

# **KHAI PHÁ DỮ LIỆU**

## **Bài 5. Phân lớp dữ liệu**

**Giáo viên: TS. Trần Mạnh Tuấn**

**Bộ môn: Hệ thống thông tin**

**Khoa: Công nghệ thông tin**

**Email: [tmtuan@tlu.edu.vn](mailto:tmtuan@tlu.edu.vn)**

**Điện thoại: 0983.668.841**

# Nội dung

- **Tổng quan**
- **Các phương pháp phân lớp dữ liệu**

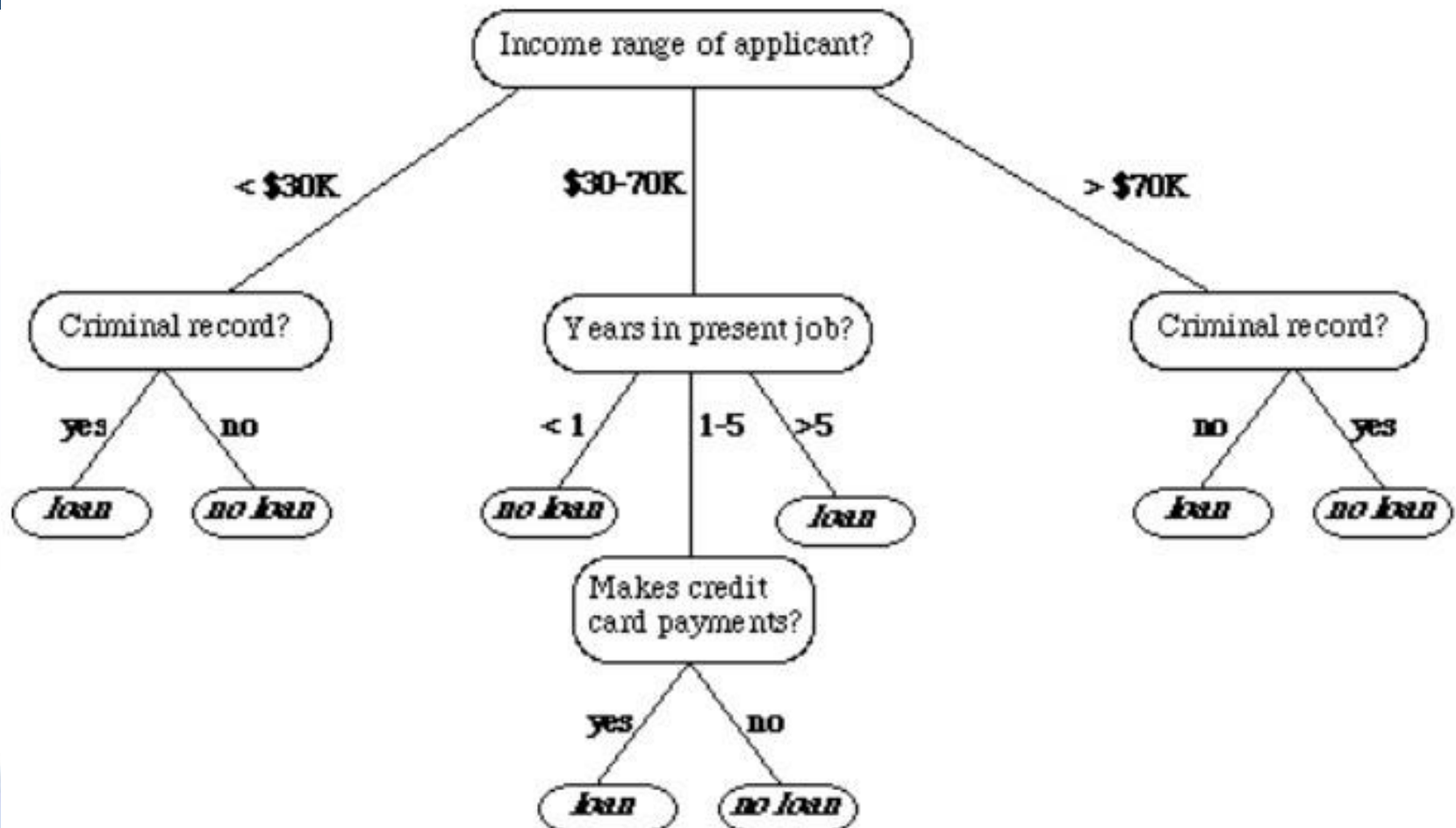
# Tổng quan

<i>Tid</i>	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???

