Lecture Three

Decision Tree
Example

Example

age	income	student	credit_rating	ys_comput
<=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

Tree induction example

- **info(D) = - p* log2p+ - p* log2p-**

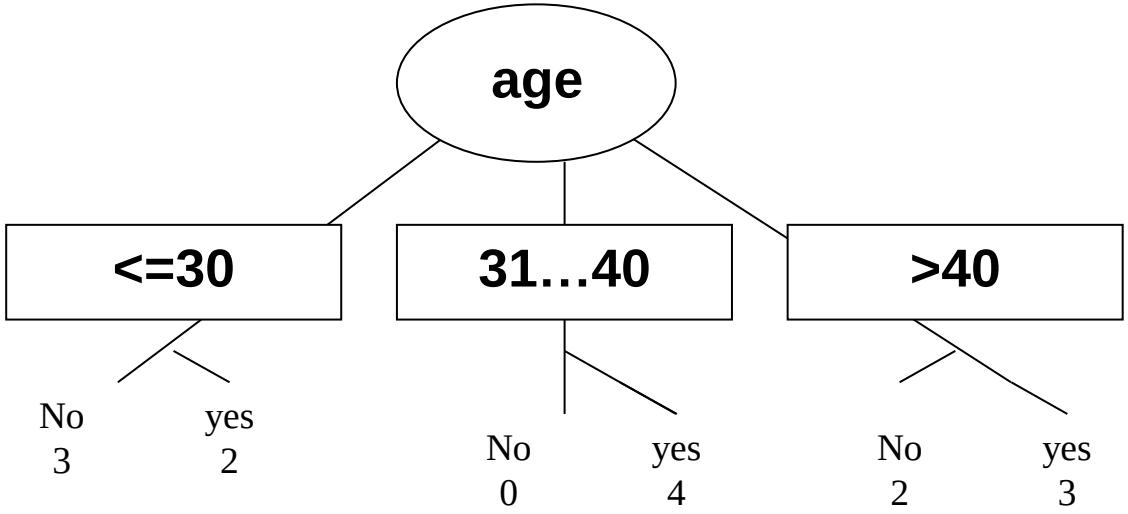
Info(D) = -((9/14)log2(9/14))-(((5/14)log2(5/14))
= 0.940

D[9+, 5-]
age

<=30 [2+,3-]

31…40 [4+,0-]

>40 [3+,2-]



$Info_{age}(D)$ = 5/14((-2/5)$\log_2$(2/5)) - ((3/5)$\log_2$(3/5))

+ 4/14 ((-4/4)$\log_2$(4/4))- ((0/4)$\log_2$(0/4))

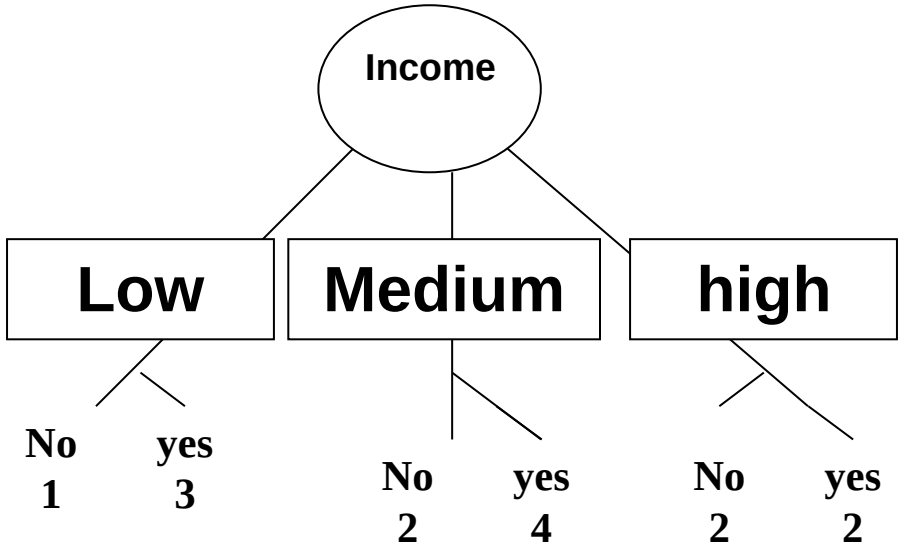
+ 5/14((-3/5)$\log_2$(3/5)) - ((2/5)$\log_2$(2/5))

= 0.694 bits

$Gain(\textbf{age}) = Info(D) - Info\ \textbf{age}(D)$ $= 0.940 - 0.694 = 0.246$ bits.

