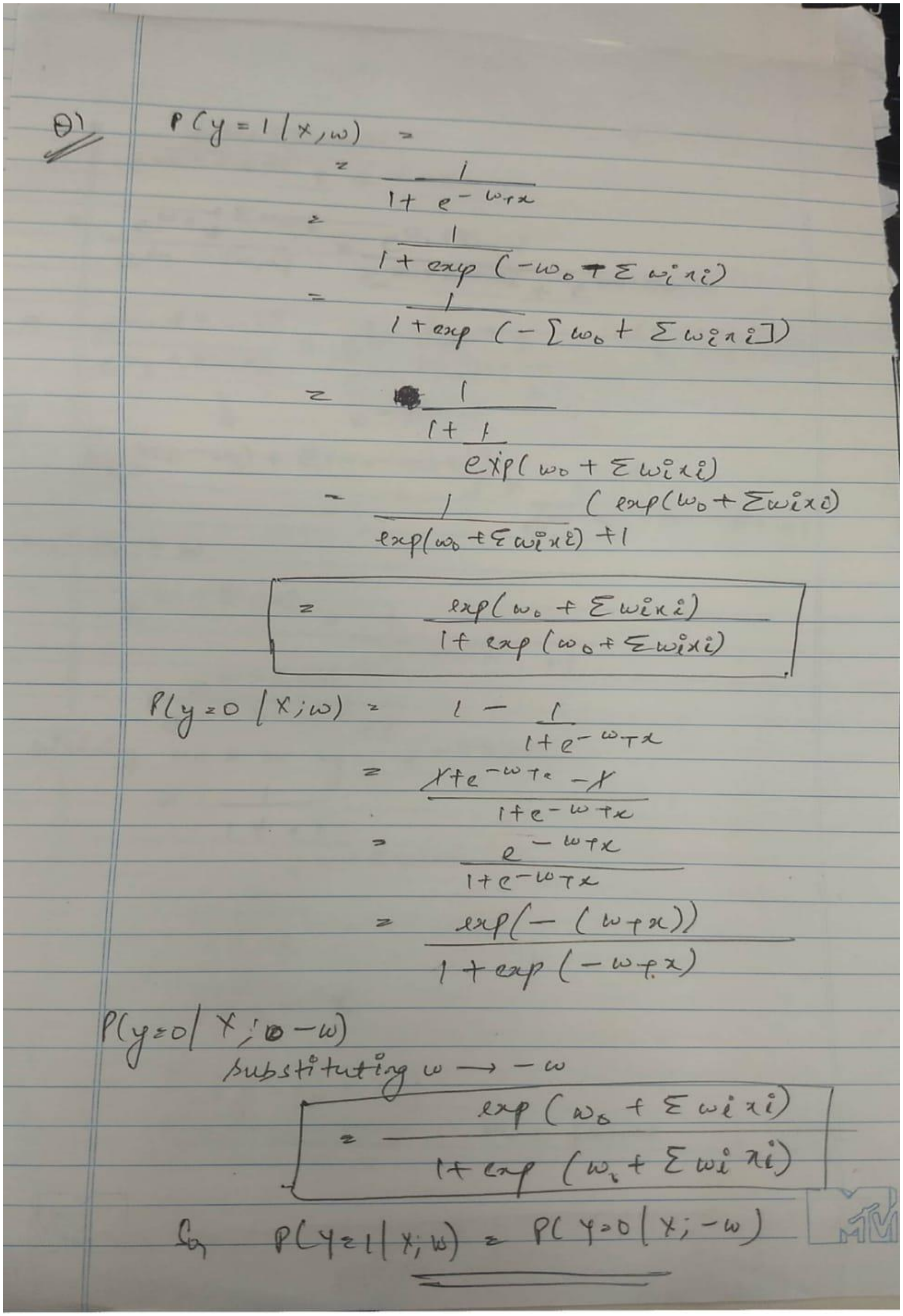
Assignment 2: MLPS(adewan@andrew.cmu.edu)

Question 1: Conceptual: Logistic Regression



Q1)

$$P(y=1|x,w) = \frac{1}{1+e^{-w_T x}}$$

$$= \frac{1}{1+\exp(-w_0 + \Sigma w_i x_i)}$$

$$= \frac{1}{1+\exp(-[w_0+\Sigma w_i x_i])}$$

$$= \frac{1}{1+\frac{1}{\exp(w_0+\Sigma w_i x_i)}}$$

$$= \frac{1}{\exp(w_0+\Sigma w_i x_i)+1} \quad (\exp(w_0+\Sigma w_i x_i)$$

$$\boxed{= \frac{\exp(w_0+\Sigma w_i x_i)}{1+\exp(w_0+\Sigma w_i x_i)}}$$

$$P(y=0|X;w) = 1 - \frac{1}{1+e^{-w_T x}}$$

$$= \frac{\cancel{1}+e^{-w_T x} - \cancel{1}}{1+e^{-w_T x}}$$

$$= \frac{e^{-w_T x}}{1+e^{-w_T x}}$$

$$= \frac{\exp(-(w_T x))}{1+\exp(-w_T x)}$$

$P(y=0|X;-w)$

substituting $w \rightarrow -w$

$$\boxed{= \frac{\exp(w_0+\Sigma w_i x_i)}{1+\exp(w_0+\Sigma w_i x_i)}}$$

So $\underline{\underline{P(y=1|X;w) = P(y=0|X;-w)}}$

(2) $P(y=0|x;w_0,w_1) = \frac{e^{w_0^T x}}{e^{w_0^T x}+e^{w_1^T x}}$

$$= \frac{e^{w_0+\Sigma w_0 x_i}}{e^{w_0+\Sigma w_0 x_i}+e^{w_1+\Sigma w_1 x_i}}$$

$$= \frac{e^{w_0+\Sigma w_0 x_i}}{e^{w_1+\Sigma w_1 x_i}} \times \frac{e^{w_1+\Sigma w_1 x_i}}{e^{w_0+\Sigma w_0 x_i}+e^{w_1+\Sigma w_1 x_i}}$$

