



ĐẠI HỌC QUỐC GIA HÀ NỘI  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

---

# KHAI PHÁ DỮ LIỆU

## Phân lớp với Cây quyết định

---

**Phan Xuân Hiếu**

Bộ môn CHTTT & KTLab,  
Khoa Công nghệ thông tin,  
Trường Đại học Công nghệ, ĐHQG HN  
Email: [hieupx@vnu.edu.vn](mailto:hieupx@vnu.edu.vn)  
URL: <http://uet.vnu.edu.vn/~hieupx>

# Nội dung bài giảng

- Cơ bản về lý thuyết thông tin (information theory)
  - Entropy
  - Entropy điều kiện (conditional entropy)
  - Information gain
- Cây quyết định (decision tree)
  - Mô hình cây quyết định
  - Thuật toán ID3
  - Ví dụ minh họa
  - Thuật toán C4.5
- Kết luận
  - Nhược điểm
  - Ưu điểm

# Nội dung bài giảng

## ■ Cơ bản về lý thuyết thông tin (information theory)

- Entropy
- Entropy điều kiện (conditional entropy)
- Information gain

## ■ Cây quyết định (decision tree)

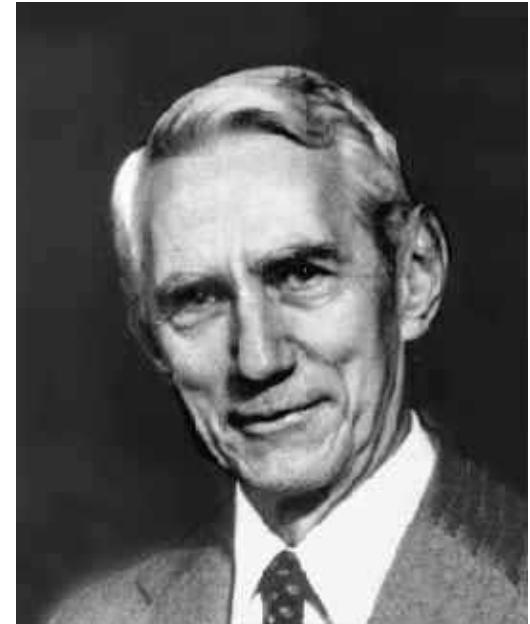
- Mô hình cây quyết định
- Thuật toán ID3
- Ví dụ minh họa
- Thuật toán C4.5

## ■ Kết luận

- Nhược điểm
- Ưu điểm

# Lý thuyết thông tin (information theory)

- Lý thuyết thông tin là một phần của
  - Toán ứng dụng (applied mathematics)
  - Kỹ nghệ điện-điện tử
  - Khoa học máy tính & kỹ nghệ truyền thông
- Có vai trò quan trọng trong:
  - Lượng hóa thông tin
  - Xử lý tín hiệu, truyền thông tin
  - Nén dữ liệu (zip; mp3, jpeg, ...)
  - Mã hóa kênh truyền
  - Suy diễn thống kê
  - Xử lý ngôn ngữ tự nhiên
  - Trò chơi & cá cược
  - .v.v.



**Claude Shannon**  
(1916-2001)

*A Mathematical Theory  
of Communication*, 1948  
Bell Labs, MIT, IAS

# Ví dụ về truyền thông tin (theo bit)

- Quan sát một biến ngẫu nhiên  $X$  nhận 1 trong 4 giá trị A, B, C, D:
  - $P(X = A) = P(X = B) = P(X = C) = P(X = D) = 0.25$
  - Ví dụ một chuỗi thông tin theo  $P(X)$ : BAACBADCDADDDA...
- Truyền dãy dữ liệu trên qua một kênh truyền mã hóa nhị phân (bit)
  - $A = 00, B = 01, C = 10, D = 11$
  - Dữ liệu: 0100001001001110110011111100...
- Cần trung bình bao nhiêu bit để truyền một đơn vị thông tin theo phân phối  $P(X)$  như trên?
  - 2 bit
- Số lượng bit trung bình nhỏ nhất để truyền dữ liệu có phụ thuộc vào phân phối của dữ liệu không?
  - Có

# Ví dụ: một phân phối khác

## ■ Ví dụ, $P(X)$ bây giờ:

- $P(X = A) = 0.5$
- $P(X = B) = 0.3$
- $P(X = C) = 0.125$
- $P(X = D) = 0.075$

## ■ Mã hóa:

- A: 0, B: 10, C: 110, D: 111

## ■ Số lượng bit trung bình để truyền dữ liệu theo $P(X)$ :

- $1 \times 0.5 + 2 \times 0.3 + 3 \times 0.125 + 3 \times 0.075 = 1.7$  (bit)
- Xem xét biểu thức và giá trị sau?

