Decision Tree

Course Overview

You are here...

Term	CDF	GCD	GCDAI	PGPDSAI
Term 1	Data Analytics with Python	Data Analytics with Python	Data Analytics with Python	Data Analytics with Python
Term 2	Data Visualization Techniques	Data Visualization Techniques	Data Visualization Techniques	Data Visualization Techniques
Term 3	EDA & Data Storytelling	EDA & Data Storytelling	EDA & Data Storytelling	EDA & Data Storytelling
		Minor Project	Minor Project	Minor Project
Term 4		Machine Learning Foundation	Machine Learning Foundation	Machine Learning Foundation
Term 5		Machine Learning Intermediate	Machine Learning Intermediate	Machine Learning Intermediate
Term 6		Machine Learning Advanced (Mandatory)	Machine Learning Advanced (Mandatory)	Machine Learning Advanced (Mandatory)
		Data Visualization with Tableau (Elective - I)	Data Visualization with Tableau (Elective - I)	Data Visualization with Tableau (Elective - I)
		Data Analytics with R (Elective - II)	Data Analytics with R (Elective - II)	Data Analytics with R (Elective - II)
		Capstone Project	Capstone Project	Capstone Project
Term 7		Bonus: Industrial ML (ML – 4 & 5)	Basics of AI, TensorFlow, and Keras	Basics of AI, TensorFlow, and Keras
Term 8			Deep Learning Foundation	Deep Learning Foundation
Term 9			NPL – I/CV – I	CV – I
Term 10			NLP – II/CV – II	NLP – I
			Capstone Project	Capstone Project
Term 11				CV – II
Term 12				NLP – II
				NLP – III + CV – III
				AutoVision & AutoNLP
				Building AI product

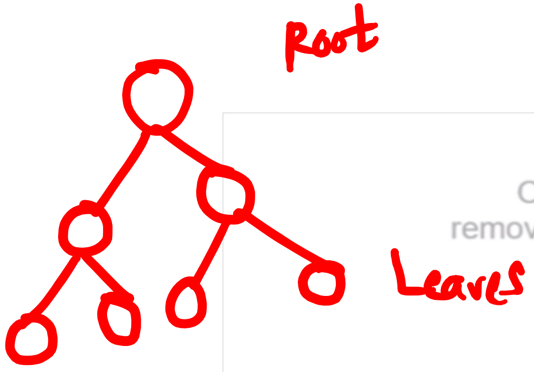
Term Context

- **Decision Tree** You are here...
- Random Forest
- Principal Component Analysis
- Naïve Bayes Classifier

Agenda

- **Terminology Related to Trees**
- Decision Tree
- Decision Tree Algorithms
- Attribute Selection Measures
- ID3 Algorithm
- Entropy & Information Gain
- Steps to Estimate Entropy & Information Gain
- CART Algorithm
- Gini Index
- Steps to estimate Gini Index
- CART – Regression Example
- Issues with Decision Trees
- Tree Pruning
- Decision Tree Applications

Terminology Related to Trees



Pruning
Opposite of Splitting, basically removing unwanted branches from the tree

Branch/SubTree
Formed by splitting the tree/node

Parent/Child Node
Root node is the parent node and all the other nodes branched from it is known as child node

Splitting
Splitting is dividing the root node/sub node into different parts on the basis of some condition.

Root Node
It represents the entire population or sample and this further gets divided into two or more homogenous sets.

Leaf Node
Node cannot be further segregated into further nodes

Agenda

- Terminology Related to Trees
- **Decision Tree**
- Decision Tree Algorithms
- Attribute Selection Measures
- ID3 Algorithm
- Entropy & Information Gain
- Steps to Estimate Entropy & Information Gain
- CART Algorithm
- Gini Index
- Steps to estimate Gini Index
- CART – Regression Example
- Issues with Decision Trees
- Tree Pruning
- Decision Tree Applications

Decision Tree

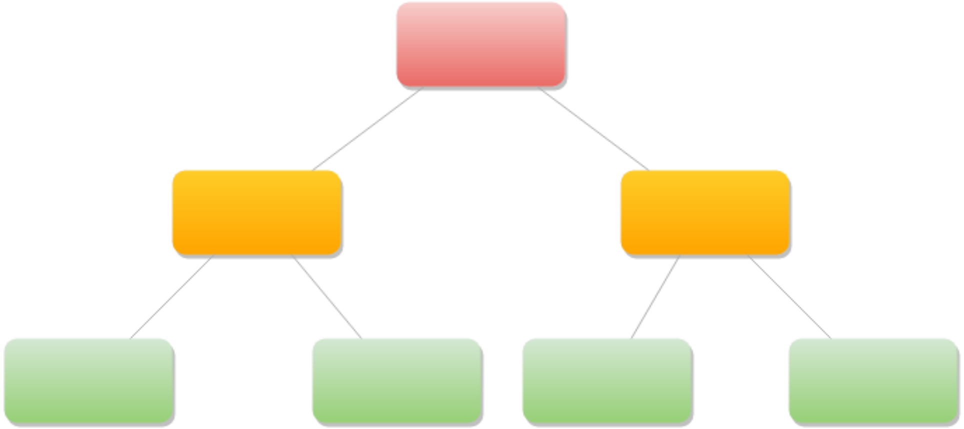
$X_1 + X_2 + \cdots + X_n \sim y$

- It is a type of **Supervised Learning** algorithm which works for both **Categorical** as well as **Continuous** data.
- A **Decision Tree** is a graphical representation of all the possible solutions based on certain conditions.
- These solutions can be seen as IF-THEN rules.

