$$i(N) = -\sum_j P(w_j) \log_2 P(w_j)$$

A5

Date ___/___/___

Total no. of samples = 10

Let num_p = no. of positive reviews in the node

Let num_n = no. of negative reviews in node

Deciding split :

Initial : num_p = 5
num_n = 10

$$\therefore i(\text{root}) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$

If we split by smell,

left child (woody)

num_p = 2
num_n = 3

$$i(lc) = -\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) = 0.9709$$

right child (fruity)

num_p = 3
num_n = 2

$$i(rc) = -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) = 0.9709$$

$$\therefore \Delta i(N) \text{ on splitting based on smell} = 1 - \frac{5}{10} \times 0.9709 - \frac{5}{10} \times 0.9709$$

$$= 0.02904$$

Also,

$$\lim_{x \to 0} x \log(x) = \lim_{x \to 0} \frac{\log(x)}{1/x} = \lim_{x \to 0} \frac{1/x}{-1/x^2} = \lim_{x \to 0} -x = 0$$

[By L'Hopital's rule]

Page No.

Date __/__/____

105

If we split based on taste

left child (sweet)

num_p = 0

num_n = 3

$i(lc) = -1 \log_2(1) = 0 - 0 = 0$

right child (salty)

num_p = 2

num_n = 2

$i(rc) = -\left(\frac{1}{2}\left(\log_2 \frac{1}{2}\right) + \frac{1}{2} \times \log_2\left(\frac{1}{2}\right)\right) = 1$

middle child (sour)

num_p = 3

num_n = 0

$i(\text{middle}) = -1 \log_2(1) - 0 - 0 = 0$

$\therefore \Delta i(N) = i(\text{root}) - \frac{3}{10} \times i(\text{middle}) - \frac{3}{10} i(lc) - \frac{4}{10} i(rc)$

on splitting based on taste $= 1 - \frac{3}{10} \times 0 - \frac{3}{10} \times 0 - \frac{4}{10} \times 1$

$= 0.6$

Let a node N be split into L nodes. Then,

$$\Delta i(N) = i(N) - \sum_{t} P_t \, i(N_t)$$

where N_i = i^{th} node (child)

P_t = fraction of nodes at N that were sent to N_t

Page No.

Date ___/___/___

Saath

If we split based on portion

Left child (small)

num-p = 4

num-n = 1

$i(l_c) = -\left(\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{4}{5}\log_2\left(\frac{4}{5}\right)\right) = 0.7219$

right child (large)

num-p = 1

num-n = 4

$i(r_c) = -\left(\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{4}{5}\log_2\left(\frac{4}{5}\right)\right) = 0.7219$

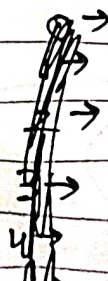
$\therefore \Delta i(N) = 1 - \frac{5}{10} \times 0.7219 - \frac{5}{10} \times 0.7219 =$ ~~0.27809~~