(multiply & divide by $e^{w_1+\Sigma w_1 x_i}$)

$$= \frac{e^{w_0+\Sigma w_0 x_i}}{e^{w_1+\Sigma w_1 x_i}} \times \frac{\cancel{e^{w_1+\Sigma w_1 x_i}}\ 1}{\frac{e^{w_0+\Sigma w_0 x_i}}{e^{w_1+\Sigma w_1 x_i}}+1}$$

$$\downarrow$$

$$\left[e^{(w_0-w_1)+\Sigma(w_0-w_1)x_i}\right] \times \frac{1}{e^{(w_0-w_1)+\Sigma(w_0-w_1)x_i}+1}$$

As $w_0 - w_1 = w$

$$e^{w+\Sigma w x_i} \times \frac{1}{e^{w+\Sigma w x_i}+1}$$

$$\frac{e^{w+\Sigma w x_i}}{e^{w+\Sigma w x_i}+1}$$

Dividing Nr & Dr by $e^{w+\Sigma w x_i}$

$$= \frac{1}{1+\frac{1}{e^{w+\Sigma w_i x_i}}}$$

$$= \frac{1}{1+e^{-(w+\Sigma w_i x_i)}}$$

$$= \frac{1}{1+e^{w^T x}}$$

$$= \underline{\underline{Pr(y=1|x;w)}}$$

Hence proved

3.

③ $$L(w) = \prod_{i=1}^{N} p(y^{(i)} | x^{(i)} ; w)$$

We can combline two terms like

$$p(y^i | x^i ; w) = p(y^{(i)} = 1 | x^{(i)} ; w)^{y^{(i)}} \; p(y^{(i)} = 0 | x^{(i)} ; w)^{1-y^{(i)}}$$

Log likehood is

$$\mathcal{L}(w) = \log L(w)$$

$$= \underset{w}{\operatorname{argmax}} \, \alpha(w) = \underset{w}{\operatorname{argmax}} \log L(w)$$

$$\operatorname{argmax} \, l(w)$$

$$= \underset{w}{\operatorname{argmax}} \log \prod_{i=1}^{N} p(y^{(i)} | x^i ; w)$$

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(y^i | x^i ; w)$$

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \log \left[p(y=1 / x^i ; w)^{y^i} \, p(y=0 | x^i ; w)^{1-y^i} \right]$$

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \log [p(y=1 / x^i ; w)^{y^i} + \sum \log [p(y=0 | x^i , w)^{1-y^i}$$

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} y^i \log [p(y=1 / x^i ; w))] + (1-y^i) \log (p(y=0 | x^i ; w)$$

$$= \underset{N}{\operatorname{argmax}} \sum_{i=1}^{N} y^i \log [p(y=1 / x^i ; w] + \log (p(y=0 | x^i ; w) - y^i \log (y=0 | x^i ; w)$$

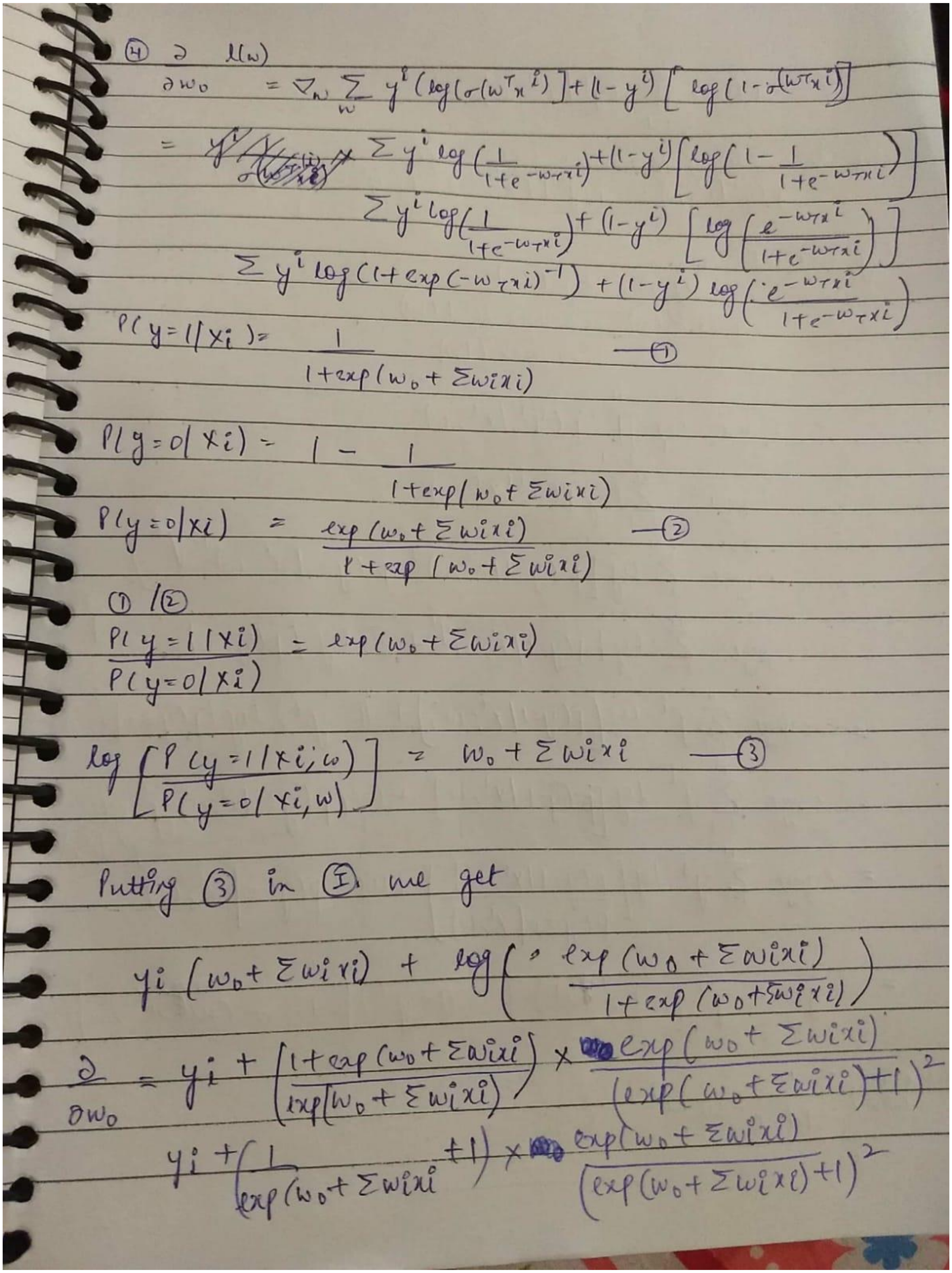
Ⓘ: $$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} y^i \log \left[\frac{p(y=1 / x^i ; w)}{p(y=0 | x^i ; w)} \right] + \log (p(y=0 | x^i ; w)$$

