Bài 4:
Phân lớp - Classification

PGS. TS. Đỗ Phúc
Khoa Hệ thống thông tin
Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

Phân lớp và dự báo

Tổng quan

- **Phân lớp là gì? Dự báo là gì?**
- **Giới thiệu cây quyết định**
- **Phân lớp kiểu Bayes**
- **Những phương pháp phân lớp khác**
- **Độ chính xác trong phân lớp**
- **Tóm tắt**

Phân lớp là gì?

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

Một số ứng dụng phân lớp tiêu biểu

- **Tín dụng**
- **Tiếp thị**
- **Chẩn đoán y khoa**
- **Phân tích hiệu quả điều trị**

Dự đoán là gì?

- **Tương tự với phân lớp**
  - o xây dựng một mô hình
  - o sử dụng mô hình để dự đoán cho những giá trị chưa biết
- **Phương thức chủ đạo: Giật lùi**
  - o hồi quy tuyến tính và nhiều cấp
  - o hồi quy không tuyến tính

Phân lớp so với dự báo

- **Phân lớp:**
  - o dự đoán các nhãn phân lớp
  - o phân lớp dữ liệu dựa trên tập huấn luyện và các giá trị trong một thuộc tính phân lớp và dùng nó để xác định lớp cho dữ liệu mới
- **Dự báo:**
  - o xây dựng mô hình các hàm giá trị liên tục
  - o dự đoán những giá trị chưa biết

Phân lớp - tiến trình hai bước

1. **Bước 1:**

   Xây dựng mô hình từ tập huấn luyện

2. **Bước 2:**

   Sử dụng mô hình - kiểm tra tính đúng đắn của mô hình và dùng nó để phân lớp dữ liệu mới

Xây dựng mô hình

Bước 1

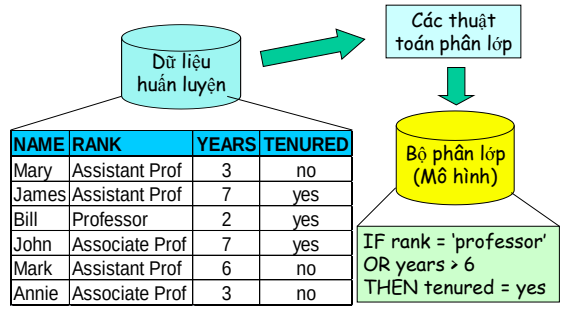
- **Mỗi bộ/mẫu dữ liệu** được phân vào một lớp được xác định trước
- Lớp của một bộ/mẫu dữ liệu được xác định bởi **thuộc tính gán nhãn lớp**
- Tập các bộ/mẫu dữ liệu huấn luyện - **tập huấn luyện** - được dùng để **xây dựng mô hình**
- Mô hình được biểu diễn bởi **các luật phân lớp**, **các cây quyết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

Sử dụng mô hình

Bước 2

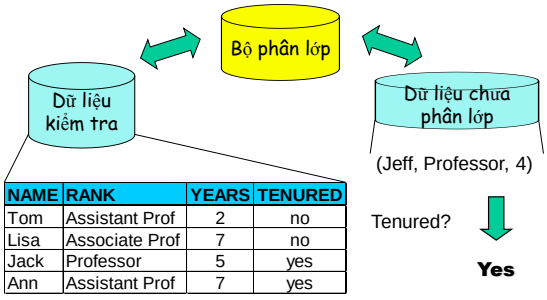
- **Phân lớp cho những đối tượng mới hoặc chưa được phân lớp**
- **Đánh giá độ chính xác của mô hình**
  - lớp biết trước của một mẫu/bộ dữ liệu đem kiểm tra được so sánh với kết quả thu được từ mô hình
  - tỉ lệ chính xác = phần trăm các mẫu/bộ dữ liệu được phân lớp đúng bởi mô hình trong số các lần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

Ví dụ: xây dựng mô hình



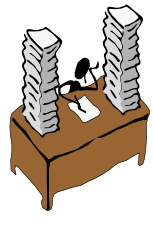
NAME	RANK	YEARS	TENURED
Mary	Assistant Prof	3	no
James	Assistant Prof	7	yes
Bill	Professor	2	yes
John	Associate Prof	7	yes
Mark	Assistant Prof	6	no
Annie	Associate Prof	3	no

Ví dụ: sử dụng mô hình



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Lisa	Associate Prof	7	no
Jack	Professor	5	yes
Ann	Assistant Prof	7	yes

