1. **The Cumulative Distribution Function $F(x)$ for some discrete random variable is given in the figure below. Please identify the corresponding Probability Mass Function $f(x)$.**

Solution 01 :-

Given,

$$F(x) = \begin{cases} 1/3 & ; -2 \leq x < 1 \\ 1/2 & ; 1 \leq x < 2 \\ 1 & ; x \geq 2 \end{cases}$$

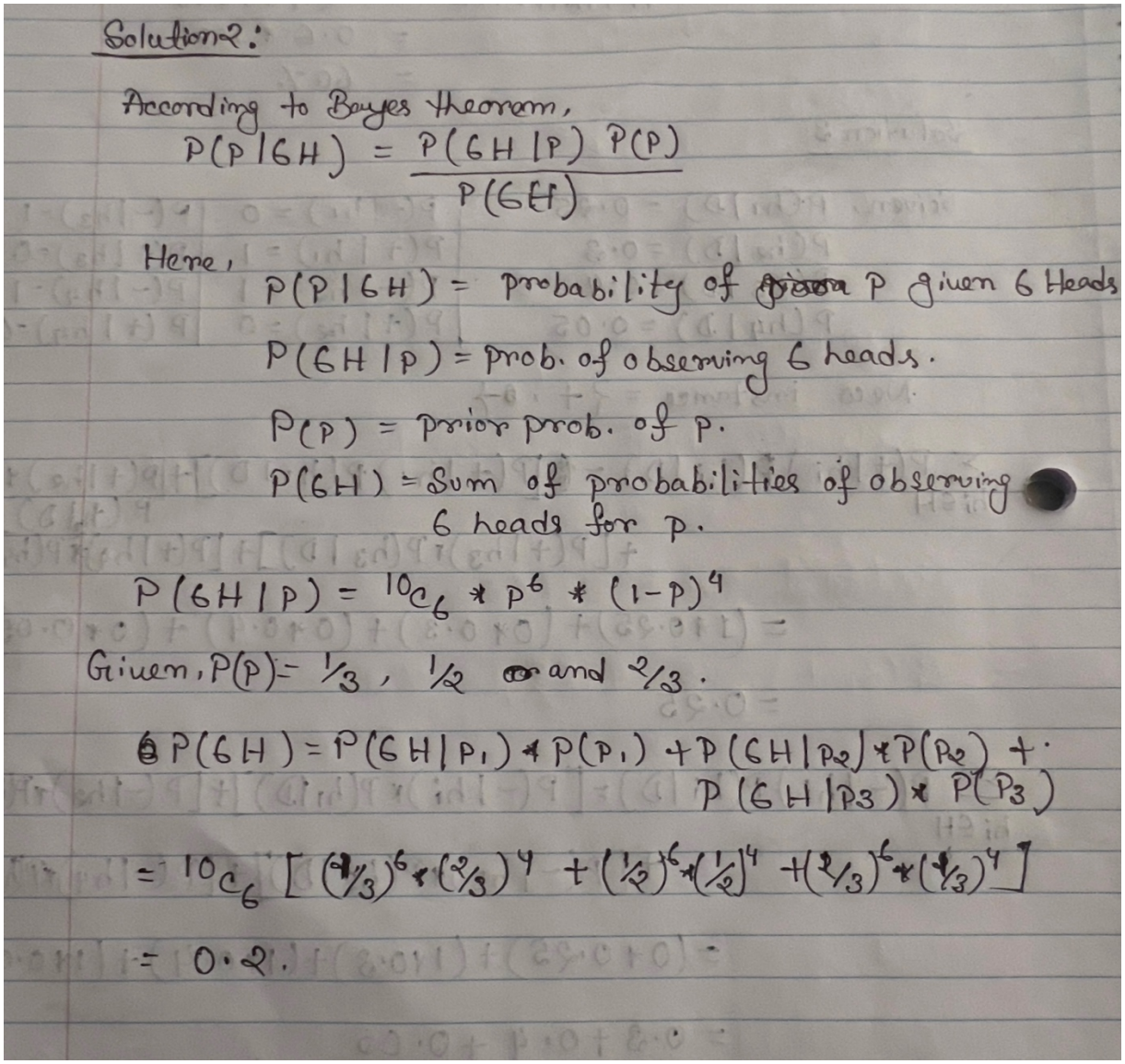
Since it is a step function, the probability mass function will be as follows :-

$$P_x(-2) = 1/3$$
$$P_x(1) = 1/2 - 1/3 = 1/6$$
$$P_x(2) = 1 - 1/2 = 1/2$$

So, Corresponding probability mass function,

$$f(x) = \begin{cases} 1/3 & ; -2 \leq x \leq 1 \\ 1/6 & ; 1 \leq x < 2 \\ 1/2 & ; x \geq 2 \end{cases}$$

2. **A coin we know nothing about has two sides {Head, Tail}, the coin is flipped 10 times, we observed 6 heads. The probability that the coin show heads is one of the following possibilities: 1/3, 1/2 or 2/3. Which is more probable? (Hint: compare *P*r ($H|p$)).**



Solution 2:

According to Bayes theorem,

$$P(P|6H) = \frac{P(6H|P)\,P(P)}{P(6H)}$$

Here,

$P(P|6H)$ = probability of P given 6 Heads

$P(6H|P)$ = prob. of observing 6 heads.