$$= \underset{w}{\operatorname{argmax}} \; y^i [\sigma(w^T x^i)] + (1-y^i) [1 - \sigma(w^T x^i)]$$

where $\sigma_{(w^T x^i)} = \dfrac{1}{1 + e^{-w^T x^i}}$

× ———— ×

4.



④ $$\frac{\partial \quad l(w)}{\partial w_0} = \nabla_w \sum_w y^i (\log(\sigma(w^T x^i)] + (1-y^i)\left[\log(1-\sigma(w^T x^i))\right]$$

$$= \cancel{y^i \frac{1}{\sigma(w^T x^i)}} \times \sum y^i \log\left(\frac{1}{1+e^{-w_T x i}}\right) + (1-y^i)\left[\log\left(1-\frac{1}{1+e^{-w_T x i}}\right)\right]$$

$$\sum y^i \log\left(\frac{1}{1+e^{-w_T x i}}\right) + (1-y^i)\left[\log\left(\frac{e^{-w_T x^i}}{1+e^{-w_T x i}}\right)\right]$$

$$\sum y^i \log(1+\exp(-w_T x i)^{-1}) + (1-y^i)\log\left(\frac{e^{-w_T x i}}{1+e^{-w_T x i}}\right)$$

$$P(y=1|x_i) = \frac{1}{1+\exp(w_0 + \sum w_i x_i)} \quad \text{—} ①$$

$$P(y=0|x_i) = 1 - \frac{1}{1+\exp(w_0 + \sum w_i x_i)}$$

$$P(y=0|x_i) = \frac{\exp(w_0 + \sum w_i x_i)}{1+\exp(w_0 + \sum w_i x_i)} \quad \text{—} ②$$

① / ②

$$\frac{P(y=1|x_i)}{P(y=0|x_i)} = \exp(w_0 + \sum w_i x_i)$$

$$\log\left[\frac{P(y=1|x_i;w)}{P(y=0|x_i,w)}\right] = w_0 + \sum w_i x_i \quad \text{—} ③$$

Putting ③ in Ⓘ we get

$$y_i(w_0 + \sum w_i x_i) + \log\left(\frac{\exp(w_0 + \sum w_i x_i)}{1+\exp(w_0 + \sum w_i x_i)}\right)$$

$$\frac{\partial}{\partial w_0} = y_i + \left(\frac{1+\exp(w_0 + \sum w_i x_i)}{\exp(w_0 + \sum w_i x_i)}\right) \times \frac{\exp(w_0 + \sum w_i x_i)}{(\exp(w_0 + \sum w_i x_i)+1)^2}$$

$$y_i + \left(\frac{1}{\exp(w_0 + \sum w_i x_i} + 1\right) \times \frac{\exp(w_0 + \sum w_i x_i)}{(\exp(w_0 + \sum w_i x_i)+1)^2}$$

$$yi + \frac{\cancel{1+e^{w_0+\Sigma w_i x_i}}}{\cancel{e^{w_0+\Sigma w_i x_i}}} \times \frac{1\,\cancel{e^{w_0+\Sigma w_i x_i}}}{(e^{w_0+\Sigma w_i x_i}+1)^{\cancel{2}}}$$

$$yi + \frac{1.}{(e^{w_0+\Sigma w_i x_i}+1)}$$

$$\boxed{\frac{\partial L(\theta)}{\partial w_0} = yi + \frac{1}{\exp(w_0+\Sigma w_i x_i)+1}}$$

$$\frac{\partial}{\partial w_j} L(\theta) = \frac{\partial}{\partial w_j}\left[y_i \log\left[\sigma(w^T x^i)\right] + 1 - y^i\left(1-\sigma(w^T x^i)\right)\right.$$

$$= \Sigma\, y^i \left(\frac{1}{\sigma(w^T x^i)} \times \sigma(w^T x^i)(1-\sigma(w^T x^i)\right) \times x^i$$

$$+ (1-y^i)\left[\frac{\cdot 1}{1-\sigma(w^T x^i)} \times -1 \times \left(\sigma^{w^T x^i}\right)(1-\sigma(w^T x^i)\right) \times x^i$$

$$= \Sigma\; y^i\left((1-\sigma^{w^T x^i})\right)x^i + (1-y^i)(-1)\left((\sigma^{w^T x^i})x^i\right)$$

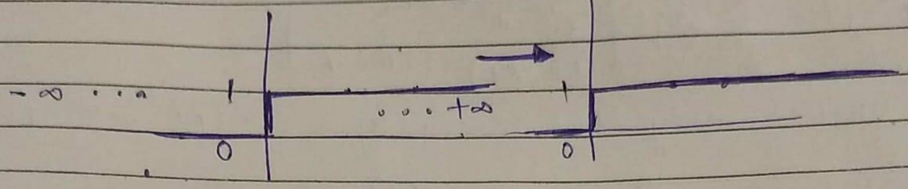
$$= \Sigma y^i x^i - y^i x^i \sigma^{w^T x^i} + (-1)\left[\sigma^{w^T x^i}\right)x^i - y^i \sigma^{w^T x^i} x^i$$

$$= \Sigma y^i x^i - y^i x^i \cancel{\sigma}(w^T x^i) - \sigma(w^T x^i) x^i + y^i x^i \cancel{\sigma}(w^T x^i)$$

$$= \Sigma y^i x^i - \sigma(w^T x^{i}) x^{(i)}$$

$$\frac{\partial L(\theta)}{w_j} = \Sigma\left[y^i - \sigma(w^T x^i)\right] x^{(i)}$$

Specifically we can keep shifting step function to left or right and still have it pass through Data points



So, the ideal parameters are of ∞ magnitude and MLE ranges from $-\infty$ to $+\infty$, and any w with descion boundary which lies between 2 seperable class gives same likelihood. Hence, there can be multiple solution ranging from $-\infty$ to $+\infty$.

