CHƯƠNG 4: PHÂN LỚP DỮ LIỆU

4.1. KHÁI NIỆM VỀ PHÂN LỚP DỮ LIỆU

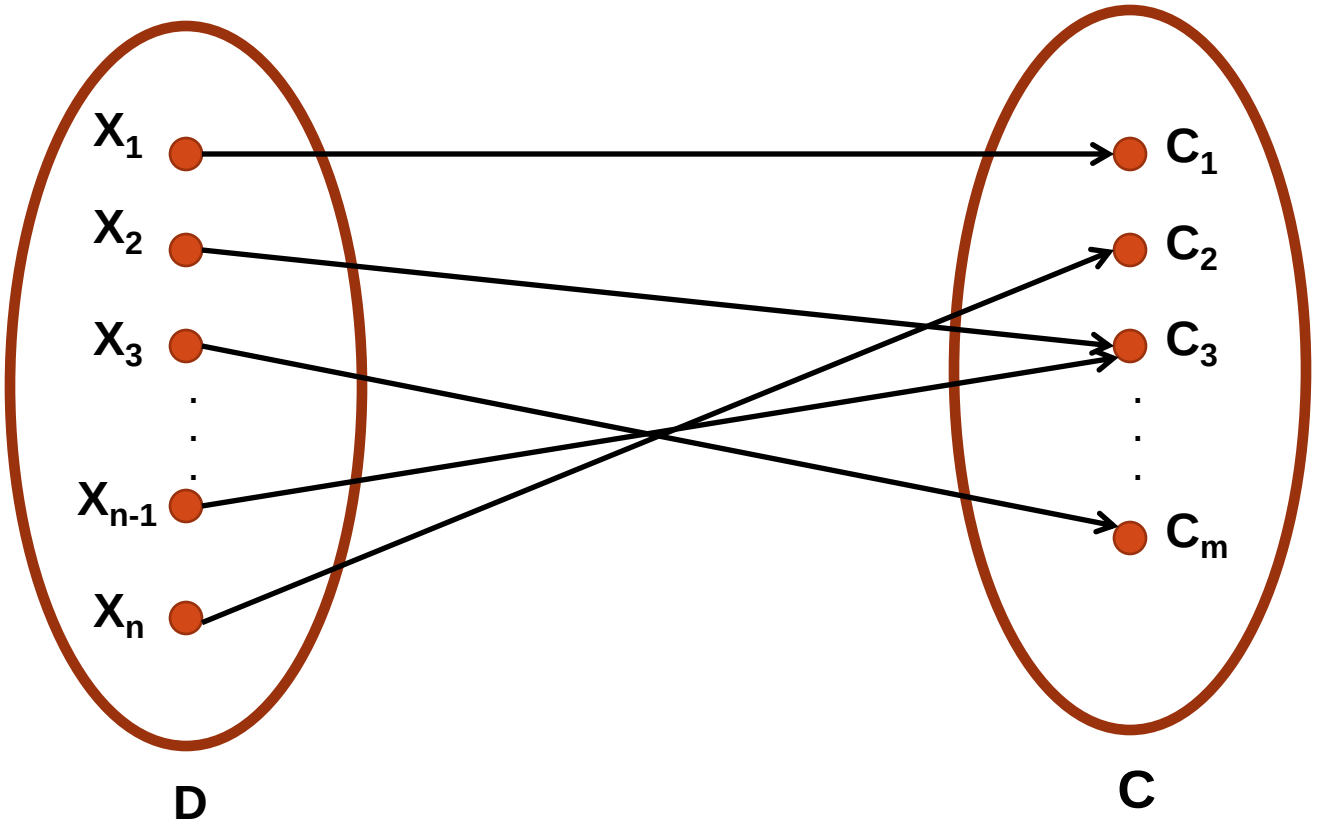
4.2. PHÂN LỚP DỰA TRÊN XÁC SUẤT CÓ ĐIỀU KIỆN
(Phân lớp Bayes – Naive Bayesian Classification)

4.3. PHÂN LỚP DỰA TRÊN CÂY QUYẾT ĐỊNH

4.1. KHÁI NIỆM VỀ PHÂN LỚP DỮ LIỆU

Cho tập các lớp C = {C_1, C_2,…, C_m} và tập dữ liệu D = {X_1, X_2 ,…, X_n}

Phân lớp dữ liệu là sự phân chia các đối tượng dữ liệu vào các lớp. Về bản chất đây quá trình ánh xạ mỗi đối tượng $X_j \in D$ tương ứng với một lớp $C_i \in C$.