$P(P)$ = prior prob. of P.

$P(6H)$ = Sum of probabilities of observing 6 heads for P.

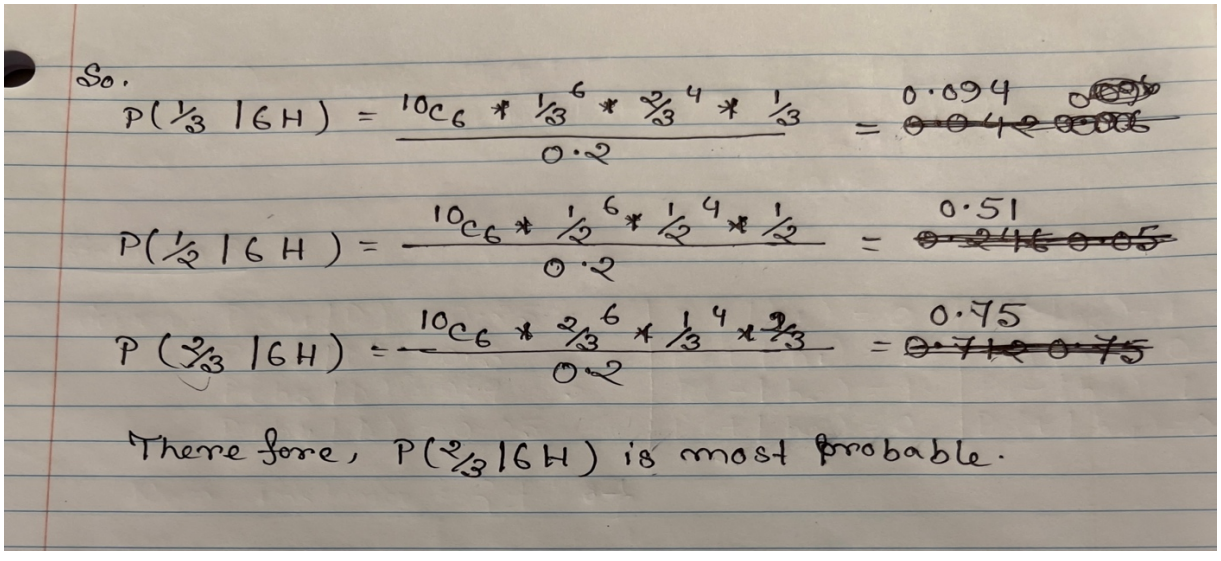
$$P(6H|P) = {}^{10}C_6 * p^6 * (1-p)^4$$

Given, $P(P) = 1/3, 1/2$ and $2/3$.

$$P(6H) = P(6H|P_1) * P(P_1) + P(6H|P_2) * P(P_2) + P(6H|P_3) * P(P_3)$$

$$= {}^{10}C_6 \left[(1/3)^6 * (2/3)^4 + (1/2)^6 * (1/2)^4 + (2/3)^6 * (1/3)^4 \right]$$

$$= 0.2$$



So.

$$P(1/3|6H) = \frac{{}^{10}C_6 * 1/3^6 * 2/3^4 * 1/3}{0.2} = 0.094$$

$$P(1/2|6H) = \frac{{}^{10}C_6 * 1/2^6 * 1/2^4 * 1/2}{0.2} = 0.51$$

$$P(2/3|6H) = \frac{{}^{10}C_6 * 2/3^6 * 1/3^4 * 2/3}{0.2} = 0.75$$

Therefore, $P(2/3|6H)$ is most probable.

3. **We have four hypotheses {*h*1,*h*2,*h*3,*h*4}. The posterior probabilities for the hypotheses are *P*(*h*1|*D*) =0.25, *P*(*h*2|*D*) =0.3, *P*(*h*3|*D*) =0.4, and *P*(*h*4|*D*)=0.05 respectively. The set of possible classification of the new instance is *V*= {+, −}. We also have: *P*(−|*h*1) =0, *P*(+|*h*1) =1, *P*(−|*h*2)=1, *P*(+|*h*2)=0, *P*(−|*h*3)=1, *P*(+|*h*3)=0, *P*(−|*h*4)=1, *P*(+|*h*4)=0. What is the result from the Bayes optimal classifier?**

