



Data Mining

# KHAI PHÁ DỮ LIỆU

## *Bài 6* Phân lớp (Classification)



Mai Xuân Hùng

# Nội dung

---



- Phân lớp là gì
- Một số ứng dụng về phân lớp
- Thuật toán phân lớp bằng cây quyết định
- Thuật toán phân lớp bằng mạng Bayes



# Phân lớp là gì

---



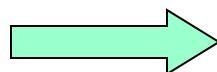
- **Mục đích:** để dự đoán những nhãn phân lớp cho các bộ dữ liệu/mẫu mới
- **Đầu vào:** một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu
- **Đầu ra:** mô hình (bộ phân lớp) dựa trên tập huấn luyện và những nhãn phân lớp



# Tình huống 1



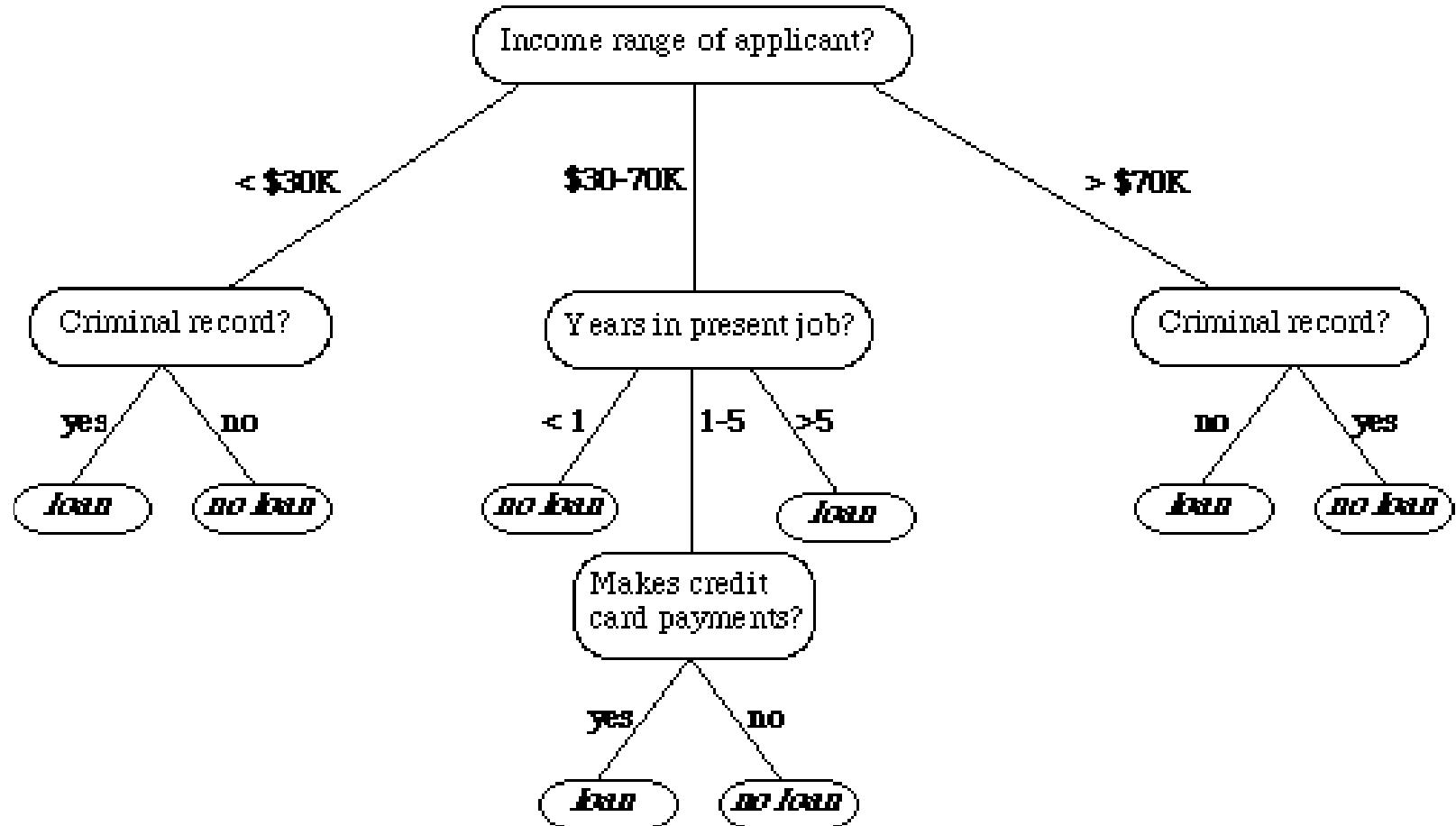
| Tid | Refund | Marital Status | Taxable Income | Evade |
|-----|--------|----------------|----------------|-------|
| 1   | Yes    | Single         | 125K           | No    |
| 2   | No     | Married        | 100K           | No    |
| 3   | No     | Single         | 70K            | No    |
| 4   | Yes    | Married        | 120K           | No    |
| 5   | No     | Divorced       | 95K            | Yes   |
| 6   | No     | Married        | 60K            | No    |
| 7   | Yes    | Divorced       | 220K           | No    |
| 8   | No     | Single         | 85K            | Yes   |
| 9   | No     | Married        | 75K            | No    |
| 10  | No     | Single         | 90K            | Yes   |



Ông A (Tid = 100) có  
khả năng trốn  
thuế???



# Tình huống 2



Với thông tin của một applicant A, xác định liệu ngân hàng có cho A vay không?



# Tình huống 3



| <b>Khóa</b> | <b>MãSV</b> | <b>MônHọc1</b> | <b>MônHọc2</b> | <b>...</b> | <b>TốtNghiệp</b> |
|-------------|-------------|----------------|----------------|------------|------------------|
| 2004        | 1           | 9.0            | 8.5            | ...        | Có               |
| 2004        | 2           | 6.5            | 8.0            | ...        | Có               |
| 2004        | 3           | 4.0            | 2.5            | ...        | Không            |
| 2004        | 8           | 5.5            | 3.5            | ...        | Không            |
| 2004        | 14          | 5.0            | 5.5            | ...        | Có               |
| ...         | ...         | ...            | ...            | ...        |                  |
| 2005        | 90          | 7.0            | 6.0            | ...        | Có               |
| 2006        | 24          | 9.5            | 7.5            | ...        | Có               |
| 2007        | 82          | 5.5            | 4.5            | ...        | Không            |
| 2008        | 47          | 2.0            | 3.0            | ...        | Không            |
| ...         | ...         | ...            | ...            | ...        | ...              |

Làm sao xác định liệu sinh viên A sẽ tốt nghiệp?



# Xây dựng mô hình

---



## Bước 1

- **Mỗi bộ/mẫu dữ liệu** được phân vào một lớp được xác định trước
- Lớp của một bộ/mẫu dữ liệu được xác định bởi **thuộc tính gán nhãn lớp**
- Tập các bộ/mẫu dữ liệu huấn luyện - **tập huấn luyện** - được dùng để **xây dựng mô hình**
- Mô hình được biểu diễn bởi **các luật phân lớp, các cây quyết định hoặc các công thức toán học**



# Sử dụng mô hình

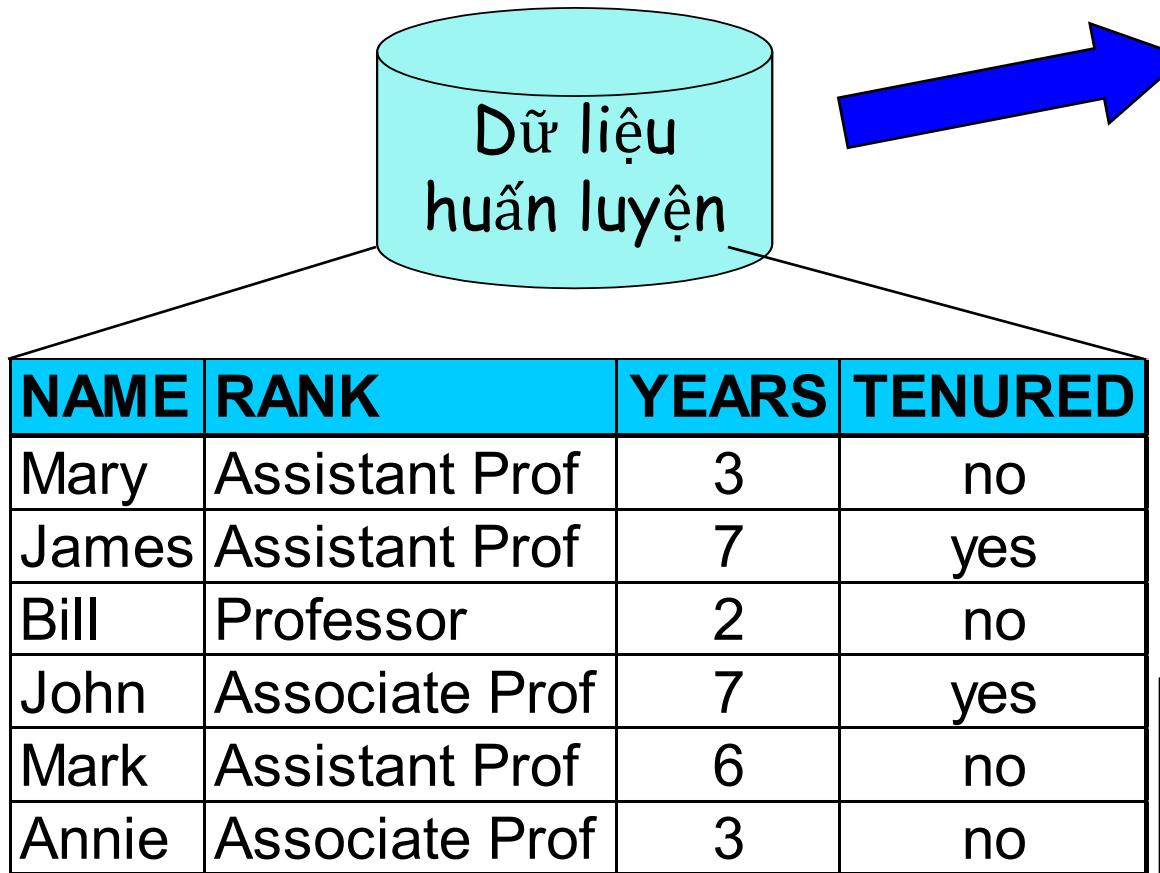


## Bước 2

- **Phân lớp cho những đối tượng mới hoặc chưa được phân lớp**
- **Đánh giá độ chính xác của mô hình**
  - lớp biết trước của một mẫu/bộ dữ liệu đem kiểm tra được so sánh với kết quả thu được từ mô hình
  - tỉ lệ chính xác = phần trăm các mẫu/bộ dữ liệu được phân lớp đúng bởi mô hình trong số các lần kiểm tra