1

2

1+2

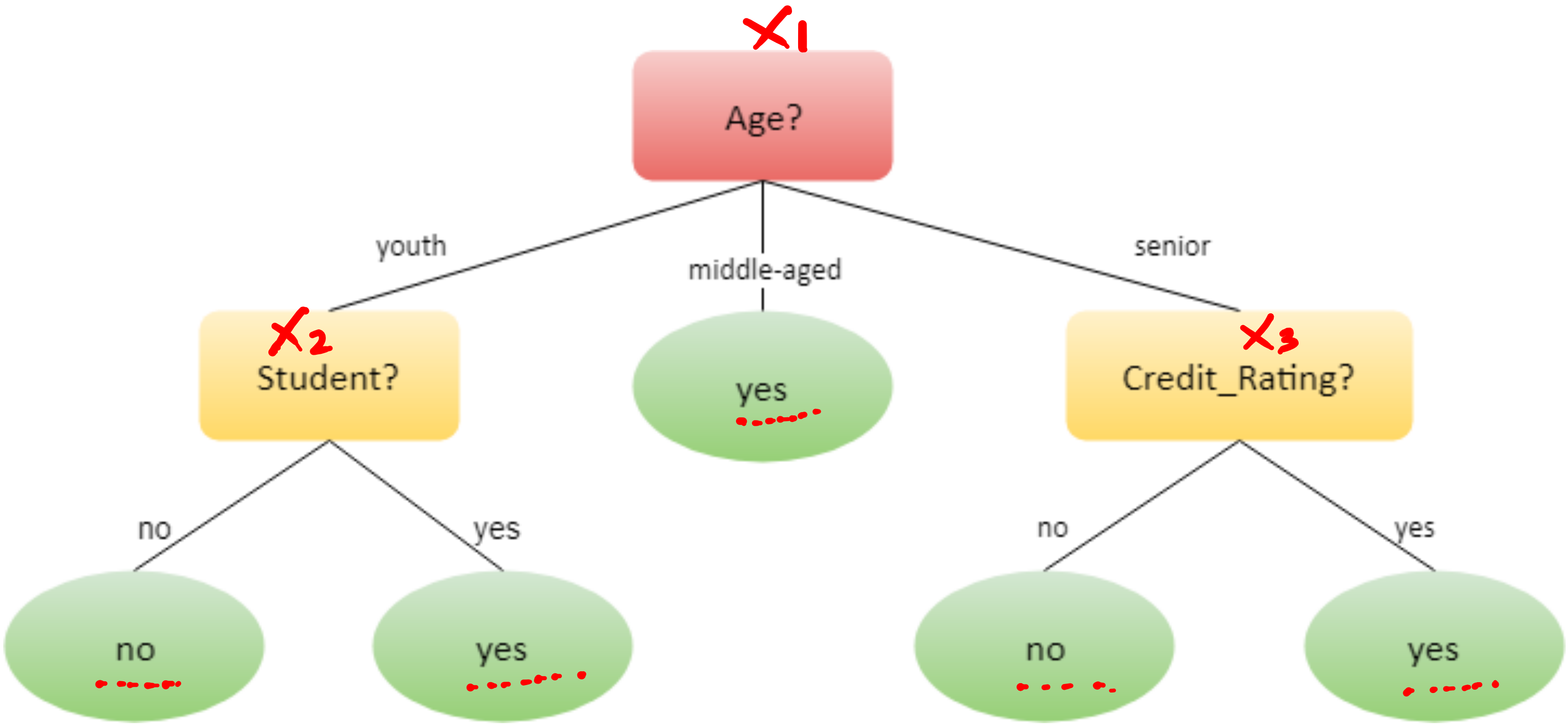


(Decision Tree)

