

## Classification : ทำนายกลุ่มของข้อมูล

- สร้างเส้นแบ่ง ที่สามารถจัดกลุ่มข้อมูล
- set of sample used for model construction **training set**
- **New data is classified based on the models** built from training set
- Predict Categorical class labels (discrete / nominal)

เส้นแบ่งที่ดีที่สุดในการจัดกลุ่มข้อมูล

จะเห็นว่า ยังมี data points บางตัวที่ไม่เป็นไปตามเส้น แต่  
This is prediction there will be some discrepancy

Training Data with class

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

Training  
Instances

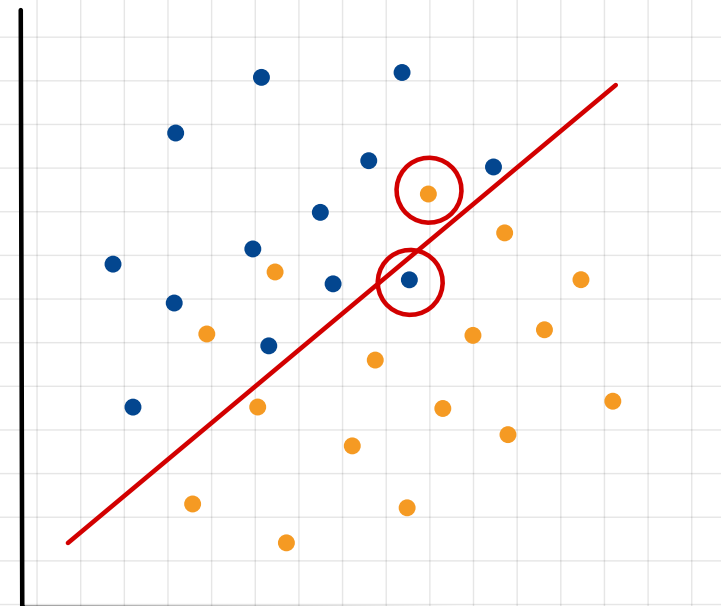
Model  
Learning

Test  
Instances

Prediction  
Model

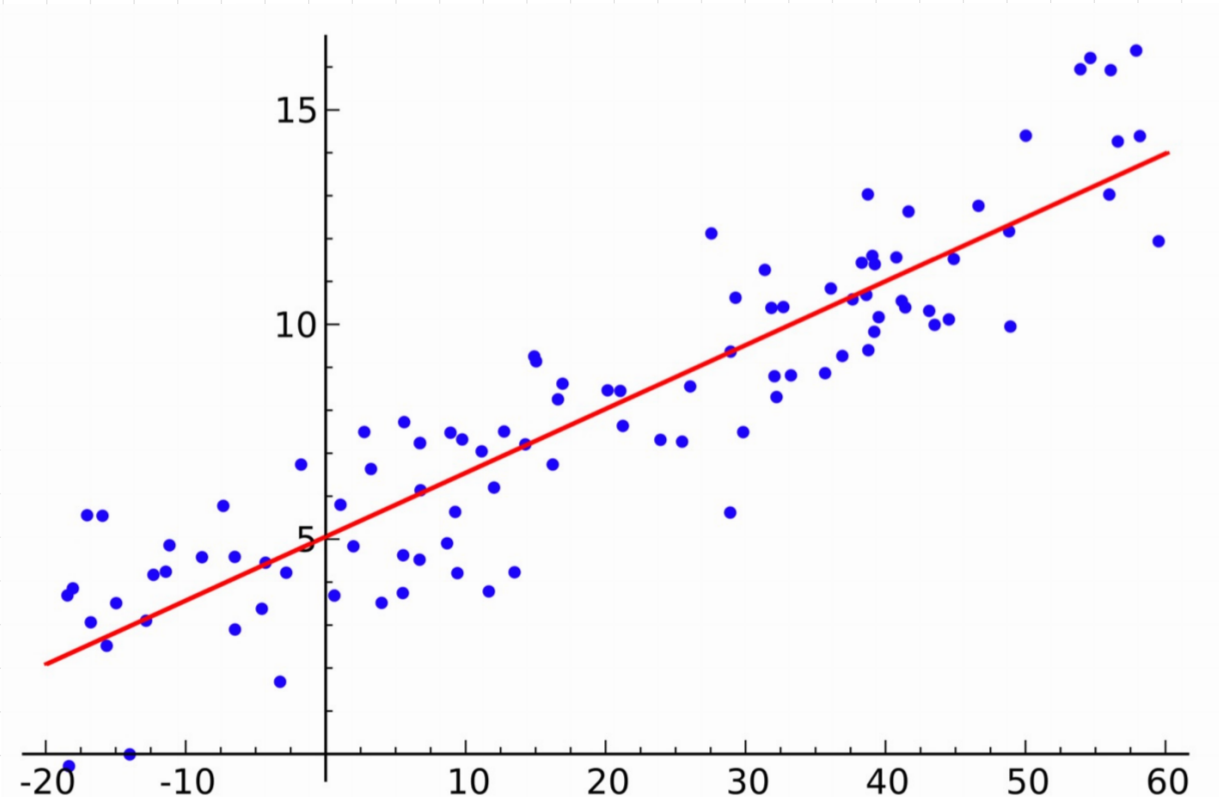
Positive

Negative



Numeric prediction : ทำหาค่าของข้อมูล

- สร้างโมเดล หรือฟังก์ชัน ที่ใช้ทำนายค่าของข้อมูล (predict unknown / missing values)
- ซึ่งก็คือ **Regression Model** หั่นแหละ สมการที่สำคัญ  $y = a + bx_1 + bx_2 + \dots + e$
- หาเส้น / ระนาบ ที่ดีที่สุด ที่ตัดผ่านจุดของข้อมูลมากที่สุด :  $f(x)$



## Decision Tree : predict class label

Information Theory : Entropy ใช้ในการทำ attributes selection measure

- เลือก the highest information gain
- มีสูตร

- Expected information (entropy) needed to classify a tuple in D:

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

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

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

- Information gained by branching on attribute A

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

RID	Age	Income	Student	Credit rating	Class: buys computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle_aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle_aged	Low	Yes	Excellent	yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle_aged	Medium	No	Excellent	Yes
13	Middle_aged	High	Yes	Fair	Yes
14	Senior	Mdium	No	Excellent	no

Step 1 : หา expected info ที่จะ classify ก่อน

จากตัวอย่าง class label: buys computer มีค่าคำตอบอยู่ 2 แบบ

- class P : buys computer = 'yes' 9 อัน
- class N : buys computer = 'no' 5 อัน

จากสูตร จะได้

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

RID	Age	Income	Student	Credit rating	Class: buys computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle_aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle_aged	Low	Yes	Excellent	yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle_aged	Medium	No	Excellent	Yes