Con số này có  
ý nghĩa gì?

$$-(0.5 \log_2 0.5 + 0.3 \log_2 0.3 + 0.125 \log_2 0.125 + 0.075 \log_2 0.075) = \mathbf{1.67636}$$

# Nội dung bài giảng

## ■ Cơ bản về lý thuyết thông tin (information theory)

- Entropy
- Entropy điều kiện (conditional entropy)
- Information gain

## ■ Cây quyết định (decision tree)

- Mô hình cây quyết định
- Thuật toán ID3
- Ví dụ minh họa
- Thuật toán C4.5

## ■ Kết luận

- Nhược điểm
- Ưu điểm

# Entropy

## ■ Biến ngẫu nhiên X:

- Có không gian mẫu  $\Omega = \{x_1, x_2, \dots, x_m\}$  với xác suất
- $P(X = x_1) = p_1, P(X = x_2) = p_2, \dots, P(X = x_m) = p_m$

## ■ Số bit trung bình nhỏ nhất để truyền một đơn vị dữ liệu theo phân phối P(X):

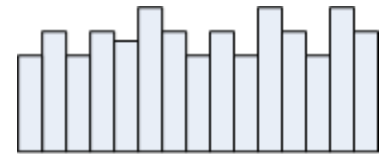
$$\begin{aligned} H(X) &= -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \dots - p_m \log_2 p_m \\ &= -\sum_{i=1}^m p_i \log_2 p_i \end{aligned}$$

## ■ H(X) là **entropy** của X

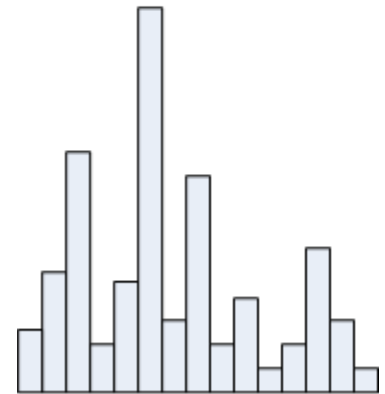
$$0 \leq H(X) \leq \log_2 m$$

## ■ Giá trị entropy:

- Lớn: phân phối P(X) gần dạng phân phối đồng nhất (uniform distribution)
- Nhỏ: phân phối P(X) xa dạng phân phối đồng nhất



Entropy lớn



Entropy nhỏ



# Entropy: một cách trực giác



# Nội dung bài giảng

## ■ Cơ bản về lý thuyết thông tin (information theory)

- Entropy
- **Entropy điều kiện (conditional entropy)**
- Information gain

## ■ Cây quyết định (decision tree)

- Mô hình cây quyết định
- Thuật toán ID3
- Ví dụ minh họa
- Thuật toán C4.5

## ■ Kết luận

- Nhược điểm
- Ưu điểm

# Entropy với điều kiện cụ thể: $H(Y|X = v)$

X	Y
Ngành	Thích chơi game
Toán	Có
Lịch sử	Không
CNTT	Có
Toán	Không
Toán	Không
CNTT	Có
Lịch sử	Không
Toán	Có

Quy ước:

$$0\log_2 0 = 0$$

- Từ bảng dữ liệu bên ta có:

- $P(Y = \text{Có}) = 0.5$
- $P(X = \text{Toán} \ \& \ Y = \text{Không}) = 0.25$
- $P(X = \text{Toán}) = 0.5$
- $P(Y = \text{Có} \mid X = \text{Lịch sử}) = 0$

- Các entropy  $H(X)$  và  $H(Y)$ :

- $H(X) = -0.5\log_2 0.5 - 0.25\log_2 0.25 - 0.25\log_2 0.25 = 1.5$
- $H(Y) = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$

- Entropy điều kiện:  $H(Y|X = v)$ :

- $H(Y|X = v)$ : là entropy của Y đối với những hàng dữ liệu mà ở đó giá trị của X là v.
- $H(Y|X = \text{Toán}) = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$
- $H(Y|X = \text{Lịch sử}) = -0\log_2 0 - 1\log_2 1 = 0$
- $H(Y|X = \text{CNTT}) = -1\log_2 1 - 0\log_2 0 = 0$

# Entropy điều kiện: $H(Y|X)$

- $H(Y|X)$  = trung bình các giá trị entropy điều kiện cụ thể  $H(Y|X = v)$

Bảng số bit trung bình nhỏ nhất để truyền Y nếu hai bên (gửi và nhận) đều biết X

$$H(Y|X) = \sum_i P(X = v_i) H(Y|X = v_i)$$

X	Y
Ngành	Thích chơi game
Toán	Có
Lịch sử	Không
CNTT	Có
Toán	Không
Toán	Không
CNTT	Có
Lịch sử	Không
Toán	Có

$v_i$	$P(X = v_i)$	$H(Y X = v_i)$
Toán	0.5	1
Lịch sử	0.25	0
CNTT	0.25	0

- Do đó:

$$H(Y|X) = 0.5 * 1 + 0.25 * 0 + 0.25 * 0 = \mathbf{0.5}$$

# Nội dung bài giảng

## ■ Cơ bản về lý thuyết thông tin (information theory)

- Entropy
- Entropy điều kiện (conditional entropy)
- **Information gain**

## ■ Cây quyết định (decision tree)

- Mô hình cây quyết định
- Thuật toán ID3
- Ví dụ minh họa
- Thuật toán C4.5

## ■ Kết luận

- Nhược điểm
- Ưu điểm

# Information Gain: $IG(Y|X)$

X	Y
Ngành	Thích chơi game
Toán	Có
Lịch sử	Không
CNTT	Có
Toán	Không
Toán	Không
CNTT	Có
Lịch sử	Không
Toán	Có

- **Định nghĩa:**

$IG(Y|X)$  là số lượng bit trung bình có thể tiết kiệm khi truyền Y mà hai đầu (gửi và nhận) đã biết X.