Decision Tree Example 1

x_1 Age, x_2 Student, x_3 Credit Rating ~ y Buys Comp

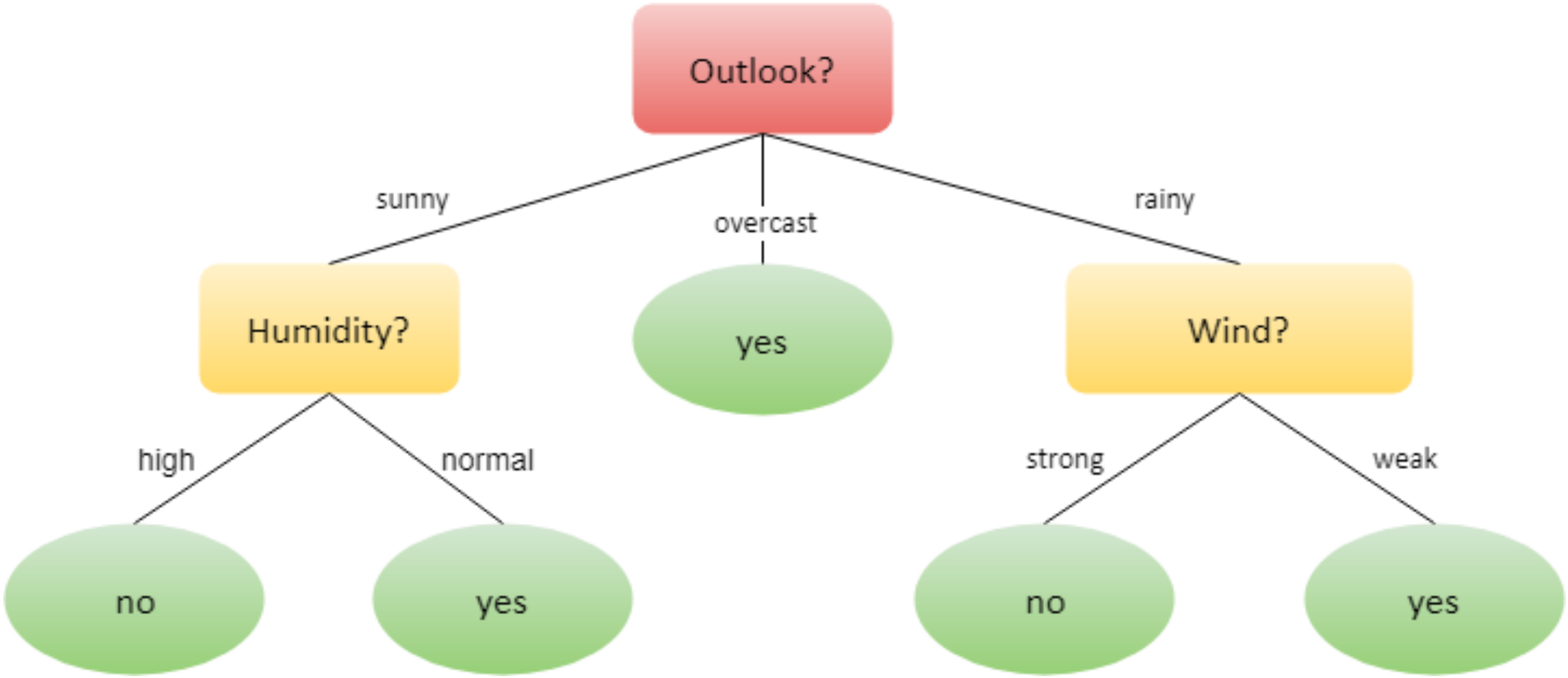
- **Buys-Computer Classification:** Based on certain factors whether a person will buy computer or not.



INSAID

Decision Tree Example 2

- **Play Golf Classification**: Whether a person will play or not based on environmental factors.



Agenda

- Terminology Related to Trees
- Decision Tree
- **Decision Tree Algorithms**
- Attribute Selection Measures
- ID3 Algorithm
- Entropy & Information Gain
- Steps to Estimate Entropy & Information Gain
- CART Algorithm
- Gini Index
- Steps to estimate Gini Index
- CART – Regression Example
- Issues with Decision Trees
- Tree Pruning
- Decision Tree Applications

Decision Tree Algorithms

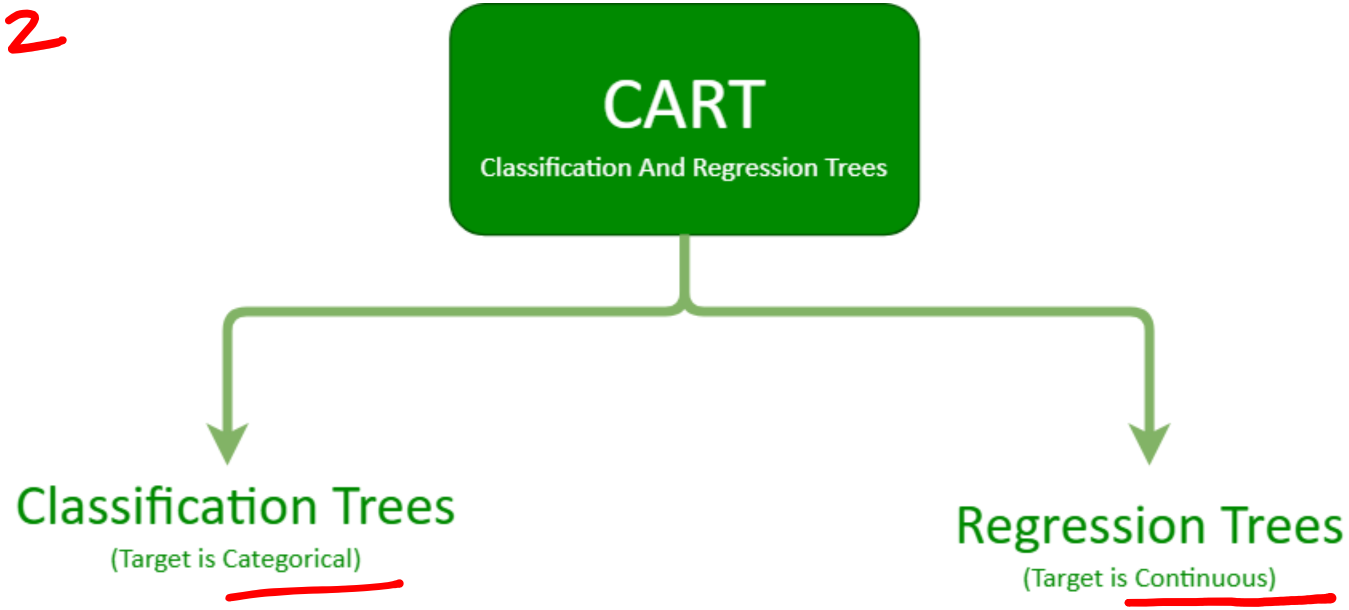
1 • ID3 – Also known as Iterative Dichotomiser 3.