13	Middle_aged	High	Yes	Fair	Yes
14	Senior	Mdium	No	Excellent	no

## Step 2: หา info ของแต่ละ attribute

- Age

	P (yes)	N (no)	
Youth (<=30)	2	3	5/14
Middle (31...40)	4	0	4/14
Senior (>40)	3	2	5/14

$$Info_{age}(D) = \frac{5}{14} \times \left( -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right)$$

$$= + \frac{4}{14} \times \left( -\frac{4}{4} \log_2 \frac{4}{4} \right) + \frac{5}{14} \times \left( -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) = 0.694 \text{ bits.}$$

## Step 3 : หา information gained ของ แต่ละ attributes

Hence, the gain in information from such a partitioning would be

$$\text{Gain}(\text{age}) = \text{Info}(D) - \text{Info}_{\text{Age}}(D)$$

$$= 0.940 - 0.694 = 0.246 \text{ bits}$$

RID	Age	Income	Student	Credit rating	Class: buys computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle_aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle_aged	Low	Yes	Excellent	yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle_aged	Medium	No	Excellent	Yes
13	Middle_aged	High	Yes	Fair	Yes
14	Senior	Mdium	No	Excellent	no

## Step 2: หา info ของแต่ละ attribute

- Income

	P (yes)	N (no)	
High	2	2	4/14
Medium	4	2	6/14
Low	3	1	4/14

$$Info_{Income}(D) = \frac{4}{14} \times \left( -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} \right)$$

$$= + \frac{6}{14} \times \left( -\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} \right) + \frac{4}{14} \times \left( -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right) = 0.911 \text{ bits.}$$

## Step 3 : หา information gained ของ แต่ละ attributes

Hence, the gain in information from such a partitioning would be

$$Gain(Info_{Income}) = Info(D) - Info_{Income}(D)$$

$$= 0.940 - 0.911 = 0.029 \text{ bits}$$



RID	Age	Income	Student	Credit rating	Class: buys computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle_aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle_aged	Low	Yes	Excellent	yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle_aged	Medium	No	Excellent	Yes
13	Middle_aged	High	Yes	Fair	Yes
14	Senior	Mdium	No	Excellent	no