# Tổng quan



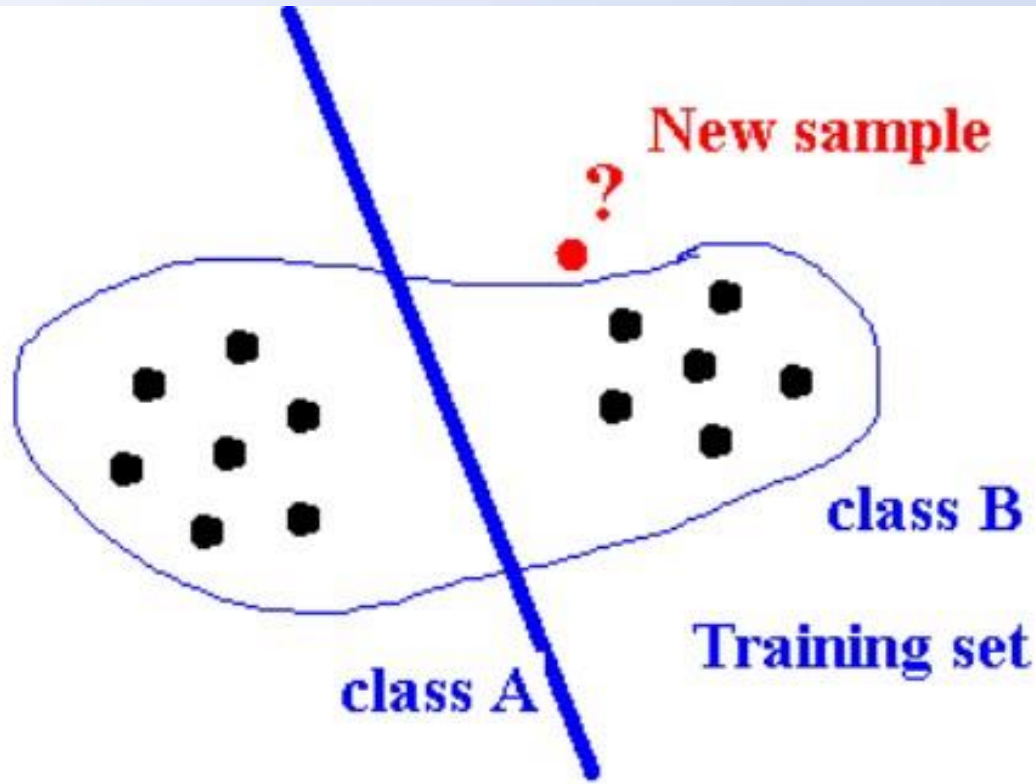
Với thông tin của một applicant A, xác định liệu ngân hàng có cho A vay không?

# Tổng quan

Khóa	MãSV	MônHọc1	MônHọc2	...	TốtNghệp
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?

# Tổng quan



Cho trước tập huấn luyện (training set), dẫn ra mô tả về class A và class B?  
Cho trước mẫu/đối tượng mới, làm sao xác định class cho mẫu/đối tượng đó?  
Liệu class đó có thực sự phù hợp/đúng cho mẫu/đối tượng đó?



# Tổng quan

Phân lớp dữ liệu (Data classification) là **xếp đối tượng DL vào một trong các lớp đã được xác định trước.**

Phân lớp gồm 2 bước:

**B ư ớ c 1 :** Xây dựng mô hình

**B ư ớ c 2 :** Vận hành mô hình.

# Tổng quan

## ***Quy trình phân lớp***

### **B1: xây dựng mô hình**

Mô tả tập các lớp xác định trước

Tập học/huấn luyện: các mẫu dành cho xây dựng mô hình.

Mỗi mẫu thuộc về 1 lớp đã định nghĩa trước.

Tìm luật phân lớp, cây quyết định hoặc công thức toán mô tả lớp.

### **B2: Vận hành mô hình**

Phân lớp các đối tượng chưa biết:

Xác định độ chính xác của mô hình, sử dụng tập dữ liệu kiểm tra độc lập.

Độ chính xác chấp nhận được -> áp dụng mô hình để phân lớp các mẫu chưa xác định được nhãn lớp.



# Tổng quan

- ◆ Các thuộc tính: liên tục và rời rạc
- ◆ Chọn một thuộc tính rời rạc đặc biệt gọi là TT *phân lớp*, các giá trị của nó gọi là *nhãn lớp*.
- ◆ Thuộc tính liên tục nhận các giá trị có thứ tự, thuộc tính rời rạc nhận các giá trị không có thứ tự.
- ◆ Các thuộc tính có thể nhận giá trị không xác định (chẳng hạn, vì những lí do khách quan ta không thể biết được giá trị của nó).

# Tổng quan

- ◆ Nhiệm vụ của quá trình phân lớp là thiết lập được **ánh xạ giữa giá trị của các thuộc tính với các nhãn lớp**.
- ◆ Mô hình biểu diễn quan hệ nói trên sau đó sẽ được dùng để xác định nhãn lớp cho các quan sát mới không nằm trong tập mẫu ban đầu.

# Tổng quan

## ◆ **Ứng dụng của phân lớp**

- ◆ Chuẩn đoán bệnh trong y học, phát hiện tiền giả, hay nghiên cứu thị trường.

## ◆ **Mô hình phân lớp (classification models):**

- ◆ Cây quyết định
- ◆ Mạng Bayes
- ◆ Mô hình thống kê,
- ◆ Mạng nơ ron,
- ◆ Giải thuật di truyền,...