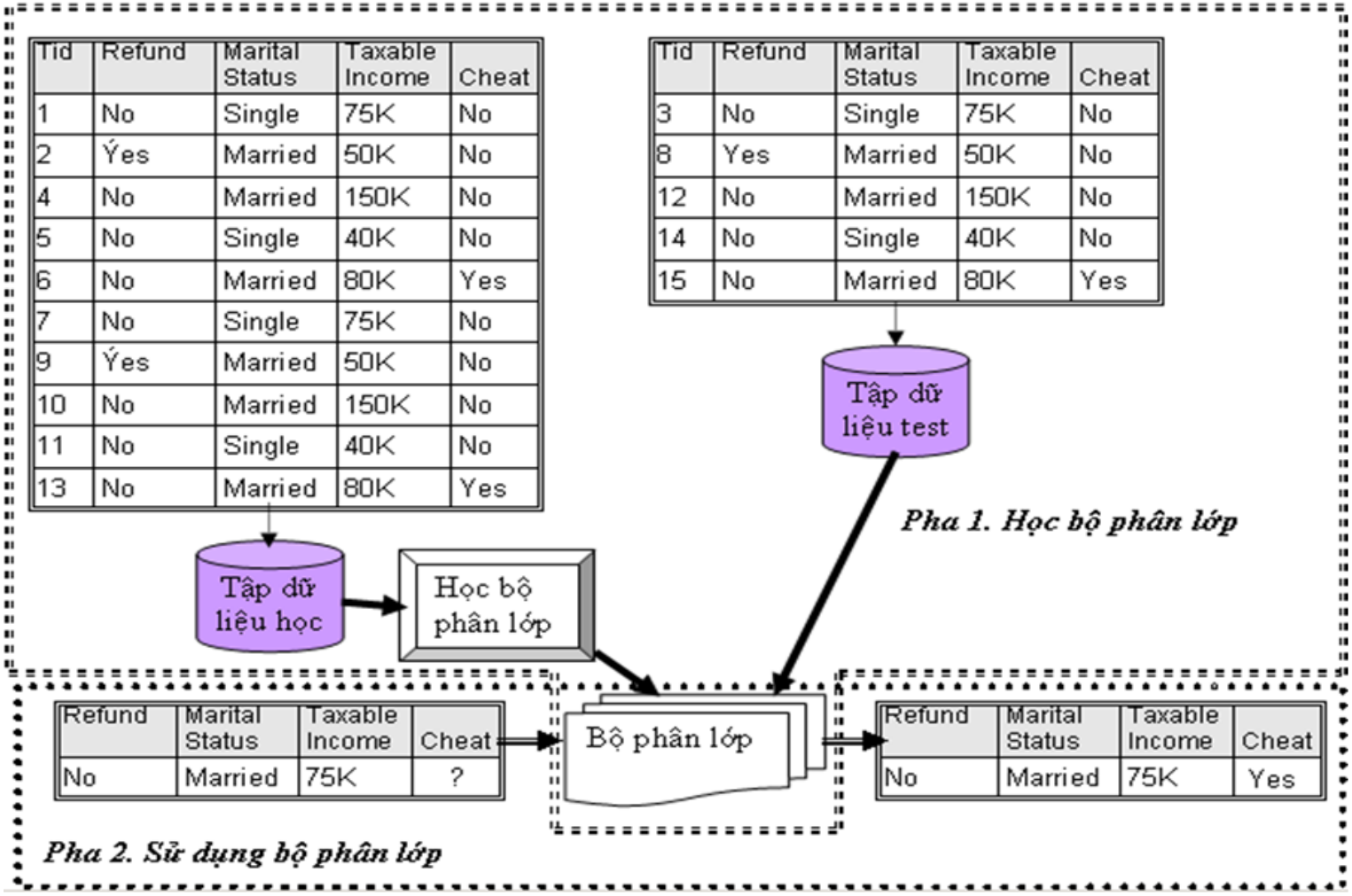


$$f: D \rightarrow C \text{ hay } c = f(X) \text{ (với } X \in D \text{ và } c \in C)$$

- Mỗi ánh xạ được gọi là một mô hình phân lớp (Classification Model).

⟹ ***Làm sao để xây dựng mô hình phân lớp?***

Thông qua quá trình huấn luyện dựa trên tập dữ liệu học (học có giám sát – supervised learning)



Xây dựng mô hình

B1: Chọn một tập ví dụ mẫu (gồm các đối tượng đã được phân lớp):

$D_{exam} = D_1 \cup D_2 \cup \ldots \cup D_m$ trong đó $D_i = \{X | (X \in D) \wedge (X \longrightarrow C_i)\}_{i=1,..,m}$

B2: Tách D_{exam} thành 02 tập:

- Tập dữ liệu học D_{train}
- Tập dữ liệu kiểm tra D_{test}

Hiển nhiên $D_{exam} = D_{train} \cup D_{test}$ và thường thì người ta tách sao cho:

$$|D_{train}| = \frac{2}{3}|D_{exam}| \qquad |D_{test}| = \frac{1}{3}|D_{exam}|$$

B3: Dùng D_{train} để xây dựng mô hình (xác định tham số). Có nhiều loại mô hình phân lớp như: Bayes, cây quyết định, luật phân lớp,…

B4: Dùng D_{test} để kiểm tra, đánh giá mô hình xây dựng được.

B5: Chọn mô hình có chất lượng nhất.

Sử dụng mô hình

Cho $X \in D$ (là tập dữ liệu chưa phân lớp) $\Longrightarrow$ Xác định lớp của X

4.2. PHÂN LỚP DỰA TRÊN XÁC SUẤT CÓ ĐIỀU KIỆN *(Naive Bayes Classifier)*

4.2.1. Xác suất có điều kiện và công thức Bayes

- Gọi X là một bộ dữ liệu (data tuple). Theo ngôn ngữ xác suất, X được xem là một biến cố (evidence).
- Gọi H là một giả thuyết (hypothesis): bộ X thuộc về lớp C_i.

