# Module : Decision Tree

## Instructor: Dr. Darshan Ingle

# 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

# Steps to estimate Entropy & Information Gain

- Calculate the expected information needed to classify a tuple in data set (D) is given by –

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

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

| | 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 |

Entropy(D) = - p(yes) x $\text{Log}_2$(p(yes)) - p(no) x $\text{Log}_2$(p(no))

Entropy(D) = - 9/14 x $\text{Log}_2$(9/14) - 5/14 x $\text{Log}_2$(5/14)

Entropy(D) = 0.94

# Steps to estimate Entropy & Information Gain Cont.

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

  This amount is measured by –

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

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

| | 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 |

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

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

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

# Steps to estimate Entropy & Information Gain Cont.

- Final step is to calculate Information Gain –

Information Gain(A) = Entropy(D) - $Entropy_A(D)$

Information Gain(age) = 0.940 – 0.629 = 0.248

# Example on Buys_Computer Data set

| | 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

$$\text{Entropy}(D) = -\frac{9}{14} \times \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \times \log_2\left(\frac{5}{14}\right) = 0.94$$

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

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

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

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

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

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

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

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

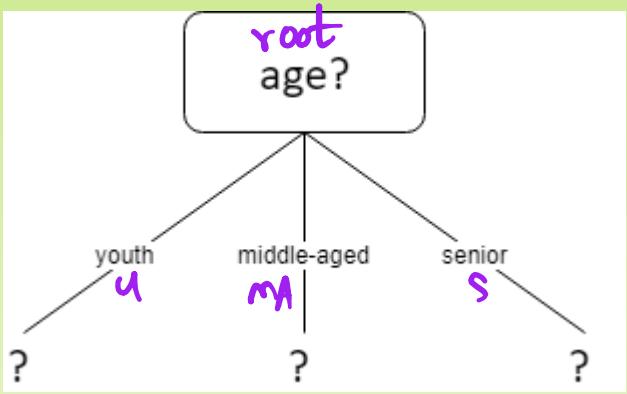
$$\text{Entropy}_{credit\_rating}(D) = \frac{|D_{fair}|}{|D|} \times \text{Entropy}(D_{fair}) + \frac{|D_{excellent}|}{|D|} \times \text{Entropy}(D_{excellent})$$