- **Như vậy:**

$$IG(Y|X) = H(Y) - H(Y|X)$$

- **Ví dụ (dữ liệu hình bên):**










$$H(Y) = 1$$

$$H(Y|X) = 0.5$$

$$IG(Y|X) = 1 - 0.5 = \mathbf{0.5}$$

# Một ví dụ khác về Information Gain

wealth values: poor rich

agegroup	10s	2507	3		$H(\text{wealth} \mid \text{agegroup} = 10s) = 0.0133271$
	20s	11262	743		$H(\text{wealth} \mid \text{agegroup} = 20s) = 0.334906$
	30s	9468	3461		$H(\text{wealth} \mid \text{agegroup} = 30s) = 0.838134$
	40s	6738	3986		$H(\text{wealth} \mid \text{agegroup} = 40s) = 0.951961$
	50s	4110	2509		$H(\text{wealth} \mid \text{agegroup} = 50s) = 0.957376$
	60s	2245	809		$H(\text{wealth} \mid \text{agegroup} = 60s) = 0.834049$
	70s	668	147		$H(\text{wealth} \mid \text{agegroup} = 70s) = 0.680882$
	80s	115	16		$H(\text{wealth} \mid \text{agegroup} = 80s) = 0.535474$
	90s	42	13		$H(\text{wealth} \mid \text{agegroup} = 90s) = 0.788941$

$H(\text{wealth}) = 0.793844$   $H(\text{wealth} \mid \text{agegroup}) = 0.709463$

$IG(\text{wealth} \mid \text{agegroup}) = 0.0843813$

Y (Tài sản)

X (Độ tuổi)

# Nội dung bài giảng

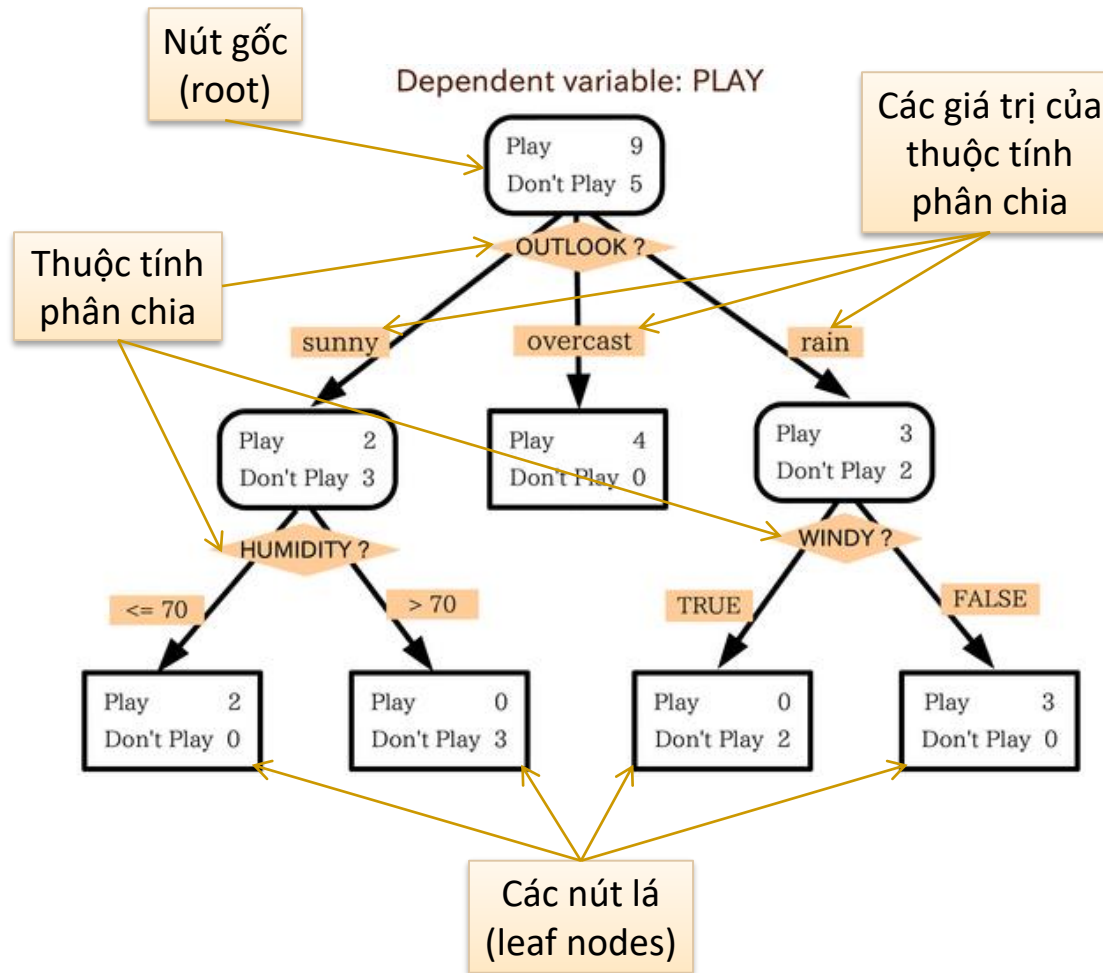
- Cơ bản về lý thuyết thông tin (information theory)
  - Entropy
  - Entropy điều kiện (conditional entropy)
  - Information gain
- **Cây quyết định (decision tree)**
  - **Mô hình cây quyết định**
  - Thuật toán ID3
  - Ví dụ minh họa
  - Thuật toán C4.5
- Kết luận
  - Nhược điểm
  - Ưu điểm