2 • CART – Also known as **Classification** and **Regression** Trees.

1

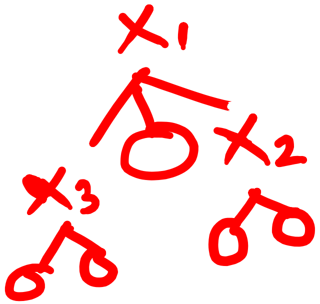
ID3

2



Agenda

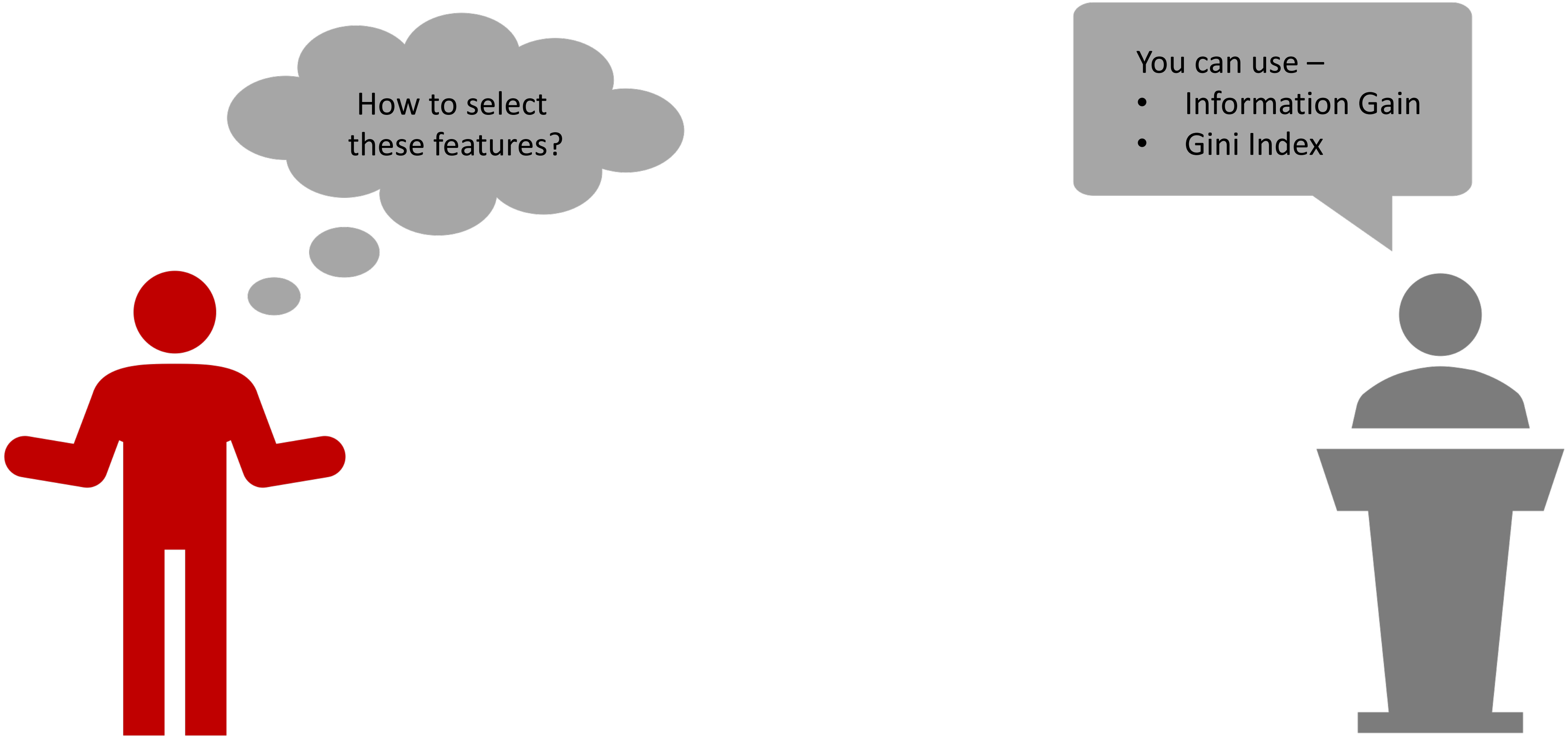
- Terminology Related to Trees
- Decision Tree
- Decision Tree Algorithms $X_1 + X_2 + X_3 + X_4$ Ny
- Column
- **Attribute Selection Measures**



