Name :- Nandish. Bakulkumar. Bhatt.
Student ID :- 20041204

Writing Assignment : 3

Ans: 1

1(a) Given that, Sigmoid function, $\sigma(\alpha) = \frac{1}{1+e^{-\alpha}}$

$$\therefore \frac{d\sigma(\alpha)}{d\alpha} = \frac{d}{d\alpha}\left[(1+e^{-\alpha})^{-1}\right]$$

$$= (-1)(1+e^{-\alpha})^{-2} \cdot \frac{d}{d\alpha}(1+e^{-\alpha})$$

$$\left[\because \frac{d[f(x)]^n}{dx} = n[f(x)]^{n-1} \cdot \frac{df(x)}{dx}\right]$$

$$\therefore \frac{d\sigma(\alpha)}{d\alpha} = \frac{(-1)}{(1+e^{-\alpha})^2} \cdot e^{-\alpha} \cdot (-1)$$

$$\left[\because \frac{de^{f(x)}}{dx} = e^{f(x)} \cdot \frac{df(x)}{dx}\right]$$

$$\therefore \frac{d\sigma(\alpha)}{d\alpha} = \frac{e^{-\alpha}}{(1+e^{-\alpha})^2}$$

$$\frac{d\sigma(\alpha)}{d\alpha} = \left[\frac{1}{1+e^{-\alpha}}\right]\left[\frac{e^{-\alpha}}{1+e^{-\alpha}}\right]$$

$$\frac{d\sigma(\alpha)}{d\alpha} = \left[\frac{1}{1+e^{-\alpha}}\right]\left[1-1+\frac{e^{-\alpha}}{1+e^{-\alpha}}\right]$$

($\because$ Adding and subtracting by 1)

$$= \left[\frac{1}{1+e^{-\alpha}}\right]\left[1 - \left(1-\frac{e^{-\alpha}}{1+e^{-\alpha}}\right)\right]$$

$$= \left[\frac{1}{1+e^{-\alpha}}\right]\left[1 - \left(\frac{1}{1+e^{-\alpha}}\right)\right]$$

$$\therefore \boxed{\frac{d\sigma(\alpha)}{d\alpha} = \sigma(\alpha).\left[1-\sigma(\alpha)\right]}$$

$\therefore$ Hence, proved.

1(b) Now, given that —

Negative Log - Likelihood (NLL) for logistic regression is given by :-

$$NLL = -\sum_{i=1}^{N} \log\left[\mu_i^{1(y_i=1)} \times (1-\mu_i)^{1(y_i=0)}\right]$$

$$= -\sum_{i=1}^{N}\left[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)\right]$$

and $\mu_i = \text{sigm}(w^T x_i) = \dfrac{1}{1+e^{-w^T x_i}}$

$\rightarrow$ Now, $NLL = -\sum_{i=1}^{N} \left[y_i \log \mu_i + (1-y_i)\log(1-\mu_i) \right]$

$$\therefore \frac{dNLL(w)}{dw} = (-1)\left\{ \sum_{i=1}^{N} \left[y_i \cdot \frac{d}{dw}(\log \mu_i) + (1-y_i)\frac{d}{dw}\left[\log(1-\mu_i)\right] \right] \right\}$$

$\Big(\because$ - y_i = const. or independent of w.

$\dfrac{d}{dx} \sum_{i=1}^{2} i\, F(x) = \sum_{i=1}^{2} i \dfrac{d}{dx}[F(x)]$ $\Big)$

$$\therefore \frac{dNLL(w)}{dw} = (-1) \sum_{i=1}^{N} \left[y_i \frac{1}{\mu_i} \frac{d\mu_i}{dw} + (1-y_i)\frac{1}{1-\mu_i} \frac{d}{dw}(1-\mu_i) \right]$$

$\Big(\because$ - $\dfrac{d}{dx}\left[\log(F(x))\right] = \dfrac{1}{F(x)} \cdot \dfrac{dF(x)}{dx}$

- μ_i is a function of w. $\Big)$

$$\therefore \frac{dNLL(w)}{dw} = (-1) \sum_{i=1}^{N} \left[\frac{y_i}{\mu_i} \frac{d\mu_i}{dw} - \frac{1-y_i}{1-\mu_i} \cdot \frac{d\mu_i}{dw} \right]$$

$$= (-1) \sum_{i=1}^{N} \left[\left(\frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i} \right) \frac{d\mu_i}{dw} \right]$$

$$= (-1) \sum_{i=1}^{N} \left[\left(\frac{y_i - \cancel{\mu_i y_i} - \mu_i + \cancel{\mu_i y_i}}{\mu_i (1-\mu_i)} \right) \cdot \frac{d\mu_i}{dw} \right]$$

$$\frac{dNLL(w)}{dw} = (-1) \sum_{i=1}^{N} \left[\left(\frac{y_i - \mu_i}{\mu_i (1-\mu_i)} \right) \cdot \frac{d\mu_i}{dw} \right] \quad \text{———} \quad (1)$$

$\longrightarrow$ Now, as proved in 1(a), for $\sigma(\alpha) = \dfrac{1}{1+e^{-\alpha}}$,

$$\frac{d\,\sigma(\alpha)}{d\alpha} = \sigma(\alpha)\left[1 - \sigma(\alpha)\right] \quad \text{———} \quad (2)$$

Whereas, $\mu_i = \text{sigm}(w^T x_i) = \dfrac{1}{1+e^{-w^T x_i}}$

μ_i is a function of $w^T x_i$ or "wx" as whole.

