Data Mining

Study Assignment Set #4

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

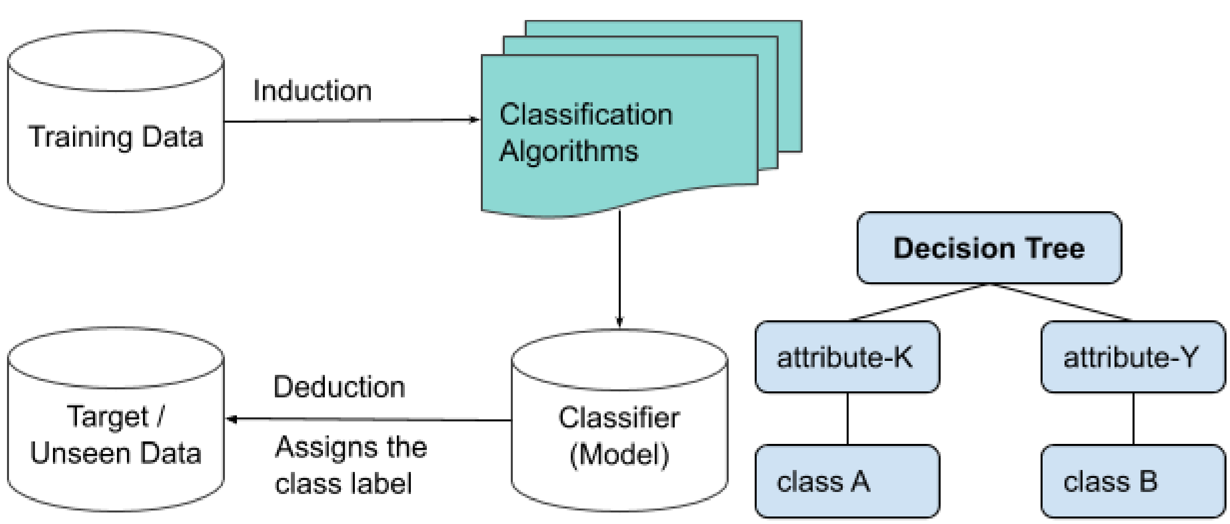
Topic: Classification of Data, Decision Trees, Gini Index

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 'Gini Index'**
- **CART (Classification and Regression Trees),** a supervised learning algorithm uses attribute selection by **Gini Index** method.

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

Basics of statistical learning learning with Decision Trees:



Some of the formulae are given as below.

Entropy

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

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

Split Information

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.

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.*

Gini Index
Gini index measures the impurity of D, a data partition or a set of training tuples as

$$Gini(D) = 1 - \sum_{j=1}^{m} p_i^2$$

where p_i is the probability of a tuple in D belongs to a class C_i (Y_i) and is estimated by $|C_{i,D}| / |D|$

m: Class labels {1 … m}

For example, m of Class label Y, buys_computer, is {yes, no}.

** The Gini Index considers a binary split for every attribute. **

How to split an attribute A = {a1, a2, a3 … av}, where a1...av are discrete values of attribute A can assume.

So, there will be 2^v possible combination of subsets.

Excluding the power set and a null set, there are 2^v - 2 possible ways to form two partitions of the data, D, based on a binary split on A.

Let's take an example:

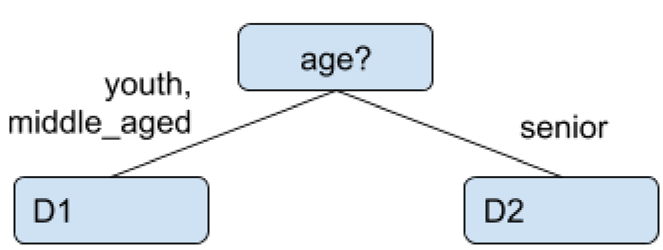
age = {youth, middle_aged, senior}

There are 8 possible ways to split on the attribute age.

	Subset S_A
1	{youth, middle_aged, senior} <-- known as Power set.
2	{youth}
3	{youth, middle_aged}
4	{middle_aged}
5	{middle_aged, senior}
6	{senior}
7	{youth, senior}
8	{} <-- known as null set

Discarding the Power Set and Null Set, we are left with 6 subsets.

