جـامعة الجـلالة
GALALA UNIVERSITY

Artificial Intelligence Science Program

Chapter 4: Learning from Examples

Gain Ratio for Attribute Selection (C4.5)

- Information gain measure is biased towards attributes with a large number of values.
- C4.5 (a successor of ID3) uses gain ratio to overcome the problem (normalization to information gain).

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

  - GainRatio(A) = Gain(A)/SplitInfo(A)
- Ex.

$$SplitInfo_{income}(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right) = 1.557$$

  - gain_ratio(income) = 0.029/1.557 = 0.019
- The attribute with the maximum gain ratio is selected as the splitting attribute

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

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

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

$$Gain(A) = Info(D) - Info_A(D)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

- Calculate ***Entropy*** of Class attribute:

buys_computer	
yes	no
9	5

$$Info(D) = I(9,5) = -\frac{9}{14}log_2\left(\frac{9}{14}\right) - \frac{5}{14}log_2\left(\frac{5}{14}\right) = \underline{\mathbf{0.9403}}$$

age	income	student	credit_rating	buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

- Calculate ***Gain Ratio*** of all other attributes:

$$-\frac{2}{5}log_2\left(\frac{2}{5}\right) - \frac{3}{5}log_2\left(\frac{3}{5}\right)$$

		Class		
		yes	no	
age	youth	2	3 ✓	5 ✓
	middle_aged	4 ✓	0	4 ✓
	senior	3 ✓	2 ✓	5 ✓
				14

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2)$$
$$= \frac{5}{14} * 0.971 + \frac{4}{14} * 0 + \frac{5}{14} * 0.971 = 0.3467 + 0 + 0.3467 = \mathbf{0.6934}$$

$$Gain(age) = Info(D) - Infoage_{(D)} = 0.9403 - 0.6934 = \mathbf{0.2469}$$

$$SplitInf0_{age}(D) = -\frac{5}{14} * \log2\left(\frac{5}{14}\right) - \frac{4}{14} * \log_2\left(\frac{4}{14}\right) - \frac{5}{14} * \log2\left(\frac{5}{14}\right) = \underline{\mathbf{1.5774}}$$

$$\mathbf{GainRatio(age)} = \frac{Gain(A)}{SplitInfo(A)} = \frac{0.246}{1.5774} = \underline{\mathbf{0.1559}}$$

		Class		
		yes	no	
income	low	3 ✓	1 ✓	4
	medium	4	2	6 ✓
	high	2 ✓	2 ✓	4 ✓
				14

$$Info_{income}(D) = \frac{4}{14}I(3,1) + \frac{6}{14}I(4,2) + \frac{4}{14}I(2,2)$$
$$= \frac{4}{14} * 0.8113 + \frac{6}{14} * 0.9183 + \frac{4}{14} * 1 = 0.2318 + 0.3935 + 0.2857 = \mathbf{0.911}$$

$$Gain(income) = 0.9403 - 0.911 = \mathbf{0.0293}$$

$$SplitInf0_{income}(D) = -\frac{4}{14} * \log2\left(\frac{4}{14}\right) - \frac{6}{14} * \log_2\left(\frac{6}{14}\right) - \frac{4}{14} * \log2\left(\frac{4}{14}\right) = \underline{\mathbf{1.5566}}$$

$$\mathbf{GainRatio(income)} = \frac{0.0293}{1.5566} = \mathbf{0.0188}$$

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

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

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

$$Gain(A) = Info(D) - Info_A(D)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

- Calculate ***Entropy*** of Class attribute:

buys_computer	
yes	no
9	5

$$Info(D) = I(9,8) = -\frac{9}{14}log_2\left(\frac{9}{14}\right) - \frac{5}{14}log_2\left(\frac{5}{14}\right) = \mathbf{0.9403}$$

age	income	student	credit_rating	buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

- Calculate ***Gain Ratio*** of all other attributes:

		Class		
		yes	no	
student	yes	6	1	7
	no	3	4	7
				14

$$Info_{student}(D) = \frac{7}{14}I(6,1) + \frac{7}{14}I(3,4)$$
$$= \frac{7}{14} * 0.5917 + \frac{7}{14} * 0.9852 = 0.2958 + 0.4926 = \mathbf{0.7884}$$

$$Gain(student) = 0.9403 - 0.7884 = \mathbf{0.1519}$$

$$SplitInf0_{student}(D) = -\frac{7}{14} * \log2\left(\frac{7}{14}\right) - \frac{7}{14} * \log_2\left(\frac{7}{14}\right) = \mathbf{1}$$

$$\mathbf{GainRatio(student)} = \frac{0.1519}{1} = \underline{\mathbf{0.1519}}$$

		Class		
		yes	no	
credit_rating	fair	6	2	8
	excellent	3	3	6
				14

$$Info_{credit\ \_\ rating}(D) = \frac{8}{14}I(6,2) + \frac{6}{14}I(3,3)$$
$$= \frac{8}{14} * 0.8113 + \frac{6}{14} * 1 \ = 0.4636 + 0.4286 = \mathbf{0.8922}$$

$$Gain(credit - rating) = 0.9403 - 0.8922 = \mathbf{0.0481}$$

$$SplitInf0_{credit-rating}(D) = -\frac{8}{14} * \log2\left(\frac{8}{14}\right) - \frac{6}{14} * \log_2\left(\frac{6}{14}\right) = \mathbf{0.9852}$$

$$\mathbf{GainRatio(credit - rating)} = \frac{0.0481}{0.9852} = \underline{\mathbf{0.0488}}$$

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

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

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

$$Gain(A) = Info(D) - Info_A(D)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

- Calculate ***Entropy*** of Class attribute:

buys_computer	
yes	no
9	5

$$Info(D) = I(9,8) = -\frac{9}{14}log_2\left(\frac{9}{14}\right) - \frac{5}{14}log_2\left(\frac{5}{14}\right) = \mathbf{0.9403}$$

- Calculate ***Gain Ratio*** of all other attributes:

		Class		
		yes	no	
student	yes	6	1	7
	no	3	4	7
				14

$$Info_{student}(D) = \frac{7}{14}I(6,1) + \frac{7}{14}I(3,4)$$
$$= \frac{7}{14} * 0.5917 + \frac{7}{14} * 0.9852 = 0.2958 + 0.4926 = \mathbf{0.7884}$$

$$Gain(student) = 0.9403 - 0.7884 = \mathbf{0.1519}$$

$$SplitInf0_{student}(D) = -\frac{7}{14} * \log2\left(\frac{7}{14}\right) - \frac{7}{14} * \log_2\left(\frac{7}{14}\right) = \mathbf{1}$$

$$\mathbf{GainRatio(student)} = \frac{0.1519}{1} = \mathbf{0.1519}$$

		Class		
		yes	no	
credit_rating	fair	6	2	8
	excellent	3	3	6
				14

$$Info_{credit\_rating}(D) = \frac{8}{14}I(6,2) + \frac{6}{14}I(3,3)$$
$$= \frac{8}{14} * 0.8113 + \frac{6}{14} * 1 = 0.4636 + 0.4286 = \mathbf{0.8922}$$