- ID3 Algorithm
- Entropy & Information Gain
- Steps to Estimate Entropy & Information Gain
- CART Algorithm
- Gini Index
- Steps to estimate Gini Index
- CART – Regression Example
- Issues with Decision Trees
- Tree Pruning
- Decision Tree Applications

INSAID

Attribute Selection Measures



Agenda

- Terminology Related to Trees
- Decision Tree
- Decision Tree Algorithms
- Attribute Selection Measures
- **ID3 Algorithm**
- Entropy & Information Gain
- Steps to Estimate Entropy & Information Gain
- CART Algorithm
- Gini Index
- Steps to estimate Gini Index
- CART – Regression Example
- Issues with Decision Trees
- Tree Pruning
- Decision Tree Applications

ID3 Algorithm

- ID3 uses **Information Gain** and **Entropy** as it attribution selection measure.

INSAID

Agenda

- Terminology Related to Trees
- Decision Tree
- Decision Tree Algorithms
- Attribute Selection Measures
- ID3 Algorithm
- **Entropy & Information Gain**
- Steps to Estimate Entropy & Information Gain
- CART Algorithm
- Gini Index
- Steps to estimate Gini Index
- CART – Regression Example
- Issues with Decision Trees
- Tree Pruning
- Decision Tree Applications

Entropy & Information Gain

- Entropy is the measure of **randomness** or **impurity in** the **data** set.
- Entropy **uses** the **concept of homogeneity**.

Things to Remember:

- **If** samples are completely **homogeneous**, **then** the **entropy** of that attribute **will be zero**.
- **If samples are equally divided**, **then entropy will be one**.
- So out of the heterogeneous options **we** need to **select** the **ones having maximum homogeneity**.

Purity vs Impurity in Data

Impurity = 0

Homogeneous Data

Impurity ≠ 0

Heterogeneous Data

Decision Tree

Home Insert Draw Design Transitions Animations Slide Show Review View

Paste Cut Copy Format New Slide Layout Reset Section

You are screen sharing Stop Share

Convert to SmartArt Picture Shapes Text Box Arrange Quick Styles Shape Fill Shape Outline

How to decide splitting criteria in case of numerical columns?

14 Advantages Disadvantages
Intuitive and easy to understand
Minimal data preparation is required
The cost of using the tree for inference is **logarithmic** in the number of data points used to train the tree
Overfitting
Prone to errors for imbalanced datasets

15 CART - Classification and Regression Trees
The logic of decision trees can also be applied to regression problems, hence the name CART

16 A fun example
https://en.akinator.com/

17 Impurity

18 Entropy
- Entropy is amount of information is needed to accurately describe the some sample.
- So if sample is homogeneous, means all the element are similar than Entropy is 0, else if sample is equally divided than entropy is maximum 1.

19 Gini index / Gini impurity
- Gini index is measure of inequality in sample.
- It has value between 0 and 1. Gini index of value 0 means sample are perfectly homogeneous and all element are similar, whereas, Gini index of value 1 means maximal inequality among elements.
- It is sum of the square of the probabilities of each class.

Impurity [0 - 1]

Pure

Max Impure

Less Imp

More Imp.

- Click to a

Imp = 0

Imp ≠ 0

Imp ≠ 0

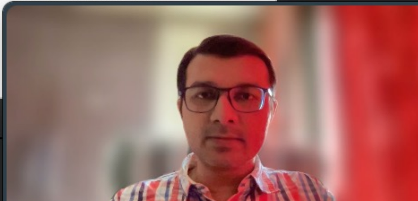
R B B B
(R+3B)

Low Medium High

Imp is Max

Trainer: Dr. Darshan Ingle

Click to add notes



Agenda

- Terminology Related to Trees
- Decision Tree
- Decision Tree Algorithms
- Attribute Selection Measures
- ID3 Algorithm
- Entropy & Information Gain
- **Steps to Estimate Entropy & Information Gain**
- CART Algorithm
- Gini Index
- Steps to estimate Gini Index
- CART – Regression Example
- Issues with Decision Trees
- Tree Pruning
- Decision Tree Applications

Math begins. Focus: Understanding formulaes & how are the values put in that formulae.

INSAID

Steps to estimate Entropy & Information Gain

$E = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$

y

- Calculate the expected information needed to classify a tuple in data set (D) is given by – $= -\sum_{i=1}^{2} p_i \cdot \log_2 p_i$

n=14 → yes 9, No 5

y → Unique values: {yes, No} (2)

$$\boxed{\text{Entropy(D)} = -\sum_{i=1}^{m} p_i \log_2(p_i)}$$

$= -\left[p_y \cdot \log_2 p_y + p_N \cdot \log_2 p_N\right]$

- We will check how many tuples are **yes** and **no in target variable** in the below data set.