# Tổng quan

## Xây dựng mô hình

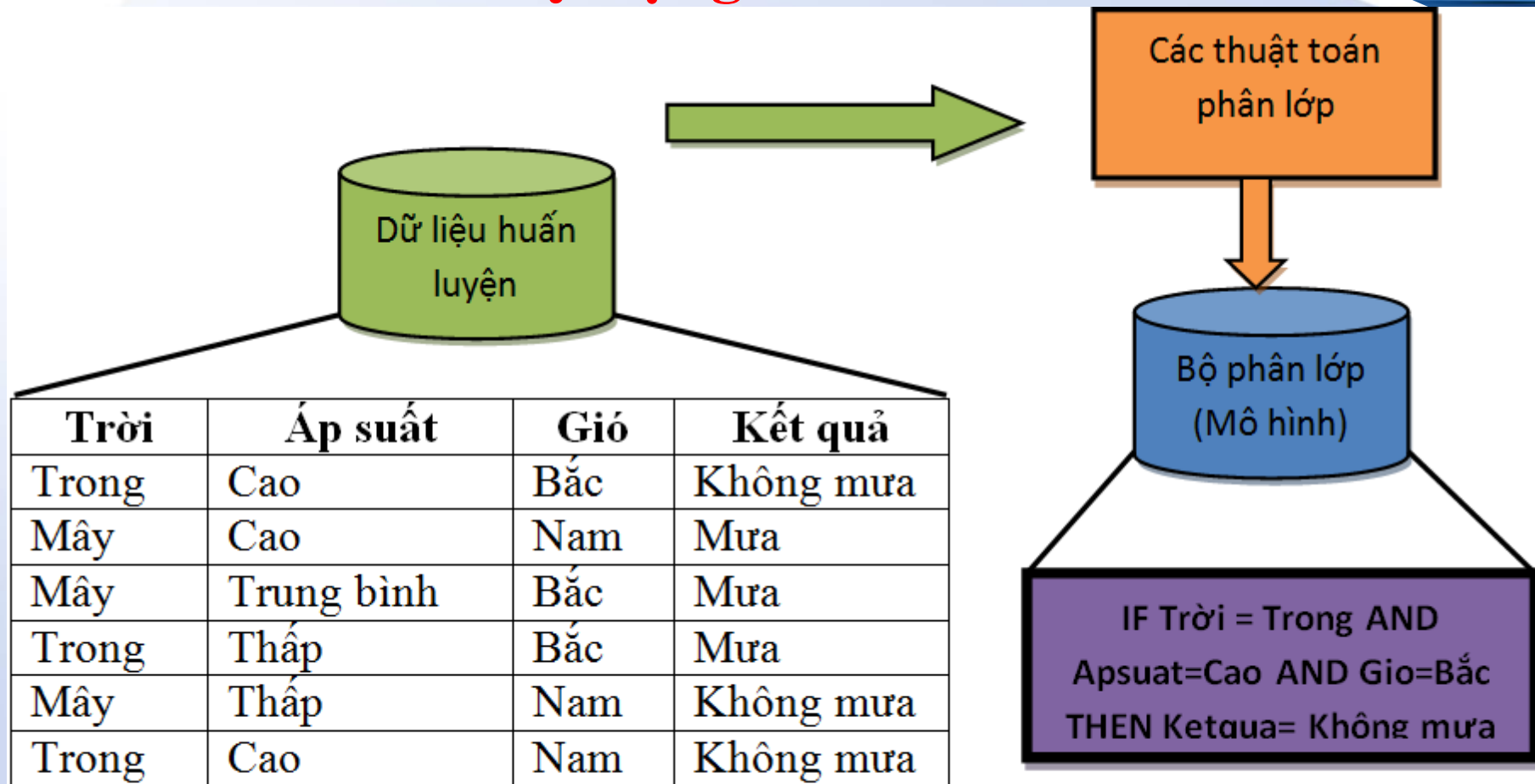
Mục tiêu mô tả một tập những lớp đã được định nghĩa trước trong đó mỗi bộ hoặc mẫu sẽ được gán về một lớp đã xác định trước bởi thuộc tính nhãn lớp.

Tập hợp những bộ được dùng để xây dựng mô hình được gọi là tập **dữ liệu học** (gọi tắt là tập học).

Mô hình được biểu diễn dưới dạng luật phân lớp, cây quyết định hoặc công thức toán học...

# Tổng quan

## Xây dựng mô hình



Hình: Xây dựng mô hình

# Tổng quan

## Vận hành mô hình

Mục đích là xác định lớp của dữ liệu trong tương lai hoặc phân lớp những đối tượng chưa biết.

Trước khi vận hành mô hình cần đánh giá độ chính xác của mô hình trong đó các **mẫu kiểm tra** (đã biết được lớp) được đem so sánh với kết quả phân lớp của mô hình.

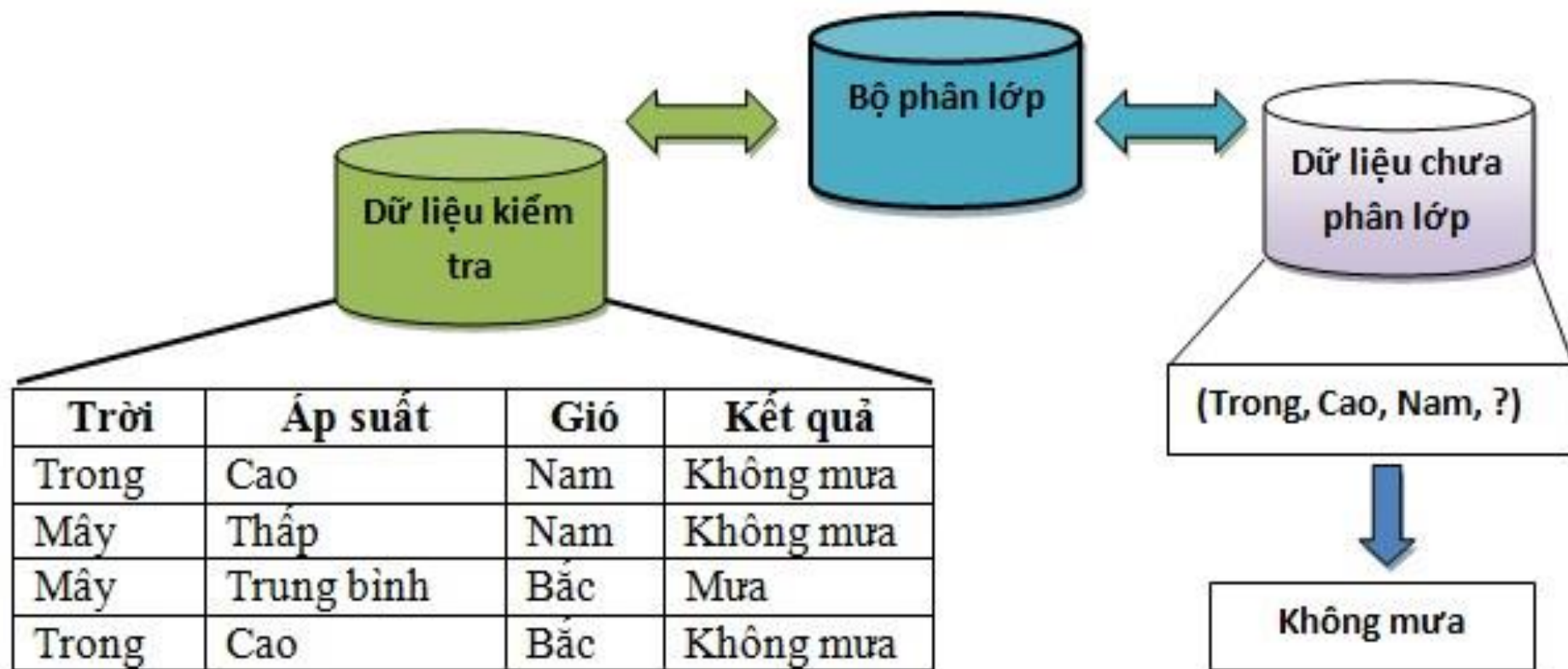
Độ chính xác là phần trăm của số mẫu kiểm tra được phân lớp đúng.