on splitting based on portion $= 0.27807$

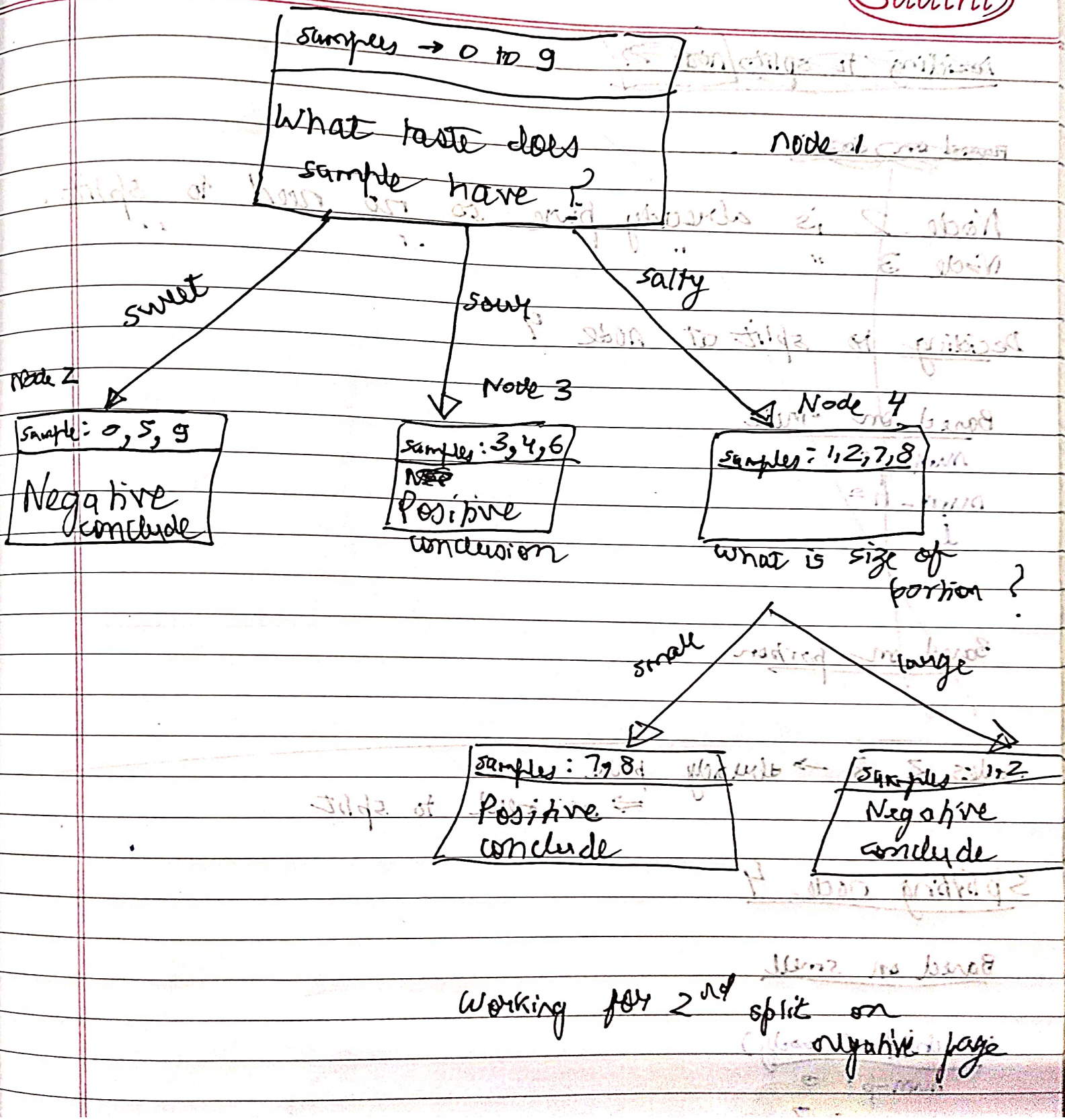
∴ We see that while exploring the information gain (impurity reduction) based on each of the 3 features, the maximum reduction of impurity happens when we split based on taste i.e, $\Delta = 0.6$.

So, assuming that the question asks to greedily choose the split, I decide to split based on taste.

~~Let samples be indexed as~~



Let samples have 0 based indexing based on sequence in question.



Working for 2nd split on negative page

nodes 2, 3 → already pure :
⇒ no need to split

Splitting node 4

Based on smell

left child (woody)

num_p = 0

num_n = 0

i(lc) = 0

right child (fruity)

num_p = 2

num_n = 2

$i(rc) = -\left(\frac{2}{4}\log_2\left(\frac{1}{2}\right) + \frac{2}{4}\log_2\left(\frac{1}{2}\right)\right) = 1$

$\Delta i = 1 - 0 - \frac{4}{4} \times 1 = 0$

Page No.

Based on portion

small (left child)

num_p = 2

num_n = 0

$i(lc) = -(0 + \frac{2}{2}\log_2(\frac{2}{2})) = 0$

right child (large)

num_p = 0

num_n = 2

$i(rc) = -(\frac{2}{2}\cdot\log_2(\frac{2}{2}) + 0) = 0$

$\Delta i = 1 - 0 - 0 = 1$

∴ Δi is greatest on splitting based on portion and hence, this is the feature we choose for split on node 4.