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

# Tabular View

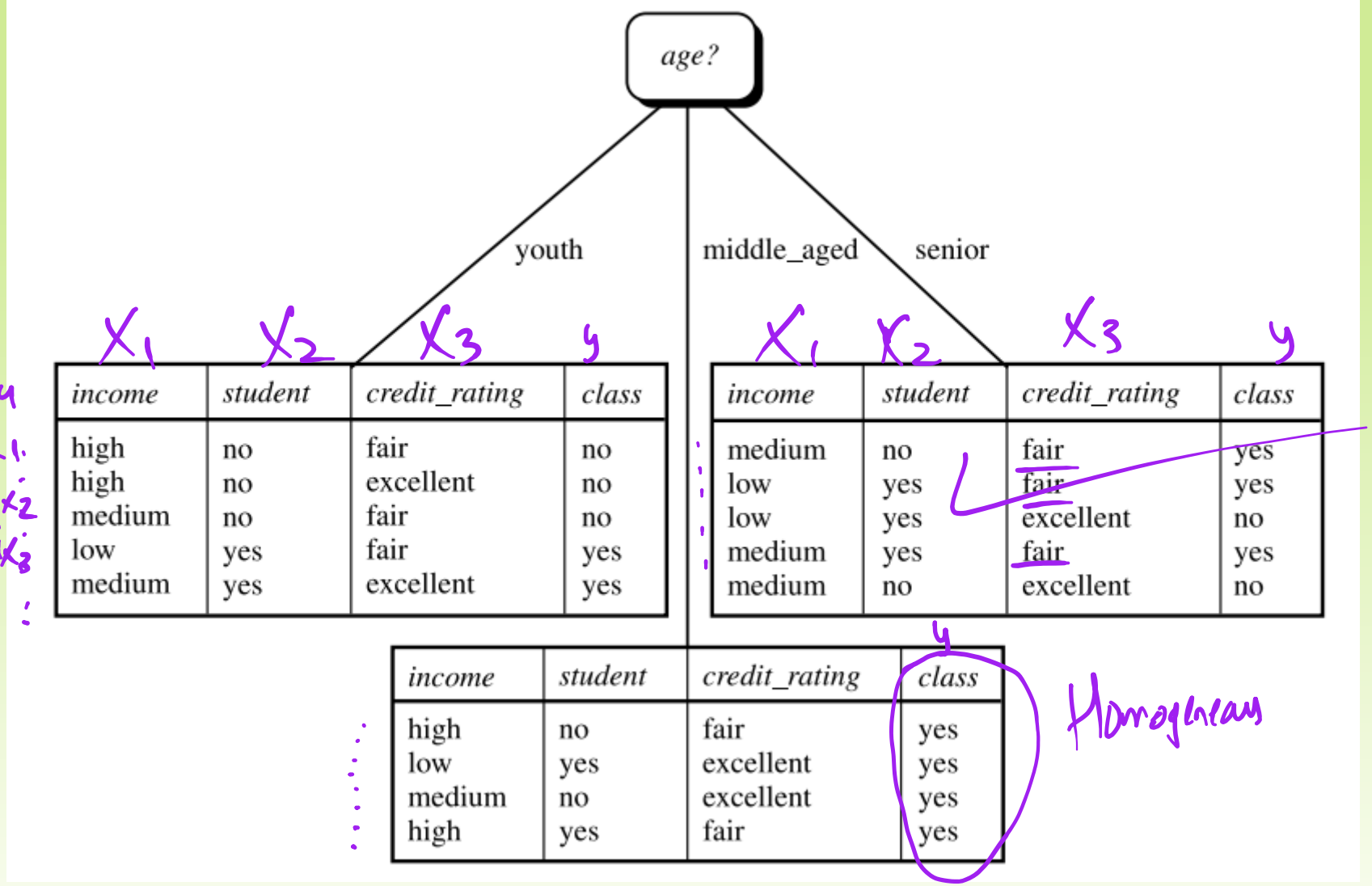
- The **age** attribute is **giving maximum information** gain. So the **root** node **will be age**.

| | A | B | C |
|---|---|---|---|
| 1 | | Entropy | Information Gain |
| 2 | Data | 0.94 | 0 |
| 3 | age | 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