Tập kiểm tra và tập học là hai tập độc lập với nhau.



# Tổng quan

## Vận hành mô hình



# Tổng quan

## ▣ Phân loại dữ liệu (classification)

- Dạng phân tích dữ liệu nhằm rút trích các mô hình mô tả các lớp dữ liệu hoặc dự đoán xu hướng dữ liệu
- Quá trình gồm hai bước:
  - ▣ Bước học (giai đoạn huấn luyện): xây dựng bộ phân loại (classifier) bằng việc phân tích/học tập huấn luyện
  - ▣ Bước phân loại (classification): phân loại dữ liệu/đối tượng mới nếu độ chính xác của bộ phân loại được đánh giá là có thể chấp nhận được (acceptable)

**$y = f(X)$**  với  $y$  là nhãn (phần mô tả) của một lớp (class) và  $X$  là dữ liệu/đối tượng

- Bước học:  $X$  trong tập huấn luyện, một trị  $y$  được cho trước với  $X \rightarrow$  xác định  $f$
- Bước phân loại: đánh giá  $f$  với  $(X', y')$  và  $X' \leftrightarrow$  mọi  $X$  trong tập huấn luyện; nếu acceptable thì dùng  $f$  để xác định  $y''$  cho  $X''$  (mới)

1

# Tổng quan

## Các giải thuật phân loại dữ liệu

- Phân loại với cây quyết định (decision tree)
- Phân loại với mạng Bayesian
- Phân loại với mạng neural
- Phân loại với k phần tử cận gần nhất (k-nearest neighbor)
- Phân loại với suy diễn dựa trên tình huống (case-based reasoning)
- Phân loại dựa trên tiến hoá gen (genetic algorithms)
- Phân loại với lý thuyết tập thô (rough sets)
- Phân loại với lý thuyết tập mờ (fuzzy sets) ...



# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

### Cây quyết định:

Gồm các nút trong biểu diễn giá trị thuộc tính,  
Các nhánh biểu diễn đầu ra của kiểm tra,  
Nút lá biểu diễn nhãn lớp.

Cây được tạo theo hai giai đoạn là tạo cây và tỉa nhánh.

### Giai đoạn tạo cây:

Bắt đầu tất cả các mẫu học đều nằm ở nút gốc,  
Sau đó các mẫu học được phân chia một cách đệ  
quy dựa trên thuộc tính được chọn.

# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

**Bước tỉa nhánh:** tìm và xóa những nhánh có phần tử không thể xếp vào lớp nào cả.

**Bước vận hành:** kiểm tra những giá trị thuộc tính của mẫu đối với các giá trị trên nhánh của cây.

# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

### Thuật toán tạo cây quyết định:

**Bước 1:** Cây được xây dựng đệ quy từ trên xuống và theo cách chia để trị.

**Bước 2:** ban đầu tất cả mẫu học đều nằm ở gốc.

**Bước 3:** Thuộc tính được phân loại (nếu là giá trị liên tục thì được rời rạc hóa)

**Bước 4:** Các mẫu học được phân chia đệ quy dựa trên thuộc tính chọn lựa.

**Bước 5:** Kiểm tra những thuộc tính được chọn dựa trên kinh nghiệm hoặc của một tiêu chuẩn thống kê.



# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

### Điều kiện dừng phân chia tập học:

Tất cả những mẫu học đối với một nút cho trước đều cùng lớp.

Không còn thuộc tính nào để phân chia tiếp.

Không còn mẫu học

# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

### Độ lợi thông tin (Information gain)

Là đại lượng dùng để chọn thuộc tính nhằm phân chia tập học.

Thuộc tính được chọn là thuộc tính có độ lợi thông tin lớn nhất.

# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

Cho hai lớp P (Positive) và N(Negative), tập học S.

Lớp P có p phần tử và lớp N có n phần tử.

Khối lượng thông tin cần để quyết định các mẫu trong S thuộc về lớp P hay lớp N được xác định bởi:

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

# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

G/S thuộc tính A được chọn để phân hoạch S thành các tập hợp  $\{S_1, S_2, \dots, S_v\}$ .

Nếu  $S_i$  chứa  $p_i$  mẫu của lớp P và  $n_i$  mẫu của lớp N thì *entropy* cần để phân loại các đối tượng trong cây con  $S_i$  là:

$$E(A) = \sum_{i=1}^v \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

Độ lợi thông tin của nhánh A là:

$$Gain(A) = I(p, n) - E(A)$$

## Thuật toán ID3

Thuật toán ID3 học trên cây quyết định do Ross Quinlan(1983) đề xuất.

### Ý tưởng:

Tạo cây quyết định bằng việc sử dụng cách tìm kiếm từ trên xuống trong tập học.

Sử dụng độ lợi thông tin để chọn thuộc tính có khả năng phân loại tốt nhất.