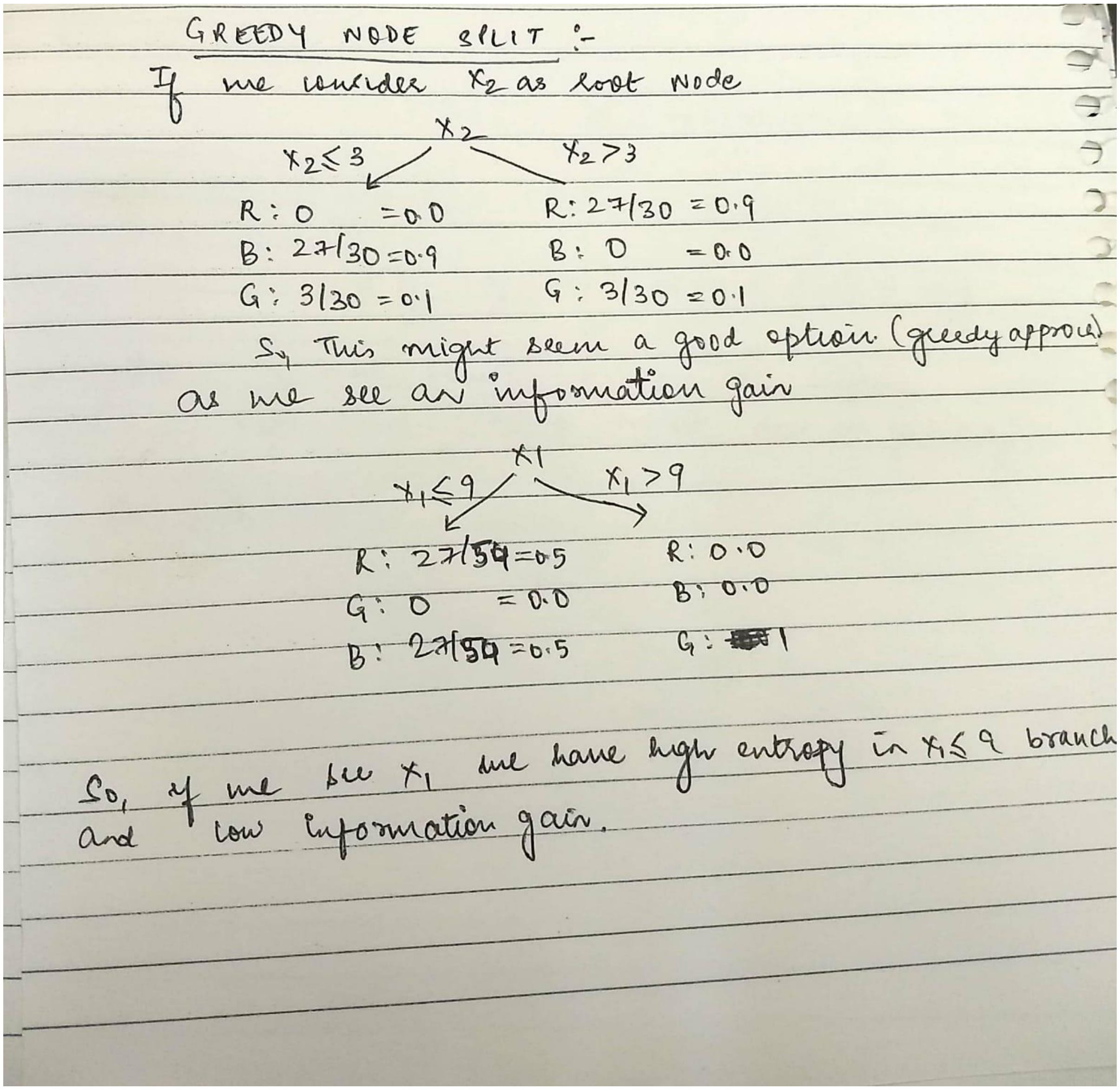
Question 2: Conceptual: Decision Trees [25 points]

1.

a) True: Yes, it is possible to achieve a 100% train accuracy if we memorize the train dataset completely in decision tree we can do this my creating tree paths for every dataset row which is worst approach.

b) False: No, a same node can be used in decision tree more than once in a different path or may be it can be used in left tree branch once and in the right branch later.

2.a)



GREEDY NODE SPLIT :-

If we consider X_2 as root Node

X_2

$X_2 \leq 3$ | $X_2 > 3$

R : 0 = 0.0 | R: 27/30 = 0.9

B: 27/30 = 0.9 | B : 0 = 0.0

G : 3/30 = 0.1 | G : 3/30 = 0.1

So, This might seem a good option (greedy approa)
as we see an information gain

X1

$X_1 \leq 9$ | $X_1 > 9$

R: 27/54 = 0.5 | R: 0.0

G: 0 = 0.0 | B: 0.0

B: 27/54 = 0.5 | G: 1

So, if we see X_1 we have high entropy in $X_1 \leq 9$ branch
and low information gain.