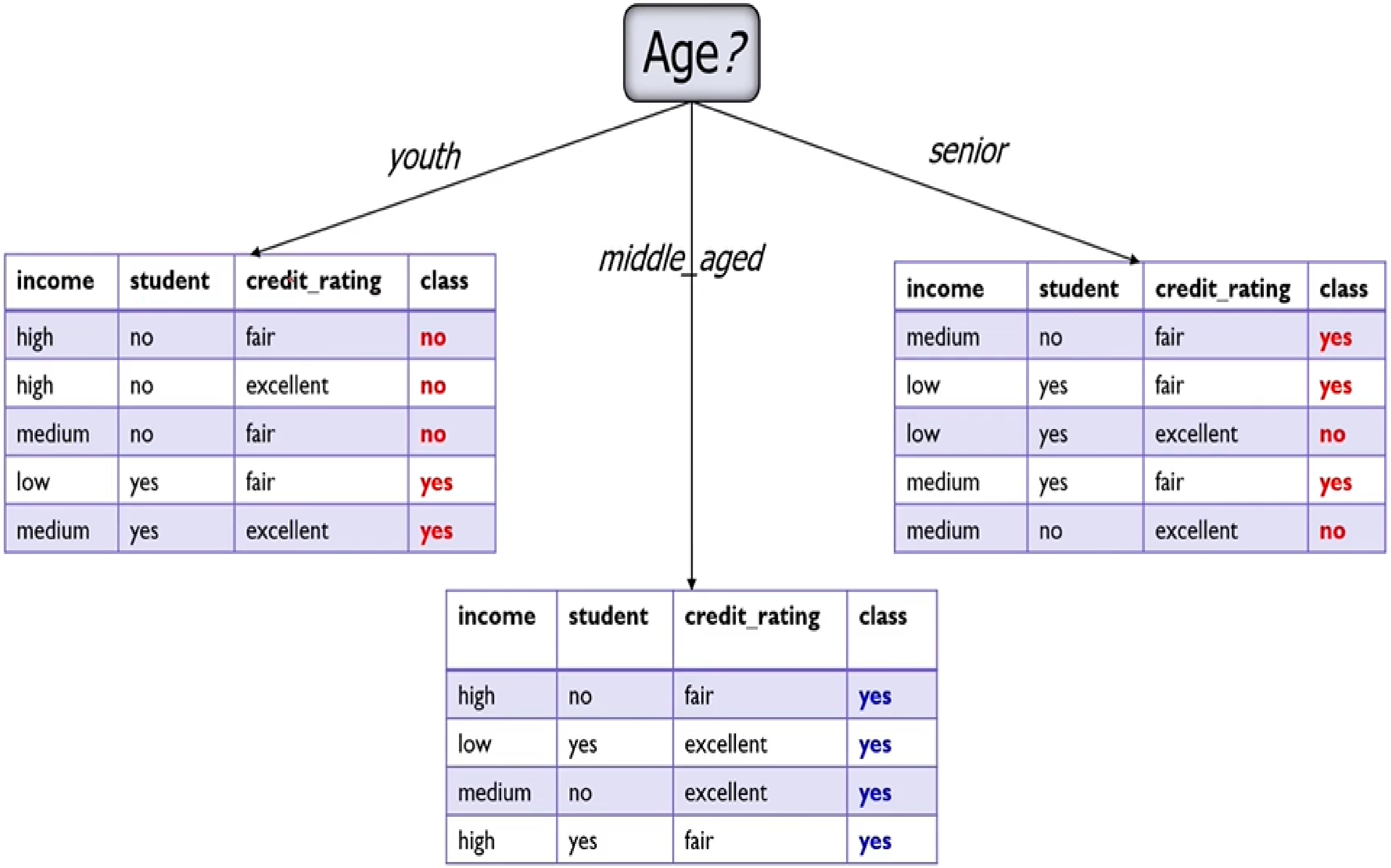
$$Gain(credit - rating) = 0.9403 - 0.8922 = \mathbf{0.0481}$$

$$SplitInf0_{credit-rating}(D) = -\frac{8}{14} * \log2\left(\frac{8}{14}\right) - \frac{6}{14} * \log_2\left(\frac{6}{14}\right) = \mathbf{0.9852}$$

$$\mathbf{GainRatio(credit - rating)} = \frac{0.0481}{0.9852} = \mathbf{0.0488}$$

age	income	student	credit_rating	buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

