_{x_1}, N_{x_2})}{N_x} + \frac{\min(N_{y_1}, N_{y_2})}{N_y}$$

$$= \frac{N_y \min(N_{x_1}, N_{x_2}) + N_x \cdot \min(N_{y_1}, N_{y_2})}{N_x \cdot N_y}$$

$$= \frac{N_x \min(N_{y_1}, N_{y_2}) + (N_M - N_x)\min(N_{x_1}, N_{x_2})}{N_x \cdot (N_M - N_x)}$$

$$L(M) = \frac{N_{m_1}}{N_m} = \frac{N_{x_1} + N_{y_1}}{N_m}$$

let $L(M) = L(X) + L(y)$

$$\frac{N_{x_1} + N_{y_1}}{N_m} = \frac{N_M \cdot \min(N_{x_1}, N_{x_2}) - [\min(N_{x_1}, N_{x_2}) - \min(N_{y_1}, N_{y_2})] \cdot N_x}{N_x \cdot (N_m - N_x)}$$

$$N_{m_1} \cdot N_x (N_m - N_x) = N_m^2 \cdot \min(N_{x_1}, N_{x_2}) - N_m \cdot N_x [\min(N_{x_1}, N_{x_2}) - \min(N_{y_1}, N_{y_2})]$$

$$N_{m_1} N_x N_m - N_{m_1} N_x^2 = N_m^2 \cdot \min(N_{x_1}, N_{x_2}) - N_m \cdot N_x [\min(N_{x_1}, N_{x_2}) - \min(N_{y_1}, N_{y_2})]$$

2.(c) cont.

~~Set: $N_{m_1} N_x N_m = N_m^2 \cdot \min(N_{x_1}, N_{x_2})$~~

~~$N_{m_1} N_x^2 = N_m \cdot N_x \cdot [\min(N_{x_1}, N_{x_2}) - \min(N_{y_1}, N_{y_2})]$~~

let $L_{xy} = L(x) + L(y)$

$$\frac{N_{m_1}}{N_m} = \frac{\min(N_{x_1}; N_{x_2})}{N_x} + \frac{\min(N_{y_1}, N_{y_2})}{N_y}$$

So the proportion doesn't need to be exactly the same. as long the sum adds up to be equal.

2.(d) $\rho = \frac{\text{Cov}(T_i(x), T_j(x))}{\sqrt{\text{Var}(T_i(x))} \cdot \sqrt{\text{Var}(T_j(x))}} = \frac{\text{Cov}(T_i(x), T_j(x))}{\sigma^2}$

$$\text{Cov}(T_i(x), T_j(x)) = \rho \cdot \sigma^2$$

$$\text{Var}\left(\frac{1}{B}\sum_{i=1}^{B} T_i(x)\right) = \frac{1}{B^2}\text{Var}\left(\sum_{i=1}^{B} T_i(x)\right) = \frac{1}{B^2}\left[\sum_{i=1}^{B}\text{Var}(T_i(x)) + 2\sum_{i<j}\text{Cov}(x_i, x_j)\right]$$

$$= \frac{1}{B^2}\left[B \cdot \sigma^2 + 2\sum_{i<j}\rho\sigma^2\right]$$

B

$\hookrightarrow = \text{Cov}(X_1, X_2) + \cdots + \text{Cov}(X_1, X_B)$
$+ \text{Cov}(X_2, X_3) + \cdots + \text{Cov}(X_2, X_B)$ } $B-1$
$+ \vdots$
$\text{Cov}(X_{B-1}, X_B)$

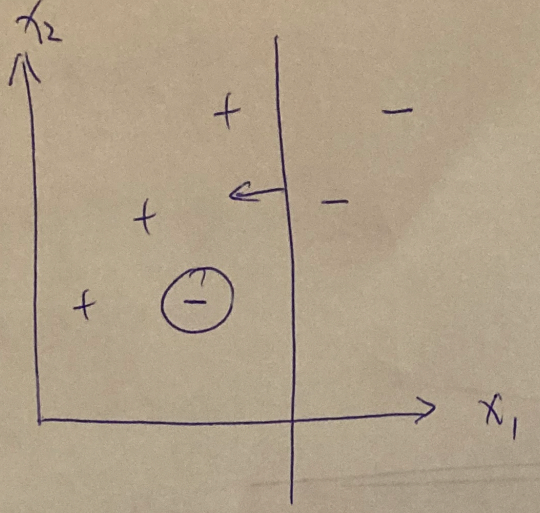
$$= \frac{1}{B^2}\left[B\sigma^2 + 2 \cdot \frac{B \cdot (B-1)}{2} \cdot \rho\sigma^2\right]$$

