Data Mining

Study Assignment Set #2

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

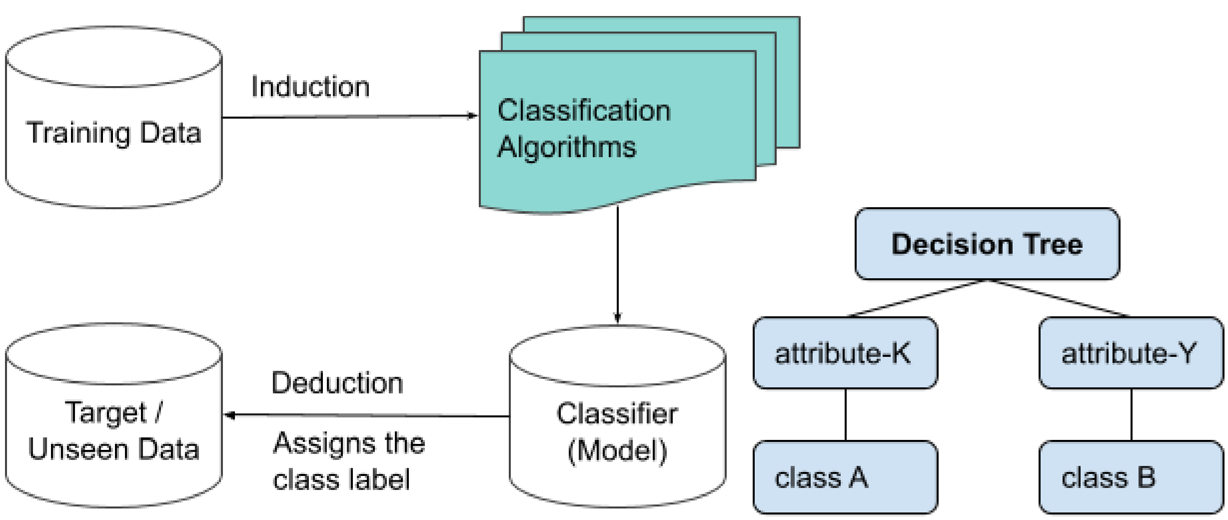
Topic: Classification of Data, Decision Trees, Information Gain

---

Classification of Data, Decision Trees	**Question 1**

Learning objectives:
- Basics of statistical learning with Decision Trees.
- Decision Tree algorithm, and attribute selection methods.

Basics of statistical learning learning with Decision Trees:



Below is the generic algorithm to generate a decision tree. It results in a decision tree with nodes split on a certain criterion.

Node **`Generate_decision_tree (Dj, attribute_list) {`**

```
1.   Create a Node N
2.   if tuple in D are all of the same class, C, then
3.       return N as a leaf node labelled with the class C;
4.   if attribute_list is empty then
        // majority voting
5.      return N as a leaf node labelled with the majority class in D;
6.   apply Attribute_selection_method (D, attribute_list) to find the best
     splitting_criterion;
7.   label node N with splitting_criterion;
8.   if splitting_attribute is discrete-valued and
        multiway splits allowed then  // not restricted to binary trees
     // remove splitting_attribute
```

```
attribute_list <- attribute_list -  splitting_attribute;
for each outcome j of splitting_criterion
// partition the tuples and grow subtrees for each partition
 let Dj be the set of data tuples in D satisfying outcome j; // a
partition
   if Dj is empty then
      attach a leaf labeled with the majority class in D to node N;
   else attach the node returned by Generate_decision_tree (Dj,
attribute_list) to node N;
  endfor
return N;
}
```

Questions:

Q	What are popular attribute selection methods for the above decision tree generation algorithm?

A	Information Gain	B	Gain Ratio
C	Gini Index	D	All of the above.