# Phát biểu lại bài toán phân lớp

- Phát biểu lại bài toán phân lớp:
  - $\mathbf{C} = \{c_1, c_2, \dots, c_K\}$ : tập K lớp
  - $\mathbf{X} = \{\mathbf{x}_i\} (i=1,2,\dots)$  là không gian các đối tượng cần phân lớp
  - Xây dựng một ánh xạ  $f: \mathbf{X} \rightarrow \mathbf{C}$
  - Ánh xạ  $f$  được gọi là mô hình phân lớp (classification model, classifier)
- Xây dựng mô hình  $f$  bằng học giám sát
  - $\mathbf{D} = \{(\mathbf{x}^1, c^1), (\mathbf{x}^2, c^2), \dots, (\mathbf{x}^N, c^N)\}$  trong đó  $\mathbf{x}^n \in \mathbf{X}$ ,  $c^n \in \mathbf{C}$  là tập dữ liệu huấn luyện (training data)
  - Huấn luyện mô hình  $f$  dựa trên tập huấn luyện  $\mathbf{D}$  sao cho  $f$  phân lớp chính xác nhất có thể
- Mô hình  $f$  có thể xây dựng theo:
  - Phương pháp Naïve Bayes
  - Phương pháp cây quyết định (decision tree) ← **trong phần này**
  - Phương pháp cực đại hóa entropy (maximum entropy classification)
  - Phương pháp máy vector hỗ trợ (support vector machines)
  - .v.v.

# Một hình cây quyết định (decision tree)



**Ross Quinlan**  
*Induction of Decision Trees*  
Machine Learning 1986

# Cấu trúc cây & các tính chất

- Nốt gốc của cây (root node):
  - Một thuộc tính điều kiện sẽ được chọn làm nốt gốc
  - Các nhánh từ nốt gốc tương ứng với các giá trị có thể của thuộc tính này
  - Nốt gốc bao hàm toàn bộ các đối tượng trong dữ liệu huấn luyện **D**
- Các nốt trong (internal nodes):
  - Mỗi nốt trong của cây có thể xem là nốt gốc của một cây con (sub-tree)
  - Mỗi nốt trong cũng tương ứng với một thuộc tính điều kiện
  - Mỗi nốt trong chỉ bao hàm những đối tượng dữ liệu thuộc một nhánh cụ thể của nốt cha.
- Các nốt lá (leaf nodes):
  - Là nốt cuối trong nhánh mà tất cả các đối tượng đều thuộc một lớp, hoặc
  - Không còn thuộc tính điều kiện nào để phân chia, hoặc
  - Không còn đối tượng dữ liệu nào để phân chia
- Cây được xây dựng theo cách chia để trị và đệ quy từ trên xuống  
**Top-down recursive divide-and-conquer manner**

# Nội dung bài giảng

- Cơ bản về lý thuyết thông tin (information theory)
  - Entropy
  - Entropy điều kiện (conditional entropy)
  - Information gain
- Cây quyết định (decision tree)
  - Mô hình cây quyết định
  - **Thuật toán ID3**
  - Ví dụ minh họa
  - Thuật toán C4.5
- Kết luận
  - Nhược điểm
  - Ưu điểm

# Các ký hiệu & thuật ngữ

- Tập dữ liệu huấn luyện:
  - $C = \{c_1, c_2, \dots, c_K\}$ : tập K lớp
  - $D = \{(\mathbf{x}^1, c^1), (\mathbf{x}^2, c^2), \dots, (\mathbf{x}^N, c^N)\}$  trong đó  $\mathbf{x}^n \in \mathbf{X}$ ,  $c^n \in C$
- Các đối tượng cần phân lớp  $x$  được biểu diễn bởi  $M$  thuộc tính
  - $\mathbf{F} = \{F_1, F_2, \dots, F_M\}$
  - Mỗi thuộc tính  $F_i \in \mathbf{F}$  có miền xác định  $\mathbf{V}^i = \{v_{1}^i, v_{2}^i, \dots, v_{p_i}^i\}$
  - Nếu một thuộc tính  $F_i$  nào đó có miền xác định liên tục,  $F_i$  trước tiên cần được rời rạc hóa để có miền giá trị như  $\mathbf{V}^i$  như trên.
- Các thuật ngữ:
  - $C$  (tập các lớp) được gọi là thuộc tính phân loại (hoặc thuộc tính đích – target attribute)
  - Các thuộc tính trong  $\mathbf{F}$  được gọi là các thuộc tính điều kiện (thuộc tính dùng để phân lớp)

# Dữ liệu minh họa

Các thuộc tính điều kiện (dùng để phân loại)

Thuộc tính cần  
phân loại  
(target attribute)

ID	Age	Income	Student	Credit_rating	Buys_computer
1	≤30	High	No	Fair	No
2	≤30	High	No	Excellent	No
3	31..40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31..40	Low	Yes	Excellent	Yes
8	≤30	Medium	No	Fair	No
9	≤30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	≤30	Medium	Yes	Excellent	Yes
12	31..40	Medium	No	Excellent	Yes
13	31..40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

$F = \{\text{Age, Income, Student, Credit\_rating}\}$ ,  $C = \text{Buys\_computer} = \{\text{Yes, No}\}$

# Thuật toán ID3

## Input:

- Examples =  $D$ , tập toàn bộ dữ liệu huấn luyện
- $C = \{c_1, c_2, \dots, c_k\}$ : tập  $K$  lớp và là thuộc tính đích
- Attributes =  $F$ , tập toàn bộ các thuộc tính điều kiện

