



## Chap4 Classification - Xjdjdjdjd

khai thác dữ liệu và ứng dụng (Trường Đại học Công nghiệp Thành phố Hồ Chí Minh)



Scan to open on Studocu

---

# Khai thác dữ liệu và ứng dụng

---

## Phân loại

# Đề cương

---

- Giới thiệu về phân loại Các
- khái niệm cơ bản
- Kỹ thuật phân loại
  - Cây quyết định Cảm ứng
  - Phương pháp phân loại Bayes
  - Phân loại dựa trên quy tắc Các
  - kỹ thuật khác
- Đánh giá và lựa chọn mô hình

# Bắt tội trốn thuế

| Mã T | Đền bù | nghệ thuật<br>Trạng thái | Tờ thuế<br>Thu nhập | Gian lận |
|------|--------|--------------------------|---------------------|----------|
| 1    | Đúng   | Đơn                      | 125K                | KHÔNG    |
| 2    | KHÔNG  | Đã cưới                  | 100K                | KHÔNG    |
| 3    | KHÔNG  | Đơn                      | 70K                 | KHÔNG    |
| 4    | Đúng   | Đã cưới                  | 120K                | KHÔNG    |
| 5    | KHÔNG  | D đã ly hôn              | 95K                 | Đúng     |
| 6    | KHÔNG  | Đã cưới                  | 60K                 | KHÔNG    |
| 7    | Đúng   | D đã ly hôn              | 220K                | KHÔNG    |
| số 8 | KHÔNG  | Đơn                      | 85K                 | Đúng     |
| 9    | KHÔNG  | Đã cưới                  | 75K                 | KHÔNG    |
| 10   | KHÔNG  | Đơn                      | 90K                 | Đúng     |

Số liệu kê khai thuế năm 2011

Tờ khai thuế mới cho năm 2012 Đây có phải là tờ khai thuế gian lận?

| Đền bù | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | Gian lận |
|--------|------------------------|-----------------------|----------|
| KHÔNG  | Đã cưới                | 80K                   | ?        |

Một ví dụ về vấn đề phân loại: tìm hiểu phương pháp phân biệt giữa các bản ghi của các loại khác nhau các lớp học (kể lừa đảo vs không gian lận)

# phân loại là gì?

-**Phân loại** là nhiệm vụ của học hỏi mục tiêu chức năng tập thuộc tính bản đồ đó đến một trong các nhãn lớp được xác định trước. Chức năng mục tiêu được biết đến như một mô hình phân loại.

categorical  
categorical  
continuous  
class

| Tid  | Đền bù | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | Gian lận |
|------|--------|------------------------|-----------------------|----------|
| 1    | Đúng   | Đơn                    | 125K                  | KHÔNG    |
| 2    | KHÔNG  | Đã cưới                | 100K                  | KHÔNG    |
| 3    | KHÔNG  | Đơn                    | 70K                   | KHÔNG    |
| 4    | Đúng   | Đã cưới                | 120K                  | KHÔNG    |
| 5    | KHÔNG  | Đã ly hôn              | 95K                   | Đúng     |
| 6    | KHÔNG  | Đã cưới                | 60K                   | KHÔNG    |
| 7    | Đúng   | Đã ly hôn              | 220K                  | KHÔNG    |
| số 8 | KHÔNG  | Đơn                    | 85K                   | Đúng     |
| 9    | KHÔNG  | Đã cưới                | 75K                   | KHÔNG    |
| 10   | KHÔNG  | Đơn                    | 90K                   | Đúng     |

Một trong những thuộc tính là **thuộc tính lớp**

Trong trường hợp này: Lừa đảo

Hai **nhãn lớp** (hoặc các lớp học): **Có (1), Không (0)**

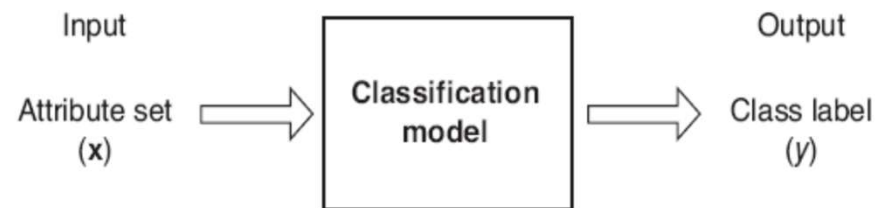


Figure 4.2. Classification as the task of mapping an input attribute set  $x$  into its class label  $y$ .

# Ví dụ về nhiệm vụ phân loại

---

- Dự đoán **khối** **u** tế bào như **nhẹ** hoặc **ác** tính
- Phân loại thẻ tín dụng **giao dịch** **BĂNG** **hợp pháp** hoặc **lừa đảo**
- Phân loại **hững** **câu** **chuyện** **mới** **BĂNG** **tài chính**, **thời tiết**, **sự** **giải trí**, **các** **môn thể thao**, **vân vân**
- Nhận dạng **thư rác** **e-mail**, web rác **trang**, **người lớn** **nội dung**
- Hiểu nếu một trang web **truy vấn** **có** **mục đích** **thương mại** hay không

# Phân loại—Quy trình hai bước

---

- **Xây dựng mô hình**: mô tả một tập hợp các lớp được xác định trước
  - Mỗi bộ/mẫu được coi là thuộc về một lớp được xác định trước, được xác định bởi **thuộc tính nhãn lớp**
  - Tập các bộ dữ liệu được sử dụng để xây dựng mô hình: tập huấn luyện
  - Mô hình được biểu diễn dưới dạng quy tắc phân loại, cây quyết định hoặc công thức toán học
- Cách sử dụng mô hình: để phân loại các đối tượng trong tương lai hoặc chưa biết
  - **Ước tính độ chính xác** của mô hình
    - Nhãn đã biết của mẫu thử được so sánh với kết quả đã phân loại từ mô hình
    - **Sự chính xác** tỷ lệ là tỷ lệ phần trăm của các mẫu thử nghiệm được phân loại chính xác theo mô hình
    - **Tập kiểm tra** độc lập với tập huấn luyện, nếu không sẽ xảy ra hiện tượng khớp quá mức. Nếu độ chính xác chấp nhận được, hãy sử dụng mô hình để **phân loại dữ liệu mới**
- Lưu ý: Nếu bộ thử nghiệm được sử dụng để chọn mô hình, nó được gọi là **bộ xác nhận (kiểm tra)**

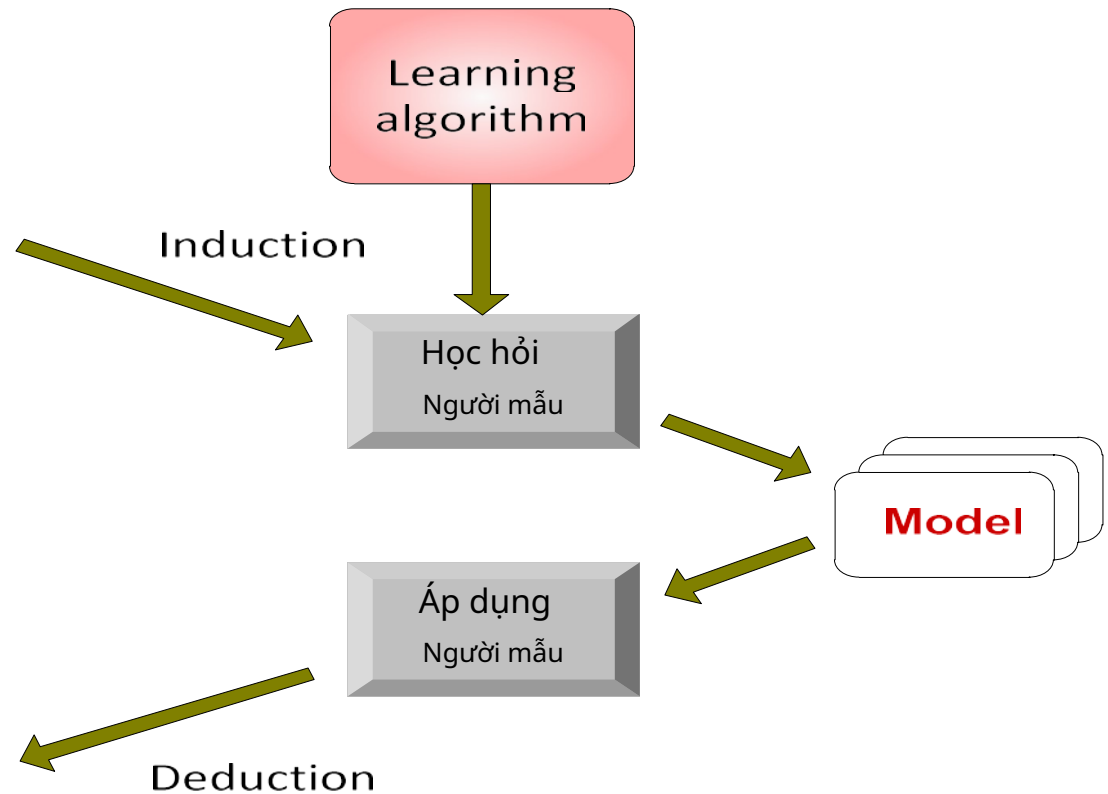
# Minh họa nhiệm vụ phân loại

| Tid  | Attrib1 | Attrib2    | Attrib3 | Lớp học |
|------|---------|------------|---------|---------|
| 1    | Đúng    | Lớn        | 125K    | KHÔNG   |
| 2    | KHÔNG   | Trung bình | 100K    | KHÔNG   |
| 3    | KHÔNG   | Bé nhỏ     | 70K     | KHÔNG   |
| 4    | Đúng    | Trung bình | 120K    | KHÔNG   |
| 5    | KHÔNG   | Lớn        | 95K     | Đúng    |
| 6    | KHÔNG   | Trung bình | 60K     | KHÔNG   |
| 7    | Đúng    | Lớn        | 220K    | KHÔNG   |
| số 8 | KHÔNG   | Bé nhỏ     | 85K     | Đúng    |
| 9    | KHÔNG   | Trung bình | 75K     | KHÔNG   |
| 10   | KHÔNG   | Bé nhỏ     | 90K     | Đúng    |

Training Set

| Tid | Attrib1 | Attrib2    | Attrib3 | Lớp học |
|-----|---------|------------|---------|---------|
| 11  | KHÔNG   | Bé nhỏ     | 55K     | ?       |
| 12  | Đúng    | Trung bình | 80K     | ?       |
| 13  | Đúng    | Lớn        | 110K    | ?       |
| 14  | KHÔNG   | Bé nhỏ     | 95K     | ?       |
| 15  | KHÔNG   | Lớn        | 67K     | ?       |

Test Set





# Đánh giá các mô hình phân loại

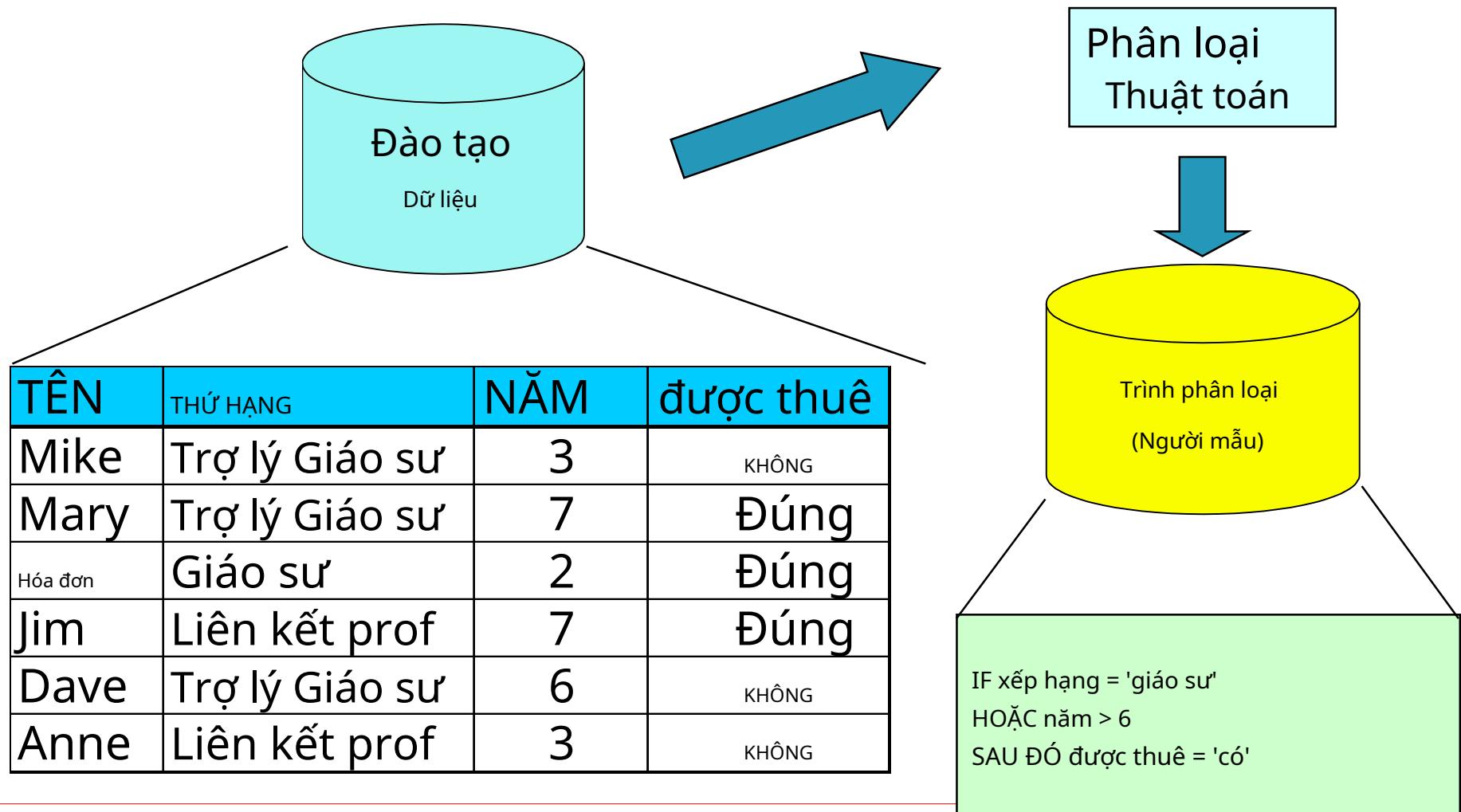
- Số lượng **hồ sơ kiểm tra** được dự đoán đúng (hoặc không chính xác) bởi mô hình phân loại
- Ma trận hỗn loạn

|             | Lớp dự đoán |          |
|-------------|-------------|----------|
|             | Lớp = 1     | Lớp = 0  |
| Lớp thực tế |             |          |
| Lớp = 1     | $f_{11}$    | $f_{10}$ |
| Lớp = 0     | $f_{01}$    | $f_{00}$ |

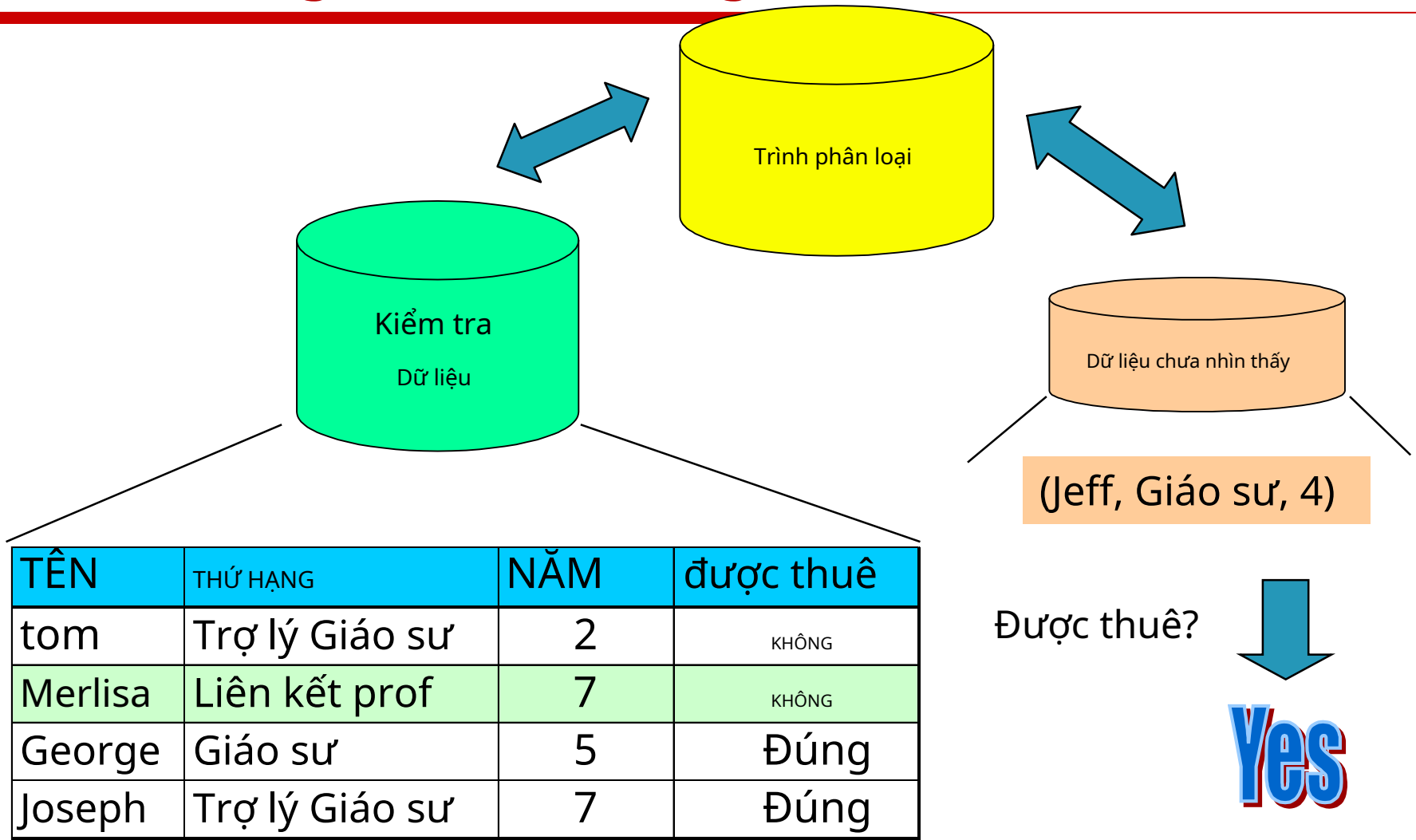
$$\text{Sự chính xác} - \frac{\text{\#dự đoán đúng}}{\text{tổng số dự đoán}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$\text{Tỷ lệ lỗi} - \frac{\text{\#dự đoán sai}}{\text{tổng số dự đoán}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

