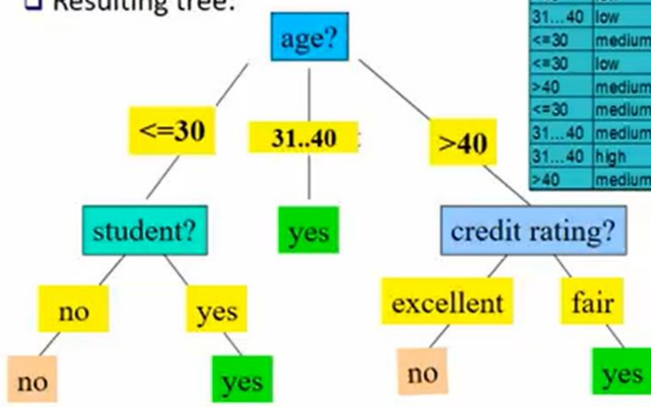
Classification

Decision Tree Induction: An Example

14

- Training data set: Buys_computer
- The data set follows an example of Quinlan's ID3 (Playing Tennis)
- Resulting tree:

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



2

classification--decision tree 4th oct 2021.mp4

Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

3:48 / 1:09:18

classification--decision tree 4th oct 2021.mp4

Attribute Selection Measures or Splitting Criterion

- The splitting criterion is determined so that, ideally, the resulting partitions at each branch are as "pure" as possible.
- A partition is pure if all the tuples in it belong to the same class.

A1	A2	A3	A4	A5	Class
X	1	Low	Yes	Khi	A
Y	1	High	No	Isl	A
X	2	Low	No	Isl	A
X	4	High	No	Lhr	B
Y	3	Low	No	Khi	A
Y	2	High	Yes	Isl	A
Y	1	Low	yes	Lhr	B

khi, isl -->A
Lhr --> B

4:04 / 1:09:18

- three popular attribute selection measures—information gain, gain ratio, and Gini index
- Information Gain - (The attribute with the highest information gain is chosen as the splitting attribute for 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
- Info(D) is just the average amount of information needed to identify the class label of a tuple in D. Info(D) is also known as the entropy of D.

- Select the attribute with the highest information gain
- Expected information (entropy) needed to classify a tuple in D: $Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$
- p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$p_{no} = 5 / 14$

$p_{yes} = 9 / 14$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

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

-0.0674

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = \boxed{0.940}$$

Information needed (after using A to split D into v partitions) to classify:

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

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's.
Thus, 5/14 I(2,3) = 5/14(-2/5 log2/5 – 3/5 log3/5)

Attribute Selection Measure: Information Gain (ID3/C4.5)

- Information gained by branching on attribute A

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

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

$$\text{Gain(age)} = 0.940 - 0.694 = 0.246$$

Attribute Selection: Information Gain

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Info income (D) = ?

Info student? (D) = ?

Infor CR (D) =?

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \qquad Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Info income (D) = 4/14 I(2,2) +6/14 I(4,2) +4/14 I(3,1)

Info income (D) = 4/14 {-2/4log(2/4)-2/4log(2/4)} +

6/14 {-4/6log(4/6)-2/6log(2/6)} +

4/14 {-3/4log(3/4)-1/4log(1/4)} = 0.911

Info student? (D) = 7/14 I(6,1) +7/14 I(3,4)

Info student? (D) = 7/14 {-6/7 log(6/7)-1/7 log(1/7)} +

7/14 {-3/7 log(3/7)-4/7 log(4/7)}

= 0.788

Info CR (D) = 8/14 I(6,2) +6/14 I(3,3)

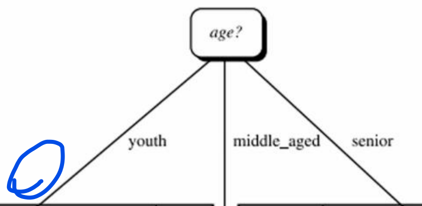
Info CR (D) = 8/14 {-6/8 log(6/8)-2/8 log(2/8)} +

6/14 {-3/6 log(3/6)-3/6 log(3/6)}

= 0.892

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

- **Age** will be selected as splitting attribute as it has the greatest gain or minimum Info



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
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
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

12

Type here to search   31°C Mostly sunny   5:19 PM 10/10/2021

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

(1) $Info_{income} = \frac{2}{5}\left\{-\frac{2}{2}\log\frac{2}{2} - \frac{0}{2}\log\frac{0}{2}\right\} + \frac{2}{5}\left\{-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right\} + \frac{1}{5}\left\{-\frac{1}{1}\log\frac{1}{1} - \frac{0}{1}\log\frac{0}{1}\right\}$