# Tabular view on subset data

Solution for **youth data**

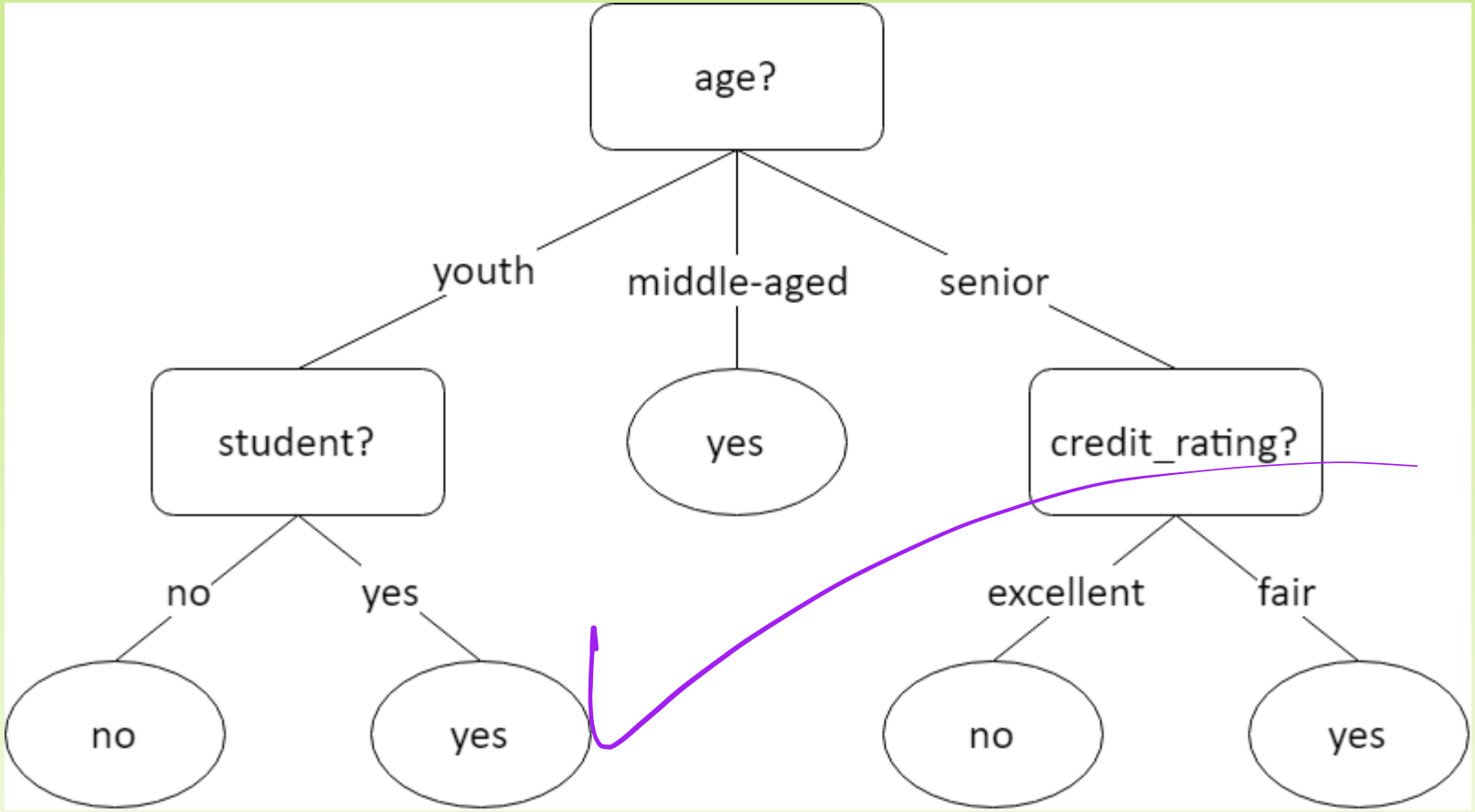
| 12 | | Entropy | Information Gain |
|---|---|---|---|
| 13 | Data | 0.97 | 0 |
| 14 | income | 0.399 | 0.571 |
| 15 | student | 0 | 0.97 |
| 16 | credit_rating | 0.948 | 0.022 |

Solution for **senior data**

| 21 | | Entropy | Information Gain |
|---|---|---|---|
| 22 | Data | 0.968 | 0 |
| 23 | income | 0.95 | 0.018 |
| 24 | student | 0.95 | 0.018 |
| 25 | credit_rating | 0 | 0.968 |