X_2 is Root node:-

$$-\frac{1}{2}\left[\overset{0}{\cancel{0\log 0}} + 0.9\log 0.9 + 0.1\log 0.1\right]$$

$$-\frac{1}{2}\left[\overset{0}{\cancel{0\log 0}} + 0.9\log 0.9 + 0.1\log 0.1\right]$$

$$H_{X_2} = -\left[0.9\log 0.9 + 0.1\log 0.1\right] = \underline{\underline{0.46}}$$

X_1 is Root Node:-

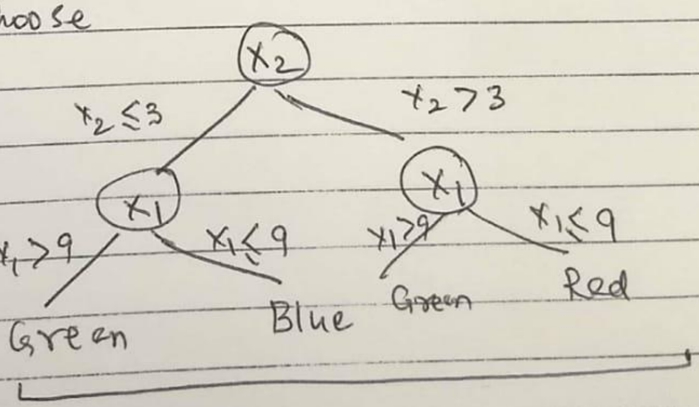
$$-0.9\left[0.5\log 0.5 + 0\cancel{\log}0 + 0.5\log 0.5\right]$$

$$-0.1\left[0\cancel{\log}0 + 0\cancel{\log}0 + 1\cancel{\log}1\right]$$

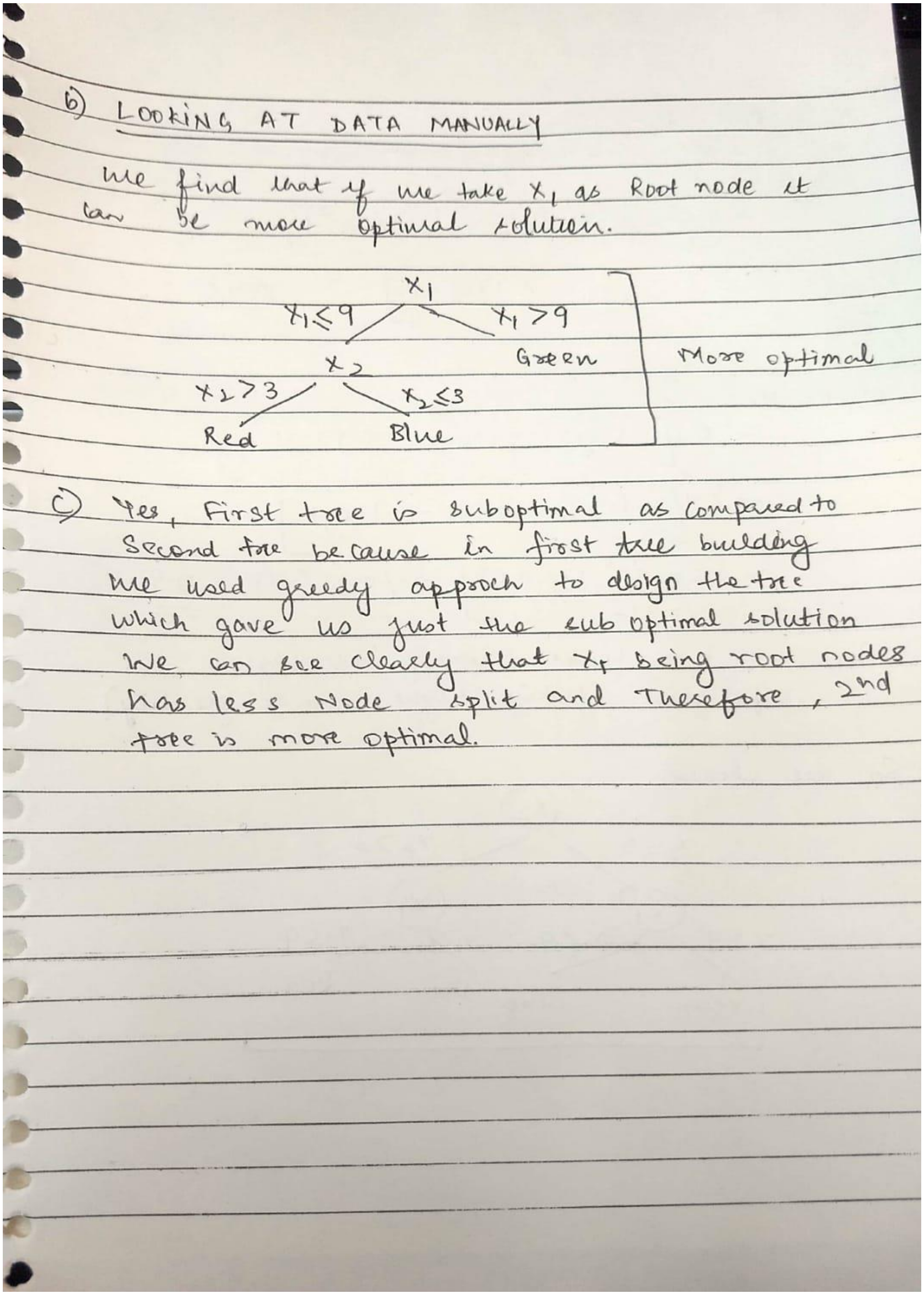
$$-0.9\left[\log(0.5)\right] \Rightarrow -(0.9)(-1) = \underline{\underline{0.9}}$$

So, X_2 has lower entropy or higher info gain (looking right now or greedily).

So, we choose



2 b)



b) LOOKING AT DATA MANUALLY

We find that if we take X_1 as Root node it can be more optimal solution.

X_1
$X_1 \leq 9$
$X_1 > 9$
X_2
Green
$X_2 > 3$
$X_2 \leq 3$
Red
Blue
More optimal

c) Yes, First tree is suboptimal as compared to Second tree because in first tree building we used greedy approch to design the tree which gave us just the sub optimal solution. We can see clearly that X_1 being root nodes has less Node split and Therefore, 2nd tree is more optimal.

3.

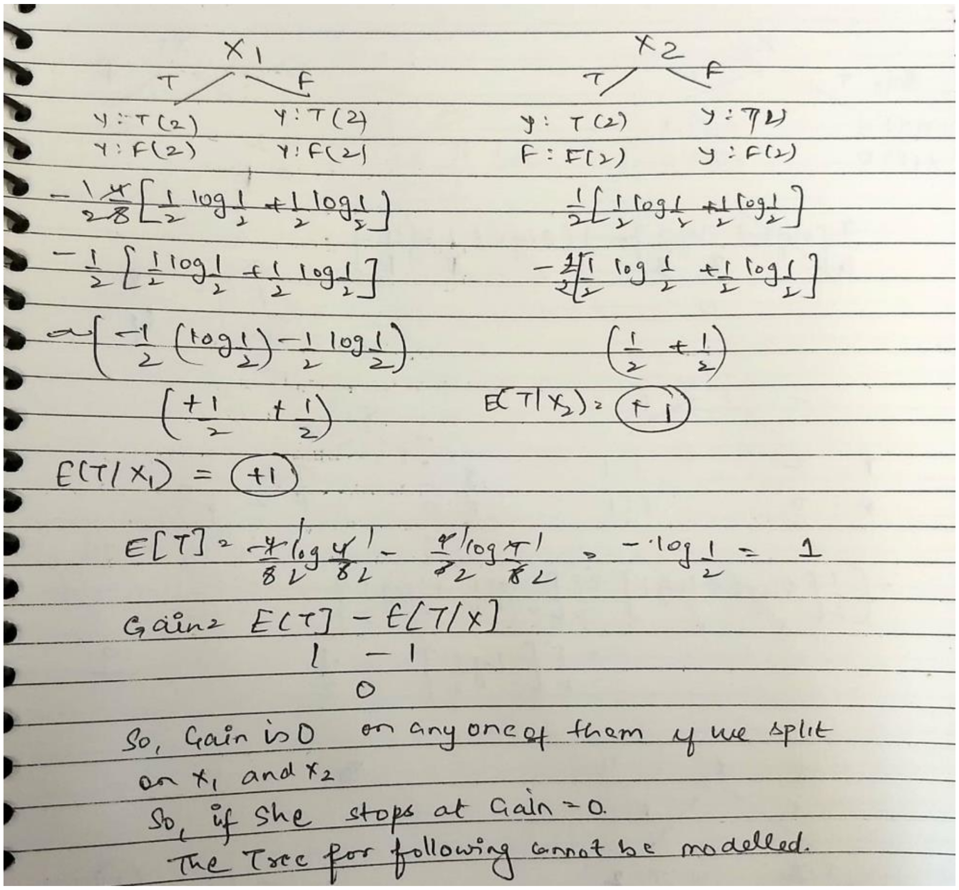
a)As she is considering splitting only on nodes where information gain is positive and not splitting on nodes which have 0 information gain.(as info gain can be either 0 or positive and cannot be negative). So she is not letting her tree grow when IG=0

The following dataset can be used.

X1	X2	y
F	T	T
T	T	T
F	T	F
T	T	F
F	F	T
T	F	T
F	F	F
T	F	F

b) The flaw is in the termination condition: Terminating when IG=0 which is incorrect.

C) Correction to the flaw is that we need to consider splitting on nodes that have 0 info gain too if that is the best possible available node split.



X1

T F

y: T(2) y: T(2)

y: F(2) y: F(2)

$-\frac{\cancel{1}4}{\cancel{2}8}\left[\frac{1}{2}\log\frac{1}{2}+\frac{1}{2}\log\frac{1}{2}\right]$

$-\frac{1}{2}\left[\frac{1}{2}\log\frac{1}{2}+\frac{1}{2}\log\frac{1}{2}\right]$

$\sim\left(-\frac{1}{2}\left(\log\frac{1}{2}\right)-\frac{1}{2}\log\frac{1}{2}\right)$

$\left(+\frac{1}{2}+\frac{1}{2}\right)$

$E(T/X_1) = \boxed{+1}$

X2

T F

y: T(2) y: T(2)

F: F(2) y: F(2)

$\frac{1}{2}\left[\frac{1}{2}\log\frac{1}{2}+\frac{1}{2}\log\frac{1}{2}\right]$

$-\frac{1}{2}\left[\frac{1}{2}\log\frac{1}{2}+\frac{1}{2}\log\frac{1}{2}\right]$

$\left(\frac{1}{2}+\frac{1}{2}\right)$

$E(T|X_2) = \boxed{+1}$

$E[T] = -\frac{4}{8}\log\frac{4}{8}-\frac{4}{8}\log\frac{4}{8} = -\log\frac{1}{2} = 1$

Gain = $E[T] - E[T/X]$

$1 - 1$

0

So, Gain is 0 on any one of them if we split on X_1 and X_2

So, if she stops at Gain = 0.

The Tree for following cannot be modelled.

4.

i) $2^{(v-1)} -1$

ii) $v(2^{(v-1)} -1)$

iii) $(v-1) (2^{(v-1)} -1)$

5.

Q5 i) QUANTIFY

Chi square Test is a Test of independence

It gives a probability of seeing the data of atleast this level of assouation if two variables were independent.

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{Expected}}$$

The larger X^2 means more they are related.

(ii) HYPERPARAMETER

MaxPChance is the hyperparameter / magic parameter that needs to be tuned to come up with a prunned tree.

(iii) Process

step1: Build the whole decison tree (unprunned)