## Step 2: หา info ของแต่ละ attribute

- Student

	P (yes)	N (no)	
Yes	6	1	7/14
No	3	4	7/14

$$Info_{Student}(D) = \frac{7}{14} \times \left( -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} \right)$$

$$+ \frac{3}{14} \times \left( -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} \right)$$

$$= 0.788$$

## Step 3 : หา information gained ของ แต่ละ attributes

Hence, the gain in information from such a partitioning would be

$$Gain_{Student} = Info(D) - Info_{Student}(D)$$

$$= 0.940 - 0.788 = 0.151 \text{ bits}$$

RID	Age	Income	Student	Credit rating	Class: buys computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle_aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle_aged	Low	Yes	Excellent	yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle_aged	Medium	No	Excellent	Yes
13	Middle_aged	High	Yes	Fair	Yes
14	Senior	Mdium	No	Excellent	no

Step 2: หา info ของแต่ละ attribute

- credit

	P (yes)	N (no)	
Fair	6	2	8/14
Excellent	3	3	6/14

$$Info_{Credit}(D) = \frac{8}{14} \times \left( -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} \right)$$

$$+ \frac{6}{14} \times \left( -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} \right)$$

$$= 0.892$$

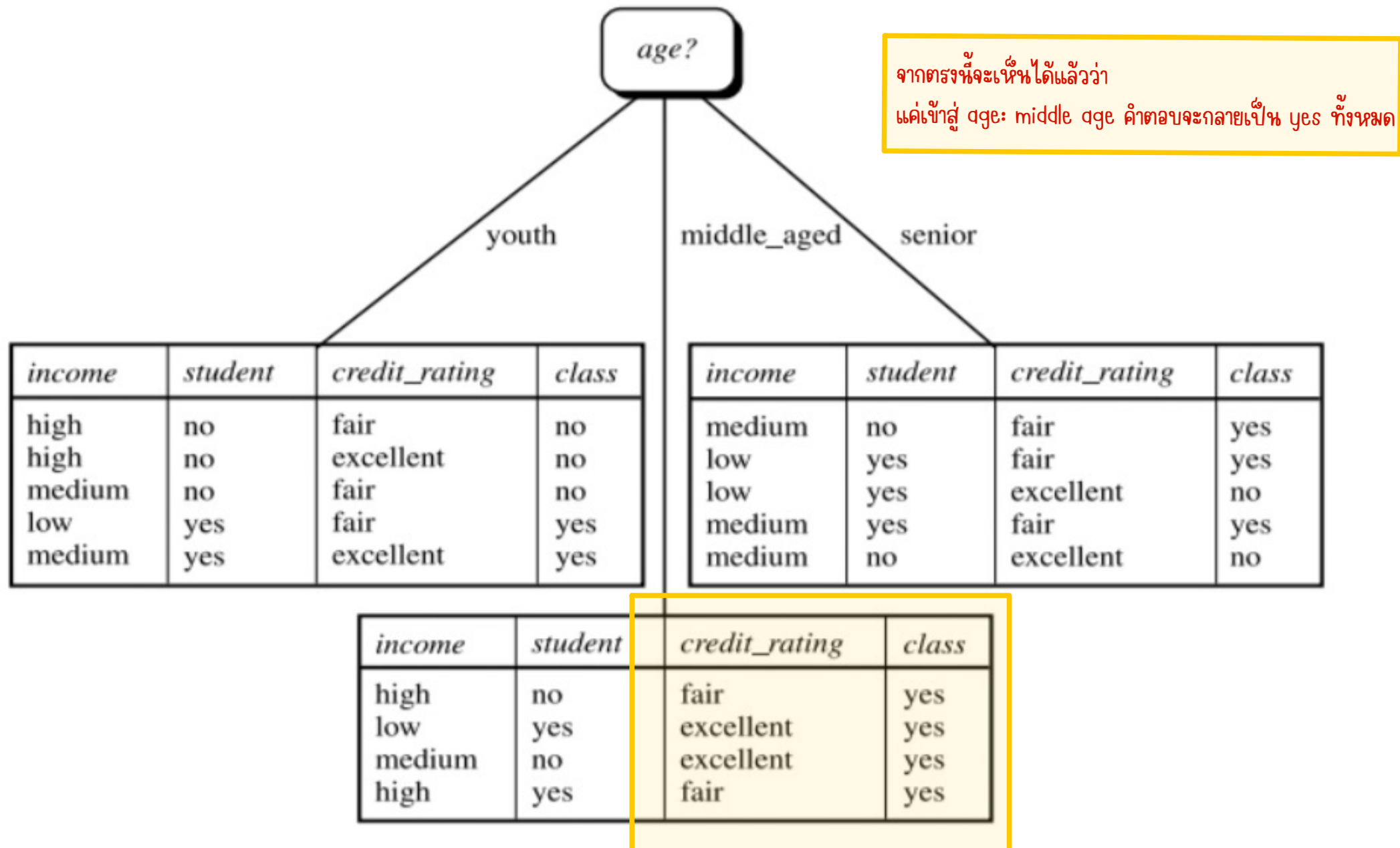
Step 3 : หา information gained ของ แต่ละ attributes