$\rightarrow$ Let's say, $a = w^T x_i$.

So, μ_i is func^n of a.

$$\therefore \frac{d\mu_i}{dw} = \frac{d\mu_i}{da} \times \frac{da}{dw} \quad \text{(Chain Rule)}$$

From previous 1(a),

$$= \mu_i[1-\mu_i] \times \frac{da}{dw} \quad \left(\because \frac{d\,\sigma(\alpha)}{d\alpha} = \sigma(\alpha)[1-\sigma(\alpha)] \quad \therefore \frac{d\mu_i(a)}{da} = \mu_i(a)[1-\mu_i(a)] \right)$$

$$= \mu_i[1-\mu_i] \cdot x_i \quad \left(\because a = w^T x_i \quad \frac{da}{dw} = x_i \frac{d(w^T)}{dw} = x_i \right)$$

$$\therefore \frac{d\mu_i}{dw} = \mu_i[1-\mu_i] \times x_i \quad \text{——— (3)}$$

$\rightarrow$ So, putting eq^n (3) into eq^n (1), we get —

Scanned by CamScanner

$$\therefore \frac{dNLL(w)}{dw} = (-1) \sum_{i=1}^{N} \left[\left(\frac{y_i - \mu_i}{\cancel{\mu_i(1-\mu_i)}} \right) \cancel{\mu_i(1-\mu_i)}\, x_i \right]$$

$$= (-1) \sum_{i=1}^{N} \left[(y_i - \mu_i)\, x_i \right]$$

$$\therefore \boxed{\frac{dNLL(w)}{dw} = \sum_{i=1}^{N} \left[(\mu_i - y_i)\, x_i \right]}$$

Scanned by CamScanner

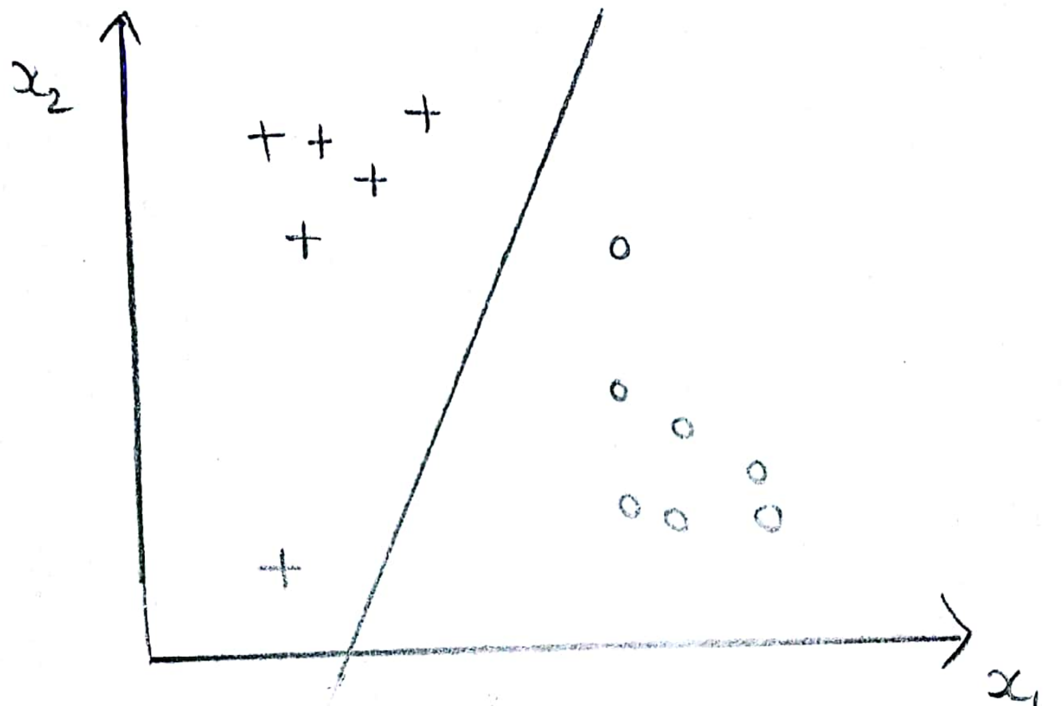
Ans : 2

2 (a) Given that, $P(y=+1 \mid x_i ; w) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$

– Fitting model by max. likelihood, minimizing the

$J(w) = -\ell(w, D_{train})$

where, $\ell(w, D_{train})$ is log likehood on training set.