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

# Decision Tree Complete View

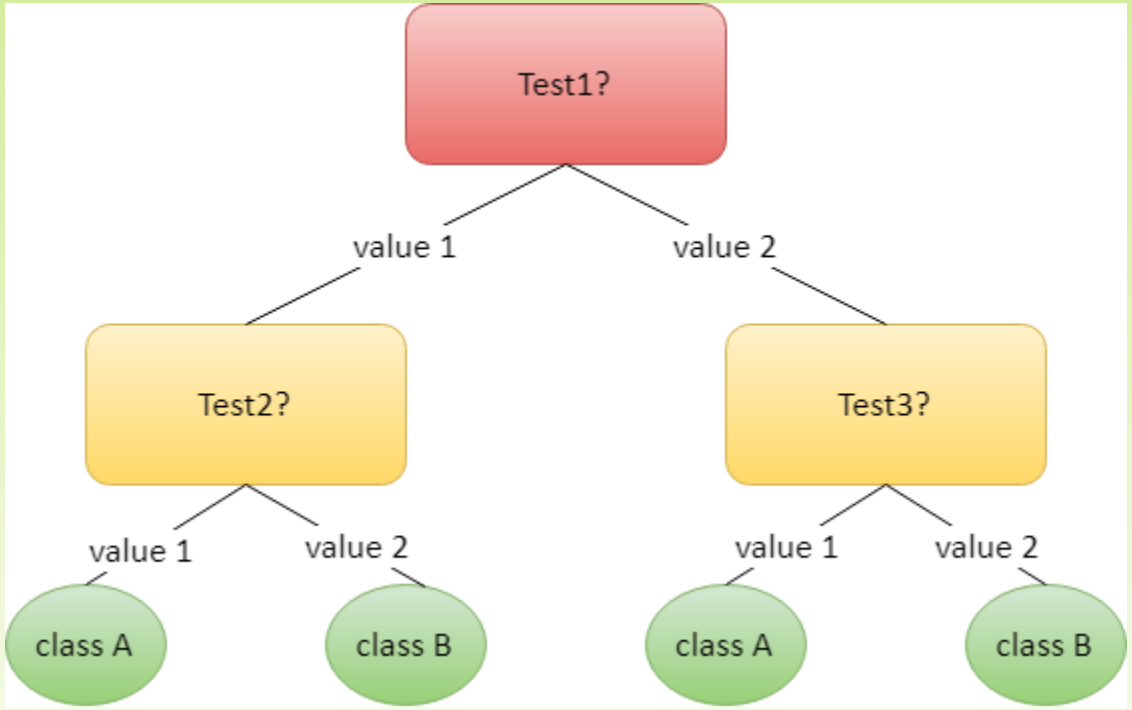



# Agenda

- 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

# CART Algorithm

- CART is known as **Classification And Regression Trees**. It has binary representation of data.
- It uses **Gini index** to find best split on the attributes.



# 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

# Gini Index

- Gini Index is an attribution selection measure to **find best split** on attributes.
- The concept is same as Information Gain except it helps to **find best split** for binary representation.
- It helps to **find maximum reduction in impurity** of data.

**Remember:** If samples are complete **homogeneous**, **then** their **Gini Index** will be **zero**.

# Steps to estimate Gini Index

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

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

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

| | 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 |

$$\text{Gini}(D) = 1 - p(\text{yes})^2 - p(\text{no})^2$$

$$\text{Gini}(D) = 1 - (9/14)^2 - (5/14)^2$$

$$\text{Gini}(D) = 0.459$$

Age: {Y, MA, S} → One vs Rest

Y vs {MA, S} — (a)
MA vs {Y, S} — (b)
S vs {Y, MA} — (c) #

3 subsets

# Steps to estimate Gini Index Cont.

- When considering a binary split, we compute a weighted sum of the impurity of each resulting partition,

  the Gini index of D given that partitioning is

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

$Gini = 1 - \sum p_i^2 = 1 - [P_{S_{Yes}}^2 + P_{S_{No}}^2] = 1 - [(\frac{3}{5})^2 + (\frac{2}{5})^2]$

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

| | 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 |

(This value is computed only for subset of split)

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

$$Gini_{age}(D) = 9/14 \times [1 - (6/9)^2 - (3/9)^2] + 5/14 \times [1 - (3/5)^2 - (2/5)^2]$$

$$Gini_{age}(D) = 0.455$$

| | Yes | No | Total |
|---|---|---|---|
| Y | 2 | 3 | 5 |
| MA | 4 | 0 | 4 |
| S | 3 | 2 | 5 |

$G = 1 - \sum p_i^2 = 1 - [P_{Y,MA}^2 + P_{Y,MA}^2]$ (Yes, No)