D[9+, 5-]
Income

- Low [3+,1-]
- Medium [4+,2-]
- High [2+,2-]

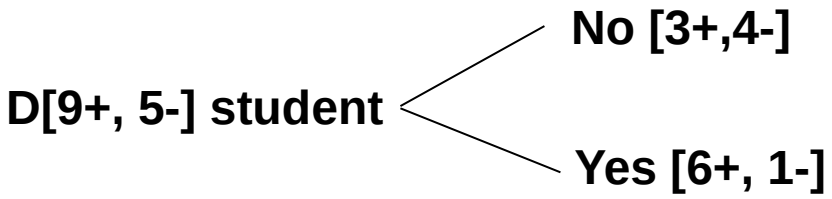


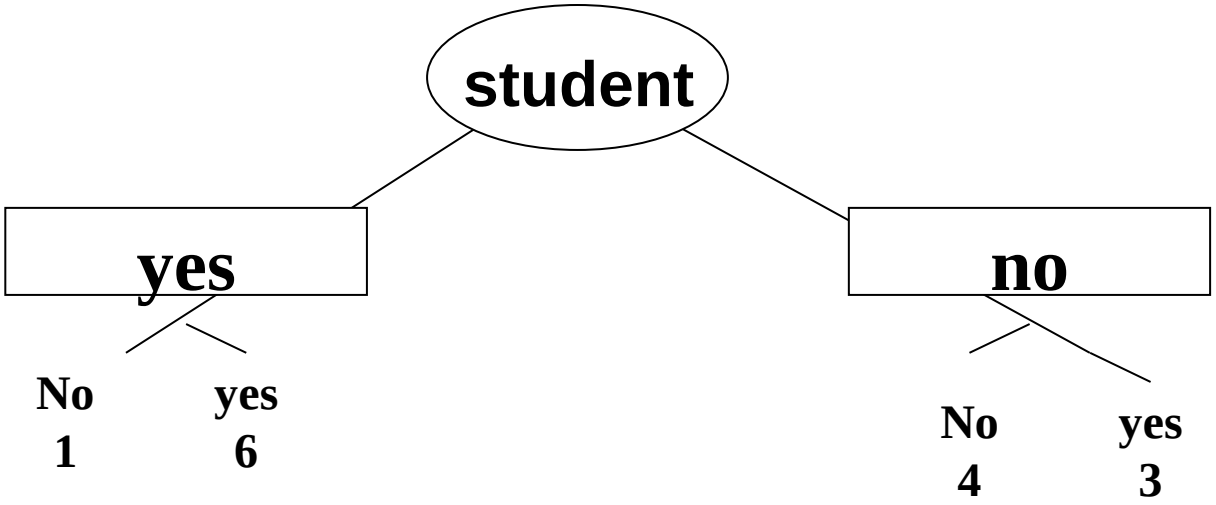
Info gain Income (D) =

$$4/14((-3/4)\log_2(3/4)) - ((1/4)\log_2(1/4))$$
$$+ 6/14((-4/6)\log_2(4/6)) - ((2/6)\log_2(2/6))$$
$$+ 4/14((-2/4)\log_2(2/4)) - ((2/4)\log_2(2/4))$$
$$= 0.91104 \text{ bits}$$

Gain Income = Info Gain (D) – Info gain Income (D)