Chuẩn bị dữ liệu

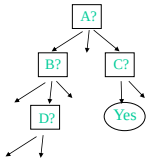


- **Làm sạch dữ liệu**
  - nhiễu
  - các giá trị trống
- **Phân tích sự liên quan** (chọn đặc trưng)
- **Biến đổi dữ liệu**

Đánh giá các phương pháp phân lớp

- **Độ chính xác**
- **Tốc độ**
- **Bền vững**
- **Quy mô lớn (scalability)**
- **Có thể biểu diễn được**
- **Dễ làm**

Qui nạp cây quyết định



Cây quyết định là một cây trong đó

- **nút trong** = một phép kiểm tra trên một thuộc tính
- **nhánh** của cây = đầu ra của một phép kiểm tra
- **nút lá** = nhãn phân lớp hoặc sự phân chia vào lớp

Tạo cây quyết định

Hai giai đoạn tạo cây quyết định:

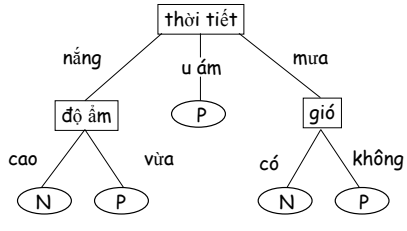
- **xây dựng cây**
  - bắt đầu, tất cả các mẫu huấn luyện đều ở gốc
  - phân chia các mẫu dựa trên các thuộc tính được chọn
  - kiểm tra các thuộc tính được chọn dựa trên một độ đo thống kê hoặc heuristic
- **thu gọn cây**
  - xác định và loại bỏ những nhánh nhiễu hoặc tách khỏi nhóm

Cây quyết định – Ví dụ tiêu biểu: play tennis?

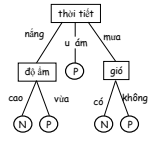
Tập huấn luyện trích từ Quinlan's ID3

STT	Thời tiết	Nhiệt độ	Độ ẩm	Gió	Lớp
1	nắng	nóng	cao	không	N
2	nắng	nóng	cao	không	N
3	u ám	nóng	cao	không	P
4	mưa	ấm áp	cao	không	P
5	mưa	mát	vừa	không	P
6	mưa	mát	vừa	có	N
7	u ám	mát	vừa	có	P
8	nắng	ấm áp	cao	không	N
9	nắng	mát	vừa	không	P
10	mưa	ấm áp	vừa	không	P
11	nắng	ấm áp	vừa	có	P
12	u ám	ấm áp	cao	có	P
13	u ám	nóng	vừa	không	P
14	mưa	ấm áp	cao	có	N

Cây quyết định thu được với ID3 (Quinlan 86)



Rút luật phân lớp từ cây quyết định



IF thời tiết=nắng
AND độ ẩm=vừa
THEN play tennis

- 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 kiện.
- Nút lá giữ quyết định phân lớp dự đoán
- Các luật tạo được dễ hiểu hơn các cây

Các thuật toán trên cây quyết định

- **Thuật toán căn bản**
  - xây dựng một cây đệ quy **phân chia và xác định** đặc tính **từ trên xuống**
  - các thuộc tính được xem là rõ ràng, rời rạc
  - tham lam (có thể có tình trạng cực đại cục bộ)
- **Nhiều dạng khác nhau**: ID3, C4.5, CART, CHAID
  - điểm khác biệt chính: tiêu chuẩn/thuộc tính phân chia, độ đo để chọn lựa

Các độ đo để lựa chọn thuộc tính

- **Độ lợi thông tin (Information gain)**
- **Gini index**
- **χ^2 – số thống kê bằng ngẫu nhiên (contingency table statistic)**
- **G- thống kê (statistic)**

Độ lợi thông tin (1/4)

- Chọn thuộc tính có **độ lợi thông tin lớn nhất** để phân tách.
- Gọi ***P*** , ***N*** là hai lớp và **S** là một tập dữ liệu có ***p*** phần tử thuộc lớp ***P*** và ***n*** phần tử thuộc lớp ***N***
- Khối lượng thông tin cần thiết để quyết định một mẫu tùy ý có thuộc về lớp ***P*** hoặc lớp ***N*** là **I(p.n)** được tính như sau:

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

Độ lợi thông tin (2/4)

- Gọi $\{S_1, S_2, \ldots, S_v\}$ là một phân hoạch của S, khi sử dụng thuộc tính A
- Với mỗi S_i chứa p_i mẫu thuộc lớp P và n_i mẫu thuộc lớp N. Entropy, hay thông tin mong muốn cần thiết để phân lớp các đối tượng trong tất cả các cây con S_i là:

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

- Thông tin có được do phân nhánh trên thuộc tính A là

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

Độ lợi thông tin – Ví dụ (3/4)

- Lớp ***P***: plays_tennis = "yes"
- Lớp ***N***: plays_tennis = "no"
- Thông tin cần thiết để phân lớp một mẫu được cho là:

$$I(p,n) = I(9,5) = 0.940$$

Độ lợi thông tin – **Ví dụ (4/4)**

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

Ta có

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

Do đó $Gain(thoitiet) = I(9,5) - E(thoitiet) = 0.246$

Tương tự

$$Gain(nhietdo) = 0.029$$
$$Gain(doam) = 0.151$$
$$Gain(gio) = 0.048$$

Những tiên chuẩn khác dùng để xây dựng cây quyết

- **Các điều kiện để ngừng phân chia**
  - o tất cả các mẫu thuộc về cùng một lớp
  - o không còn thuộc tính nào nữa để phân chia
  - o không còn mẫu nào để phân lớp
- **Chiến lược rẽ nhánh**
  - o nhị phân và *k*-phân
  - o các thuộc tính rời rạc, rõ ràng và các thuộc tính liên tục
- **Luật đánh nhãn**: một nút lá được đánh nhãn vào một lớp mà phần lớn các mẫu tại nút này thuộc về lớp đó

Overfitting trong phân lớp bằng cây quyết định

- **Cây tạo được có thể overfit dữ liệu huấn luyện**
  - o quá nhiều nhánh
  - o độ chính xác kém cho những mẫu chưa biết
- **Lý do overfit**
  - o dữ liệu nhiễu và tách rời khỏi nhóm
  - o dữ liệu huấn luyện quá ít
  - o các giá trị tối đa cục bộ trong tìm kiếm tham lam (greedy search)

Cách nào để tránh overfitting?

Hai hướng:

- **rút gọn trước:** ngừng sớm
- **rút gọn sau:** loại bỏ bớt các nhánh sau khi xây xong toàn bộ cây

Phân lớp trong các cơ sở dữ liệu lớn

- **Tính scalability**: phân lớp các tập dữ liệu có hàng triệu mẫu và hàng trăm thuộc tính với tốc độ chấp nhận được
- **Tại sao sử dụng cây quyết định trong khai thác dữ liệu?**
  - o tốc độ học tương đối nhanh hơn các phương pháp khác
  - o có thể chuyển đổi thành các luật phân lớp đơn giản và dễ hiểu
  - o có thể dùng các truy vấn SQL phục vụ truy cập cơ sở dữ liệu
  - o độ chính xác trong phân lớp

Các phương pháp sử dụng cây quyết định trong các nghiên cứu về khai phá dữ liệu

- **SLIQ** (EDBT'96 — Mehta et al.)
- **SPRINT** (VLDB'96 — J. Shafer et al.)
- **PUBLIC** (VLDB'98 — Rastogi & Shim)
- **RainForest** (VLDB'98 — Gehrke, Ramakrishnan & Ganti)

Phân lớp Bayes: Tại sao? (1)

- **Học theo xác suất**:
  - o Tính các xác suất rõ ràng cho các giả thiết
- **Có tăng trưởng**:
  - o Mỗi mẫu huấn luyện có thể tăng/giảm dần khả năng đúng của một giả thiết
  - o Tri thức ưu tiên có thể kết hợp với dữ liệu quan sát

Phân lớp Bayes: Tại sao? (2)

- **Dự đoán theo xác suất**:
  - Dự đoán nhiều giả thiết, trọng số cho bởi khả năng xảy ra của chúng
- **Chuẩn**:
  - Ngay cả khi các phương pháp Bayes khó trong tính toán, chúng vẫn có thể cung cấp một chuẩn để tạo quyết định tới ưu so những phương pháp khác

Phân lớp Bayes

- Bài toán phân lớp có thể hình thức hóa bằng **xác suất a-posteriori:**

  P($C|X$**)** = xác suất mẫu $X=<x_1,...,x_k>$ thuộc về lớp C
- Ví dụ

  P(class=***N*** | outlook=sunny,windy=true,…**)**
- **Ý tưởng:** gán cho mẫu X nhãn phân lớp là C sao cho **P(**$C|X$**)** là lớn nhất