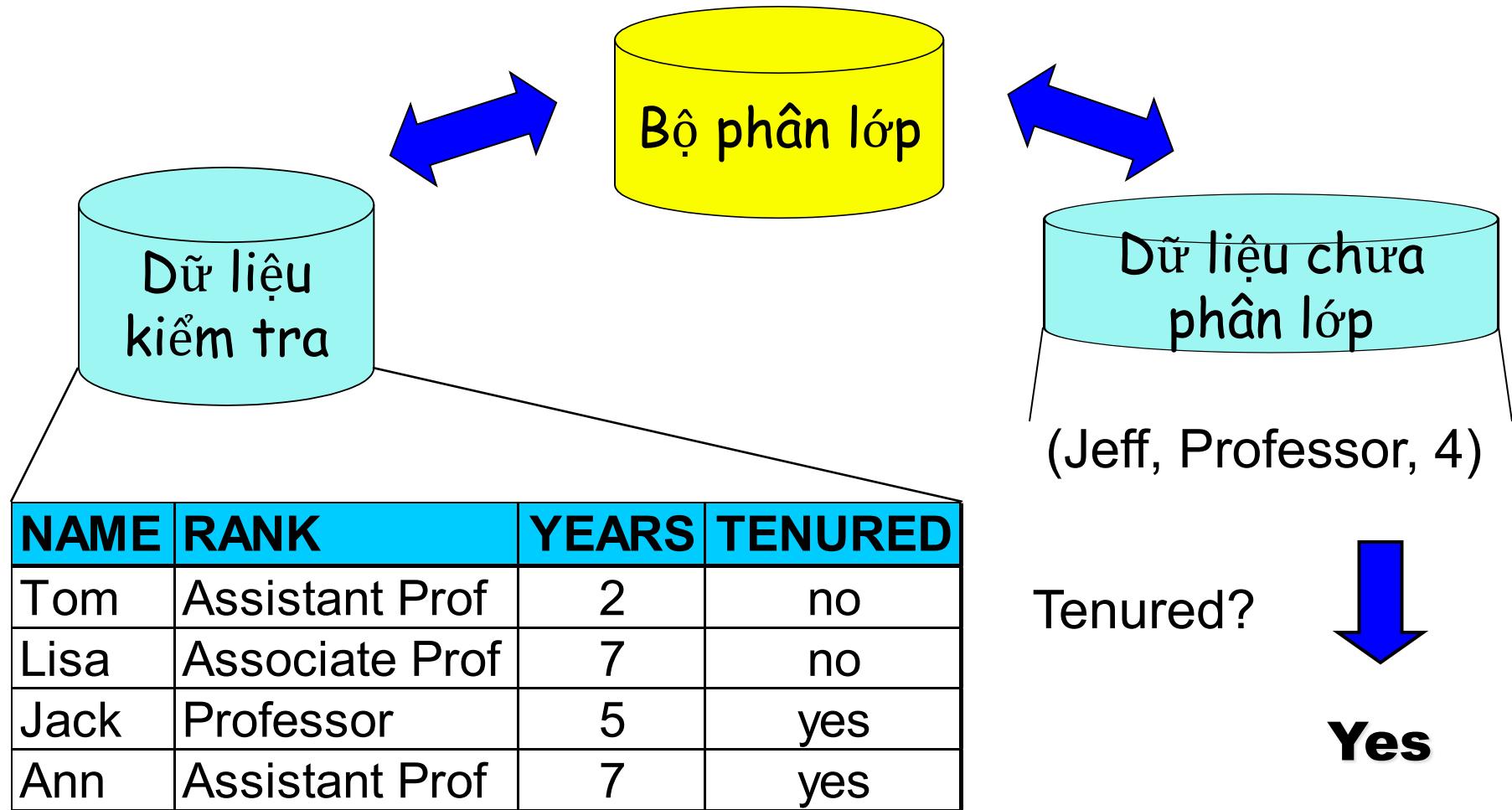
# Ví dụ: xây dựng mô hình



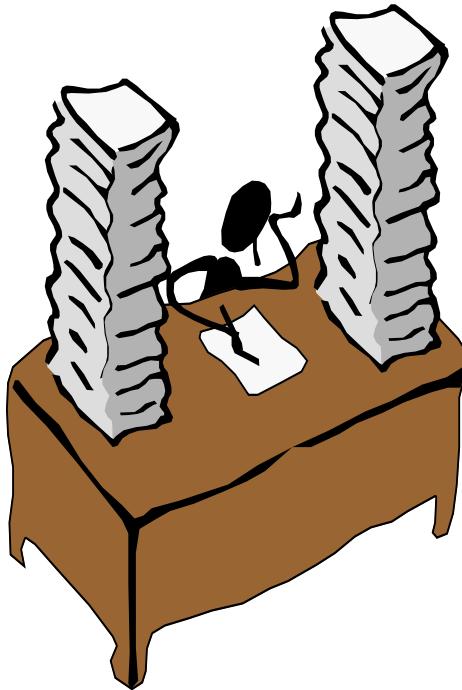
IF rank =  
'professor'  
OR years > 6  
THEN tenured = yes



# Ví dụ: sử dụng mô hình



# Chuẩn bị dữ liệu



- **Làm sạch dữ liệu**
  - nhiều
  - các giá trị trống
- **Phân tích sự liên quan** (chọn đặc trưng)
- **Biến đổi dữ liệu**

# Đánh giá các phương pháp phân lớp

---



- Độ chính xác
- Tốc độ
- Bền vững
- Co dãn (scalability)
- Có thể biểu diễn được
- Dễ làm

# Các thuật toán phân lớp dữ liệu

---



## Thuật toán phân lớp bằng cây quyết định

- Thuật toán Quinlan
- Thuật toán ID3



# Định nghĩa cây quyết định

---



- Cây quyết định là một kiểu mô hình dự báo
- Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định
- Phương tiện có tính mô tả dành cho việc tính toán các xác suất có điều kiện
- Sự kết hợp của các kỹ thuật toán học và tính toán nhằm hỗ trợ việc mô tả, phân loại và tổng quát hóa một tập dữ liệu cho trước



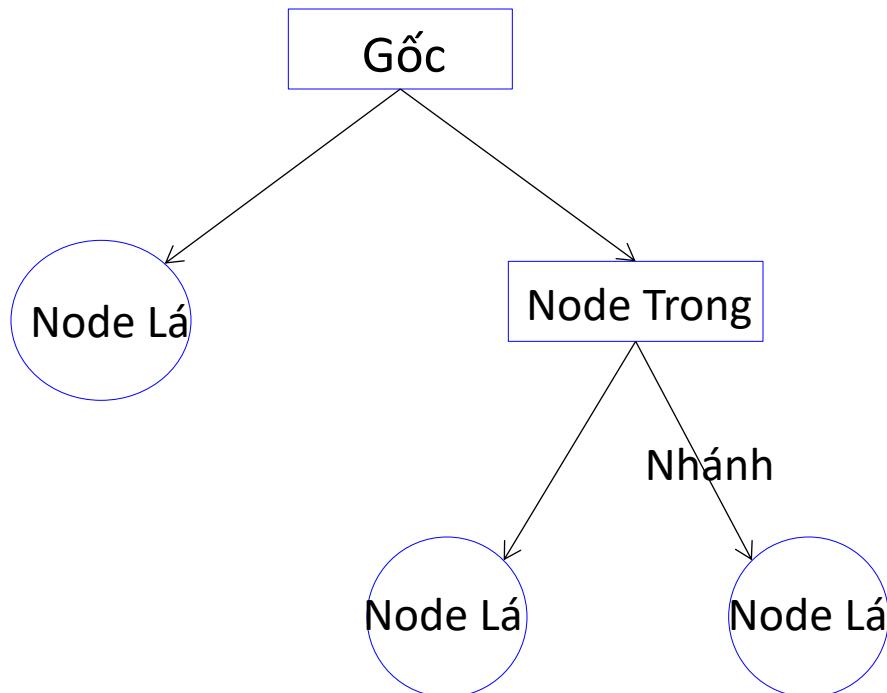
# Định nghĩa cây quyết định



- Cây quyết định là một cấu trúc phân cấp của các nút và các nhánh
  - 3 loại nút trên cây:
    - Nút gốc
    - Nút nội bộ: mang tên thuộc tính của CSDL
    - Nút lá: mang tên lớp  $C_i$
  - Nhánh: mang giá trị có thể của thuộc tính
- Cây quyết định được sử dụng trong phân lớp bằng cách duyệt từ nút gốc của cây cho đến khi đụng đến nút lá, từ đó rút ra lớp của đối tượng cần xét



# Hình dạng cây Quyết định



# Ví dụ



David là quản lý của một câu lạc bộ đánh golf nổi tiếng. Anh ta đang có rắc rối chuyện các thành viên đến hay không đến. Có ngày ai cũng muốn chơi golf nhưng số nhân viên câu lạc bộ lại không đủ phục vụ. Có hôm, không hiểu vì lý do gì mà chẳng ai đến chơi, và câu lạc bộ lại thừa nhân viên.

Mục tiêu của David là tối ưu hóa số nhân viên phục vụ mỗi ngày bằng cách dựa theo thông tin dự báo thời tiết để đoán xem khi nào người ta sẽ đến chơi golf. Để thực hiện điều đó, anh cần hiểu được tại sao khách hàng quyết định chơi và tìm hiểu xem có cách giải thích nào cho việc đó hay không.

Vậy là trong hai tuần, anh ta thu thập thông tin về: Trời (outlook) (nắng (sunny), nhiều mây (overcast) hoặc mưa (raining)). Nhiệt độ (temperature) bằng độ F. Độ ẩm (humidity). Có gió mạnh (wind) hay không.

Và tất nhiên là số người đến chơi golf vào hôm đó. David thu được một bộ dữ liệu gồm 14 dòng và 5 cột.



