AMA546 Statistical Data Mining

Q2 Part2

1. The dataset is shown in the Table 1. Please write down the formula to calculate the information gain when splitting the data by attributes A and B (no need to calculate the final result), and explain the role of calculating information gain in classification algorithms.

A	B	Class Label
T	F	*
T	T	*
T	T	*
T	F	#
T	T	*
F	F	#
F	F	#
F	F	#
T	T	#
T	T	#

Table 1

Solution:

Based on the dataset, we can count the frequency of class labels in different attributes.

		Class Label		
		*	#	Total
A	T	4	3	7
	F	0	3	3
	Total	4	6	10

		Class Label		
		*	#	Total
B	T	3	2	5
	F	1	4	5
	Total	4	6	10

Then the information gain can be calculated based on the tables above.

$$\text{Information Gain}(A) = H(\frac{4}{10}, \frac{6}{10}) - \frac{7}{10}H(\frac{4}{7}, \frac{3}{7}) - \frac{3}{10}H(1,0)$$
$$= \left(-\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10}\right) - \frac{7}{10}\left(-\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}\right) - 0$$

$$\text{Information Gain}(B) = H(\frac{4}{10}, \frac{6}{10}) - \frac{4}{10}H(\frac{3}{4}, \frac{1}{4}) - \frac{6}{10}H(\frac{4}{6}, \frac{2}{6})$$
$$= \left(-\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10}\right) - \frac{4}{10}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) - \frac{6}{10}\left(-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right)$$

Information gain is a commonly used metric in classification algorithms for **selecting the most informative features** that contribute to the classification of data points. It measures the **reduction in entropy or uncertainty of a target variable after splitting the data based on a given attribute**. It helps to identify the most relevant features, which in turn can lead to more accurate and efficient models.

2. The Table 2 shows a set of labeled tuples randomly selected from a customer database. In this example, each attribute is discrete-valued. The class label attribute, buys computer, has two different values (i.e., yes, no), so there are two different classes (i.e., m = 2). Let the C1 class correspond to yes and the C2 class correspond to no.

RID	Age	Income	Student	Credit_rating	Class
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	yes

Table 2

(a) Calculate the information gain of attribute 'income'.

(b) Calculate the information gain ratio of attribute 'income'.

(c) Calculate the Gini index of attribute 'income'.

Solution:

(a) Note that the formula for calculating information gain is:

$$\text{Information Gain}(S, A) = \text{H}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \cdot \text{H}(S_v)$$

where $\text{H}(S)$ is the entropy of the set S and $\text{H}(S_v)$ is the entropy of the subset S_v created by splitting S on the attribute A with value v. $|\cdot|$ refers to the size of the set. The formula for entropy in the context of Information Gain is:

$$H(X) = -p_i \log_2 p_i$$

where p_i is the probability of the i-th outcome.

The table below gives the income distribution by class:

		Income			
		High	Medium	Low	Total
class	yes	2	5	3	10
	no	2	1	1	4
	Total	4	6	4	14

Table 3: Class count of attribute Imcome

To calculate the information gain of attribute "income", we need to first calculate the entropy of the target attribute:

$$\begin{aligned} H(Class) &= -\frac{10}{14} \log_2 \frac{10}{14} - \frac{4}{14} \log_2 \frac{4}{14} \\ &= 0.8631 \end{aligned}$$

Next, we need to calculate the entropy of the "income" attribute:

$$H(Income = High) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1$$

$$H(Income = Medium) = -\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6} = 0.6500$$

$$H(Income = Low) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$$

Finally, we can calculate the information gain of "income" as:

$$\text{Information Gain}(Income) = H(Class) - \frac{4}{14}H(Income = High) - \frac{6}{14}H(Income = Medium) - \frac{4}{14}H(Income = Low)$$
$$= 0.0670$$

(b) **Information Gain Ratio** is calculated as the **ratio of the information gain of an attribute to its intrinsic information**. It helps to overcome the bias for the attribute with many outcomes by normalizing the information gain using the intrinsic information of the attribute. It is calculated as follows:

$$\text{Information Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Information}}$$

where **Information Gain** is the difference between the entropy of the parent node and the weighted average of the entropy of the child nodes after the split, and **Split Information** is a measure of the amount of uncertainty in the split.

The formula for Split Information is:

$$\text{Split Information} = -\sum_{i=1}^{n}\frac{N_i}{N}\log_2\frac{N_i}{N}$$

where N_i is the number of examples that belong to the i-th child node, and N is the total number of examples. A detailed introduction can be found here.