→ It is observed from graph, data are separated linearly. Thus, logistic regression will find a line that will fit the data perfect.

→ From fig., with the given ~~data~~ data and estimation of model, Classification error on training set = 0.

~~these~~

→ Line is actually not unique. It is indicating decision boundary. (we can even change it normally, i.e, twitch it little bit)

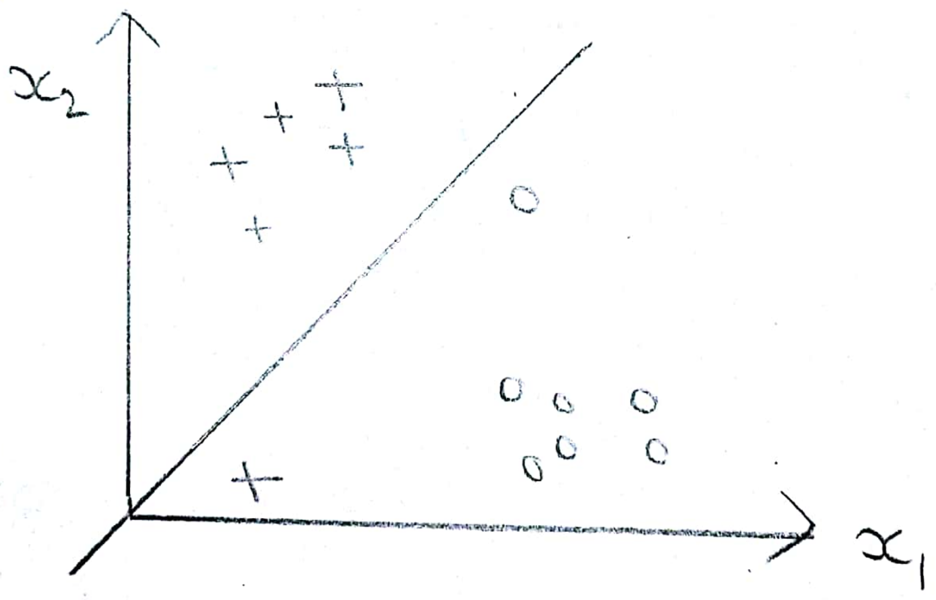
2(b) Now, by regularizing only w_0 and all others parameters are un-regularized. Then, the training error increases. ~~So~~ Then, the boundary will eventually go through origin.

→ As, $w_0 = 0$, point (0, 0) origin will must be on the decision boundary.
Since, at that point,

$$\sigma(w_0 + w_1 x_1 + w_2 x_2) = \sigma(0) = 0.5$$

→ So, regularized logistic regression will find best decision boundary as plotted which passes through (0,0).

→ It will make only one mistake on training data.



→ Moreover, due to this reason in regularized logistic or linear regression, we generally don't penalize the bias term.
(i.e, the weight which corresponding to feature which is always 1)

2(c)

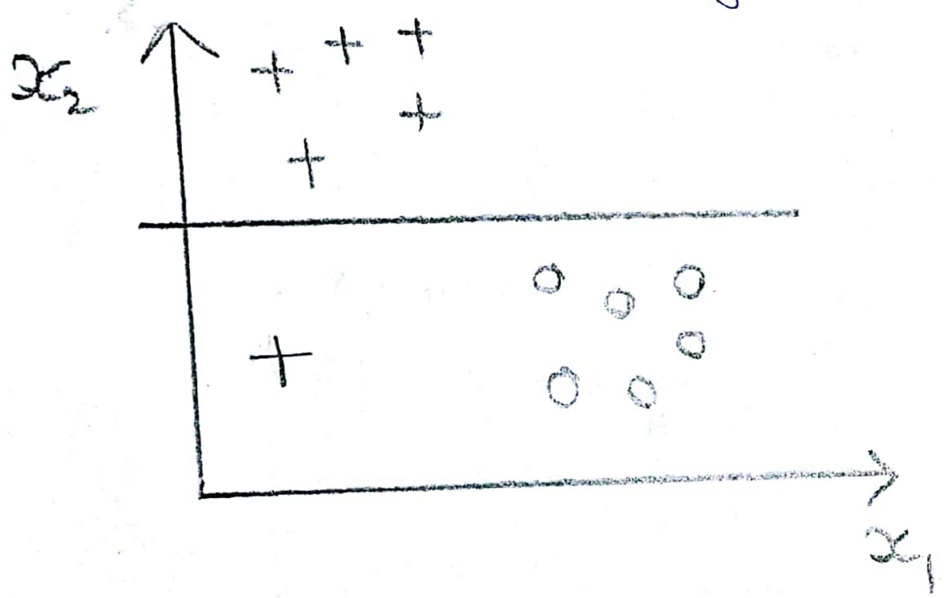
Minimizing, $J_1(w) = -l(w, D_{train}) + \lambda w_1^2$

So, by heavily regularizing w_1, the resulting boundary can rely less and less on values of x_1 and thus, it becomes more horizontal.