⟹ **Cần xác định P(H|X)**: ***xác suất xảy ra H khi đã xuất hiện X (hay xác suất để X thuộc về lớp C_i nếu như đã biết các thuộc tính của X).***

Nhãn phân lớp

X được xác định thông qua tập giá trị của các thuộc tính

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	medium	no	excellent	no

- **P(H|X) là xác suất có điều kiện của H đối với X (xác xuất xảy ra H khi biết X xảy ra).**

  Ví dụ: *X = (age=35 years old, income=$40,000),*

  H = (buy_computer=Yes)

  P(H|X) = P(by_computer=yes | age=35 years old, income=$40,000)

  ⟹Xác suất để một người 35 tuổi có thu nhập $40,000 mua máy tính

- **P(X|H) là xác suất có điều kiện của X đối với H (xác suất xảy ra X khi biết H xảy ra).**

  ***Ví dụ*:**

  P(X|H) = P(age=35 years old, income=$40,000|buy_computer=yes)

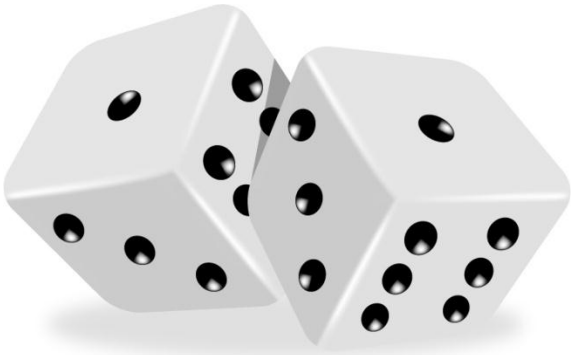
  ⟹ Xác suất để một người mua máy tính có độ tuổi là 35 và thu nhập là $40,000.

- **P(X) là xác suất tiên nghiệm của X.**

  ***Ví dụ*:** *P(X) = P(age=35 years old,income=$40,000) ⟹ Xác suất để tìm thấy trong tập dữ liệu đang xét một người có độ tuổi là 35 và thu nhập là $40,000.*

- **P(H) là xác suất tiên nghiệm của H.**

  ***Ví dụ*:** *P(H) = P(buy_computer=Yes) ⟹ Xác suất mua máy tính của khách hàng nói chung (không quan tâm đến độ tuổi hay thu nhập)*



Thomas Bayes
(1702 – 1761)

Công thức Bayes:

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

T. Bayes.

4.2.2. Phân lớp dữ liệu dựa trên xác suất có điều kiện (phân lớp Bayes)

Bộ phân lớp Bayes hoạt động như sau:

1. *Cho D là tập dữ liệu học gồm các bộ và nhãn lớp tương ứng (đã được phân lớp). Mỗi bộ được biểu diễn bởi một vector n chiều $X = (x_1, x_2, \ldots, x_n)$ trong đó x_i là giá trị tương ứng với thuộc tính A_i ($i = 1, 2, \ldots, n$). Tập $D_i = \{X | (X \in D) \wedge (X \rightarrow C_i)\}$ là tập các bộ trong D thuộc về lớp C_i.*
2. *Giả sử có m lớp $C_1, C_2, \ldots, C_m$. Bộ X được dự đoán là thuộc về lớp C_i khi và chỉ khi: $P(C_i|X) > P(C_j|X)$ với mọi $j \neq i$ và $1 \leq j \leq m$ (X thuộc về lớp mà xác suất có điều kiện khi biết X là lớn nhất) $\Longrightarrow$ Đi tìm lớp C_i trong số m lớp sao cho $P(C_i|X)$ là lớn nhất.*
3. *P(X) là giống nhau với tất cả các lớp nên theo công thức Bayes thì $P(C_i|X)$ lớn nhất tương ứng với tích $P(X|C_i)P(C_i)$ lớn nhất $\Longrightarrow$ Đi tìm C_i sao cho tích $P(X|C_i)P(C_i)$ là lớn nhất ($i = 1, 2, \ldots, m$).*
4. *Ta có thể tính:*

$$P(C_i) = \frac{|D_i|}{|D|}$$

và nếu coi n thuộc tính của X là độc lập thì:

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) P(x_2 \mid C_i) \ldots P(x_n \mid C_i)$$

$$P(x_k \mid C_i) = \frac{\left|\{X' \mid (X'(A_k) = x_k) \Lambda (X' \in D_i)\}\right|}{|D_i|}$$

Chú ý:

- **Nếu không tính được $P(C_i)$ thì có thể coi $P(C_1) = P(C_2) = \ldots = P(C_m)$ và bài toán quy về tìm lớp C_i trong số m lớp sao cho $P(X|C_i)$ có giá tri lớn nhất.**
- **Nếu tồn tại $P(x_k|C_i) = 0$ thì có thể áp dụng hiệu chỉnh Lapace và công thức tính của $P(x_k|C_i)$ được hiệu chỉnh như sau:**

$$P(x_k \mid C_i) = \frac{\left|\left\{X' \mid (X'(A_k) = x_k) \Lambda (X' \in D_i)\right\}\right| + 1}{\left|D_i\right| + q}$$

q: số giá trị khác nhau của A_k

Ví dụ:

Cho tập dữ liệu học gồm các bộ dữ liệu đã được phân lớp như sau:

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	medium	no	excellent	no

Áp dụng phân lớp Bayes hãy dự đoán bộ dữ liệu

$$X = (age = youth, income = medium, student = yes, credit\_rating = fair)$$

thuộc lớp nào?