Each subset, S_A, can be considered as a binary test for attribute A of the form "$A \in S_A$?"



In the above example, {youth, middle_aged} splits the D into 2 partitions namely D1 and D2. Compute a weighted sum of the impurity of each resulting partition.

$$Gini(D) = 1 - \sum_{j=1}^{m} p_i^2$$

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D1) + \frac{|D_2|}{|D|} Gini(D2)$$

Finding Gini Index on every attribute (and possible binary splits) is required to determine the best split by considering the '**Lowest Gini Index**'.

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

Continue to the next question.

Classification of Data	Question 2

Learning objectives:

- Basics of statistical learning with Decision Trees.
- Decision Tree algorithm, and attribute selection methods.
- **Attribute selection by 'Gini Index'**
- **CART (Classification and Regression Trees),** a supervised learning algorithm uses attribute selection by **Gini Index** method.

Prerequisites:

- Study Assignment Set #1 (Conditional probability)
- Study Assignment Set #2 (Entropy, Information, Information Gain).
- Study Assignment Set #3 (Entropy, Information, Information Gain, Gain Ratio).
- Study Assignment Set #4 (Question 1).

An online computer store uses a Decision Tree classifier with '**Gini Index**' 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

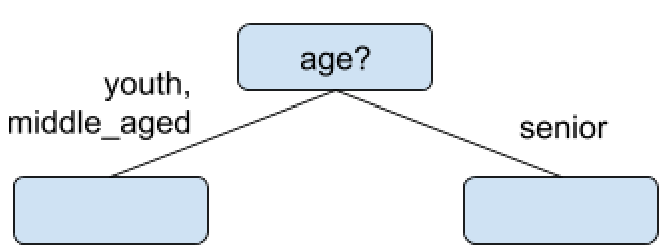
Let's find the Gini Index on age.
age = {youth, middle_aged, senior}
There are 8 possible ways to split on the attribute age.

	Subset S_A
1	{youth, middle_aged, senior} <-- known as Power set.
2	{youth}
3	{youth, middle_aged}
4	{middle_aged}
5	{middle_aged, senior}
6	{senior}
7	{youth, senior}
8	{} <-- known as null set

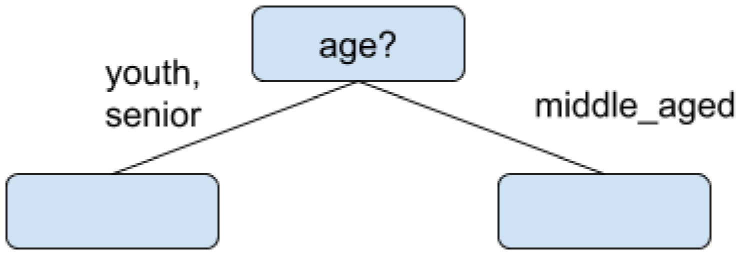
Discarding the Power Set and Null Set, we are left with 6 subsets.

Questions:

<table>
<tr><td>A</td><td>Find
- Gini (D) for the class label attribute, Y = buys_computer = {yes, no}.
m: {yes, no}</td></tr>
<tr><td>B</td><td>- $Gini_{age \in \{ youth, middle\_aged\}}$ (D)
It means the split is on the subset {youth, middle_aged}
m: class label attribute, Y or C = buys_computer = {yes, no}.

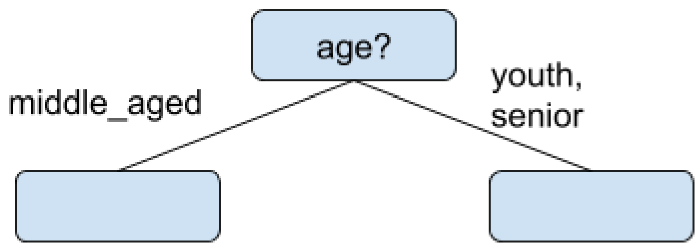


</td></tr>
<tr><td>C</td><td>- $Gini_{age \in \{ youth, senior\}}$ (D)
It means the split is on the subset {youth, senior}
m: class label attribute, Y or C = buys_computer = {yes, no}.</td></tr>
</table>

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


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

- Gini$_{age \in \{middle_aged\}}$ (D)
It means the split is on the subset {middle_aged}
m: class label attribute, Y or C = buys_computer = {yes, no}.