→ Also, training data can be seperated with zero training error with horizontal linear separator.

→ From fig.., it is clear that,

Classification errors ~~on training set~~ = 2.

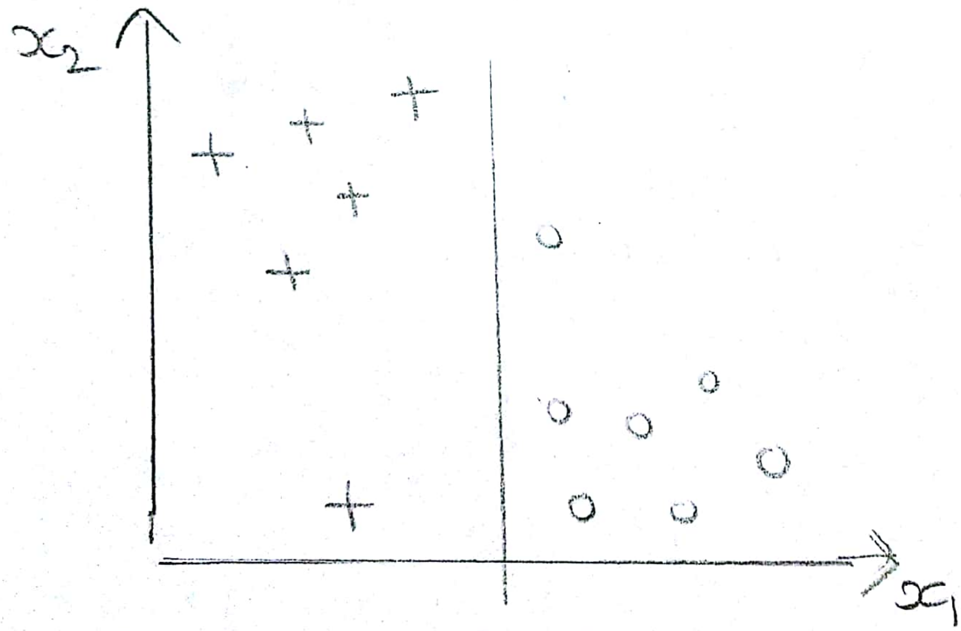


2(d)

→ Now, if we regularize only w_2 parameter, then the resulting boundary will rely less and less on value of x_2 and therefore, it becomes more vertical.

→ So, the decision boundary will become vertical line.

→ Classification error on training set = 0.


x_2
x_1

Ans: 3

Now, $T = 3$.

Credit	Term	Income	α				y	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
Good	3 yrs.	High	$\frac{1}{9}$	~~0.5485~~	0.05	0.0245	Safe	Safe	Safe	Safe
Good	5 yrs.	Low	$\frac{1}{9}$	~~0.5485~~	0.05	0.1011	Risky	Risky	Safe	Risky
Good	5 yrs.	High	$\frac{1}{9}$	~~0.2078~~	0.207	0.1014	Safe	Risky	Safe	Risky
Bad	5 yrs.	High	$\frac{1}{9}$	~~0.5485~~	0.05	0.0245	Risky	Risky	Risky	Risky
Bad	3 yrs.	~~High~~ Low	$\frac{1}{9}$	~~0.5485~~	0.05	0.1011	Safe	Safe	Risky	Safe
Good	5 yrs.	Low	$\frac{1}{9}$	~~0.5485~~	0.05	0.1011	Risky	Risky	Safe	Risky
Bad	3 yrs	High	$\frac{1}{9}$	~~0.2078~~	0.207	0.1014	Risky	Safe	Risky	Safe
Bad	5 yrs.	Low	$\frac{1}{9}$	~~0.5485~~	0.05	0.0245	Risky	Risky	Risky	Risky
Good	3 yrs.	High	$\frac{1}{9}$	~~0.5485~~	0.05	0.0245	Safe	Safe	Safe	Safe