Solution 3

Given, $P(h_1|D) = 0.25$, $P(h_2|D) = 0.3$, $P(h_3|D) = 0.4$, $P(h_4|D) = 0.05$

$P(-|h_1) = 0$, $P(+|h_1) = 1$, $P(-|h_2) = 1$, $P(+|h_2) = 0$

$P(-|h_3) = 1$, $P(+|h_3) = 0$, $P(-|h_4) = 1$, $P(+|h_4) = 0$

New instance = {+, −}

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D) = [P(+|h_1)P(h_1|D)] + [P(+|h_2) * P(h_2|D)] + [P(+|h_3) * P(h_3|D)] + [P(+|h_4) * P(h_4|D)]$$

$$= (1 * 0.25) + (0 * 0.3) + (0 * 0.4) + (0 * 0.05)$$

$$= 0.25$$

$$\sum_{h_i \in H} P(-|h_i)P(h_i|D) = [P(-|h_1) * P(h_1|D)] + [P(-|h_2) * P(h_2|D)] + [P(-|h_3) * P(h_3|D)] + [P(-|h_4) * P(h_4|D)]$$

$$= (0 * 0.25) + (1 * 0.3) + (1 * 0.4) + (1 * 0.05)$$

$$= 0.3 + 0.4 + 0.05$$

$$= 0.75$$

Hence, $\underset{v_j \in \{+,-\}}{\operatorname{argmax}} \sum_{h_j \in H} P(v_j|h_i)P(h_i|D) =$ −

4. **Using the Boston housing dataset (you can find the dataset on Blackboard). Implement linear regression on the training dataset, and report the intercept, slope, and R^2 values. Use these values to discuss the quality of the model.**

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

```
#load dataset and remove unwanted column
df = pd.read_csv('/home/ssultana/ML Class/.ipynb_checkpoints/housing.csv')
df.drop (['Unnamed: 0'], axis=1, inplace = True)
```

```
#Separate independent features and target
X = df.drop(['medv'], axis=1)
y = df['medv']
```

```
#Split the dataset into 80:20 ratio
x_train, x_test, y_train, y_test = train_test_split (X, y, test_size=0.2, random_state=0)
```

```
#Create model and fit it with training dataset
model = LinearRegression ()
model.fit (x_train, y_train)
```

```
LinearRegression()
```

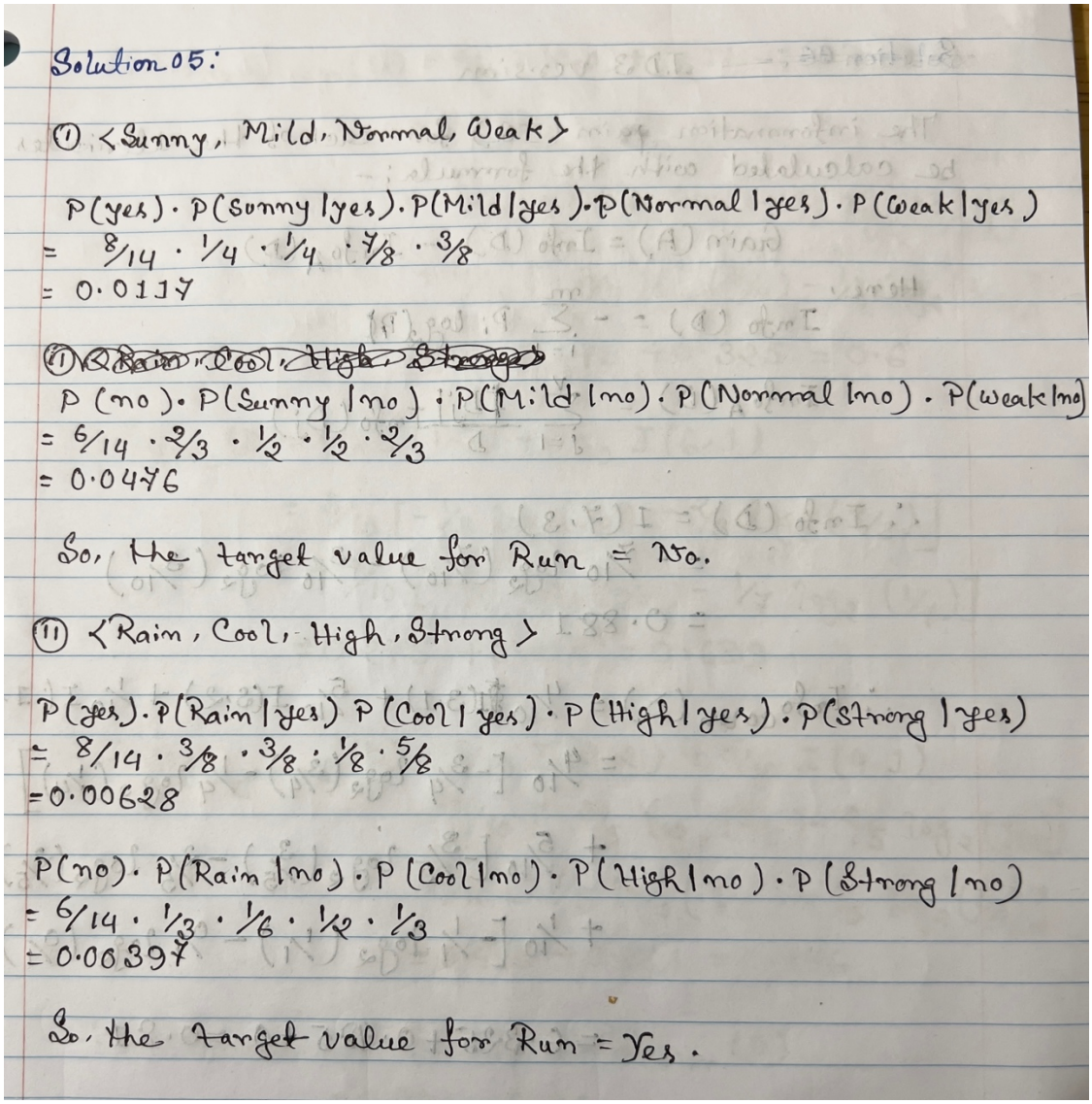
```
#Report the intercept, slope, and R^2 values of the model
print ("Slope = ", model.coef_)
print ("\nY_intercept = ", model.intercept_)
print ("\nR^2 Score = ", model.score(x_train, y_train))
```

```
Slope =  [-1.16530283e-01  8.05236516e-02 -7.37965959e-02  3.09472360e+00
 -6.79859508e+00  4.09752674e+00 -5.85764956e-03 -1.60277921e+00
  1.23420304e-01 -1.32770089e-02 -5.46758402e-01]

Y_intercept =  18.20299859220848

R^2 Score =  0.7359157332407895
```