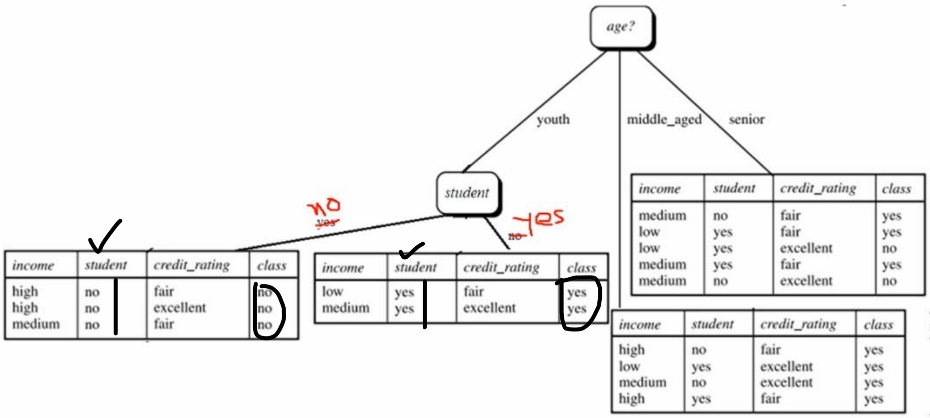
$= \boxed{0.4}$ 0.54

(2) $Info_{student} = \frac{3}{5}\left\{-\frac{3}{3}\log\frac{3}{3} - 0\right\} + \frac{2}{5}\left\{-\frac{2}{2}\log\frac{2}{2} - 0\right\}$

$\boxed{= 0}$ 0.940

(3) $Info_{CR} = \frac{3}{5}\left\{-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right\} + \frac{2}{5}\left\{-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right\}$

$= \boxed{0.951}$ -0.011

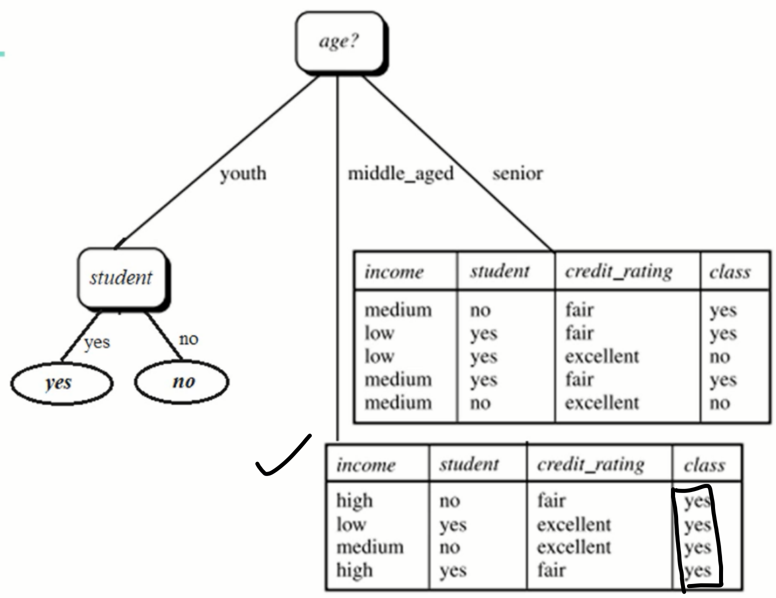


income	student	credit_rating	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no

income	student	credit_rating	class
low	yes	fair	yes
medium	yes	excellent	yes

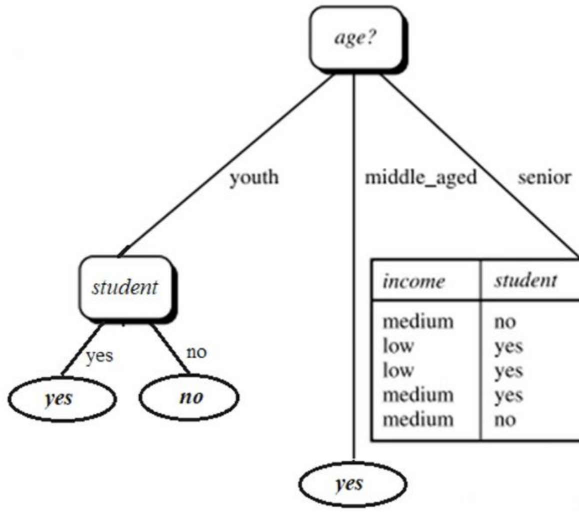
income	student	credit_rating	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

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

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

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

1. $$\text{Info}_{\text{income}} = \frac{3}{5}\left\{-\overset{Y}{\frac{2}{3}}\log\frac{2}{3} - \overset{N}{\frac{1}{3}}\log\frac{1}{3}\right\} + \frac{2}{5}\left\{-\overset{Y}{\frac{1}{2}}\log\frac{1}{2} - \frac{1}{2}\overset{N}{\log}\frac{1}{2}\right\} = \boxed{0.951}$$

0.011

2. $$\text{Info}_{\text{student}} = \frac{2}{5}\left\{-\overset{Y}{\frac{1}{2}}\log\frac{1}{2} - \frac{1}{2}\overset{N}{\log}\frac{1}{2}\right\} + \frac{3}{5}\left\{-\overset{Y}{\frac{2}{3}}\log\frac{2}{3} - \frac{1}{3}\overset{N}{\log}\frac{1}{3}\right\} = \boxed{0.951}$$

"

3. $$\text{Info}_{CR} = \frac{3}{5}\left\{\overset{Y}{\frac{3}{3}}\log\frac{3}{3} - \overset{N}{\frac{0}{3}}\right\} + \frac{2}{5}\left\{\overset{N}{\frac{2}{2}}\log\frac{2}{2} \overset{Y}{-} 0\right\} \boxed{= 0}$$

0.94