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

Answers:

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

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

$$Gini(D) = 1 - \sum_{j=1}^{m} p_i^2$$

Gini (D) = 1 - (9/14)^2 - (5/14)^2
= **0.459**
</td></tr>
<tr><td>B</td><td>

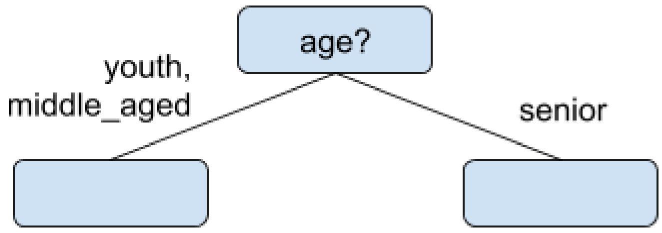
Let's select age as a splitting attribute.
D: Training data set.
Class label Y: buys_computer = {yes, no}.
A: age
v$_{age}$: {youth, middle_aged, senior}

From the data set D,

<table>
<tr><td></td><td>age = youth</td><td>age = middle_aged</td><td>age = senior</td><td></td></tr>
<tr><td>buys_computer = yes</td><td>2</td><td>4</td><td>3</td><td>SUM = 9</td></tr>
<tr><td>buys_computer = no</td><td>3</td><td>0</td><td>2</td><td>SUM = 5</td></tr>
<tr><td></td><td>SUM = 5</td><td>SUM = 4</td><td>SUM = 5</td><td></td></tr>
</table>

Gini$_{age \in \{youth, middle_aged\}}$ (D)
</td></tr>
</table>

It means the split is on the subset {youth, middle_aged}
m: class label attribute, Y or C = buys_computer = {yes, no}.



D1 is the partition created by the attribute age with a subset {youth, middle_aged}.
D2 is the partition created by the attribute age that are not in the subset {youth, middle_aged}.
Please note, it is a binary split.

$$Gini(D) = 1 - \sum_{j=1}^{m} p_i^2$$

Gini (D1) = 1 - (6/9)^2 - (3/9)^2 = 0.4444
Gini (D2) = 1- (3/5)^2 - (2/5)^2 = 0.48

$$Gini_A(D) = \frac{|D_1|}{|D|}Gini(D1) + \frac{|D_2|}{|D|}Gini(D2)$$

$Gini_{age \in \{\text{youth, middle\_aged}\}}$ (D) = (9/14)*(0.4444) + (5/14)*0.48 = **0.4571**

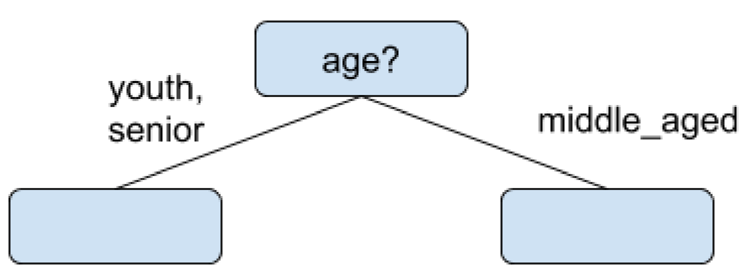
C

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}

From the data set D,

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

$Gini_{age \in \{\text{youth, senior}\}}$ (D)
It means the split is on the subset {youth, seir}
m: class label attribute, Y or C = buys_computer = {yes, no}.



D1 is the partition created by the attribute age with a subset {youth, senior}.
D2 is the partition created by the attribute age that are not in the subset {youth, senior}.
Please note, it is a binary split.

	$$Gini(D) = 1 - \sum_{j=1}^{m} p_i^2$$ Gini (D1) = 1 - (5/10)^2 - (5/10)^2 = 0.5 Gini (D2) = 1- (4/4)^2 - 0 = 0 $$Gini_A(D) = \frac{\lvert D_1 \rvert}{\lvert D \rvert} Gini(D1) + \frac{\lvert D_2 \rvert}{\lvert D \rvert} Gini(D2)$$ $Gini_{age \in \{youth, senior\}}$ (D) = (10/14)*(0.5) + (4/14)*0 = **0.3571**
D	Please practice the assignment.