5. **The hypothesis space is defined in the following table, each hypothesis is represented as a pair of 4-tuples. Please use the Naïve Bayes classifier to predict the target value *Run* for the instances: (15pts)**

   i) <Sunny, Mild, Normal, Weak>

   ii) <Rain, Cool, High, Strong>



Solution 05:

(i) <Sunny, Mild, Normal, Weak>

$P(yes) \cdot P(Sunny|yes) \cdot P(Mild|yes) \cdot P(Normal|yes) \cdot P(Weak|yes)$
$= 8/14 \cdot 1/4 \cdot 1/4 \cdot 7/8 \cdot 3/8$
$= 0.0117$

(ii) ~~<Rain, Cool, High, Strong>~~

$P(no) \cdot P(Sunny|no) \cdot P(Mild|no) \cdot P(Normal|no) \cdot P(Weak|no)$
$= 6/14 \cdot 2/3 \cdot 1/2 \cdot 1/2 \cdot 2/3$
$= 0.0476$

So, the target value for Run = No.

(ii) <Rain, Cool, High, Strong>

$P(yes) \cdot P(Rain|yes) \cdot P(Cool|yes) \cdot P(High|yes) \cdot P(Strong|yes)$
$= 8/14 \cdot 3/8 \cdot 3/8 \cdot 1/8 \cdot 5/8$
$= 0.00628$

$P(no) \cdot P(Rain|no) \cdot P(Cool|no) \cdot P(High|no) \cdot P(Strong|no)$
$= 6/14 \cdot 1/3 \cdot 1/6 \cdot 1/2 \cdot 1/3$
$= 0.00397$

So, the target value for Run = Yes.

6.

Consider the hypothesis space defined over instances shown below, we characterize each hypothesis (apple taste) by 4-tuples. Please hand trace the CART classifier to build a decision tree, then predict the target value Taste=Sweet/Tart for the following instances:

a) <Red, High, Some, No>
b) <Red, Low, Some, Yes>
c) <Yellow, Low, Some, No>
d) <Green, High, None, No>
e) <Green, Mid, Some, Yes>

Now suppose the actual taste of the five apples above are actually "Sweet, Sweet, Sweet, Tart, Tart", what is the accuracy of the decision tree? Please show all the steps and include the corresponding confusion matrix for accuracy calculation. (**20pts**)