$= -p_y \cdot \log_2 p_y - p_N \cdot \log_2 p_N$

$= -\frac{9}{14} \cdot \log_2 \frac{9}{14} - \frac{5}{14} \cdot \log_2 \frac{5}{14}$

	A (X_1)	B (X_2)	C (X_3)	D (X_4)	E (y)
1	age	income	student	credit_rating	buys_computer
2	youth	high	no	fair	no
3	youth	high	no	excellent	no
4	middle-aged	high	no	fair	yes
5	senior	medium	no	fair	yes
6	senior	low	yes	fair	yes
7	senior	low	yes	excellent	no
8	middle-aged	low	yes	excellent	yes
9	youth	medium	no	fair	no
10	youth	low	yes	fair	yes
11	senior	medium	yes	fair	yes
12	youth	medium	yes	excellent	yes
13	middle-aged	medium	no	excellent	yes
14	middle-aged	high	yes	fair	yes
15	senior	medium	no	excellent	no

Entropy(D) = - p(yes) x Log_2(p(yes)) - p(no) x Log_2(p(no))

Entropy(D) = - 9/14 x Log_2(9/14) - 5/14 x Log_2(5/14)

Entropy(D) = 0.94

Steps to estimate Entropy & Information Gain Cont.

n=14

Youth : Unique values {Yes, No}

X1
AGE:

- How much more information would we still need (after the partitioning) to arrive at an exact classification?

This amount is measured by –

	Yes	No	Total
Youth	2	3	5
MA	4	0	4
Senior	3	2	5

$$\text{Entropy}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Entropy}(D_j)$$

$$E_{youth} = -\sum_{i=1}^{2} p_i \cdot \log_2 p_i$$

$$= -[P_Y \cdot \log_2 P_Y + P_N \cdot \log_2 P_N]$$

- We will check how many tuples are yes and no in target variable along with that particular predictor variable.

$$= -P_Y \cdot \log_2 P_Y - P_N \cdot \log_2 P_N$$

$$= -\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{3}{5} \cdot \log_2 \frac{3}{5}$$

X1   y

	A	B	C	D	E
1	age	income	student	credit_rating	buys_computer
2	youth	high	no	fair	no
3	youth	high	no	excellent	no
4	middle-aged	high	no	fair	yes
5	senior	medium	no	fair	yes
6	senior	low	yes	fair	yes
7	senior	low	yes	excellent	no
8	middle-aged	low	yes	excellent	yes
9	youth	medium	no	fair	no
10	youth	low	yes	fair	yes
11	senior	medium	yes	fair	yes
12	youth	medium	yes	excellent	yes
13	middle-aged	medium	no	excellent	yes
14	middle-aged	high	yes	fair	yes
15	senior	medium	no	excellent	no

Ex1

$$\text{Entropy}_{age}(D) = \frac{|D_{youth}|}{|D|} \times \text{Entropy}(D_{youth}) + \frac{|D_{middle\text{-}aged}|}{|D|} \times \text{Entropy}(D_{middle\text{-}aged}) + \frac{|D_{senior}|}{|D|} \times \text{Entropy}(D_{senior})$$

$$\text{Entropy}_{age}(D) = 5/14 \times [-2/5 \, \text{Log}_2(2/5) - 3/5 \, \text{Log}_2(3/5)] + 4/14 \times [-4/4 \, \text{Log}_2(4/4)] + 5/14 \times [-3/5 \, \text{Log}_2(3/5) - 2/5 \, \text{Log}_2(2/5)]$$

$$\text{Entropy}_{age}(D) = 0.629$$

INSAID

Steps to estimate Entropy & Information Gain Cont.

- Final step is to calculate Information Gain –

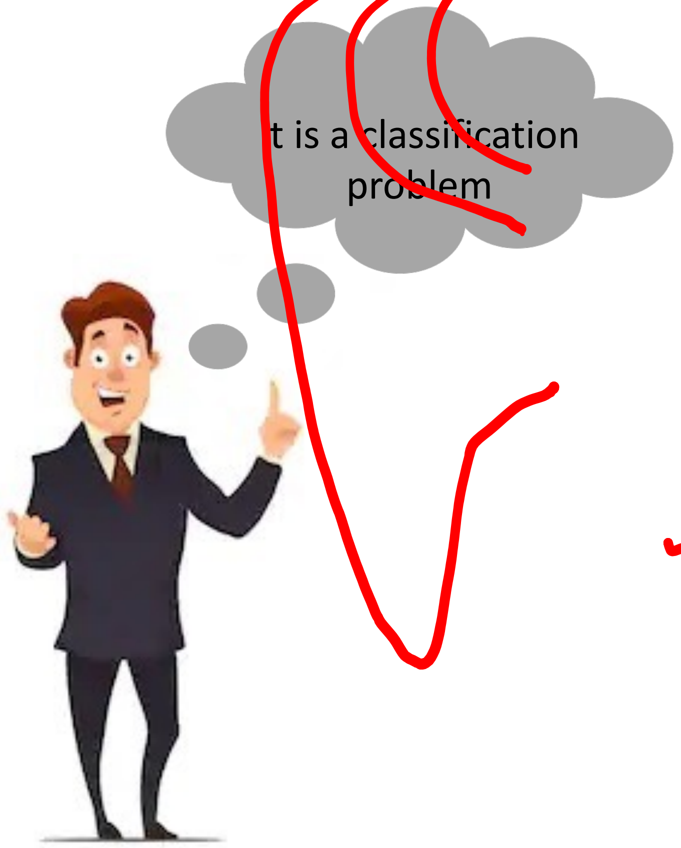
$$\text{Information Gain}(A) = \text{Entropy}(D) - \text{Entropy}_A(D)$$