# Ví dụ



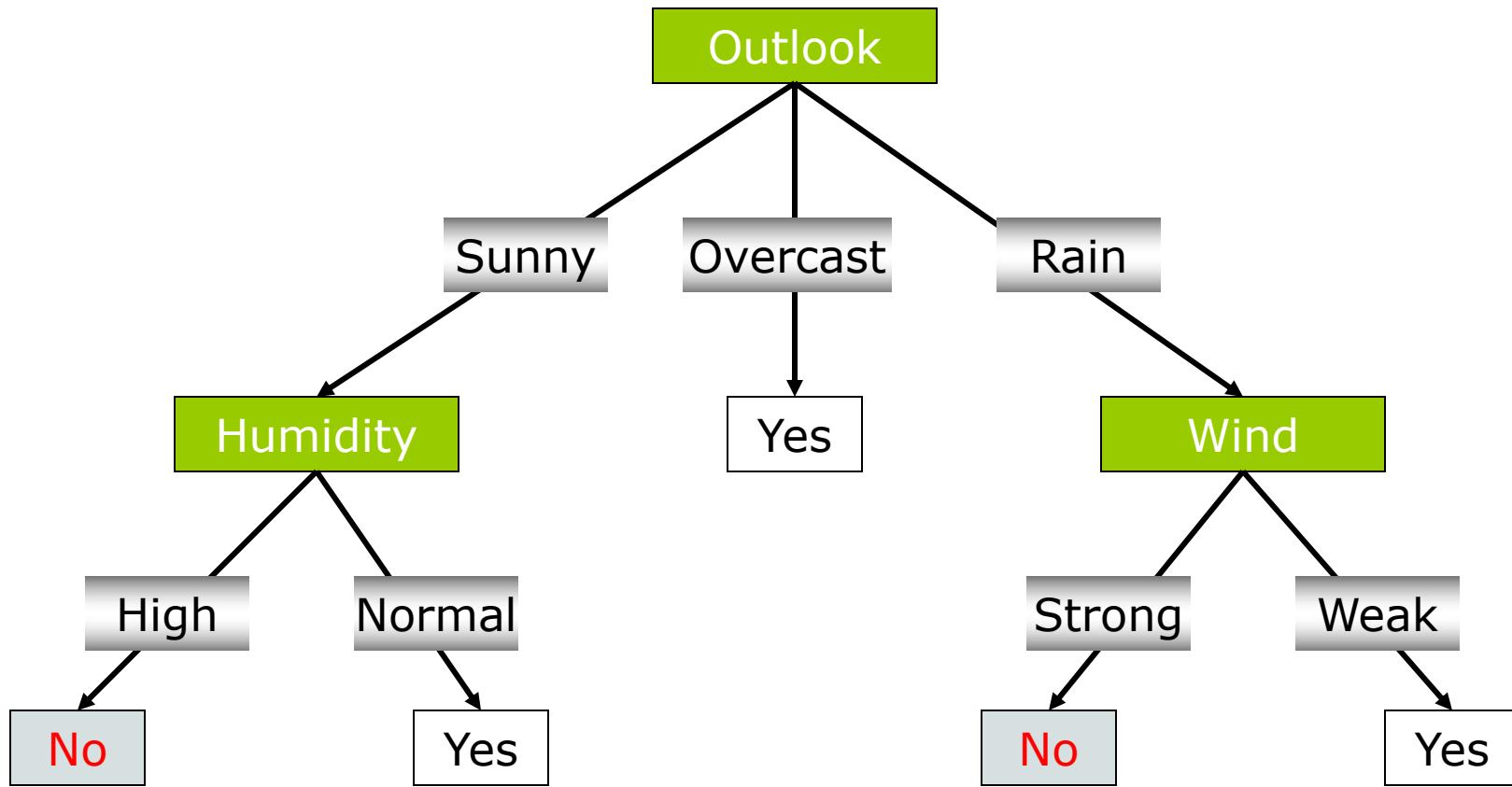
| Day | Outlook  | Temp | Humidity | Wind   | Play? |
|-----|----------|------|----------|--------|-------|
| 1   | Sunny    | Hot  | High     | Weak   | No    |
| 2   | Sunny    | Hot  | High     | Strong | No    |
| 3   | Overcast | Hot  | High     | Weak   | Yes   |
| 4   | Rainy    | Mild | High     | Weak   | Yes   |
| 5   | Rainy    | Cool | Normal   | Weak   | Yes   |
| 6   | Rainy    | Cool | Normal   | Strong | No    |
| 7   | Overcast | Cool | Normal   | Strong | Yes   |
| 8   | Sunny    | Mild | High     | Weak   | No    |
| 9   | Sunny    | Cold | Normal   | Weak   | Yes   |
| 10  | Rainy    | Mild | Normal   | Weak   | Yes   |
| 11  | Sunny    | Mild | Normal   | Strong | Yes   |
| 12  | Overcast | Mild | High     | Strong | Yes   |
| 13  | Overcast | Hot  | Normal   | Weak   | Yes   |
| 14  | Rainy    | Mild | High     | Strong | No    |



# Ví dụ



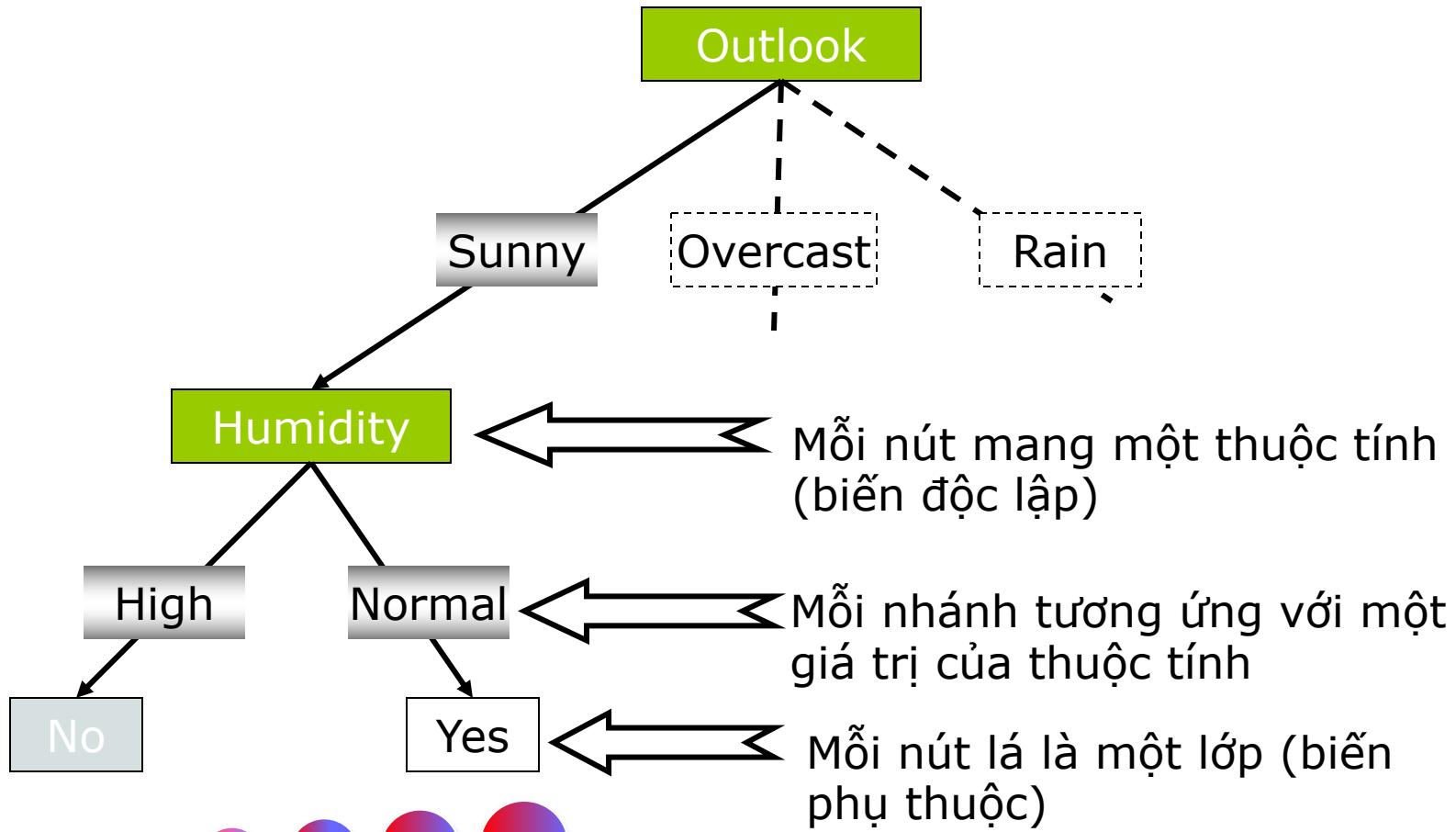
Kiểm tra khi nào chơi golf, khi nào không chơi