# Một số phương pháp phân lớp

## Thuật toán ID3

Thuật toán ID3 (S, D, A)

Input: Tập học S, Thuộc tính quyết định D, tập thuộc tính A

Output: Nút gốc của Cây quyết định

Begin

- Tạo 'Nút\_gốc' cho cây quyết định
- If tất cả các mẫu học của S đều có trị của D là P, trả về cây có một nút duy nhất là Nút\_gốc với nhãn P.
- If tất cả mẫu học của S đều có trị của D là N, trả về cây có một nút duy nhất là "Nút\_gốc" với nhãn N
- If A là rỗng, trả về cây có nút duy nhất là Nút\_gốc với nhãn là trị phổ biến nhất của D trong tập mẫu
- Else Begin
  - o Gọi X là thuộc tính của A phân lớp S tốt nhất// tính độ lợi
  - o Gán X vào thuộc tính quyết định D của Nút\_gốc
  - o For each trị v của X
    - Thêm một nhánh cây mới dưới Nút\_gốc ứng với X=v
    - Gọi Sv là tập con của v trị của X là v
    - If Sv rỗng
      - Thêm dưới nhánh mới này, một nút lá có nhãn là trị phổ biến nhất của thuộc tính quyết định trong S.

Else

Thêm cây con vào dưới nhánh mới này  
ID3(Sv, D, A-{X})

End

Return Nút\_gốc



# Một số phương pháp phân lớp

## Thuật toán ID3

### Ví dụ: Minh họa thuật toán ID3.

Sử dụng dữ liệu “chơi tennis” trong bảng sau:

Các thuộc tính và miền giá trị bao gồm:

Thuộc tính Thời tiết có miền giá trị {Nắng, U\_ám, Mưa}

Thuộc tính Nhiệt độ có miền giá trị {Nóng, Mát, Ấm\_áp}

Thuộc tính Độ ẩm có miền giá trị {Cao, Vừa}

Thuộc tính Gió có miền giá trị {Có, Không}

Thuộc tính Lớp có miền giá trị {P,N}

# Một số phương pháp phân lớp

## Thuật toán ID3

Bảng : Tập dữ liệu học “chơi Tennis”

Thời tiết	Nhiệt độ	Độ ẩm	Gió	Lớp
Nắng	Nóng	Cao	Không	N
Nắng	Nóng	Cao	Không	N
U_ám	Nóng	Cao	Không	P
Mưa	Ấm áp	Cao	Không	P
Mưa	Mát	Vừa	Không	P
Mưa	Mát	Vừa	Có	N
U_ám	Mát	Vừa	Có	P
Nắng	Ấm_áp	Cao	Không	N
Nắng	Mát	Vừa	Không	P
Mưa	Ấm_áp	Vừa	Không	P
Nắng	Ấm_áp	Vừa	Có	P
U_ám	Ấm_áp	Cao	Có	P
U_ám	Nóng	Vừa	Không	P
Mưa	Ấm_áp	Cao	Có	N

# Một số phương pháp phân lớp

## Thuật toán ID3

Tính Entropy cho thuộc tính Thời tiết:

Thời tiết	$p_i$	$n_i$	$I(p_i, n_i)$
Nắng	2	3	0.971
U ám	4	0	0
Mưa	3	2	0.971

$$E(\text{Thời tiết}) = (5/14)I(2,3) + (4/14)I(4,0) + (5/14)I(3,2) = 0.694$$

$$\text{Gain}(\text{thời tiết}) = I(9,5) - E(\text{thời tiết}) = 0.246$$

Tương tự tính được các Gain khác

$$\text{Gain}(\text{Nhiệt độ}) = 0.029;$$

$$\text{Gain}(\text{Độ ẩm}) = 0.151$$

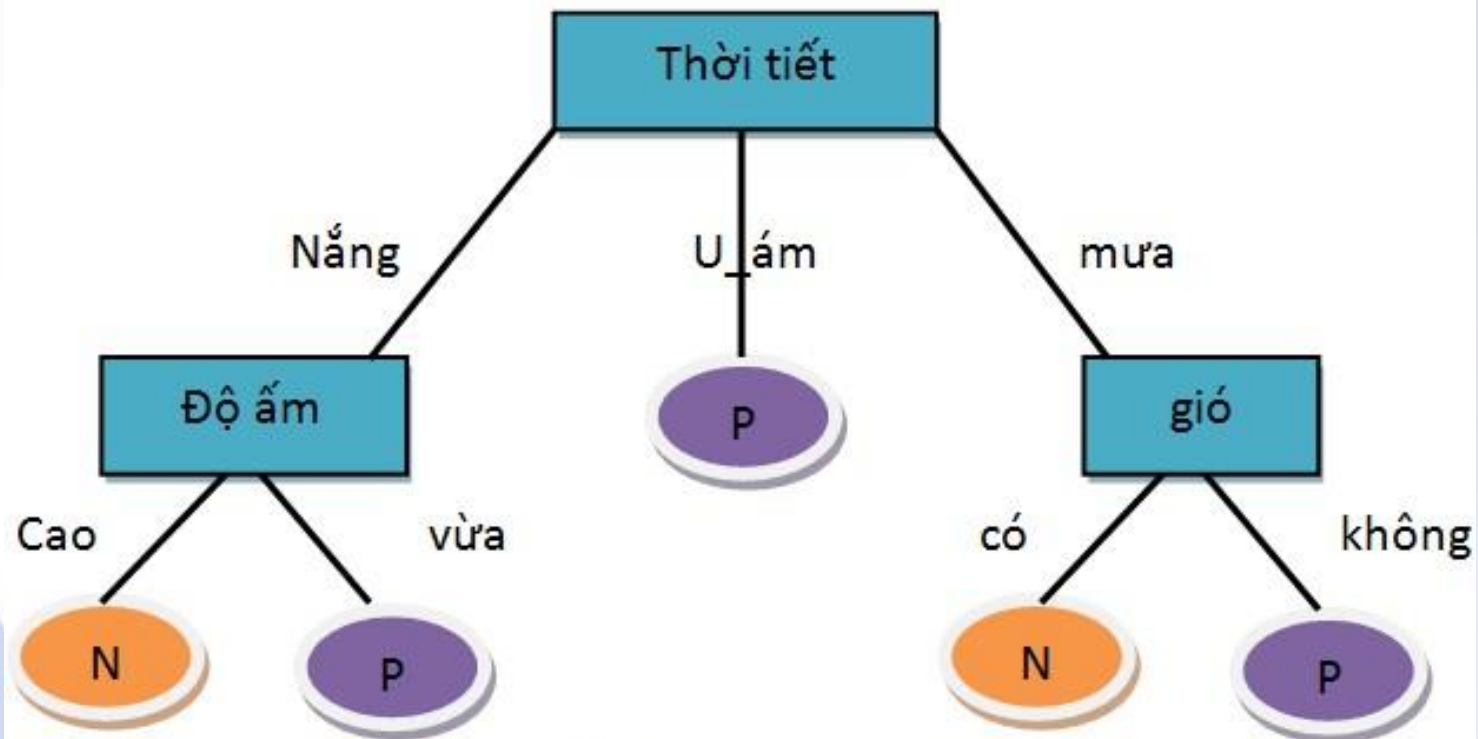
$$\text{Gain}(\text{gió}) = 0.048$$

# Một số phương pháp phân lớp

## Thuật toán ID3

Chọn thuộc tính có Gain lớn nhất là “thời tiết”

Áp dụng ID3 cho mỗi nút con của nút gốc này cho đến khi đạt đến nút lá hoặc nút có entropy=0.



Hình: Cây quyết định

# Một số phương pháp phân lớp

## Thuật toán ID3

### **Rút luật từ cây quyết định:**

Mỗi một đường dẫn từ gốc đến lá trong cây tạo thành một luật.

Mỗi cặp giá trị thuộc tính trên một đường dẫn tạo nên một sự liên kết.

Nút lá giữa quyết định phân lớp dự đoán

Các luật tạo được dễ hiểu hơn các cây

If thời tiết=Nắng AND Độ ẩm = Vừa THEN Chơi tennis

## Thuật toán ID3

### Nhược điểm của ID3:

ID3 hết khả năng phân chia tại một nút.

ID3 đòi hỏi số mẫu học lớn.

Khả năng khắc phục nhiễu của tập học là rất quan trọng khi ứng dụng thuật giải ID3.

Nếu có nhiễu và tập học không lớn thì ID3 có thể dẫn đến kết quả sai.



# Một số phương pháp phân lớp

## Thuật toán ID3

### Mở rộng của ID3:

ID3 được mở rộng cho trường hợp tập mẫu có thuộc tính liên tục. Lúc đó cần phân tích thuộc tính liên tục thành một tập rời rạc các khoảng.

Đối với các mẫu học có một số thuộc tính chưa có giá trị được thực hiện bằng cách gán trị thông dụng nhất của thuộc tính hoặc gán khả năng có thể có với từng giá trị khả dĩ.

## Thuật toán C4.5

C4.5 là phiên bản của ID3 trên một số khía cạnh sau:

Trong bước xây dựng cây, chỉ tạo mô hình dựa trên các bản ghi đã xác định đầy đủ giá trị thuộc tính.

Trong bước vận hành cây quyết định, có thể phân loại những bản ghi có những giá trị thuộc tính chưa biết bằng việc ước lượng xác suất những kết quả có khả năng xảy ra.

# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

### Độ đo Information Gain

- Dựa trên lý thuyết thông tin (information theory) của Claude Shannon về giá trị (nội dung thông tin) của tin
- Thuộc tính tương ứng với information gain lớn nhất sẽ được chọn làm splitting attribute cho node N.
  - Node N là node hiện tại cần phân hoạch các phần tử trong D.
  - Splitting attribute đảm bảo sự trùng lặp (impurity)/ngẫu nhiên (randomness) ít nhất giữa các phân hoạch tạo được.
  - Cách tiếp cận này giúp tối thiểu số phép thử (test) để phân loại một phần tử.

# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

### Độ đo Information Gain

- Lượng thông tin cần để phân loại một phần tử trong  $D$  (= Entropy của  $D$ ):  $\text{Info}(D)$ 
  - ▣  $p_i$ : xác suất để một phần tử bất kỳ trong  $D$  thuộc về lớp  $C_i$  với  $i = 1..m$
  - ▣  $C_{i,D}$ : tập các phần tử của lớp  $C_i$  trong  $D$

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

$$p_i = |C_{i,D}| / |D|$$



# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

### Độ đo Information Gain

- Lượng thông tin cần để phân loại một phần tử trong D dựa trên thuộc tính A:  $\text{Info}_A(D)$ 
  - ▣ Thuộc tính A dùng phân tách D thành v phân hoạch  $\{D_1, D_2, \dots, D_j, \dots, D_v\}$ .
  - ▣ Mỗi phân hoạch  $D_j$  gồm  $|D_j|$  phần tử trong D.
  - ▣ Lượng thông tin này sẽ cho biết mức độ trùng lặp giữa các phân hoạch, nghĩa là một phân hoạch chứa các phần tử từ một lớp hay nhiều lớp khác nhau.
  - ▣ Mong đợi:  $\text{Info}_A(D)$  càng nhỏ càng tốt.

$$\text{Info}_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} * \text{Info}(D_j)$$

# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

### Độ đo Information Gain

- Information gain chính là độ sai biệt giữa trị thông tin  $\text{Info}(D)$  ban đầu (trước phân hoạch) và trị thông tin mới  $\text{Info}_A(D)$  (sau phân hoạch với  $A$ ).