Here,

$Gini(S) = 1 - [(7/10)^2 + (3/10)^2]$; we have, Sweet = 7, Tart = 3

$= 0.42$

1st iteration

~~$Gini_{Color}(Red) = 1 - ($~~

$Gini_{Red}(color) = 1 - [(3/4)^2 + (1/4)^2]$

$= 0.75$

$Gini_{yellow}(color) = 1 - [(3/5)^2 + (2/5)^2]$

$= 0.72$

$Gini_{Green}(color) = 1 - (1/1)^2$

$= 0$

$\therefore Gini(Color) = (0.75 * 4/10) + (0.72 * 5/10) + 0$

$= 0.3 + .36 = 0.66$

$Gini_{High}(Crispiness) = 1 - (3/3)^2$

$= 0$

$Gini_{Mid}(Crispiness) = 1 - [(2/3)^2 + (1/3)^2]$

$= 0.44$

$Gini_{Loco}(Crispiness) = 1 - [(3/4)^2 + (1/4)^2]$

$= 0.375$

$Gini(Crispiness) = 0 + (0.44 * 3/10) + (0.375 * 4/10)$

$= 0 + 0.132 + 0.15$

$= 0.282$

$Gini_{None}(Spot) = 1 - [(6/7)^2 + (1/7)^2]$

$= 0.24$

$Gini_{Some}(Spot) = 1 - [(2/3)^2 + (1/3)^2]$

$= 0.67$

$Gini(Spot) = (0.24 * 7/10) + (0.67 * 3/10)$

$= 0.372$

$Gini_{yes}(Fragment) = 1 - [(3/5)^2 + (2/5)^2]$

$= 0.72$

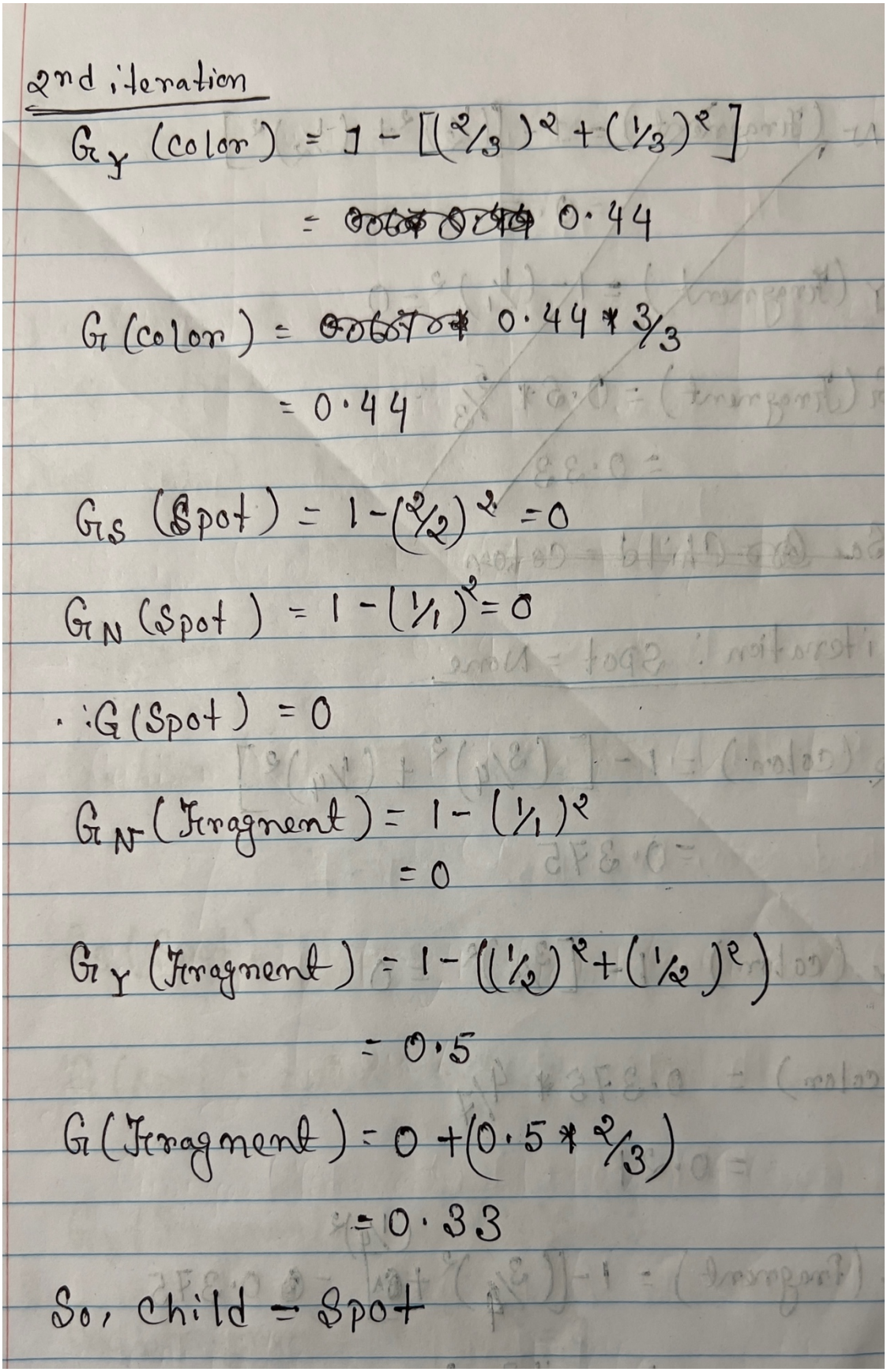
$Gini_{No}(Fragment) = 1 - [(4/5)^2 + (1/5)^2]$