# Ví dụ



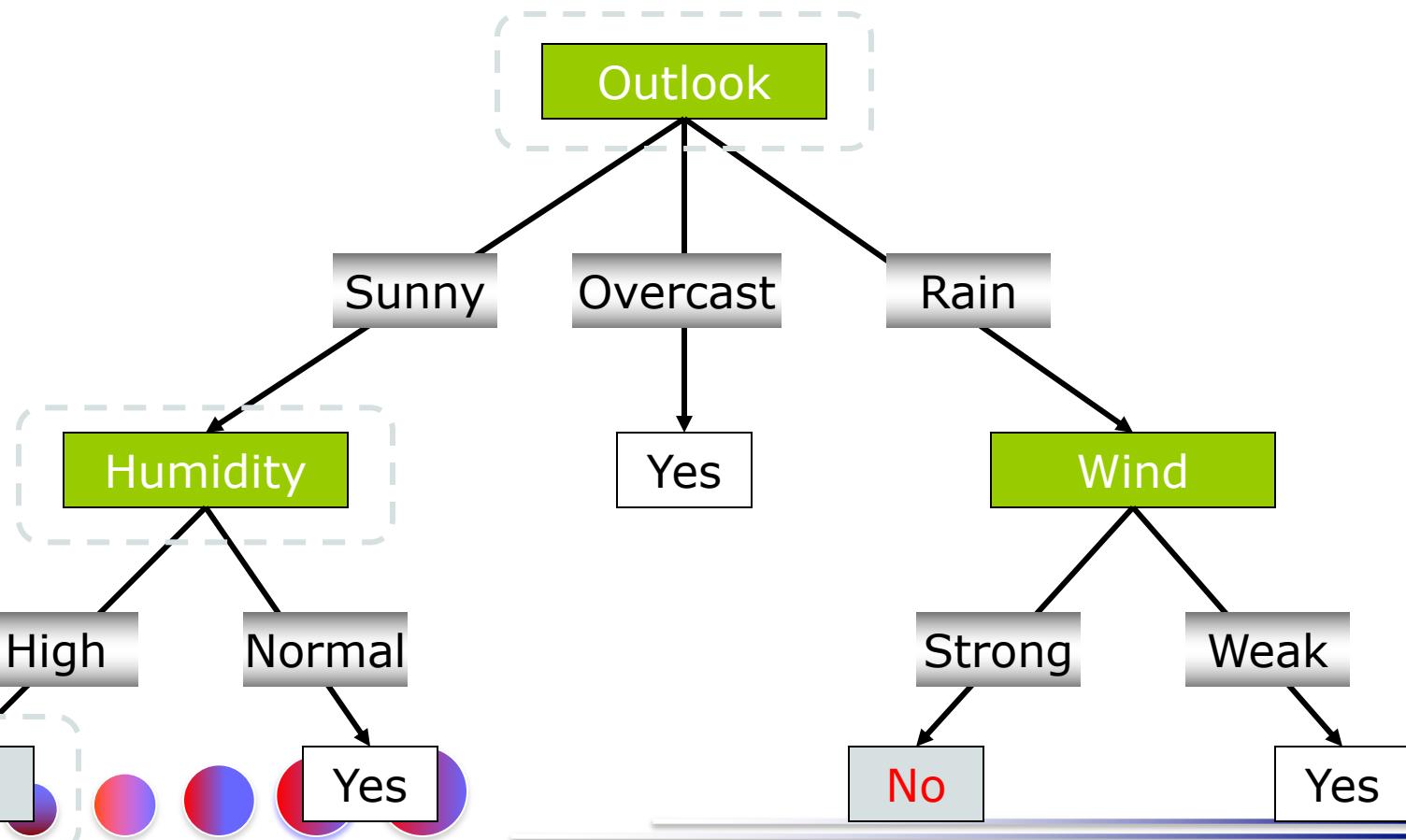
Kiểm tra khi nào chơi golf, khi nào không chơi



# Duyệt cây quyết định



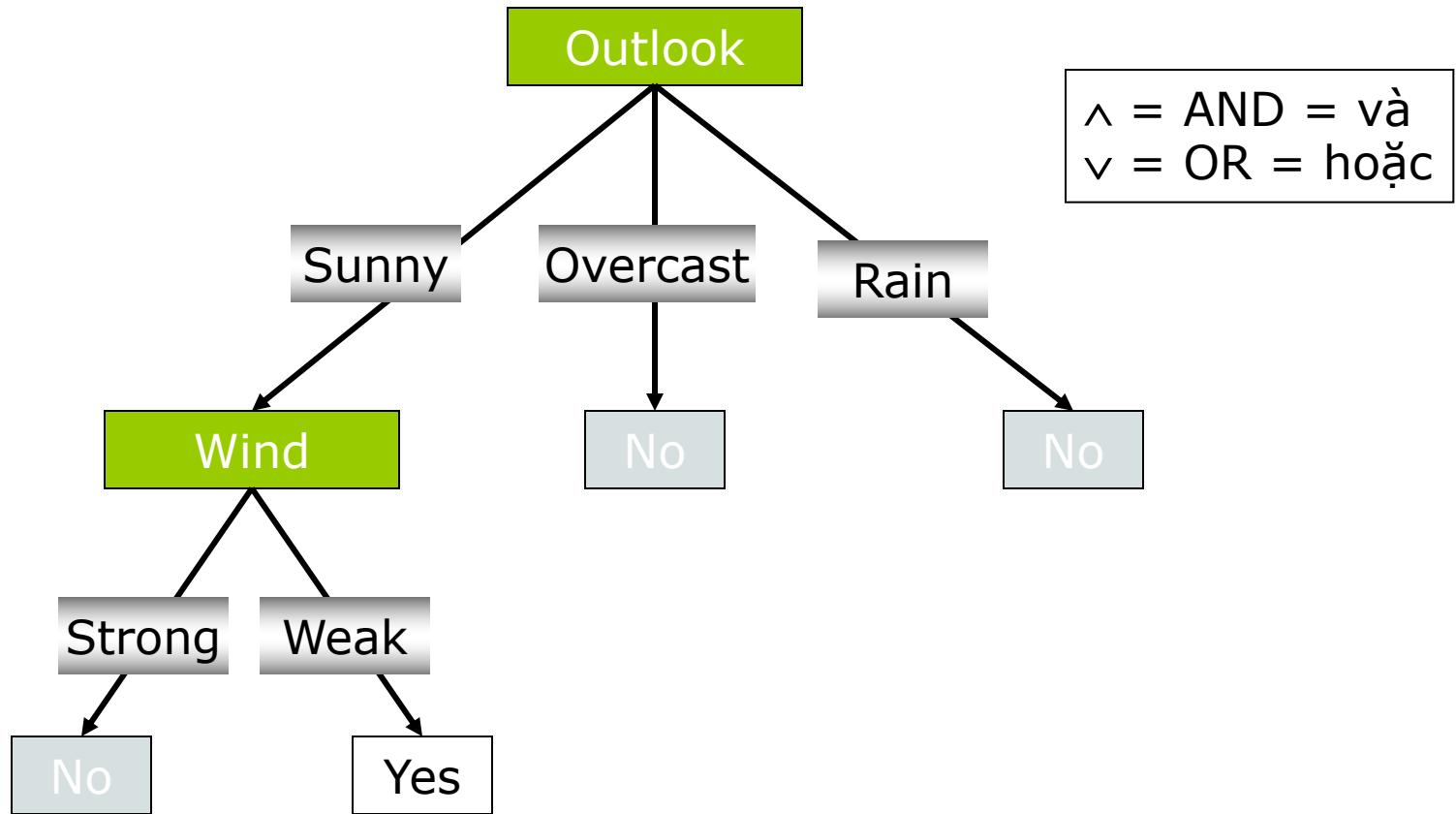
| Day | Outlook | Temp | Humidity | Wind | Play? |
|-----|---------|------|----------|------|-------|
| 1   | Sunny   | Hot  | High     | Weak | No    |