$n = 14$

$= 1 - [(\frac{6}{9})^2 + (\frac{3}{9})^2]$ (Yes)

# Steps to estimate Gini Index Cont.

- The reduction in impurity that would be incurred by a binary split on a discrete or continuous-valued attribute A is

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

$$\Delta Gini(age) = 0.459 - 0.455 = 0.004$$

(This value is computed only for subset of split)

# Example on Buys_Computer Data set

| | 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

$Gini(D) = 1 - (9/14)^2 - (5/14)^2 = 0.459$

A $Gini_{age} \in$ {{youth, middle-aged} or {senior}, {youth, senior} or {middle-aged}, {middle-aged, senior} or {youth}}

c b a

$Gini_{age} \in$ {youth, middle-aged} or {senior}

$Gini_{age} = 9/14 \times Gini(D1) + 5/14 \times Gini(D2)$

$Gini_{age} = 9/14 \times [1 - (6/9)^2 - (3/9)^2] + 5/14 \times [1 - (3/5)^2 - (2/5)^2]$

$Gini_{age} = 0.455$

$\Delta Gini(age) = 0.459 - 0.455 = 0.004$

$Gini_{age} \in$ {youth, senior} or {middle-aged}

$Gini_{age} = 10/14 \times Gini(D1) + 4/14 \times Gini(D2)$

$Gini_{age} = 10/14 \times [1 - (5/10)^2 - (5/10)^2] + 4/14 \times [1 - (4/4)^2]$

$Gini_{age} = 0.357$

(Optimal for binary split)

$\Delta Gini(age) = 0.459 - 0.357 = 0.102$

$Gini_{age} \in$ {middle-aged, senior} or {youth}

$Gini_{age} = 9/14 \times Gini(D1) + 5/14 \times Gini(D2)$

$Gini_{age} = 9/14 \times [1 - (7/9)^2 - (2/9)^2] + 5/14 \times [1 - (2/5)^2 - (3/5)^2]$

$Gini_{age} = 0.393$

$\Delta Gini(age) = 0.459 - 0.393 = 0.066$

# Calculation Work Cont.

$X_2$

B $Gini_{income} \in$ {{low, medium} or {high}, {low, high} or {medium}, {medium, high} or {low}}

(c) (b) (a)

(c)

$Gini_{income} \in$ {low, medium} or {high}

$Gini_{income} = 10/14 \times Gini(D1) + 4/14 \times Gini(D2)$

$Gini_{income} = 10/14 \times [1 - (7/10)^2 - (3/10)^2] + 4/14 \times [1 - (2/4)^2 - (2/4)^2]$

$Gini_{income} = 0.443$

(Optimal for binary split)

$\Delta Gini(income) = 0.459 - 0.443 = 0.016$

(b)

$Gini_{income} \in$ {low, high} or {medium}

$Gini_{income} = 8/14 \times Gini(D1) + 6/14 \times Gini(D2)$

$Gini_{income} = 8/14 \times [1 - (5/8)^2 - (3/8)^2] + 6/14 \times [1 - (4/6)^2 - (2/6)^2]$

$Gini_{income} = 0.458$

$\Delta Gini(income) = 0.459 - 0.458 = 0.001$

(a)

$Gini_{income} \in$ {medium, high} or {low}

$Gini_{income} = 8/14 \times Gini(D1) + 6/14 \times Gini(D2)$

$Gini_{income} = 10/14 \times [1 - (5/10)^2 - (5/10)^2] + 4/14 \times [1 - (3/4)^2 - (1/4)^2]$

$Gini_{income} = 0.450$

$\Delta Gini(income) = 0.459 - 0.450 = 0.009$

# Calculation Work Cont.

$X_3$

C $Gini_{student} \in \{yes\}$ or $\{no\}$

$Gini_{student} = 7/14 \times Gini(D1) + 7/14 \times Gini(D2)$

$Gini_{student} = 7/14 \times [1 - (6/7)^2 - (1/7)^2] + 7/14 \times [1 - (3/7)^2 - (4/7)^2]$

