Data Mining

Study Assignment Set #3

Prepared by: P V S Maruthi Rao | pvsmaruthirao@wilp.bits-pilani.ac.in

Reference Books:
Introduction to Data Mining by Tan P. N., Steinbach M and Kumar V.
Pearson Education, 2006

Data Mining: Concepts and Techniques, Second Edition by Jiawei Han and Micheline Kamber
Morgan Kaufmann Publishers, 2006

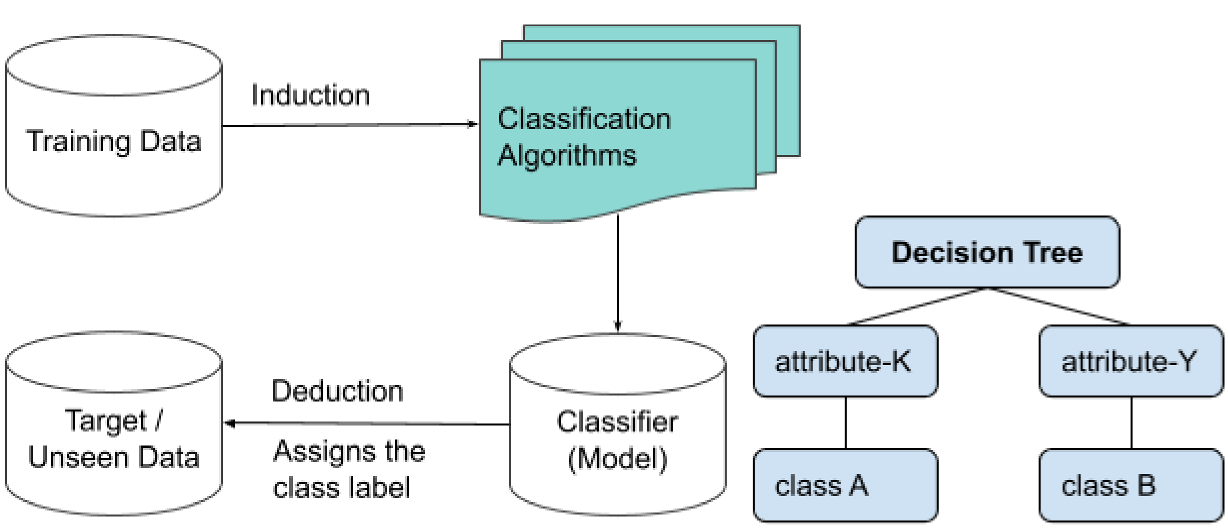
Topic: Classification of Data, Decision Trees, Gain Ratio

Classification of Data, Decision Trees	Question 1

Learning objectives:
- Basics of statistical learning with Decision Trees.
- Decision Tree algorithm, and attribute selection methods.
- **Attribute selection by 'Gain Ratio'**
- **C4.5,** a supervised learning algorithm proposed by J Ross Quinlan, uses attribute selection by Gain Ratio method. C4.5 is a successor of ID3, a supervised learning algorithm proposed by J Ross Quinlan, that uses 'Information Gain' as an attribute selection method.

Prerequisites:
- Study Assignment Set #1 (Conditional probability)
- Study Assignment Set #2 (Entropy, Information, Information Gain).

Basics of statistical learning learning with Decision Trees:



Some of the formulae are given as below.

$$Info\,(D) \;=\; -\sum_{i=1}^{m} p_i log_2(p_i)$$

D: Training Data
m: Distinct values of the class label attribute.
p_i: non-zero probability that an **attribute tuple** in D belongs to a **class Y_i** and is estimated by $|Y_i, D| / |D|$

** $P(Y_i \mid D) = P(Y_i, D) / P(D) = |Y_i, D| / |D|$ **

[Some use C_i for class.]

How much more information would we still need (after partitioning) to arrive at an exact classification? Measure $Info_A(D)$ for attribute A as below.

$$Info_A\,(D) \;=\; \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

$Info_A(D)$ is the expected information required to classify a tuple from D based on the partition by the attribute A. The smaller the information (still) required, the greater the purity of the partition.

Gain (A) = Info (D) - $Info_A(D)$
Gain (A) is an indication of how much would be gained by branching on A (attribute A).

** Branch on the attribute that gives highest gain **

The C4.5 supervised learning algorithm applies a kind of **normalization to information gain** using a “**split information**” value defined as below.

$$SplitInfo_A\,(D) \;=\; -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2 \frac{|D_j|}{|D|}$$

v is a set of possible partitions on split attribute A.