# Biểu thức luận lý



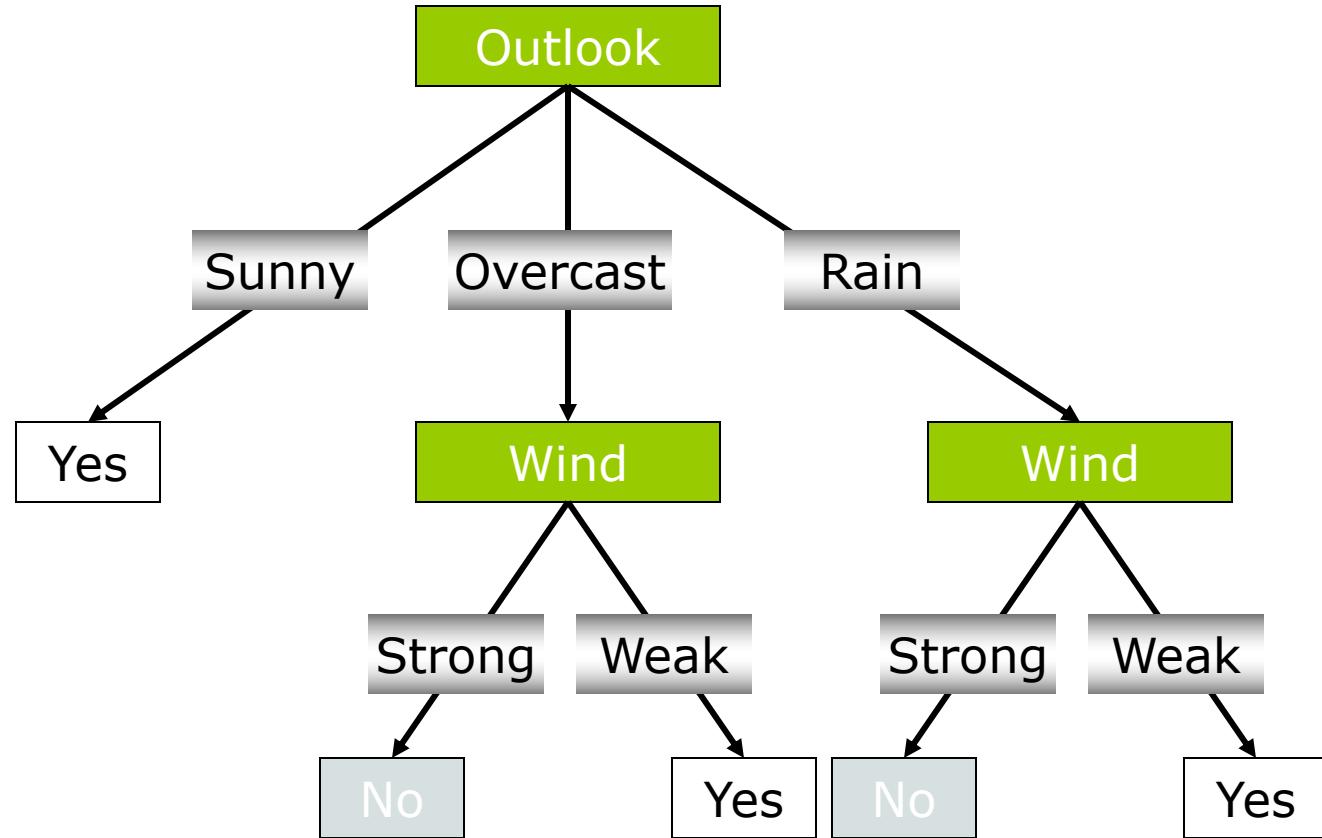
Outlook=Sunny  $\wedge$  Wind=Weak



# Biểu thức luận lý



$\text{Outlook} = \text{Sunny} \vee \text{Wind} = \text{Weak}$



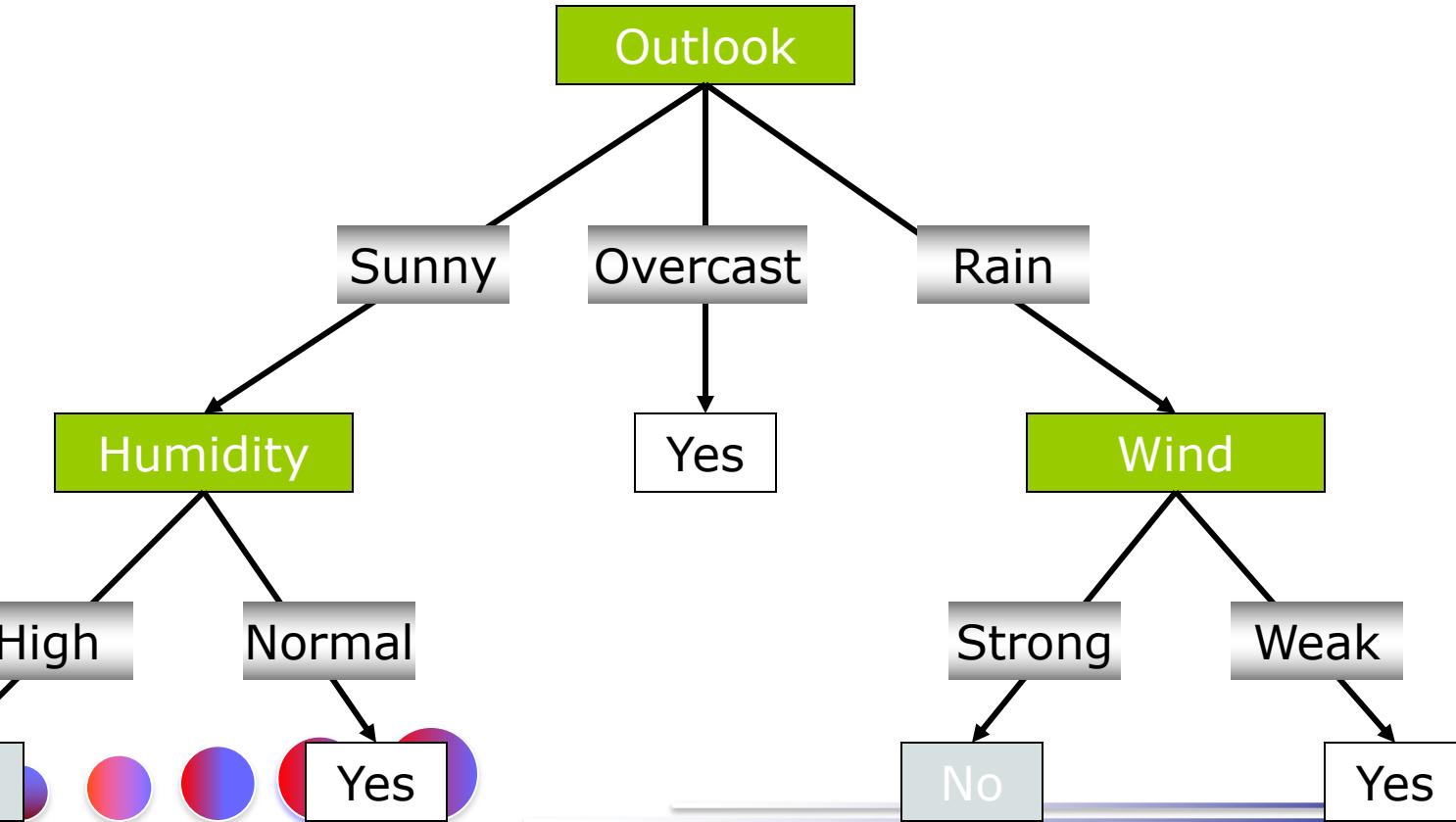
# Biểu thức luận lý



(Outlook=Sunny  $\wedge$  Humidity=Normal)

✓ Outlook=Overcast

✓ (Outlook=Rain  $\wedge$  Wind=Weak)



# Xây dựng cây quyết định

---



- Cây được thiết lập từ trên xuống dưới
- Rời rạc hóa các thuộc tính dạng phi số
- Các mẫu huấn luyện nằm ở gốc của cây
- Chọn một thuộc tính để phân chia thành các nhánh. Thuộc tính được chọn dựa trên độ đo thống kê hoặc độ đo heuristic
- Tiếp tục lặp lại việc xây dựng cây quyết định cho các nhánh



# Xây dựng cây quyết định

---



- Điều kiện dừng

- Tất cả các mẫu rơi vào một nút thuộc về cùng một lớp (nút lá)
- Không còn thuộc tính nào có thể dùng để phân chia mẫu nữa
- Không còn lại mẫu nào tại nút



# Thuật toán Quinlan

---



- Chọn thuộc tính nào có số lượng vector đơn vị nhiều

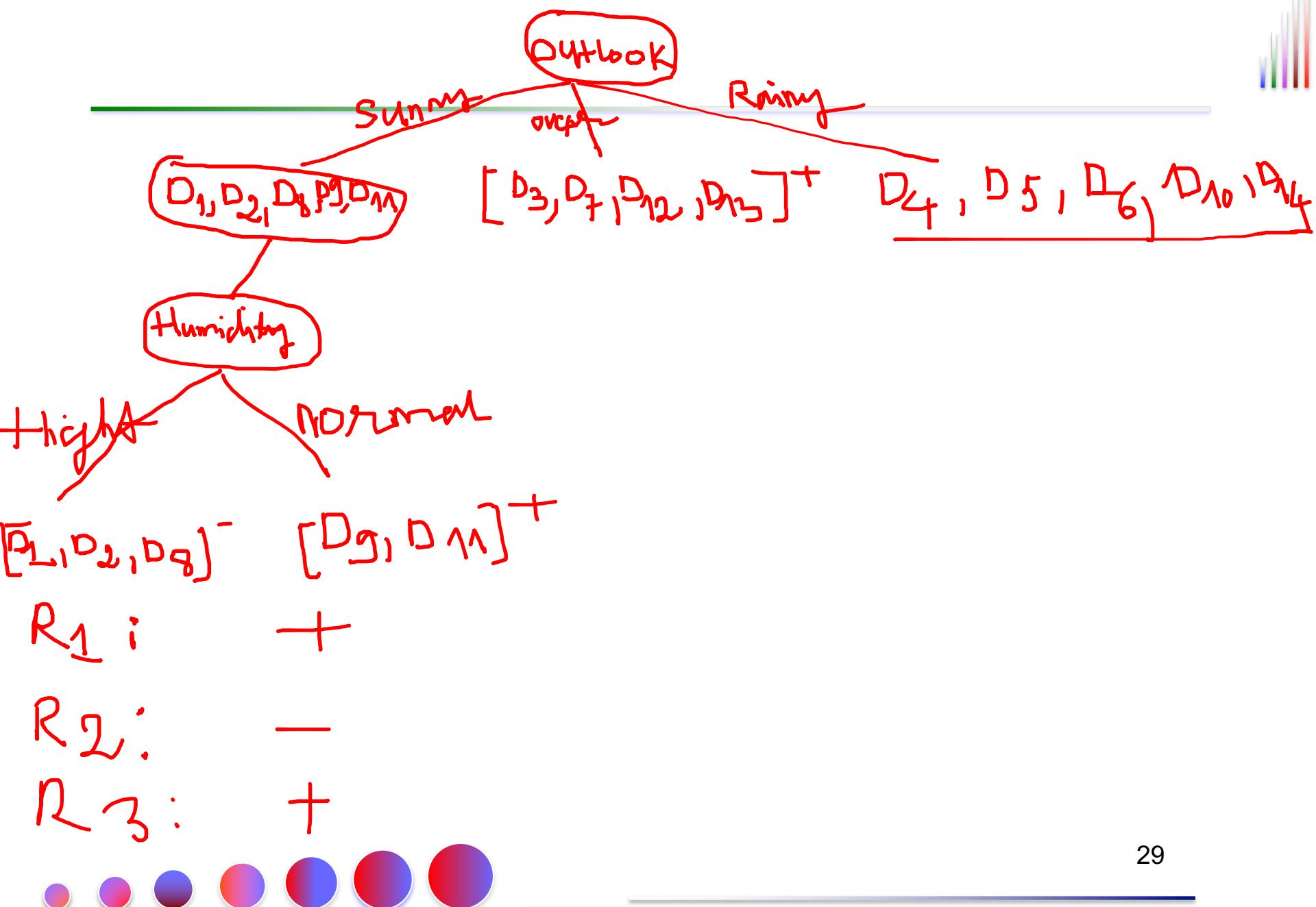


# Ví dụ



| DAY | Outlook  | Temp | Humidity | Windy | Play |
|-----|----------|------|----------|-------|------|
| D1  | Sunny    | Hot  | High     | False | No   |
| D2  | Sunny    | Hot  | High     | True  | No   |
| D3  | Overcast | Hot  | High     | False | Yes  |
| D4  | Rainy    | Mild | High     | False | Yes  |
| D5  | Rainy    | Cool | Normal   | False | Yes  |
| D6  | Rainy    | Cool | Normal   | True  | No   |
| D7  | Overcast | Cool | Normal   | True  | Yes  |
| D8  | Sunny    | Mild | High     | False | No   |
| D9  | Sunny    | Cool | Normal   | False | Yes  |
| D10 | Rainy    | Mild | Normal   | False | Yes  |
| D11 | Sunny    | Mild | Normal   | True  | Yes  |
| D12 | Overcast | Mild | High     | True  | Yes  |
| D13 | Overcast | Hot  | Normal   | False | Yes  |
| D14 | Rainy    | Mild | High     | True  | No   |





# Thuộc tính: Outlook

---



- $V(\text{outlook} = \text{sunny}) = (T(\text{sunny}, \text{No}), T(\text{sunny}, \text{Yes})) = (3/5, 2/5)$
- $V(\text{outlook} = \text{overcast}) = (T(\text{overcast}, \text{No}), T(\text{overcast}, \text{Yes})) = (0/4, 4/4) = \mathbf{(0,1)}$
- $V(\text{outlook} = \text{rainy}) = (T(\text{rainy}, \text{No}), T(\text{rainy}, \text{Yes})) = (2/5, 3/5)$



# Thuộc tính: Temp

---



- $V(Temp = \text{hot}) = (T(\text{hot}, \text{No}), T(\text{hot}, \text{Yes})) = (2/4, 2/4)$
- $V(Temp = \text{mild}) = (T(\text{mild}, \text{No}), T(\text{mild}, \text{Yes})) = (2/6, 4/6)$
- $V(Temp = \text{cool}) = (T(\text{cool}, \text{No}), T(\text{cool}, \text{Yes})) = (1/4, 3/4)$