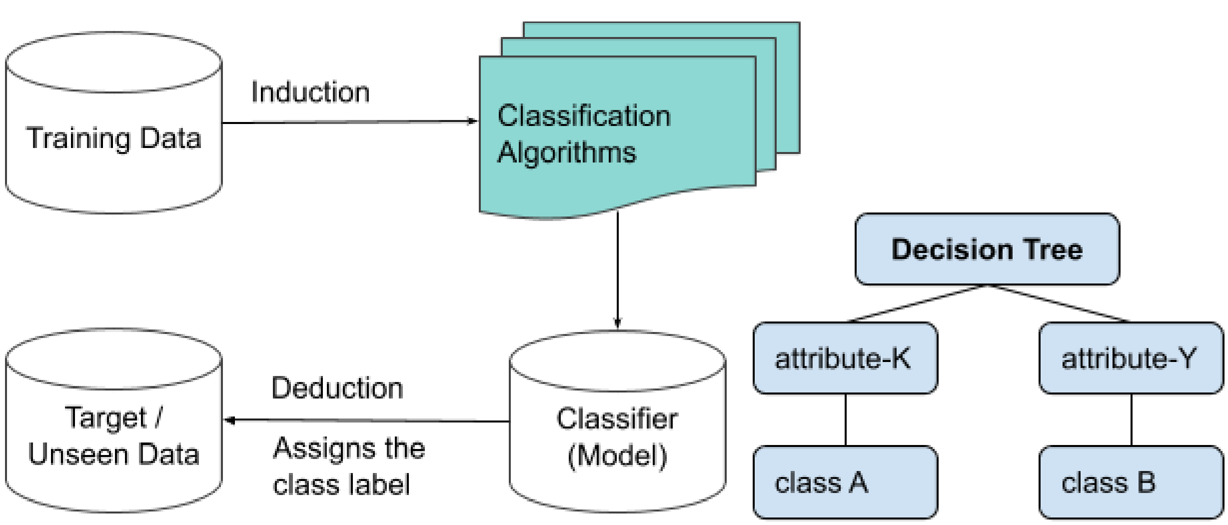
Answer	**D**
Remarks	

Classification of Data	Question 2

Learning objectives:

- Basics of statistical learning with Decision Trees.
- Generating Decision Tree (Refer to Question #1)
- Understanding the concept of Entropy.

Basics of statistical learning learning with Decision Trees:



Use of Information Gain for attribute selection.

- Attribute with highest information gain is chosen as the splitting attribute for a node N
- This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or “impurity” in these partitions.

Entropy:

The expected information needed to classify a tuple in D is given by

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

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.]

Example:
An online computer store uses a Decision Tree classifier with 'Info Gain' as a method of attribute selection method.

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}.

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}. 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.

MCQ:

A		B	
C		D	

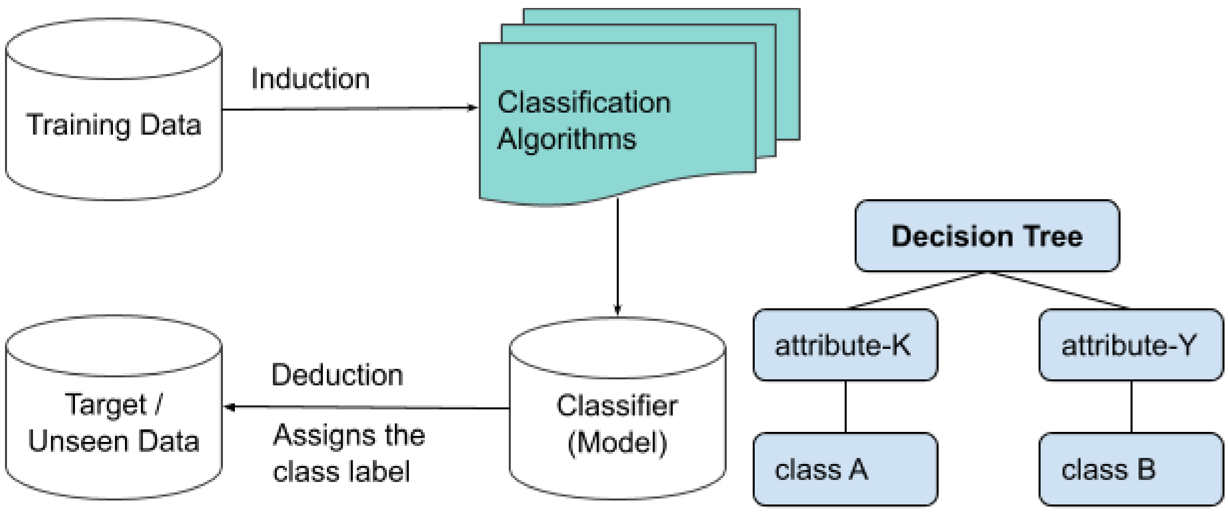
Answer	
Remarks	

Classification of Data	Question 3

Learning objectives:

- Basics of statistical learning with Decision Trees.
- Generating Decision Tree (Refer to Question #1)
- Understanding the concept of Entropy. (Refer to Question #2)
- Understanding attribute selection ('Info Gain').