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$



# Một số phương pháp phân lớp

## Phân lớp bằng cây quyết định

Xây dựng cây quyết định từ cơ sở dữ liệu huấn luyện AllElectronics

- Dùng độ đo Information Gain
  - Dùng độ đo Gain Ratio
  - Dùng độ đo Gini Index
- Các cây quyết định học được giống nhau???
- Tiến hành đánh giá và phân loại với các cây quyết định học được

# Một số phương pháp phân lớp

## Phân lớp bằng Bayesian

### Sự phân hoạch và công thức Bayes

Cho  $H_1, H_2, \dots, H_n$  là một phân hoạch không gian mẫu  $M$  và  $A$  là biến cố bất kỳ trong  $M$ .

Ta có: 
$$P(A) = \sum_{i=1}^n P(H_i)P(A/H_i)$$

Các xác suất  $P(H_i)$  và  $P(A/H_i)$  được biết trước gọi là xác suất tiên nghiệm (apriori probability).

$P(H_i/A)$  là xác suất hậu nghiệm (a posterior probability).

$$P(H_i / A_i) = \frac{P(H_i)P(A / H_i)}{P(A)}$$

# Một số phương pháp phân lớp

## Phân lớp bằng Bayesian

### Bộ phân lớp Naïve Bayes

Cho  $V_1, V_2, \dots, V_m$  là phân hoạch không gian mẫu  $V$ , mỗi  $V_i$  là một lớp.

Không gian các thể hiện  $X$  gồm các thể hiện được mô tả bởi tập thuộc tính  $A_1, A_2, \dots, A_n$ .

Không gian các thể hiện  $X$  tập học. Khi có thể hiện mới với giá trị  $\langle a_1, a_2, \dots, a_n \rangle$ , bộ phân lớp sẽ xuất giá trị hàm phân lớp  $f(x)$  là một trong các  $V_i$ .

# Một số phương pháp phân lớp

## Phân lớp bằng Bayesian

Lấy giá trị có xác suất cao nhất  $V_{MAP}$  cho thể hiện mới (MAP - Maximun A Posterior).

$$V_{MAP} = \max_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n)$$

Sử dụng Bayes, ta có:

$$V = \max_{v_j \in V} \frac{P(v_j)P(a_1, a_2, \dots, a_n | v_j)}{P(a_1, a_2, \dots, a_n)} = \max_{v_j \in V} P(v_j)P(a_1, a_2, \dots, a_n | v_j)$$

# Một số phương pháp phân lớp

## Phân lớp bằng Bayesian

Tính  $P(v_j)$  bằng cách đếm số lần xuất hiện của giá trị đích trong  $v_j$  trên tập học.

Tính  $P(a_1, a_2, \dots, a_n)$ : G/S các thuộc tính là độc lập.  
Xác suất của một thể hiện quan sát được  $\langle a_1, a_2, \dots, a_n \rangle$  trên mỗi lớp  $v_j$  là tích các khả năng của từng thuộc tính riêng biệt trên  $v_j$ .

$$P(a_1, a_2, \dots, a_n | v_j) = \prod_i P(a_i | v_j)$$

# Một số phương pháp phân lớp

## Phân lớp bằng Bayesian

Viết lại công thức (NB - Naive Bayes):

$$V_{NB} = \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

Bộ phân lớp Bayes liên quan đến bước học trong đó  $P(v_j)$  và  $P(a_1, a_2, \dots, a_n)$  được tính dựa trên tập học.

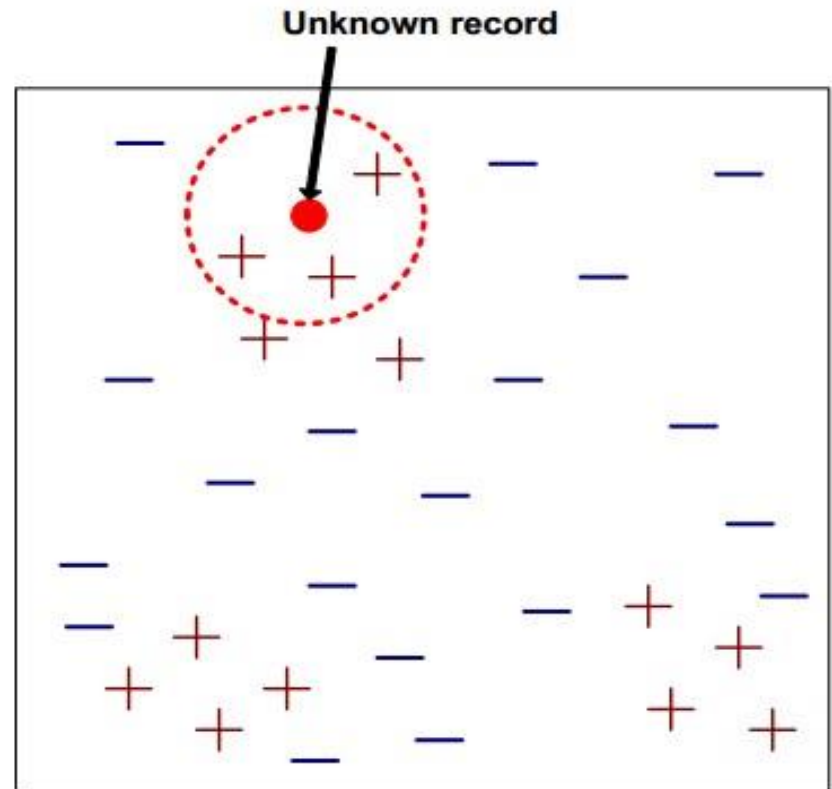


# Một số phương pháp phân lớp

## Phân lớp bằng KNN

Phân loại k-nn (k-nearest neighbor)