= 0.940 – 0.91194 = 0.029 bits





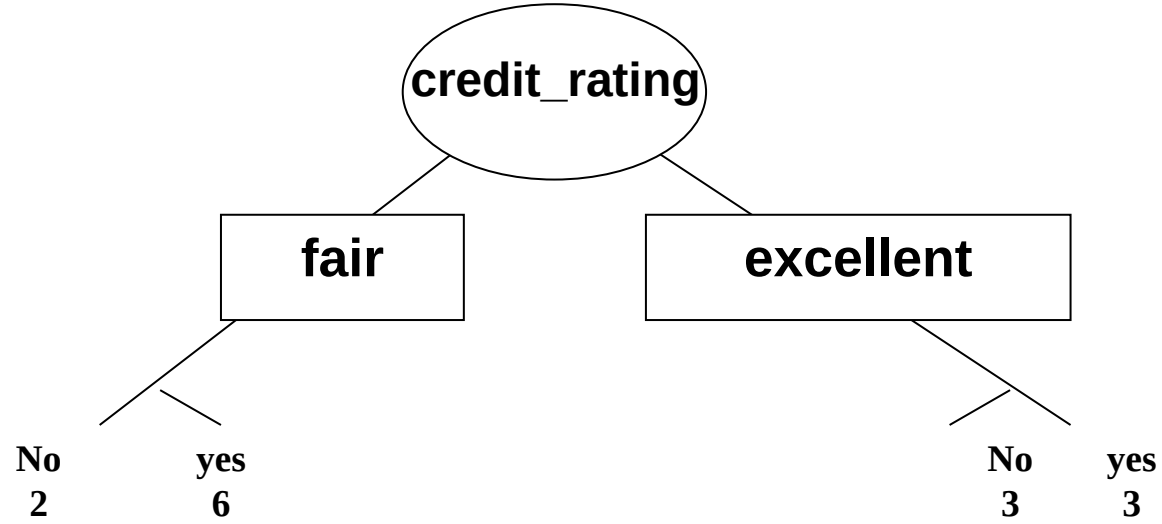
Info gain student(D) = $7/14((-4/7)\log_2(4/7)) - ((3/7)\log_2(3/7))$

$+ 7/14\ ((-6/7)\log_2(6/7)) - ((1/7)\log_2(1/7))$

$= 0.7884$ **bits.**

Gain student = InfoGain (D) – InfoGain student (D)

$= 0.940 - 0.7884 = 0.151$ bits.



Info gain credit_rating (D) == $8/14((-6/8)\log_2(6/8)) - ((2/8)\log_2(2/8))$

$+ 6/14 ((-3/6)\log_2(3/6)) - ((3/6)\log_2(3/6))$

= 0.892 bits.

Gain = InfoGain (D) – Info gain credit_rating (D)

= 0.940 – 0.892 = 0.048 bits.

age	income	student	credit_rating	comp
<=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

Gain = 0.246 = 0.029 = 0.151 =0.048

Gain Ratio for Attribute Selection (C4.5)

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

split info بشوف عندي كام متغير في كل عمود وبحسب علي اساسه

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

D[14]
age

- <=30 [5]
- 31…40 [4]
- >40 [5]

$$SplitInfo_{Age}(D) = -\frac{5}{14} \times \log_2\left(\frac{5}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{5}{14} \times \log_2\left(\frac{5}{14}\right) = 1.5774$$

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

D[14]
Income

- Low [4]
- Medium [6]
- High [4]

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

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

D[14]
student

- No [7]
- Yes [7]

$$SplitInfo_{student}(D) = -\frac{7}{14} \times \log_2\left(\frac{7}{14}\right) - \frac{7}{14} \times \log_2\left(\frac{7}{14}\right) = 1$$

$$SplitInfo_{credit\_rating}(D) = -\sum_{i=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

D[14]
credit_rating

- Fair [8]
- Excellent [6]

$$SplitInfo_{credit\_rating}(D) = -\frac{8}{14} \times \log_2\left(\frac{8}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) = 0.9852$$

age	income	student	credit_rating	_com
<=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

Gain = 0.246 = 0.029 = 0.151 =0.048

split Info = 1.5774 = 1.577 = 1 =0.9852

Gain Ratio(A) = Gain(A)/Split Info(A)

	Age	Income	Student	credit rating
Gain	0.246	0.029	0.151	0.048
Split Info	1.5774	1.577	1	0.9852

Gain Ratio(Age) = Gain(age)/Split Info(age)

Gain Ratio(Age) = 0.246/1.5774 =0.1559

Gain Ratio(Income) = Gain(income)/Split Info(income)

Gain Ratio(income) = 0.029/1.577 =0.019

Gain Ratio(A) = Gain(A)/Split Info(A)

	Age	Income	Student	credit rating
Gain	0.246	0.029	0.151	0.048
Split Info	1.5774	1.577	1	0.9852

Gain Ratio(Student) = Gain(Student)/Split Info(Student)