# Thuộc tính: Humidity

---

- $V(\text{Humidity} = \text{high})$   
=( $T(\text{high},\text{No}), T(\text{high},\text{Yes})$ ) = ( 4/7 ,3/7)
- $V(\text{Humidity} = \text{normal})$   
=( $T(\text{normal},\text{No}), T(\text{normal},\text{Yes})$ ) = ( 1/7  
,6/7)



# Thuộc tính: Wind

---

- $V(\text{wind} = \text{false})$   
=  $(T(\text{false}, \text{No}), T(\text{false}, \text{Yes})) = (2/8, 6/8)$
- $V(\text{wind} = \text{true}) = (T(\text{true No}), T(\text{true Yes})) = (3/6, 3/6)$
- **Như vậy, thuộc tính Outlook có số vector đơn vị nhiều nhất nên sẽ được phân hoạch**



# Nhận xét

---

- Sau khi phân hoạch theo Outlook xong, chỉ có phân hoạch theo Outlook (sunny) và (rain) là còn chứa kết quả là Yes và No nên ta sẽ tiếp tục phân hoạch tập này. Ta sẽ thực hiện thao tác tính vector đặc trưng tương tự đối với các thuộc tính còn lại (*Temp, Humidity, Wind*). Trong phân hoạch (sunny), (rain) , tập dữ liệu của chúng ta còn lại là :



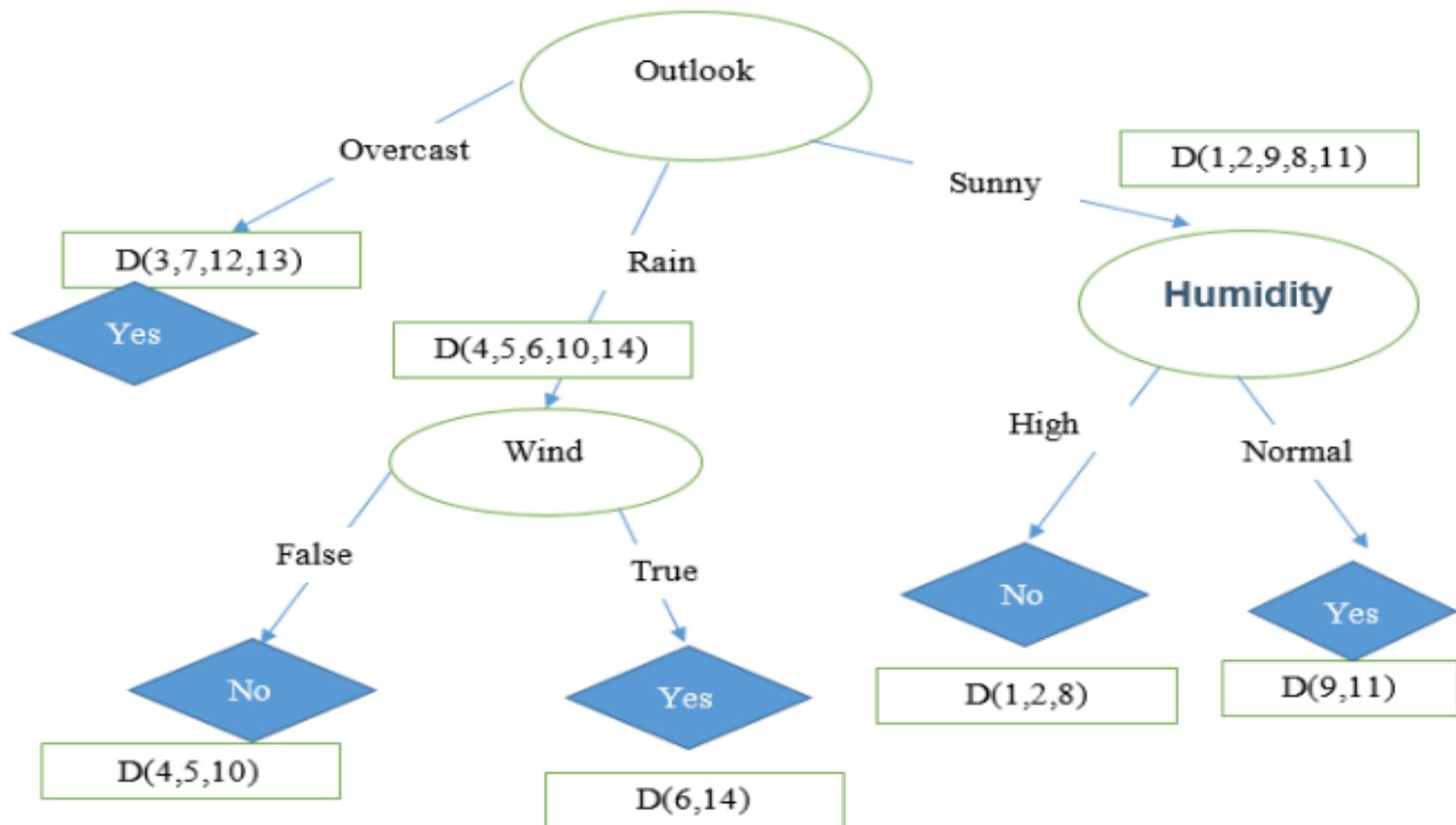
# Các mẫu còn lại trong sunny (TH1)

| DAY | TEMP | HUMIDITY | WIND  | PLAY |
|-----|------|----------|-------|------|
| D1  | HOT  | HIGH     | FALSE | NO   |
| D2  | HOT  | HIGH     | TRUE  | NO   |
| D8  | MILD | HIGH     | FALSE | NO   |
| D9  | COOL | NORMAL   | FALSE | YES  |
| D11 | MILD | NORMAL   | TRUE  | YES  |

# Các mẫu còn lại trong rain (TH2)

| DAY | TEMP | HUMIDITY | WIND  | PLAY |
|-----|------|----------|-------|------|
| D4  | MILD | HIGH     | FALSE | YES  |
| D5  | COOL | NORMAL   | FALSE | YES  |
| D6  | COOL | NORMAL   | TRUE  | NO   |
| D10 | MILD | NORMAL   | FALSE | YES  |
| D14 | MILD | HIGH     | TRUE  | NO   |

# Tính toán tương tự ta có kq



# Thuật toán ID3

---



- Giải thuật ID3 (gọi tắt là ID3) Được phát triển đồng thời bởi Quinlan trong AI và Breiman, Friedman, Olsen và Stone trong thống kê



# Lựa chọn thuộc tính



- Độ đo để lựa chọn thuộc tính: Thuộc tính được chọn là thuộc tính có lợi nhất cho quá trình phân lớp (tạo ra cây nhỏ nhất)
- Có 2 độ đo thường dùng
  - 1. Độ lợi thông tin (Information gain)
    - Giả sử tất cả các thuộc tính dạng phi số
    - Có thể biến đổi để áp dụng cho thuộc tính số
  - 2. Chỉ số Gini (Gini index)
    - Giả sử tất cả các thuộc tính dạng số
    - Giả sử tồn tại một vài giá trị có thể phân chia giá trị của từng thuộc tính
    - Có thể biến đổi để áp dụng cho thuộc tính phi số



# Độ lợi thông tin (Information gain)

---

- $S$ : số lượng tập huấn luyện
- $S_i$ : số các mẫu của  $S$  nằm trong lớp  $C_i$  với  $i = \{1, \dots, m\}$
- Thông tin cần biết để phân lớp một mẫu

$$I(S_1, S_2, \dots, S_m) = - \sum_{i=1}^m \frac{S_i}{S} \log_2 \frac{S_i}{S}$$



# Độ lợi thông tin (tt)



- Cho  $P$  và  $N$  là hai lớp và  $S$  là một tập dữ liệu có  $p$  phần tử lớp  $P$  và  $n$  phần tử lớp  $N$
- Khối lượng thông tin cần thiết để quyết định một mẫu tùy ý có thuộc về lớp  $P$  hay  $N$  hay không là

$$I(p,n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$



# Độ lợi thông tin



- Thuộc tính A có các giá trị  $\{a_1, a_2, \dots, a_n\}$
- Dùng thuộc tính A để phân chia tập huấn luyện thành  $n$  tập con  $\{S_1, S_2, \dots, S_n\}$
- $S_{ij}$  : số mẫu của lớp  $C_i$  thuộc tập con  $S_j$  ( $A=a_j$ )
- Entropy của thuộc tính A:

$$E(A) = \sum_{j=1}^n \frac{S_{1j} + \dots + S_{mj}}{S} I(S_{1j}, \dots, S_{mj})$$

- Độ lợi thông tin dựa trên phân nhánh bằng thuộc tính A:

$$G(A) = I(S_1, S_2, \dots, S_m) - E(A)$$

- Tại mỗi cấp, chúng ta chọn thuộc tính **có độ lợi lớn nhất** để phân nhánh cây hiện tại



# Ví dụ



| Day | Outlook  | Temp | Humidity | Wind   | Play? |
|-----|----------|------|----------|--------|-------|
| 1   | Sunny    | Hot  | High     | Weak   | No    |
| 2   | Sunny    | Hot  | High     | Strong | No    |
| 3   | Overcast | Hot  | High     | Weak   | Yes   |
| 4   | Rain     | Mild | High     | Weak   | Yes   |
| 5   | Rain     | Cool | Normal   | Weak   | No    |
| 6   | Rain     | Cool | Normal   | Strong | Yes   |
| 7   | Overcast | Cool | Normal   | Weak   | Yes   |
| 8   | Sunny    | Mild | High     | Weak   | Yes   |
| 9   | Sunny    | Cool | Normal   | Weak   | Yes   |
| 10  | Rain     | Mild | Normal   | Strong | Yes   |
| 11  | Sunny    | Mild | Normal   | Strong | Yes   |
| 12  | Overcast | Mild | High     | Strong | Yes   |
| 13  | Overcast | Hot  | Normal   | Weak   | Yes   |
| 14  | Rain     | Mild | High     | Strong | No    |

# Độ lợi thông tin, ví dụ

---



- Ta có
  - $S = 14$
  - $m = 2$
  - $C_1 = \text{"Yes"}, C_2 = \text{"No"}$
  - $S_1 = 10, S_2 = 4$

$$I(S_1, S_2) = I(10, 4) = -\frac{10}{14} \log_2 \frac{10}{14} - \frac{4}{14} \log_2 \frac{4}{14} = 0.863$$