- Cho trước tập dữ liệu huấn luyện D với các lớp, phân loại record/object X vào các lớp dựa vào k phần tử tương tự với X nhất (dùng luật số đông: majority vote)



# Một số phương pháp phân lớp

## Phân lớp bằng KNN

### ▣ Chọn độ đo

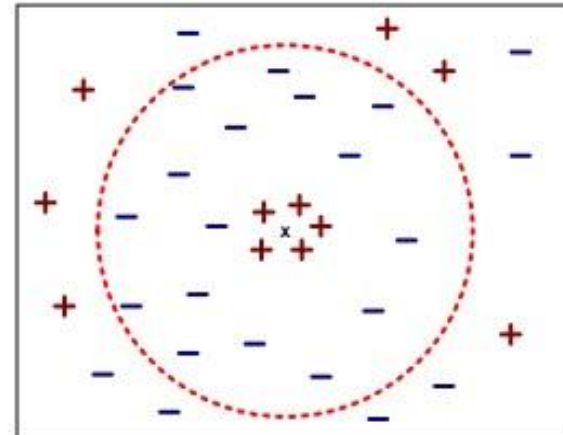
- Độ đo Euclidean

$$d(p, q) = \sqrt{\sum_i (p_i - q_i)^2}$$

### ▣ Chọn trị k

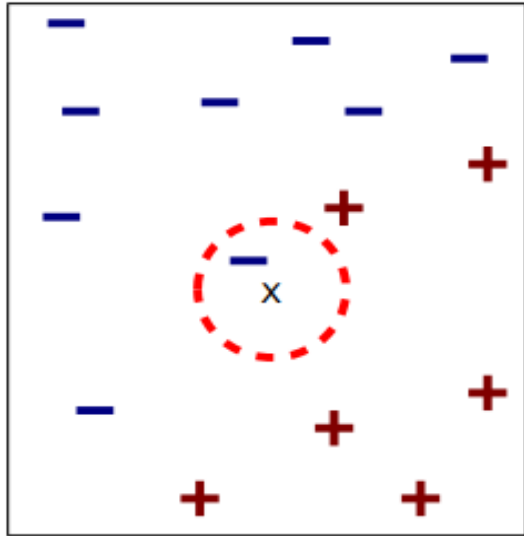
- Nếu k quá nhỏ thì kết quả dễ bị ảnh hưởng bởi nhiễu.
- Nếu k quá lớn thì nhiều phần tử láng giềng chọn được có thể đến từ các lớp khác.

k quá lớn!



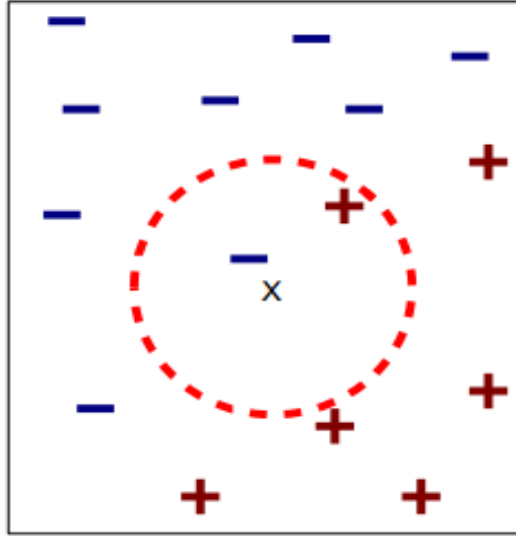
# Một số phương pháp phân lớp

## Phân lớp bằng KNN



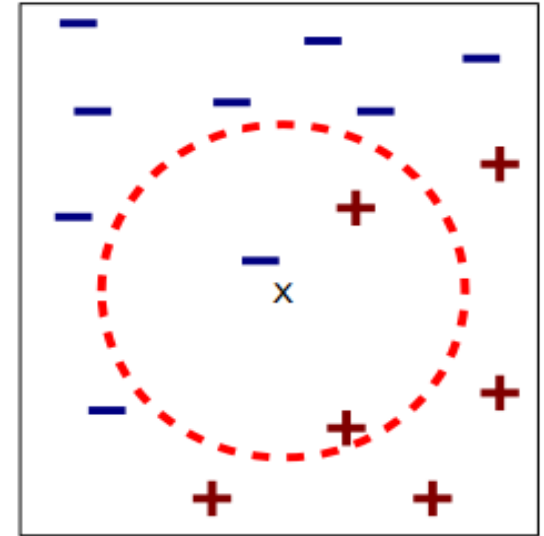
(a) 1-nearest neighbor

$X \in \text{MINUS}$



(b) 2-nearest neighbor

$X \in \text{MINUS}$   
hay  
 $X \in \text{PLUS ?}$



(c) 3-nearest neighbor

$X \in \text{PLUS}$

# Một số phương pháp phân lớp

## Phân lớp bằng KNN

Học có giám sát (supervised learning)

- Quá trình học và phân lớp dựa trực tiếp trên cá thể cần được phân lớp (instance-based).
  - ▣ Lazy learning → time-consuming nếu training data lớn
- Kết quả phân lớp phụ thuộc:
  - ▣ Độ đo khoảng cách để xác định sự tương tự.
    - Bị ảnh hưởng bởi các thuộc tính (chiều) không phù hợp
  - ▣ Trị  $k$ , số phần tử láng giềng →  $k \leq |D|^{1/2}$ 
    - Trả về phần tử mean của  $k$  phần tử láng giềng gần nhất
      - bị ảnh hưởng bởi nhiễu

**Trao đổi, câu hỏi?**