# Quy trình phân loại (1): Xây dựng mô hình



# CP(2): Sử dụng mô hình trong dự đoán



# Các vấn đề liên quan đến phân loại và dự đoán

---

## Vấn đề (1): Chuẩn bị dữ liệu

- Làm sạch dữ liệu
  - Tiền xử lý dữ liệu để giảm nhiễu và xử lý các giá trị bị thiếu
- Phân tích mức độ liên quan (lựa chọn tính năng)
  - Loại bỏ các thuộc tính không liên quan hoặc dư thừa
- Chuyển đổi dữ liệu
  - Tổng quát hóa và/hoặc chuẩn hóa dữ liệu

# Các vấn đề liên quan đến phân loại và dự đoán (tiếp)

---

## Vấn đề (2): Đánh giá phương pháp phân loại

- Độ chính xác dự đoán
- Tốc độ và khả năng mở rộng
  - Thời gian xây dựng mô hình
  - Thời gian sử dụng mô hình
- Độ bền
  - xử lý tiếng ồn và các giá trị bị thiếu
- hiệu quả trong cơ sở dữ liệu lưu trữ trên đĩa
- Khả năng giải thích:
  - sự hiểu biết và hiểu biết sâu sắc được cung cấp bởi mô hình
- Sự tốt đẹp của các quy tắc
  - kích thước cây quyết định
  - sự cô đọng của các quy tắc phân loại

# Kỹ thuật phân loại

---

- Phương pháp dựa trên cây quyết định
  - Phương pháp phân loại Bayes
  - Phương pháp dựa trên quy tắc
  - k-hàng xóm gần nhất (kNN)
  - Máy vectơ hỗ trợ (SVM)
  - Phân loại theo lan truyền ngược
  - Máy vectơ hỗ trợ (SVM)
  - Thuật toán di truyền
  - Cách tiếp cận tập thô
  - Các phương pháp tiếp cận tập mờ
  - ....

# Kỹ thuật phân loại

---

## -Phương pháp dựa trên cây quyết định

- Phương pháp phân loại Bayes Phương
- pháp dựa trên quy tắc
- k-hàng xóm gần nhất (kNN) Máy
- vectơ hỗ trợ (SVM) Phân loại
- theo lan truyền ngược Máy
- vectơ hỗ trợ (SVM) Thuật toán di
- truyền
- Cách tiếp cận tập thô
- Các phương pháp tiếp cận tập mờ
- ....

# Phân loại theo quy nạp cây quyết định

- Cây quyết định
  - Cấu trúc cây giống như biểu đồ luồng Nút bên
  - trong biểu thị kiểm tra trên một thuộc tính Nhánh
  - đại diện cho kết quả của kiểm tra
  - Các nút lá đại diện cho nhãn lớp hoặc phân bố lớp
- Việc tạo cây quyết định bao gồm hai giai đoạn
  - Xây dựng cây
    - Lúc đầu, tất cả các ví dụ huấn luyện đều ở gốc
    - Ví dụ về phân vùng đệ quy dựa trên các thuộc tính đã
  - chọn
    - Xác định và loại bỏ các nhánh phản ánh tiếng ồn hoặc các ngoại lệ
- Sử dụng cây quyết định: Phân loại một mẫu chưa biết
  - Kiểm tra các giá trị thuộc tính của mẫu dựa trên cây quyết định



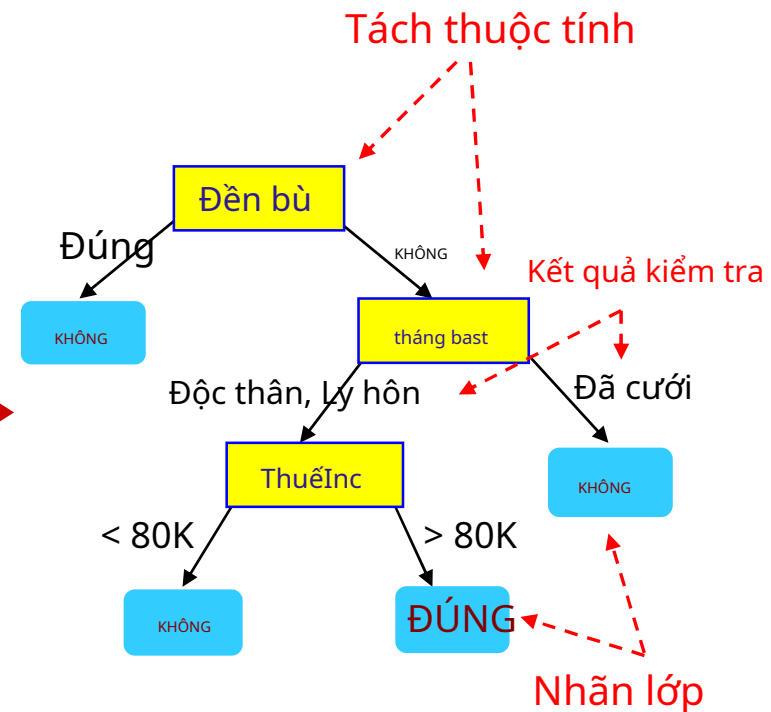
# Ví dụ về cây quyết định

categorical categorical continuous class

| Tid  | Đền bù | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | Gian lận |
|------|--------|------------------------|-----------------------|----------|
| 1    | Đúng   | Đơn                    | 125K                  | KHÔNG    |
| 2    | KHÔNG  | Đã cưới                | 100K                  | KHÔNG    |
| 3    | KHÔNG  | Đơn                    | 70K                   | KHÔNG    |
| 4    | Đúng   | Đã cưới                | 120K                  | KHÔNG    |
| 5    | KHÔNG  | Đã ly hôn              | 95K                   | Đúng     |
| 6    | KHÔNG  | Đã cưới                | 60K                   | KHÔNG    |
| 7    | Đúng   | Đã ly hôn              | 220K                  | KHÔNG    |
| số 8 | KHÔNG  | Đơn                    | 85K                   | Đúng     |
| 9    | KHÔNG  | Đã cưới                | 75K                   | KHÔNG    |
| 10   | KHÔNG  | Đơn                    | 90K                   | Đúng     |

Dữ liệu đào tạo

Hướng dẫn

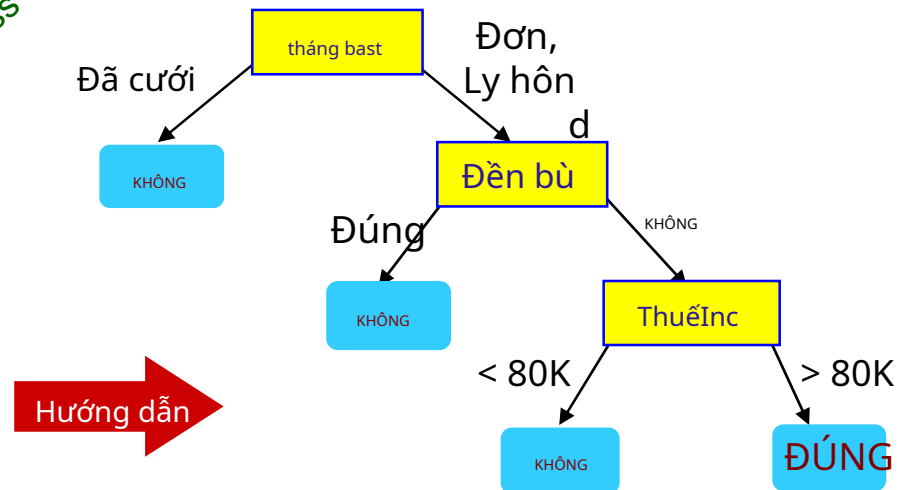


Model: Cây quyết định

# Một ví dụ khác về cây quyết định

| Tid  | Đền bù | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | Gian lận |
|------|--------|------------------------|-----------------------|----------|
| 1    | Đúng   | Đơn                    | 125K                  | KHÔNG    |
| 2    | KHÔNG  | Đã cưới                | 100K                  | KHÔNG    |
| 3    | KHÔNG  | Đơn                    | 70K                   | KHÔNG    |
| 4    | Đúng   | Đã cưới                | 120K                  | KHÔNG    |
| 5    | KHÔNG  | Đã ly hôn              | 95K                   | Đúng     |
| 6    | KHÔNG  | Đã cưới                | 60K                   | KHÔNG    |
| 7    | Đúng   | Đã ly hôn              | 220K                  | KHÔNG    |
| số 8 | KHÔNG  | Đơn                    | 85K                   | Đúng     |
| 9    | KHÔNG  | Đã cưới                | 75K                   | KHÔNG    |
| 10   | KHÔNG  | Đơn                    | 90K                   | Đúng     |

categorical  
categorical  
continuous  
class



Có thể có nhiều cây phù hợp với cùng một dữ liệu!

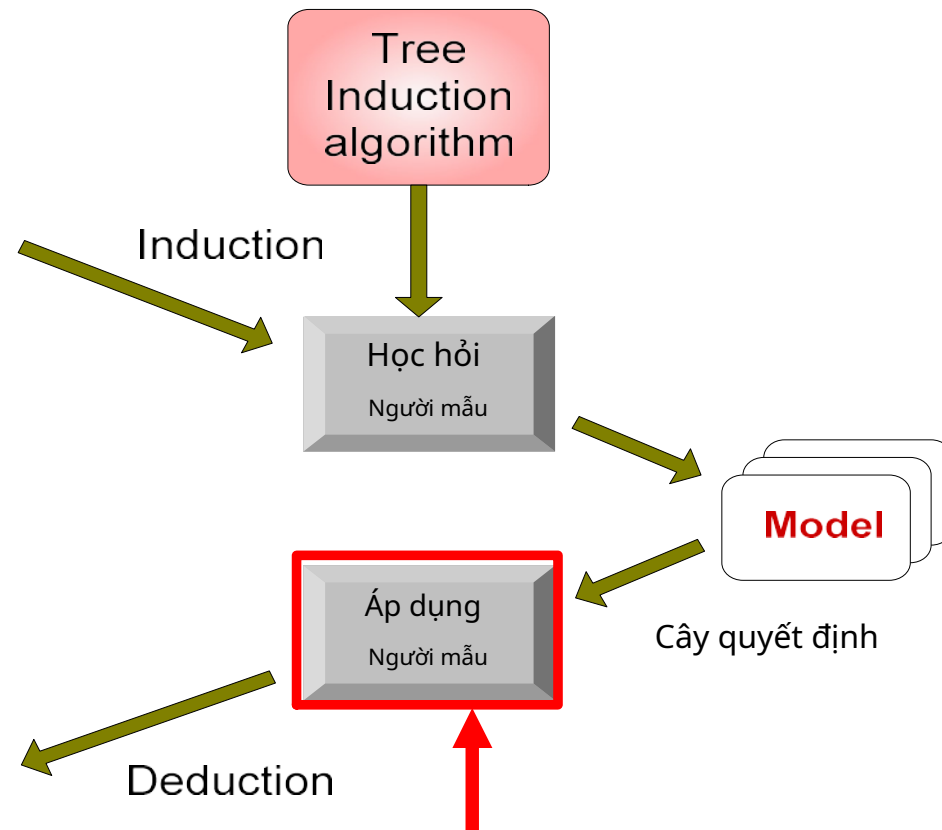
# Nhiệm vụ phân loại cây quyết định

| Tid  | Attrib1 | Attrib2    | Attrib3 | Lớp học |
|------|---------|------------|---------|---------|
| 1    | Đúng    | Lớn        | 125K    | KHÔNG   |
| 2    | KHÔNG   | Trung bình | 100K    | KHÔNG   |
| 3    | KHÔNG   | Bé nhỏ     | 70K     | KHÔNG   |
| 4    | Đúng    | Trung bình | 120K    | KHÔNG   |
| 5    | KHÔNG   | Lớn        | 95K     | Đúng    |
| 6    | KHÔNG   | Trung bình | 60K     | KHÔNG   |
| 7    | Đúng    | Lớn        | 220K    | KHÔNG   |
| số 8 | KHÔNG   | Bé nhỏ     | 85K     | Đúng    |
| 9    | KHÔNG   | Trung bình | 75K     | KHÔNG   |
| 10   | KHÔNG   | Bé nhỏ     | 90K     | Đúng    |

Training Set

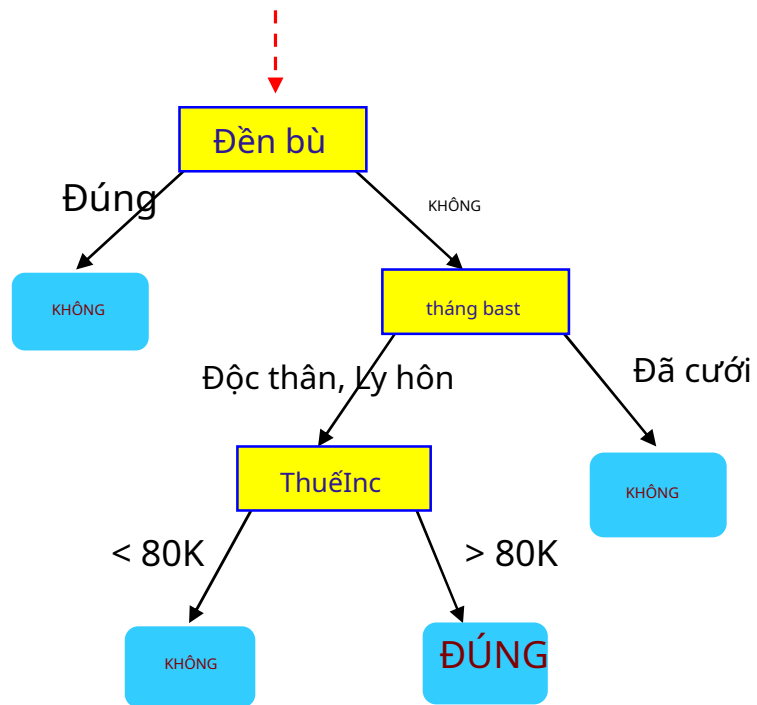
| Tid | Attrib1 | Attrib2    | Attrib3 | Lớp học |
|-----|---------|------------|---------|---------|
| 11  | KHÔNG   | Bé nhỏ     | 55K     | ?       |
| 12  | Đúng    | Trung bình | 80K     | ?       |
| 13  | Đúng    | Lớn        | 110K    | ?       |
| 14  | KHÔNG   | Bé nhỏ     | 95K     | ?       |
| 15  | KHÔNG   | Lớn        | 67K     | ?       |

Test Set



# Áp dụng mô hình để kiểm tra dữ liệu

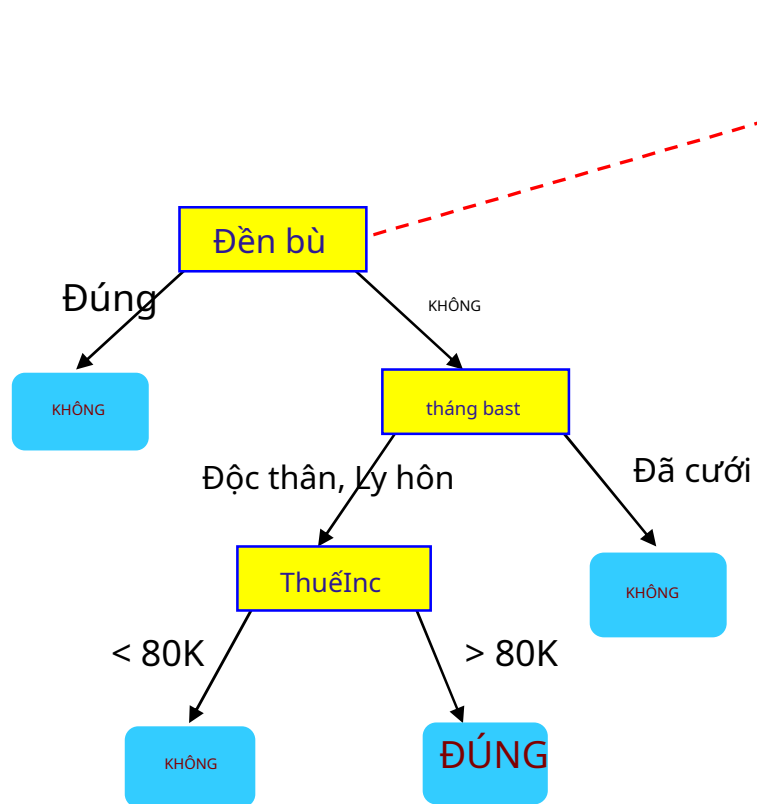
Bắt đầu từ gốc cây.