Step2: once fully grown start to prune

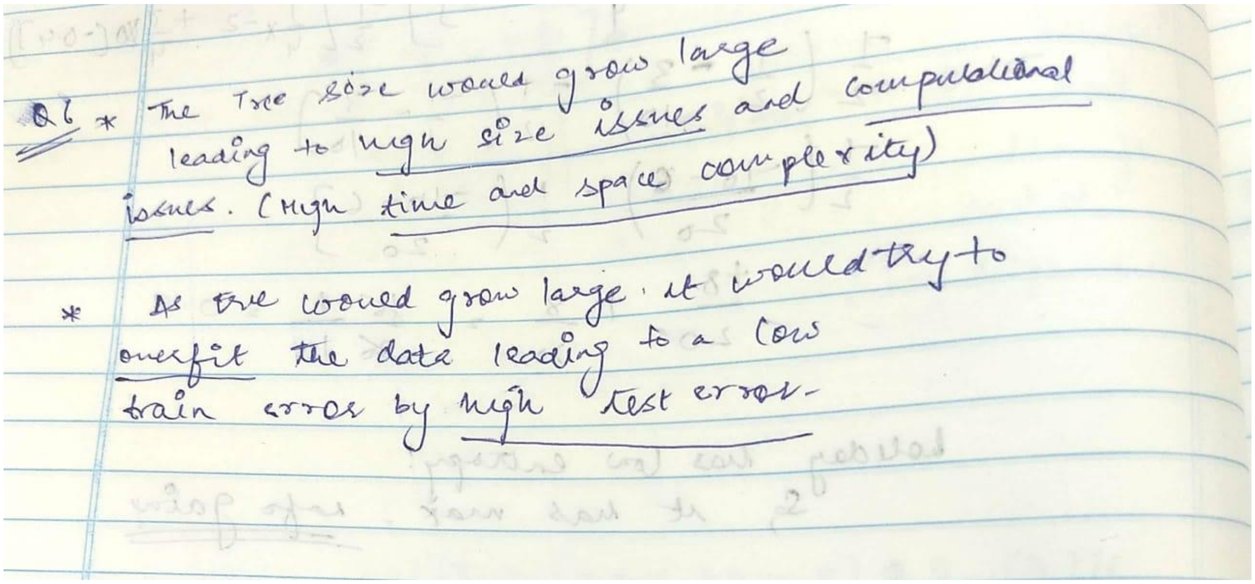
Down → up. (i) Delete splits where P_{chance} > Maxpchance

(ii) continue deleting untill all nodes have P_{chance} <= Maxp chance.

Maxp chance is the threshold p chance value and all nodes should have pchance < or equal to MaxPchance

'Pchance' is nothing but the probability of chance. of seeing data if 2 variable are independent and is calculated from X^2.

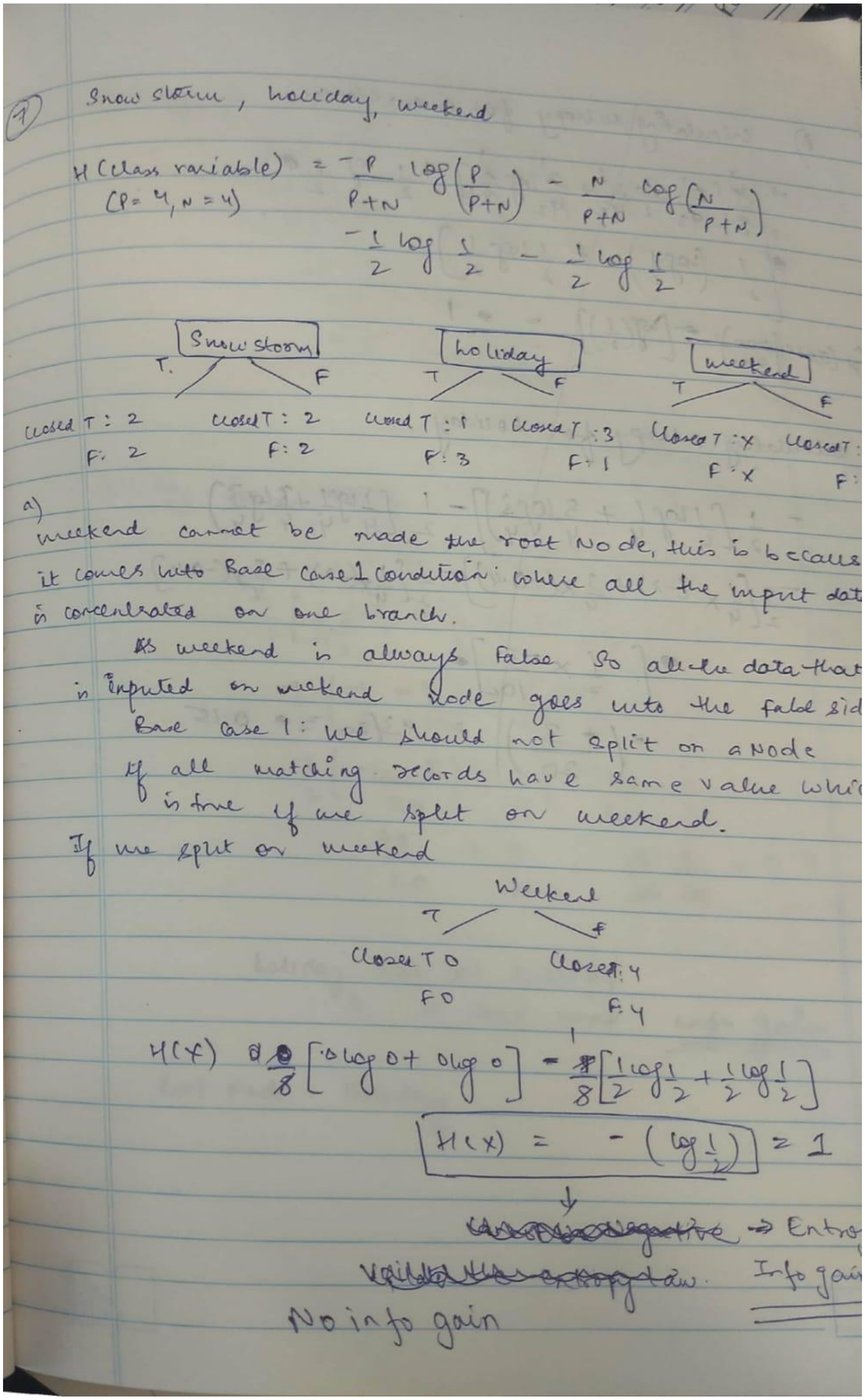
6.



Q6 * The Tree size would grow large leading to high size issues and computational issues. (High time and space complexity)

* As tree would grow large. it would try to overfit the data leading to a low train error by high test error.

7.



(7) Snow storm, holiday, weekend

H (class variable) $= -\frac{P}{P+N}\log\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log\left(\frac{N}{P+N}\right)$

(P = 4, N = 4)

$-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$

Snow storm

T. F

closed T : 2
F: 2

closed T : 2
F: 2

holiday

T F

closed T : 1
F: 3

closed T : 3
F: 1

weekend

T F

closed T : x
F : x

closed T :
F :

a)
weekend cannot be made the root Node, this is becaus it comes into Base case 1 condition: where all the input dat is concentrated on one branch.