$Gini_{student} = 0.368$ $\Delta Gini(student) = 0.459 - 0.368 = 0.091$

$X_4$

D $Gini_{credit\_rating} \in \{excellent\}$ or $\{fair\}$

$Gini_{credit\_rating} = 8/14 \times Gini(D1) + 6/14 \times Gini(D2)$

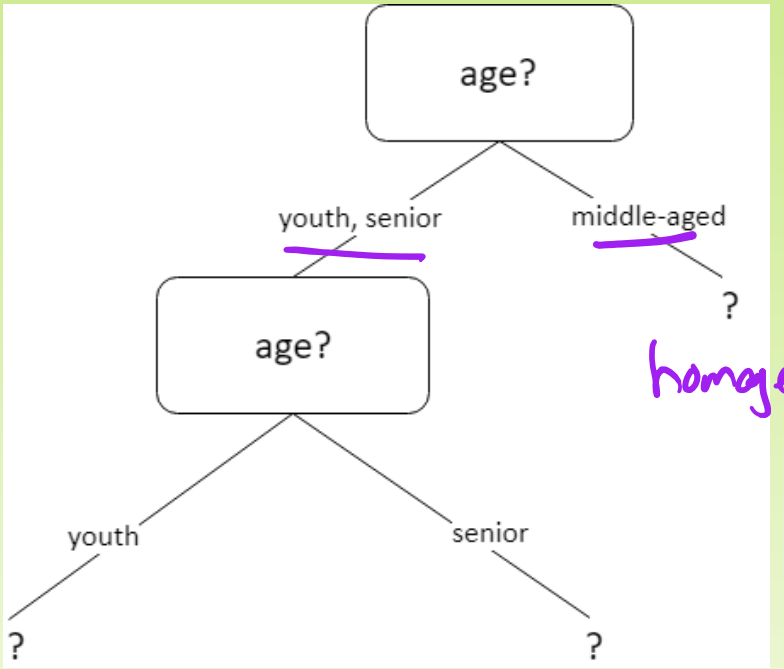
$Gini_{credit\_rating} = 8/14 \times [1 - (6/8)^2 - (2/8)^2] + 6/14 \times [1 - (3/6)^2 - (3/6)^2]$

$Gini_{credit\_rating} = 0.428$ $\Delta Gini(credit\_rating) = 0.459 - 0.428 = 0.031$

# Tabular Representation

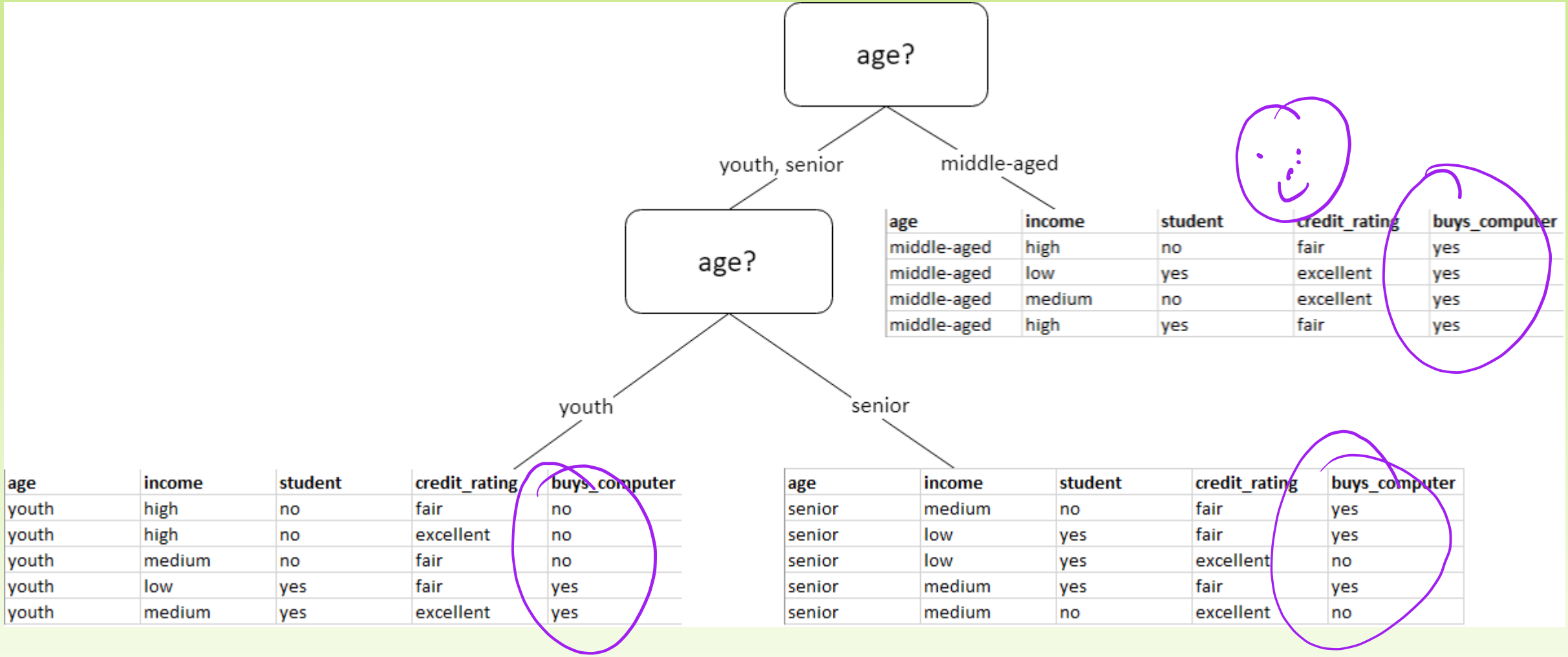
- The **age** attribute is **giving Reduction in Impurity**. So the **root** node **will be age**.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | **Gini Index** | **Reduction in Impurity** | |
| 2 | Data | 0.459 | 0 | |
| 3 | age | 0.455 | 0.004 | |
| 4 | | 0.357 | 0.102 | (Highest Reduction) |
| 5 | | 0.393 | 0.066 | |
| 6 | income | 0.443 | 0.016 | |
| 7 | | 0.458 | 0.001 | |
| 8 | | 0.45 | 0.009 | |
| 9 | student | 0.091 | 0.091 | |
| 10 | credit_rating | 0.428 | 0.031 | |




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

# Next View of Representation




| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| middle-aged | high | no | fair | yes |
| middle-aged | low | yes | excellent | yes |
| middle-aged | medium | no | excellent | yes |
| middle-aged | high | yes | fair | yes |

| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| youth | high | no | fair | no |
| youth | high | no | excellent | no |
| youth | medium | no | fair | no |
| youth | low | yes | fair | yes |
| youth | medium | yes | excellent | yes |

| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| senior | medium | no | fair | yes |
| senior | low | yes | fair | yes |
| senior | low | yes | excellent | no |
| senior | medium | yes | fair | yes |
| senior | medium | no | excellent | no |

# Tabular view on subset data

Solution for **youth** data

| | Gini Index | Reduction in Impurtiy |
|---|---|---|
| Data | 0.48 | 0 |
| income | 0.266 | 0.214 |
| | 0.466 | 0.014 |
| | 0.3 | 0.18 |
| student | 0 | 0.48 |
| credit_rating | 0.466 | 0.014 |

Solution for **senior** data

| | Gini Index | Reduction in Impurity |
|---|---|---|
| Data | 0.48 | 0 |
| income | 0.466 | 0.014 |
| student | 0.466 | 0.014 |
| credit_rating | 0 | 0.48 |

- **Note**: Gini Index for middle-aged data is zero.

# CART Complete View