Dữ liệu thử nghiệm

| Hoàn tiền hôn nhân | Chịu thuế | Thu nhập | Gian lận |
|--------------------|-----------|----------|----------|
| Trạng thái         | Thu nhập  |          |          |
| KHÔNG              | Đã cưới   | 80K      | ?        |

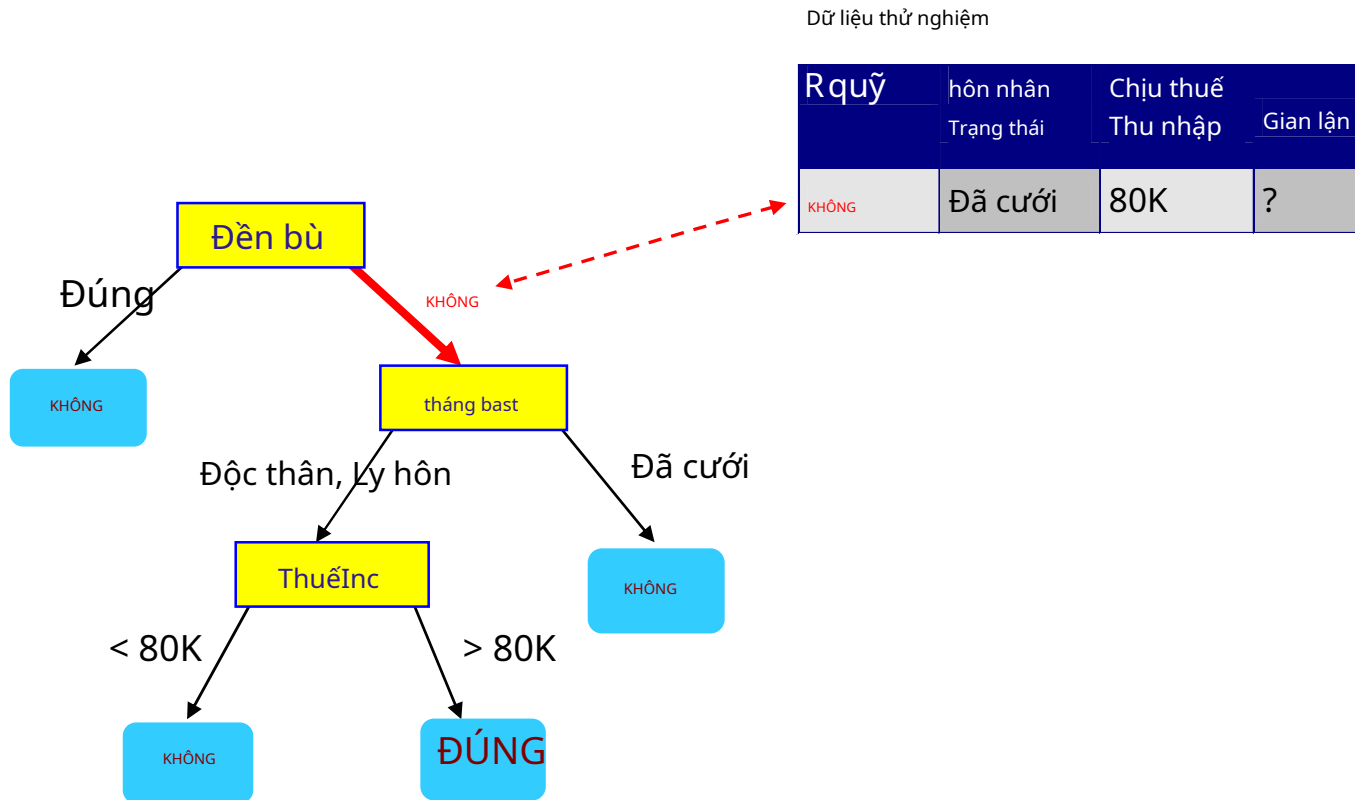
# Áp dụng mô hình để kiểm tra dữ liệu



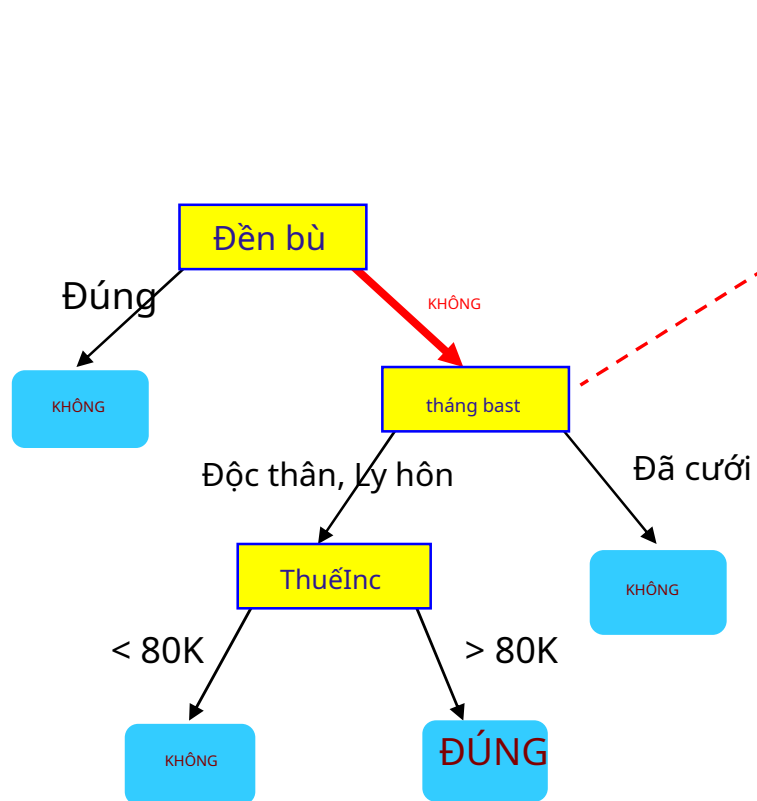
Dữ liệu thử nghiệm

| Rquỹ  | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | Gian lận |
|-------|------------------------|-----------------------|----------|
| KHÔNG | Đã cưới                | 80K                   | ?        |

# Áp dụng mô hình để kiểm tra dữ liệu



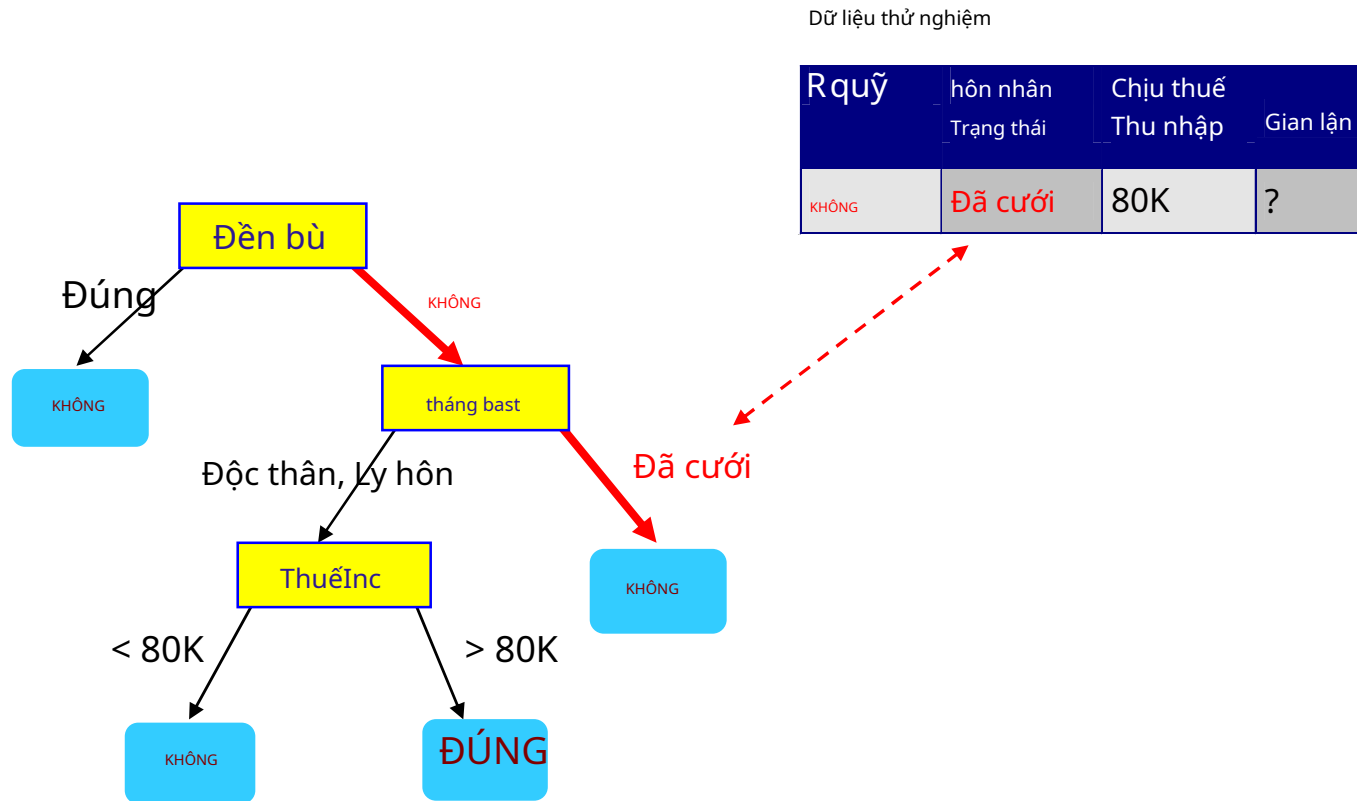
# Áp dụng mô hình để kiểm tra dữ liệu



Dữ liệu thử nghiệm

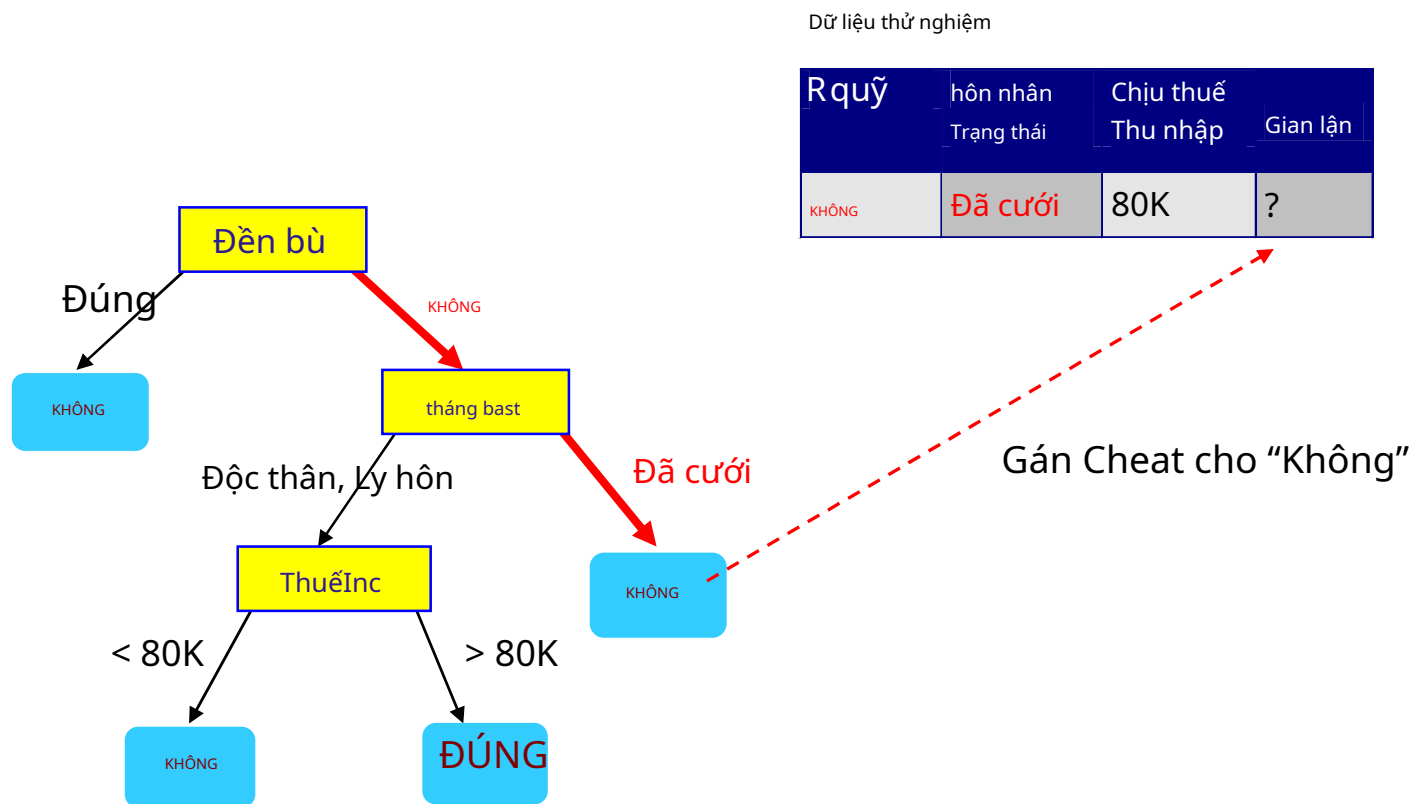
| Rquỹ  | hôn nhân   | Chịu thuế | Gian lận |
|-------|------------|-----------|----------|
|       | Trạng thái | Thu nhập  |          |
| KHÔNG | Đã cưới    | 80K       | ?        |

# Áp dụng mô hình để kiểm tra dữ liệu





# Áp dụng mô hình để kiểm tra dữ liệu



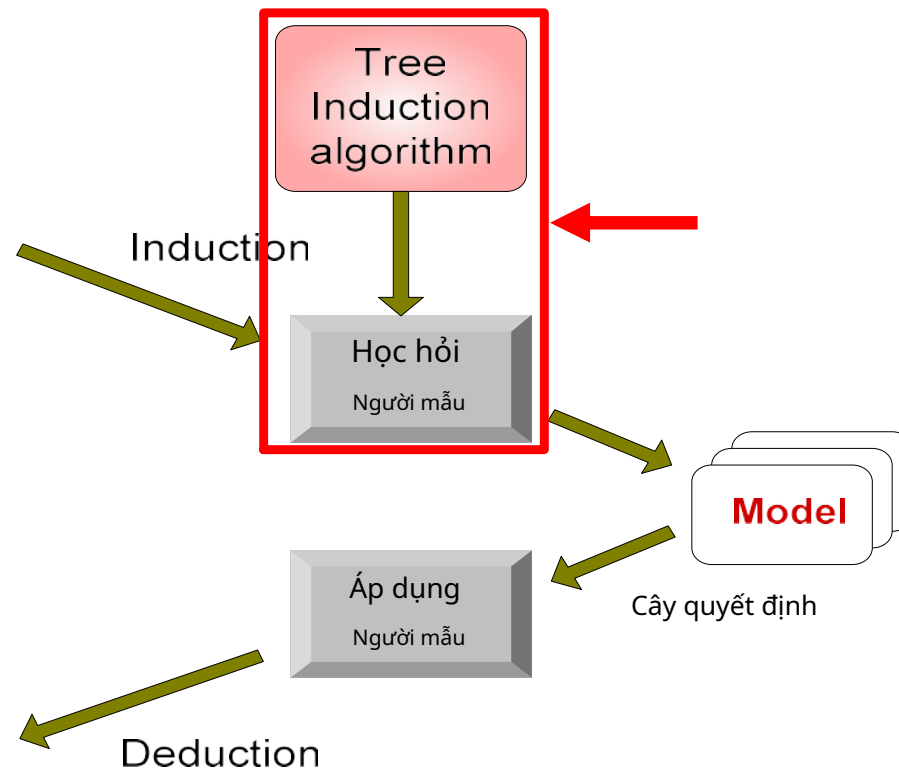
# Nhiệm vụ phân loại cây quyết định

| Tid  | Attrib1 | Attrib2    | Attrib3 | Lớp học |
|------|---------|------------|---------|---------|
| 1    | Đúng    | Lớn        | 125K    | KHÔNG   |
| 2    | KHÔNG   | Trung bình | 100K    | KHÔNG   |
| 3    | KHÔNG   | Bé nhỏ     | 70K     | KHÔNG   |
| 4    | Đúng    | Trung bình | 120K    | KHÔNG   |
| 5    | KHÔNG   | Lớn        | 95K     | Đúng    |
| 6    | KHÔNG   | Trung bình | 60K     | KHÔNG   |
| 7    | Đúng    | Lớn        | 220K    | KHÔNG   |
| số 8 | KHÔNG   | Bé nhỏ     | 85K     | Đúng    |
| 9    | KHÔNG   | Trung bình | 75K     | KHÔNG   |
| 10   | KHÔNG   | Bé nhỏ     | 90K     | Đúng    |

Training Set

| Tid | Attrib1 | Attrib2    | Attrib3 | Lớp học |
|-----|---------|------------|---------|---------|
| 11  | KHÔNG   | Bé nhỏ     | 55K     | ?       |
| 12  | Đúng    | Trung bình | 80K     | ?       |
| 13  | Đúng    | Lớn        | 110K    | ?       |
| 14  | KHÔNG   | Bé nhỏ     | 95K     | ?       |
| 15  | KHÔNG   | Lớn        | 67K     | ?       |