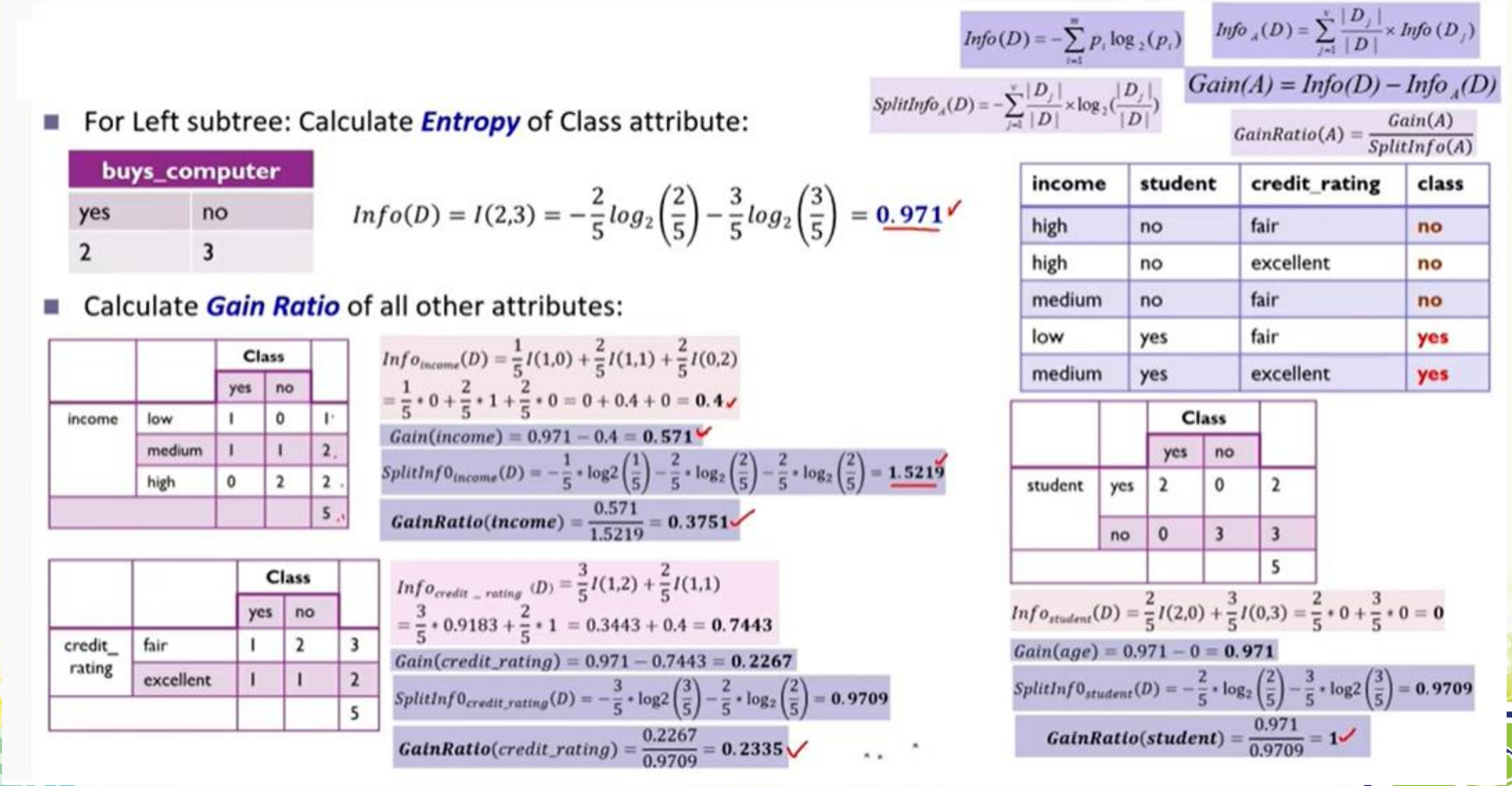
- As, the Gain Ratio of “age” is highest,
- So “**age**” is the best attribute & becomes the root node of the decision tree.



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

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

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



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

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

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

$$Gain(A) = Info(D) - Info_A(D)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

- For Left subtree: Calculate ***Entropy*** of Class attribute:

buys_computer	
yes	no
2	3

$$Info(D) = I(2{,}3) = -\frac{2}{5}log_2\left(\frac{2}{5}\right) - \frac{3}{5}log_2\left(\frac{3}{5}\right) = \underline{0.971}\checkmark$$

- Calculate ***Gain Ratio*** of all other attributes:

		Class		
		yes	no	
income	low	1	0	1
	medium	1	1	2
	high	0	2	2
				5

$$Info_{income}(D) = \frac{1}{5}I(1{,}0) + \frac{2}{5}I(1{,}1) + \frac{2}{5}I(0{,}2)$$
$$= \frac{1}{5} * 0 + \frac{2}{5} * 1 + \frac{2}{5} * 0 = 0 + 0.4 + 0 = \mathbf{0.4}\checkmark$$

$$Gain(income) = 0.971 - 0.4 = \mathbf{0.571}\checkmark$$

$$SplitInf0_{income}(D) = -\frac{1}{5} * \log2\left(\frac{1}{5}\right) - \frac{2}{5} * \log_2\left(\frac{2}{5}\right) - \frac{2}{5} * \log_2\left(\frac{2}{5}\right) = \underline{\mathbf{1.5219}}\checkmark$$

$$\mathbf{GainRatio(income)} = \frac{0.571}{1.5219} = \mathbf{0.3751}\checkmark$$

		Class		
		yes	no	
credit_ rating	fair	1	2	3
	excellent	1	1	2
				5

$$Info_{credit\_rating}(D) = \frac{3}{5}I(1{,}2) + \frac{2}{5}I(1{,}1)$$
$$= \frac{3}{5} * 0.9183 + \frac{2}{5} * 1 = 0.3443 + 0.4 = \mathbf{0.7443}$$

$$Gain(credit\_rating) = 0.971 - 0.7443 = \mathbf{0.2267}$$

$$SplitInf0_{credit\_rating}(D) = -\frac{3}{5} * \log2\left(\frac{3}{5}\right) - \frac{2}{5} * \log_2\left(\frac{2}{5}\right) = \mathbf{0.9709}$$

$$\mathbf{GainRatio}(credit\_rating) = \frac{0.2267}{0.9709} = \mathbf{0.2335}\checkmark$$

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

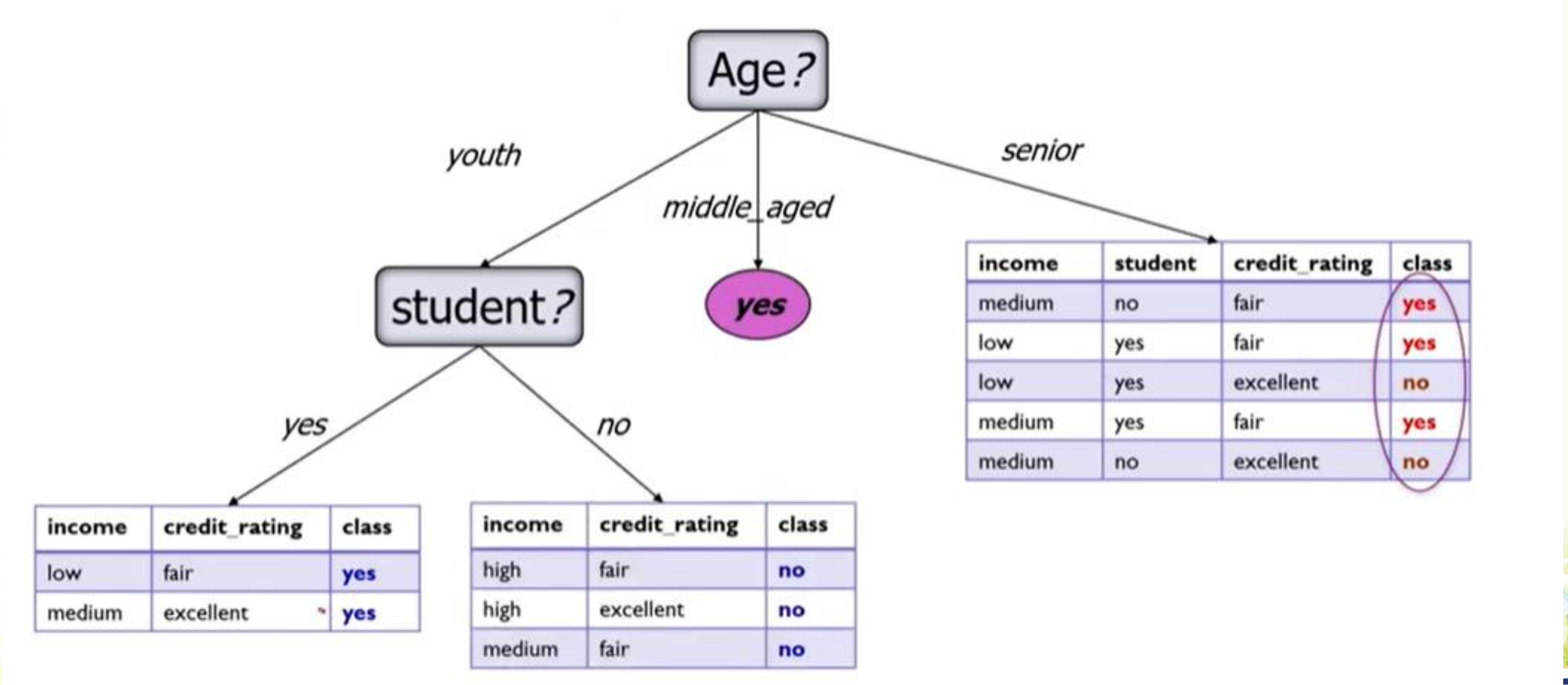
		Class		
		yes	no	
student	yes	2	0	2
	no	0	3	3
				5

$$Info_{student}(D) = \frac{2}{5}I(2{,}0) + \frac{3}{5}I(0{,}3) = \frac{2}{5} * 0 + \frac{3}{5} * 0 = \mathbf{0}$$

$$Gain(age) = 0.971 - 0 = \mathbf{0.971}$$

$$SplitInf0_{student}(D) = -\frac{2}{5} * \log_2\left(\frac{2}{5}\right) - \frac{3}{5} * \log2\left(\frac{3}{5}\right) = \mathbf{0.9709}$$

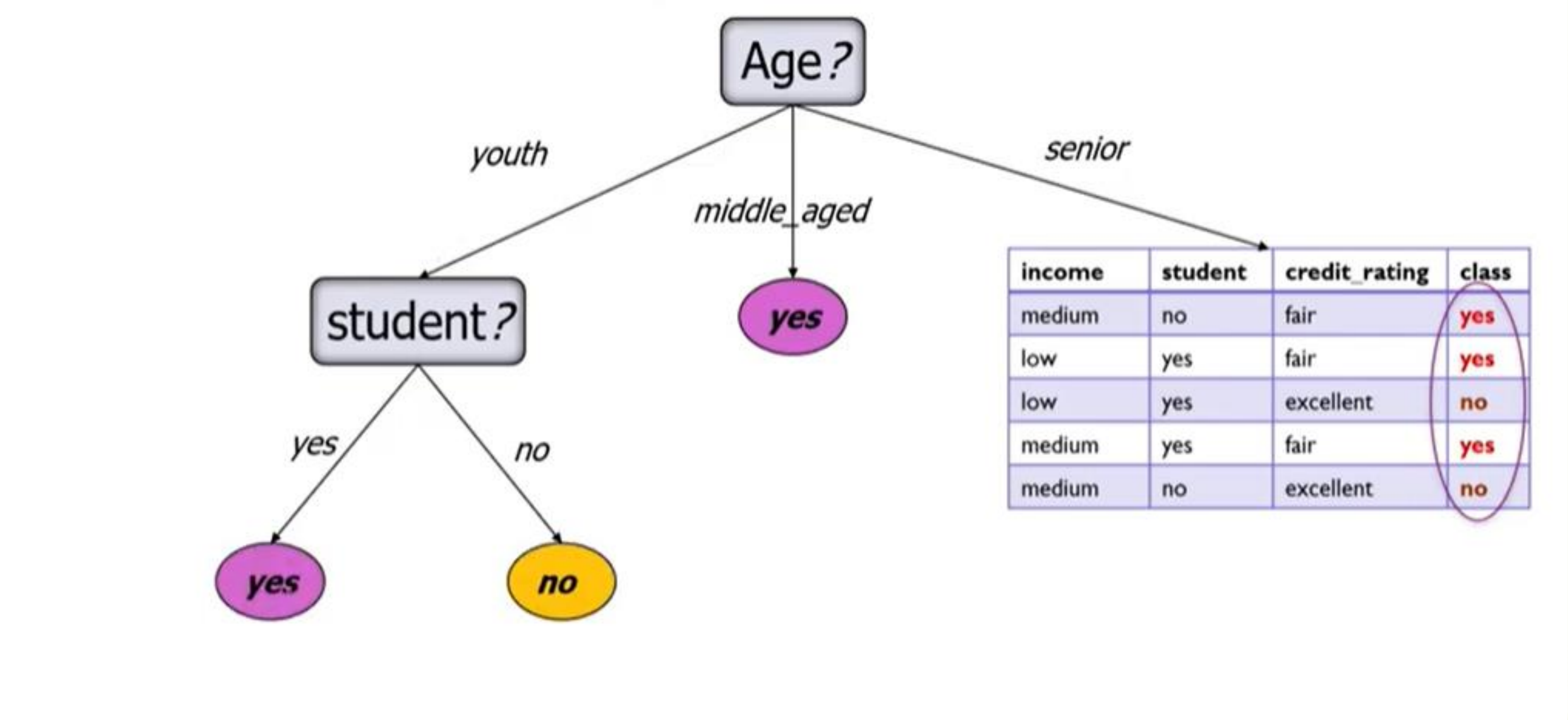
$$\mathbf{GainRatio(student)} = \frac{0.971}{0.9709} = \mathbf{1}\checkmark$$



