



University of Asia Pacific

Department of CSE

Name: Rashik Rahman

Reg ID: 17201012

Year: 4th

Semester: 1st

Course Code: CSE 427

Course Title: Machine learning

Date: 17.02.2021

"During Examination and upload time I will not take any help from anyone. I will give my exam all by myself."

Answer to the Q.No.1

Formula of gain : $-P_1 \log_2 P_1 - P_2 \log_2 P_2$

$$\therefore \text{Gain over total dataset} = -\frac{4}{7} \log_2 \frac{4}{7} - \frac{3}{7} \log_2 \frac{3}{7}$$

Gain:

$$= -\frac{4}{7} (-0.81) - \frac{3}{7} (-1.22)$$

$$= 0.463 + 0.523 = 0.986$$

For Turmon size:

Feature Attribute	positive	negative	gain
Small	1	2	$-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3}$ $= -\frac{1}{3} (-1.59) - \frac{2}{3} (-0.59)$ $= 0.53 + 0.39 = 0.923$
Medium	1	1	0.923 $-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}$ $= -\frac{1}{2} (-1) - \frac{1}{2} (-1)$ $= \frac{1}{2} + \frac{1}{2} = 1$
Large	2	0	$-\frac{2}{2} \log_2 \frac{2}{2} - 0 \log_2 0$ $= 0$

17201012
②

∴ Information gain of Tumor size = gain (total dataset)

$$= \left\{ \frac{3}{7} \text{gain}(\text{small}) + \frac{2}{7} \text{gain}(\text{medium}) + \frac{2}{7} \text{gain}(\text{large}) \right\}$$

$$= 0.986 - \left(\frac{3}{7} \times 0.923 + \frac{2}{7} \times 1 + \frac{2}{7} \times 0 \right)$$

$$= 0.986 - (0.396 + 0.286) = 0.304$$

For stage:

Attribute	positive	negative	gain ()
A	0	1	$-\frac{0}{1} \log_2 0 - \frac{1}{1} \log_2 1 = 0$
B	3	1	$-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.31 + 0.5$ $= 0.81$
C	1	1	$-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$

$$\text{Info. gain of stage} = 0.986 - \left\{ \frac{1}{7} \text{gain}(A) + \frac{4}{7} \text{gain}(B) + \frac{2}{7} \text{gain}(C) \right\}$$

$$= 0.986 - \left(\frac{1}{7} \times 0 + \frac{4}{7} \times 0.81 + \frac{2}{7} \times 1 \right)$$

$$= 0.986 - (0.46 + 0.29)$$

$$= 0.240$$

for smoken:

Attribute	positive	negative	gain()
Yes	4	0	0
No	0	3	0

98% label count
0 so gain 0

Info.

$$\therefore \text{Gain for smoke} = 0.986 - \left(\frac{4}{7} \text{gain(Yes)} + \frac{3}{7} \text{gain(No)} \right) \\ = 0.986$$

Info

$$\therefore \text{Gain(tumor size)} = 0.309$$

$$\text{Info gain(stage)} = 0.240$$

$$\text{Info gain(smoke)} = 0.986$$

17201012

(4)

Answer to the Q. No. 2

1	1	A (3,2)
1	2	
2	2	
3	1	B (5,5)
4	4	
4	5	
5	6	
6	7	

Centroid positions of both cluster in next iteration are following.

$$\text{For cluster center } A = \left(\frac{1+1+2+3}{4}, \frac{1+2+2+1}{4} \right) \\ = \left(\frac{7}{4}, \frac{6}{4} \right) = (1.75, 1.5)$$

$$\text{For cluster center } B = \left(\frac{4+4+5+6}{4}, \frac{4+5+6+7}{4} \right) \\ = (4.75, 5.5)$$

17201012
⑤

So the new positions of both centroids
are, $A = (1.75, 1.5)$, ~~B~~ $B = (4.75, 5.5)$

A