Basics of statistical learning learning with Decision Trees:



Use of Information Gain for attribute selection.

- Attribute with highest information gain is chosen as the splitting attribute for a node N
- This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions.

Entropy:

The expected information needed to classify a tuple in D is given by

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

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.]

Usually, Info (D) is the entropy of the Root node. Our objective is to find an appropriate 'attribute' to perform a split on D.

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) - $\mathbf{Info_A(D)}$
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 **

Example:
An online computer store uses a Decision Tree classifier with 'Info Gain' as a method of attribute selection method.

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	Find Info_{age}(D), Gain (age)
C	Find Info_{income}(D), Gain (income)
D	Find $\text{Info}_{student}$(D), Gain (student)
E	Find $\text{Info}_{credit\_rating}$(D), Gain (credit_rating)

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}. 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	Find Info_{age}(D), Gain (age) $$Info_A\,(D) = \sum_{j=1}^{v} \frac{\|D_j\|}{\|D\|} \times Info(D_j)$$ Attribute age = {youth, middle_aged, senior}. So, if we split D on **age**, then it will partition D into 3 partitions. Therefore v = 3.

<table>
<tr><td></td><td>

$$Info_{age} = \frac{|D_{youth}|}{|D|} \times Info(D_{youth}) + \frac{|D_{middle\_aged}|}{|D|} \times Info(D_{middle\_aged}) + \frac{|D_{senior}|}{|D|} \times Info(D_{senior})$$

Calculating Info(D_{youth}) ----------- (1)
D_{youth} has a total 5 tuples (training data).
m: Distinct values of the class label attribute = 2. buys_computer has two distinct value {yes, no}.
p(buys_computer = yes | D_{youth}) = 2/5
p(buys_computer = no | D_{youth}) = 3/5
Info (D_{youth}) = -2/5 $\log_2$ (2/5) - 3/5 $\log_2$ (3/5) bits

Calculating Info(D_{middle_aged}) ------------ (2)
D_{middle_aged} has a total 4 tuples (training data).
m: Distinct values of the class label attribute = 2. buys_computer has two distinct value {yes, no}.
p(buys_computer = yes | D_{middle_aged}) = 4/4
p(buys_computer = no | D_{middle_aged}) = 0
Info (D_{middle_aged}) = -4/4 $\log_2$ (4/4) bits

Calculating Info(D_{senior}) --------------- (3)
D_{senior} has a total 5 tuples (training data).
m: Distinct values of the class label attribute = 2. buys_computer has two distinct value {yes, no}.
p(buys_computer = yes | D_{senior}) = 3/5
p(buys_computer = no | D_{senior}) = 2/5
Info (D_{middle_aged}) = -3/5 $\log_2$ (3/5) -2/5 $\log_2$ (2/5) bits

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.

</td></tr>
<tr><td>C</td><td>

Please do the numerical calculation of Gain (income).

$$Info_{income} = \frac{|D_{high}|}{|D|} \times Info(D_{high}) + \frac{|D_{medium}|}{|D|} \times Info(D_{medium}) + \frac{|D_{low}|}{|D|} \times Info(D_{low})$$

</td></tr>
<tr><td>D</td><td>

Please do the numerical calculation of Gain (student).

$$Info_{student} = \frac{|D_{yes}|}{|D|} \times Info(D_{yes}) + \frac{|D_{no}|}{|D|} \times Info(D_{no})$$

</td></tr>
<tr><td>E</td><td>

Please do the numerical calculation of Gain (credit_rating).

$$Info_{credit\_rating} = \frac{|D_{fair}|}{|D|} \times Info(D_{fair}) + \frac{|D_{excellent}|}{|D|} \times Info(D_{excellent})$$

</td></tr>
</table>

MCQ:

A		B	

C		D	

Answer	
Remarks	