



Học viện Công nghệ Bưu chính Viễn thông
Khoa Công nghệ thông tin 1

Nhập môn trí tuệ nhân tạo

Giới thiệu học máy

Nguyễn Thị Mai Trang

maitrangnguyen.pt@gmail.com Telegram: @maitrangnguyen



Nội dung

-
- Giới thiệu
 - Học cây quyết định
 - Phân loại Bayes đơn giản
 - Học dựa trên ví dụ



Tài liệu tham khảo

1. N. Nilsson. Introduction to machine learning
<http://ai.stanford.edu/people/nilsson/mlbook.html>
2. T. Mitchell. Machine learning. McGraw-Hill, 1997.
3. E. Alpaydin. Introduction to machine learning. MIT Press, 2004.
4. M. Mohri, A. Rostamizadeh, A. Talwalkar. Foundations of Machine Learning. MIT Press, 2012.



Công cụ và dữ liệu

- Bộ công cụ Weka
 - <http://www.cs.waikato.ac.nz/~ml/weka>
- Kho dữ liệu mẫu UC Irvine
 - <http://www.ics.uci.edu/~mlearn/ML/Repository.html>

Một số Ứng dụng của học máy (1/3)

- Những ứng dụng khó lập trình theo cách thông thường do không tồn tại hoặc khó giải thích kinh nghiệm, kỹ năng của con người
 - Nhận dạng chữ viết, âm thanh, hình ảnh
 - Lái xe tự động, thám hiểm sao Hoả
- Chương trình máy tính có khả năng thích nghi: lời giải thay đổi theo thời gian hoặc theo tình huống cụ thể
 - Chương trình trợ giúp cá nhân
 - Định tuyến mạng

Một số ứng dụng của học máy (2/3)

- Khai phá (phân tích) dữ liệu
 - Hồ sơ bệnh án → tri thức y học
 - Dữ liệu bán hàng → quy luật kinh doanh



Một số ứng dụng của học máy (3/3)

Hầu hết các ứng dụng trí tuệ nhân tạo ngày nay có sử dụng học máy

...

- Web search
- Speech recognition
- Handwriting recognition
- Machine translation
- Information extraction
- Document summarization
- Question answering
- Spelling correction
- Image recognition
- 3D scene reconstruction
- Human activity recognition
- Autonomous driving
- Music information retrieval
- Automatic composition
- Social network analysis

...

- Product recommendation
- Advertisement placement
- Smart-grid energy optimization
- Household robotics
- Robotic surgery
- Robot exploration
- Spam filtering
- Fraud detection
- Fault diagnostics
- AI for video games
- Character animation
- Financial trading
- Protein folding
- Medical diagnosis
- Medical imaging

...

...



Học máy là gì?

- Học:
 - ...thu thập kiến thức hoặc kỹ năng...

Học máy là khả năng của chương trình máy tính sử dụng kinh nghiệm, quan sát, hoặc dữ liệu trong quá khứ để cải thiện công việc của mình trong tương lai thay vì chỉ thực hiện theo đúng các quy tắc đã được lập trình sẵn



Học máy là gì?

□ Học:

- ...thu thập kiến thức hoặc kỹ năng...
- “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience $E.$ ” Tom Mitchell (1997)

□ Học máy:

- Giải quyết vấn đề từ kinh nghiệm
- ...được thực hiện bởi chương trình máy tính có khả năng:
 - Thực hiện công việc T tốt hơn
 - Theo tiêu chí P
 - Nhờ sử dụng dữ liệu mẫu hoặc kinh nghiệm E

Ví dụ

- Học đánh cờ
 - T : đánh cờ
 - P : số ván thắng
 - E : kinh nghiệm tự chơi
- Học nhận dạng chữ
 - T : nhận dạng chữ cái từ ảnh
 - P : phần trăm chữ nhận dạng đúng
 - E : ảnh số của chữ và chữ tương ứng
- Dịch máy
 - T : dịch một câu tiếng Anh sang tiếng Việt
 - P : độ đo dịch máy (ví dụ số câu đúng, số mệnh đề đúng,...)
 - E : cặp câu tiếng Anh và tiếng Việt tương ứng

- công việc T
- tiêu chí P
- kinh nghiệm E

Vấn đề cần quan tâm (1/2)

- **Kinh nghiệm cụ thể như thế nào?**
 - Kinh nghiệm **trực tiếp** và **gián tiếp**
 - Trực tiếp: trạng thái cụ thể + nước đi đúng tương ứng
 - Gián tiếp: toàn bộ ván cờ và kết quả
 - **Có giám sát** (hướng dẫn) và **không giám sát**
 - Có giám sát
 - Không giám sát
 - Bán giám sát
- Cần phải **học cái gì? Biểu diễn kiến thức** học được thế nào?
 - Tri thức cần học được biểu diễn như một **hàm đích**, cần lựa chọn hàm đích cụ thể
 - Ví dụ đánh cờ:
 - Chọn_nước đi: *trạng thái* → *nước đi*
 - Điểm_số: *trạng thái* → *điểm số*

Vấn đề cần quan tâm (2/2)

- Sử dụng thuật toán gì để học?
 - Sử dụng hàm
 - VD: $điểm_số = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \dots$
 - Sử dụng các luật
 - Sử dụng mạng nơ ron
 - Sử dụng cây quyết định
 - Sử dụng các mô hình xác suất
 - ...



Một số khái niệm

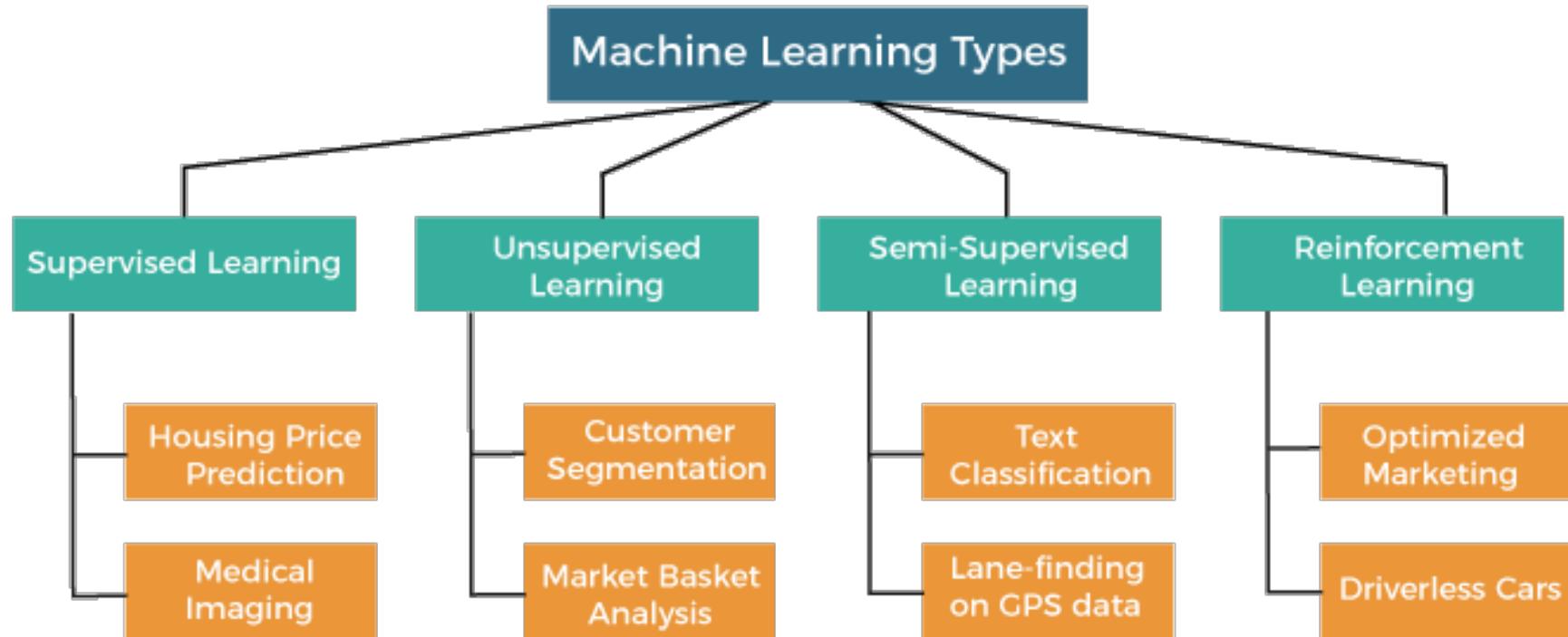
- **Mẫu**, hay ví dụ (samples): là đối tượng cần xử lý (ví dụ phân loại)
 - Ví dụ: khi lọc thư rác thì mỗi thư là một mẫu
- Mẫu thường được mô tả bằng tập thuộc tính hay **đặc trưng** (features)
 - Ví dụ: trong chuẩn đoán bệnh, thuộc tính là triệu chứng của người bệnh, và các tham số khác như chiều cao, cân nặng, ...
- **Nhãn** phân loại (label): thể hiện loại của đối tượng mà ta cần dự đoán
 - Ví dụ: nhãn phân loại thư rác có thể là “rác” hoặc “bình thường”

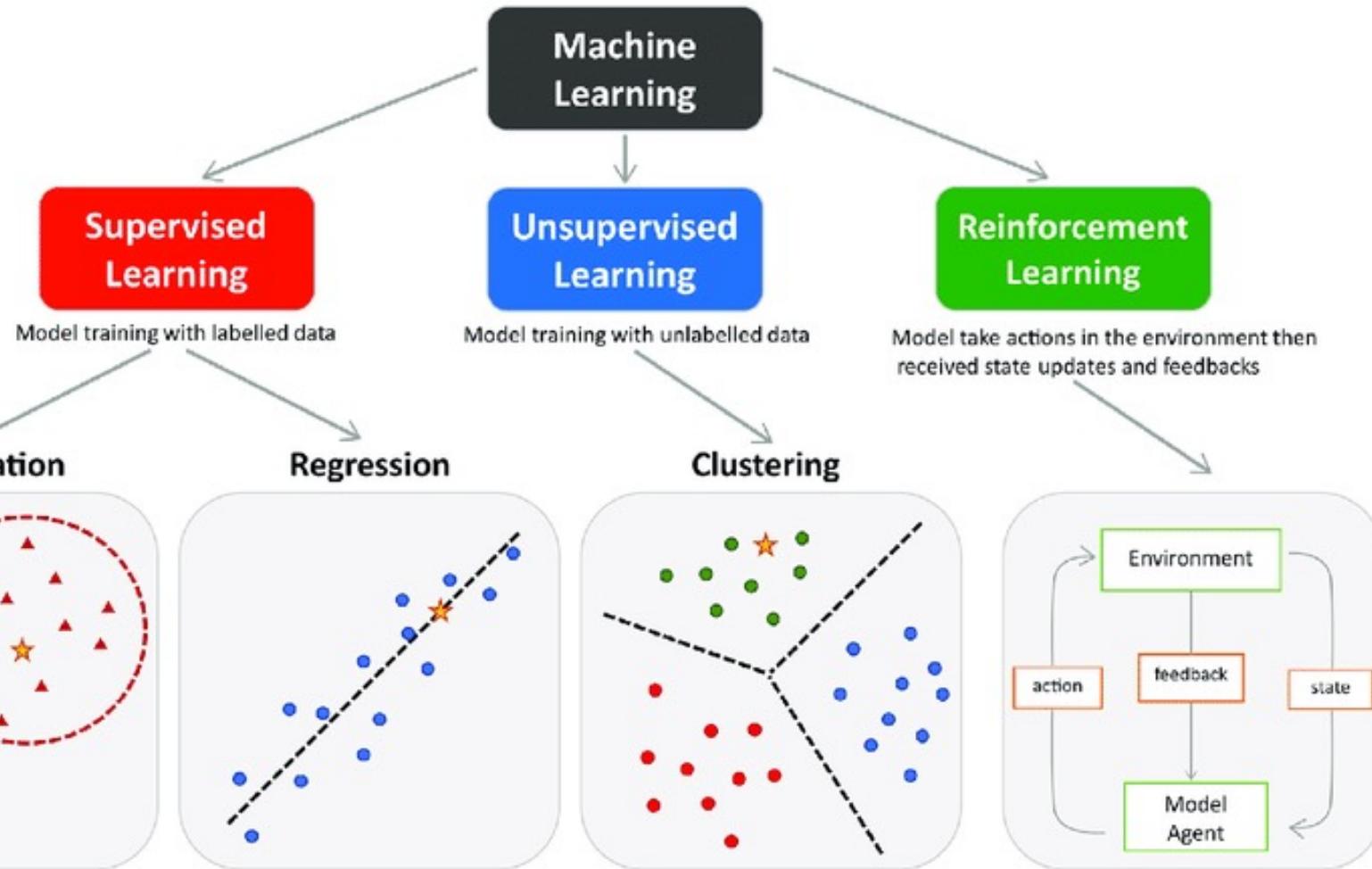


Một số dạng học máy phổ biến

- **Học có giám sát** (supervised learning)
 - Phân lớp (classification)
 - Hồi quy (regression)
- **Học không giám sát** (unsupervised learning)
 - Học luật kết hợp (association)
 - Phân cụm (clustering)
- **Học bán giám sát** (semi-supervised learning)
- **Học tăng cường** (reinforcement learning)

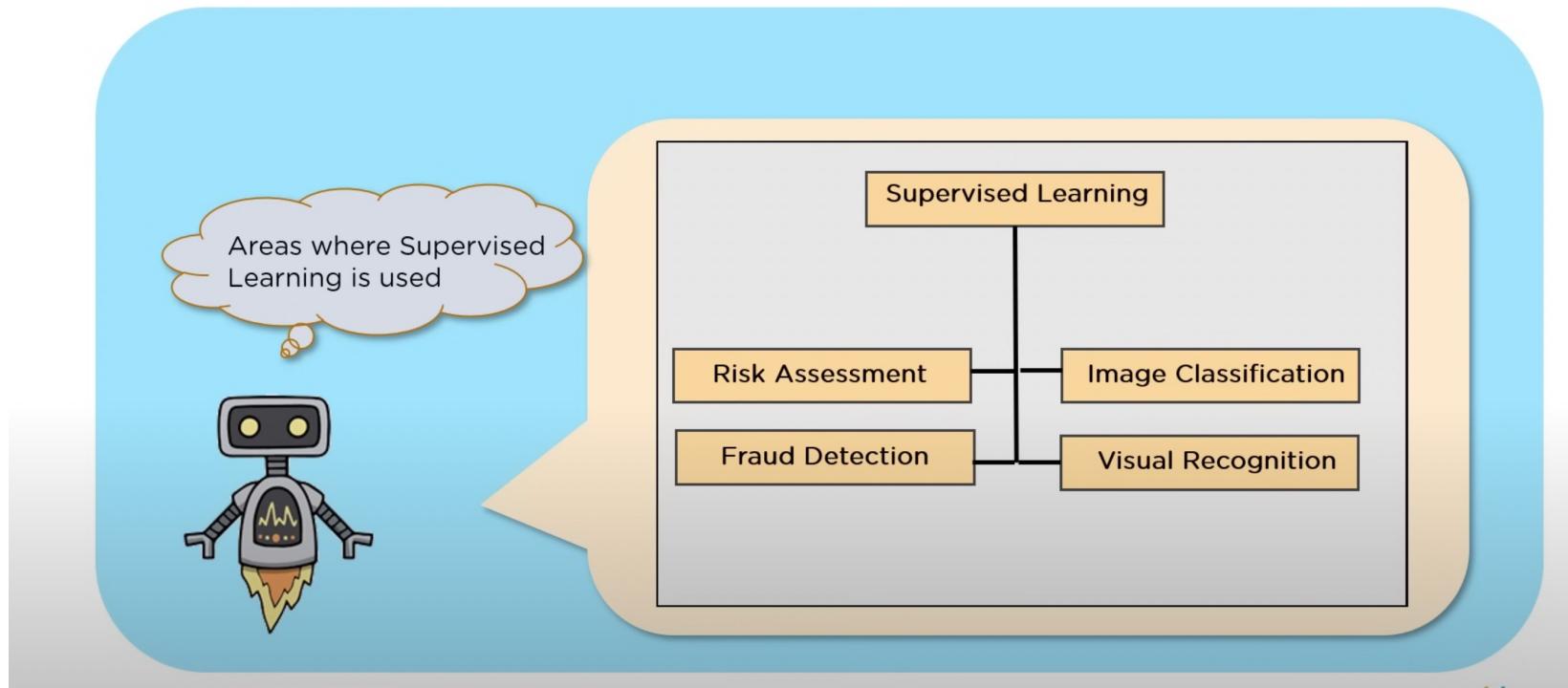
Một số dạng học máy phổ biến



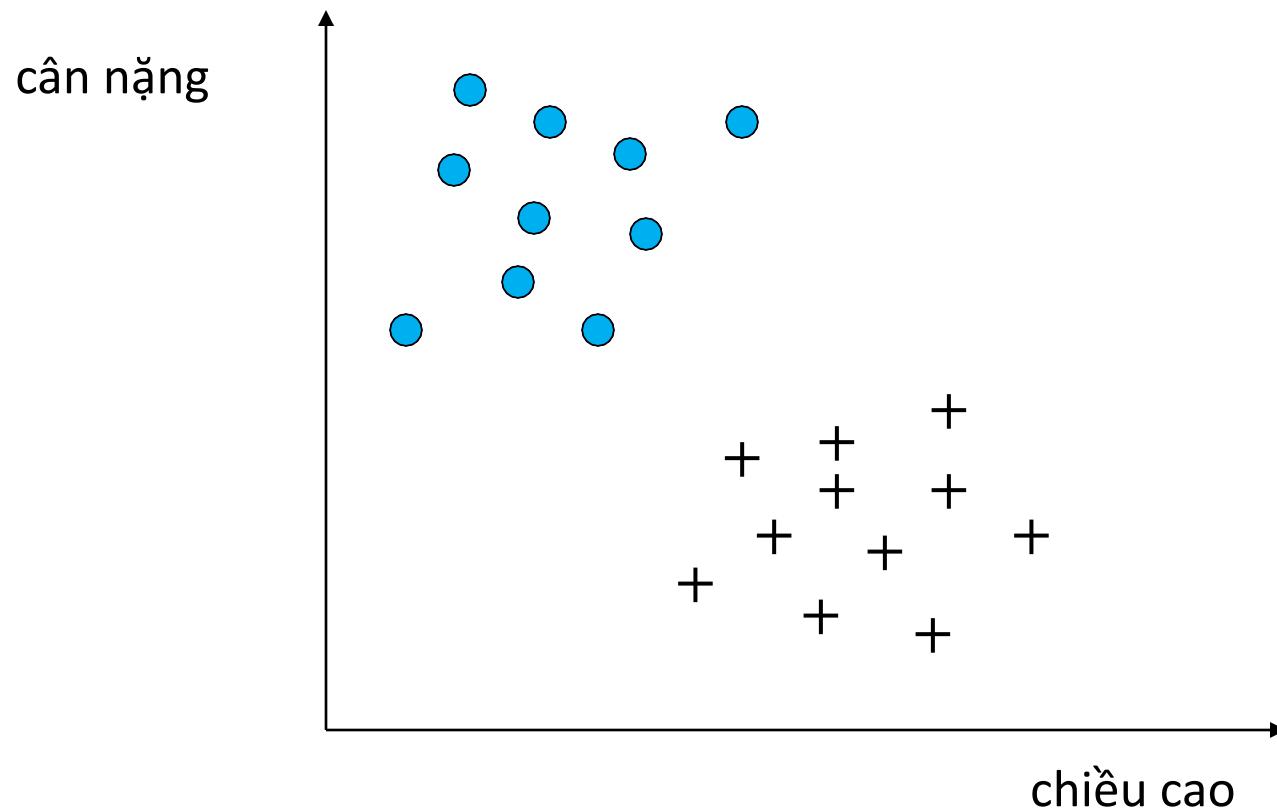


Học có giám sát (supervised learning)