Gain Ratio(Student) = 0.151/1=0.151

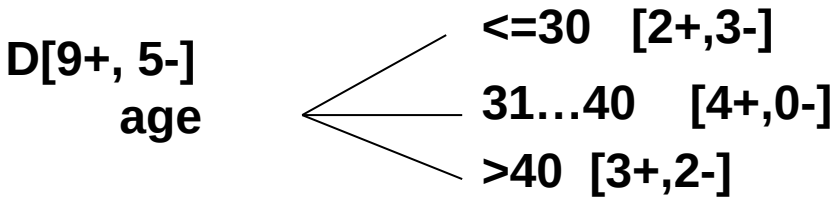
Gain Ratio(credit_rating) = Gain(credit_rating)/Split Info(credit_rating)

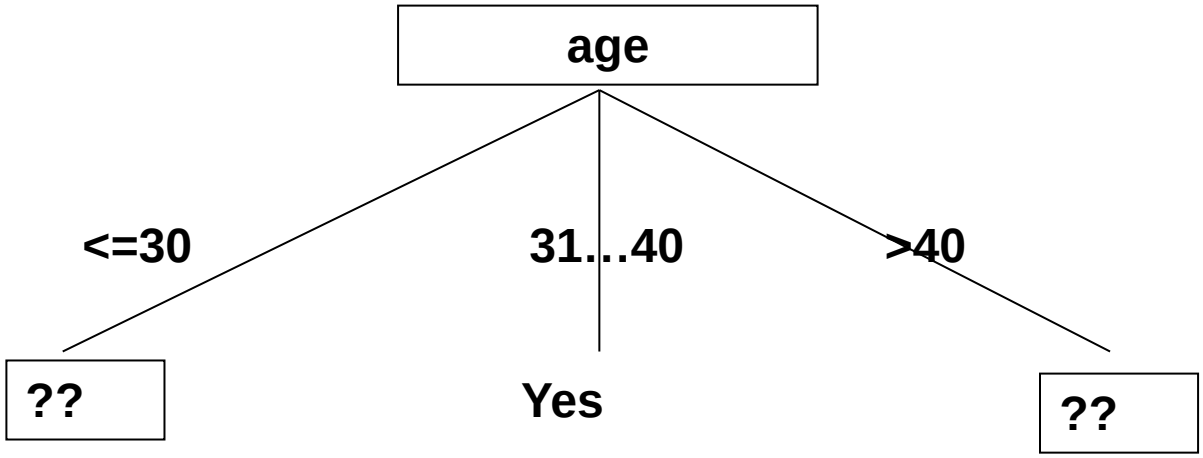
Gain Ratio(credit_rating) = 0.048/0.9852=0.048782

Gain Ratio(A) = Gain(A)/Split Info(A)

	Age	Income	Student	credit rating
Gain	0.246	0.029	0.151	0.048
Split Info	1.5774	0.926	1	0.9852
Gain Ratio	0.1559	0.019	0.151	0.048782

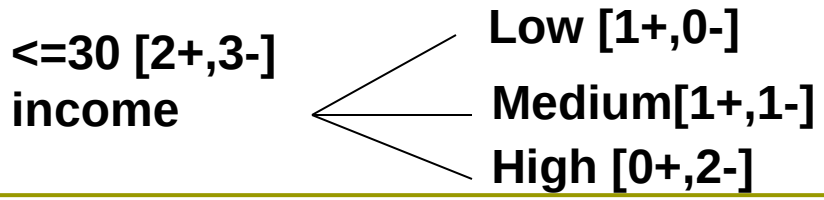
Age attribute with the maximum gain ratio is selected as the splitting attribute


D[9+, 5-]
age
<=30 [2+,3-]
31…40 [4+,0-]
>40 [3+,2-]

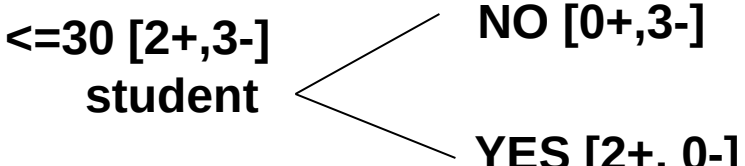

age
<=30
31…40
>40
??
Yes
??