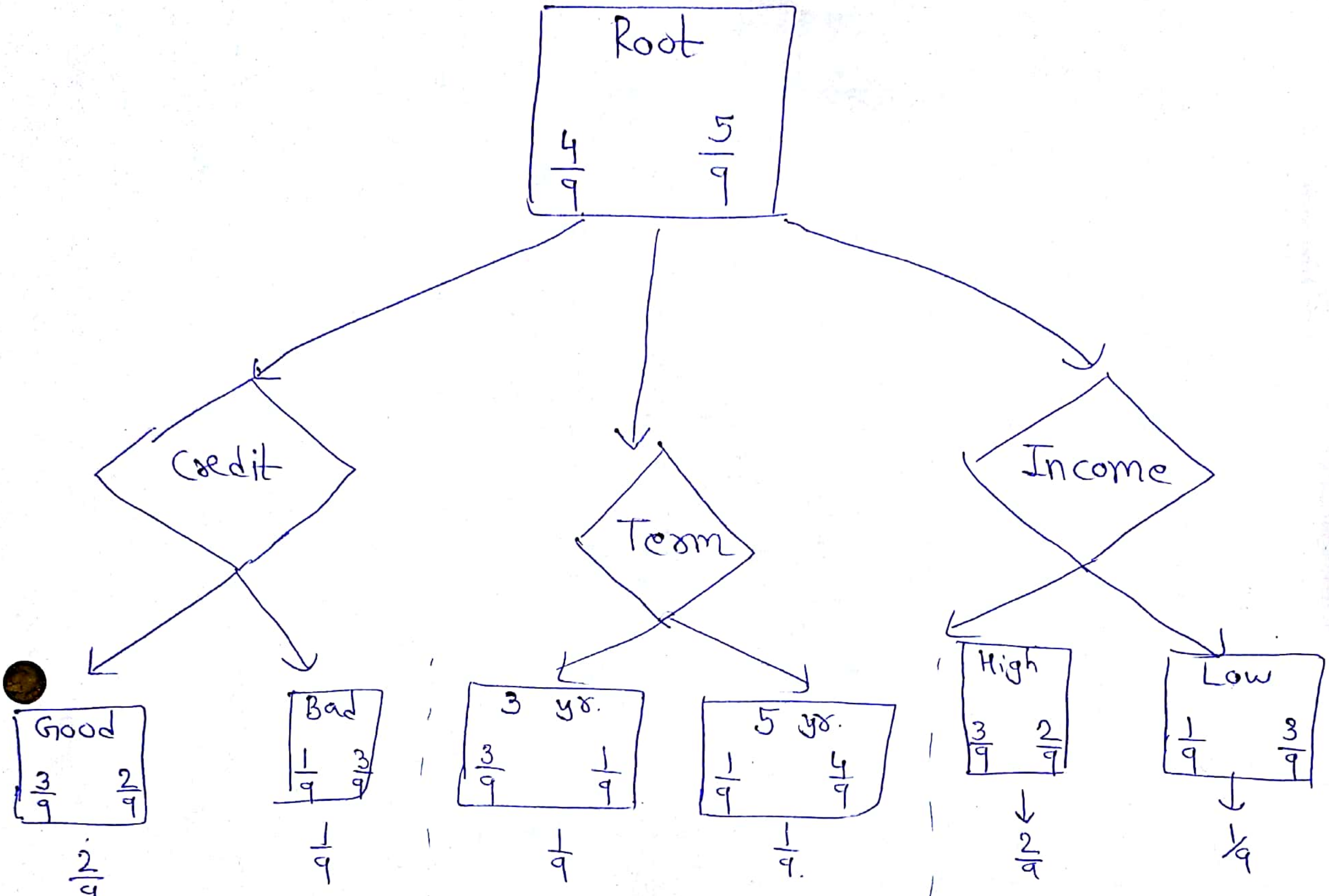
$\boxed{\text{For } t=1}$ $N = 9$

$\therefore \alpha_i = \frac{1}{N} = \frac{1}{9}.$

Firstly, same weights for all points.

So, making trees as follows:—

PTO



$\therefore$ Error $= \dfrac{2/9 + 1/9}{9/9} = 3/9$

Error $= \dfrac{1/9 + 1/9}{9/9} = 2/9$

Error $= \dfrac{2/9 + 1/9}{9/9} = 3/9$

So, "Term" is winner for t = 1.

→ So, for Term majorities are → 3 yr. (3/9, 1/9) → Safe

→ 5 yr. (1/9, 4/9) → Risky.

→ So, all with term -3 yr are predicted to be safe and all with term of 5 yr. are predicted to be risky.

→ Thus, weighted error$[f_1]$ = $\dfrac{\sum_{i=1}^{N} \alpha_i (y_i \neq \hat{y}_i)}{\sum_{i=1}^{N} \alpha_i}$

$$= \frac{1/9 + 1/9}{1} = \frac{2}{9}$$

$$\therefore \hat{w}_1 = \frac{1}{2} \ln \left[\frac{1 - \text{weighted error}[f_1]}{\text{weighted error}[f_1]} \right] = \frac{1}{2} \ln \left[\frac{7/9}{2/9} \right]$$

$\boxed{\hat{w}_1 = 0.6263}$ ← coefficient

→ Now, recomputing weights such that,

$$\alpha_i \leftarrow \begin{cases} \alpha_i e^{-\hat{w}_t}, & \text{if } y_i = \hat{y}_i \\ \alpha_i e^{+\hat{w}_t}, & \text{if } y_i \neq \hat{y}_i \end{cases}$$