As weekend is always False So all the data that is inputed on weekend Node goes into the false sid

Base case 1: we should not split on a Node if all matching records have same value whic is true if we split on weekend.

If we split on weekend

Weekend

T F

Closed T 0
F 0

Closed T: 4
F: 4

$H(x)\ \frac{0}{8}\left[0\log 0 + 0\log 0\right] - \frac{\overset{1}{\cancel{8}}}{8}\left[\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right]$

$\boxed{H(x) = -\left(\log\frac{1}{2}\right)} = 1$

↓

~~negative~~ → Entro

~~[illegible] entropy draw.~~ Info gain

No info gain

entropy for snow storm

$$-\frac{4}{8}\left[\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right] - \frac{4}{8}\left[\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right]$$

$$\frac{-1}{2}\left(\frac{1}{2}\times -1 + \frac{1}{2}\times -1\right) - \frac{1}{2}\left(\frac{-1}{2} + \frac{-1}{2}\right)$$

$$\frac{-1}{2}\left(\frac{-1}{2} - \frac{1}{2}\right) - \frac{1}{2}(-1)$$

$$\frac{-1}{2}(-1) - \frac{1}{2}(-1)$$

$$\therefore 1 + 1 = 2$$

entropy for holiday

$$\frac{-1}{2}\left[\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}\right] - \frac{1}{2}\left[\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}\right]$$

$$\frac{-1}{2}\left[\frac{1}{\cancel{4}_{2}}\times(-\cancel{2}) + \frac{3}{4}\times(-0.4]\right] - \frac{1}{2}\left[\frac{1}{4}\times -2 + \frac{3}{4}\times(-0.4)\right]$$

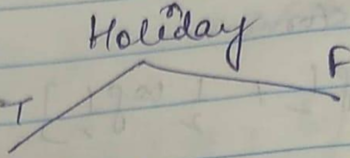
$$\frac{-1}{2}\left(\frac{-1}{2} - \frac{3}{10}\right) - \frac{1}{2}\left(\frac{-1}{2} - \frac{3}{10}\right)$$

$$\frac{-1}{2}\left(\frac{-10-6}{20}\right) - \frac{1}{2}\left(\frac{-10-6}{20}\right]$$

$$\frac{+8}{20} + \frac{8}{20} = \frac{\cancel{16}}{\cancel{20}}\frac{8}{10} = 0.8$$

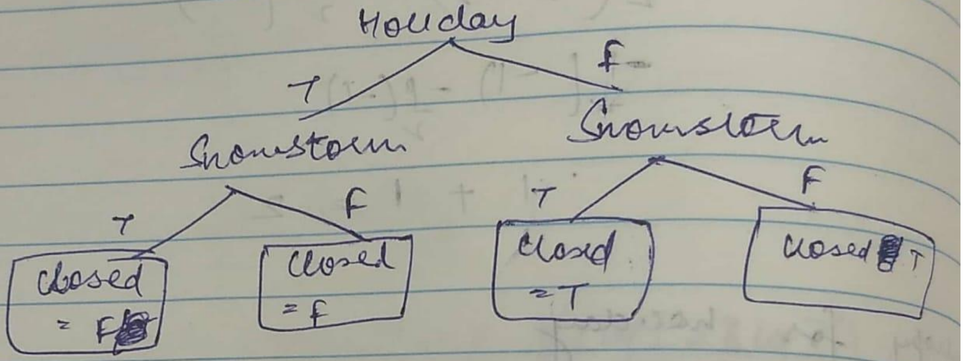
holiday has low entropy.
So, it has max. info gain

Root Node : Holiday



As weekend split does not offer any infor

So, Snowstorm split is only option



Question3: Model Evaluation and Model Selection(Conceptual)

1)

a) False:

The best value of hyperparameter is not the average of hyper parameter values for each fold. But to evaluate the best possible values of hyperparameter we use techniques like cross validation and Grid search that exhaustively searches for best possible value of hyperparameter by plugging various values on a hyperparameter grid and then reports the value that produces best evaluation metric value (like accuracy score)

b) True

c) False:

No, train accuracy score can never give us a true picture of test accuracy score. A very low train accuracy score can still have a high-test accuracy and vic-a-versa.

d) True

e) True:

Yes, if the model is trained on the training data and it leads to high train accuracy if as the model gets trained on the dataset. Then cross validation accuracy will be significantly dropped as it reports the true accuracy of the model without any memorization.

f) False:

Cross validation accuracy basically measures the accuracy of the model by taking various folds of data but still it is the part of training set as a result it reports an unbiased and true train accuracy for the model. This cross-validation accuracy is in fact higher than test accuracy as test data has unobservable characteristics which the model has not been perfectly trained on which adversely affects test accuracy.

g) False:

Numerical hyperparameter are difficult to optimize but ranging over various values to calculate accuracy and keep on incrementing the hyper parameter in some step size and recording the accuracy we will find a turning point(sweet-spot) where the accuracy will be the maximum and above which the model starts getting overfitting. This is the point of low variance and bias which we can say to be optimal sweet spot.

2.

Apart from accuracy matrix we can use AUC ROC: In this the data is split in train, validation and test data, the AUC scores are obtained for all validation sets in a cross validation procedure and mean validation Auc score is used for giving the validation performance metrics(of AUC score).Similarly test Auc score is calculated on held out test data.

3.

She Is doing wrong methodology, she is using test errors and looking on the performance of test errors and plotting it, she is tuning her hyperparameter by increasing the depth of tree as per the results obtained from test error.

This is a wrong approach and can lead to wrong and biased estimates. Test error should be used not for model selection and model training purposes but only for model evaluation once and for all.

STEPS SHE SHOULD FOLLOW:

She should take different values of D at a larger space like D {5,10,15,20,25,30,35….100}

Now cross validate on all depths and report the results and plot it.

She might not find exact 78 as the right depth but she can surely decide on 75 and get little high-test error as compared to most optimal depth size of 78.

But this model will be more robust and accurate and test data would not be touched in training time.