Assignment On

Data Mining (MCSE 642)

Assignment – 01

Submitted to

Dr. Md Tauhid Bin Iqbal

Submitted by:

Mohammed Morad Hossen

Student ID: MCE 079 055 32

Dept of Computer Science and Engineering

Stamford University Bangladesh

Name: Mohammed Monica Hossen
ID: MCE 079 055 32

Income

Income in a nominal feature. It can be High, mediam, Low. I will summarize the final decision for Income feature

Income	Yes	NO	Number of Instances
High	2	2	4
Medium	4	2	6
Low	3	1	4

$$Gini(Income = High) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 1 - 0.25 - 0.25$$
$$= 0.5$$

$$Gini(Income = Medium) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = \text{~~+ 0.11 -~~}$$
$$= 1 - 0.44 - 0.11 = 0.45$$

$$Gini(Income = Low) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 1 - 0.563 - 0.063$$
$$= 0.374$$

Then we will calculate weighted sum of Gini index for Income feature

$$= \frac{4}{14}(0.5) + \frac{6}{14}(0.45) + \left(\frac{4}{14}\right)(0.374)$$
$$= 0.143 + 0.193 + 0.107$$
$$= 0.443$$

STUDENT

Student in a nominal feature. It can be Yes, NO. I will summarize the final decision for student feature

Student	Yes	NO	Number of Instances
Yes	6	1	7
NO	3	4	7

Gini(student = yes) = $1-(6/7)^2-(1/7)^2 = 1-0.73-0.02$

$= 0.25$

Gini(student = no) = $1-(3/7)^2-(4/7)^2 = 1-0.18-0.33$

$= 0.49$

Then we will calculate weighted sum of Gini Index for ~~Income~~ student feature

$= \cancel{(7/7)(0.25)+(7/7)}$

$= (7/14)(0.25) + (7/14)\,0.49$

$= 0.125 + 0.245 = 0.37$

AGE

Age is a nominal feature. It can be <=30, 31..40, >40
I will summarize the final decision for Age feature.

Age	Yes	No	Number of Instances
<=30	2	3	5
31..40	4	0	4
>40	3	2	5

Gini(Age = <=30) = $1-(2/5)^2-(3/5)^2 = 1-0.16-0.36$

$= 0.48$

Gini(Age = 31..40) = $1-(4/4)^2-(0/4)^2 = 1-1-0 = 0$

Gini(Age > 40) = $1-(3/5)^2-(2/5)^2 = 1-0.36-0.16$

$= 0.48$

~~Credit-Rating : Credit Rating is a nominal feature. It can be~~

Then we will calculate sum of GINI Index for Age feature = $(5/14)\times 0.48 + (4/14)\times 0 + (5/14)\,0.48$

$= 0.171 + 0.171 = 0.34$

Credit_Rating

Credit Rating is a nominal feature. It can be Fair, Excellent. I will summarize the final decision for Credit Rating.

Credit_Rating	Yes	NO	Number of Instances.
Fair	6	2	8
Excellent	3	3	6

$$Gini(C.R = Fair) = 1 - (6/8)^2 - (2/8)^2 = 1 - 0.56 - 0.063$$
$$= 0.38$$

$$Gini(C.R\ Excellent) = 1 - (3/6)^2 - (3/6)^2 = 1 - 0.25 - 0.25$$
$$= 0.50$$

Then we will calculate weighted sum of GINI index for Credit Rating

$$= (8/14) \times 0.38 + (6/14)(0.5)$$
$$= 0.22 + 0.21 = 0.43$$

Feature	GINI Index
Age	0.34
Income	0.443
Student	0.37
Credit_Rating	0.43

Here Age is the winner feature because its cost is lowest (0.34). The Age attribute will be considered as the root node.

Calculating Entropy

$$E(S) = \sum_{i=1}^{c} - P_i(\log_2 P)$$

Buy Computer	
Yes	NO
9	5

Entropy (Buy_Computer)

$$= -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\left(\log_2 5/14\right)$$

$$= -0.64 \times (-0.64) - 0.36(-1.47)$$

$$= 0.41 + 0.53 = 0.94$$

Income	Buyes Buys Computer		Numbe of Instances
	Yes	NO	
High	2	2	4
Medium	4	2	6
Low	3	1	4

Entropy (Income, High) $= -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log\left(\frac{2}{4}\right)$

$$= -0.5(-1) - (0.5)(-1)$$

$$= 0.5 + 0.5 = 1$$

Entropy (Income, Medium) $= -4/6\log_2(4/6) - \frac{2}{6}\log_2(2/6)$

$$= 0.67(-0.58) - 0.33(-1.6)$$

$$= 0.39 + 0.53$$

$$= 0.92$$

Entropy (Income, Low) $= -\frac{3}{4}\log(3/4) - 1/4\log_2(1/4)$

$$= 0.75(-0.42) - 0.25(-2)$$

$$= 0.32 + 0.5 = 0.82$$

Entropy (Buy Computer, Income)