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

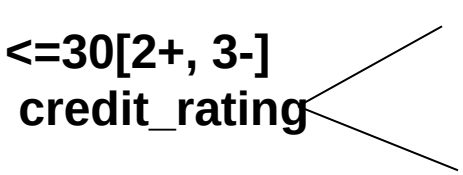
$$Info(\mathbf{<=30}) = -2/5(\log_2(2/5)) - 3/5(\log_2(3/5)) = 0.97$$



$$Gain(income) = 0.97 - 1/5[-1/1(\log_2(1/1))-0/1(\log_2(0/1))] - 2/5[-1/2(\log_2(1/2))-1/2(\log_2(1/2))] - 2/5[-0/2(\log_2(0/2))-2/2(\log_2(2/2))] = 0.97 - 0.4 = 0.57$$



$$Gain(student) = 0.97 - 3/5[-0/3(\log_2(0/3))-3/3(\log_2(3/3))] - 2/5[-2/2(\log_2(2/2))-0/2(\log_2(0/2))] = 0.97 - 0 = \underline{0.97}$$



$$Gain(credit\_rating) = 0.97 - 3/5[-1/3(\log_2(1/3))-2/3(\log_2(2/3))] - 3/5[-1/2(\log_2(1/2))-1/2(\log_2(1/2))] = 0.97 - 0.96 = 0.02$$

age	income	student	credit_rating	_com
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

Gain = *0.57* = *0.97* = *0.02*

$$SplitInfo_{Income}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$
<=30

D[5]
Income
- Low [1]
- Medium [2]
- High [2]

$$= -\frac{1}{5} \times \log_2\left(\frac{1}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right) = \mathbf{0.993}$$

$$SplitInfo_{student}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$
<=30

D[5]
student
- No [3]
- Yes [2]

$$-\frac{2}{5} \times \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \times \log_2\left(\frac{3}{5}\right) = \mathbf{0.970}$$

$$SplitInfo_{credit\_rating}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$
<=30

D[5]
credit_rating
- Fair [3]
- Excellent[2]

$$= -\frac{3}{5} \times \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right) = \mathbf{0.970}$$

age	income	student	credit_rating	_com
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

Gain = ***0.57*** = ***0.97*** = ***0.02***

split Info = ***0.993*** = **0.970** = **0.970**

Gain Ratio(A) = Gain(A)/Split Info(A)

	Age	Income	Student	credit rating
Gain	<= 30	0.057	0.97	0.02
Split Info		0.993	0.97	0.97

Gain Ratio(Income <=30) = Gain(income)/Split Info(income)

Gain Ratio(Income <=30) = 0.057/0.993 =0.0574

Gain Ratio(A) = Gain(A)/Split Info(A)

	Age	Income	Student	credit rating
Gain	<= 30	0.057	0.97	0.02
Split Info		0.993	0.97	0.97

Gain Ratio(Student <= 30) = Gain(Student)/Split Info(Student)

Gain Ratio(Student <= **30) =** *0.97* **/***0.97* **=1**

Gain Ratio(credit_rating <= 30) = Gain(credit_rating)/Split Info(credit_rating)

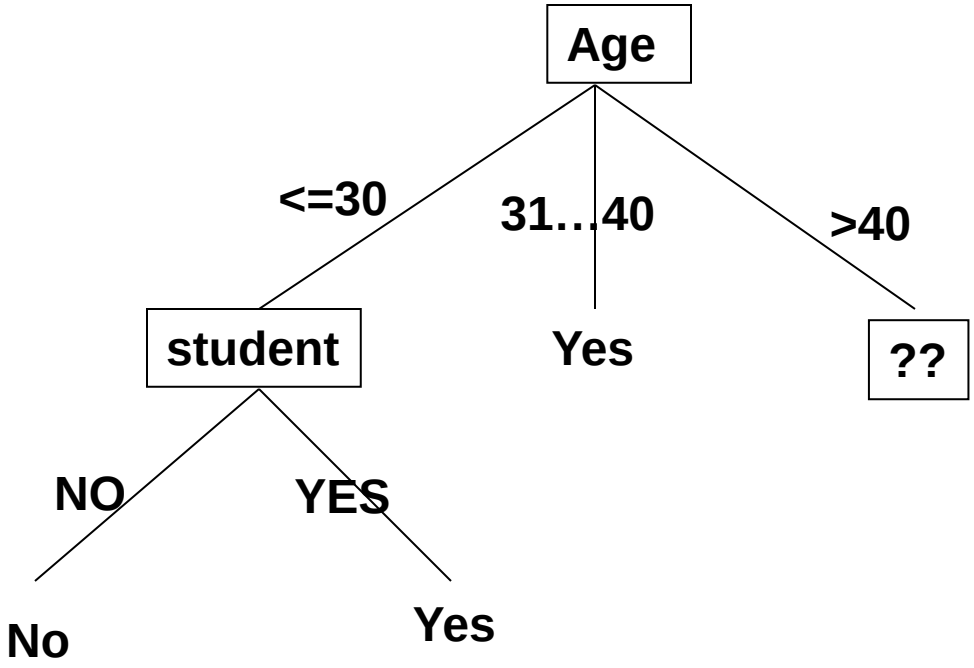
Gain Ratio(credit_rating) = 0.02/0.97=0.0206