$$\text{Information Gain}(age) = 0.940 - 0.629 = 0.248$$

Example on Buys_Computer Data set

X_2 X_3 X_4

	A	B	C	D	E
1	age	income	student	credit_rating	buys_computer
2	youth	high	no	fair	no
3	youth	high	no	excellent	no
4	middle-aged	high	no	fair	yes
5	senior	medium	no	fair	yes
6	senior	low	yes	fair	yes
7	senior	low	yes	excellent	no
8	middle-aged	low	yes	excellent	yes
9	youth	medium	no	fair	no
10	youth	low	yes	fair	yes
11	senior	medium	yes	fair	yes
12	youth	medium	yes	excellent	yes
13	middle-aged	medium	no	excellent	yes
14	middle-aged	high	yes	fair	yes
15	senior	medium	no	excellent	no



- **Predictors** are: age, income, student, credit_rating. **Target variable** is buys_computer.

Calculation Work

E_y

$$\text{Entropy}(D) = -\,9/14 \times \log_2(9/14) - 5/14 \times \log_2(5/14) = 0.94$$

E_{x_1}

$$\text{Entropy}_{age}(D) = \frac{|D_{youth}|}{|D|} \times \text{Entropy}(D_{youth}) + \frac{|D_{middle}|}{|D|} \times \text{Entropy}(D_{middle}) + \frac{|D_{senior}|}{|D|} \times \text{Entropy}(D_{senior})$$

$$\text{Entropy}_{age}(D) = \frac{5}{14}\left[-\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5)\right] + \frac{4}{14}\left[-\frac{4}{4}\log_2(4/4)\right] + \frac{5}{14}\left[-\frac{3}{5}\log_2(3/5) - \frac{2}{5}\log_2(2/5)\right]$$

$$\text{Entropy}_{age}(D) = 0.629$$

E_{x_2}

$$\text{Entropy}_{income}(D) = \frac{|D_{high}|}{|D|} \times \text{Entropy}(D_{high}) + \frac{|D_{medium}|}{|D|} \times \text{Entropy}(D_{medium}) + \frac{|D_{low}|}{|D|} \times \text{Entropy}(D_{low})$$

$$\text{Entropy}_{income}(D) = \frac{4}{14}\left[-\frac{2}{4}\log_2(2/4) - \frac{2}{4}\log_2(2/4)\right] + \frac{6}{14}\left[-\frac{4}{6}\log_2(4/6) - \frac{2}{6}\log_2(2/6)\right] + \frac{4}{14}\left[-\frac{3}{4}\log_2(3/4) - \frac{1}{4}\log_2(1/4)\right]$$

$$\text{Entropy}_{income}(D) = 0.908$$

E_{x_3}

$$\text{Entropy}_{student}(D) = \frac{|D_{yes}|}{|D|} \times \text{Entropy}(D_{yes}) + \frac{|D_{no}|}{|D|} \times \text{Entropy}(D_{no})$$

$$\text{Entropy}_{student}(D) = 0.786$$

E_{x_4}

$$\text{Entropy}_{credit\_rating}(D) = \frac{|D_{fair}|}{|D|} \times \text{Entropy}(D_{fair}) + \frac{|D_{excellent}|}{|D|} \times \text{Entropy}(D_{excellent})$$

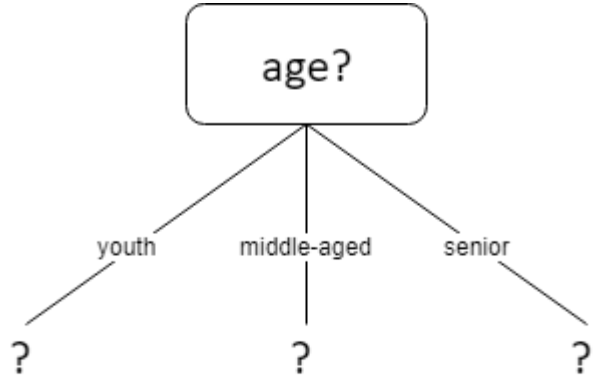
$$\text{Entropy}_{credit\_rating}(D) = 0.89$$

INSAID

Tabular View

- The **age** attribute is **giving maximum information** gain. So the **root** node **will be age**.