To calculate the information gain ratio of attribute "income", we first need to calculate the split information:

$$\text{Split Information}(Income) = -\frac{4}{14}\log_2\frac{4}{14} - \frac{6}{14}\log_2\frac{6}{14} - \frac{4}{14}\log_2\frac{4}{14}$$
$$= 1.557$$

Next, we can calculate the information gain ratio of "income" as:

$$\text{Information Gain Ratio}(Income) = \frac{\text{Information Gain}(Income)}{\text{Split Information}(Income)} = \frac{0.0670}{1.557} = 0.0431$$

(c) We need to calculate the Gini index of the "income" attribute by weighting the Gini index of each possible value:

$$\text{Gini}(Income) = \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) + \frac{6}{14}\left(1 - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right)$$
$$= 0.3690$$

3. Given the data in Table 4, we hope to use the Naive Bayes classifier to predict the class label of an unknown principle. The data tuple is described by the attributes "age", "income", "student", and "credit rating", and the class label attribute "buys computer" has two different values. Please use the Naive Bayes method to classify X. Use the Laplace Estimation to avoid zero conditional probabilities.

$$X = (Age = youth, Income = medium, Student = yes, Credit\_rating = Fair)$$

Solution:

Here we use *Laplace smoothing*. We add 1 in the numerator and number of classes in the denominator to avoid zero probabilities.

The prior probability of each class:

$$P(class = yes) = \frac{10+1}{14+2} = \frac{11}{16}$$
$$P(class = no) = \frac{4+1}{14+2} = \frac{5}{16}$$

The conditional probability of each attribute given class label:

RID	Age	Income	Student	Credit_rating	Class
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	yes

Table 4

		Total	Age = youth	Income = medium	Student = yes	Credit rating = Fair
Class	yes	10	2	5	6	6
	no	4	3	1	1	2
		P(C)	P(Age = youth|C)	P(Income = medium|C)	P(Student = yes|C)	P(Credit rating = Fair|C)
Class	yes	11/16	3/13	6/13	7/12	7/12
	no	5/16	4/7	2/7	2/6	3/6

$$P(age = youth \mid class = yes) = \frac{2+1}{10+3} = \frac{3}{13}$$

$$P(age = youth \mid class = no) = \frac{3+1}{4+3} = \frac{4}{7}$$

$$P(income = medium \mid class = yes) = \frac{5+1}{10+3} = \frac{6}{13}$$

$$P(income = medium \mid class = no) = \frac{1+1}{4+3} = \frac{2}{7}$$

$$P(student = yes \mid class = yes) = \frac{6+1}{10+2} = \frac{7}{12}$$

$$P(student = yes \mid class = no) = \frac{1+1}{4+2} = \frac{2}{6}$$

$$P(credit = fair \mid class = yes) = \frac{6+1}{10+2} = \frac{7}{12}$$

$$P(credit = fair \mid class = no) = \frac{2+1}{4+2} = \frac{3}{6}$$

Then

$P(class = yes \mid Age = youth, Income = medium, Student = yes, Credit\_rating = Fair)$
$\propto P(Age = youth, Income = medium, Student = yes|, Credit\_rating = Fair \mid class = yes)$
$= P(age = youth \mid class = yes)P(income = medium \mid class = yes)P(student = yes \mid class = yes)P(credit = fair \mid class = yes)P(class = yes)$
$= \frac{3}{13}\frac{6}{13}\frac{7}{12}\frac{7}{12}\frac{11}{16}$
$= 0.0249$

$P(class = no \mid Age = youth, Income = medium, Student = yes, Credit\_rating = Fair)$
$\propto P(Age = youth, Income = medium, Student = yes|, Credit\_rating = Fair \mid class = no)$
$= P(age = youth \mid class = no)P(income = medium \mid class = no)P(student = no \mid class = no)P(credit = fair \mid class = no)P(class = no)$
$= \frac{4}{7}\frac{2}{7}\frac{2}{6}\frac{3}{6}\frac{5}{16}$
$= 0.0085$

Therefore, the predicted label of X is **yes**.

4. The tuples in the Table 5 have been sorted in descending order of the probability values returned by the classifier. For each tuple:

Tuple	Class	Probability
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.6
6	P	0.55
7	N	0.53
8	P	0.52
9	N	0.51
10	N	0.4

Table 5

(a) Calculate true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

(b) Calculate true positive rate (TPR) and false positive rate (FPR).

(c) Plot the ROC curve for the data.

Solution:

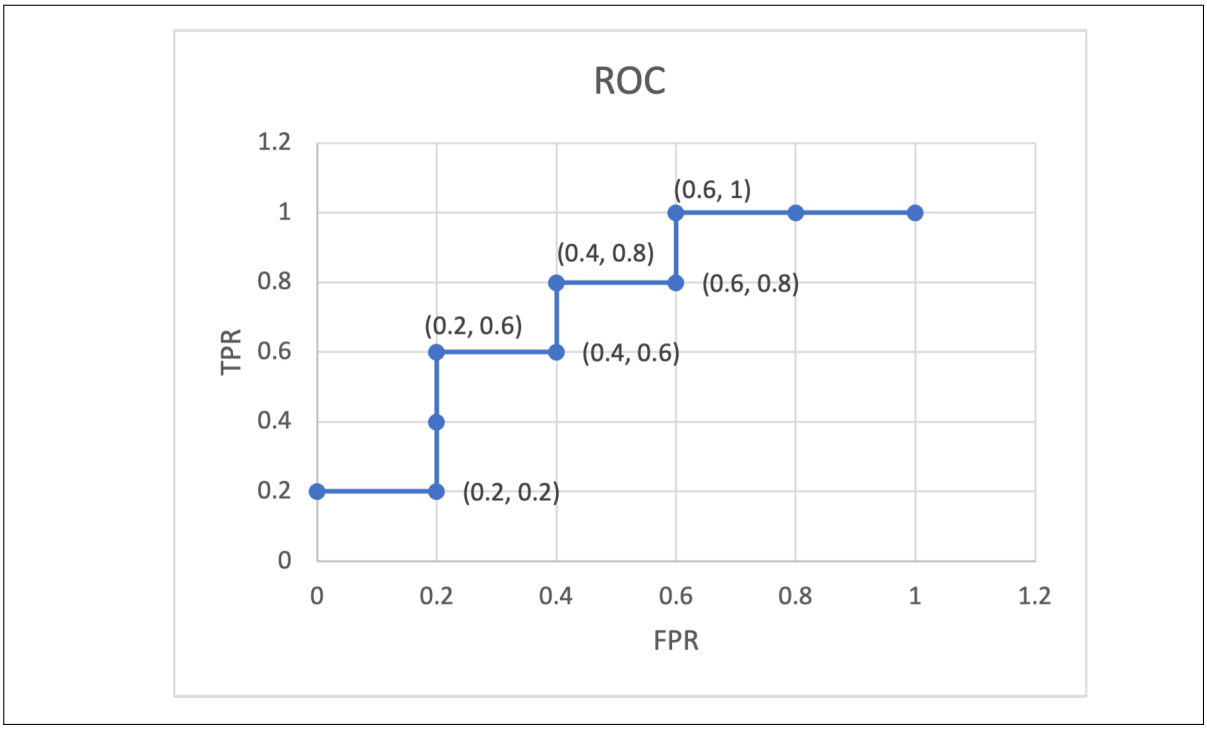
(a) The probability column in Table 5 indicates the likelihood of the tuple in the class P.

If we set the probability threshold between 0.95 and 0.85, then only tuple 1 is been predicted to belong to class P. The rest tuples are predicted to belong to class N. Therefore, all 5 "N" tuples are labeled "N" and only 1 "P" tuple is labeled "P" (tuple 1), with 4 "P" tuple labeled "N". Thus we have TP=1, FP=0, TN=5 and FN=4. The $TPR = \frac{TP}{TP+FN} = 0.2$ and $FPR = \frac{FP}{FP+TN} = 0$.

When we move the probability threshold between 0.85 to 0.78, then tuple 2 is also been predicted to belong to class P. In this way, 4 "N" tuples are labeled "N" with 1 "N" tuples are labeled "P" (tuple 2), 1 "P" tuple is labeled "P" (tuple 1), with 4 "P" tuple labeled "N". Thus we have TP=1, FP=1, TN=4 and FN=4. The $TPR = \frac{TP}{TP+FN} = 0.2$ and $FPR = \frac{FP}{FP+TN} = 0.2$.

	TP	FP	TN	FN	TPR	FPR
P	1	0	5	4	0.2	0
N	1	1	4	4	0.2	0.2
P	2	1	4	3	0.4	0.2
P	3	1	4	2	0.6	0.2
N	3	2	3	2	0.6	0.4
P	4	2	3	1	0.8	0.4
N	4	3	2	1	0.8	0.6
P	5	3	2	0	1	0.6
N	5	4	1	0	1	0.8
N	5	5	0	0	1	1

(b) The sample ROC curve is plotted below:



2.5) For the first clustering, the central coordinates of the two clusters are as follows:

Clusters	Centre coordinates	
	X_1	X_2
(A, B)	1	1
(C, D)	-2	-3

Step 2: Calculate the Euclidean square distance of a sample to the centre of each class and then assign that sample to the nearest class. For classes where samples have changed, recalculate their centre coordinates in preparation for the next clustering step. Start by calculating the squared distances from A to the two clusters:

$$d^2(A,(AB)) = (4-1)^2 + (2-1)^2 = 10$$

$$d^2(A,(CD)) = (4+2)^2 + (2+3)^2 = 61$$

Since the distance from A to (A, B) is less than the distance to (C, D), A does not need to be reassigned. Calculate the squared distance from B to both classes:

$$d^2(B,(AB)) = (-2-1)^2 + (0-1)^2 = 10$$

$$d^2(B,(CD)) = (-2+2)^2 + (0+3)^2 = 9$$

Since the distance from B to (A, B) is greater than the distance to (C, D), B is to be assigned to the class (C, D), giving the new clusters (A) and (B, C, D). Update the centre coordinates as shown in the table below.

Clusters	Centre coordinates	
	$\overline{X}_1$	$\overline{X}_2$
(A)	4	2
(B, C, D)	-2	-2

Step 3: Each sample was checked again to determine if it needed to be re-clustered. The squared distance from each sample to each centre is calculated and the results are shown in the table below.

Clusters	Square of the distance from the sample to the centre			
	A	B	C	D
A	0	40	41	89
(B, C, D)	40	2	5	5

By far, each sample has been assigned to the class closest to the centre, so this concludes the clustering process. The final clustering result for $K = 2$ is that A is alone in one cluster, and B, C and D are clustered in one cluster.