# Độ lợi thông tin – Ví dụ (2)



Tính entropy cho thuộc tính *thời tiết*:

| Outlook  | p | q | I(p, q) |
|----------|---|---|---------|
| sunny    | 3 | 2 | 0.971   |
| Overcast | 4 | 0 | 0       |
| rain     | 3 | 2 | 0.971   |

Ta có

$$E(\text{Outlook}) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

Do đó  $\text{Gain}(\text{Outlook}) = I(10,4) - E(\text{Outlook}) = 0.169$



# Độ lợi thông tin, ví dụ

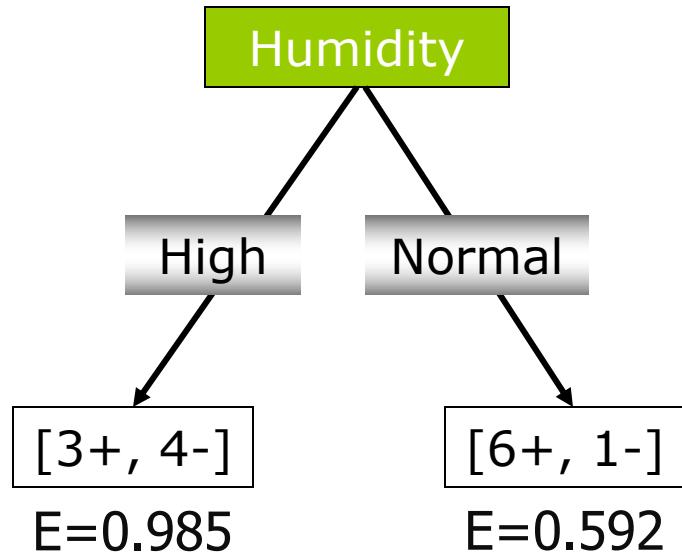
---



$$E(\text{temp}) = 4/14 * I(2,2) + 6/14 * I(5,1) + 4/14 * I(1,3) = 0.796$$



# Độ lợi thông tin, ví dụ



$$-\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.985$$

$$-\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} = 0.592$$

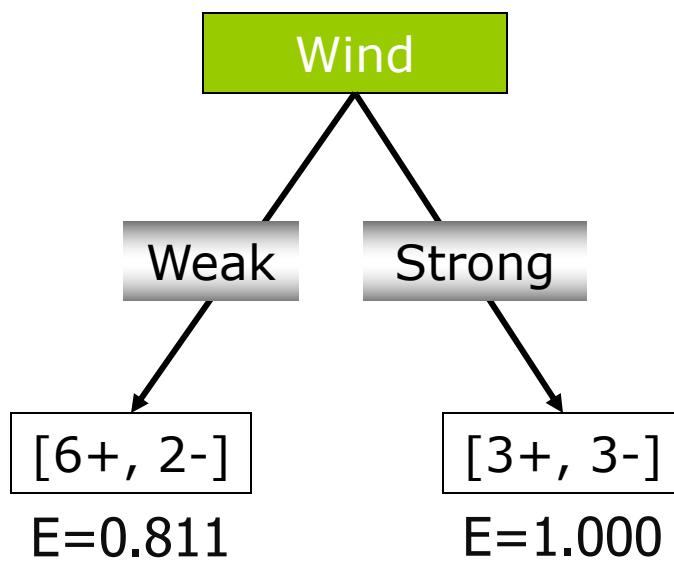
Gain(S, Humidity)

$$= 0.940 - (7/14)*0.985 - (7/14)*0.592$$

$$= 0.151$$

Ghi chú: Để tính  $\log_2 5$  bằng máy tính điện tử, nhấn: 5 log / 2 log =

# Độ lợi thông tin, ví dụ

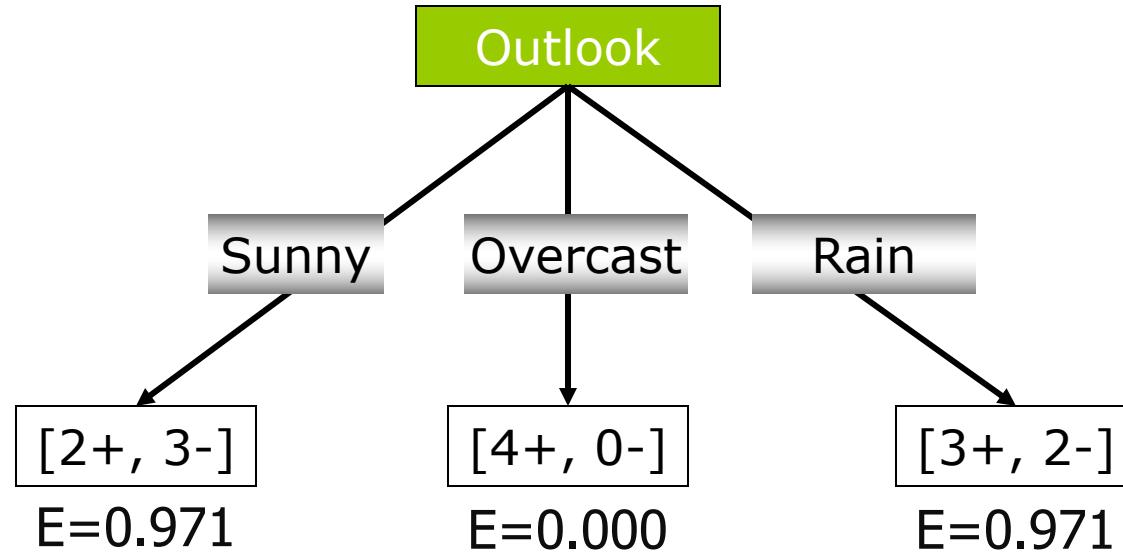


$$-\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0.811$$

$$-\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} = 1.000$$

$$\begin{aligned}\text{Gain}(S, \text{Wind}) &= 0.940 - (8/14)*0.811 - (6/14)*1.000 \\ &= 0.048\end{aligned}$$

# Độ lợi thông tin, ví dụ



$$\text{Gain}(S, \text{Wind}) = 0.048$$

$$\text{Gain}(S, \text{Humidity}) = 0.151$$

$$\begin{aligned}\text{Gain}(S, \text{Outlook}) &= 0.940 - (5/14)*0.971 \\ &\quad - (4/14)*0.0 - (5/14)*0.0971 \\ &= \mathbf{0.247}\end{aligned}$$



# Ví dụ chỉ mục Gini



|     | Outlook  | Temperature | Humidity | Wind   | Play ball |
|-----|----------|-------------|----------|--------|-----------|
| D1  | Sunny    | Hot         | High     | Weak   | No        |
| D2  | Sunny    | Hot         | High     | Strong | No        |
| D3  | Overcast | Hot         | High     | Weak   | Yes       |
| D4  | Rainy    | Mild        | High     | Weak   | Yes       |
| D5  | Rainy    | Cool        | Normal   | Weak   | Yes       |
| D6  | Rainy    | Cool        | Normal   | Strong | No        |
| D7  | Overcast | Cool        | Normal   | Strong | Yes       |
| D8  | Sunny    | Mild        | High     | Weak   | No        |
| D9  | Sunny    | Cool        | Normal   | Weak   | Yes       |
| D10 | Rainy    | Mild        | Normal   | Weak   | Yes       |
| D11 | Sunny    | Mild        | Normal   | Strong | Yes       |
| D12 | Overcast | Mild        | High     | Strong | Yes       |
| D13 | Overcast | Hot         | Normal   | Weak   | Yes       |
| D14 | Rainy    | Mild        | High     | Strong | No        |



# Ví dụ chỉ mục Gini



- $\text{Gini}(D) = 1 - (9/14)^2 - (5/14)^2 = 0.459$
- Tìm chỉ mục Gini cho từng thuộc tính

$$\text{Gini}_{\text{Outlook}}(D) = \frac{5}{14} * \text{Gini}(S_{\text{sunny}}) + \frac{4}{14} * \text{Gini}(S_{\text{overcast}}) + \frac{5}{14} * \text{Gini}(S_{\text{rainy}}) = 0.343$$

Với

- $\text{Gini}(S_{\text{sunny}}) = 0.48$  // 2Yes, 3No
- $\text{Gini}(S_{\text{overcast}}) = 0$  // 4Yes, 0No
- $\text{Gini}(S_{\text{rainy}}) = 0.48$  // 3Yes, 2No



# Bài tập

---



- Tính Gini cho các thuộc tính

✓ Gini<sub>Temperature</sub>(D)

✓ Gini<sub>Humidity</sub>(D)

✓ Gini<sub>Wind</sub>(D)



# Ví dụ chỉ mục Gini



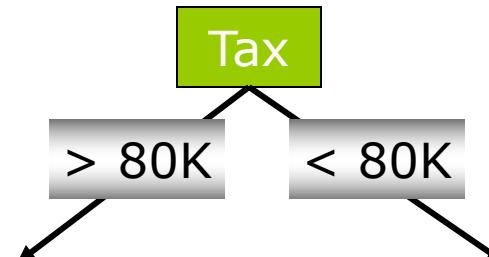
- Với chỉ mục Gini
  - 1.  $\text{Gini}_{\text{Outlook}}(D) = 0.343$
  - 2.  $\text{Gini}_{\text{Temperature}}(D) = 0.405$
  - 3.  $\text{Gini}_{\text{Humidity}}(D) = 0.734$
  - 4.  $\text{Gini}_{\text{Wind}}(D) = 0.875$
- Thuộc tính có giá trị chỉ mục Gini nhỏ nhất sẽ được chọn để phân nhánh: **Outlook**

# Phân chia thuộc tính có giá trị liên tục



- Dựa trên một giá trị nếu muốn phân chia nhị phân
- Dựa trên vài giá trị nếu muốn có nhiều nhánh
- Với mỗi giá trị tính các mẫu thuộc một lớp theo dạng  $A_{\leq v}$  và  $A_{> v}$
- Cách chọn giá trị  $v$  đơn giản: với mỗi giá trị  $v$  trong CSDL đều tính Gini của nó và lấy giá trị có Gini nhỏ nhất  $\rightarrow$  kém hiệu quả

| TID | Refund | Marital  | Tax  | Cheat |
|-----|--------|----------|------|-------|
| 1   | Yes    | Single   | 125K | No    |
| 2   | No     | Married  | 100K | No    |
| 3   | No     | Single   | 70K  | No    |
| 4   | Yes    | Married  | 120K | No    |
| 5   | No     | Divorced | 95K  | Yes   |
| 6   | No     | Married  | 60K  | No    |
| 7   | Yes    | Divorced | 220K | No    |
| 8   | No     | Single   | 85K  | Yes   |
| 9   | No     | Married  | 75K  | No    |
| 10  | No     | Single   | 90K  | Yes   |