	A	B	C
1		**Entropy**	**Information Gain**
2	Data	0.94	0
3	age (root)	0.629	0.248 (Max)
4	income	0.908	0.032
5	student	0.786	0.154
6	credit_rating	0.89	0.05



- **But how should I choose next attribute?**
  - Repeat the step we've done so far on the subset of data.

Next View of Representation

X_1

age?

youth

X_1	X_2	X_3	y
income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

middle_aged

			y
income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

Homogenous (Impurity=0)

senior

X_1	X_2	X_3	y
income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

Tabular view on subset data

Solution for **youth data**

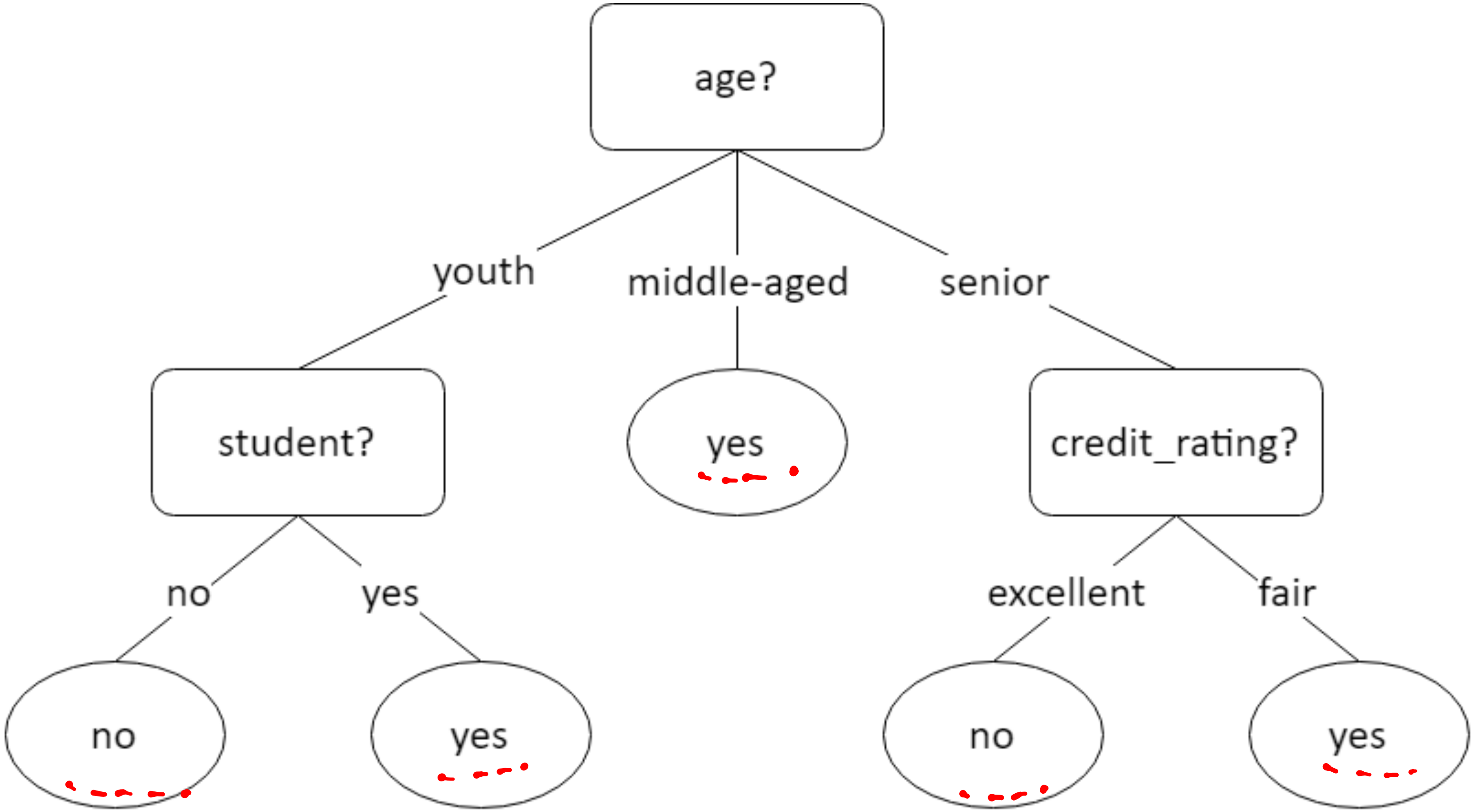
12		Entropy	Information Gain
13	Data y	0.97	0
14	income x_1	0.399	= 0.571
15	student x_2	0	= 0.97 (max)
16	credit_rating x_3	0.948	= 0.022

Solution for **senior data**

21		Entropy	Information Gain
22	Data y	0.968	0
23	income x_1	0.95	= 0.018
24	student x_2	0.95	= 0.018
25	credit_rating x_3	0	= 0.968 (max)

- **Note**: Entropy for middle-aged data is zero.

Decision Tree Complete View



INSAID

Thanks for watching