Hence, the gain in information from such a partitioning would be

$$\begin{aligned} \text{Gain}_{Credit} &= \text{Info}(D) - \text{Info}_{Credit}(D) \\ &= 0.940 - 0.892 = 0.048 \text{ bits} \end{aligned}$$



มาถึงตรงนี้ เราจะได้ Root Node : Age เพราะ มี highest information gain



age?

youth

middle\_aged

senior

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

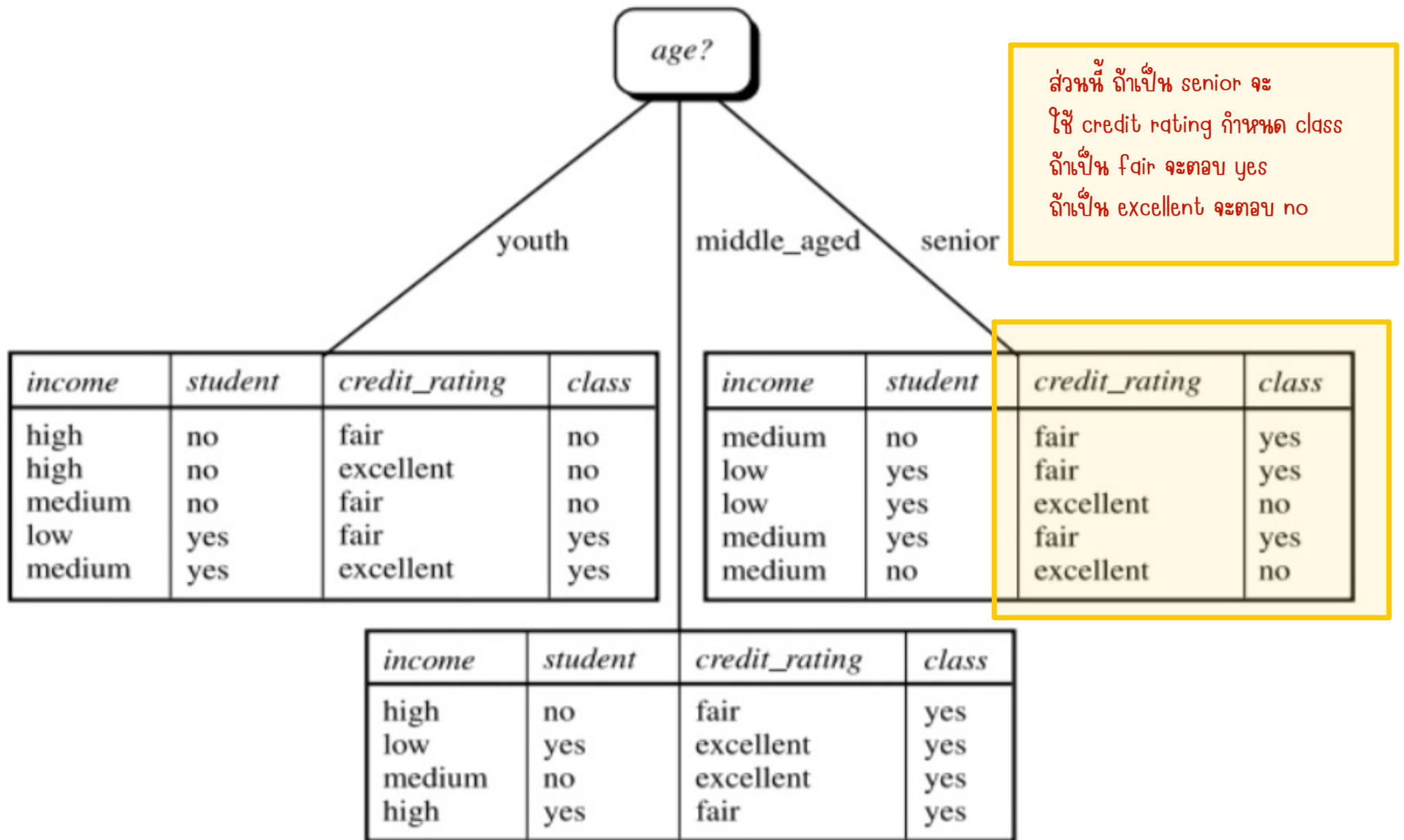
income	student	credit_rating	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

จากตารางนี้จะเห็นได้แล้วว่า

Age : youth ถ้าเป็น student

คำตอบจะตอบ yes ถ้าไม่เป็น จะตอบ no

จึงใช้ student node เป็นตัวกำหนด class



สุดท้าย จะทำให้ได้ Decision tree ดังต่อไปนี้

- Resulting tree:

