# Chương 4 : Phân loại và dự báo

Trịnh Anh Phúc [1]

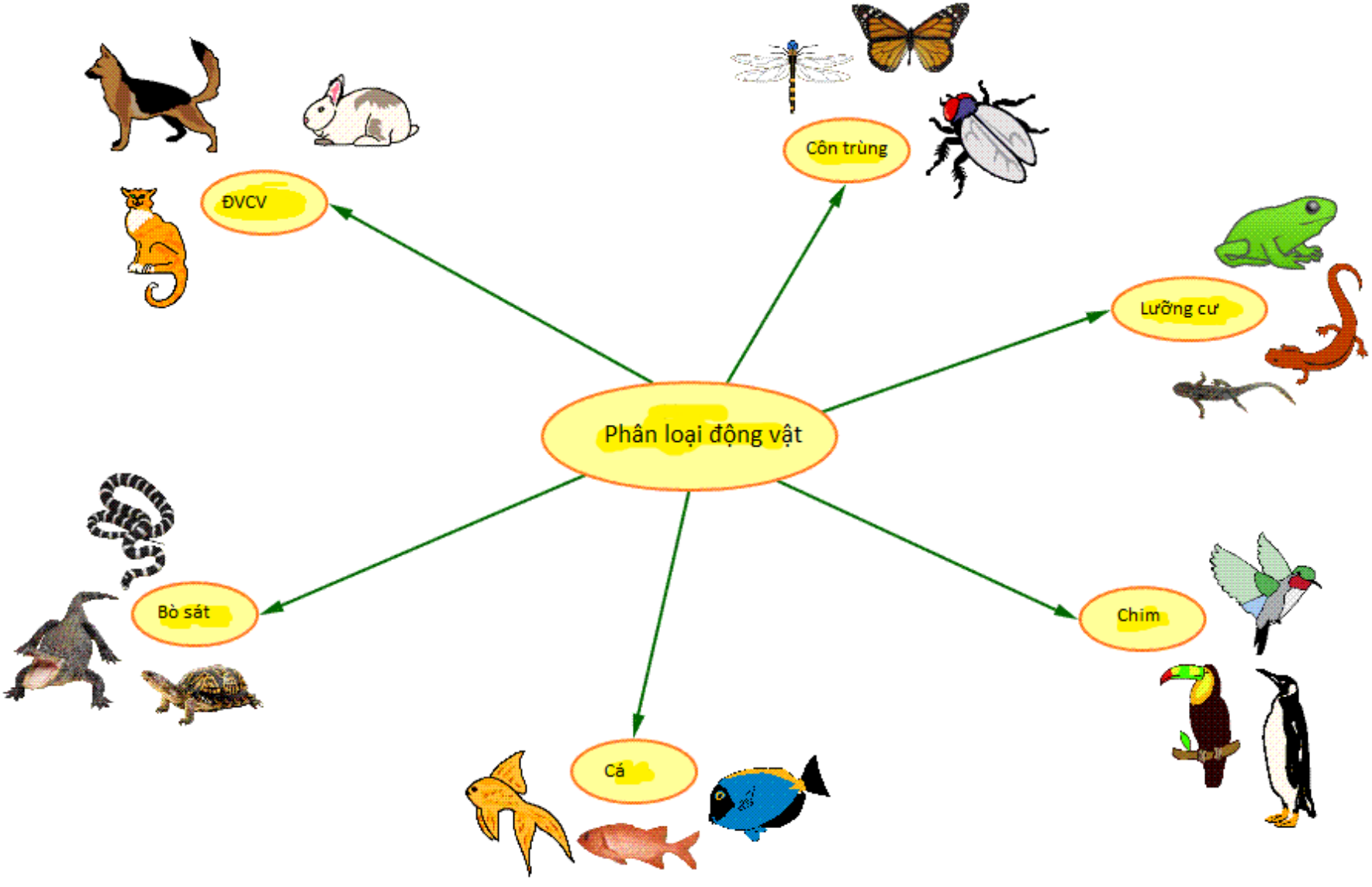
[1]Bộ môn Khoa Học Máy Tính, Viện CNTT & TT,
Trường Đại Học Bách Khoa Hà Nội

Ngày 20 tháng 8 năm 2014

# Giới thiệu

1. Bài toán phân loại
   - Định nghĩa
   - Các mô hình dùng giải bài toán phân loại
2. Mô hình phân loại trực tiếp Bay ét (Naive Bayes Classifier)
   - Định lý Bay ét
   - Mô hình
   - Ví dụ
3. Mô hình phân loại k-hàng xóm gần nhất (k-Nearest Neighbor Classifier)
   - Tiền đề khởi phát mô hình
   - Mô hình
   - Ví dụ
4. Mô hình cây quyết định (Decision Tree Classifier)
   - Mô hình
   - Thông tin thêm - information gain
   - Xây dựng cây quyết định

Hình: Phân loại động vật sống


Phân loại động vật
ĐVCV
Côn trùng
Lưỡng cư
Bò sát
Cá
Chim

## Định nghĩa về bài toán phân loại

Bài toán phân loại là bài toán xác định đối tượng quan sát thuộc về nhóm (lớp) các đối tượng đã được phân biệt, đã được nhận dạng hay có hiểu biết trước đó. Như vậy có ba đặc tính đi kèm với bài toán phân loại

- Phân biệt (Differentiated)
- Nhận biết (Recognized)
- Hiểu biết (Understood)

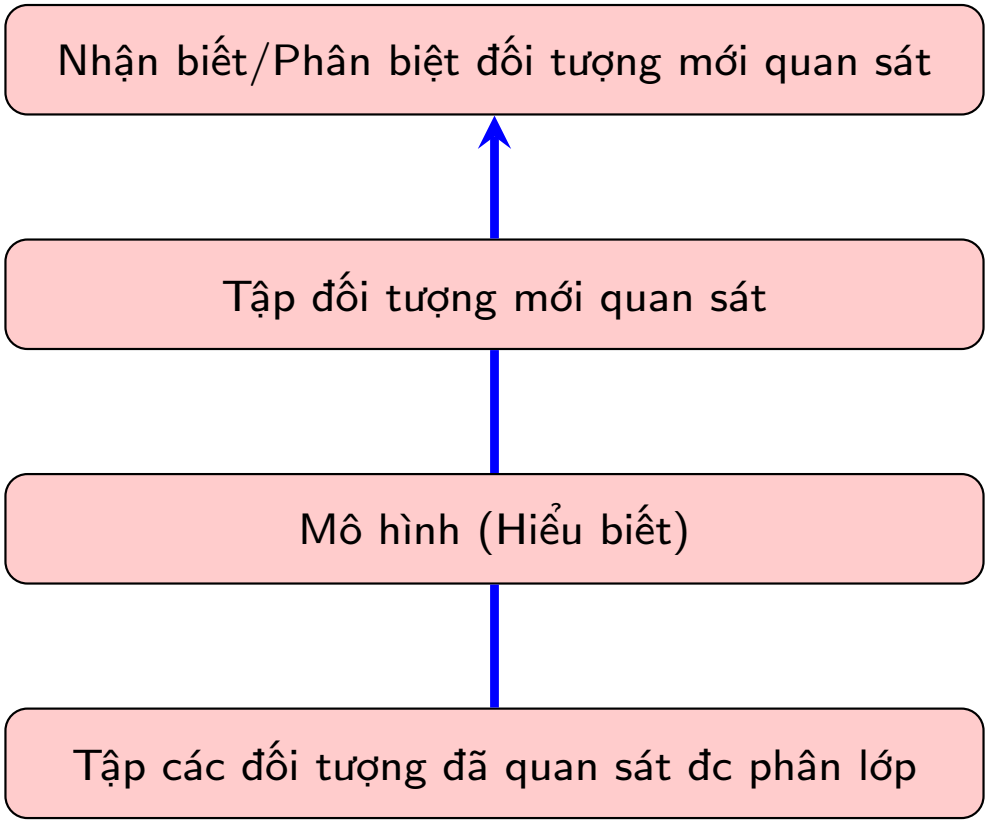
# Định nghĩa về bài toán phân loại

## Mô hình phân loại

Được dùng trong bài toán phân loại để thực hiện các vai trò :

- Hiểu biết lớp các đối tượng thông qua tập đã được quan sát trước đó
- Khi có đối tượng mới được quan sát, phân biệt được nó với các đối tượng đã quan sát
- Nhận biết đối tượng thuộc một nhóm (lớp) nào đã được quan sát trước đó

# Định nghĩa về bài toán phân loại




Hình: Tiến trình xử lý bài toán phân loại dựa trên mô hình

# Định nghĩa về bài toán phân loại

Có rất nhiều mô hình dùng để phân loại

- Sử dụng luật xác suất có điều kiện theo luật Bay ét
  - *Trực tiếp Bay ét* (NaiveBayes)
  - Mạng bay ét (BayesNet)
- Sử dụng hàm quyết định
  - Hàm logistic $f(x) = \frac{1}{1+e^{-x}}$
  - Hàm tuyến tính (Perceptron, SVM) $f(x) = w \cdot x + b$
  - Mạng RBF $f(x) = exp(-\frac{|x|^2}{\sigma^2})$
- Sử dụng cấu trúc cây
  - *Cây quyết định (Decision tree)*
  - Cây ngẫu nhiên (Random tree)
  - Rừng ngẫu nhiên (Random forest)
- Sử dụng phân loại dựa trên đối tượng (Instanced-Based Classifier)
  - *hàng xóm gần nhất* (Nearest Neigboors)

# Định nghĩa về bài toán phân loại

Do điều kiện của học phần, ta sẽ giới hạn giới thiệu các mô hình phân loại sau

1. Trực tiếp Bay ét (Naive Bayes Classifier)
2. k-hàng xóm gần nhất (k-Nearest Neighbors Classifier)
3. Cây quyết định (Decision Tree Classifier)

# Dành cho trả lời câu hỏi

# Phân loại trực tiếp Bay ét

## Định lý Bay ét

- Cho hai biến ngẫu nhiên $X, Y$
- Xác xuất hợp hai biến $Pr(X = x, Y = y)$
- Xác suất có điều kiện $Pr(Y = y|X = x)$
- Quan hệ giữa hai xác suất

$$Pr(X, Y) = Pr(X|Y)Pr(Y) = Pr(Y|X)Pr(X)$$

- **Định lý Bay ét** được phát biểu như sau

$$Pr(Y|X) = \frac{Pr(X|Y)Pr(Y)}{Pr(X)}$$

# Phân loại trực tiếp Bay ét

## Ứng dụng vào bài toán phân loại

- X là các biến thuộc tính
- Y là biến nhãn lớp
- Y phụ thuộc các biến thuộc tính X một cách không tất định (có tính xác suất)
- Vậy ta có thể nắm bắt sự phụ thuộc này thông qua mối quan hệ giữa xác suất hậu nghiệm $Pr(Y|X)$ và xác suất tiền nghiệm $Pr(Y)$ bởi công thức Bay ét

# Phân loại trực tiếp Bay ét

## Mô hình

Gồm có hai pha

1. Pha hiểu biết : Sử dụng tập các đối tượng đã được quan sát và phân lớp để tính xác suất hậu nghiệm $Pr(Y|X)$
2. Pha phân biệt/nhận biết : Với đối tượng mới quan sát $X'$ ta lấy lớp $Y'$ sao cho xác suất hậu nghiệm $Pr(Y'|X')$ lớn nhất có thể

# Phân loại trực tiếp Bay ét

| TT | X={outlook, | temperature, | humidity, | windy} | Y={play} |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | false | no |
| 2 | sunny | hot | high | true | no |
| 3 | overcast | hot | high | false | yes |
| 4 | rainy | mild | high | false | yes |
| 5 | rainy | cool | normal | false | yes |
| 6 | rainy | cool | normal | true | no |
| 7 | overcast | cool | normal | true | yes |
| 8 | sunny | mild | high | false | no |

- X'={outlook=rainy,temperature=cool,humidity=high,windy=true}
- Với Y' ta xác định max{ Pr(play=yes|X'), Pr(play=no|X')} ⇒ chơi tennis

# Phân loại trực tiếp Bay ét

Quay trở lại công thức Bay ét

$$Pr(Y|X) = \frac{Pr(X|Y)Pr(Y)}{Pr(X)}$$

- Do Pr(X) là hằng số nên có thể bỏ mặc
- Pr(Y) là xác suất từng lớp của tập các đối tượng đã được quan sát, e.g. Y={play=yes, play=no}
- Còn lại ta cần biết Pr(X|Y) ? ⇒ giả thuyết để tính

# Phân loại trực tiếp Bay ét

Giả sử các thuộc tính là *độc lập có điều kiện* [1] - trên $Y = y$, ta có công thức tính xác suất có điều kiện

$$Pr(X|Y = y) = \prod_{i=i}^{d} Pr(X_i|Y = y)$$

trong đó $d$ là số các thuộc tính trong $X = \{X_1, X_2, .., X_d\}$ thế vào công thức Bay ét

$$Pr(Y|X) = \frac{\prod_{i=i}^{d} Pr(X_i|Y)Pr(Y)}{Pr(X)}$$

[1] Định nghĩa : Cho ba biến ngẫu nhiên $X, Y, Z$, ký hiệu $X \perp\!\!\!\perp Y|Z$ nghĩa là hai biến $X, Y$ độc lập có điều kiện với biến Z, ta có $P(X, Y|Z) = P(X|Z)P(Y|Z)$

# Phân loại trực tiếp Bay ét

| TT | X={outlook, | temperature, | humidity, | windy} | Y={play} |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | false | no |
| 2 | sunny | hot | high | true | no |
| 3 | overcast | hot | high | false | yes |
| 4 | rainy | mild | high | false | yes |
| 5 | rainy | cool | normal | false | yes |
| 6 | rainy | cool | normal | true | no |
| 7 | overcast | cool | normal | true | yes |
| 8 | sunny | mild | high | false | no |

Với X'={outlook=rainy, temperature=cool, humidity=high, windy=true}

$$P(\text{play=yes}|X') = \frac{P(\text{rainy}|\text{yes})P(\text{cool}|\text{yes})P(\text{high}|\text{yes})P(\text{true}|\text{yes})P(\text{play=yes})}{P(X)}$$

$$\approx 2/4 \times 2/4 \times 2/4 \times 1/4 \times 4/8$$

# Phân loại trực tiếp Bay ét

| TT | X={outlook, | temperature, | humidity, | windy} | Y={play} |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | false | no |
| 2 | sunny | hot | high | true | no |
| 3 | overcast | hot | high | false | yes |
| 4 | rainy | mild | high | false | yes |
| 5 | rainy | cool | normal | false | yes |
| 6 | rainy | cool | normal | true | no |
| 7 | overcast | cool | normal | true | yes |
| 8 | sunny | mild | high | false | no |

Với X'={outlook=rainy, temperature=cool, humidity=high, windy=true}

$$P(\text{play=no}|X') = \frac{P(\text{rainy}|\text{no})P(\text{cool}|\text{no})P(\text{high}|\text{no})P(\text{true}|\text{yes})P(\text{play=no})}{P(X)}$$

$$\approx 1/4 \times 1/4 \times 3/4 \times 2/4 \times 4/8$$

# Phân loại trực tiếp Bay ét

Vậy để xác định lớp Y' được lựa chọn,

$$\max \{P(\text{play=yes}|X'), P(\text{play=no}|X')\}$$

theo ví dụ trên

$$P(\text{play=yes}|X') \approx 2/4 \times 2/4 \times 2/4 \times 1/4 \times 4/8$$
$$P(\text{play=no}|X') \approx 1/4 \times 1/4 \times 3/4 \times 2/4 \times 4/8$$

Vậy P(play=yes|X')=57% còn P(play=no|X')=43% thì xác suất để chơi tennis sẽ cao hơn

# Phân loại trực tiếp Bay ét

## Nhận xét chung

- Chấp nhận dữ liệu bị mất giá trị
- Dễ dàng loại bỏ nhiều dữ liệu
- Thường dùng cho các dữ liệu có trường thuộc tính dạng rời rạc
- Trong trường hợp dữ liệu có trường thuộc tính dạng số, phải thêm giả thuyết cho phân bố tương ứng

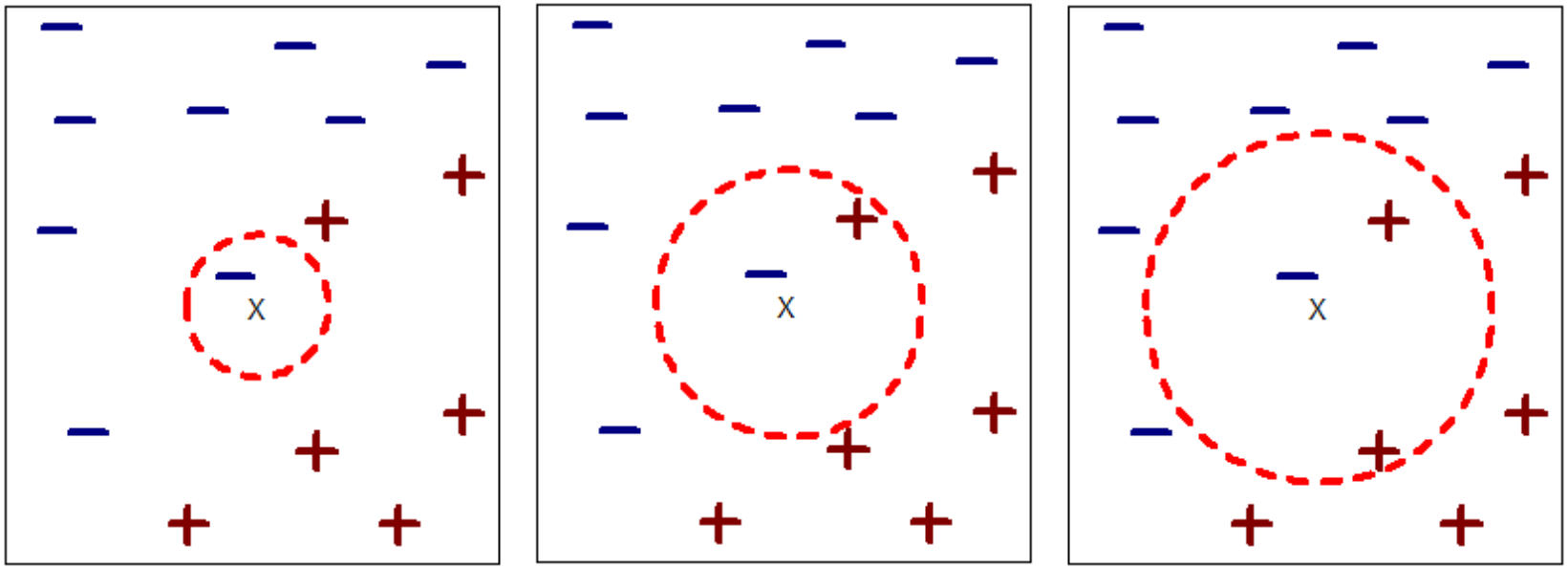
# Dành cho trả lời câu hỏi

# Mô hình phân loại k-hàng xóm gần nhất

## Tiền đề khởi phát mô hình

Một đối tượng mới quan sát sẽ được xếp cùng nhóm (lớp) với các đối tượng đã quan sát nêu "khoảng cách" giữa chúng là nhỏ nhất có thể. Từ 'khoảng cách' được hiểu theo nghĩa rộng là độ đo sự giống nhau của đối tượng mới quan sát với các thành viên của nhóm.

# Mô hình phân loại k-hàng xóm gần nhất




a) 1-hàng xóm gần nhất b) 2-hàng xóm gần nhất c) 3-hàng xóm gần nhất

Hình: Hình trên minh họa tiền đề của mô hình k- hàng xóm gần nhất với dữ liệu biểu diễn đối tượng mới quan sát được đánh dấu x còn các đối tượng đã biết được chia thành hai nhóm (đánh dấu +,-)

# Mô hình phân loại k-hàng xóm gần nhất

## Mô hình

Xét dữ liệu x biểu diễn đối tượng mới quan sát

- xác định k-hàng xóm gần nhất theo "khoảng cách"
  - khoảng cách Euclide
  - khoảng cách Hamming
  - khoảng cách Jaccard [2]
- xác định thuộc nhóm (lớp) theo sự phân lớp của k-hàng xóm gần nhất
  - nếu đa số k-hàng xóm gần nhất thuộc lớp đó
  - nếu trọng số được đánh theo khoảng cách, lớp được chọn sẽ có tổng trọng số nhỏ nhất

[2] $d_J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$ trong đó $A, B$ là tập giá trị thuộc tính biểu diễn hai đối tượng

# Mô hình phân loại k-hàng xóm gần nhất

| TT | X={outlook, | temperature, | humidity, | windy} | Y={play} |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | false | no |
| 2 | sunny | hot | high | true | no |
| 3 | overcast | hot | high | false | yes |
| 4 | rainy | mild | high | false | yes |
| 5 | rainy | cool | normal | false | yes |
| 6 | rainy | cool | normal | true | no |
| 7 | overcast | cool | normal | true | yes |
| 8 | sunny | mild | high | false | no |

x={outlook=rainy, temperature=cool, humidity=high, windy=true}

- k=1
- khoảng cách Hamming $d_{\mathcal{H}}$
- đa số k-hàng xóm

# Mô hình phân loại k-hàng xóm gần nhất

| TT | X={outlook, | temperature, | humidity, | windy} | Y={play} | $d_{\mathcal{H}}$ |
|---|---|---|---|---|---|---|
| 1 | sunny | hot | high | false | no | 3 |
| 2 | sunny | hot | high | true | no | 2 |
| 3 | overcast | hot | high | false | yes | 3 |
| 4 | rainy | mild | high | false | yes | 2 |
| 5 | rainy | cool | normal | false | yes | 2 |
| 6 | rainy | cool | normal | true | no | 1 |
| 7 | overcast | cool | normal | true | yes | 2 |
| 8 | sunny | mild | high | false | no | 3 |

x={outlook=rainy, temperature=cool, humidity=high, windy=true} vậy ta có lớp tương ứng 1-hàng xóm gần nhất nhãn là (play=no) $\Rightarrow$ không chơi tennis

# Mô hình phân loại k-hàng xóm gần nhất

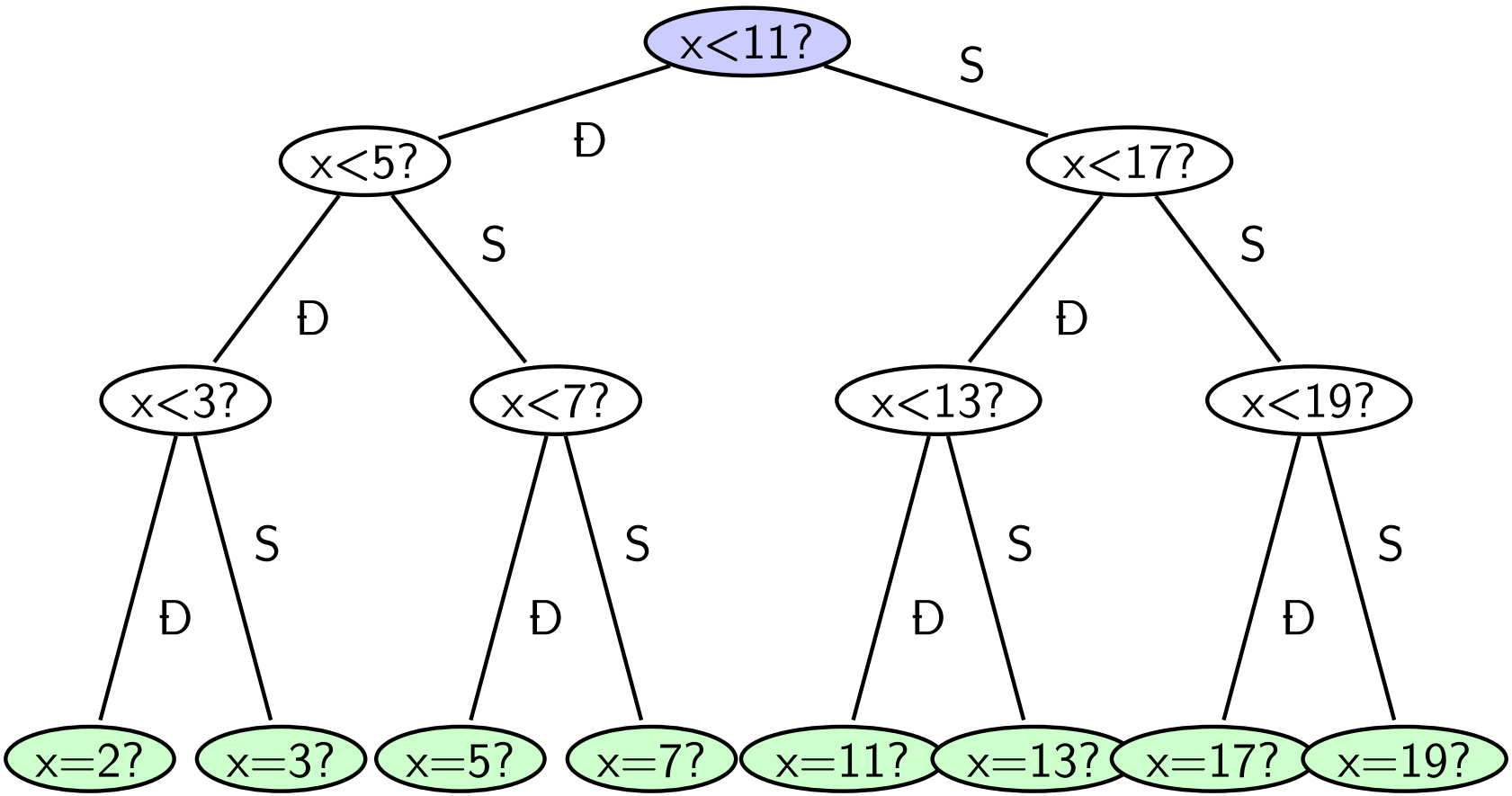
## Nhận xét

- Phu thuộc vào khoảng cách được chọn
- Có thể dùng được với các trường thuộc tính dạng giá trị số (numeric)
- Phải quét toàn bộ dữ liệu để tìm k-hàng xóm gần nhất
- Tuy vậy, lại có tính địa phương vì chỉ chọn k-hàng xóm thường $k << n$
- Giá trị k được xác định thường dựa vào cảm tính

# Dành cho trả lời câu hỏi

# Mô hình cây quyết định

Tìm kiếm nhị phân A = (2,3,5,7,11,13,17,19)




Đ : Đúng, S: Sai

# Mô hình cây quyết định

## Mô tả

Cây quyết định dùng trong KPDL được ứng dụng như sau

- Tập dữ liệu biểu diễn các đối tượng được khởi tạo tại nút gốc
- Quyết định được đưa ra thông qua phép duyệt từ gốc đến lá
- Các nút trong tương ứng với quyết định ứng giá trị trường thuộc tính
- Nhánh cây biểu diễn đầu ra của quyết định hay tập con dữ liệu được phân chia tương ứng quyết định nút cha
- Nút lá biểu diễn các nhãn lớp

# Mô hình cây quyết định

## Tạo cây quyết định gồm hai pha

1. Xây dựng cây
   - Khởi đầu, mọi dữ liệu biểu diễn tập các đối tượng đã quan sát đều xuất phát từ gốc
   - Chia tập các dữ liệu thành các phần nhỏ tương ứng quyết định ứng giá trị trường thuộc tính
2. Xén cành cây
   - Xác định và loại bỏ các nhánh tương ứng dữ liệu nhiễu hay không liên quan
3. Dùng để phân loại
   - Duyệt qua quyết định các thuộc tính thông qua nút trên cây đến nút lá để xác định nhãn lớp tương ứng

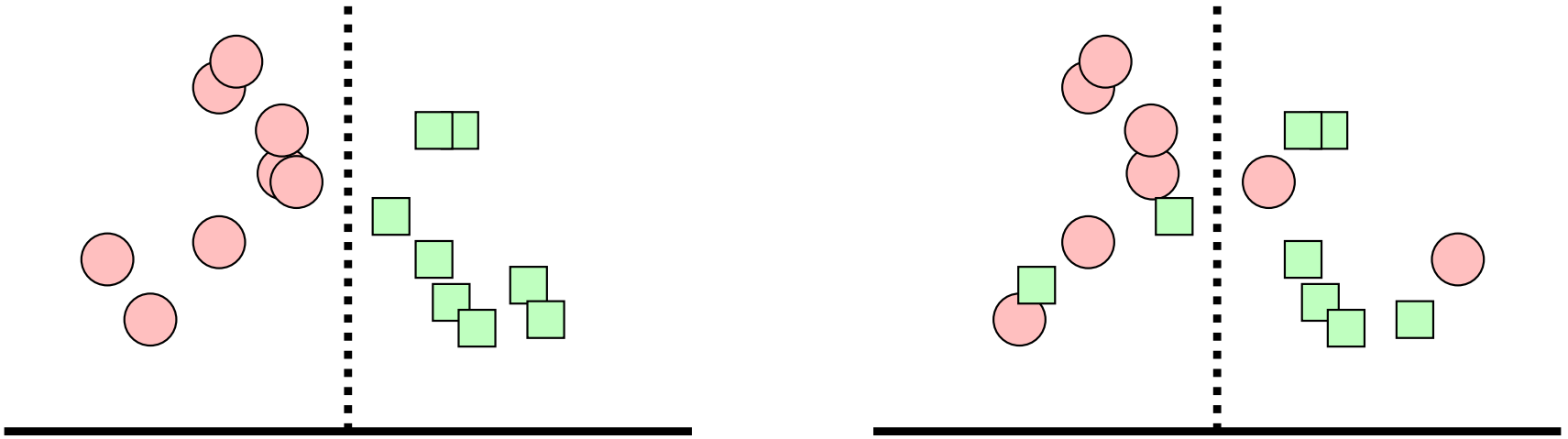
# Mô hình cây quyết định

## Ý nghĩa của thông tin thêm - information gain

Trong quá trình xây dựng cũng như xén cây quyết định trong bài toán phân loại, ta luôn dựa vào độ đo thông tin thêm - information gain.

# Mô hình cây quyết định

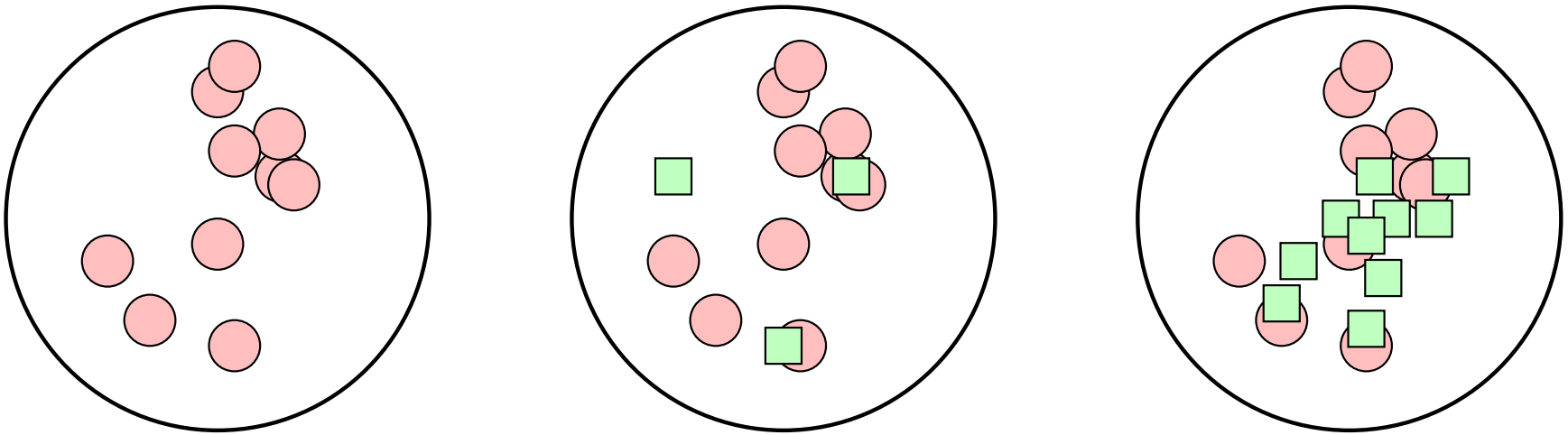
Kiểm tra xem quyết định của các nút trong cây có nhiều thông tin không ?



Hình: Các điểm dữ liệu được biểu diễn trên mặt phẳng được phân chia bởi đường quyết định nét đứt. Có hai kiểu dữ liệu tương ứng với hai lớp tô mầu hồng hình tròn và mầu xanh nhạt hình vuông.

# Mô hình cây quyết định

Entropy được dùng để đo độ không sạch của một nhóm dữ liệu

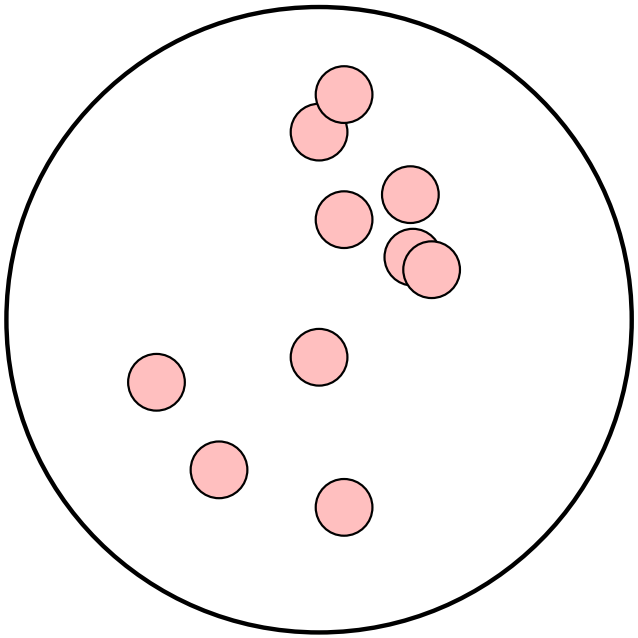


Công thức của entropy như sau

$$entropy = -p(c_1)\log_2 p(c_1) - p(c_2)\log_2 p(c_2)$$

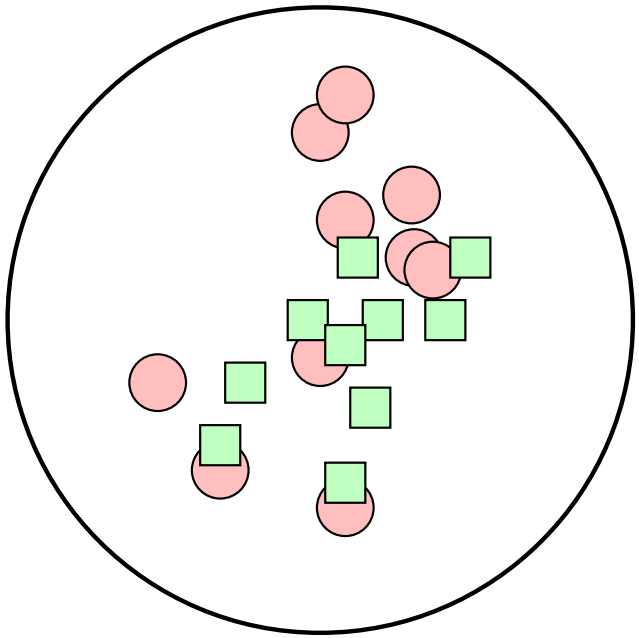
Vậy độ không sạch $0 \leq entropy \leq 1$, càng gần giá trị 1 càng không sạch và ngược lại.

# Mô hình cây quyết định



- Độ đo entropy của nhóm dữ liệu cùng thuộc một lớp

$$entropy = -1 \log_2 1 = 0$$



- Độ đo entropy của nhóm dữ liệu có số các đối tượng thuộc hai lớp bằng nhau

$$entropy = -1/2 \log_2 1/2 - 1/2 \log_2 1/2 = 1$$

# Mô hình cây quyết định

## Ý nghĩa về độ đo thông tin thêm - information gain

- Cần xác định thuộc tính hữu ích cho việc phân loại tập các dữ liệu biểu diễn đối tượng ?
- Độ đo thông tin thêm nói cho ta biết sự quan trọng của thuộc tính
- Ta sẽ dùng nó để sắp xếp lại các quyết định tương ứng thuộc tính cần phân chia trên cây