The Gain Ratio:

$$GainRatio(A) \;=\; \frac{Gain(A)}{SplitInfo_A(D)}$$

A is an Attribute, D is a training data set.

** Select the attribute with highest ‘Gain Ratio’ **

If Split Info is approaching zero, the gain ratio is unstable. So a constraint is added to avoid this, whereby the information gain of the test selected must be large - at least as great as the average gain over all tests examined.

Note: *When a calculation, system or subsystem behavior is tending towards unstable, then* *design a constraint to avoid such instability.*

Classification of Data	Question 2

Learning objectives:

- Basics of statistical learning with Decision Trees.
- Decision Tree algorithm, and attribute selection methods.
- **Attribute selection by 'Gain Ratio'**
- **C4.5,** a supervised learning algorithm proposed by J Ross Quinlan, uses attribute selection by Gain Ratio method. C4.5 is a successor of ID3, a supervised learning algorithm proposed by J Ross Quinlan, that uses 'Information Gain' as an attribute selection method.

Prerequisites:

- Study Assignment Set #1 (Conditional probability)
- Study Assignment Set #2 (Entropy, Information, Information Gain).
- Study Assignment Set #3 (Question 1).

An online computer store uses a Decision Tree classifier with '**Gain Ratio**' as a method of attribute selection method. Please see the Question #1 above for Gain Ratio.

Let X is a set of attributes of the registered user.
X = {id, age, income, student, credit_rating}

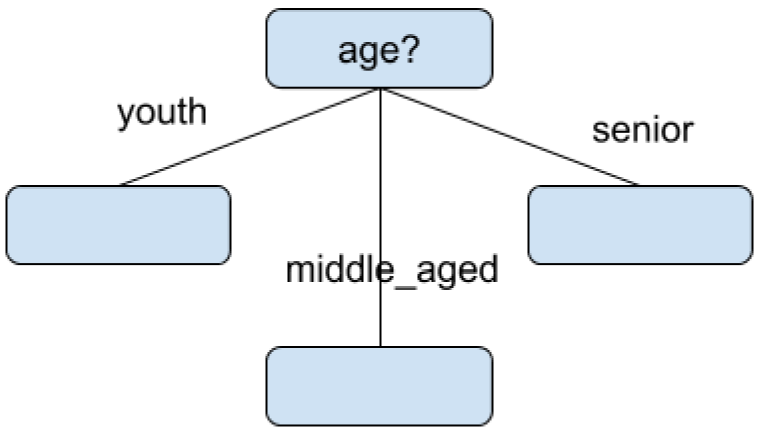
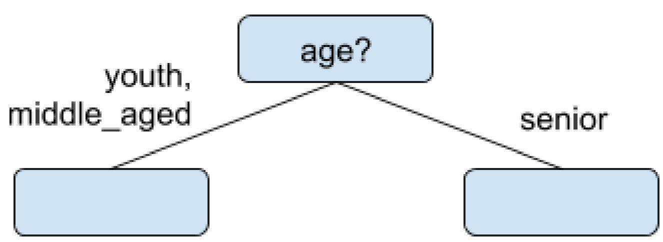
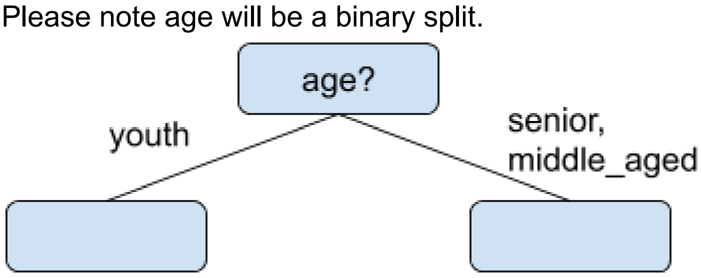
Let Y is the class variable
Y = buys_computer = {yes, no}

The **training dataset**, D, is as below.

id	**age**	**income**	**student**	**credit_rating**	**buys_computer**
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	no

Questions:

A	Find - Info (D) (Entropy of the D) for the class label attribute, Y = buys_computer = {yes, no}.
B	- $Info_{age}(D)$ - Gain (age) # Let age is chosen as a splitting attribute. - Gain Ratio (age = {youth, middle_aged, senior}). Apply v = {youth, middle_aged, senior} in the SplitInfo formula. 
C	- Gain (age) # Let age is chosen as a splitting attribute. Please note, the split is {{youth, middle_aged}, senior}. It means, we have to calculate Gain (age = {{youth, middle_aged}, senior}). - Gain Ratio (age = {{youth, middle_aged}, senior}). Apply v = {youth, middle_aged}, senior} in the SplitInfo formula. Please note age will be a binary split. 
D	- Gain (age) # Let age is chosen as a splitting attribute. Please note, the split is {youth, {middle_aged, senior}}. It means, we have to calculate Gain (age = {youth, {middle_aged, senior}}). - Gain Ratio (age = {youth, {middle_aged, senior}}). Apply v = {youth, {middle_aged, senior}} in the SplitInfo formula. Please note age will be a binary split. 