$= 0.32$

$Gini(Fragment) = (0.72 * 5/10) + (0.32 * 5/10)$

$= 0.52$

So, Root = Gini (Crispiness)



2nd iteration

$G_Y(\text{color}) = 1 - [(2/3)^2 + (1/3)^2]$

$= 0.44$

$G(\text{color}) = 0.44 * 3/3$

$= 0.44$

$G_S(\text{Spot}) = 1 - (2/2)^2 = 0$

$G_N(\text{Spot}) = 1 - (1/1)^2 = 0$

$\therefore G(\text{Spot}) = 0$

$G_N(\text{Fragment}) = 1 - (1/1)^2$

$= 0$

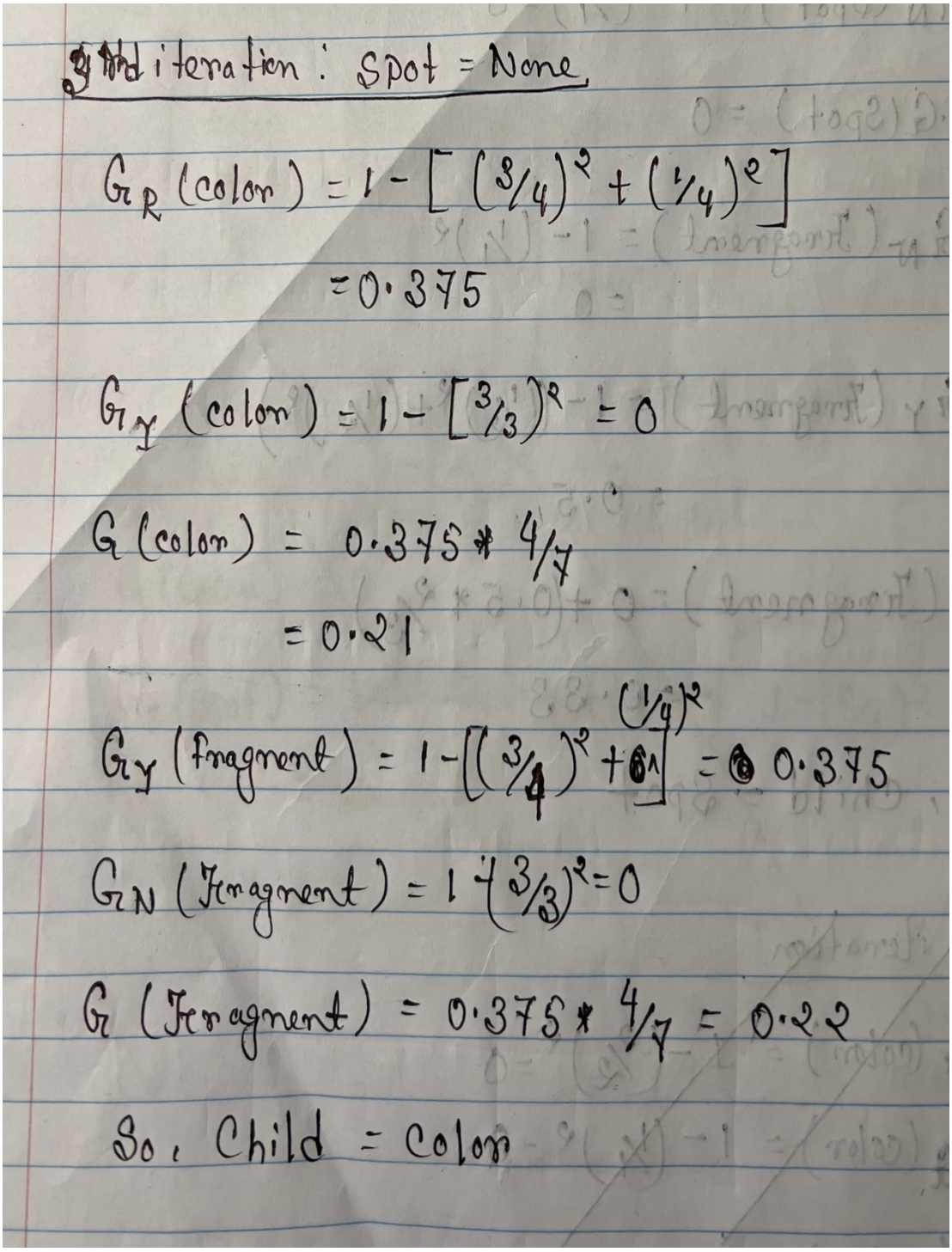
$G_Y(\text{Fragment}) = 1 - ((1/2)^2 + (1/2)^2)$