Có 02 lớp dữ liệu tương ứng với buys_computer = yes và buys_computer = no

$$P(buys\_computer = yes) = 9/14 = 0.643$$
$$P(buys\_computer = no) = 5/14 = 0.357$$

$$P(age = youth \mid buys\_computer = yes) = 2/9 = 0.222$$
$$P(age = youth \mid buys\_computer = no) = 3/5 = 0.600$$
$$P(income = medium \mid buys\_computer = yes) = 4/9 = 0.444$$
$$P(income = medium \mid buys\_computer = no) = 2/5 = 0.400$$
$$P(student = yes \mid buys\_computer = yes) = 6/9 = 0.667$$
$$P(student = yes \mid buys\_computer = no) = 1/5 = 0.200$$
$$P(credit\_rating = fair \mid buys\_computer = yes) = 6/9 = 0.667$$
$$P(credit\_rating = fair \mid buys\_computer = no) = 2/5 = 0.400$$

Suy ra:
$$P(X|buys\_computer = yes) = P(age = youth \mid buys\_computer = yes) \times P(income = medium \mid buys\_computer = yes) \times P(student = yes \mid buys\_computer = yes) \times P(credit\_rating = fair \mid buys\_computer = yes)$$
$$= 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044.$$

Tương tự: $P(X|buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019$

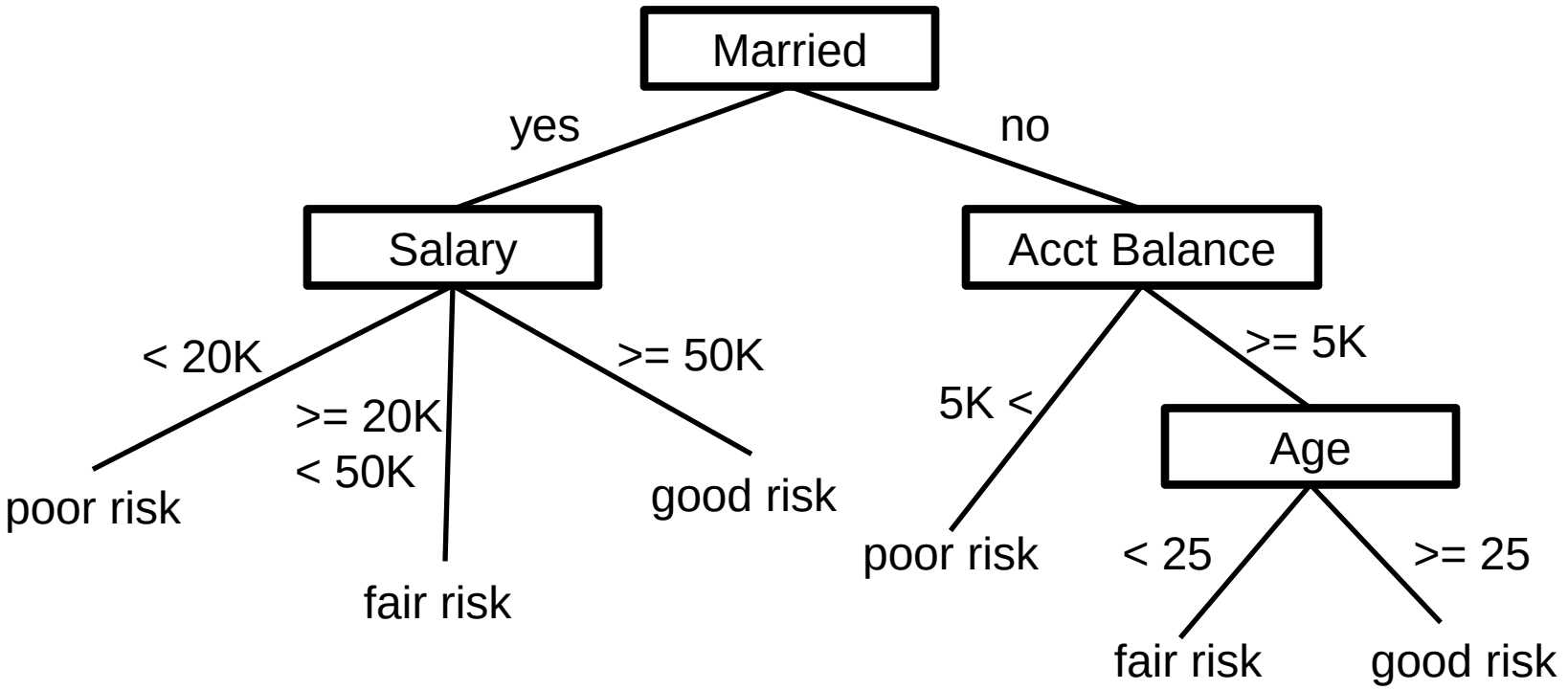
$$P(X|buys\_computer = yes)P(buys\_computer = yes) = 0.044 \times 0.643 = 0.028$$
$$P(X|buys\_computer = no)P(buys\_computer = no) = 0.019 \times 0.357 = 0.007$$