$= P(\text{High})E(2,2) + P(\text{medium})E(4,2) + P(\text{low})E(3,1)$

$= \frac{4}{14} \times 1 + \frac{6}{14} \times 0.92 + 4/14 \times 0.82$

$= 0.29 + 0.4 + 0.23 = 0.92$

Information Gain (Buy Computer, Income)

$= E(\text{Buy Computer}) - E(\text{Buy Computer, Income})$

$= 0.94 - 0.92$

$= 0.02$

Student

Student	Buy Computer		Number of Instances
	Yes	NO	
Yes	6	1	7
NO	3	4	7

Entropy (student, yes) $= -\frac{6}{7}\log_2(6/7) - \frac{1}{7}(\log_2 1/7)$

$= -0.86(-0.22) - 0.14(-2.84)$

$= 0.19 + 0.4 = 0.59$

Entropy (student, NO) $= -\frac{3}{7}\log_2(3/7) - \frac{4}{7}\log_2(4/7)$

$= -0.43(-1.22) - 0.57(-0.81)$

$= 0.52 + 0.46 = 0.98$

Entropy (student, Buy computer) $= P(\text{yes})E(6,1) + P(\text{NO})E(3,4)$

$= 7/14 \times 0.59 + 7/14\ (0.98)$

$= 0.3 + 0.49 = 0.79$

Information Gain (Buy Computer, Student)

= E (Buy Computer) - E (Buy Computer, Student)

= 0.94 − 0.79 = 0.15

AGE

Age	Buy Computer		Number of Instances
	yes	NO	
<=30	2	3	5
31..40	4	0	4
>40	3	2	5

$$\text{Entropy (Age, <=30)} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= -0.4(-1.32) - 0.6(-0.74)$$

$$= 0.53 + 0.44 = 0.97$$

$$\text{Entropy (Age, 31..40)} = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log\left(\frac{0}{4}\right)$$

$$= -1(0) - 0 = 0$$

$$\text{Entropy (Age, >40)} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$$

$$= -0.6(-0.74) - 0.4(-1.32)$$

$$= 0.44 + 0.53 = 0.97$$

$$\text{Entropy (Age, Buy computer)} = P(<=30)\,E(2,3) + P(31..40)\,E(4,0) + P(>40)\,E(3,2)$$

$$= \frac{5}{14}(0.97) + \frac{4}{14}(0) + \frac{5}{14}(0.97)$$

$$= 0.35 + 0.35 = 0.7$$

Information Gain (Buy computer, Age)

= E(Buy Computer) - E(Buy Computer, Age)

= 0.94 − 0.7 = 0.24

Credit_Rating

Credit-Rating	Buy Computer		Number of Instances
	Yes	NO	
fair	6	2	8
Excellent	3	3	6

$$\text{Entropy}(C.R, \text{fair}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$$
$$= -0.75(-0.42) - 0.25(-2)$$
$$= 0.33 + 0.5 = 0.83$$

$$\text{Entropy}(C.R, \text{Excellent}) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$
$$= -0.5(-1) - (0.5)(-1)$$
$$= 0.5 + 0.5 = 1$$

$$\text{Entropy}(C.R, \text{Buy computer}) = \frac{8}{14}(0.83) + \frac{6}{14}(1)$$
$$= 0.47 + 0.43 = 0.9$$

$$\text{Infonmain Grain}(\text{Buy computer}, \text{Credit-Rating}) = E(\text{Buy computer}) - E(\text{Buy computer}, C.R)$$
$$= 0.94 - 0.9 = 0.04$$

feature	Information Gain
Age	0.24
Income	0.02
Student	0.15
Credit-Rating	0.04

Age feature produce highest score. So Age will be considered on the root node.

Gain Ration (Decision, Age)

$$= \frac{\text{I.Gain (Decision, Age)}}{\text{SplitInfo (Decision, Age)}}$$

$$= \frac{0.24}{0.7} = 0.34$$

Gain Ration (Decision, Income)

$$= \frac{\text{I.G(Decision, Income)}}{\text{SplitInfo (Decision, Income)}}$$

$$= \frac{0.02}{0.92} = 0.02$$

Gain Ratio (Decision, Student)

$$= \frac{\text{I.G (Decision, Student)}}{\text{SplitInfo (Decision, Student)}}$$

$$= \frac{0.15}{0.79} = 0.19$$

Gain Ratio = (Decision, Credit_Rating)

$$= \frac{I.G\ (Decision,\ C.R)}{SplitInfo\ (Decision,\ CR)}$$

$$= \frac{0.04}{0.9} = 0.04$$

Feature	Gain Ratio
Age	0.34
Income	0.02
Student	0.19
Credit Rating	0.04

Here Age feature has highest gain Ratio and has been selected as root node.