$= 0.5$

$G(\text{Fragment}) = 0 + (0.5 * 2/3)$

$= 0.33$

So, child = Spot



3rd iteration : Spot = None

$G_R(\text{color}) = 1 - [(3/4)^2 + (1/4)^2]$

$= 0.375$

$G_Y(\text{color}) = 1 - [3/3)^2 = 0$

$G(\text{color}) = 0.375 * 4/7$

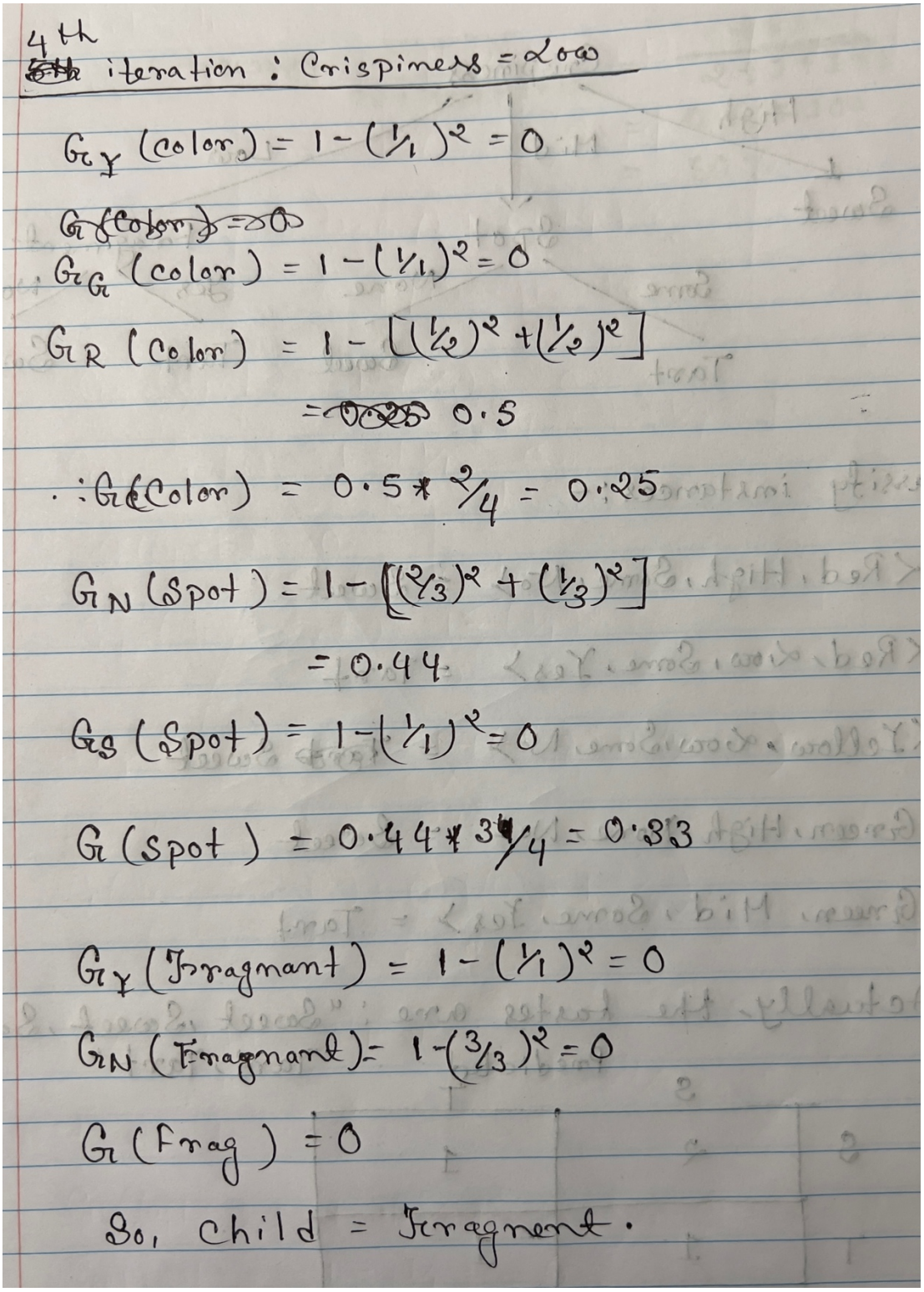
$= 0.21$

$G_Y(\text{fragrent}) = 1 - [(3/4)^2 + (1/4)^2] = 0.375$

$G_N(\text{Fragrent}) = 1 - (3/3)^2 = 0$

$G(\text{Fragrent}) = 0.375 * 4/7 = 0.22$

So, Child = Color



4th iteration : Crispiness = Low

$G_Y(\text{color}) = 1 - (1/1)^2 = 0$

$G_G(\text{color}) = 1 - (1/1)^2 = 0$

$G_R(\text{color}) = 1 - [(1/2)^2 + (1/2)^2]$

$= 0.5$

$\therefore G(\text{Color}) = 0.5 * 2/4 = 0.25$

$G_N(\text{Spot}) = 1 - [(2/3)^2 + (1/3)^2]$

$= 0.44$

$G_S(\text{Spot}) = 1 - (1/1)^2 = 0$

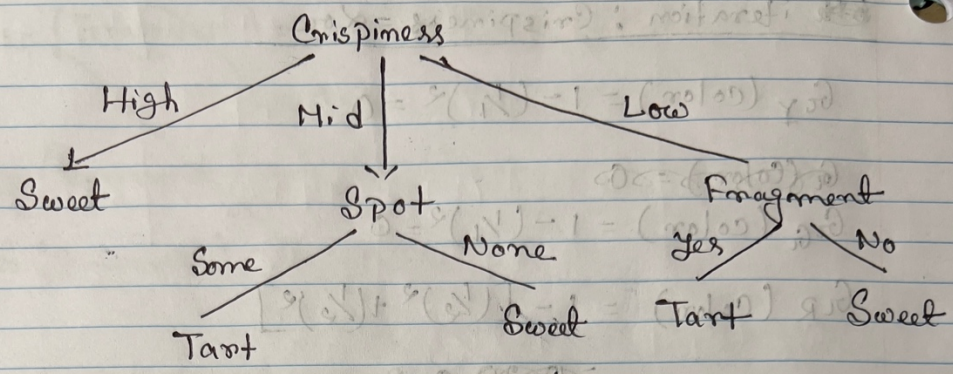
$G(\text{Spot}) = 0.44 * 3/4 = 0.33$

$G_Y(\text{Fragnant}) = 1 - (1/1)^2 = 0$

$G_N(\text{Fragnant}) = 1 - (3/3)^2 = 0$

$G(\text{Frag}) = 0$

So, Child = Fragnent.



Classify instances :-

(a) <Red, High, Some, No> = Sweet

(b) <Red, Low, Some, Yes> = Tart

(c) <Yellow, Low, Some, No> = ~~Tart~~ Sweet

(d) <Green, High, None, No> = Sweet

(e) <Green, Mid, Some, Yes> = Tart

Actually, the tastes are : "Sweet, Sweet, Sweet, Tart, Tart"

True :

		Predicted S	Predicted T
	S	2	1
	T	1	1

So TP = 2, TN = 1, FP = 1, FN = 1

$$\therefore \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{2+1}{2+1+1+1}$$

$$= 0.6 * 100$$

$$= 60\ \%.$$

7. **In a simple neural network shown below, the corresponding weights and biases are given. Please discuss in your own words the general function of a neural network, and the advantages/disadvantages of a deeper network structure compared to a shallow network structure.**

   An unknown bias value *b*3 is initialized as 0. Use backpropagation to update *b*3 twice. Include all the calculation steps. Use the learning rate=0.1.