Test Set



# Thuật toán tạo cây quyết định

---

- Thuật toán cơ bản (thuật toán tham lam)
  - Cây được xây dựng theo kiểu **Cách phân chia và chinh phục đệ quy từ trên xuống**
  - Lúc đầu, tất cả các ví dụ huấn luyện đều ở gốc
  - Các thuộc tính có tính phân loại (nếu có giá trị liên tục, chúng sẽ được rời rạc hóa trước)
  - Các ví dụ được phân vùng đệ quy dựa trên các thuộc tính đã chọn
  - Các thuộc tính kiểm tra được chọn trên cơ sở thước đo heuristic hoặc thống kê (ví dụ: **thu được thông tin**)

# Cảm ứng cây quyết định

---

- Vấn đề
  - Làm cách nào để **Phân loại** một nút lá
    - Chỉ định **lớp đa số**
    - Nếu lá trống, hãy gán **lớp mặc định** – lớp có mức độ phổ biến cao nhất.
  - Xác định cách phân chia các bản ghi
    - Làm cách nào để chỉ định điều kiện kiểm tra thuộc tính?
    - Thuộc tính nào sẽ được sử dụng trong phân chia nút chia
      - Làm thế nào để xác định sự phân chia tốt nhất?
    - Chúng ta nên sử dụng chia 2 chiều hay chia nhiều chiều?
- Xác định thời điểm dừng chia tách

# Thuật toán tạo cây quyết định

---

## -Điều kiện dừng phân vùng

- Tất cả các mẫu cho một nút nhất định đều thuộc cùng một lớp
- Không còn thuộc tính nào để phân vùng thêm nữa -  
biểu quyết đa số được sử dụng để phân loại lá
- Không còn mẫu nào

# Biện pháp lựa chọn thuộc tính

---

## - Thu được thông tin(ID3/C4.5)

- Tất cả các thuộc tính được coi là phân loại Có thể được
- sửa đổi cho các thuộc tính có giá trị liên tục

## - chỉ số Gini(GIỎ HÀNG, SLIQ, SPRINT.))

- Tất cả các thuộc tính được giả định có giá trị liên tục
- Giả sử tồn tại một số giá trị phân chia có thể có cho mỗi thuộc tính Có thể cần các công cụ
- khác, chẳng hạn như phân cụm, để có được các giá trị phân chia có thể có Có thể được sửa
- đổi cho các thuộc tính phân loại

## Biện pháp lựa chọn thuộc tính: Tăng thông tin (ID3/C4.5)

- Chọn thuộc tính có mức tăng thông tin cao nhất
- Cho phép  $P_{T\hat{o}i}$  là xác suất để một bộ tùy ý trong  $D$  thuộc lớp  $C_{T\hat{o}i}$ , ước tính bởi  $|C_{T\hat{o}i,D}|/|D|$
- **Thông tin dự kiến** (entropy) cần thiết để phân loại một bộ dữ liệu trong  $D$ :  

$$\text{Thông tin}(D) = -\sum_{T\hat{o}i} P_{T\hat{o}i} \log_2(P_{T\hat{o}i})$$
- **Thông tin** cần thiết (sau khi dùng  $A$  để chia  $D$  thành  $v$  phân vùng) để phân loại  $D$ :

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

- **Thông tin thu được** bằng cách phân nhánh trên thuộc tính  $A$

$$\text{Đạt được}(A) = \text{Thông tin}(D) - \sum_{j=1}^v \frac{|D_j|}{|D|} \text{Thông tin}(D_j)$$

# Tập dữ liệu đào tạo

| 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            |



# Lựa chọn thuộc tính bằng tính toán tăng thông tin

- Lớp P: buys\_computer = "có"
- Lớp N: buy\_computer = "không"
- $|D| = 14$ ,  $|C_{P,D}| = 9$ ,  $|C_{N,D}| = 5$

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

- Tính toán tôităng thông tin từ tuổi:

| tuổi    | P <sub>Tôi</sub> | N <sub>Tôi</sub> | Tôi (p <sub>Tôi</sub> , N <sub>Tôi</sub> ) |
|---------|------------------|------------------|--|
| <=30    | 2                | 3                | 0,971                                      |
| 30...40 | 4                | 0                | 0  |
| > 40    | 3                | 2                | 0,971                                      |

$$I(2,3) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.971$$

$$I(4,0) = -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} = 0$$

$$I(3,2) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971$$

$$\begin{aligned} \text{Thông tin tuổi}(\mathcal{D}) &= \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) \\ &+ \frac{5}{14} I(3,2) = 0.694 \end{aligned}$$

$\frac{5}{14} I(2,3)$  nghĩa là "tuổi  $\leq 30$ " có 5 out  
gồm 14 mẫu, trong đó có 2 mẫu có và 3  
mẫu không.

Kể từ đây

$$\text{Tăng(tuổi)} = \text{Thông tin}(\mathcal{D}) - \text{Thông tin}_{\text{tuổi}}(\mathcal{D}) = 0.246$$

# Lựa chọn thuộc tính bằng tính toán tăng thông tin

---

- Lớp P: buys\_computer = “có”
- Lớp N: buy\_computer = “không”
- $Tôi(p,n) = I(9, 5) = 0,940$
- Tính cái tôi tăng thông tin vì:
  - Tuổi = 0,25
  - Thu nhập = ?
  - Sinh viên = ?
  - tín dụng\_xếp hạng = ?

age?

$\leq 30$

31..40

$> 40$

| income | student | credit_rating | buys_computer |
|--------|---------|---------------|---------------|
| high   | no      | fair          | no            |
| high   | no      | excellent     | no            |
| medium | no      | fair          | no            |
| low    | yes     | fair          | yes           |
| medium | yes     | excellent     | yes           |

| income | student | credit_rating | buys_computer |
|--------|---------|---------------|---------------|
| medium | no      | fair          | yes           |
| low    | yes     | fair          | yes           |
| low    | yes     | excellent     | no            |
| medium | yes     | fair          | yes           |
| medium | no      | excellent     | no            |

| income | student | credit_rating | buys_computer |
|--------|---------|---------------|---------------|
| high   | no      | fair          | yes           |
| low    | yes     | excellent     | yes           |
| medium | no      | excellent     | yes           |
| high   | yes     | fair          | yes           |

age?

Tiếp tục với  
bảng con 1 –D1

$\leq 30$

31..40

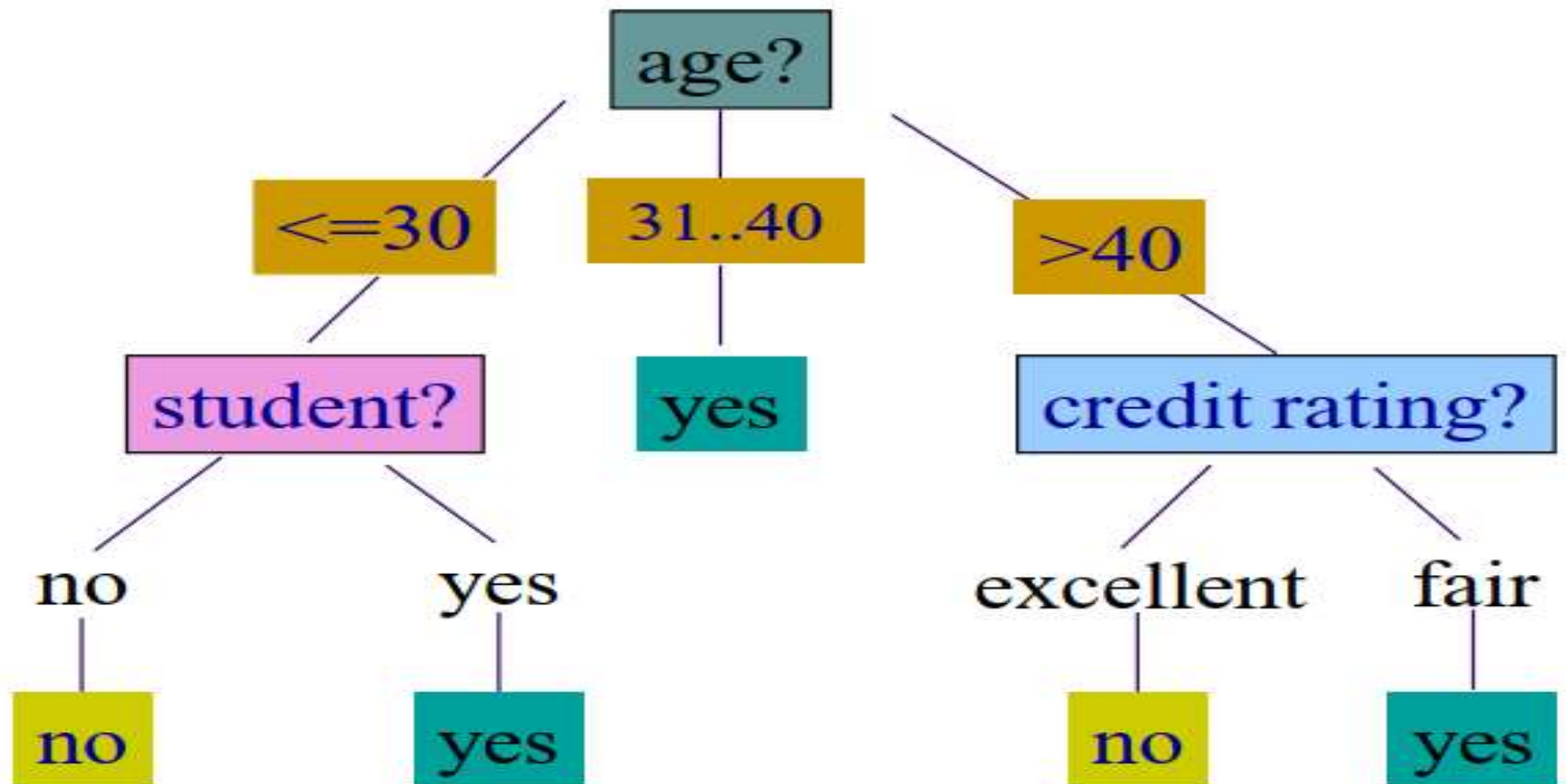
$> 40$

Tiếp tục với  
bảng con 2 –D2

| income | student | credit_rating | buys_computer |
|--------|---------|---------------|---------------|
| high   | no      | fair          | no            |
| high   | no      | excellent     | no            |
| medium | no      | fair          | no            |
| low    | yes     | fair          | yes           |
| medium | yes     | excellent     | yes           |

| income | student | credit_rating | buys_computer |
|--------|---------|---------------|---------------|
| medium | no      | fair          | yes           |
| low    | yes     | fair          | yes           |
| low    | yes     | excellent     | no            |
| medium | yes     | fair          | yes           |
| medium | no      | excellent     | no            |

yes



# Tính toán thu được thông tin cho các thuộc tính có giá trị liên tục

- Đặt thuộc tính A là thuộc tính có giá trị liên tục
- Phải xác định **điểm chia tốt nhất** cho một
  - Sắp xếp giá trị A theo thứ tự tăng dần
  - Thông thường, điểm giữa giữa mỗi cặp giá trị liên kề được coi là có thể điểm chia đôi
    - $(Môi + a_{môi+1})/2$  là trung điểm giữa các giá trị của  $a_{môi}$  và  $a_{môi+1}$
- Điểm với yêu cầu thông tin dự kiến tối thiểu với A được chọn làm điểm phân chia cho A
- Tách ra:
  - D1 là tập hợp các bộ dữ liệu trong D thỏa mãn  $A > \text{điểm phân tách}$  và D2 là tập hợp các bộ dữ liệu trong D thỏa mãn  $A < \text{điểm phân tách}$



# Tỷ lệ tăng cho lựa chọn thuộc tính (C4.5)

- Thước đo độ lợi thông tin thiên về các thuộc tính có số lượng giá trị lớn
- C4.5 (phiên bản kế thừa của ID3) sử dụng tỷ lệ khuếch đại để khắc phục vấn đề (chuẩn hóa thành tăng thông tin)

$$\text{Thông tin phân chia}_{\text{MOT}(D)} = - \sum_{j=1}^v \frac{|D_j|}{|D|} \log_2 \left( \frac{|D_j|}{|D|} \right)$$

- $\text{GainRatio}(A) = \text{Gain}(A) / \text{SplitInfo}(A)$

- Bán tại.

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

- Tỷ lệ lãi(thu nhập) =  $0,029 / 1,557 = 0,019$
- Thuộc tính có tỷ lệ khuếch đại tối đa được chọn làm thuộc tính tách

# Chỉ số Gini (GIỎI, IBM IntelligenceMiner)

- Nếu một tập dữ liệu  $D$  chứa các ví dụ từ lớp học, chỉ số gini,  $gini(D)$  được định nghĩa là
 
$$gini(D) = 1 - \sum_{j=1}^N p_j^2$$

Ở đây  $p_j$  là tần số tương đối của lớp  $j$  TRONG  $D$

- Nếu một tập dữ liệu  $D$  được chia trên  $A$  thành hai tập con  $D_1$  và  $D_2$ , các chỉ số gini mục lục  $gini(D)$  được định nghĩa là

$$gini(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

- Giảm tạp chất:

$$-gini(A) - gini(D) - gini(D)$$

- Thuộc tính cung cấp giá trị nhỏ nhất (hoặc mức giảm tạp chất lớn nhất) được chọn để phân chia nút (cần liệt kê tất cả các điểm phân chia có thể có cho từng thuộc tính)



# Tính toán chỉ số Gini

- Bán tại. D có 9 bộ trong buys\_computer = "có" và 5 trong "không"

$$\text{gini}(D) = 1 - \frac{9^2 + 5^2}{14^2} = 0,459$$

- Giả sử thuộc tính thu nhập D chia thành 10 trong D<sub>1</sub>: {thấp, trung bình} và 4 ở D<sub>2</sub>

$$\begin{aligned} \text{rượu gini thu nhập} &= \text{Gini}(D_1) \cdot \frac{10}{14} + \text{Gini}(D_2) \cdot \frac{4}{14} \\ &= \frac{10}{14} \left( 1 - \left( \frac{7}{10} \right)^2 - \left( \frac{3}{10} \right)^2 \right) + \frac{4}{14} \left( 1 - \left( \frac{2}{4} \right)^2 - \left( \frac{2}{4} \right)^2 \right) \\ &= 0,443 \\ &= \text{Gini}_{\text{income} \in \{\text{high}\}}(D). \end{aligned}$$

Gini<sub>{Cao thấp}</sub> là 0,458; Gini<sub>{Trung bình khá}</sub> là 0,450. Vì vậy, phân chia trên {low, medium} (và {high}) vì nó có chỉ số Gini thấp nhất

- Tất cả các thuộc tính được giả định có giá trị liên tục
- Có thể cần các công cụ khác, ví dụ: phân cụm, để có được các giá trị phân chia có thể
- Có thể được sửa đổi cho các thuộc tính phân loại

# So sánh lựa chọn thuộc tính n Biện pháp

- Ba biện pháp nói chung đều cho kết quả tốt nhưng
  - Đạt được thông tin:
    - thiên về các thuộc tính đa giá trị Tỷ lệ
  - đạt được:
    - có xu hướng thích sự phân chia không cân bằng trong đó một phân vùng nhỏ hơn nhiều so với các phân vùng khác
- Chỉ số Gini:
  - thiên về các thuộc tính đa giá trị
  - gặp khó khăn khi số lớp đông
  - có xu hướng ưu tiên các thử nghiệm dẫn đến các phân vùng có kích thước bằng nhau và độ tinh khiết trong cả hai phân vùng

# Trang bị quá mức và cắt tỉa cây

---

- Trang bị quá mức : Cây cảm ứng có thể phù hợp quá mức với dữ liệu huấn luyện
  - Quá nhiều nhánh, một số nhánh có thể phản ánh sự bất thường do nhiễu hoặc ngoại lệ Độ chính xác kém
  - đối với các mẫu không nhìn thấy được
- Hai cách tiếp cận để tránh trang bị quá mức
  - Cắt tỉa trước : Dừng việc xây dựng cây sớm không phân chia một nút nếu điều này sẽ dẫn đến việc đo mức độ tốt giảm xuống dưới ngưỡng
    - Khó chọn ngưỡng thích hợp
  - cắt tỉa sau : Xóa cành từ một cây “đã trưởng thành”—lấy một chuỗi các cây được cắt tỉa dần dần
    - Sử dụng một tập dữ liệu khác với dữ liệu huấn luyện để quyết định “cây được cắt tỉa tốt nhất”

# Những cải tiến đối với việc tạo ra cây quyết định cơ bản

---

- Cho phép thuộc tính có giá trị liên tục
  - Tự động xác định các thuộc tính có giá trị rời rạc mới phân chia giá trị thuộc tính liên tục thành một tập hợp các khoảng rời rạc
- Xử lý giá trị thuộc tính bị thiếu
  - Gán giá trị chung nhất của thuộc tính Gán
  - xác suất cho từng giá trị có thể
- Xây dựng thuộc tính
  - Tạo các thuộc tính mới dựa trên các thuộc tính hiện có được biểu diễn thừa
  - thớt. Điều này làm giảm sự phân mảnh, lặp lại và sao chép

## Trình phân loại dựa trên quy tắc

---

- Phân loại bản ghi bằng cách sử dụng tập hợp các quy tắc “nếu...thì...”
- Luật lệ: (Điều kiện) - y
  - Ở đâu
    - Tình trạng là sự kết hợp của các bài kiểm tra về thuộc
    - tính y là nhãn lớp
  - Ví dụ về quy tắc phân loại:
    - IF (tuổi = tuổi trẻ) VÀ (sinh viên = có) THEN (mua\_máy tính = có)
    - IF (Nhóm máu=Ấm) - (Đẻ trứng=Có) - Chim
    - (Thu nhập chịu thuế < 50K) - (Hoàn tiền=Có) - Trốn tránh=Không