Answers:

A	D has a total 14 tuples (training data). m: Distinct values of the class label attribute = 2. buys_computer has two distinct value {yes, no}.

C	p(buys_computer = yes | D) = 9/14 p(buys_computer = no | D) = 5/14 $$Info\,(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$$ Info (D) = -9/14 $\log_2$ (9/14) - 5/14 $\log_2$ (5/14) = **0.940 bits.** It is also known as Entropy of D.
B	Let's select age as a splitting attribute. D: Training data set. Class label Y: buys_computer = {yes, no}. A: age v_{age}: {youth, middle_aged, senior} Info (D) = -9/14 $\log_2$ (9/14) - 5/14 $\log_2$ (5/14) = **0.940 bits.** It is also known as Entropy of D. Info (D_{age}) = (5/14 * Info(D_{youth})) + (4/14 * Info(D_{middle_aged})) + (5/14 * Info(D_{senior})) Gain (age) = Info (D) - Info (D_{age}) = 0.246 bits. (From the Assignment Set #2 OR you may calculate here) $$SplitInfo_A\,(D) = -\sum_{j=1}^{v} \frac{\|D_j\|}{\|D\|} \times log_2 \frac{\|D_j\|}{\|D\|}$$ $$SplitInfo_A\,(D) = -\frac{\|D_{youth}\|}{\|D\|} \times log_2 \frac{\|D_{youth}\|}{\|D\|} - \frac{\|D_{middle\_aged}\|}{\|D\|} \times log_2 \frac{\|D_{middle\_aged}\|}{\|D\|} - \frac{\|D_{senior}\|}{\|D\|} \times log_2 \frac{\|D_{senior}\|}{\|D\|}$$ $SplitInfo_{age}$ (D) = -5/14$\log_2$(5/14) - 4/14$\log_2$(4/14) - 5/14$\log_2$(5/14) = 1.5447 bits $$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$ GainRaio (age) = Gain (age) / $SplitInfo_{age}$ (D) GainRaio (age) = 0.246 bits / 1.5447 bits = **0.1559**
C	Let's select age as a splitting attribute. D: Training data set. Class label Y: buys_computer = {yes, no}. A: age v_{age}: {{youth, middle_aged}, senior} Info (D) = -9/14 $\log_2$ (9/14) - 5/14 $\log_2$ (5/14) = **0.940 bits.**

It is also known as Entropy of D.

From the data set D,

	age = youth	age = middle_aged	age = senior	
buys_computer = yes	2	4	3	SUM = 9
buys_computer = no	3	0	2	SUM = 5
	SUM = 5	SUM = 4	SUM = 5	

Info (D_{age}) = (9/14 * Info($D_{youth,\ middle\_aged}$)) + (5/14 * Info(D_{senior}))
= 9/14 (-6/9 $\log_2$ (6/9) - 3/9$\log_2$(3/9)) + 5/14 (-3/5$\log_2$ (3/5) - 2/5$\log_2$(2/5))
= **0.9371**

Gain (age) = Info (D) - Info (D_{age})
= 0.940 - 0.9371
= 0.0029

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|}$$

Let's select age as a splitting attribute.
D: Training data set.
A: age
v_{age}: {{youth, middle_aged}, senior}

$$SplitInfo_A(D) = -\frac{|D_{youth,\ middle\_aged}|}{|D|} \times \log_2 \frac{|D_{youth,\ middle\_aged}|}{|D|} - \frac{|D_{senior}|}{|D|} \times \log_2 \frac{|D_{senior}|}{|D|}$$

$SplitInfo_{age}$ (D) = -9/14$\log_2$(9/14) - 5/14$\log_2$(5/14)
= 0.9403 bits

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$

GainRaio (age) = Gain (age) / $SplitInfo_{age}$ (D)

GainRaio (age) = 0.0029 bits / 0.9403 bits
= **0.0031**

D	Follow the answer B) above and calculate the Gain Ratio for the spilt age = {youth, {middle_aged, senior}}.

Remarks	In the above example, the root node is split on age. The tree gets added with new nodes or split partitions recursively.