## ID3(Examples, C, Attributes)

- Tạo nốt gốc (Root) cho cây.
- Nếu tất cả đối tượng  $x \in \text{Examples}$  có cùng một lớp  $c_k$ , trả về nốt gốc Root với nhãn  $c_k$ .
- Nếu không còn thuộc tính điều kiện nào (Attributes =  $\emptyset$ ), trả về nốt gốc Root với nhãn  $c_k$  nào đó xuất hiện nhiều nhất trong Examples.
- Nếu không thì (else):
  - Chọn  $F_i \in \text{Attributes}$  là thuộc tính **phân lớp tốt nhất** cho tập Examples làm nốt gốc Root.
  - Đối với mỗi giá trị  $v_j^i (\in V^i)$  của  $F_i$ :
    - Thêm một nhánh dưới nốt gốc Root tương ứng với  $F_i = v_j^i$ .
    - Examples( $v_j^i$ ) là tập các đối tượng thuộc Examples có  $F_i = v_j^i$ .
    - Nếu Examples( $v_j^i$ ) =  $\emptyset$ : thêm một nốt lá (leaf node) dưới nhánh này với nhãn  $c_k$  nào đó phổ biến nhất trong Examples.

Ngược lại (else): dưới nhánh này thêm một cây con **ID3**(Examples( $v_j^i$ ), C, Attributes – { $F_i$ })
- Trả về nốt gốc Root.

Đánh giá bằng tiêu chí nào?  
Information Gain?

# Information Gain (IG)

## ■ Ký hiệu:

- Thuộc tính đích  $C = \{c_1, c_2, \dots, c_K\}$ : tập  $K$  lớp
- Tập dữ liệu huấn luyện  $\mathbf{D} = \{(\mathbf{x}^1, c^1), (\mathbf{x}^2, c^2), \dots, (\mathbf{x}^N, c^N)\}$  trong đó  $\mathbf{x}^n \in \mathbf{X}$ ,  $c^n \in C$

## ■ Entropy của $C$ trên tập $\mathbf{D}$ , ký hiệu là $H_{\mathbf{D}}(C)$ , được tính như sau:

- Ký hiệu  $p_k$  là xác suất để một đối tượng  $\mathbf{x}$  thuộc lớp  $c_k$ .  $p_k$  có thể ước lượng từ dữ liệu huấn luyện:  $p_k = |\mathbf{D}_k| / |\mathbf{D}|$  (với  $\mathbf{D}_k \subseteq \mathbf{D}$  là tập các đối tượng  $\mathbf{x}$  thuộc lớp  $c_k$ )

$$H_{\mathbf{D}}(C) = - \sum_{k=1}^K p_k \log_2 p_k = - \sum_{k=1}^K \frac{|\mathbf{D}_k|}{|\mathbf{D}|} \log_2 \frac{|\mathbf{D}_k|}{|\mathbf{D}|}$$

## ■ Chọn một thuộc tính $F_i \in \mathbf{F} = \{F_1, F_2, \dots, F_M\}$ để phân chia $\mathbf{D}$ : khi đó entropy của $C$ (trên $\mathbf{D}$ ) điều kiện $F_i$ , ký hiệu là $H_{\mathbf{D}}(C | F_i)$ , được tính:

- Thuộc tính  $F_i$  có miền xác định  $\mathbf{V}^i = \{v_{1}^i, v_{2}^i, \dots, v_{p_i}^i\}$
- Gọi  $\mathbf{D}_j \subseteq \mathbf{D}$  là tập các đối tượng  $\mathbf{x}$  có thuộc tính  $F_i = v_j^i$ .

$$H_{\mathbf{D}}(C | F_i) = \sum_{j=1}^{P_i} \frac{|\mathbf{D}_j|}{|\mathbf{D}|} H_{\mathbf{D}_j}(C | F_i = v_j^i)$$

## ■ Khi đó:

$$IG_{\mathbf{D}}(C | F_i) = H_{\mathbf{D}}(C) - H_{\mathbf{D}}(C | F_i)$$



## Information Gain (IG) – cont'd

$$H_{\mathbf{D}}(C) = -\sum_{k=1}^K p_k \log_2 p_k = -\sum_{k=1}^K \frac{|\mathbf{D}_k|}{|\mathbf{D}|} \log_2 \frac{|\mathbf{D}_k|}{|\mathbf{D}|}$$

$$H_{\mathbf{D}}(C | F_i) = \sum_{j=1}^{P_i} \frac{|\mathbf{D}_j|}{|\mathbf{D}|} H_{\mathbf{D}_j}(C | F_i = v_j^i)$$

$$IG_{\mathbf{D}}(C | F_i) = H_{\mathbf{D}}(C) - H_{\mathbf{D}}(C | F_i)$$

Thuộc tính **phân lớp tốt nhất**:

Là thuộc tính F đem lại giá trị  $IG_{\mathbf{D}}(C|F)$  **lớn nhất**

Đây là tiêu chuẩn để chọn lựa thuộc tính  $F_i$  trong thuật toán xây dựng cây quyết định ID3.