income	credit_rating	class
low	fair	yes
medium	excellent	yes

income	credit_rating	class
high	fair	no
high	excellent	no
medium	fair	no

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



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

Gain Ratio [C4.5] - Example

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

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

$$Gain(A) = Info(D) - Info_A(D)$$

- For Right subtree: Calculate ***Entropy*** of Class attribute:

buys_computer	
yes	no
3	2

$$Info(D) = I(3,2) = -\frac{3}{5}log_2\left(\frac{3}{5}\right) - \frac{2}{5}log_2\left(\frac{2}{5}\right) = \mathbf{0.971}$$ ✓

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

- Calculate ***Gain Ratio*** of all other attributes:

		Class		
		yes	no	
income	low	1	1	2
	medium	2	1	3
	high	0	0	0
				5

$$Info_{income}(D) = \frac{2}{5}I(1,1) + \frac{3}{5}I(2,1)$$
$$= \frac{2}{5} \cdot 1 + \frac{3}{5} \cdot 0.9183 = 0.4 + 0.551 = \mathbf{0.951}$$

$$Gain(income) = 0.971 - 0.951 = \mathbf{0.02}$$

$$SplitInf0_{income}(D) = -\frac{2}{5} \cdot \log2\left(\frac{2}{5}\right) - \frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right) = \mathbf{0.9709}$$

$$\mathbf{GainRatio(income)} = \frac{0.02}{0.9709} = \mathbf{0.0205}$$ ✓

		Class		
		yes	no	
credit_ rating	fair	3	0	3
	excellent	0	2	2
				5

$$Info_{credit\_rating}(D) = \frac{3}{5}I(3,0) + \frac{2}{5}I(0,2) = \frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 0 = \mathbf{0}$$

$$Gain(credit\_rating) = 0.971 - 0 = \mathbf{0.971}$$

$$SplitInf0_{credit\_rating}(D) = -\frac{3}{5} \cdot \log2\left(\frac{3}{5}\right) - \frac{2}{5} \cdot \log_2\left(\frac{2}{5}\right) = \mathbf{0.9709}$$