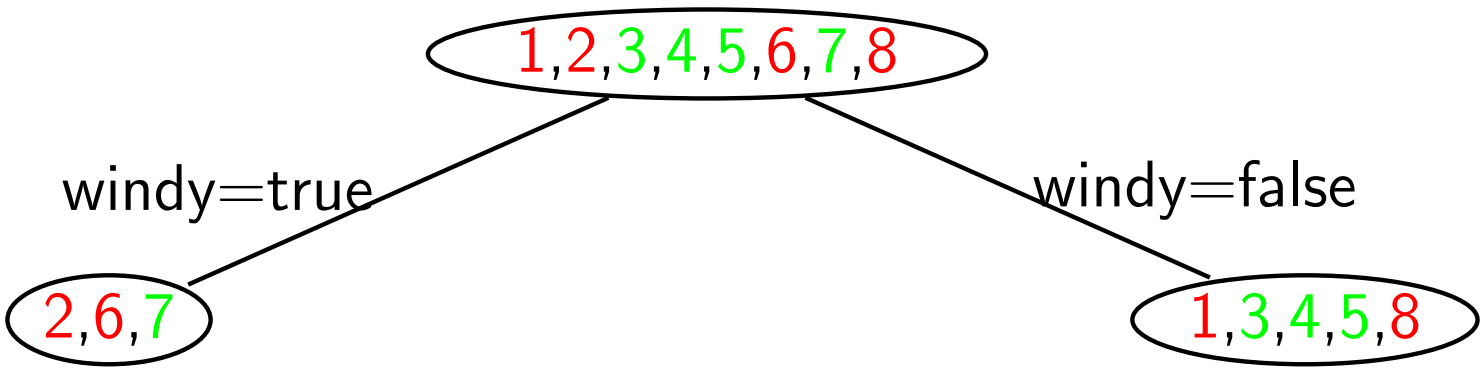
# Mô hình cây quyết định

| TT | outlook, | temperature, | humidity, | windy | play |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | false | no |
| 2 | sunny | hot | high | true | no |
| 3 | overcast | hot | high | false | yes |
| 4 | rainy | mild | high | false | yes |
| 5 | rainy | cool | normal | false | yes |
| 6 | rainy | cool | normal | true | no |
| 7 | overcast | cool | normal | true | yes |
| 8 | sunny | mild | high | false | no |

Phân chia tập 8 dữ liệu theo trường thuộc tính windy gồm hai giá trị windy={true, false}

# Mô hình cây quyết định

Hình: Quyết định dùng thuộc tính windy phân chia tập 8 dữ liệu. Các dữ liệu đánh số mầu đỏ (play=no) còn số mầu xanh (play=yes)




- entropy(cha) $= -4/8 \log_2 4/8 - 4/8 \log_2 4/8 = 1$
- entropy(con trái) $= -2/3 \log_2 2/3 - 1/3 \log_2 1/3 \approx 0.9256$
- entropy(con phải) $= -2/5 \log_2 2/5 - 3/5 \log_2 3/5 \approx 0.9709$
- IG = entropy(cha) - 3/8 entropy(con trái) - 5/8 entropy(con phải) $= 1 - 0.375 \times 0.9256 - 0.625 \times 0.9709 = 0.0462$

# Giải thuật Hunt xây dựng cây quyết định

- **Đầu vào** : $\mathbf{X}_t$ là tập các dữ liệu có nhãn lớp $\mathbf{y} = \{y_1, y_2, ..., y_c\}$ tại nút $t$
- **Đầu ra** : Cây quyết định dùng để phân loại $\mathbf{X}_t$

**Procedure** HuntAlgorithm($\mathbf{X}_t, t$)

1. **if**(mọi dữ liệu $\mathbf{x} \in \mathbf{X}_t$ thuộc cùng một lớp)**then** tạo nút $t$ thành nút lá
2. **else**
3. Xác định thuộc tính phân chia tập $\mathbf{X}_t \equiv \mathbf{X}_{t1} \cup \mathbf{X}_{t2} \cup ... \cup \mathbf{X}_{tm}$
4. **for**(mỗi $\mathbf{X}_{ti}$ với $i = \overline{1, m}$) **do**
5. Tạo nút con của nút $t$ là nút $ti$
6. HuntAlgorithm($\mathbf{X}_{ti}, ti$) // Gọi đệ qui
7. **endfor**
8. **endif**

**End**

# Giải thuật Hunt xây dựng cây quyết định

## Nhận xét

- Lời gọi ban đầu HuntAlgorithm($\mathbf{X}$,$r$) với $r$ là nút gốc còn $\mathbf{X}$ tập toàn bộ dữ liệu ban đầu
- Thuộc tính phân chia thành $\mathbf{X}_t \equiv \mathbf{X}_{t1} \cup \mathbf{X}_{t2} \cup ... \cup \mathbf{X}_{tm}$ nên thường là $m$ giá trị rời rạc
- Việc xác định trường thuộc tính phân chia, bước 3, tại mỗi mức của cây quyết định thì ta dùng độ đo thông tin thêm - Information Gain
- Để xác định $t$ đủ điều kiện thành lá, bước 1, có thể thêm khi tất cả các dữ liệu có cùng giá trị thuộc tính (giải thích đây là dạng dữ liệu gì ?)

# Mô hình cây quyết định

## Ưu điểm của cây quyết định

- Chi phí không cao khi xây dựng cây
- Rất nhanh khi xác định dữ liệu biểu diễn đối tượng mới thuộc lớp nào (đường đi từ gốc đến lá)
- Đường đi từ gốc đến lá tương ứng mệnh đề suy dẫn (gần gũi luật kết hợp)
- Với cây có kích thước nhỏ, ta dễ hình dung được bài toán phân loại
- Hạn chế với trường thuộc tính số, có nhiễu

# Mô hình cây quyết định

## Cải tiến của cây quyết định

- Sử dụng độ đo khác [3]
- Cần giảm kích thước thông qua xén cây, trường hợp mọi phân chia đến mức mọi dữ liệu thuộc cùng lớp không cần thiết. Có hai hướng tiếp cận
  - Phát triển cây quyết định đến tận cùng sau đó duyệt dưới lên thay nhánh cây con bằng nút lá (xén sau)
  - Xác định độ không sạch $\epsilon > 0$ đủ nhỏ nếu $entropy(\mathbf{X}_t) \leq \epsilon$ thì tạo nút $t$ thành nút lá luôn (xén đồng thời)

[3] Quinlan, J. R., (1986). Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers

# Chương trình Weka

## Yêu cầu

- Tải tập dữ liệu bank-data.arff
- Chạy giải thuật 1NN, kNN và J48
- Giải thích các tham số của các giải thuật