$\Longrightarrow$X thuộc lớp dữ liệu tương ứng với buys_computer = yes

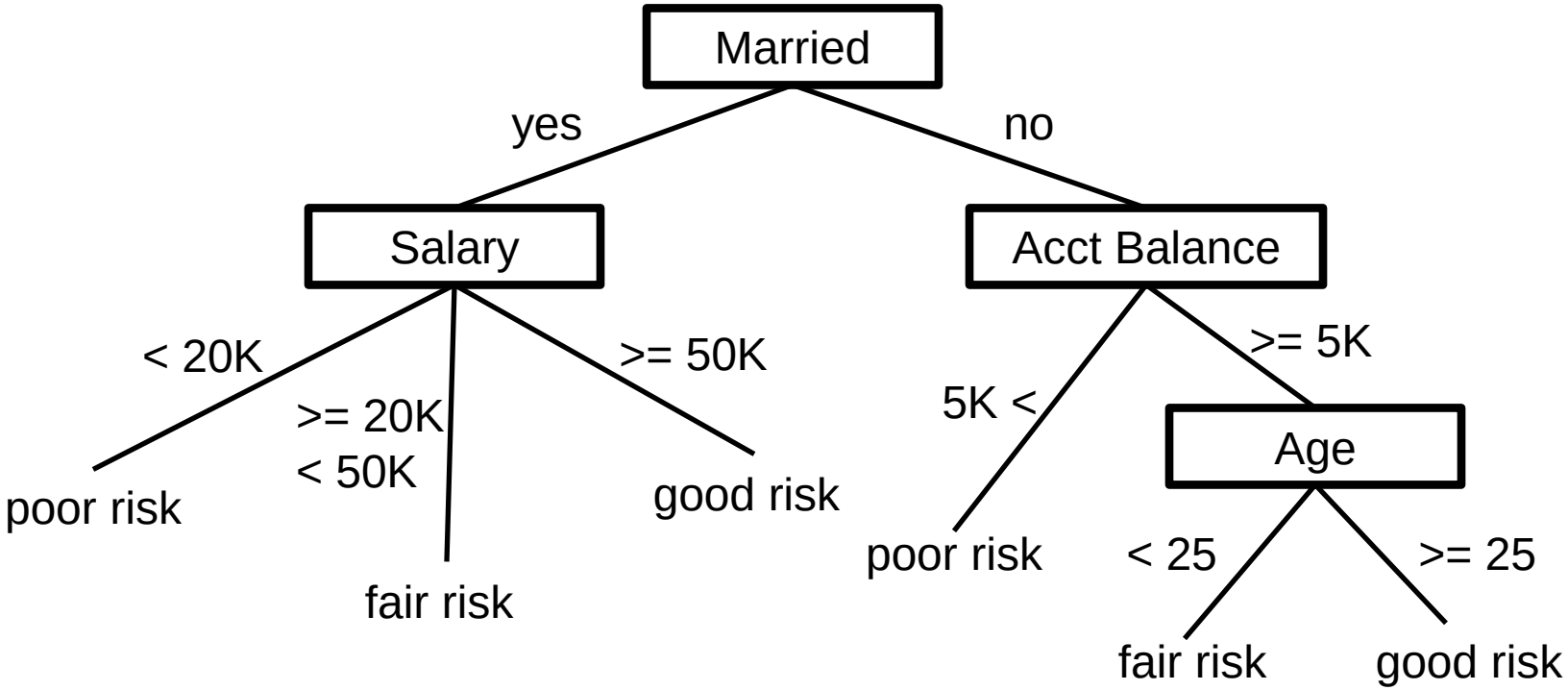
4.3. PHÂN LỚP DỰA TRÊN CÂY QUYẾT ĐỊNH

4.3.1. Mô hình phân lớp cây quyết định

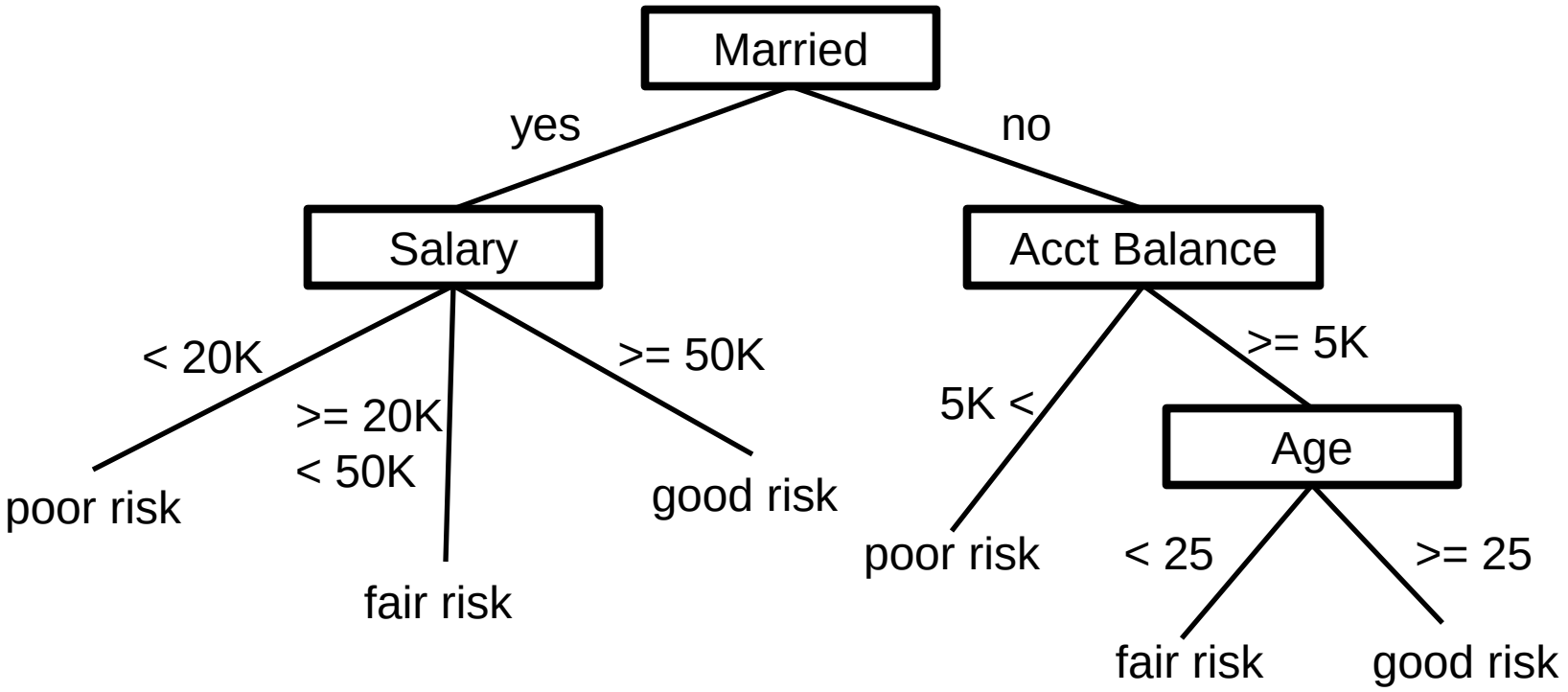
- Cây quyết định (decision tree) là một mô hình phân lớp điển hình.
- Cây quyết định bao gồm:
  - **Các nút trong**: biểu diễn cho một thuộc tính được kiểm thử (test).
  - **Các nút lá**: nhãn/mô tả của một lớp (class label).
  - **Nhánh**: xuất phát từ một nút trong, phản ánh kết quả của một phép thử trên thuộc tính tương ứng.