$$\mathbf{GainRatio}(credit\_rating) = \frac{0.971}{0.9709} = \mathbf{1}$$

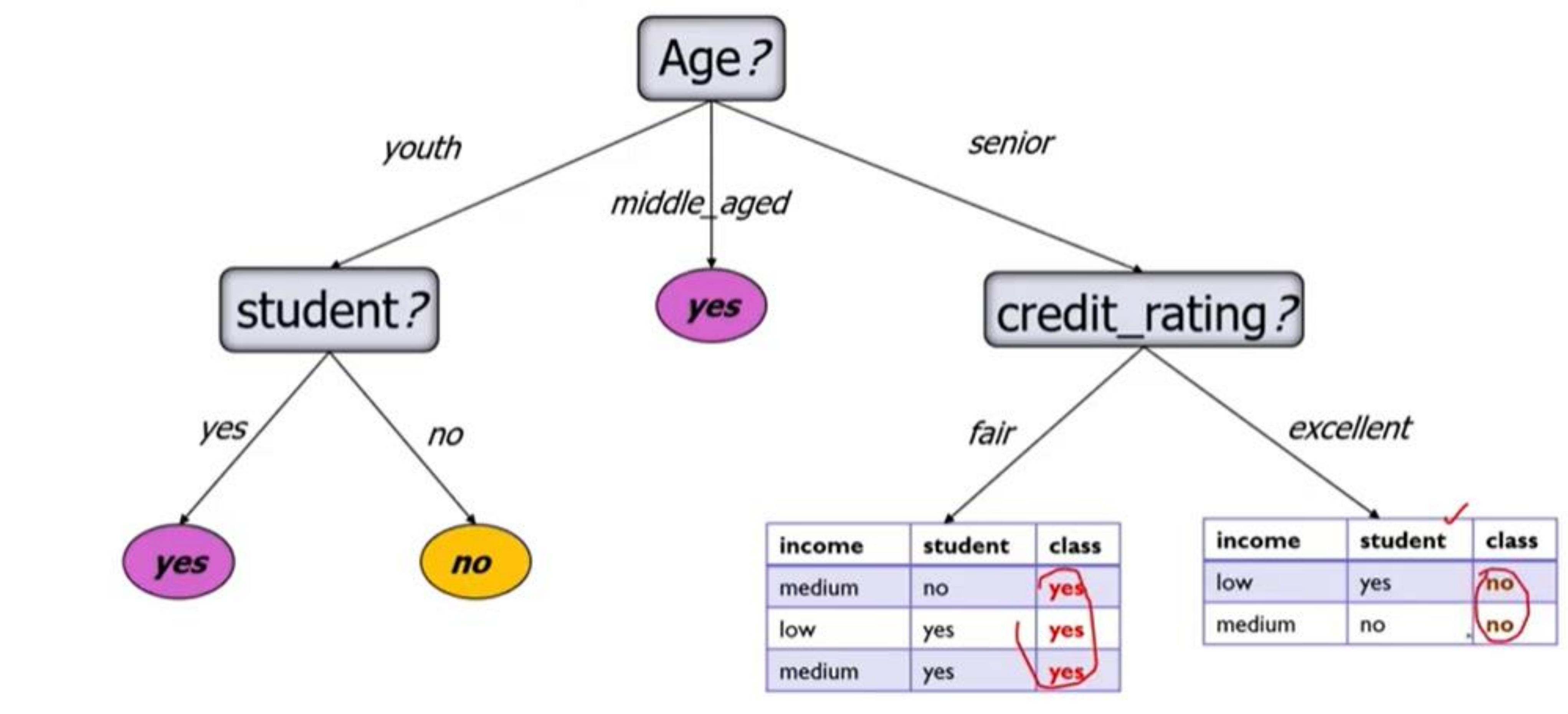
		Class		
		yes	no	
student	yes	2	1	3
	no	1	1	2
				5

$$Info_{student}(D) = \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1)$$
$$= \frac{3}{5} \cdot 0.9183 + \frac{2}{5} \cdot 1 = 0.551 + 0.4 = \mathbf{0.951}$$

$$Gain(age) = 0.971 - .951 = \mathbf{0.02}$$

$$SplitInf0_{student}(D) = -\frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \cdot \log2\left(\frac{2}{5}\right) = \mathbf{0.9709}$$

$$\mathbf{GainRatio(student)} = \frac{0.02}{0.9709} = \mathbf{0.0205}$$



income	student	class
medium	no	yes
low	yes	yes
medium	yes	yes

income	student ✓	class
low	yes	no
medium	no	no

What is the decision for

- X=[age, income, student, cradit]=[15,low,no,excellent]
- X=[age, income, student, cradit]=[40,low,no,excellent]

Comparing Attribute Selection Measures

- The two measures, in general, return good results but
  - **Information gain**:
    - biased towards multivalued attributes
  - **Gain ratio**:
    - tends to prefer unbalanced splits in which one partition is much smaller than the others