# Nội dung bài giảng

- Cơ bản về lý thuyết thông tin (information theory)
  - Entropy
  - Entropy điều kiện (conditional entropy)
  - Information gain
- Cây quyết định (decision tree)
  - Mô hình cây quyết định
  - Thuật toán ID3
  - **Ví dụ minh họa**
  - Thuật toán C4.5
- Kết luận
  - Nhược điểm
  - Ưu điểm











# Ví dụ minh họa

Các thuộc tính điều kiện (dùng để phân loại)

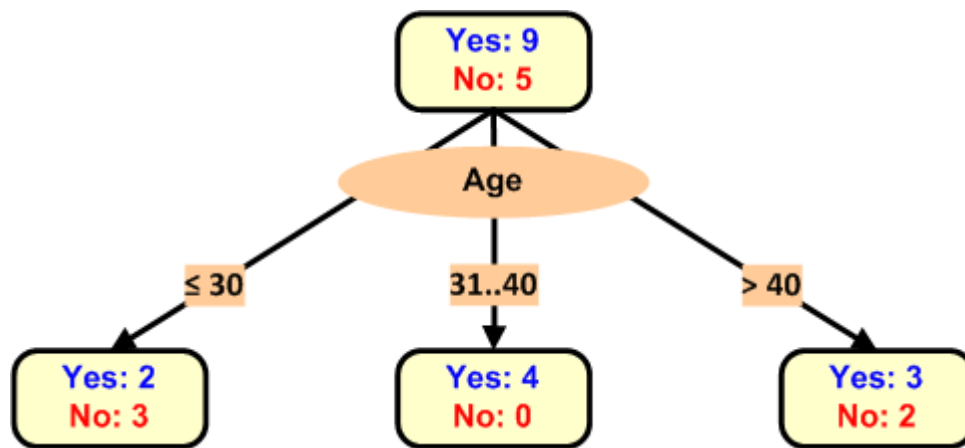
Thuộc tính cần  
phân loại  
(target attribute)

ID	Age	Income	Student	Credit_rating	Buys_computer
1	≤30	High	No	Fair	No
2	≤30	High	No	Excellent	No
3	31..40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31..40	Low	Yes	Excellent	Yes
8	≤30	Medium	No	Fair	No
9	≤30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	≤30	Medium	Yes	Excellent	Yes
12	31..40	Medium	No	Excellent	Yes
13	31..40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

# Information gain theo từng thuộc tính

Thuộc tính	Giá trị	Số lượng theo giá trị	Phân phối theo giá trị	Information Gain
<b>Age</b>				
	$\leq 30$	(2, 3)		$H(\text{Buys\_computer})$ $H(\text{Buys\_computer} \mid \text{Age})$ $0.9403 - 0.6935 = \mathbf{0.2468}$ $IG(\text{Buys\_computer} \mid \text{Age})$
	31..40	(4, 0)		
	>40	(3, 2)		
Income				
	High	(2, 2)		$0.9403 - 0.9111 = 0.0292$
	Medium	(4, 2)		
	Low	(3, 1)		
Student				
	Yes	(6, 1)		$0.9403 - 0.7885 = 0.1518$
	No	(3, 4)		
Credit_rating				
	Fair	(6, 2)		$0.9403 - 0.8922 = 0.0481$
	Excellent	(3, 3)		