→ So, $e^{-\hat{w}_1} = 0.5345$ and $e^{+\hat{w}_1} = 1.8706$

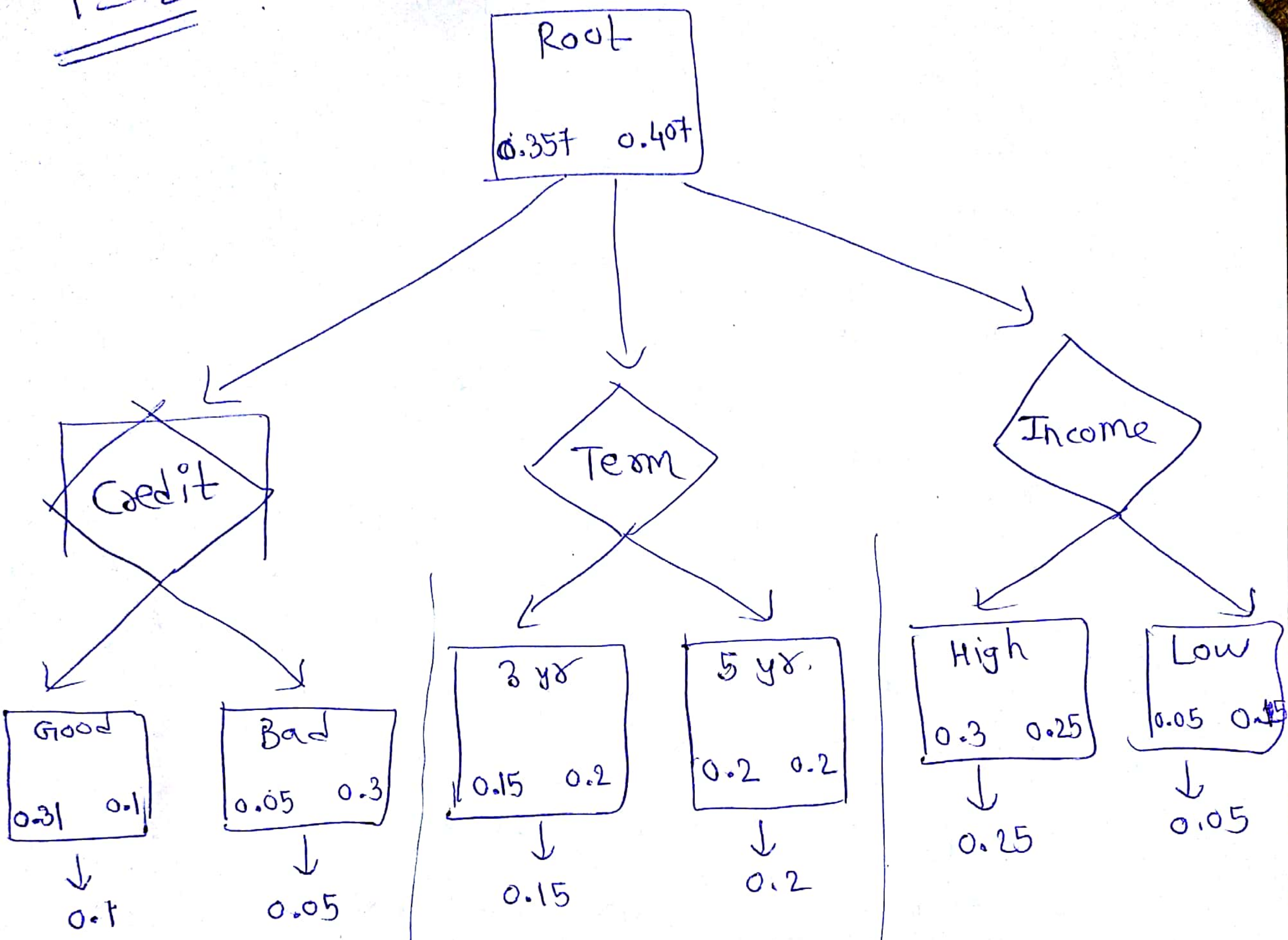
→ Since, α_q is same for all $t=1$.

∴ So, updating weights –

$$\alpha e^{-\hat{w}_1} = 0.059 \cong 0.05 \quad \text{and} \quad \alpha e^{+\hat{w}_1} = 0.207 \cong 0.2$$

→ Now, we have updated weights. Thus, this will form another decision stump for these new values of α.