# Trình phân loại dựa trên quy tắc (Ví dụ)

| Tên               | Nhóm máu | Sinh con | Có thể bay | Sống trong môi trường nước | Lớp học           |
|-------------------|----------|----------|------------|----------------------------|-------------------|
| nhân loại         | ấm       | Đúng     | KHÔNG      | KHÔNG                      | động vật có vú    |
| py thon           | lạnh lẽo | KHÔNG    | KHÔNG      | KHÔNG                      | đại diện gạch s   |
| sa lm on          | lạnh lẽo | KHÔNG    | KHÔNG      | Đúng                       | này anh ấy        |
| cá voi            | ấm       | Đúng     | KHÔNG      | Đúng                       | động vật có vú    |
| con ếch           | lạnh lẽo | KHÔNG    | KHÔNG      | Thỉnh thoảng               | động vật lưỡng cư |
| kom odo           | lạnh lẽo | KHÔNG    | KHÔNG      | KHÔNG                      | đại diện gạch s   |
| con dơi           | ấm       | Đúng     | Đúng       | KHÔNG                      | động vật có vú    |
| p igeon           | ấm       | KHÔNG    | Đúng       | KHÔNG                      | b ird s           |
| con mèo           | ấm       | Đúng     | KHÔNG      | KHÔNG                      | động vật có vú    |
| leopa rd sha rk   | lạnh lẽo | Đúng     | KHÔNG      | Đúng                       | này anh ấy        |
| con rùa           | lạnh lẽo | KHÔNG    | KHÔNG      | Thỉnh thoảng               | đại diện gạch s   |
| chim cánh cụt     | ấm       | KHÔNG    | KHÔNG      | Thỉnh thoảng               | b ird s           |
| bạn có biết không | ấm       | Đúng     | KHÔNG      | KHÔNG                      | động vật có vú    |
| ôi tôi            | lạnh lẽo | KHÔNG    | KHÔNG      | Đúng                       | này anh ấy        |
| kỳ nhông          | lạnh lẽo | KHÔNG    | KHÔNG      | Thỉnh thoảng               | động vật lưỡng cư |
| g ila m on s te   | lạnh lẽo | KHÔNG    | KHÔNG      | KHÔNG                      | đại diện gạch s   |
| rp la typu s      | ấm       | KHÔNG    | KHÔNG      | KHÔNG                      | động vật có vú    |
| con cú            | ấm       | KHÔNG    | Đúng       | KHÔNG                      | b ird s           |
| lâm lph trong     | ấm       | Đúng     | KHÔNG      | Đúng                       | động vật có vú    |
| chim ưng          | ấm       | KHÔNG    | Đúng       | KHÔNG                      | b ird s           |

R1: (Sinh con = không) - (Có thể bay = có) - Chim  
 R2: (Sinh con = không) - (Sống dưới nước = có) - Cá  
 R3: (Sinh con = có) - (Nhóm máu = ấm) - Động vật có vú  
 R4: (Để con = không) - (Có thể bay = không) - Bò sát  
 R5: (Sống dưới nước = đôi khi) - Động vật lưỡng cư

# Ứng dụng phân loại dựa trên quy tắc

- Một quy tắc **bao gồm** một ví dụ nếu các thuộc tính của thể hiện thỏa mãn điều kiện của quy tắc

R1: (Sinh con = không) - (Có thể bay = có) - Chim  
R2: (Sinh con = không) - (Sống dưới nước = có) - Cá  
R3: (Sinh con = có) - (Nhóm máu = ấm) - Động vật có vú  
R4: (Đẻ con = không) - (Có thể bay = không) - Bò sát  
R5: (Sống dưới nước = đôi khi) - Động vật lưỡng cư

| Tên      | Nhóm máu | sinh nhật | Có thể bay | Sống trong môi trường nước | Lớp học |
|----------|----------|-----------|------------|----------------------------|---------|
| chim ưng | ấm       | KHÔNG     | Đúng       | KHÔNG                      | ?       |
| con gấu  | ấm       | Đúng      | KHÔNG      | KHÔNG                      | ?       |

Quy tắc R1 bao trùm chim ưng => Chim

Quy tắc R3 áp dụng cho gấu xám => Động vật có vú

# Phạm vi quy tắc và độ chính xác

## Đánh giá một quy tắc: phủ sóng và sự chính xác

- Phạm vi của một quy tắc:
  - Phân số các bản ghi thỏa mãn tiền đề của một quy tắc
    - $N_{\text{bao gồm}} = \#$  bộ dữ liệu được bao phủ bởi R
    - $N_{\text{chính xác}} = \text{Số bộ dữ liệu được phân loại chính xác bởi R}$

phạm vi bảo hiểm (R) =  $n_{\text{bao gồm}} / |D|$  /\* D: tập dữ liệu huấn luyện \*/

### - Độ chính xác của quy tắc:

- Phân số các bản ghi thỏa mãn tiền đề cũng thỏa mãn hệ quả của một quy tắc

độ chính xác(R) =  $n_{\text{chính xác}} / N_{\text{bao gồm}}$

| Tid  | Đền bù | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | Lớp học |
|------|--------|------------------------|-----------------------|---------|
| 1    | Đúng   | Đơn                    | 125K                  | KHÔNG   |
| 2    | KHÔNG  | Đã cưới                | 100K                  | KHÔNG   |
| 3    | KHÔNG  | Đơn                    | 70K                   | KHÔNG   |
| 4    | Đúng   | Đã cưới                | 120K                  | KHÔNG   |
| 5    | KHÔNG  | Đã ly hôn              | 95K                   | Đúng    |
| 6    | KHÔNG  | Đã cưới                | 60K                   | KHÔNG   |
| 7    | Đúng   | Đã ly hôn              | 220K                  | KHÔNG   |
| số 8 | KHÔNG  | Đơn                    | 85K                   | Đúng    |
| 9    | KHÔNG  | Đã cưới                | 75K                   | KHÔNG   |
| 10   | KHÔNG  | Đơn                    | 90K                   | Đúng    |

(Trạng thái=Đơn)-KHÔNG

Độ che phủ = 40%, Độ chính xác = 50%



## Trình phân loại dựa trên quy tắc hoạt động như thế nào?

R1: (Sinh con = không) - (Có thể bay = có) - Chim  
R2: (Sinh con = không) - (Sống dưới nước = có) - Cá  
R3: (Sinh con = có) - (Nhóm máu = ấm) - Động vật có vú  
R4: (Đẻ con = không) - (Có thể bay = không) - Bò sát

R5: (Sống dưới nước = đôi khi) - Động vật lưỡng cư

| Tên        | Nhóm máu | sinh nhật | Có thể bay | Sống trong môi trường nước | Lớp học |
|------------|----------|-----------|------------|----------------------------|---------|
| vong linh  | ấm       | Đúng      | KHÔNG      | KHÔNG                      | ?       |
| con rùa    | lạnh lẽo | KHÔNG     | KHÔNG      | Thỉnh thoảng               | ?       |
| cá mập chó | lạnh lẽo | Đúng      | KHÔNG      | Đúng                       | ?       |

Vượn cáo kích hoạt quy tắc R3 nên được xếp vào loại động vật có vú. Rùa kích hoạt cả R4 và R5

Một con cá mập dogfish không gây ra quy tắc nào

## Trình phân loại dựa trên quy tắc hoạt động như thế nào?

---

- Quy tắc loại trừ lẫn nhau
  - Trình phân loại chứa các quy tắc loại trừ lẫn nhau nếu các quy tắc đó độc lập với nhau
  - Mỗi bản ghi được bao phủ bởi nhiều nhất một quy tắc
- Quy tắc đầy đủ
  - Trình phân loại có phạm vi bao phủ toàn diện nếu nó tính đến mọi kết hợp có thể có của các giá trị thuộc tính
  - Mỗi bản ghi được bao phủ bởi ít nhất một quy tắc

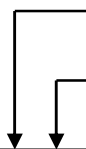
## Trình phân loại dựa trên quy tắc hoạt động như thế nào?

---

- Nếu có nhiều hơn một quy tắc được kích hoạt, cần giải quyết xung đột
  - Thứ tự kích thước: gán mức ưu tiên cao nhất cho các quy tắc kích hoạt có yêu cầu “khó khăn nhất” (nghĩa là với hầu hết các bài kiểm tra thuộc tính)
  - Sắp xếp theo lớp: thứ tự giảm dần của chi phí phổ biến hoặc phân loại sai cho mỗi lớp
  - Thứ tự dựa trên quy tắc (danh sách quyết định): các quy tắc được sắp xếp thành một danh sách ưu tiên dài, theo một số thước đo về chất lượng quy tắc hoặc bởi các chuyên gia
- Bản ghi có thể không kích hoạt bất kỳ quy tắc nào
  - Sử dụng một lớp mặc định

# Trình phân loại dựa trên quy tắc hoạt động như thế nào?

R1: (Sinh con = không) - (Có thể bay = có) - Chim  
R2: (Sinh con = không) - (Sống dưới nước = có) - Cá  
R3: (Sinh con = có) - (Nhóm máu = ấm) - Động vật có vú  
R4: (Đẻ con = không) - (Có thể bay = không) - Bò sát  
R5: (Sống dưới nước = đôi khi) - Động vật lưỡng cư



| Tên     | Nhóm máu | sinh nhật | Có thể bay | Sống trong môi trường nước | Lớp học |
|---------|----------|-----------|------------|----------------------------|---------|
| con rùa | lạnh lẽo | KHÔNG     | KHÔNG      | Thỉnh thoảng               | ?       |

# Quy tắc phân loại tòa nhà

---

## -Phương pháp trực tiếp:

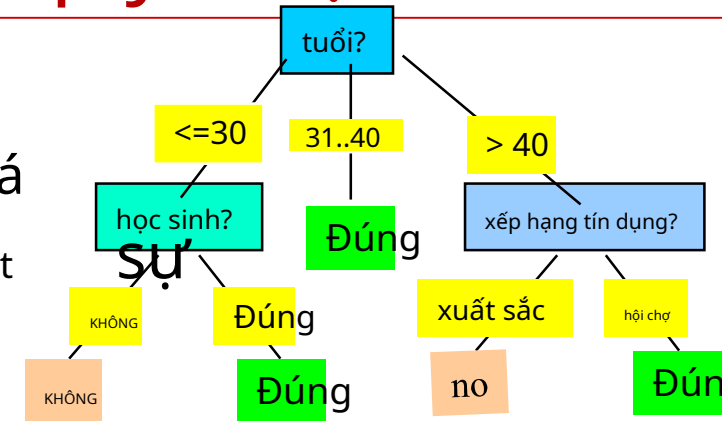
- Trích xuất quy tắc trực tiếp từ dữ liệu Ví dụ:
- ILA, RIPPER, CN2, Holte's 1R

## -Phương pháp gián tiếp:

- Trích xuất các quy tắc từ các mô hình phân loại khác (ví dụ: cây quyết định, mạng lưới thần kinh, v.v.).
- Ví dụ: Quy tắc C4.5

# Trích xuất quy tắc phân loại từ cây quyết định

- Biểu diễn tri thức dưới dạng **NẾU-THÌ** quy tắc
- Một quy tắc được tạo cho mỗi đường dẫn từ gốc đến lá
- Mỗi cặp thuộc tính-giá trị dọc theo một đường dẫn tạo thành một liên kết
- Nút lá chứa dự đoán lớp Các quy tắc dễ hiểu hơn đối với con người
- Ví dụ: Trích xuất quy tắc từ mua\_máy tính cây quyết định



NẾU NHƯ  $tuổi = "<=30"$  VÀ  $sinh\ viên = "không"$  SAU ĐÓ  $buy\_computer = "không"$

NẾU NHƯ  $tuổi = "<=30"$  VÀ  $sinh\ viên = "có"$  SAU ĐÓ  $buy\_computer = "có"$  NẾU NHƯ  $tuổi = "31...40"$  SAU ĐÓ  $buy\_computer = "có"$

NẾU NHƯ  $tuổi = ">40"$  VÀ  $credit\_rated = "xuất\ sắc"$  SAU ĐÓ  $mua\_máy\ tính = "Đúng"$

NẾU NHƯ  $tuổi = ">40"$  VÀ  $credit\_rated = "công\ bằng"$  SAU ĐÓ  $buy\_computer = "không"$

# Bài tập

| Quang cảnh | Nhiệt độ | Độ ẩm | Sức gió | Chơi tennis |
|------------|----------|-------|---------|-------------|
| Nắng       | Nóng     | Cao   | Yếu     | Không       |
| Nắng       | Nóng     | Cao   | Mạnh    | Không       |
| Mây        | Nóng     | Cao   | Yếu     | Có          |
| Mưa        | TB       | Cao   | Yếu     | Có          |
| Mưa        | Lạnh     | BT    | Yếu     | Có          |
| Mưa        | Lạnh     | BT    | Mạnh    | Không       |
| Mây        | Lạnh     | BT    | Mạnh    | Có          |
| Nắng       | TB       | Cao   | Yếu     | Không       |
| Nắng       | Lạnh     | BT    | Yếu     | Có          |
| Mưa        | TB       | BT    | Yếu     | Có          |
| Nắng       | TB       | BT    | Mạnh    | Có          |
| Mây        | TB       | Cao   | Mạnh    | Có          |
| Mây        | Nóng     | BT    | Yếu     | Có          |
| Mưa        | TB       | Cao   | Mạnh    | Không       |

# Bài tập

---

- Xây dựng cây quyết định với thước đo Information Gain
- Trích xuất các luật từ cây quyết định
- Xác định nhãn lớp cho mẫu mới sau:

| Quang cảnh | Nhiệt độ | Độ ẩm | Sức gió | Chơi quần vợt |
|------------|----------|-------|---------|---------------|
| mưa rào    | bệnh lao | BT    | Mạnh    | ?             |
| Nắng       | bệnh lao | Cao   | Mạnh    | ?             |



# Trích xuất ule từ dữ liệu đào tạo

---

- Thuật toán bao phủ tuần tự: Trích xuất các quy tắc trực tiếp từ dữ liệu huấn luyện
- Các thuật toán bao phủ tuần tự điển hình: FOIL, ILA, AQ, CN2, RIPPER
- Các quy tắc được học tuần tự, mỗi cái cho một lớp  $C$  nhất định. Tôi sẽ bao gồm nhiều bộ dữ liệu của  $C$  tôi nhưng không có (hoặc một vài) bộ dữ liệu của các lớp khác
- Các bước:
  - Các quy tắc được học lần lượt
  - Mỗi lần học một quy tắc, các bộ chứa quy tắc đó sẽ bị xóa
  - Quá trình lặp lại trên các bộ còn lại trừ khi điều kiện chấm dứt, ví dụ: khi không có thêm ví dụ đào tạo nào hoặc khi chất lượng của quy tắc được trả về dưới ngưỡng do người dùng chỉ định
- Comp. w. quy nạp cây quyết định: học một bộ quy tắc đồng thời

# Phương pháp trực tiếp: Che phủ tuần tự

---

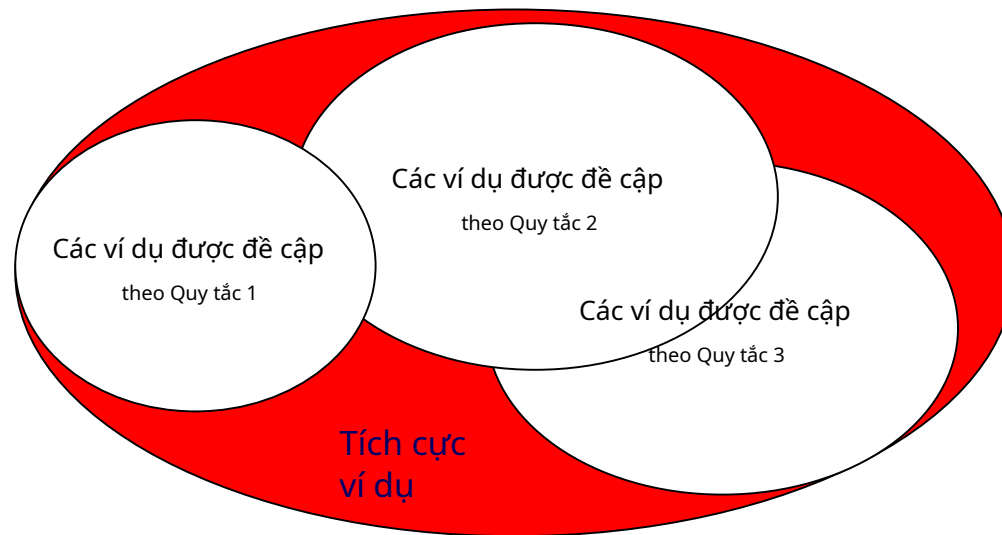
1. Bắt đầu từ một quy tắc trống
2. Phát triển quy tắc bằng cách sử dụng chức năng Tìm hiểu một quy tắc
3. Xóa hồ sơ đào tạo nằm trong phạm vi quy định
4. Lặp lại Bước (2) và (3) cho đến khi đáp ứng tiêu chí dừng