# Ví dụ: xây dựng cây



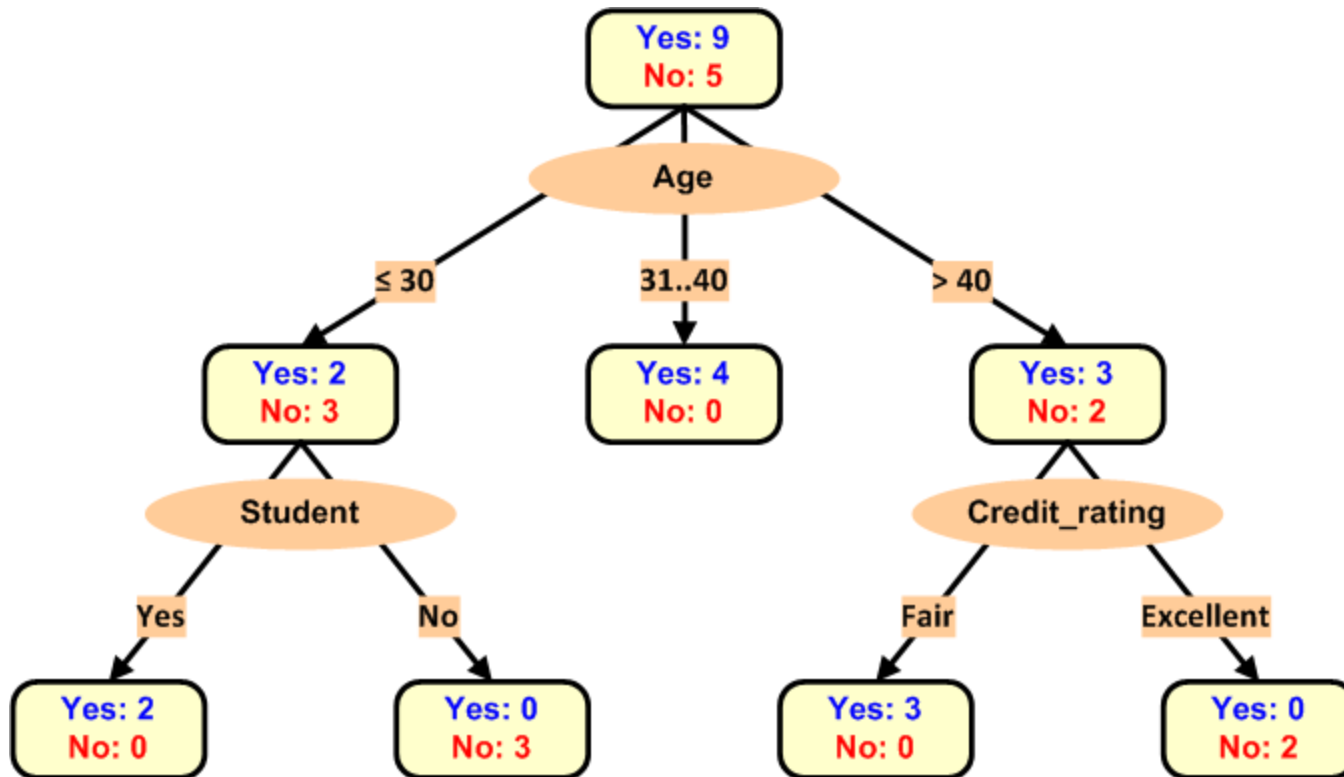
$D_{\text{Age}=\leq 30}$

Thuộc tính	Giá trị	Phân phối	Information Gain
Income	High	(0, 2)	$0.971 - 0.4 = 0.571$
	Medium	(1, 1)	
	Low	(1, 0)	
<b>Student</b>	Yes	(2, 0)	$0.971 - 0 = \mathbf{0.971}$
	No	(0, 3)	
Credit_rating	Fair	(1, 2)	$0.971 - 0.951 = 0.02$
	Excellent	(1, 1)	

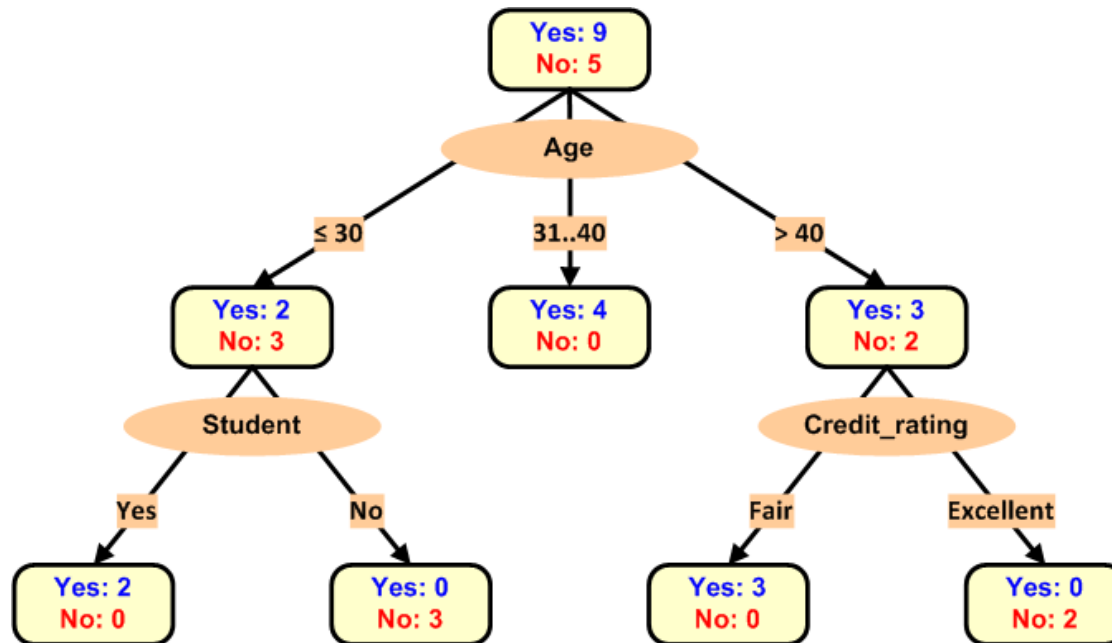
$D_{\text{Age}=> 40}$

Thuộc tính	Giá trị	Phân phối	Information Gain
Income	High		$0.971 - 0.951 = 0.02$
	Medium	(2, 1)	
	Low	(1, 1)	
Student	Yes	(2, 1)	$0.971 - 0.951 = 0.02$
	No	(1, 1)	
<b>Credit_rating</b>	Fair	(3, 0)	$0.971 - 0 = \mathbf{0.971}$
	Excellent	(0, 2)	

## Ví dụ: xây dựng cây (cont'd)



# Phân lớp với cây quyết định



*IF Age  $\leq 30$  AND Student = Yes THEN Buys\_computer = Yes*  
*IF  $31 \leq \text{Age} \leq 40$  THEN Buys\_computer = Yes*  
*IF Age  $> 40$  AND Credit\_rating = Excellent THEN Buys\_computer = No*

# Nội dung bài giảng

- Cơ bản về lý thuyết thông tin (information theory)
  - Entropy
  - Entropy điều kiện (conditional entropy)
  - Information gain
- Cây quyết định (decision tree)
  - Mô hình cây quyết định
  - Thuật toán ID3
  - Ví dụ minh họa
  - **Thuật toán C4.5**
- Kết luận
  - Nhược điểm
  - Ưu điểm



# Thuật toán C4.5

## Một số cải tiến của thuật toán C4.5 so với ID3:

- Sử dụng Gain Ratio (thay vì Information Gain) để chọn thuộc tính phân chia trong quá trình xây dựng cây
- Xử lý tốt cả hai dạng thuộc tính: rời rạc và liên tục
- Xử lý dữ liệu không đầy đủ (thiếu một số giá trị tại một số thuộc tính).
  - C4.5 cho phép các thuộc tính-giá trị bị thiếu có thể thay bằng dấu hỏi (?)
  - Những giá trị bị thiếu không được xem xét khi tính toán Information Gain và Gain Ratio
- Cắt tỉa cây sau khi xây dựng
  - Loại bỏ những nhánh cây không thực sự ý nghĩa (thay bằng nốt lá)