- Có thể dễ dàng chuyển đổi từ mô hình **cây quyết định** sang mô hình **luật phân lớp** bằng cách: đi từ nút gốc cho tới nút lá, mỗi đường đi tương ứng với một luật phân lớp.



1. **If** (Married = yes) And (Salary > 20K) **Then** Class = poor risk
2. **If** (Married = yes) And (50K > Salary >= 20K) **Then** Class = fair risk
3. **If** (Married = yes) And (Salary >= 50K) **Then** Class = good risk
4. **If** (Married = no) And (Acct Balance < 5K) **Then** Class = poor risk
5. **If** (Married = no) And (Acct Balance >= 5K) And (Age < 25) **Then** Class = fair risk
6. **If** (Married = no) And (Acct Balance >= 5K) And (Age >= 25) **Then** Class = good risk



Name	Age	Married	Salary	Acct Balance	Class
Alice	19	yes	30K	6K	?
Pike	28	no	60K	7K	?
Tom	35	yes	10K	10K	?
Peter	24	no	20K	8K	?
Lucas	40	no	20K	3K	?

Name	Age	Married	Salary	Acct Balance	Class
Alice	19	yes	30K	6K	fair risk
Pike	28	no	60K	7K	good risk
Tom	35	yes	10K	10K	poor risk
Peter	24	no	20K	8K	fair risk
Lucas	40	no	20K	3K	poor risk

4.3.2. Các độ đo sử dụng trong phân lớp

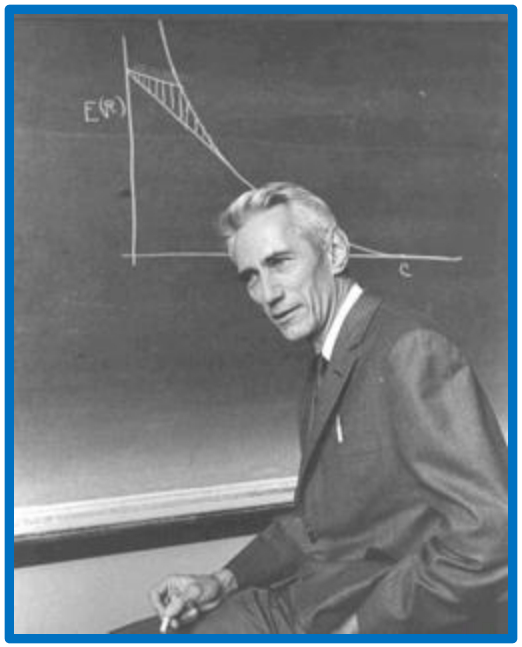
A. Entropy của tập dữ liệu

Là lượng thông tin cần để phân loại một phần tử trong tập dữ liệu D. Ký hiệu là Infor(D).

Gọi:

p_i: xác suất để một phần tử bất kỳ trong D thuộc về lớp C_i (i=1, 2,…, m).

D_i: Tập các phần tử trong D thuộc về lớp C_i.



Claude Elwood Shannon (1916 – 2001)

$$p_i = \frac{|D_i|}{|D|}$$

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

B. Entropy của dữ liệu ứng với một thuộc tính

Là lượng thông tin cần để phân loại một phần tử trong tập dữ liệu D dựa trên thuộc tính A. Ký hiệu là $Infor_A(D)$.

- **Thuộc tính A dùng để phân tách D thành v phân hoạch (tập con) là $D_1, D_2, \ldots, D_v$.**
- **Mỗi phân hoạch D_j có $|D_j|$ phần tử.**
- **Lượng thông tin này sẽ cho biết mức độ trùng lặp giữa các phân hoạch, nghĩa là một phân hoạchchứa các phần tử từ một hay nhiều lớp khác nhau.**

$\Longrightarrow$ Mong đợi: $Infor_A(D)$ càng nhỏ càng tốt.

$$Infor_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Infor(D_j)$$

C. Độ lợi thông tin (Information Gain)

- ***Mục tiêu: Tối thiểu hóa lượng thông tin cần thiết để phân lớp các các mẫu dữ liệu (tối thiểu hóa số lượng các điều kiện kiểm tra cần thiết để phân lớp một bản ghi mới).***

Độ lợi thông tin ứng với thuộc tính A (*ký hiệu Gain(A)*) chính là *độ sai biệt* giữa Entropy ban đầu của tập dữ liệu (trước phân hoạch) và Entropy của dữ liệu ứng với thuộc tính A (sau khi phân hoạch bởi A).

$$Gain(A) = Infor(D) - Infor_A(D)$$

4.3.3. Giải thuật ID3 xây dựng cây quyết định

Input: Tập dữ liệu học Records gồm m đối tượng (bản ghi) $R_1, R_2, \ldots, R_m$.