# Phân chia thuộc tính có giá trị liên tục



- Cách chọn giá trị v hiệu quả:
  - Sắp xếp các giá trị tăng dần
  - Chọn giá trị trung bình của từng giá trị của thuộc tính để phân chia và tính chỉ số gini
  - Chọn giá trị phân chia có chỉ số gini thấp nhất

|               |       | Taxable Income |       |       |       |       |       |       |       |       |       |     |   |   |
|---------------|-------|----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-----|---|---|
|               |       | 60             | 70    | 75    | 85    | 90    | 95    | 100   | 120   | 125   | 220   |     |   |   |
| Sorted Values | →     | 55             | 65    | 72    | 80    | 87    | 92    | 97    | 110   | 122   | 172   | 230 |   |   |
|               | →     | <=             | >     | <=    | >     | <=    | >     | <=    | >     | <=    | >     | <=  | > |   |
| Yes           | 0     | 3              | 0     | 3     | 0     | 3     | 1     | 2     | 2     | 1     | 3     | 0   | 3 | 0 |
| No            | 0     | 7              | 1     | 6     | 2     | 5     | 3     | 4     | 3     | 4     | 3     | 4   | 4 | 3 |
| Gini          | 0.420 | 0.400          | 0.375 | 0.343 | 0.417 | 0.400 | 0.300 | 0.343 | 0.375 | 0.400 | 0.420 |     |   |   |



# Biến đổi cây quyết định thành luật

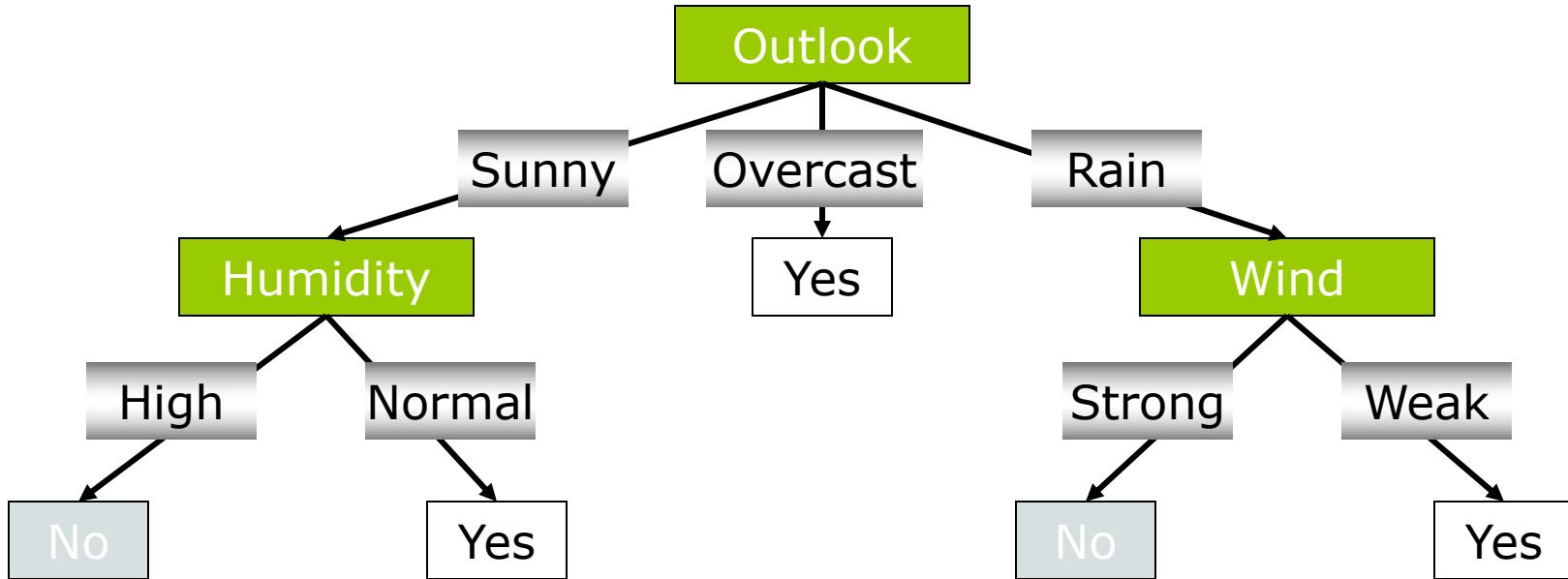
---



- Biểu diễn tri thức dưới dạng luật IF-THEN
- Mỗi luật tạo ra từ mỗi đường dẫn từ gốc đến lá
- Mỗi cặp giá trị thuộc tính dọc theo đường dẫn tạo nên phép kết (phép AND – và)
- Các nút lá mang tên của lớp



# Biến đổi cây quyết định thành luật



R<sub>1</sub>: If (Outlook=Sunny)  $\wedge$  (Humidity=High) Then Play=No

R<sub>2</sub>: If (Outlook=Sunny)  $\wedge$  (Humidity=Normal) Then Play=Yes

R<sub>3</sub>: If (Outlook=Overcast) Then Play=Yes

R<sub>4</sub>: If (Outlook=Rain)  $\wedge$  (Wind=Strong) Then Play=No

R<sub>5</sub>: If (Outlook=Rain)  $\wedge$  (Wind=Weak) Then Play=Yes



# Ưu điểm của cây quyết định

---



- Cây quyết định dễ hiểu
- Việc chuẩn bị dữ liệu cho một cây quyết định là cơ bản hoặc không cần thiết
- Cây quyết định có thể xử lý cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thẻ loại
- Cây quyết định là một mô hình hộp trắng
- Có thể thẩm định một mô hình bằng các kiểm tra thống kê
- Cây quyết định có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn



# Bài tập



- Cho bảng quyết định như sau:

| <b>Attr_A</b> | <b>Attr_B</b> | <b>Attr_C</b> | <b>Attr_D</b> |
|---------------|---------------|---------------|---------------|
| T             | F             | T             | T             |
| T             | F             | T             | T             |
| F             | T             | T             | F             |
| T             | T             | T             | T             |
| T             | T             | F             | F             |
| T             | T             | F             | F             |
| T             | F             | F             | T             |
| F             | T             | F             | F             |
| F             | F             | T             | T             |



# Giải



- Tính độ lợi thông tin

| Attr    | $p_i$       | $n_i$  | $I(p_i, n_i)$ | $R(a)$ | Gain(<br>a) |
|---------|-------------|--------|---------------|--------|-------------|
| Initial |             |        |               |        |             |
| Attr_A  | 4<br>T<br>F | 2<br>2 | 0.92<br>0.92  | 0.92   | 0.07        |
| Attr_B  | 1<br>T<br>F | 4<br>0 | 0.72<br>0     | 0.4    | 0.59        |
| Attr_C  | 4<br>T<br>F | 1<br>3 | 0.72<br>0.81  | 0.76   | 0.23        |



# Giải (tt)



Dữ liệu ID3 sau khi tách nhánh theo thuộc tính Attr\_B vì thuộc tính Attr\_B có độ lợi lớn nhất.

| Attr_A | Attr_C | Attr_D |
|--------|--------|--------|
| F      | T      | F      |
| T      | T      | T      |
| T      | F      | F      |
| T      | F      | F      |
| F      | F      | F      |



# Giải (tt)



| Attr    | $p_i$ | $n_i$ | $I(p_i, n_i)$ | $R(a)$ | Gain(a) |
|---------|-------|-------|---------------|--------|---------|
| Initial |       |       |               |        |         |
| Attr_A  | 1     | 2     | 0.92          | 0.552  | 0.17    |
| T       | 0     | 2     | 0             |        |         |
| F       |       |       |               |        |         |
| Attr_C  | 1     | 1     | 1.0           | 0.40   | 0.32    |
| T       | 0     | 3     | 0             |        |         |
| F       |       |       |               |        |         |



# Giải



Dữ liệu ID3 sau khi tách nhánh theo thuộc tính Attr\_C  
vì thuộc tính Attr\_C có độ lợi lớn nhất

| <b>Attr_A</b> | <b>Attr_D</b> |
|---------------|---------------|
| F             | F             |
| T             | T             |

# Bài tập



| age     | income | Region | credit_rating | Buy Mobile |
|---------|--------|--------|---------------|------------|
| <20     | high   | USA    | Low           | no         |
| <20     | high   | USA    | High          | no         |
| 21...50 | high   | USA    | Low           | yes        |
| >50     | medium | USA    | Low           | yes        |
| >50     | low    | PK     | Low           | yes        |
| >50     | low    | PK     | High          | no         |
| 21...50 | low    | PK     | High          | yes        |
| <20     | medium | USA    | Low           | no         |
| <20     | low    | PK     | Low           | yes        |
| >50     | medium | PK     | Low           | yes        |
| <20     | medium | PK     | High          | yes        |
| 21...50 | medium | USA    | High          | yes        |
| 21...50 | high   | PK     | Low           | yes        |
| >50     | medium | USA    | High          | no         |



# Sample Experience Table



| Example | Attributes |         |          |       | Target |
|---------|------------|---------|----------|-------|--------|
|         | Hour       | Weather | Accident | Stall |        |
| D1      | 8 AM       | Sunny   | No       | No    | Long   |
| D2      | 8 AM       | Cloudy  | No       | Yes   | Long   |
| D3      | 10 AM      | Sunny   | No       | No    | Short  |
| D4      | 9 AM       | Rainy   | Yes      | No    | Long   |
| D5      | 9 AM       | Sunny   | Yes      | Yes   | Long   |
| D6      | 10 AM      | Sunny   | No       | No    | Short  |
| D7      | 10 AM      | Cloudy  | No       | No    | Short  |
| D8      | 9 AM       | Rainy   | No       | No    | Medium |
| D9      | 9 AM       | Sunny   | Yes      | No    | Long   |
| D10     | 10 AM      | Cloudy  | Yes      | Yes   | Long   |
| D11     | 10 AM      | Rainy   | No       | No    | Short  |
| D12     | 8 AM       | Cloudy  | Yes      | No    | Long   |
| D13     | 9 AM       | Sunny   | No       | No    | Medium |