# Thuật toán che phủ tuần tự

trong khi (còn đủ bộ mục tiêu)

tạo ra một quy tắc

loại bỏ các bộ mục tiêu tích cực thỏa mãn quy tắc này



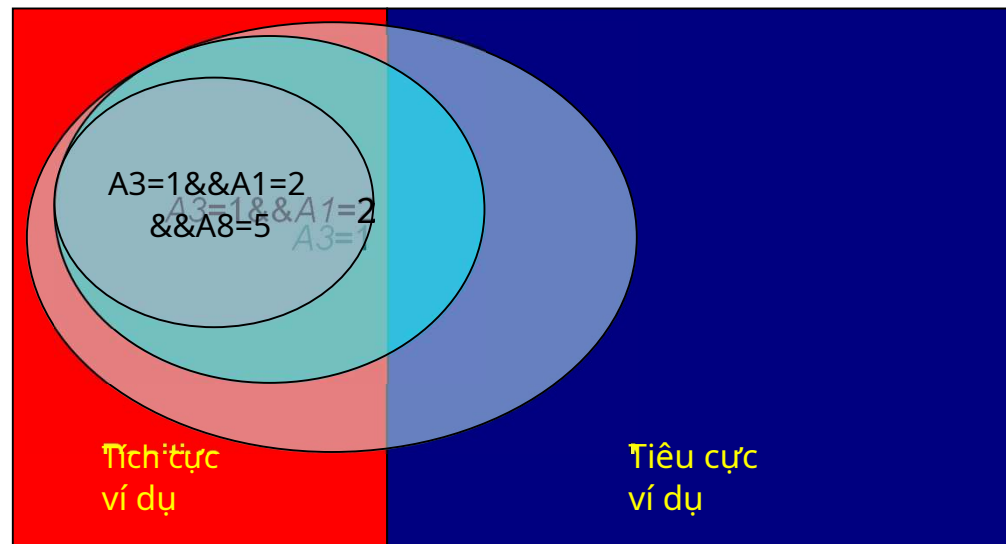
# Tạo quy tắc

- Để tạo ra một quy tắc trong

khi(ĐÚNG VẬY)

tìm vị ngữ tốt nhấtP

nếu nhữnăg lá(p) >ngưỡnăgsau đốthêm vàoPtheo quy định hiện tại  
khácp há vớ



# Ưu điểm của bộ phân loại dựa trên quy tắc

---

- Có đặc điểm khá giống với cây quyết định
  - Có tính biểu cảm cao như cây quyết định Dễ diễn
  - giải (nếu quy tắc được sắp xếp theo lớp) Hiệu suất
  - tương đương với cây quyết định
    - Có thể xử lý các thuộc tính dư thừa và không liên quan. Tương tác
    - giữa các biến có thể gây ra sự cố (ví dụ: sự cố X-OR)
- Phù hợp hơn để xử lý các lớp mất cân bằng
- Khó xử lý các giá trị bị thiếu trong tập kiểm tra hơn

# Thuật toán học quy nạp- ILA

---

- M.Tolun, 1998, ILA – thuật toán học quy nạp được sử dụng để tạo ra một bộ quy tắc phân loại
- Quy tắc dạng “IF-THEN”
- chia bảng 'T' chứa m ví dụ thành n bảng con.
-

## Thuật toán ILA

---

- Bước 1: Chia bảng "T" chứa m ví dụ thành n bảng con ( $t_1, t_2, \dots, t_n$ ). Một bảng cho mỗi giá trị có thể có của thuộc tính lớp. (lặp lại các bước 2-8 cho mỗi bảng phụ)
- Bước 2: Khởi tạo tổ hợp thuộc tính count 'j' = 1
- Bước 3: Đối với bảng con đang thực hiện công việc, chia danh sách thuộc tính thành các tổ hợp riêng biệt, mỗi tổ hợp có thuộc tính riêng biệt 'j'
- Đối với mỗi tổ hợp thuộc tính, đếm số lần xuất hiện của các giá trị thuộc tính xuất hiện dưới cùng tổ hợp thuộc tính trong các hàng không được đánh dấu của bảng con đang xem xét, đồng thời không xuất hiện dưới cùng tổ hợp thuộc tính của bảng con khác. -những cái bàn. Gọi kết hợp đầu tiên có số lần xuất hiện tối đa là kết hợp tối đa 'MAX'.

# Thuật toán ILA

---

- Bước 5: Nếu 'MAX' == null, tăng 'j' lên 1 và chuyển sang Bước 3.
- Bước 6: Đánh dấu tất cả các hàng của bảng con nơi làm việc, trong đó xuất hiện giá trị 'MAX' là đã phân loại
- Thêm một quy tắc (thuộc tính IF = "XYZ" -> THEN quyết định là CÓ/ KHÔNG) cho R có phía bên trái sẽ có tên thuộc tính 'MAX' với các giá trị của chúng được phân tách bằng AND và phía bên phải của nó chứa giá trị thuộc tính quyết định được liên kết với bảng con
- Bước 8: Nếu tất cả các hàng được đánh dấu là đã phân loại thì chuyển sang xử lý bảng con khác và chuyển sang Bước 2. Ngược lại, chuyển sang Bước 4. Nếu không có bảng phụ nào, hãy thoát ra với bộ quy tắc thu được cho đến lúc đó



# Ví dụ

| No | Size | Color       | Shape | Decision |
|----|------|-------------|-------|----------|
| 1  | Vừa  | Xanh dương  | Hộp   | Yes      |
| 2  | Nhỏ  | đỏ          | Nón   | No       |
| 3  | Nhỏ  | đỏ          | Cầu   | Yes      |
| 4  | Lớn  | đỏ          | Nón   | No       |
| 5  | Lớn  | Xanh lá cây | Trụ   | Yes      |
| 6  | Lớn  | đỏ          | Trụ   | No       |
| 7  | Lớn  | Xanh lá cây | Cầu   | Yes      |

|       |         |             |           |            |
|-------|---------|-------------|-----------|------------|
| KHÔNG | Kích cỡ | Màu sắc     | Hình dạng | Phán quyết |
| 1     | Medium  | Xanh lá cây | Cầu       | ?          |
| 2     | Nhỏ     | Màu đỏ      | nón       | ?          |

# Bài tập 1

Cho một dữ liệu huấn luyện về người mua máy tính, áp dụng thuật toán ID3 và ILA để xây dựng mô hình phân loại

| 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            |

Xác định nhãn lớp cho các mẫu:

1.  $X_{new}$  = (tuổi  $\leq 30$ , thu nhập = cao, sinh viên = có, xếp hạng tín chỉ = khá)
2.  $X_{new}$  = (tuổi  $> 40$ , thu nhập = cao, sinh viên = không, xếp hạng tín chỉ = khá)

# Bài tập 2

Cho một dữ liệu huấn luyện về chơi tennis hay không, áp dụng thuật toán ID3 và ILA để xây dựng mô hình phân loại:

| Day | Outlook  | Temperature | Humidity | Windy  | PlayTennis |
|-----|----------|-------------|----------|--------|------------|
| D1  | Sunny    | Hot         | High     | Weak   | No         |
| D2  | Sunny    | Hot         | High     | Strong | No         |
| D3  | Overcast | Hot         | High     | Weak   | Yes        |
| D4  | Rain     | Mild        | High     | Weak   | Yes        |
| D5  | Rain     | Cool        | Normal   | Weak   | Yes        |
| D6  | Rain     | Cool        | Normal   | Strong | No         |
| D7  | Overcast | Cool        | Normal   | Strong | Yes        |
| D8  | Sunny    | Mild        | High     | Weak   | No         |
| D9  | Sunny    | Cool        | Normal   | Weak   | Yes        |
| D10 | Rain     | Mild        | Normal   | Weak   | Yes        |
| D11 | Sunny    | Mild        | Normal   | Strong | Yes        |
| D12 | Overcast | Mild        | High     | Strong | Yes        |
| D13 | Overcast | Hot         | Normal   | Weak   | Yes        |
| D14 | Rain     | Mild        | High     | Strong | No         |

Xác định nhãn lớp cho các mẫu:

- Xnew = <Triển vọng=nắng, Nhiệt độ = mát mẻ, Độ ẩm = cao, Gió = mạnh>
- Xnew = <Triển vọng = u ám, Nhiệt độ = mát mẻ, Độ ẩm = cao, Gió = mạnh>

# Phân loại Bayes

- Bộ phân loại thống kê : thực hiện dự đoán xác suất, tức là dự đoán xác suất thành viên của lớp
- Sự thành lập: Dựa trên Định lý Bayes.
- Hiệu suất: Một bộ phân loại Bayesian đơn giản, bộ phân loại Bayes ngây thơ, có hiệu suất tương đương với cây quyết định và các bộ phân loại mạng thần kinh được chọn
- Học xác suất : Tính toán xác suất rõ ràng cho giả thuyết, một trong những cách tiếp cận thực tế nhất đối với một số loại vấn đề học tập nhất định
- Tăng dần : Mỗi ví dụ huấn luyện có thể tăng/giảm dần xác suất một giả thuyết là đúng. Kiến thức trước đây có thể được kết hợp với dữ liệu quan sát được.
- Dự đoán xác suất : Dự đoán nhiều giả thuyết, có trọng số bằng xác suất của chúng
- Tiêu chuẩn : Ngay cả khi các phương pháp Bayes khó tính toán, chúng vẫn có thể cung cấp tiêu chuẩn cho việc ra quyết định tối ưu mà các phương pháp khác có thể đo lường được

# Định lý Bayes: Cơ bản

---

- Cho phép  $X$  là một mẫu dữ liệu ("chứng cứ"): nhãn lớp chưa xác định
- Giả sử  $H$  là giả thuyết  $X$  thuộc lớp  $C$
- Phân loại nhằm xác định  $P(H | X)$ , (xác suất hậu nghiệm), xác suất mà giả thuyết giữ được với mẫu dữ liệu được quan sát  $X$
- $P(H)$  (Xác suất trước), xác suất ban đầu
  - Ví dụ,  $X$  sẽ mua máy tính, không phân biệt tuổi tác, thu nhập, ...
- $P(X)$ : xác suất dữ liệu mẫu được quan sát
- $P(X | H)$  (khả năng), xác suất quan sát được mẫu  $X$ , cho rằng giả thuyết giữ
  - Ví dụ: Cho rằng  $X$  sẽ mua máy tính, vấn đề là vậy.  $X$  là 31..40, thu nhập trung bình

# Định lý Bayes

- Cho dữ liệu huấn luyện  $X$ , xác suất hậu nghiệm của một giả thuyết  $H$ ,  $P(H | X)$ , tuân theo định lý Bayes

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

- Một cách không chính thức, điều này có thể được viết là  
hậu nghiệm = khả năng x trước / bằng chứng
- Dự đoán  $X$  thuộc về  $C_2$  nếu xác suất  $P(C_{\text{Tôi}} | X)$  là cao nhất trong số tất cả  $P(C_k | X)$  cho tất cả các lớp học
- Khó khăn thực tế: đòi hỏi kiến thức ban đầu về nhiều xác suất, chi phí tính toán đáng kể

# Hướng tới bộ phân loại Naïve Bayes

- Cho D là tập huấn luyện gồm các bộ dữ liệu và các nhãn lớp liên quan của chúng và mỗi bộ dữ liệu được biểu thị bằng một vectơ thuộc tính  $nDX = (x_1, x_2, \dots, x_N)$
- Giả sử có  $i$  lớp  $C_1, C_2, \dots, C_i$ .
- Việc phân loại nhằm rút ra giá trị hậu nghiệm tối đa, tức là  $P(C) \text{ tối đa} | X$  Điều
- này có thể được suy ra từ định lý Bayes

$$P(C_i | X) = \frac{P(X | C_i) P(C_i)}{P(X)}$$

- Vì  $P(X)$  là hằng số cho tất cả các lớp nên chỉ  $P(C_i | X) = P(X | C_i) P(C_i)$  cần phải được tối đa hóa

# Nguồn gốc của bộ phân loại Naïve Bayes

- Một giả định đơn giản hóa: các thuộc tính độc lập có điều kiện (nghĩa là không có mối quan hệ phụ thuộc giữa các thuộc tính):  

$$P(X | C_i) = P(x_1 | C_i) \cdot P(x_2 | C_i) \cdot \dots \cdot P(x_N | C_i)$$
- Điều này làm giảm đáng kể chi phí tính toán: Chỉ tính phân phối lớp
- Nếu một  $x_k$  là phân loại,  $P(x_k | C_{\text{Tôi}})$  là số bộ trong  $C_{\text{Tôi}}$  có giá trị  $x_k$  cho một  $x_k$  chia cho  $|C_{\text{nhận dạng}}|$  (số bộ của  $C_{\text{Tôi}}$  ở D)
- Nếu một  $x_k$  có giá trị liên tục,  $P(x_k | C_{\text{Tôi}})$  thường được tính toán dựa trên phân bố Gaussian với giá trị trung bình  $\mu$  và độ lệch chuẩn  $\sigma$

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

và  $P(x_k | C_{\text{Tôi}}) = \frac{1}{|C_{\text{Tôi}}|} \sum_{C_{\text{Tôi}}} g(x_k, \mu, \sigma)$



# Bộ phân loại Bayes

-Coi mỗi thuộc tính và nhãn lớp là các biến ngẫu nhiên

| T    | R quý | hôn nhân   | Chịu thuế |            |
|------|-------|------------|-----------|------------|
|      |       | Trạng thái | Thu nhập  | trốn tránh |
| 1    | Đúng  | Đơn        | 125K      | KHÔNG      |
| 2    | KHÔNG | Đã cưới    | 100K      | KHÔNG      |
| 3    | KHÔNG | Đơn        | 70K       | KHÔNG      |
| 4    | Đúng  | Đã cưới    | 120K      | KHÔNG      |
| 5    | KHÔNG | Đã ly hôn  | 95K       | Đúng       |
| 6    | KHÔNG | Đã cưới    | 60K       | KHÔNG      |
| 7    | Đúng  | Đã ly hôn  | 220K      | KHÔNG      |
| số 8 | KHÔNG | Đơn        | 85K       | Đúng       |
| 9    | KHÔNG | Đã cưới    | 75K       | KHÔNG      |
| 10   | KHÔNG | Đơn        | 90K       | Đúng       |

Trốn tránh C

Không gian sự kiện: {Có, Không}  $P(C) = (0,3, 0,7)$

Hoàn tiền A<sub>1</sub>

Không gian sự kiện: {Có, Không}  $P(A_1) = (0,3,0,7)$

Trạng Thái Vỡ A<sub>2</sub>

Không gian sự kiện: {Độc thân, Đã kết hôn, Đã ly hôn}  $P(A_2) = (0,4,0,4,0,2)$

Thu nhập chịu thuế A<sub>3</sub>

Không gian sự kiện: R

$P(A_3) \sim \text{Bình thường}(-,-)$

# Bộ phân loại Bayes

- Cách phân loại bản ghi mới  $X = ('Có', 'Đơn', 80K)$

| T<br>nhận dạng | R<br>quỹ | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | trốn tránh |
|----------------|----------|------------------------|-----------------------|------------|
| 1              | Đúng     | Đơn                    | 125K                  | KHÔNG      |
| 2              | KHÔNG    | Đã cưới                | 100K                  | KHÔNG      |
| 3              | KHÔNG    | Đơn                    | 70K                   | KHÔNG      |
| 4              | Đúng     | Đã cưới                | 120K                  | KHÔNG      |
| 5              | KHÔNG    | Đã ly hôn              | 95K                   | Đúng       |
| 6              | KHÔNG    | Đã cưới                | 60K                   | KHÔNG      |
| 7              | Đúng     | Đã ly hôn              | 220K                  | KHÔNG      |
| số 8           | KHÔNG    | Đơn                    | 85K                   | Đúng       |
| 9              | KHÔNG    | Đã cưới                | 75K                   | KHÔNG      |
| 10             | KHÔNG    | Đơn                    | 90K                   | Đúng       |

Tìm lớp có xác suất cao nhất cho các giá trị vectơ.

Xác suất hậu thế tối đa ước lượng:

- Tìm giá trị  $c$  cho lớp học  $C$  tối đa hóa  $P(C=c | X)$

Làm thế nào để chúng tôi ước tính  $P(C | X)$  cho các giá trị khác nhau của  $C$ ?

- Chúng tôi muốn ước tính  $P(C=Có | X)$
- Và  $P(C=Không | X)$

# Bộ phân loại Bayes

- Để xác định rõ xác suất:
  - Hãy xem xét từng thuộc tính và nhãn lớp như biến ngẫu nhiên
  - Xác suất được xác định từ dữ liệu

| T    | nhân dạng | R quĩ | nghệ thuật  | Chịu thuế |            |
|------|-----------|-------|-------------|-----------|------------|
|      |           |       | Trạng thái  | Thu nhập  | trốn tránh |
| 1    |           | Đúng  | Đơn         | 125K      | KHÔNG      |
| 2    | KHÔNG     |       | Đã cưới     | 100K      | KHÔNG      |
| 3    | KHÔNG     |       | Đơn         | 70K       | KHÔNG      |
| 4    |           | Đúng  | Đã cưới     | 120K      | KHÔNG      |
| 5    | KHÔNG     |       | D đã ly hôn | 95K       | Đúng       |
| 6    | KHÔNG     |       | Đã cưới     | 60K       | KHÔNG      |
| 7    |           | Đúng  | D đã ly hôn | 220K      | KHÔNG      |
| số 8 | KHÔNG     |       | Đơn         | 85K       | Đúng       |
| 9    | KHÔNG     |       | Đã cưới     | 75K       | KHÔNG      |
| 10   | KHÔNG     |       | Đơn         | 90K       | Đúng       |

Trốn tránh C  
Không gian sự kiện: {Có, Không}  $P(C) = (0,3, 0,7)$

Hoàn tiền A<sub>1</sub>  
Không gian sự kiện: {Có, Không}  $P(A_1) = (0,3,0,7)$

Trạng Thái Vỡ A<sub>2</sub>  
Không gian sự kiện: {Độc thân, Đã kết hôn, Đã ly hôn}  $P(A_2) = (0,4,0,4,0,2)$

Thu nhập chịu thuế A<sub>3</sub>  
Không gian sự kiện: R  
 $P(A_3) \sim \text{Bình thường}(-,-,2)$   
 $\mu = 104$ : trung bình mẫu,  $\sigma^2=1874$ : mẫu var

# Ví dụ

## -Ghi

$X = (\text{Hoàn tiền} = \text{Có}, \text{Trạng thái} = \text{Độc thân}, \text{Thu nhập} = 80K)$

-Đối với lớp học  $C = \text{'Né tránh'}$ , chúng tôi muốn tính toán:

$P(C = \text{Có} | X)$  và  $P(C = \text{Không} | X)$

- Chúng tôi tính toán:

- $P(C = \text{Có} | X) = P(C = \text{Có}) * P(\text{Hoàn tiền} = \text{Có} | C = \text{Có})$ 
  - \*  $P(\text{Tình trạng} = \text{Độc thân} | C = \text{Có})$
  - \*  $P(\text{Thu nhập} = 80K | C = \text{Có})$
- $P(C = \text{Không} | X) = P(C = \text{Không}) * P(\text{Hoàn tiền} = \text{Có} | C = \text{Không})$ 
  - \*  $P(\text{Tình trạng} = \text{Độc thân} | C = \text{Không})$
  - \*  $P(\text{Thu nhập} = 80K | C = \text{Không})$

# Làm thế nào để ước tính xác suất từ dữ liệu?

phân loại  
phân loại  
tiếp diễn  
lớp học

-Xác suất ưu tiên của lớp:

$$P(C = c) = \frac{N_c}{N}$$

ví dụ,  $P(C = \text{Không}) = 7/10$ ,

$P(C = \text{Có}) = 3/10$

-Đối với thuộc tính rời rạc:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

Ở đây  $N_{a,c}$  là số trường hợp có thuộc tính  $A_i = a$  và thuộc về lớp  $c$

-Ví dụ:

$P(\text{Tình trạng}=\text{Đã kết hôn} | \text{Không}) = 4/7$   
 $P(\text{Hoàn tiền}=\text{Có} | \text{Có})=0$

| T         | R     | hôn nhân   | Chịu thuế |            |
|-----------|-------|------------|-----------|------------|
| nhân dạng | quỹ   | Trạng thái | Thu nhập  | trốn tránh |
| 1         | Đúng  | Đơn        | 125K      | KHÔNG      |
| 2         | KHÔNG | Đã cưới    | 100K      | KHÔNG      |
| 3         | KHÔNG | Đơn        | 70K       | KHÔNG      |
| 4         | Đúng  | Đã cưới    | 120K      | KHÔNG      |
| 5         | KHÔNG | Đã ly hôn  | 95K       | Đúng       |
| 6         | KHÔNG | Đã cưới    | 60K       | KHÔNG      |
| 7         | Đúng  | Đã ly hôn  | 220K      | KHÔNG      |
| số 8      | KHÔNG | Đơn        | 85K       | Đúng       |
| 9         | KHÔNG | Đã cưới    | 75K       | KHÔNG      |
| 10        | KHÔNG | Đơn        | 90K       | Đúng       |

# Làm thế nào để ước tính xác suất từ dữ liệu?

- **Vì tiếp diễn** thuộc tính:
  - rời rạc hóa phạm vi vào thùng
    - một thuộc tính thứ tự trên mỗi thùng vi
    - phạm giả định về tính độc lập
  - **Chia hai chiều:**  $(A < v)$  hoặc  $(A > v)$ 
    - chỉ chọn một trong hai phần tách làm thuộc tính
- mới **Ước tính mật độ xác suất:**
  - Giả sử thuộc tính theo sau một **phân phối bình thường**
  - Sử dụng dữ liệu để ước tính các tham số phân phối (nghĩa là **nghĩa là -Và độ lệch chuẩn -**)
  - Khi đã biết phân bố xác suất, chúng ta có thể sử dụng nó để ước tính xác suất có điều kiện  $P(A_{tôi} | c)$

# Làm thế nào để ước tính xác suất từ dữ liệu?

| T         | R     | hôn nhân   | Chiếu thuế |            |
|-----------|-------|------------|------------|------------|
| nhân dạng | quỹ   | Trạng thái | Thu nhập   | trốn tránh |
| 1         | Đúng  | Đơn        | 125K       | KHÔNG      |
| 2         | KHÔNG | Đã cưới    | 100K       | KHÔNG      |
| 3         | KHÔNG | Đơn        | 70K        | KHÔNG      |
| 4         | Đúng  | Đã cưới    | 120K       | KHÔNG      |
| 5         | KHÔNG | Đã ly hôn  | 95K        | Đúng       |
| 6         | KHÔNG | Đã cưới    | 60K        | KHÔNG      |
| 7         | Đúng  | Đã ly hôn  | 220K       | KHÔNG      |
| số 8      | KHÔNG | Đơn        | 85K        | Đúng       |
| 9         | KHÔNG | Đã cưới    | 75K        | KHÔNG      |
| 10        | KHÔNG | Đơn        | 90K        | Đúng       |

Thuộc tính rời rạc:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$ : số trường hợp có thuộc tính  $A_i = a$  và thuộc về lớp

$c$

$N_c$ : số phiên bản của lớp  $c$

# Làm thế nào để ước tính xác suất từ dữ liệu?

phân loại      phân loại      tiếp diễn      lớp học

| T<br>nhận dạng | R<br>quỹ | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | trốn tránh |
|----------------|----------|------------------------|-----------------------|------------|
| 1              | Đúng     | Đơn                    | 125K                  | KHÔNG      |
| 2              | KHÔNG    | Đã cưới                | 100K                  | KHÔNG      |
| 3              | KHÔNG    | Đơn                    | 70K                   | KHÔNG      |
| 4              | Đúng     | Đã cưới                | 120K                  | KHÔNG      |
| 5              | KHÔNG    | Đã ly hôn              | 95K                   | Đúng       |
| 6              | KHÔNG    | Đã cưới                | 60K                   | KHÔNG      |
| 7              | Đúng     | Đã ly hôn              | 220K                  | KHÔNG      |
| số 8           | KHÔNG    | Đơn                    | 85K                   | Đúng       |
| 9              | KHÔNG    | Đã cưới                | 75K                   | KHÔNG      |
| 10             | KHÔNG    | Đơn                    | 90K                   | Đúng       |

Thuộc tính rời rạc:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$ : số trường hợp có thuộc tính  $A_i = a$  và thuộc về lớp  $c$

$N_c$ : số phiên bản của lớp  $c$

$$P(\text{Hoàn tiền} = \text{Có} | \text{KHÔNG}) = 3/7$$



# Làm thế nào để ước tính xác suất từ dữ liệu?

phân loại      phân loại      tiếp diễn      lớp học

| T<br>nhận dạng | R<br>quỹ | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | trốn tránh |
|----------------|----------|------------------------|-----------------------|------------|
| 1              | Đúng     | Đơn                    | 125K                  | KHÔNG      |
| 2              | KHÔNG    | Đã cưới                | 100K                  | KHÔNG      |
| 3              | KHÔNG    | Đơn                    | 70K                   | KHÔNG      |
| 4              | Đúng     | Đã cưới                | 120K                  | KHÔNG      |
| 5              | KHÔNG    | Đã ly hôn              | 95K                   | Đúng       |
| 6              | KHÔNG    | Đã cưới                | 60K                   | KHÔNG      |
| 7              | Đúng     | Đã ly hôn              | 220K                  | KHÔNG      |
| số 8           | KHÔNG    | Đơn                    | 85K                   | Đúng       |
| 9              | KHÔNG    | Đã cưới                | 75K                   | KHÔNG      |
| 10             | KHÔNG    | Đơn                    | 90K                   | Đúng       |

Thuộc tính rời rạc:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$ : số trường hợp có thuộc tính  $A_i = a$  và thuộc về lớp  $c$

$N_c$ : số phiên bản của lớp  $c$

$$P(\text{Hoàn tiền} = \text{Có} | \text{Đúng}) = 0$$

# Làm thế nào để ước tính xác suất từ dữ liệu?

phân loại      phân loại      tiếp diễn      lớp học

| T<br>nhận dạng | R quỹ | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | trốn tránh |
|----------------|-------|------------------------|-----------------------|------------|
| 1              | Đúng  | Đơn                    | 125K                  | KHÔNG      |
| 2              | KHÔNG | Đã cưới                | 100K                  | KHÔNG      |
| 3              | KHÔNG | Đơn                    | 70K                   | KHÔNG      |
| 4              | Đúng  | Đã cưới                | 120K                  | KHÔNG      |
| 5              | KHÔNG | Đã ly hôn              | 95K                   | Đúng       |
| 6              | KHÔNG | Đã cưới                | 60K                   | KHÔNG      |
| 7              | Đúng  | Đã ly hôn              | 220K                  | KHÔNG      |
| số 8           | KHÔNG | Đơn                    | 85K                   | Đúng       |
| 9              | KHÔNG | Đã cưới                | 75K                   | KHÔNG      |
| 10             | KHÔNG | Đơn                    | 90K                   | Đúng       |

Thuộc tính rời rạc:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$ : số trường hợp có thuộc tính  $A_i = a$  và thuộc về lớp  $c$

$N_c$ : số phiên bản của lớp  $c$

$$P(\text{Trạng thái}=\text{Độc thân} | \text{KHÔNG}) = 2/7$$

# Làm thế nào để ước tính xác suất từ dữ liệu?

phân loại      phân loại      tiếp diễn      lớp học

| T<br>nhận dạng | R<br>quỹ | hôn nhân<br>Trạng thái | Chịu thuế<br>Thu nhập | trốn tránh |
|----------------|----------|------------------------|-----------------------|------------|
| 1              | Đúng     | Đơn                    | 125K                  | KHÔNG      |
| 2              | KHÔNG    | Đã cưới                | 100K                  | KHÔNG      |
| 3              | KHÔNG    | Đơn                    | 70K                   | KHÔNG      |
| 4              | Đúng     | Đã cưới                | 120K                  | KHÔNG      |
| 5              | KHÔNG    | Đã ly hôn              | 95K                   | Đúng       |
| 6              | KHÔNG    | Đã cưới                | 60K                   | KHÔNG      |
| 7              | Đúng     | Đã ly hôn              | 220K                  | KHÔNG      |
| số 8           | KHÔNG    | Đơn                    | 85K                   | Đúng       |
| 9              | KHÔNG    | Đã cưới                | 75K                   | KHÔNG      |
| 10             | KHÔNG    | Đơn                    | 90K                   | Đúng       |

Thuộc tính rời rạc:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$ : số trường hợp có thuộc tính  $A_i = a$  và thuộc về lớp  $c$

$N_c$ : số phiên bản của lớp  $c$

$$P(\text{Trạng thái} = \text{Độc thân} | \text{Đúng}) = 2/3$$

# Làm thế nào để ước tính xác suất từ dữ liệu?

| T    | nhận dạng | R quý | hôn nhân   | Chịu thuế | trốn tránh |
|------|-----------|-------|------------|-----------|------------|
|      |           |       | Trạng thái | Thu nhập  |            |
| 1    |           | Đúng  | Đơn        | 125K      | KHÔNG      |
| 2    |           | KHÔNG | Đã cưới    | 100K      | KHÔNG      |
| 3    |           | KHÔNG | Đơn        | 70K       | KHÔNG      |
| 4    |           | Đúng  | Đã cưới    | 120K      | KHÔNG      |
| 5    |           | KHÔNG | Đã ly hôn  | 95K       | Đúng       |
| 6    |           | KHÔNG | Đã cưới    | 60K       | KHÔNG      |
| 7    |           | Đúng  | Đã ly hôn  | 220K      | KHÔNG      |
| số 8 |           | KHÔNG | Đơn        | 85K       | Đúng       |
| 9    |           | KHÔNG | Đã cưới    | 75K       | KHÔNG      |
| 10   |           | KHÔNG | Đơn        | 90K       | Đúng       |

- Phân phối bình thường:

$$P(A_{\text{tối}} \text{ một} | c_j) = \frac{1}{\sqrt{2\pi} \sigma_{ij}} e^{-\frac{(\text{Một} - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- Một cho mỗi  $(a, ci)$  đôi

- Vì  $Lớp = Không$

- ý nghĩa mẫu  $\mu = 110$

- phương sai mẫu  $\sigma^2 = 2975$

- Vì  $Thu \text{ nhập} = 80$

$$P(\text{Thu nhập} = 80 | \text{KHÔNG}) = \frac{1}{\sqrt{2\pi} \cdot 54.54} e^{-\frac{(80 - 110)^2}{2 \cdot 2975}} = 0,0062$$

# Làm thế nào để ước tính xác suất từ dữ liệu?

| T    | nhận dạng | R     | quỹ  | hôn nhân   | Chịu thuế | thu nhập | trốn tránh |
|------|-----------|-------|------|------------|-----------|----------|------------|
|      |           |       |      | Trạng thái |           |          |            |
| 1    |           |       | Đúng | Đơn        | 125K      |          | KHÔNG      |
| 2    |           | KHÔNG |      | Đã cưới    | 100K      |          | KHÔNG      |
| 3    |           | KHÔNG |      | Đơn        | 70K       |          | KHÔNG      |
| 4    |           |       | Đúng | Đã cưới    | 120K      |          | KHÔNG      |
| 5    |           | KHÔNG |      | Đã ly hôn  | 95K       |          | Đúng       |
| 6    |           | KHÔNG |      | Đã cưới    | 60K       |          | KHÔNG      |
| 7    |           |       | Đúng | Đã ly hôn  | 220K      |          | KHÔNG      |
| số 8 |           | KHÔNG |      | Đơn        | 85K       |          | Đúng       |
| 9    |           | KHÔNG |      | Đã cưới    | 75K       |          | KHÔNG      |
| 10   |           | KHÔNG |      | Đơn        | 90K       |          | Đúng       |

- Phân phối bình thường:

$$P(A_{\text{Tôi}} = \text{một} \mid c_j) = \frac{1}{\sqrt{2\pi} \sigma_j} e^{-\frac{(\text{Một} - \mu_j)^2}{2\sigma_j^2}}$$

- Một cho mỗi  $(a, c_i)$  đôi

- Vì  $L_{\text{Ớp}} = C_0$

- ý nghĩa mẫu  $\mu = 90$

- phương sai mẫu  $\sigma^2 = 25$

- Vì  $\text{Thu nhập} = 80$

$$P(\text{Thu nhập} = 80 \mid \text{Đúng}) = \frac{1}{\sqrt{2\pi} (5)} e^{-\frac{(80 - 90)^2}{2(25)}} = 0,01$$

# Ví dụ

## -Ghi

X = (Hoàn tiền = Có, Trạng thái = Độc thân, Thu nhập = 80K)

- Chúng tôi tính toán:

$$\begin{aligned} - P(C = \text{Có} | X) &= P(C = \text{Có}) * P(\text{Hoàn tiền} = \text{Có} | C = \text{Có}) \\ &\quad * P(\text{Tình trạng} = \text{Độc thân} | C = \text{Có}) \\ &\quad * P(\text{Thu nhập} = 80K | C = \text{Có}) \\ &= 3/10 * 0 * 2/3 * 0,01 = 0 \end{aligned}$$

$$\begin{aligned} - P(C = \text{Không} | X) &= P(C = \text{Không}) * P(\text{Hoàn tiền} = \text{Có} | C = \text{Không}) \\ &\quad * P(\text{Tình trạng} = \text{Độc thân} | C = \text{Không}) \\ &\quad * P(\text{Thu nhập} = 80K | C = \text{Không}) \\ &= 7/10 * 3/7 * 2/7 * 0,0062 = 0,0005 \end{aligned}$$

# Ví dụ về Trình phân loại Naïve Bayes

-Tạo Trình phân loại Naïve Bayes, về cơ bản có nghĩa là tính toán **đếm**:

Tổng số hồ sơ:  $N = 10$

Lớp số:

Số lượng hồ sơ: 7

Hoàn trả thuộc tính:

Có: 3

Số 4

Thuộc tính Tình trạng hôn nhân:

Đơn: 2

Đã ly hôn: 1

Đã kết hôn: 4

Thu nhập thuộc tính:

có nghĩa là: 110

phương sai: 2975

Lớp Có:

Số lượng hồ sơ: 3

Hoàn trả thuộc tính:

Có: 0

Số 3

Thuộc tính Tình trạng hôn nhân:

Đơn: 2

Đã ly hôn: 1

Đã kết hôn: 0

Thu nhập thuộc tính:

có nghĩa là: 90

phương sai: 25

# Ví dụ về Trình phân loại Naïve Bayes

Đưa ra một bản ghi thử nghiệm:  $X = (\text{Hoàn tiền} = \text{Có}, \text{Trạng thái} = \text{Độc thân}, \text{Thu nhập} = 80\text{K})$

naive Bayes Classifier:

$$P(\text{Refund}=\text{Yes}|\text{No}) = 3/7$$

$$P(\text{Refund}=\text{No}|\text{No}) = 4/7$$

$$P(\text{Refund}=\text{Yes}|\text{Yes}) = 0$$

$$P(\text{Refund}=\text{No}|\text{Yes}) = 1$$

$$P(\text{Marital Status}=\text{Single}|\text{No}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{No}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{No}) = 4/7$$

$$P(\text{Marital Status}=\text{Single}|\text{Yes}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{Yes}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{Yes}) = 0$$

For taxable income:

If class=No:    sample mean=110  
                         sample variance=2975

If class=Yes:    sample mean=90  
                         sample variance=25

$$-P(X|\text{Lớp}=\text{Không}) = P(\text{Hoàn tiền}=\text{Có}|\text{Lớp}=\text{Không})$$

$$-P(\text{Đã kết hôn}|\text{Lớp}=\text{Không})$$

$$-P(\text{Thu nhập}=120\text{K}|\text{Lớp}=\text{Không}) = \\ 3/7 * 2/7 * 0,0062 = 0,00075$$

$$-P(X|\text{Lớp}=\text{Có}) = P(\text{Hoàn tiền}=\text{Không}|\text{Hạng}=\text{Có})$$

$$-P(\text{Đã kết hôn}|\text{Lớp}=\text{Có})$$

$$-P(\text{Thu nhập}=120\text{K}|\text{Hạng}=\text{Có}) \\ = 0 * 2/3 * 0,01 = 0$$

- $P(\text{Không}) = 0,3, P(\text{Có}) = 0,7$  Vì  $P(X|$

$\text{Không})P(\text{Không}) > P(X|\text{Có})P(\text{Có})$  Do đó

$$P(\text{Không}|X) > P(\text{Có}|X)$$

$\Rightarrow \text{Lớp} = \text{Không}$



# Ví dụ về Trình phân loại Naïve Bayes

Đưa ra một bản ghi thử nghiệm: X - (Hoàn tiền- Không, Đã kết hôn, Thu nhập-120K)

naive Bayes Classifier:

$$P(\text{Refund}=\text{Yes}|\text{No}) = 3/7$$

$$P(\text{Refund}=\text{No}|\text{No}) = 4/7$$

$$P(\text{Refund}=\text{Yes}|\text{Yes}) = 0$$

$$P(\text{Refund}=\text{No}|\text{Yes}) = 1$$

$$P(\text{Marital Status}=\text{Single}|\text{No}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{No}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{No}) = 4/7$$

$$P(\text{Marital Status}=\text{Single}|\text{Yes}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{Yes}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{Yes}) = 0$$

For taxable income:

If class=No:     sample mean=110  
                         sample variance=2975

If class=Yes:     sample mean=90  
                         sample variance=25

$$-P(X|\text{Lớp}=\text{Không}) = P(\text{Hoàn tiền}=\text{Không}|\text{Lớp}=\text{Không})$$

$$-P(\text{Đã kết hôn}|\text{Lớp}=\text{Không})$$

$$-P(\text{Thu nhập}=120\text{K}|\text{Lớp}=\text{Không}) =$$

$$4/7 - 4/7 - 0,0072 = 0,0024$$

$$-P(X|\text{Lớp}=\text{Có}) = P(\text{Hoàn tiền}=\text{Không}|\text{Hạng}=\text{Có})$$

$$-P(\text{Đã kết hôn}|\text{Lớp}=\text{Có})$$

$$-P(\text{Thu nhập}=120\text{K}|$$

Lớp=Có)

$$= 1 - 0 - 1,2 \cdot 10^{-9} = 0$$

$$P(\text{Không}) = 0,3, P(\text{Có}) = 0,7$$

$$\text{Vì } P(X|\text{Không})P(\text{Không}) > P(X|\text{Có})P(\text{Có})$$

$$\text{Do đó } P(\text{Không}|X) > P(\text{Có}|X)$$

=> **Lớp = Không**

# Tránh vấn đề không có xác suất

- Dự đoán Naïve Bayesian yêu cầu từng thăm dò có điều kiện. là khác không. Nếu không thì, vấn đề được dự đoán sẽ bằng không

$$P(X | C_i) = \prod_{k=1}^N P(x_k | C_i)$$

- Bán tại. Giả sử một tập dữ liệu có 1000 bộ dữ liệu, thu nhập = thấp (0), thu nhập = trung bình (990) và thu nhập = cao (10)
- Sử dụng Hiệu chỉnh Laplacian (hoặc công cụ ước tính Laplacian)
  - Thêm 1 vào mỗi trường hợp  $\text{Prob}(\text{thu nhập} = \text{thấp}) = 1/1003$   $\text{Prob}(\text{thu nhập} = \text{trung bình}) = 991/1003$   $\text{Prob}(\text{thu nhập} = \text{cao}) = 11/1003$
  - Vấn đề “đã sửa”. ước tính gần với các đối tác “không được điều chỉnh” của chúng

# Ví dụ về Trình phân loại Naïve Bayes

Đưa ra một bản ghi thử nghiệm:  $X = (\text{Hoàn tiền} = \text{Có}, \text{Trạng thái} = \text{Độc thân}, \text{Thu nhập} = 80\text{K})$

Với chức năng làm mịn Laplace

naive Bayes Classifier:

$$P(\text{Refund}=\text{Yes}|\text{No}) = 4/9$$

$$P(\text{Refund}=\text{No}|\text{No}) = 5/9$$

$$P(\text{Refund}=\text{Yes}|\text{Yes}) = 1/5$$

$$P(\text{Refund}=\text{No}|\text{Yes}) = 4/5$$

$$P(\text{Marital Status}=\text{Single}|\text{No}) = 3/10$$

$$P(\text{Marital Status}=\text{Divorced}|\text{No}) = 2/10$$

$$P(\text{Marital Status}=\text{Married}|\text{No}) = 5/10$$

$$P(\text{Marital Status}=\text{Single}|\text{Yes}) = 3/6$$

$$P(\text{Marital Status}=\text{Divorced}|\text{Yes}) = 2/6$$

$$P(\text{Marital Status}=\text{Married}|\text{Yes}) = 1/6$$

For taxable income:

If class=No:    sample mean=110  
                         sample variance=2975

If class=Yes:    sample mean=90  
                         sample variance=25

$$P(X|\text{Lớp}=\text{Không}) = P(\text{Hoàn tiền}=\text{Không}|\text{Lớp}=\text{Không})$$

$$-P(\text{Đã kết hôn}|\text{Lớp}=\text{Không})$$

$$-P(\text{Thu nhập}=120\text{K}|\text{Lớp}=\text{Không})$$

$$= 4/9 - 3/10 - 0,0062 = 0,00082$$

$$P(X|\text{Lớp}=\text{Có}) = P(\text{Hoàn tiền}=\text{Không}|\text{Hạng}=\text{Có})$$

$$-P(\text{Đã kết hôn}|\text{Lớp}=\text{Có})$$

$$-P(\text{Thu nhập}=120\text{K}|\text{Lớp}=\text{Có})$$

Lớp=Có)

$$= 1/5 - 3/6 - 0,01 = 0,001$$

- $P(\text{Không}) = 0,7, P(\text{Có}) = 0,3$

- $P(X|\text{Không})P(\text{Không}) = 0,0005$

- $P(X|\text{Có})P(\text{Có}) = 0,0003$

=>Lớp = Không

# Trình phân loại Naïve Bayes: Tập dữ liệu đào tạo

Lớp học:

C1:buys\_computer = 'có'

C2:buys\_computer = 'không'

Dữ liệu cần phân loại:

X = (tuổi <=30,

Thu nhập = trung bình,

Sinh viên = có

Tín dụng\_xếp hạng = Khá)

| tuổi    | thu nhập   | tudent | credit_rating | buys computer |
|---------|------------|--------|---------------|---------------|
| <=30    | cao        | KHÔNG  | hội chợ       | KHÔNG         |
| <=30    | cao        | KHÔNG  | xuất sắc      | KHÔNG         |
| 31...40 | cao        | KHÔNG  | hội chợ       | Đúng          |
| > 40    | trung bình | KHÔNG  | hội chợ       | Đúng          |
| > 40    | thấp       | Đúng   | hội chợ       | Đúng          |
| > 40    | thấp       | Đúng   | xuất sắc      | KHÔNG         |
| 31...40 | thấp       | Đúng   | xuất sắc      | Đúng          |
| <=30    | trung bình | KHÔNG  | hội chợ       | KHÔNG         |
| <=30    | thấp       | Đúng   | hội chợ       | Đúng          |
| > 40    | trung bình | Đúng   | hội chợ       | Đúng          |
| <=30    | trung bình | Đúng   | xuất sắc      | Đúng          |
| 31...40 | trung bình | KHÔNG  | xuất sắc      | Đúng          |
| 31...40 | cao        | Đúng   | hội chợ       | Đúng          |
| > 40    | trung bình | KHÔNG  | xuất sắc      | KHÔNG         |

# Trình phân loại Naïve Bayesian: Nhận xét

---

- Thuận lợi
  - Dễ để thực hiện
  - Kết quả tốt đạt được trong hầu hết các trường hợp
- Nhược điểm
  - Giả định: lớp độc lập có điều kiện, do đó mất độ chính xác Trên thực tế, tồn tại sự phụ thuộc giữa các biến
    - Ví dụ: bệnh viện: bệnh nhân: Hồ sơ: tuổi, tiền sử gia đình, v.v. Triệu chứng: sốt, ho, v.v., Bệnh tật: ung thư phổi, tiểu đường, v.v.
    - Sự phụ thuộc giữa những điều này không thể được mô hình hóa bởi Trình phân loại Naïve Bayesian
- Làm thế nào để giải quyết những sự phụ thuộc này?
  - Mạng lưới niềm tin Bayes

# Bài tập 1

Cho một dữ liệu huấn luyện về người mua máy tính, áp dụng thuật toán ID3 và ILA để xây dựng mô hình phân loại

| 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            |

Xác định nhãn lớp cho các mẫu:

1.  $X_{new}$  = (tuổi  $\leq 30$ , thu nhập = cao, sinh viên = có, xếp hạng tín chỉ = khá)
2.  $X_{new}$  = (tuổi  $> 40$ , thu nhập = cao, sinh viên = không, xếp hạng tín chỉ = khá)

# Bài tập 2

Cho một dữ liệu huấn luyện về chơi tennis hay không, áp dụng thuật toán ID3 và ILA để xây dựng mô hình phân loại:

| Day | Outlook  | Temperature | Humidity | Windy  | PlayTennis |
|-----|----------|-------------|----------|--------|------------|
| D1  | Sunny    | Hot         | High     | Weak   | No         |
| D2  | Sunny    | Hot         | High     | Strong | No         |
| D3  | Overcast | Hot         | High     | Weak   | Yes        |
| D4  | Rain     | Mild        | High     | Weak   | Yes        |
| D5  | Rain     | Cool        | Normal   | Weak   | Yes        |
| D6  | Rain     | Cool        | Normal   | Strong | No         |
| D7  | Overcast | Cool        | Normal   | Strong | Yes        |
| D8  | Sunny    | Mild        | High     | Weak   | No         |
| D9  | Sunny    | Cool        | Normal   | Weak   | Yes        |
| D10 | Rain     | Mild        | Normal   | Weak   | Yes        |
| D11 | Sunny    | Mild        | Normal   | Strong | Yes        |
| D12 | Overcast | Mild        | High     | Strong | Yes        |
| D13 | Overcast | Hot         | Normal   | Weak   | Yes        |
| D14 | Rain     | Mild        | High     | Strong | No         |

Xác định nhãn lớp cho các mẫu:

- Xnew = <Triển vọng=nắng, Nhiệt độ = mát mẻ, Độ ẩm = cao, Gió = mạnh>
- Xnew = <Triển vọng = u ám, Nhiệt độ = mát mẻ, Độ ẩm = cao, Gió = mạnh>

# Đánh giá và lựa chọn mô hình

---

- Số liệu đánh giá: Làm thế nào chúng ta có thể đo lường độ chính xác? Các số liệu khác cần xem xét?
- Sử dụng **bộ kiểm tra xác nhận** của các bộ dữ liệu được gắn nhãn lớp thay vì tập huấn luyện khi đánh giá độ chính xác
- Các phương pháp ước tính độ chính xác của bộ phân loại:
  - Phương pháp giữ lại, lấy mẫu con ngẫu nhiên
  - Xác thực chéo
  - Khởi động
- So sánh các phân loại:
  - Khoảng tin cậy
  - Phân tích chi phí-lợi ích và đường cong ROC



# Số liệu đánh giá phân loại : Ma trận hỗn loạn

Ma trận hỗn loạn:

| Lớp thực tế\Lớp dự đoán | $C_1$                 | $\bar{C}_1$          |
|-------------------------|-----------------------|----------------------|
| $C_1$                   | Tích cực thực sự (TP) | Âm tính giả (FN)     |
| $\bar{C}_1$             | Dương tính giả (FP)   | Âm tính thực sự (TN) |

Ví dụ về Ma trận nhầm lẫn:

| Lớp thực tế\Dự đoán<br>lớp học | mua_máy tính<br>= vâng | mua_máy tính<br>= không | Tổng cộng |
|--------------------------------|------------------------|-------------------------|-----------|
| mua_máy tính = có              | 6954                   | 46                      | 7000      |
| mua_máy tính = không           | 412                    | 2588                    | 3000      |
| Tổng cộng                      | 7366                   | 2634                    | 10000     |

- Được cho  $t$  lớp học, một mục nhập,  $CM_{t \times i, j}$  trong một ma trận hỗn loạn cho biết số bộ dữ liệu trong lớp  $T_i$  được bộ phân loại dán nhãn là lớp  $j$
- Có thể có thêm hàng/cột để cung cấp tổng số

# Số liệu đánh giá phân loại: Độ chính xác, tỷ lệ lỗi, độ nhạy và độ đặc hiệu

|     |    |    |        |
|-----|----|----|--------|
| A\P | C  | ❖C |        |
| C   | TP | FN | P      |
| ❖C  | FP | TN | N      |
|     | P' | N' | Tất cả |

- Độ chính xác của phân loại, hoặc tỷ lệ nhận dạng: tỷ lệ phần trăm của các bộ dữ liệu trong tập kiểm tra được phân loại chính xác  
 $\text{Độ chính xác} = (TP + TN) / \text{Tất cả}$
- tỷ lệ lỗi:  $1 - \text{độ chính xác}$ , hoặc  
 $\text{Tỷ lệ lỗi} = (FP + FN) / \text{Tất cả}$

## Vấn đề mất cân bằng lớp:

- Một lớp có thể hiếm, ví dụ như gian lận hoặc dương tính với HIV
- Có ý nghĩa đa số lớp tiêu cực và thiểu số của tầng lớp tích cực
- Nhạy cảm: Tỷ lệ nhận dạng tích cực thực sự  
 $\text{Độ nhạy} = TP / P$
- Tính đặc hiệu: Tỷ lệ nhận dạng âm tính thực sự  
 $\text{Độ đặc hiệu} = TN / N$

Các số liệu đánh giá của bộ phân loại: Độ chính xác và khả năng thu hồi cũng như các thước đo F

- Độ chính xác:độ chính xác – bao nhiêu % bộ dữ liệu được bộ phân loại gán nhãn là dương thực sự là dương

$$precision = \frac{TP}{TP + FP}$$

- Nhớ lại:tính đầy đủ – bao nhiêu % bộ dữ liệu dương được bộ phân loại gán nhãn là dương?

$$recall = \frac{TP}{TP + FN}$$

- Điểm tuyệt đối là 1,0
- Mối quan hệ nghịch đảo giữa độ chính xác và thu hồi

- Fđo lường (F1 hoặc F-điểm):ý nghĩa hài hòa của độ chính xác và thu hồi,

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

- $F_\beta$ :thước đo trọng số của độ chính xác và thu hồi
- gán trọng lượng gấp  $\beta$  lần để thu hồi cũng như độ chính xác

$$F_\beta = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

# Số liệu đánh giá phân loại: Ví dụ

| Lớp thực tế\Lớp dự đoán | ung thư = có | ung thư = không | Tổng cộng | Sự công nhận(%)       |
|-------------------------|--------------|-----------------|-----------|-----------------------|
| ung thư = có            | 90           | 210             | 300       | 30:00 (nhạy cảm)      |
| ung thư = không         | 140          | 9560            | 9700      | 98,56 (tính đặc hiệu) |
| Tổng cộng               | 230          | 9770            | 10000     | 96,40 (sự chính xác)  |

-Độ chính xác  $= 90/230 = 39,13\%$

Nhớ lại  $= 90/300 = 30,00\%$

## Đánh giá độ chính xác của trình phân loại: Phương pháp xác thực chéo và xác thực chéo

- Phương pháp giữ lại
  - Dữ liệu đã cho được phân chia ngẫu nhiên thành hai bộ độc lập
    - Tập huấn luyện (ví dụ: 2/3) để xây dựng mô hình
    - Bộ kiểm tra (ví dụ: 1/3) để ước tính độ chính xác Lấy
  - mẫu ngẫu nhiên : một biến thể của việc nắm giữ
    - Lặp lại lần giữ k lần, độ chính xác = trung bình. về độ chính xác thu được
- Xác thực chéo (k-gấp, trong đó k = 10 là phổ biến nhất)
  - Phân chia ngẫu nhiên dữ liệu thành k loại trừ lẫn nhau tập hợp con, mỗi tập hợp có kích thước xấp xỉ bằng nhau
  - Tại mỗi lần lặp thứ, sử dụng  $D_{\text{Tôi}}$  làm tập kiểm tra và các tập khác làm tập huấn luyện Bỏ đi
  - một lần : k nếp gấp ở đâu  $k = \#$  bộ dữ liệu, cho dữ liệu có kích thước nhỏ
  - \* CR phân tầng ờ ss-xác thực\* : nếp gấp được phân tầng sao cho lớp dist. trong mỗi lần là khoảng. giống như trong dữ liệu ban đầu