Applications of Supervised Learning

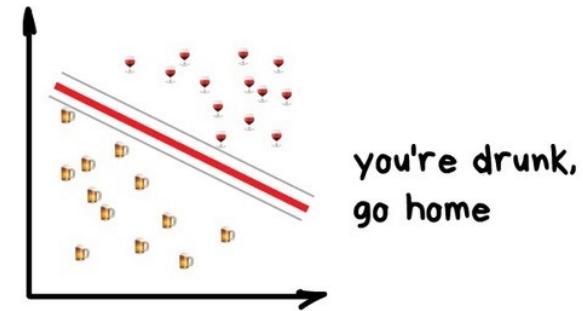
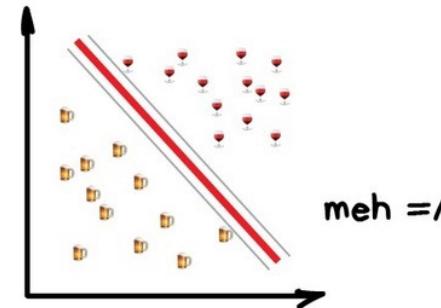
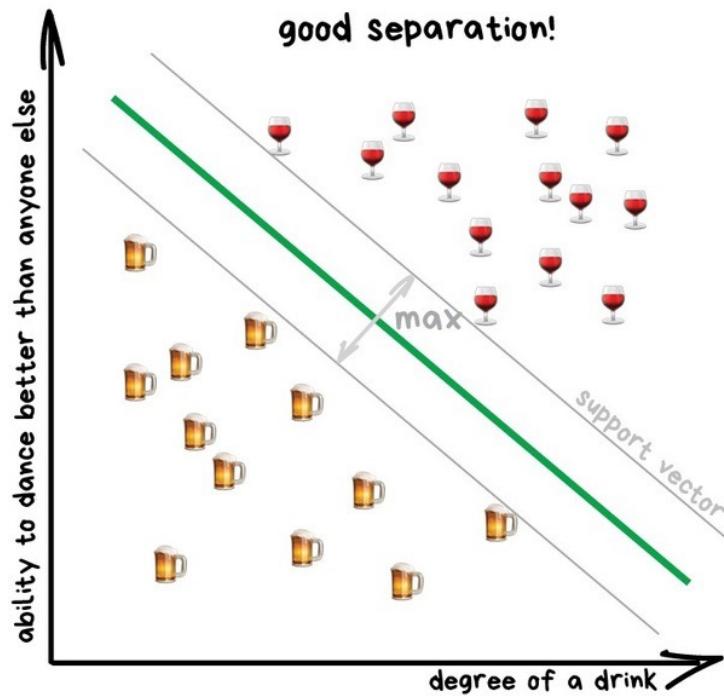


Phân lớp



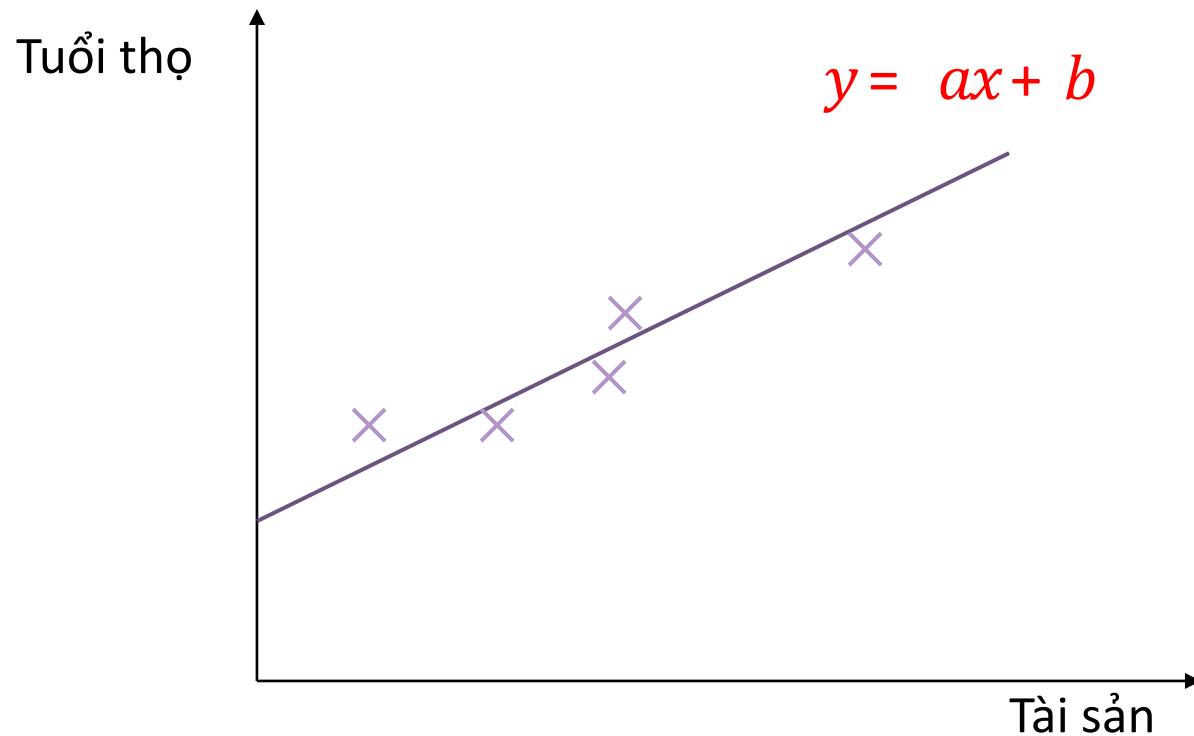
Bài toán phân loại rượu bia

SEPARATE TYPES OF ALCOHOL



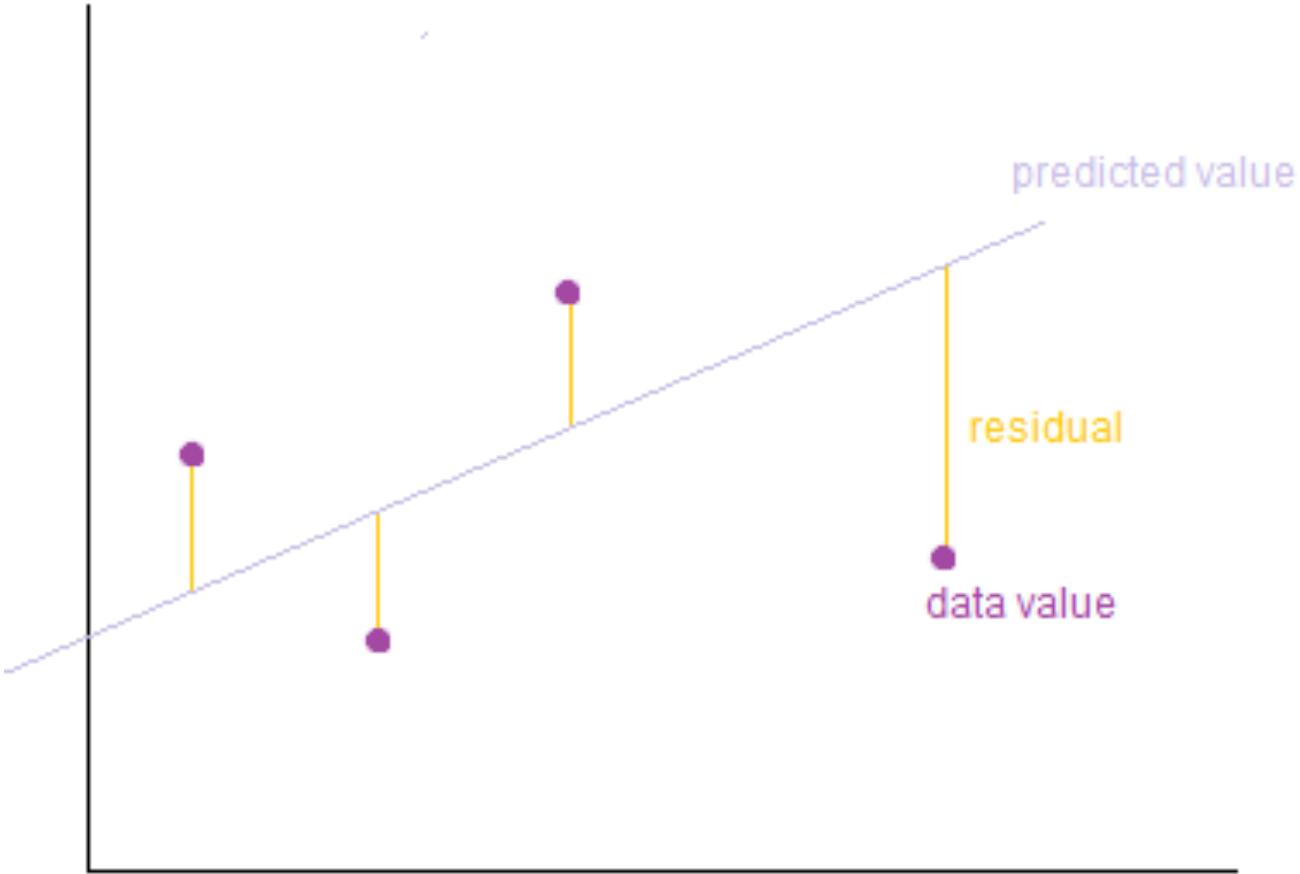
SUPPORT VECTOR MACHINE

Hồi quy (regression)

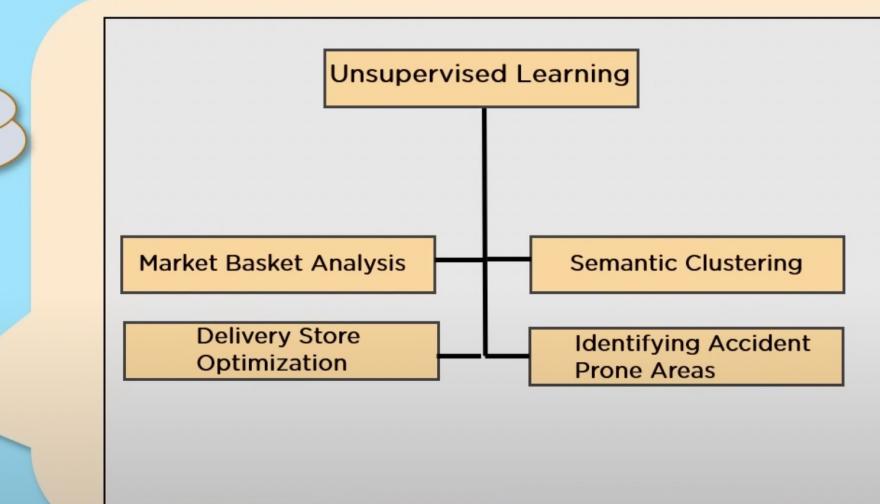
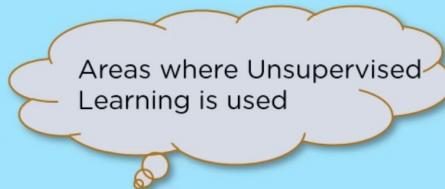


Ứng dụng: dự đoán giá cả, lái xe,...

Sai số (*residuals*)



Applications of Unsupervised Learning





Học luật kết hợp

□ Ví dụ

- Phân tích giao dịch, mua bán (hóa đơn mua hàng)

⁇ $P(Y|X)$

- Xác suất người mua hàng X còn mua hàng Y

□ Ví dụ luật kết hợp

- Người mua bánh mì thường mua bơ
- Người mua lạc rang thường mua bia



Phân cụm

- Nhóm những trường hợp tương tự với nhau
- Không có giá trị đầu ra
- Ứng dụng
 - Phân cụm khách hàng, phân cụm sinh viên
 - Phân đoạn ảnh
 - Thiết kế vi mạch



Học tăng cường

- Kinh nghiệm không được cho trực tiếp dưới dạng đầu vào / đầu ra
- Hệ thống nhận được một giá trị thưởng (reward) là kết quả cho một chuỗi hành động nào đó
- Thuật toán cần học cách hành động để cực đại hóa giá trị thưởng
- Ví dụ: học đánh cờ
 - Hệ thống không được chỉ cho nước đi nào là hợp lý cho từng tình huống cụ thể
 - Chỉ biết kết quả thắng thua sau một chuỗi nước đi



Nội dung

-
- Giới thiệu
 - Học cây quyết định (decision tree learning)
 - Phân loại Bayes đơn giản
 - Học dựa trên ví dụ

Dữ liệu huấn luyện

Ngày	Trời	Nhiệt độ	Độ ẩm	Gió	Chơi tennis
D1	nắng	nóng	cao	yếu	không
D2	nắng	nóng	cao	mạnh	không
D3	u ám	nóng	cao	yếu	có
D4	mưa	trung bình	cao	yếu	có
D5	mưa	lạnh	bình thường	yếu	có
D6	mưa	lạnh	bình thường	mạnh	không
D7	u ám	lạnh	bình thường	mạnh	có
D8	nắng	trung bình	cao	yếu	không
D9	nắng	lạnh	bình thường	yếu	có
D10	mưa	trung bình	bình thường	yếu	có
D11	nắng	trung bình	bình thường	mạnh	có
D12	u ám	trung bình	cao	mạnh	có
D13	u ám	nóng	bình thường	yếu	có
D14	mưa	trung bình	cao	mạnh	không

thuộc tính

nhãn

Ngày	Trời	Nhiệt độ	Độ ẩm	Gió	Chơi tennis
D1	nắng	nóng	cao	yếu	không
D2	nắng	nóng	cao	mạnh	không
D3	u ám	nóng	cao	yếu	có
D4	mưa	trung bình	cao	yếu	có
D5	mưa	lạnh	bình thường	yếu	có
D6	mưa	lạnh	bình thường	mạnh	không
D7	u ám	lạnh	bình thường	mạnh	có
D8	nắng	trung bình	cao	yếu	không
D9	nắng	lạnh	bình thường	yếu	có
D10	mưa	trung bình	bình thường	yếu	có
D11	nắng	trung bình	bình thường	mạnh	có
D12	u ám	trung bình	cao	mạnh	có
D13	u ám	nóng	bình thường	yếu	có
D14	mưa	trung bình	cao	mạnh	không

mẫu



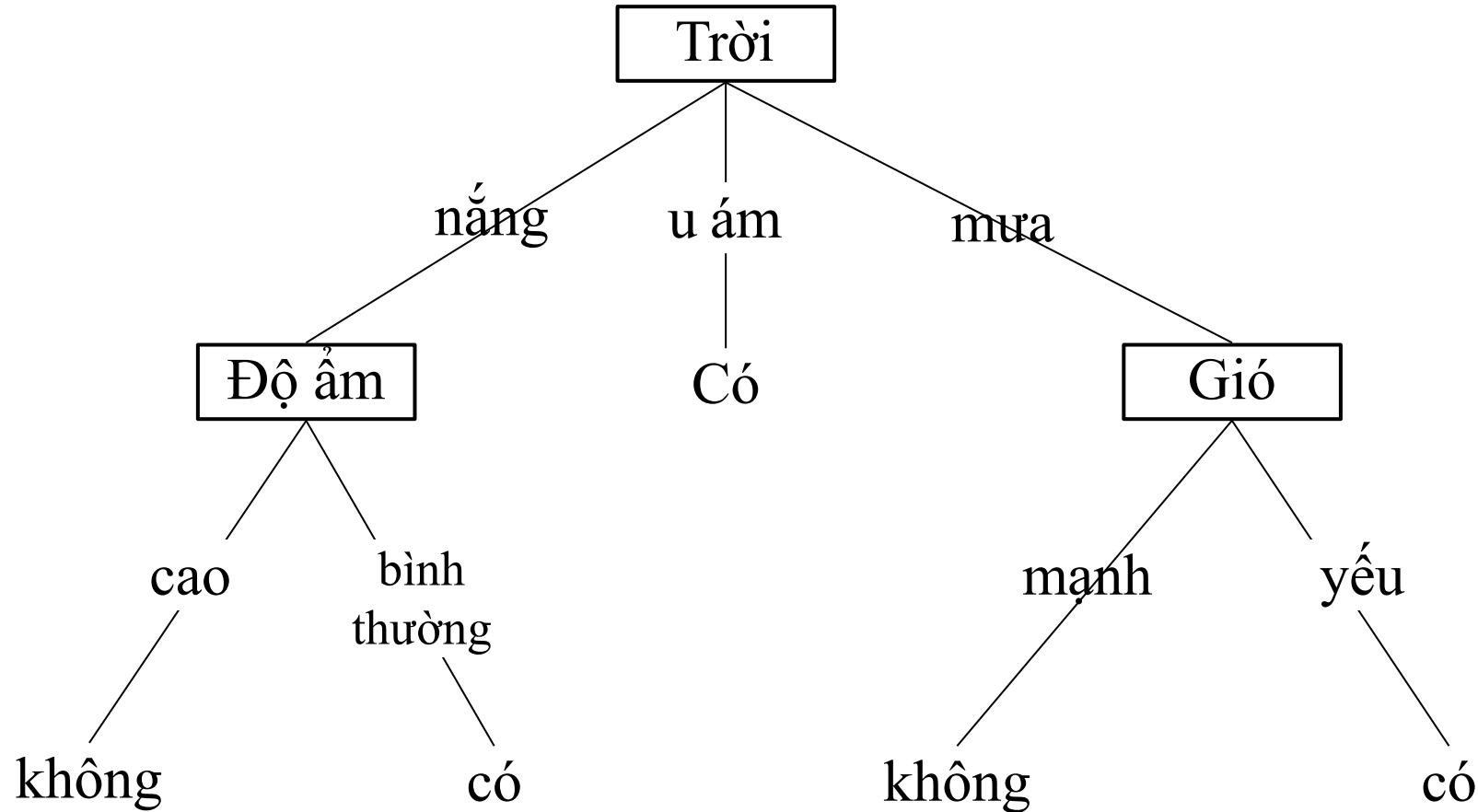
Dữ liệu

- n mẫu huấn luyện, mỗi mẫu là một cặp $\langle x, y \rangle$
 - x là vector các thuộc tính
 - y là nhãn phân loại, $y \in C$ (tập các nhãn)
- Ví dụ mẫu D4

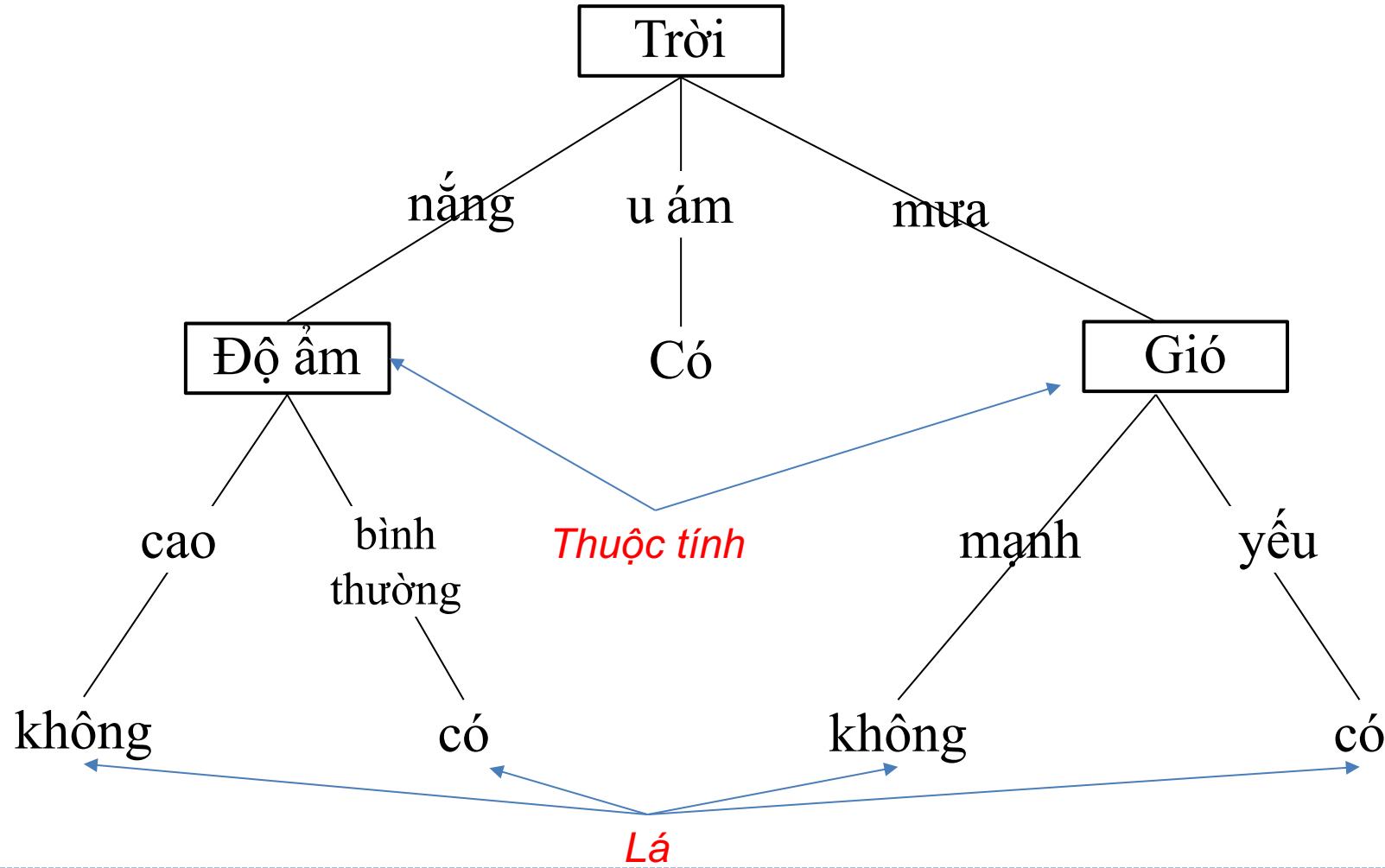
$x = (\text{mưa}, \text{trung bình}, \text{cao}, \text{yếu})$

$y = \text{có}$

Ví dụ cây quyết định



Ví dụ cây quyết định



Cây quyết định là gì?

□ Là mô hình phân loại có dạng cây

- Mỗi nút trung gian (không phải lá) ứng với một phép kiểm tra thuộc tính, mỗi nhánh của nút ứng với một giá trị của thuộc tính tại nút đó
- Mỗi nút lá ứng với một nhãn phân loại

□ Quá trình phân loại thực hiện như sau

- Mẫu phân loại đi từ gốc cây xuống dưới
- Tại mỗi nút trung gian, thuộc tính tương ứng với nút được kiểm tra, tùy giá trị thuộc tính, mẫu được chuyển xuống nhánh tương ứng
- Khi tới nút lá, mẫu được nhận nhãn phân loại của nút

Biểu diễn dưới dạng quy tắc

- Cây quyết định có thể biểu diễn tương đương dưới dạng các quy tắc logic
- Mỗi cây là tuyển của các quy tắc, mỗi quy tắc bao gồm các phép hội
- Ví dụ

(Trời = nắng \wedge Độ ẩm = bình_thường)

\vee (Trời = u_ám)

\vee (Trời = mưa \wedge Gió = yếu)



Học cây quyết định

- Cây quyết định được học (xây dựng) từ dữ liệu huấn luyện
- Với mỗi bộ dữ liệu có thể xây dựng nhiều cây quyết định
 - Chọn cây nào?
- Quá trình học là quá trình tìm kiếm cây quyết định phù hợp với dữ liệu huấn luyện
 - Cho phép phân loại đúng dữ liệu huấn luyện

Thuật toán ID3

- Xây dựng lần lượt các nút của cây bắt đầu từ gốc
- Thuật toán
 - **Khởi đầu:** nút hiện thời là nút gốc chứa toàn bộ tập dữ liệu huấn luyện
 - Tại nút hiện thời *n*, lựa chọn thuộc tính
 - Chưa được sử dụng ở nút tổ tiên
 - Cho phép phân chia tập dữ liệu hiện thời thành các tập con **một cách tốt nhất**
 - Với mỗi giá trị thuộc tính được chọn thêm một nút con bên dưới
 - Chia các ví dụ ở nút hiện thời về các nút con theo giá trị thuộc tính được chọn
 - **Lặp** (đệ quy) cho tới khi
 - Tất cả các thuộc tính đã được sử dụng ở các nút phía trên, hoặc
 - Tất cả ví dụ tại nút hiện thời có cùng nhãn phân loại
 - Nhãn của nút được lấy theo đa số nhãn của ví dụ tại nút hiện thời

Lựa chọn thuộc tính tại mỗi nút thế nào?



Tiêu chuẩn chọn thuộc tính của ID3

- Tại mỗi nút n
 - Tập (con) dữ liệu ứng với nút đó
 - Cần lựa chọn thuộc tính cho phép phân chia tập dữ liệu tốt nhất
- Tiêu chuẩn:
 - Dữ liệu sau khi phân chia càng đồng nhất càng tốt
 - Đo bằng độ tăng thông tin (Information Gain - IG)
 - **Chọn thuộc tính có độ tăng thông tin lớn nhất**
 - IG dựa trên entropy của tập (con) dữ liệu

Entropy

- ▶ Trường hợp tập dữ liệu S có 2 loại nhãn: đúng (+) hoặc sai (-)

$$\text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

p_+ : % số mẫu đúng, p_- : % số mẫu sai

- ▶ Trường hợp tổng quát: có C loại nhãn

$$\text{Entropy}(S) = \sum_{i=1}^C -p_i \log_2 p_i$$

p_i : % ví dụ của S thuộc loại i

- ▶ Ví dụ

$$\begin{aligned}\text{Entropy}([9^+, 5^-]) &= -(9/14) \log_2(9/14) - (5/14) \log_2(5/14) \\ &= 0.94\end{aligned}$$

$\text{Entropy}(S)$ hay viết là $H(S)$

Độ tăng thông tin IG

Với tập (con) mẫu S và thuộc tính A

$$IG(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Trong đó:

$values(A)$: tập các giá trị của A

S_v là tập con của S bao gồm các mẫu có giá trị của A bằng v

$|S|$ số phần tử của S



Ví dụ: tính IG(S, Trời)

Ta tính toán entropy của thuộc tính "Chơi tennis" như sau:

- Không chơi tennis: 5/14

- Chơi tennis: 9/14

$$H(S) = -(5/14)*\log_2(5/14) - (9/14)*\log_2(9/14) = 0.940$$

Tiếp theo, ta cần tính toán entropy cho từng giá trị của thuộc tính "Trời" (Nắng, U ám, Mưa):

- Nắng:

- Chơi tennis: 2/5

- Không chơi tennis: 3/5

- $H(S \text{ nắng}) = -(2/5)*\log_2(2/5) - (3/5)*\log_2(3/5) = 0.971$

- Mưa:

- Không chơi tennis: 3/5

- Chơi tennis: 2/5

- $H(S \text{ mưa}) = -(3/5)*\log_2(3/5) - (2/5)*\log_2(2/5) = 0.971$

- U ám:

- Không chơi tennis: 0

- Chơi tennis: 4/4

- $H(S \text{ u ám}) = 0$

$$IG(S, \text{Trời}) = 0.940 - [(4/14)*0 + (5/14)*0.971 + (5/14)*0.971] = 0.247$$



Ví dụ tính $IG(S, \text{Gió})$

► Tính $IG(S, \text{Gió})$

$values(\text{Gió}) = \{\text{yếu}, \text{mạnh}\}$

$$S = [9+, 5-], H(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{\text{yếu}} = [6+, 2-], H(S_{\text{yếu}}) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0.811$$

$$S_{\text{mạnh}} = [3+, 3-], H(S_{\text{mạnh}}) = -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} = 1$$

$$\begin{aligned} IG(S, \text{Gió}) &= H(S) - \frac{8}{14} H(S_{\text{yếu}}) - \frac{6}{14} H(S_{\text{mạnh}}) \\ &= 0.94 - \frac{8}{14} 0.811 - \frac{6}{14} 1 \\ &= 0.048 \end{aligned}$$

Ví dụ tính IG

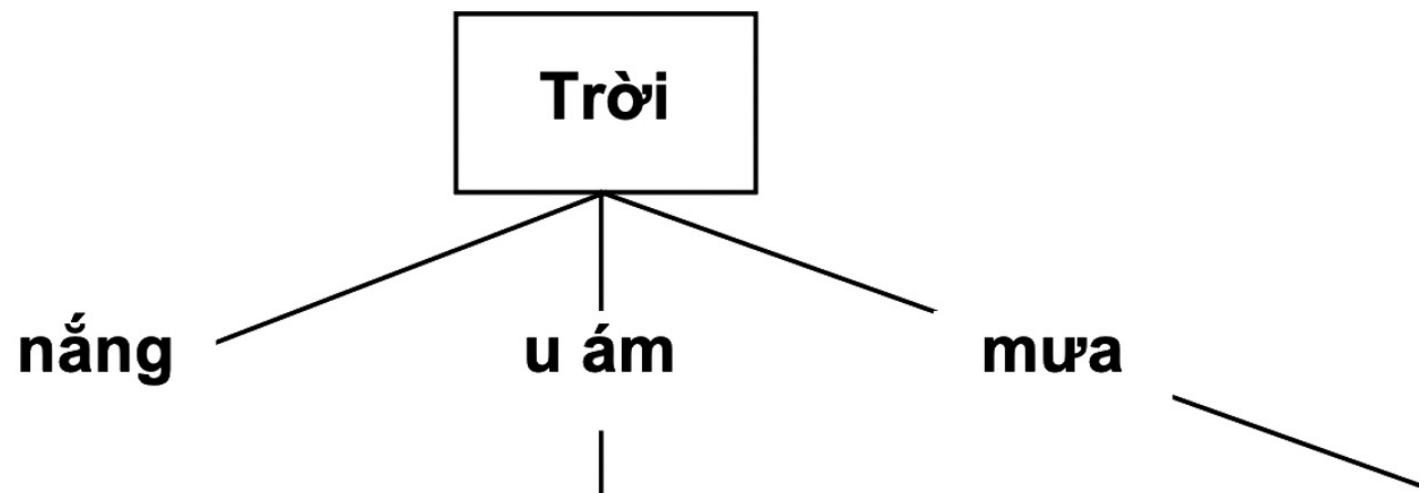
Tính tương tự với ba thuộc tính còn lại:

$$IG(S, \text{Trời}) = 0.247$$

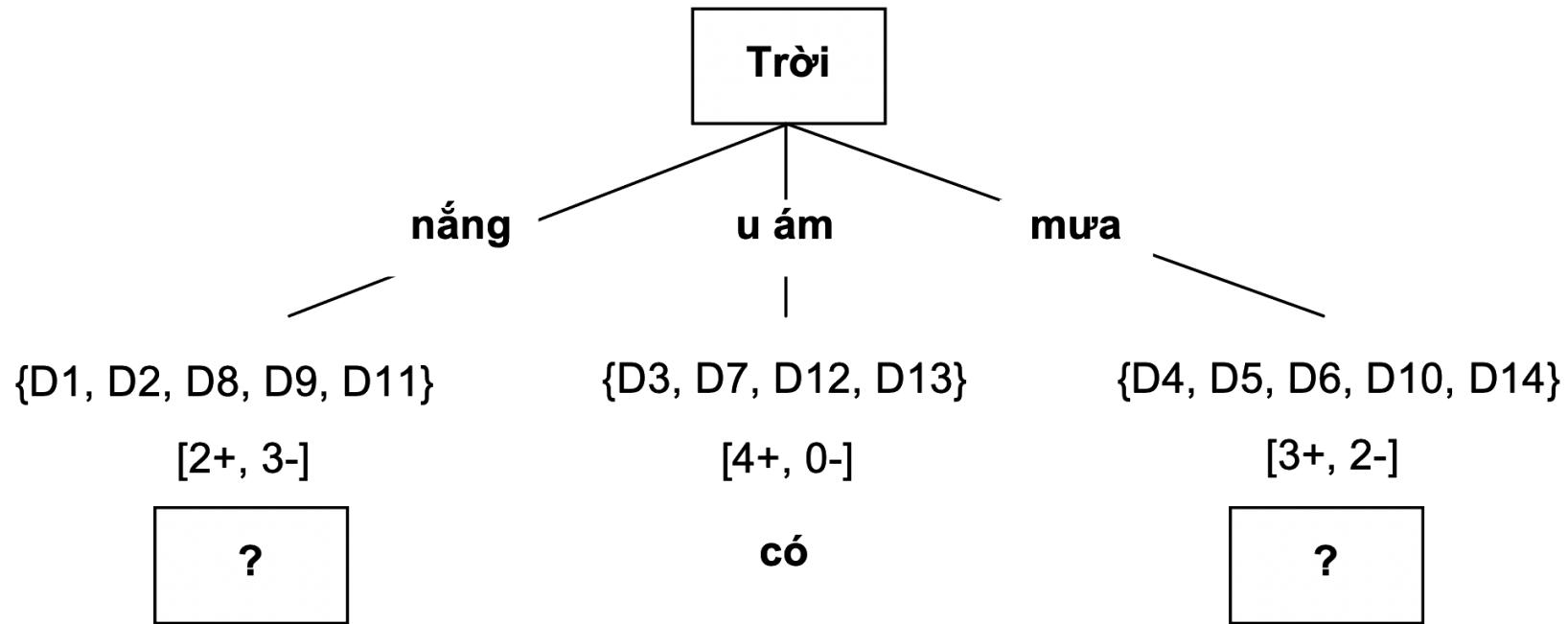
$$IG(S, \text{Độ ẩm}) = 0.151$$

$$IG(S, \text{Nhiệt độ}) = 0.029$$

$$IG(S, \text{Gió}) = 0.048$$



{D1, D2, ..., D14}
[9+, 5-]



Các đặc điểm của ID3

- ID3 là thuật toán tìm kiếm cây quyết định phù hợp với dữ liệu huấn luyện
- Tìm kiếm theo kiểu tham lam, bắt đầu từ cây rỗng
- Hàm đánh giá là độ tăng thông tin
- ID3 có khuynh hướng (bias) lựa chọn cây đơn giản
 - Ít nút
 - Các thuộc tính có độ tăng thông tin lớn nằm gần gốc

Training error và Test error (1/2)

▶ Training error (lỗi huấn luyện)

- Là lỗi đo được trên tập dữ liệu huấn luyện
- Thường đo bằng sự sai khác giữa giá trị tính toán của mô hình và giá trị thực của dữ liệu huấn luyện
- Trong quá trình học ta cố gắng làm giảm tới mức tối thiểu lỗi huấn luyện

▶ Test error (lỗi kiểm tra)

- Là lỗi đo được trên tập dữ liệu kiểm tra
- Là cái ta thực sự quan tâm

Làm sao ta có thể tác động tới hiệu quả của mô hình trên tập dữ liệu kiểm tra khi ta chỉ quan sát được tập dữ liệu huấn luyện?

Training error và Test error (2/2)

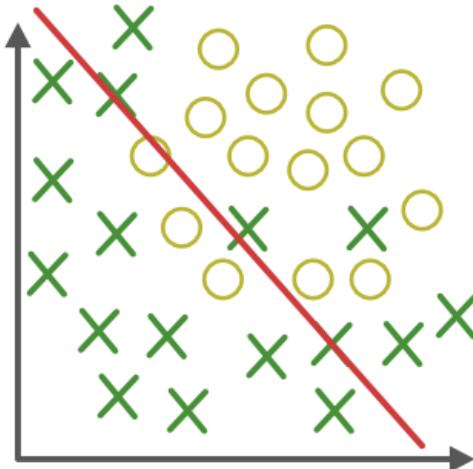
► i.i.d assumptions (independent, identically distributed)

- Giả thiết rằng các mẫu dữ liệu (cả ở tập huấn luyện và tập kiểm tra) là **độc lập**, và các tập dữ liệu huấn luyện và kiểm tra có **cùng phân phối**
- Nếu ta cố định các tham số của mô hình thì lỗi huấn luyện và lỗi kiểm tra sẽ bằng nhau
 - Trong quá trình huấn luyện tham số được tối ưu theo lỗi huấn luyện, do đó **lỗi kiểm tra thường lớn hơn lỗi huấn luyện**

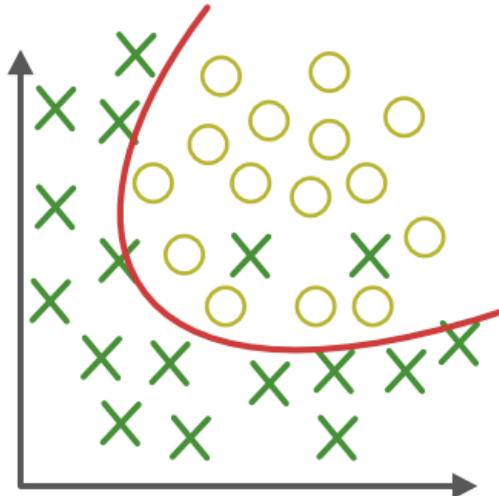
► Hai yếu tố đánh giá độ tốt của một thuật toán học máy

- Khả năng giảm thiểu lỗi huấn luyện
- Khả năng giảm thiểu khoảng cách giữa lỗi huấn luyện và lỗi kiểm tra

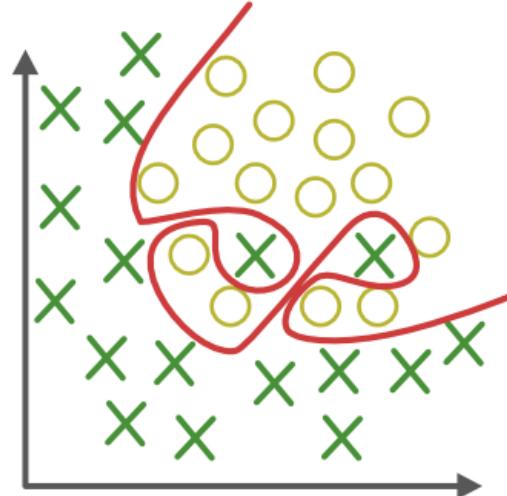
Underfitting và Overfitting



Under-fitting
(too simple to
explain the variance)



Appropriate-fitting



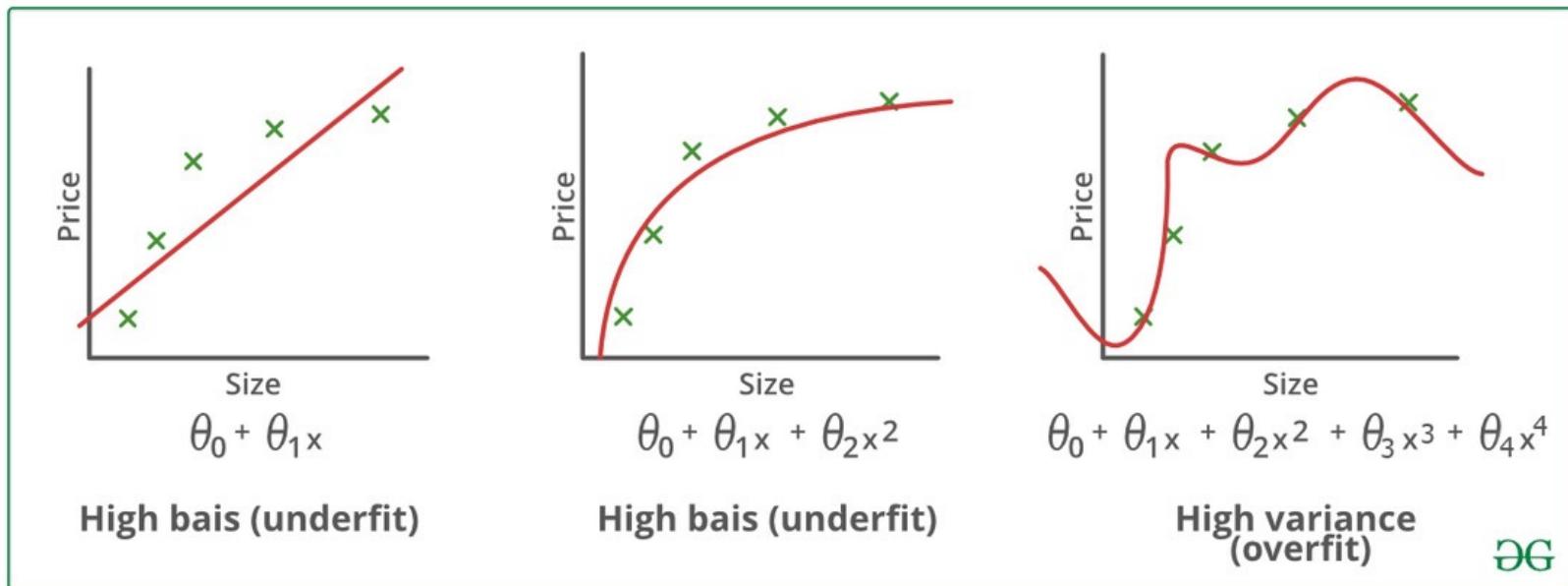
Over-fitting
(forcefitting--too
good to be true)
DG

Underfitting: dưới vừa; Overfitting: quá vừa
Capacity: Khả năng của mô hình

Overfitting

Lý do cho việc Overfitting như sau:

- Phương sai cao(*variance*) và sai lệch thấp (*bias*)
- Mô hình quá phức tạp
- Kích thước dữ liệu huấn luyện





Chống quá vừa bằng cách tỉa cây

- Chia dữ liệu thành hai phần
 - Huấn luyện
 - Kiểm tra
- Tạo cây đủ lớn trên dữ liệu huấn luyện
- Tính độ chính xác của cây trên tập kiểm tra
- Loại bỏ cây con sao cho kết quả trên dữ liệu kiểm tra được cải thiện nhất
- Lặp lại cho đến khi không còn cải thiện được kết quả nữa



Chống quá vừa dữ liệu bằng cách tỉa luật (C4.5)

- Biến đổi cây thành các luật
- Tỉa mỗi luật độc lập với các luật khác
 - Bỏ một số phần trong vế trái của luật
- Sắp xếp các luật sau khi tỉa theo mức độ chính xác của luật

Sử dụng thuộc tính có giá trị liên tục

- Tạo ra những thuộc tính rời rạc mới
- Ví dụ, với thuộc tính liên lục A , tạo ra thuộc tính rời rạc Ac như sau
 - o $Ac = true$ nếu $A > c$
 - o $Ac = false$ nếu $A \leq c$
- Xác định ngưỡng c thế nào?
 - o Thường chọn sao cho Ac đem lại độ tăng thông tin lớn nhất
- Có thể chia thành nhiều khoảng với nhiều ngưỡng

Các độ đo khác

- ▶ Độ đo Information Gain (IG) ưu tiên thuộc tính có nhiều giá trị, ví dụ, thuộc tính ngày sẽ có độ tăng thông tin cao nhất
- ▶ Thông tin chia

$$SplitInformation(S, A) = - \sum_{i=1}^c \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

- ▶ Tiêu chuẩn đánh giá thuộc tính

$$GainRatio = \frac{InformationGain(S, A)}{SplitInformation(S, A)}$$



Nội dung

- Giới thiệu
- Học cây quyết định
- Phân loại Bayes đơn giản (Naïve Bayes classification)
- Học dựa trên ví dụ

Phương pháp phân loại Bayes (1/2)

- ▶ Trong giai đoạn huấn luyện ta có một tập mẫu, mỗi mẫu được cho bởi cặp $\langle x_i, y_i \rangle$, trong đó
 - x_i là vector đặc trưng (thuộc tính)
 - y_i là nhãn phân loại, $y_i \in C$ (C là tập các nhãn)
- ▶ Sau khi huấn luyện xong, bộ phân loại cần dự đoán nhãn y cho mẫu mới $x = \langle x_1, x_2, \dots, x_n \rangle$

$$y = \operatorname{argmax}_{c_j \in C} P(c_j | x_1, x_2, \dots, x_n)$$

- ▶ Sử dụng quy tắc Bayes

$$\begin{aligned} y &= \operatorname{argmax}_{c_j \in C} \frac{P(x_1, x_2, \dots, x_n | c_j) P(c_j)}{P(x_1, x_2, \dots, x_n)} \\ &= \operatorname{argmax}_{c_j \in C} P(x_1, x_2, \dots, x_n | c_j) P(c_j) \end{aligned}$$

Phương pháp phân loại Bayes (2/2)

Tần xuất quan sát thấy nhãn c_j trên tập dữ liệu D:

$$\frac{\text{count}(c_j)}{|D|}$$

$$y = \operatorname{argmax}_{c_j \in C} P(x_1, x_2, \dots, x_n | c_j) P(c_j)$$

Sử dụng giả thiết về tính độc lập (**Đơn giản!!!**)

$$P(x_1, x_2, \dots, x_n | c_j) = P(x_1 | c_j) P(x_2 | c_j) \dots P(x_n | c_j)$$

Số lần xuất hiện x_i cùng với c_j chia
cho số lần xuất hiện c_j : $\frac{\text{count}(x_i, c_j)}{\text{count}(c_j)}$



Huấn luyện

Quá trình huấn luyện hay học Bayes đơn giản là quá trình tính các xác suất $P(cj)$ và các xác suất điều kiện $P(xi | cj)$

$$P(c_j) = \frac{\text{Số mẫu có nhãn là } c_j}{\text{Tổng số mẫu trong tập huấn luyện}}$$

$$P(x_i | c_j) = \frac{\text{Số mẫu có giá trị thuộc tính } X_i = x_i \text{ và nhãn là } c_j}{\text{Số mẫu có nhãn là } c_j}$$





Ví dụ

- ▶ Xác định nhãn phân loại cho mẫu sau

$<Trời = nắng, Nhiệt độ = trung bình, Độ ẩm = cao, Gió = mạnh >$

$$y = \operatorname{argmax}_{c \in \{có, không\}} P(Trời = nắng|c)P(Nhiệt độ = trung bình|c) \\ P(Độ ẩm = cao|c)P(Gió = mạnh|c)P(c)$$



Ví dụ

Thay số liệu của bài toán vào công thức Bayes đơn giản, ta có:

$$\begin{aligned} c_{NB} &= \operatorname{argmax}_{c_j} P(c_j) \prod_{i \in C} P(x_i | c_j) \\ &= \operatorname{argmax}_{c_j \in \{co, khong\}} P(\text{Trời=nắng} | c_j) P(\text{Nh. đđ=t. bình} | c_j) P(\text{Đđ ảm=cao} | c_j) P(\text{Gió=mạnh} | c_j) P(c_j) \end{aligned}$$

Do c_j có thể nhận hai giá trị, ta cần tính 10 xác suất. Các xác suất $P(\text{có})$ và $P(\text{không})$ được tính bằng tần suất “có” và “không” trên dữ liệu huấn luyện.

$$P(\text{có}) = 9/14 = 0,64$$

$$P(\text{không}) = 5/14 = 0,36$$

Các xác suất điều kiện cũng được tính từ dữ liệu huấn luyện, ví dụ ta có:

$$P(\text{Đđ ảm = cao} | \text{có}) = 3/9 = 0,33$$

$$P(\text{Đđ ảm = cao} | \text{không}) = 4/5 = 0,8$$





Thay các xác suất thành phần vào công thức Bayes đơn giản, ta được:

$$\begin{aligned} P1 &= P(\text{có})P(\text{nắng| có})P(\text{trung bình| có})P(\text{cao| có})P(\text{mạnh| có}) \\ &= 9/14 \cdot 2/9 \cdot 4/9 \cdot 3/9 \cdot 3/9 \\ &= 0.007 \end{aligned}$$

$$\begin{aligned} P2 &= P(\text{không})P(\text{nắng| không})P(\text{trung bình| không})P(\text{cao| không})P(\text{mạnh| không}) \\ &= 5/14 \cdot 3/5 \cdot 2/5 \cdot 4/5 \cdot 3/5 \\ &= 0.041 \end{aligned}$$

$$y = \operatorname{argmax} \{P1, P2\}$$

Vậy mẫu được phân loại là không chơi.





Bài tập 1

Cho tập dữ liệu như bảng bên dưới, trong đó Chiều cao, Dân tộc, Tính tình là các thuộc tính, f là nhãn phân loại.

Số TT	f	Chiều cao	Dân tộc	Tính tình
1	+	Cao	Lào	Vui vẻ
2	+	Thấp	Việt	Vui vẻ
3	-	Cao	Việt	Ôn hòa
4	-	Thấp	Thái	Vui vẻ
5	-	Cao	Thái	Vui vẻ
6	-	Cao	Lào	Ôn hòa

- Xây dựng cây quyết định sử dụng thuật toán ID3.
- Sử dụng phương pháp phân loại Bayes đơn giản tìm nhãn cho mẫu (**Chiều cao = “Thấp”, Dân tộc = “Việt”, Tính tình = “Ôn hòa”**).





Nội dung

-
- Giới thiệu
 - Học cây quyết định
 - Phân loại Bayes đơn giản
 - Học dựa trên ví dụ (Instance based learning)

Nguyên tắc chung

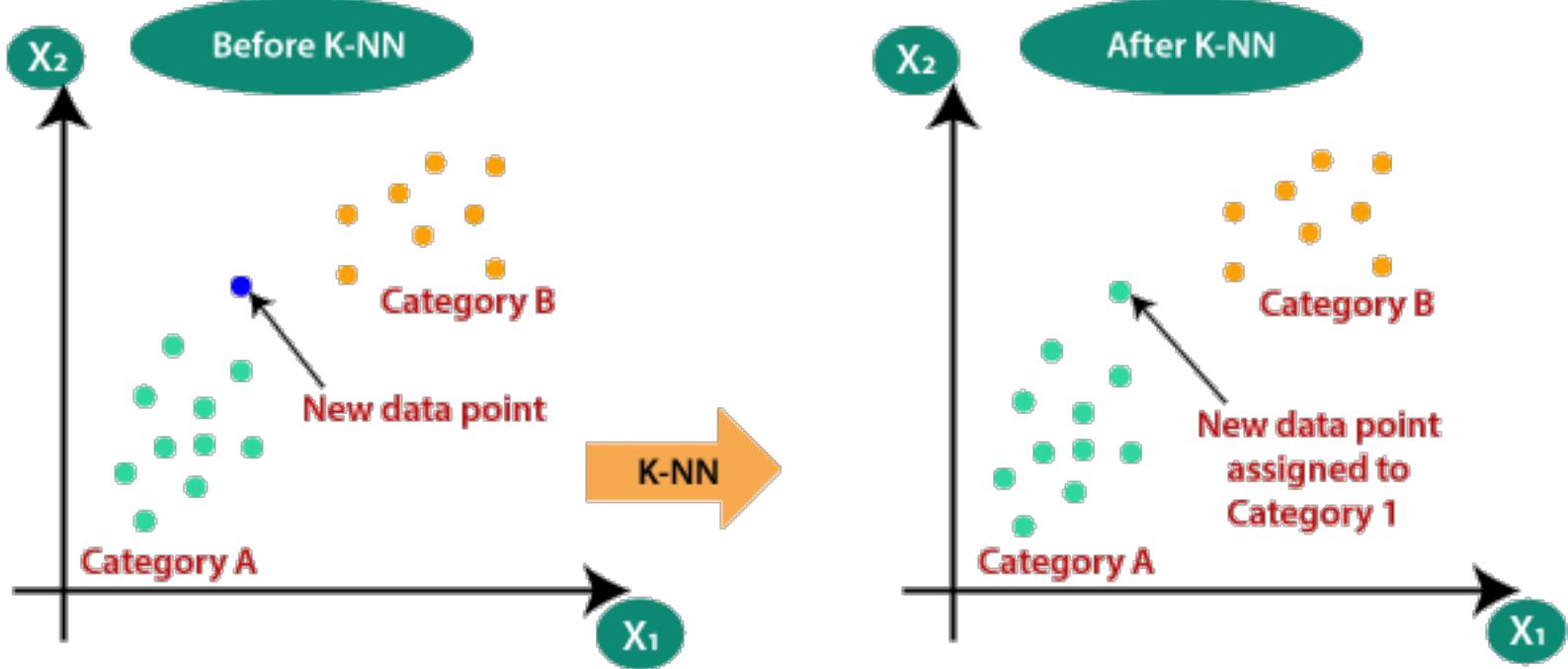
- Không xây dựng mô hình
- Chỉ lưu lại các mẫu huấn luyện
- Xác định nhãn cho mẫu mới dựa trên những mẫu giống mẫu mới nhất
- Gọi là học lười (lazy learning) / học bằng các nhớ (memory-based learning)

Thuật toán k hàng xóm gần nhất

K-láng giềng gần nhất (k -nearest neighbors, viết tắt là k -NN) là phương pháp tiêu biểu nhất của học dựa trên ví dụ.