$$= \frac{\sigma^2}{B} + \frac{B-1}{B}\rho\sigma^2$$

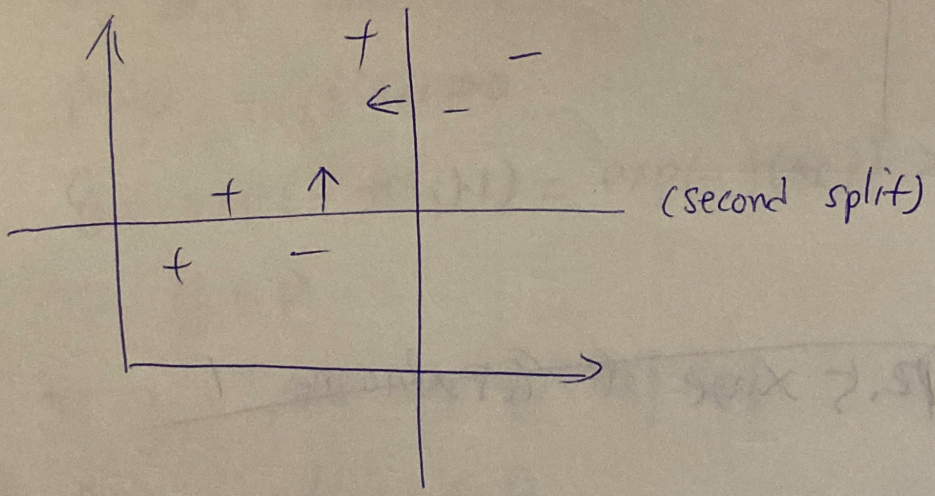
$$= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

3.

(a)



(b)



(c) $\hat{w}_2 < \hat{w}_1 < 0.5$

Weighted error of t_2 is smaller than t_1

so $\hat{w}_2 < \hat{w}_1$

4. (a)

$$\mathbb{1}\{F(x_i) \neq y_i\}$$

$$= \mathbb{1}(F(x_i) \cdot y_i > 0)$$

for a data point i

$$\exp(-f(x_i) y_i) \triangleq q$$

$$\mathbb{1}(F(x_i) \cdot y_i > 0)$$

$$= \mathbb{1}(\text{sign}\{f(x_i)\} \cdot y_i > 0) \triangleq p$$

if $y_i = 1$ and $\text{sign}\{f(x_i)\} = 1$

$\Rightarrow p = 0, \quad f(x_i) > 0$

$q = \exp(-f(x_i) \cdot 1) = \exp(-f(x_i)) > 0$

$\therefore q > p$

if $y_i = 1$ and $\text{sign}\{f(x_i)\} = -1$

$\Rightarrow p = 1 \quad f(x_i) < 0$

$q = \exp(-f(x_i) \cdot 1) \geq \exp(0) = 1$

$\therefore q \geq p$

if $y_i = -1$ and $\text{sign}\{f(x_i)\} = -1$

$\Rightarrow p = 0, \quad f(x_i) < 0$

$q = \exp(-f(x_i) \cdot -1) > 0$ by definition of exp

$\therefore q > p$

if $y_i = -1$ and $\text{sign}\{f(x_i)\} = 1$

$\Rightarrow p = 1 \quad f(x_i) > 0 \quad q = \exp(-f(x_i) \cdot -1) \geq \exp(0) = 1 = p$

4.(a) continue.

$\therefore$ for any i. $\quad p \le q$

$\therefore \frac{1}{n}\sum_{i=1}^{n} p_i \le \frac{1}{n}\sum_{i=1}^{n} q$.

4.(b)

$$Z_t = \sum_{i=1}^{n} a_{i,t} \exp(-\hat{w}_t f_t(x_i) y_i)$$

let $\exp(-\hat{w}_t f_t(x_i) y_i) \triangleq \exp_{i,t}()$

$$Z_t = \sum_{i=1}^{n} a_{it} \cdot \exp_{i,t}()$$

$$\prod_{t=1}^{T} Z_t = Z_1 \cdot Z_2 \cdot \cdots \cdot Z_{t-1} \cdot Z_t$$

$$= Z_1 \cdot Z_2 \cdots Z_{t-1} \cdot \left(\sum_{i=1}^{n} a_{it} \exp_{i,t}()\right)$$

$$= Z_1 \cdot Z_2 \cdots \cancel{Z_{t-1}} \cdot \left(\sum_{i=1}^{n} \frac{a_{i,t-1} \cdot \exp_{i,t-1}()}{\cancel{Z_{t-1}}} \cdot \exp_{i,t}()\right)$$

$$\Rightarrow \prod_{t=1}^{T} Z_t = \sum_{i=1}^{n} a_{i,0} \cdot \exp_{i,1}() \cdot \exp_{i,2}() \cdot \cdots \exp_{i,t}()$$

$a_{i,0} = \frac{1}{n}$ as the unweighted dataset

$$\therefore \prod_{t=1}^{T} Z_t = \sum_{i=1}^{n} \frac{1}{n} \exp_{i,1}() \cdot \exp_{i,2}() \cdots \exp_{i,t}().$$

$$= \sum_{i=1}^{n} \frac{1}{n} \exp(-\hat{w}_1 f_1(x_i) y_i - \hat{w}_2 f_2(x_i) y_i - \cdots \hat{w}_t f_t(x_i) y_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \exp(-f(x_i) \cdot y_i)$$

4 (c) (i)

$$\frac{\partial z_t}{\partial \varepsilon_t} = (1-\varepsilon_t)\cdot(-\exp(-\hat{w_t})) + \varepsilon_t \cdot \exp(\hat{w_t})$$

Set $\frac{\partial z_t}{\partial \varepsilon_t} = 0$

$$(1-\varepsilon_t)\cdot\exp(-\hat{w_t}) = \varepsilon_t \exp(\hat{w_t})$$

$$\exp(\hat{w_t})^2 = \frac{1-\varepsilon_t}{\varepsilon_t}$$

$$\exp(\hat{w_t}) = \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}$$

$$z_t^{opt} = (1-\varepsilon_t)\exp(-\hat{w_t}) + \varepsilon_t \exp(\hat{w_t})$$

$$= \frac{1-\varepsilon_t}{\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}} + \varepsilon_t \cdot \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}$$

$$= \sqrt{(1-\varepsilon_t)(\varepsilon_t)} + \sqrt{\varepsilon_t(1-\varepsilon_t)}$$

$$= 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$

(c) (ii) by definition $0 \le \varepsilon_t \le 1$

$$\Rightarrow 0 \le \frac{1}{2} - r_t \le 1$$

$$\Rightarrow -\frac{1}{2} \le r_t < \frac{1}{2}$$

$$\Rightarrow \frac{1}{4} \ge r_t^2 \ge 0 \quad \Rightarrow 1 \ge 4r_t^2 \ge 0$$

~~$\Rightarrow 0 \le 1-4r_t^2 \le 1$~~

$$z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)} = 2\sqrt{\frac{1}{4} - r_t^2} = \sqrt{1-4r_t^2}$$

$$\log(z_t) = \frac{1}{2}\log(1-4r_t^2) \le \frac{1}{2}\cdot(-4r_t^2) = -2r_t^2$$

$$\therefore z_t \le \exp(-2r_t^2)$$

4.(4) (iii)

$$E_{training} \le \frac{1}{n}\sum_{i=1}^{n} \exp(-f(x_i) y_i)$$

$$= \prod_{t=1}^{T} Z_t$$

$$= Z_1 \cdot Z_2 \cdot \cdots Z_t$$

$$\le \exp(-2\gamma_1^2) \cdot \exp(-2\gamma_2^2) \cdot \cdots \exp(-2\gamma_t^2)$$

$$= \exp(-2\gamma_1^2 - 2\gamma_2^2 - \cdots - 2\gamma_t^2)$$

$\because \gamma_t > \gamma$ for all t and $\gamma > 0$

$\Rightarrow \gamma_t^2 > \gamma$

$\Rightarrow -\gamma_t^2 < \gamma^2$

$$< \exp(\underbrace{-2\gamma^2 - 2\gamma^2 - \cdots 2\gamma^2}_{t})$$

$$= \exp(-2T\gamma^2)$$

for $T \ge 1$

When $T = 0$, $E_{training} \le 1 = \exp(0)$

$\therefore E_{training} \le \exp(-2T\gamma^2)$

(5) (a) for a single data point.

let $a_1 = w^{[1]T} \cdot x + b_1$ $\left(b_1 \triangleq \begin{bmatrix} w_{0_1}^{[1]} \\ w_{0_2}^{[1]} \\ w_{0_3}^{[1]} \end{bmatrix}\right)$

$h = \sigma(a_1)$

let $a_2 = w^{[2]T} \cdot h + b_2$ $(b_2 \triangleq w_0^{[2]})$

$o = \sigma(a_2)$

$$\frac{\partial o}{\partial a_2} = \sigma(a_2) \cdot (1 - \sigma(a_2))$$

$$\frac{\partial a_2}{\partial h} = w^{[2]} \quad , \quad \frac{\partial a_2}{\partial h_2} = w_2^{[2]}$$

$$\frac{\partial h_2}{\partial a_1} = \sigma(a_1)(1 - \sigma(a_1))$$

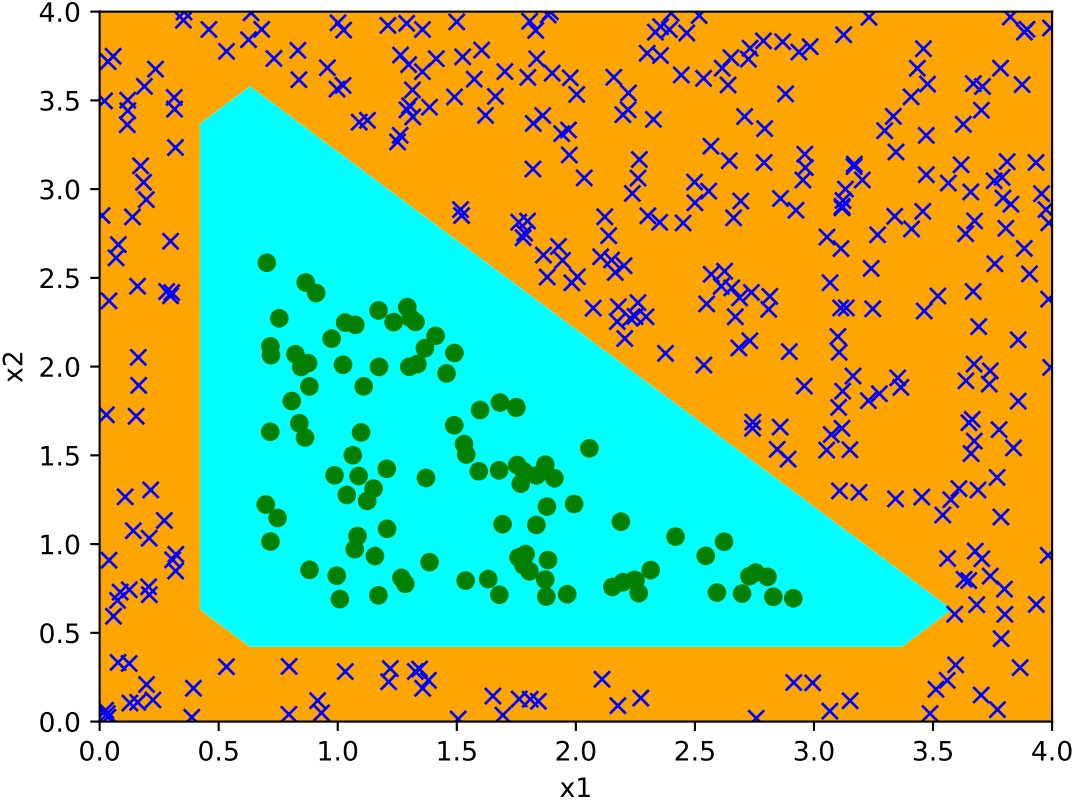
$$\frac{\partial h}{\partial a_1} = \begin{bmatrix} \sigma(a_{11})(1 - \sigma(a_{11})) \\ \sigma(a_{12})(1 - \sigma(a_{12})) \\ \sigma(a_{13})(1 - \sigma(a_{13})) \end{bmatrix} \quad , \quad \frac{\partial h_2}{\partial a_{1,2}} = \sigma(a_{12})(1 - \sigma(a_{12}))$$

$$\frac{\partial a_{12}}{\partial w_{1,2}^{[1]}} = x_1$$

$$\frac{\partial \ell}{\partial o} = 2(o - y)$$

$$\frac{\partial \ell}{\partial w_{1,2}^{[1]}} = \frac{\partial \ell}{\partial o} \cdot \frac{\partial o}{\partial a_2} \cdot \frac{\partial a_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial a_{12}} \cdot \frac{\partial a_{12}}{\partial w_{1,2}^{[1]}}$$

$$= 2(o - y) \cdot \sigma(a_2)(1 - \sigma(a_2)) \cdot w_2^{[2]} \cdot \sigma(a_{12})(1 - \sigma(a_{12})) \cdot x_1$$


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
x2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
x1

5.(c)

no.

linear function can only result in ~~linear regression~~. representing linear relationship among the data.

the training data is non-linear

thus cannot be represented by linear function.