T=2

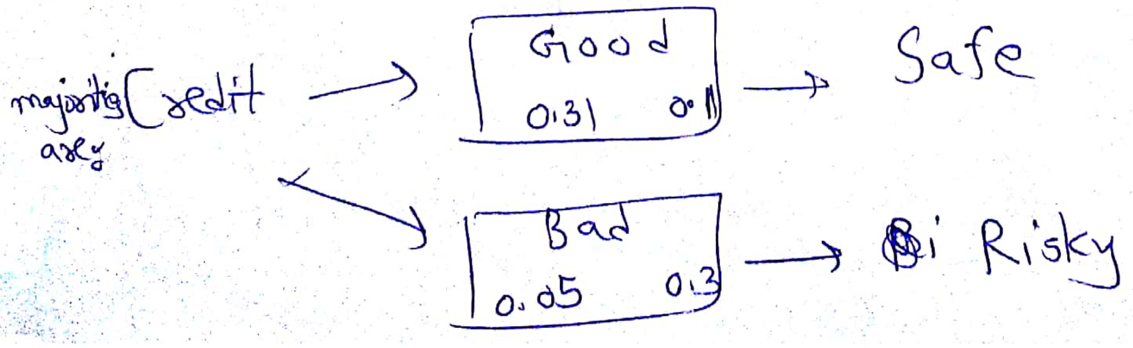


$$\text{Error} = \frac{0.15}{0.764} = 0.1963$$

$$\text{Error} = \frac{0.35}{0.764} = 0.4581$$

$$\text{Error} = \frac{0.30}{0.764} = 0.3926$$

→ So, Credit has lowest error, so it is winner for T=2



So,

weighted error for $f_2 = \frac{0.05 + 0.05 + 0.05}{0.764}$

$= 0.1963$

$\therefore \hat{w_2} = \frac{1}{2} \ln \left[\frac{1 - weighted\_error}{weighted\_error} \right]$