# Gain Ratio

$$H_{\mathbf{D}}(C) = -\sum_{k=1}^K p_k \log_2 p_k = -\sum_{k=1}^K \frac{|\mathbf{D}_k|}{|\mathbf{D}|} \log_2 \frac{|\mathbf{D}_k|}{|\mathbf{D}|}$$

$$H_{\mathbf{D}}(C | F_i) = \sum_{j=1}^{P_i} \frac{|\mathbf{D}_j|}{|\mathbf{D}|} H_{\mathbf{D}_j}(C | F_i = v_j^i)$$

$$IG_{\mathbf{D}}(C | F_i) = H_{\mathbf{D}}(C) - H_{\mathbf{D}}(C | F_i)$$











- Splitting entropy của thuộc tính  $F_i$ , ký hiệu  $SE(F_i)$ :

$$SE_{\mathbf{D}}(F_i) = -\sum_{j=1}^{P_i} \frac{|\mathbf{D}_j|}{|\mathbf{D}|} \log_2 \frac{|\mathbf{D}_j|}{|\mathbf{D}|}$$

- Khi đó, Gain Ratio, ký hiệu  $GR_{\mathbf{D}}(C | F_i)$ , được xác định:

$$GR_{\mathbf{D}}(C | F_i) = \frac{IG_{\mathbf{D}}(C | F_i)}{SE_{\mathbf{D}}(F_i)}$$

# Information Gain & Gain Ratio

Thuộc tính	Giá trị	Số lượng theo giá trị	Phân phối theo giá trị	Information Gain	Gain ratio
Age	≤30	(2, 3)		0.2468	$0.2468 / 1.5774 = 0.1565$
	31..40	(4, 0)			
	>40	(3, 2)			
Income	High	(2, 2)		0.0292	$0.0292 / 1.5567 = 0.0187$
	Medium	(4, 2)			
	Low	(3, 1)			
Student	Yes	(6, 1)		0.1518	$0.1518 / 1 = 0.1518$
	No	(3, 4)			
Credit_rating	Fair	(6, 2)		0.0481	$0.0481 / 0.9852 = 0.04882$
	Excellent	(3, 3)			

# Ý nghĩa của Gain Ratio

- Tiêu chí Information Gain thường “ưu tiên” chọn những thuộc tính có nhiều giá trị (miền xác định lớn)
  - Ví dụ: thuộc tính “Số thẻ tín dụng”
- Splitting entropy,  $SE_D(F_i)$  sẽ lớn khi thuộc tính  $F_i$  có nhiều giá trị. Điều này giúp:
  - Làm giảm Gain Ratio của thuộc tính có nhiều giá trị
  - Làm tăng Gain Ratio của thuộc tính có ít giá trị
- Ý nghĩa khác:
  - Giảm vấn đề “quá khớp” (overfitting)

# Thuật toán C4.5, See5/C5.0 – Các cải tiến khác

## C4.5 - Xem thêm:

- J. Ross Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publisher, 1993.
- J. Ross Quinlan, Improved use of continuous attributes in C4.5, Journal of Artificial Intelligence Research (JAIR), 1996.

## See5/C5.0:

- Cải tiến C4.5:
  - Tăng tốc độ
  - Quản lý và sử dụng bộ nhớ tốt hơn
  - Cây nhỏ gọn hơn
  - Hỗ trợ boosting
  - .v.v.
- Thương mại hóa, không có đặc tả cụ thể, có bản single-threaded free trên Linux

# Nội dung bài giảng

- Cơ bản về lý thuyết thông tin (information theory)
  - Entropy
  - Entropy điều kiện (conditional entropy)
  - Information gain
- Cây quyết định (decision tree)
  - Mô hình cây quyết định
  - Thuật toán ID3
  - Ví dụ minh họa
  - Thuật toán C4.5
- **Kết luận**
  - Nhược điểm
  - Ưu điểm

# Ưu - Nhược điểm

## ■ Nhược điểm:

- ❑ Không đảm bảo xây dựng được cây tối ưu
- ❑ Có thể overfitting (tạo ra cây quá phức tạp, quá khớp với dữ liệu huấn luyện)
- ❑ Thường ưu tiên thuộc tính có nhiều giá trị (khắc phục phần nào bằng Gain Ratio)

## ■ Ưu điểm:

- ❑ Mô hình dễ hiểu và dễ giải thích: cây  $\leftrightarrow$  luật
- ❑ Cần ít dữ liệu huấn luyện
- ❑ Có thể xử lý tốt với dữ liệu số (rời rạc/liên tục) và dữ liệu hạng mục (categorical)
- ❑ Mô hình dạng “white box” (khác với “black box” như mạng nơ ron chẳng hạn)
- ❑ Xây dựng cây nhanh
- ❑ Phân lớp nhanh

# Tổng kết

## ■ Lý thuyết thông tin:

- Entropy
- Entropy điều kiện
- Information gain

## ■ Cây quyết định:

- Mô hình cây quyết định
- Thuật toán ID3
- Những cải tiến trong thuật toán C4.5 so với ID3
- Gain Ratio

## ■ Ứng dụng:

- Xử lý tốt cho dữ liệu bảng biểu với số thuộc tính không quá lớn
- Không phù hợp khi số lượng thuộc tính bùng nổ (ví dụ dữ liệu văn bản, hình ảnh, âm thanh, video, .v.v.)