Gain Ratio(A) = Gain(A)/Split Info(A)

	Age	Income	Student	credit rating
Gain	<= 30	0.057	0.97	0.02
Split Info		0.993	0.97	0.97
GainRatio		*0.0574*	*1*	*0.0206*

Student **attribute with the maximum gain ratio is selected as the splitting attribute**

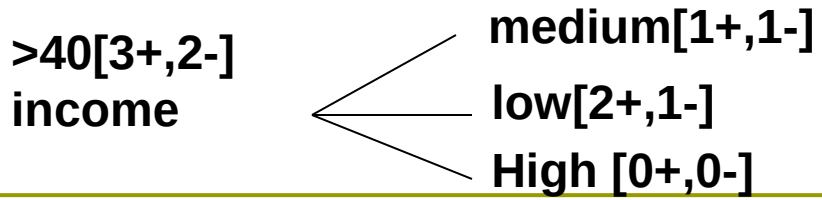
Tree induction example



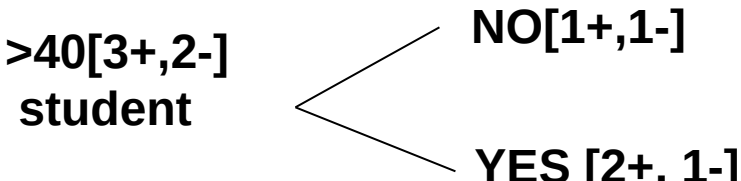
Tree induction example

age	income	student	credit_rating	_com
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
>40	medium	yes	fair	yes
>40	medium	no	excellent	no

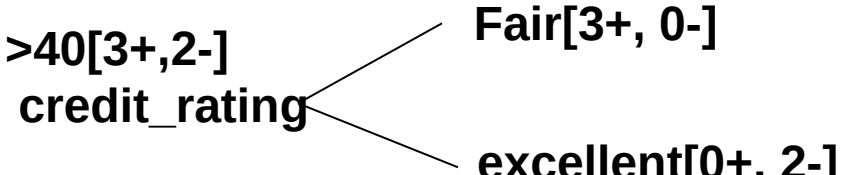
$$Info(>40) = -3/5(log2(3/5))-2/5(log_2(2/5))$$
$$= 0.97$$



$$Gain(\text{income}) = 0.97 - 2/5[-1/2(\log_2(1/2))-1/2(\log_2(1/2))] - 3/5[-2/3(\log_2(2/3))-1/3(\log_2(1/3))] - 0/5[-0/0(\log_2(0/0))-0/0(\log_2(0/0))] = 0.97 - 0.75 = 0.22$$



$$Gain(\text{student}) = 0.97 - 2/5[-1/2(\log_2(1/2))-1/2(\log_2(1/2))] - 3/5[-2/3(\log_2(2/3))-1/3(\log_2(1/3))] = 0.97 - 0.43 = 0.54$$



$$Gain(\text{credit\_rating}) = 0.97 - 3/5[-3/3(\log_2(3/3))-0/3(\log_2(0/3))] - 2/5[-0/2(\log_2(0/2))-2/2(\log_2(2/2))] = 0.97 - 0 = \underline{0.97}$$

age	income	student	credit_rating	_com
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
>40	medium	yes	fair	yes
>40	medium	no	excellent	no

Gain = *0.22* = *0.54* = *0.97*

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

>40[3+,2-]
income

- medium[1+,1-]
- low[2+,1-]
- High [0+,0-]

$$-\frac{2}{5} \times \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \times \log_2\left(\frac{3}{5}\right) = \mathbf{0.970}$$

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

>40[3+,2-]
student

- NO[1+,1-]
- YES [2+, 1-]

$$-\frac{2}{5} \times \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \times \log_2\left(\frac{3}{5}\right) = \mathbf{0.970}$$