Part 1:

A neural network imitates human brain functions and uses interconnected neurons to identify patterns in data. Layers of neurons perform specific computations, extracting abstract features from input data, and the output layer generates predictions.

One advantage of a deeper network structure is that it can learn more complex and abstract representations of the input data. Deeper networks can identify subtle patterns and dependencies that may not be captured by a shallow network. They are also more robust to noise and can generalize better to new, unseen data.

However, deeper networks are more computationally expensive to train, require more data to prevent overfitting, and are more prone to vanishing or exploding gradients, which can make training difficult. They may also be more prone to overfitting if not designed properly, and the complexity of the model may make it harder to interpret the learned representations.

Part 2:

Assume,

$p_1 = w_1.x + b_1$,

$p_2 = w_2.x + b_2$,

$y_1 = \ln(1+e^{p1})$,

$y_2 = \ln(1+e^{p2})$,

$y_3 = y_1.w_3 + y_2.w_4$,

$y_{predict} = y_3 + b_3$,

step size = d SSR / d b_3 = $-2(y_{observed} - y_{predict})$

learning rate = 0.1

First Propagation	**Second Propagation**	**Third Propagation**
At x=0,	**At x=0.5 (now, b_3= 0.501)**	**At x=1 (now, b_3= 0.915)**
p_1 = -1.4	p_1 = 0.25	p_1 = 1.9

$p_2 = 0.5$ $y_1 = 0.2204$ $y_2 = 0.974$ $y_3 = -2.504$ $y_{predict} = -2.504$ step size = -5.008 $b_3^* = b_3 - (\text{learning rate. step size}) = 0.501$	$p_2 = -1.25$ $y_1 = 0.826$ $y_2 = 0.252$ $y_3 = -1.57$ $y_{predict} = -1.069$ step size = -4.138 $b_3^* = b_3 - (\text{learning rate. step size}) = 0.915$	$p_2 = -3$ $y_1 = 2.039$ $y_2 = 0.048$ $y_3 = -2.55$ $y_{predict} = -1.635$ step size = -11.27 $b_3^* = b_3 - (\text{learning rate. step size}) = 2.042$