Tập thuộc tính Attributes gồm m thuộc tính $A_1, A_2, \ldots, A_n$.

Output: Mô hình cây quyết định.

procedure Build_tree(Records, Attributes)
begin
Tạo nút N;
if (tất cả các bản ghi thuộc về một lớp C_i nào đó) **then**
begin
N.Label = C_i;
return N;
end;
if (Attributes = $\varnothing$) **then**
begin
Tìm lớp C_j mà phần lớn các bản ghi r ∈ Records thuộc về lớp đó.
N.Label = C_j;
return N;
end;
Chọn A_i ∈ Attribute sao cho Gain(A_i)→max;
N.Label = A_i;
for each giá trị v_i đã biết của A_i **do**
begin
Thêm một nhánh mới vào nút N ứng với $A_i = v_j$;
S_j = Tập con của Records có $A_i = v_j$;
if (S_j = $\varnothing$) **then**
Thêm một nút lá L với nhãn là lớp mà phần lớn các bản ghi r ∈ Records thuộc về lớp đó;
Return L;
else
Thêm vào nút được trả về bởi Build_Tree(S_j, Attribute \{A_i});
end ;
end;

Phương pháp lựa chọn thuộc tính

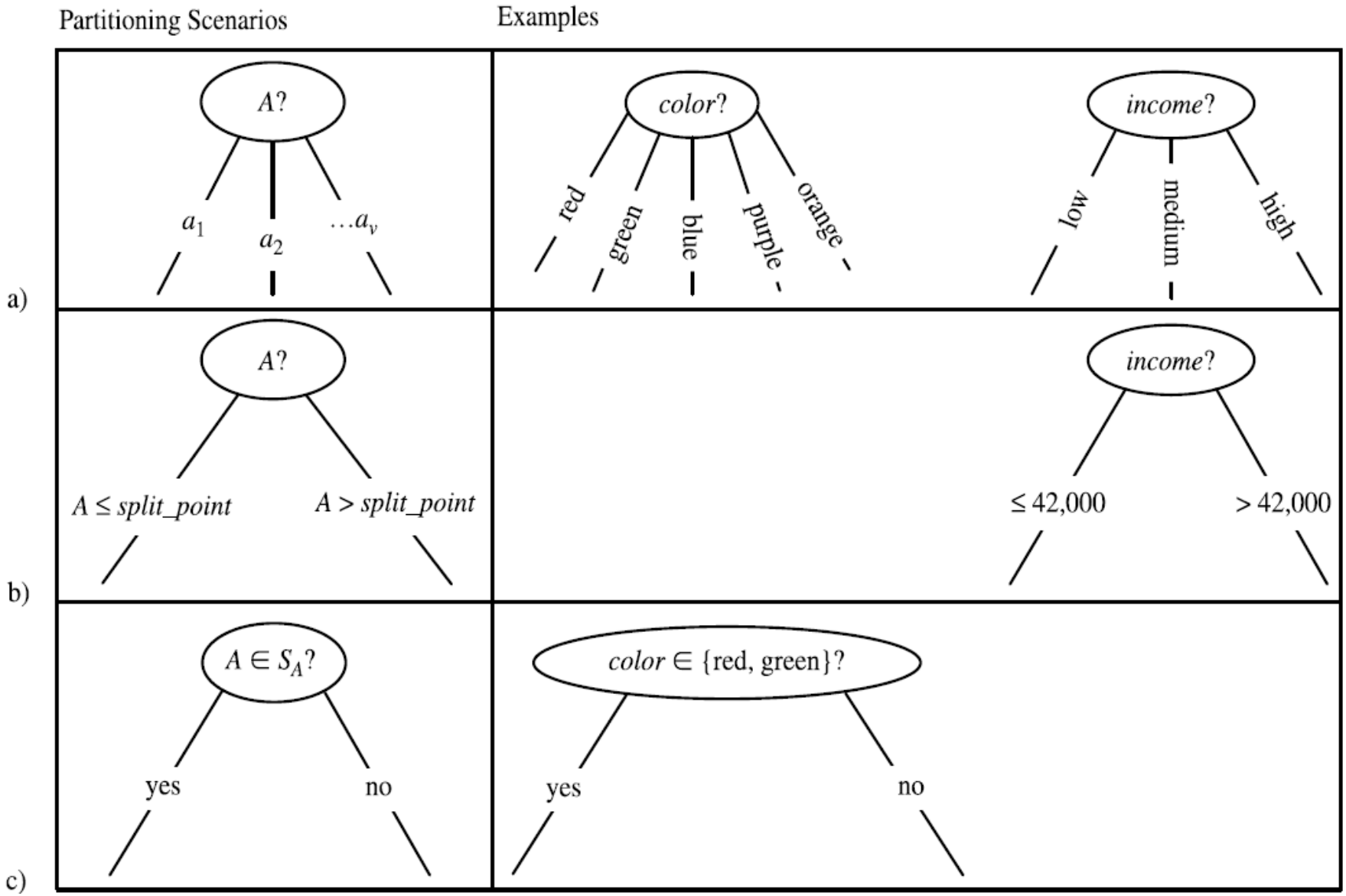
Dùng heuristic để chọn tiêu chí rẽ nhánh tại một nút: Phân hoạch tập dữ liệu học D thành các phân hoạch con với các nhãn phù hợp:

- Xếp hạng mỗi thuộc tính.
- Thuộc tính được chọn để rẽ nhánh là thuộc tính có trị số điểm (score) là lớn nhất.
- Độ đo để chọn thuộc tính phân tách (splitting attribute) là Information Gain (được xây dựng dựa trên lý thuyết thông tin của Claude Elwood Shannon).