$= \frac{1}{2} \ln \left[\frac{1 - 0.1963}{0.1963} \right]$

$= 0.7047.$

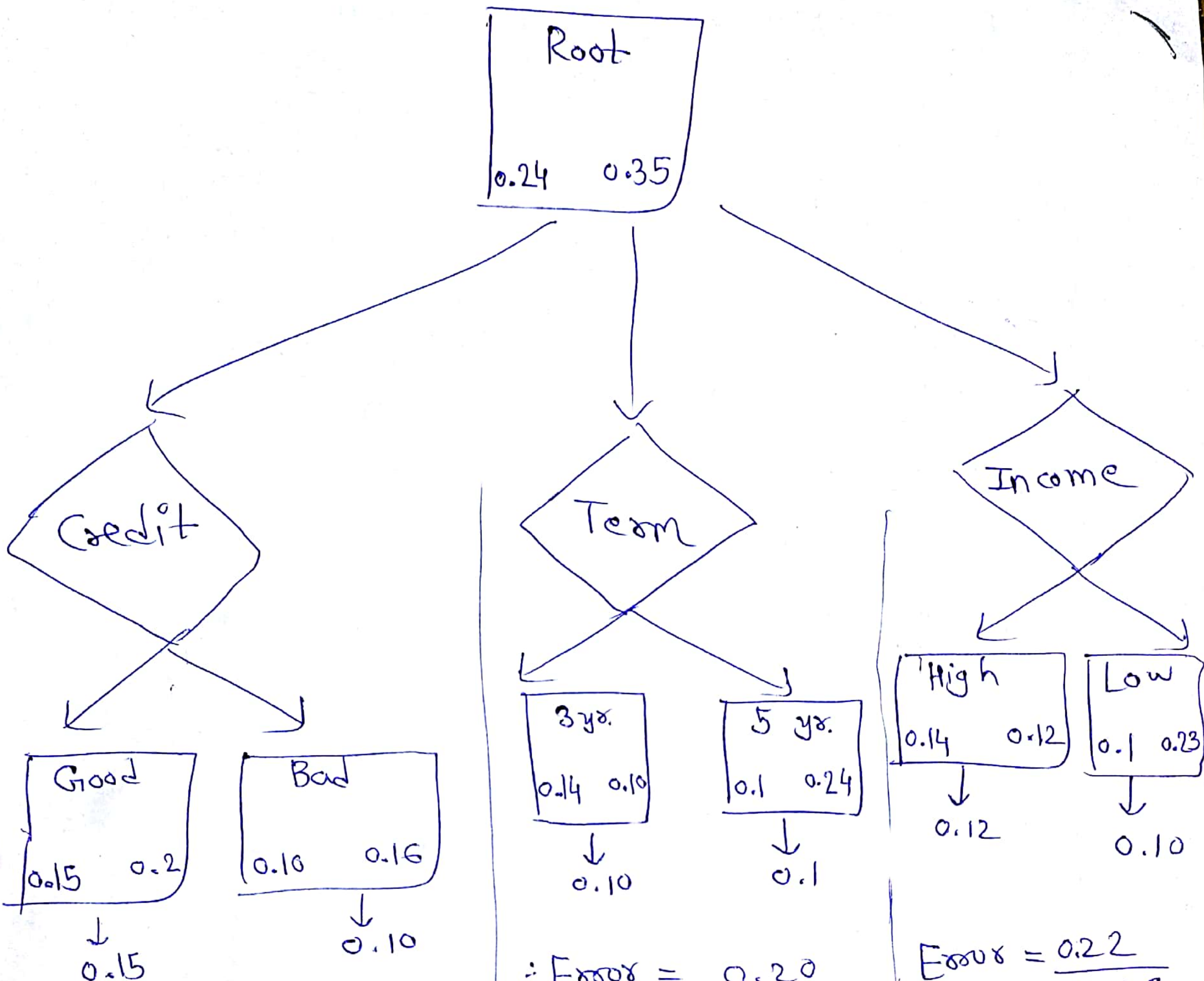
Recomputing weights, so,

$e^{-\hat{w_2}} = 0.4942$ $\quad (y_i = \hat{y}_i)$

$e^{+\hat{w_2}} = 2.0232.$ $\quad (y_i \neq \hat{y}_i)$

→ So, we have updated the values of α. So, we need to form one more decision stump of $T=3$.

$\therefore$ $T=3$



$\therefore$ Error $= \dfrac{0.25}{0.59}$

$= 0.4237$

$\therefore$ Error $= \dfrac{0.20}{0.59}$

$= 0.33$

Error $= \dfrac{0.22}{0.59}$

$= 0.37$

$\therefore$ So, lowest classification error $= 0.33$ is of Term.

$\rightarrow$ So, Term is winner for $T=3$.

So; for Term majorities are:

$\rightarrow$ $\boxed{\begin{matrix} 3\text{ yr.} \\ 0.14 \quad 0.10 \end{matrix}}$ $\rightarrow$ safe

$\rightarrow$ $\boxed{\begin{matrix} 5\text{ yr.} \\ 0.1 \quad 0.24 \end{matrix}}$ $\rightarrow$ Risky

→ So, weighted error of $f_3 = \frac{0.1 + 0.1.}{0.59}$

$= 0.33$

$\therefore \hat{W}_3 = \frac{1}{2} \ln \left[\frac{1 - \text{weighted\_error}}{\text{weighted\_error}} \right]$

$= \frac{1}{2} \ln \left[\frac{1 - 0.33}{0.33} \right]$

$\therefore \boxed{\hat{W}_3 = 0.3540}$

→ So, recomputing weights, $e^{-\hat{W}_3} = 0.7018$ | $e^{+\hat{W}_3} = 1.4247$

old weights	update weights
0.0245	$0.0245 \times 0.7018 = 0.017$
0.1011	$0.1011 \times 0.7018 = 0.071$
0.1014	$0.1014 \times 1.4247 = 0.14247$
0.0245	$0.0245 \times 0.7018 = 0.017$
0.1011	$0.1011 \times 0.7018 = 0.071$
0.1011	$0.1011 \times 0.7018 = 0.071$
0.1014	$0.1014 \times 1.4247 = 0.144$
0.0245	$0.0245 \times 0.7018 = 0.017$
0.0245	$0.0245 \times 0.7018 = 0.017$

→ So, finally,

$$\hat{y} = \text{sign}\left(\sum_{t=1}^{T} \hat{w}_t f_t(x)\right).$$

$$= \text{sign}\left(\sum_{t=1}^{3} \hat{w}_t f_t(x)\right)$$

$$= \text{sign}\left[\hat{w}_1 f_1(x) + \hat{w}_2 f_2(x) + \hat{w}_3 f_3(x)\right]$$

So, ~~= sign [(0.6263)~~ for each datapoint,

$\hat{y}$

$0.6263(+1) + 0.70(+1) + 0.35(+1) = 1.6763 \Rightarrow$ Safe

$0.6263(-1) + 0.70(+1) + 0.35(-1) = (-0.2763) \Rightarrow$ Risky

$0.6263(-1) + 0.70(+1) + 0.35(-1) = (-0.2763) \Rightarrow$ Risky

$0.6263(-1) + 0.70(-1) + 0.35(-1) = (-1.6763) \Rightarrow$ Risky

$0.6263(+1) + 0.70(-1) + 0.35(+1) = 0.2763 \Rightarrow$ Safe

$0.6263(-1) + 0.70(+1) + 0.35(-1) = (-0.2763) \Rightarrow$ Risky

$0.6263(+1) + 0.70(-1) + 0.35(+1) = 0.2763 \Rightarrow$ Safe

$0.6263(-1) + 0.70(-1) + 0.35(-1) = (-1.6763) \Rightarrow$ ~~Safe~~ Risky

$0.6263(+1) + 0.70(+1) + 0.35(+1) = +1.6763 \Rightarrow$ Safe

→ Thus, we can say from true obs^{n} and predicted obs^{n} of target variable are same for all, except 3rd and 7th, which has incorrect classification

Scanned by CamScanner