Data Mining (CS4038)

Quiz No.4

Solution

Roll No: _____ **Section:** B **Date:01-04-2024**

Question No.1 **(10 marks)**

Consider the following dataset of Passed and Failed exams with three attributes: studied(Y/N), slept(Y/N) and cheated(Y/N) and answer the questions given below. Compute the Gini Index for all of these attributes and identify which is best attribute to serve as root node.

Student ID	Studied	Slept	Cheated	Result
1	Yes	No	No	Passed
2	Yes	No	Yes	Failed
3	No	Yes	No	Failed
4	Yes	Yes	Yes	Failed
5	Yes	Yes	No	Passed
6	No	No	Yes	Failed
7	No	No	No	Failed
8	No	Yes	Yes	Failed

$$Gini(t) = 1 - \sum_{i=0}^{c-1} P_i(t)^2$$

Result

Passed 2

Filed 6

8

$$Gini(Result) = 1 - \left(\frac{2}{8}\right)^2 - \left(\frac{6}{8}\right)^2$$

$$= \boxed{0.375}$$

Studied

	Yes	No
Passed	2	0
Failed	2	4
	4	4

~~Gini(Studied|Yes)~~ +3

$$Gini(Yes/studied) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$$

$$= 0.5$$

$$Gini(No/studied) = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2$$

$$= 0$$

$$Gini(Studied) = \frac{4}{8}(0.5) + \frac{4}{8}(0)$$

$$= \boxed{0.25}$$

Gini(Yes/slept) $= 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2$ +3

$= 0.375$

Gini(No/slept) $= 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2$

$= 0.375$

Gini(slept) $= \frac{4}{8}(0.375) + \frac{4}{8}(0.375)$

$= 0.1875 + 0.1875 = \boxed{0.375}$

slept	yes	No
passed	1	1
Failed	$\frac{3}{4}$	$\frac{3}{4}$

Gini(Yes/cheated) $= 1 - (\frac{0}{4})^2 - (\frac{4}{4})^2$ +3

$= 0$

Gini(No/cheated) $= 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2$

$= 0.5$

Gini(cheated) $= \frac{4}{8}(0) + \frac{4}{8}(0.5)$

$= \boxed{0.25}$

cheated	yes	No
passed	0	2
Failed	$\frac{4}{4}$	$\frac{2}{4}$

+1

As the Gini Index value of studied and cheated attributes is the lowest and same So, ~~both~~ anyone of these 2 attributes can serve as Root node in Decision tree.

Data Mining (CS4038)

Quiz No.4

Roll No: _____ **Section:** A **Date:01-04-2024**

Question No.1 **(10 marks)**

Consider the following dataset of Passed and Failed exams with three attributes: studied(Y/N), slept(Y/N) and cheated(Y/N) and answer the questions given below. Compute gain ratio of all these attributes and identify which is the best attribute to serve as root node.

Student ID	Studied	Slept	Cheated	Result
1	Yes	No	No	Passed
2	Yes	No	Yes	Failed
3	No	Yes	No	Failed
4	Yes	Yes	Yes	Failed
5	Yes	Yes	No	Passed
6	No	No	Yes	Failed
7	No	No	No	Failed
8	No	Yes	Yes	Failed

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} P_i(t) \log_2 P_i(t)$$

Result

Passed 2

Failed 6

8

$$\text{Entropy}(\text{Result}) = -\frac{2}{8}\log_2\left(\frac{2}{8}\right) - \frac{6}{8}\log_2\left(\frac{6}{8}\right)$$

$$= 0.5 + 0.3112$$

$$= \boxed{0.8112}$$

+3

Studied	yes	no
passed	2	0
Failed	2	4
	4	4

$$\text{Entropy}(\text{yes}|\text{studied}) = -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= 1$$

$$\text{Entropy}(\text{no}|\text{studied}) = -\frac{0}{4}\log\left(\frac{0}{4}\right) - \frac{4}{4}\log\left(\frac{4}{4}\right)$$

$$= 0$$

$$\text{Entropy}(\text{studied}) = \frac{4}{8}(1) + \frac{4}{8}(0)$$

$$= 0.5$$

$$\text{Gain}_{\text{info}} = \text{Ent}(\text{Result}) - \text{Ent}(\text{studied})$$

$$= 0.8112 - 0.5 = 0.3112$$

$$\text{splitinfo} = -\sum_{i=1}^{k} \frac{n_i}{n}\log_2\frac{n_i}{n} = -\frac{4}{8}\log_2\frac{4}{8} + \frac{4}{8}\log_2\frac{4}{8} = 1$$

$$\text{Gain}_{\text{Ratio}} = \frac{\text{Gain info}}{\text{Split info}} = \frac{0.3112}{1} = \boxed{0.3112}$$

$Ent(Yes|slept) = -\frac{1}{4}\log_2(\frac{1}{4}) - \frac{3}{4}\log_2(\frac{3}{4})$ +3

$= 0.8112$

$Ent(no|slept) = -\frac{1}{4}\log_2(\frac{1}{4}) - \frac{3}{4}\log_2(\frac{3}{4})$

$= 0.8112$

slept

	yes	no
passed	1	1
Failed	3/4	3/4

$Ent(slept) = \frac{4}{8}(0.8112) + \frac{4}{8}(0.8112)$

$= 0.8112$

$Gain_{info} = 0.8112 - 0.8112$

$= 0$

$Split_{info} = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1$

$Gain_{Ratio} = \frac{0}{1} = \boxed{0}$

+3

cheated

	yes	no
passed	0	2
Failed	4/4	2/4

$Ent(Yes|cheated) = -\frac{0}{4}\log_2(\frac{0}{4}) - \frac{4}{4}\log_2(\frac{4}{4})$

$= 0$

$Ent(No|cheated) = -\frac{2}{4}\log_2(\frac{2}{4}) - (\frac{2}{4})\log_2(\frac{2}{4})$

$= 0.5 + 0.5 = 1$

$Ent(cheated) = \frac{4}{8}(0) + \frac{4}{8}(1) = 0.5$

$Gain_{info} = 0.8112 - 0.5 = 0.3112$

$Split_{info} = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1$

$Gain_{Ratio} = \frac{0.3112}{1} = \boxed{0.3112}$

+1

So, As the Gain Ratio of Studied and cheated attributes is highest and same so anyone of these attributes can serve as root node for Decision tree.