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

>40[3+,2-]
credit_rating

- Fair[3+, 0-]
- excellent[0+, 2-]

$$= -\frac{3}{5} \times \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right) = \mathbf{0.970}$$

age	income	student	credit_rating	_com
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
>40	medium	yes	fair	yes
>40	medium	no	excellent	no

Gain = *0.57* = *0. 54* = *0.97*

split Info = *0.970* = **0.970** = **0.970**

Gain Ratio(A) = Gain(A)/Split Info(A)

	Age	Income	Student	credit rating
Gain	> 40	0.057	0.54	0.97
Split Info		0.970	0.97	0.97

Gain Ratio(Income > 40) = Gain(income)/Split Info(income)

Gain Ratio(Income > 40) = 0.057/0.970 =0.05876

Gain Ratio(A) = Gain(A)/Split Info(A)

	Age	Income	Student	credit rating
Gain	> 40	0.057	0.54	0.97
Split Info		0.970	0.97	0.97

Gain Ratio(Student > 40) = Gain(Student)/Split Info(Student)

Gain Ratio(Student > 40**) = 0.54 /0.97 =0.5567**

Gain Ratio(credit_rating > 40) = Gain(credit_rating)/Split Info(credit_rating)

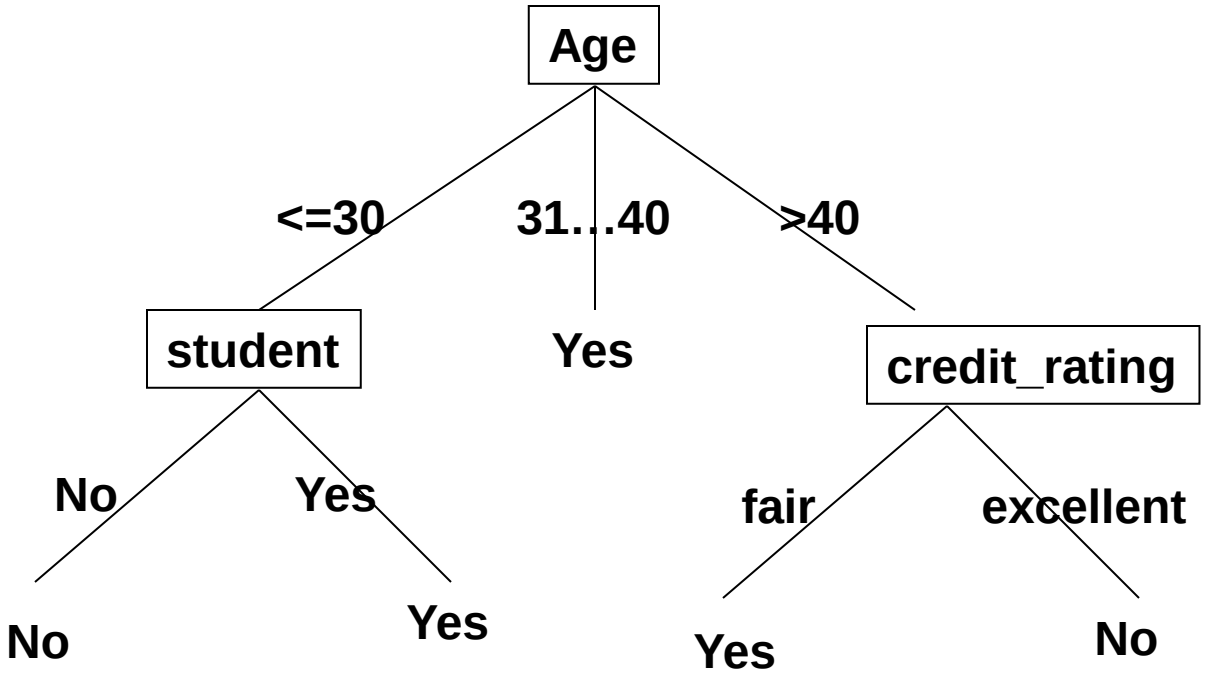
Gain Ratio(credit_rating) = 0.97/0.97= 1

Gain Ratio(A) = Gain(A)/Split Info(A)

	Age	Income	Student	credit rating
Gain	> 40	0.057	0.54	0.97
Split Info		0.970	0.97	0.97
GainRatio		0.05876	0.5567	1

credit rating attribute with the maximum gain ratio is selected as the splitting attribute

Tree induction example




Thank you