Cụ thể: **Thuộc tính có giá trị Information Gain lớn nhất sẽ được chọn làm thuộc tính phân nhánh cho nút N.**

- ❖ Nút N là nút hiện tại cần phân hoạch các phần tử trong D.
- ❖ Thuộc tính phân hoạch đảm bảo sự trùng lắp ngẫu nhiên ít nhất giữa các phân hoạch tạo được.

⟹ ***Giúp tối thiểu số phép thử (test) cần để phân loại một phần tử.***


Partitioning Scenarios
Examples
a)
A?
a_1
a_2
$\ldots a_v$
color?
red
green
blue
purple
orange
income?
low
medium
high
b)
A?
$A \leq split\_point$
$A > split\_point$
income?
≤ 42,000
> 42,000
c)
$A \in S_A$?
yes
no
color ∈ {red, green}?
yes
no

***Ví dụ 1*: Cho tập dữ liệu học:**

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	medium	no	excellent	no

Tính toán tương tự:

$$Gain(income) = 0.029 \text{ bits}$$

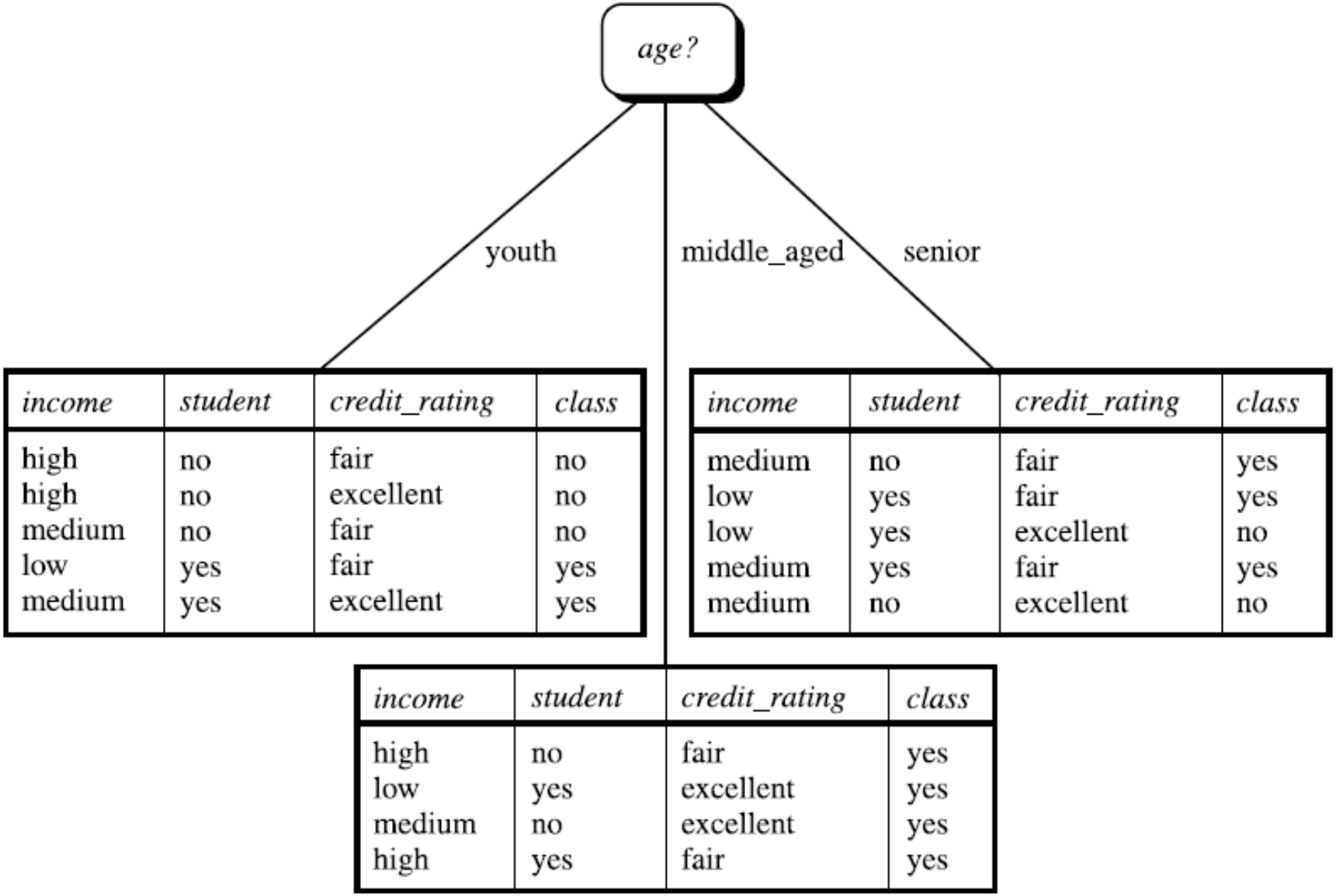
$$Gain(student) = 0.151 \text{ bits}$$

$$Gain(credit\_rating) = 0.048 \text{ bits}$$

⟹ Chọn age là thuộc tính phân tách

$$Info_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) = 0.694 \text{ bits.}$$

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



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

Q & A