- Chọn k mẫu giống mẫu cần phân loại nhất, gọi là k hàng xóm
- Gán nhãn phân loại cho mẫu chỉ sử dụng thông tin của k hàng xóm này
 - o Ví dụ lấy theo đa số trong số k hàng xóm
- Chọn hàng xóm thế nào?

K-NN



Tính khoảng cách

- ▶ Giả sử mẫu x có giá trị thuộc tính là $\langle a_1(x), a_2(x), \dots, a_n(x) \rangle$, thuộc tính là số thực
- ▶ Khoảng cách giữa hai mẫu x_i và x_j là khoảng cách Euclidean

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^n (a_l(x_i) - a_l(x_j))^2}$$



Ưu nhược điểm k-NN

Ưu điểm

- 1.Thuật toán đơn giản, dễ dàng triển khai.
- 2.Độ phức tạp tính toán nhỏ.
- 3.Xử lý tốt với tập dữ liệu nhiễu

Nhược điểm

- 1.Với K nhỏ dễ gấp nhiễu dẫn tới kết quả đưa ra không chính xác
- 2.Cần nhiều thời gian để thực hiện do phải tính toán khoảng cách với tất cả các đối tượng trong tập dữ liệu.
- 3.Cần chuyển đổi kiểu dữ liệu thành các yếu tố định tính.



Thuật toán k -NN với trường hợp phân loại

Giai đoạn học (huấn luyện)

Lưu các mẫu huấn luyện có dạng $\langle x, f(x) \rangle$ vào cơ sở dữ liệu

Giai đoạn phân loại

Đầu vào: tham số k

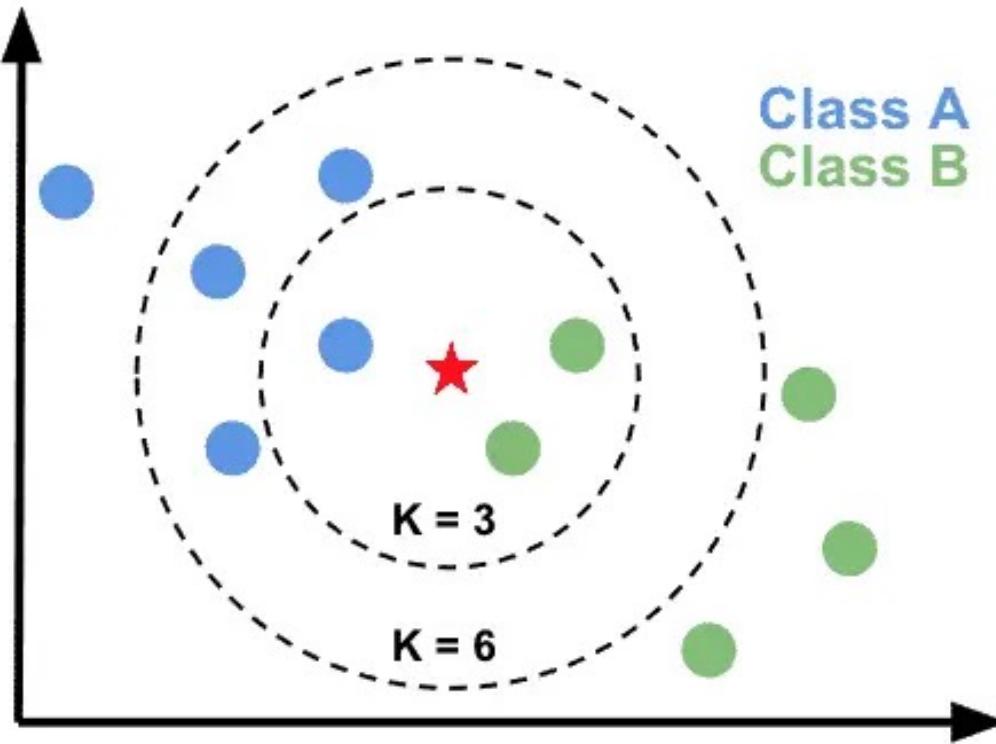
Với mẫu x cần phân loại:

1. Tính khoảng cách $d(x, x_i)$ từ x tới tất cả mẫu x_i trong cơ sở dữ liệu

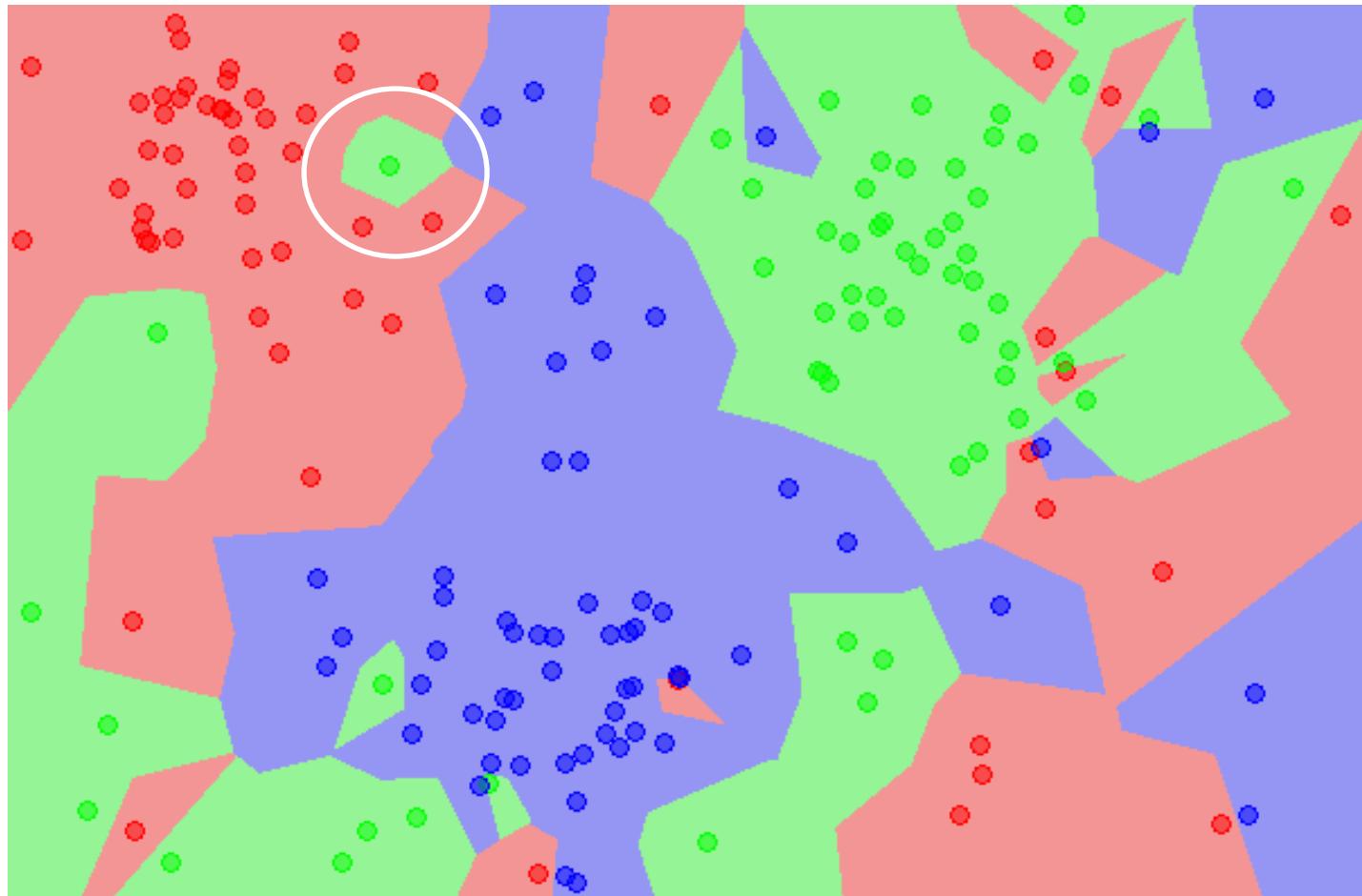
2. Tìm k mẫu có $d(x, x_i)$ nhỏ nhất, giả sử k mẫu đó là x_1, x_2, \dots, x_k .

3. Xác định nhãn phân loại $f'(x)$ là nhãn chiếm đa số trong tập $\{x_1, x_2, \dots, x_k\}$

Thuật toán k -NN với trường hợp phân loại



Thuật toán k -NN với trường hợp hồi quy



Bản đồ của 1-NN

Thuật toán k -NN với trường hợp Hồi quy (regression)

- Mỗi mẫu x có thể nhận phân loại $f(x)$ với $f(x)$ là một số thực.
- Thuật toán k-NN ở bài toán phân loại có thể thay đổi dễ dàng cho bài toán hồi quy này bằng cách thay bước đánh số 3 trong thuật toán.

$$f'(x) = \frac{\sum_{i=1}^k f(x_i)}{k}$$

- Nếu $k = 1$, giá trị hàm $f'(x)$ được chọn bằng giá trị hàm f của mẫu gần nhất.
- k trong khoảng từ 5 đến 10 là phù hợp