Tính xác suất a-posteriori

- **Định lý Bayes**:

  P($C|X$**) = P(**$X|C$**)·P(**C**) / P(**X**)**
- **P(**X**)** là hằng số cho tất cả các lớp
- **P(**C**)** = tần số liên quan của các mẫu thuộc lớp **C**
- *Chọn lớp C* sao cho **P(**$C|X$*)* lớn nhất = *Chọn lớp C* sao cho **P(**$X|C$**)·P(**C**)** lớn nhất

Phân lớp Naïve Bayesian

- Thừa nhận Naïve: **sự độc lập thuộc tính**

  P($x_1,...,x_k|C$**) = P(**$x_1|C$**)·…·P(**$x_k|C$**)**
- Nếu thuộc tính thứ i là **rời rạc**:
  P($x_i|C$**)** được ước lượng bởi tần số liên quan của các mẫu có giá trị x_i cho thuộc tính thứ i trong lớp C
- Nếu thuộc tính thứ i là **liên tục**:
  P($x_i|C$**)** được ước lượng thông qua một hàm mật độ Gaussian
- Tính toán dễ dàng trong cả hai trường hợp

Phân lớp Naïve Bayes – Ví dụ (1)

- Ước lượng **P(**$x_i|C$**)**

P(p**) = 9/14**
P(n**) = 5/14**

Thời tiết	
P(nắng | p) = 2/9	P(nắng | n) = 3/5
P(u ám | p) = 4/9	P(u ám | n) = 0
P(mưa | p) = 3/9	P(mưa | n) = 2/5
Nhiệt độ	
P(nóng | p) = 2/9	P(nóng | n) = 2/5
P(ấm áp | p) = 4/9	P(ấm áp | n) = 2/5
P(mát | p) = 3/9	P(mát | n) = 1/5

Độ ẩm	
P(cao | p) = 3/9	P(cao | n) = 4/5
P(vừa | p) = 6/9	P(vừa | n) = 1/5
Gió	
P(có | p) = 3/9	P(có | n) = 3/5
P(không | p) = 6/9	P(không | n) = 2/5

Phân lớp Naïve Bayesian – Ví dụ (2)

- Phân lớp X:
  - một mẫu chưa thấy X = <*mưa, nóng, cao, không*>
  - **P(**$X|p$**)·P(**p**) =**
    P(*mưa*$|p$**)·P(***nóng*$|p$**)·P(***cao*$|p$**)·P(***không*$|p$**)·P(**p**) =**
    3/9·2/9·3/9·6/9·9/14 = 0.010582
  - **P(**$X|n$**)·P(**n**) =**
    P(*mưa*$|n$**)·P(***nóng*$|n$**)·P(***cao*$|n$**)·P(***không*$|n$**)·P(**n**) =**
    2/5·2/5·4/5·2/5·5/14 = **0.018286**
  - Mẫu X được phân vào lớp n (không chơi tennis)

Phân lớp Naïve Bayesian – giả thuyết độc lập

- … làm cho có thể tính toán
- … cho ra bộ phân lớp tối ưu khi thỏa yêu cầu
- … nhưng yêu cầu ít khi được thỏa trong thực tế vì các thuộc tính (các biến) thường có liên quan với nhau.
- Những cố gắng khắc phục điểm hạn chế này:
  - **Các mạng Bayes (Bayesian networks)**, kết hợp lý luận Bayes với các mối quan hệ nhân quả giữa các thuộc tính
  - **Các cây quyết định**, lý luận trên một thuộc tính tại một thời điểm, xét những thuộc tính quan trọng nhất trước

Các phương pháp phân lớp khác

- **Mạng Neural**
- **Phân lớp k láng giềng gần nhất**
- **Suy luận dựa vào trường hợp**
- **Thuật toán di truyền**
- **Hướng tập thô**
- **Các hướng tập mờ**

Độ chính xác trong phân lớp

Ước lượng tỉ lệ sai:

- **Phân hoạch**: huấn luyện và kiểm tra (những tập dữ liệu lớn)
  - dùng hai tập dữ liệu độc lập , tập huấn luyện (2/3), tập kiểm tra (1/3)
- **Kiểm tra chéo** (những tập dữ liệu vừa)
  - chia tập dữ liệu thành *k* mẫu con
  - sử dụng *k-1* mẫu con làm tập huấn luyện và một mẫu con làm tập kiểm tra --- kiểm tra chép *k* thành phần
- **Bootstrapping**: xóa đi một - leave-one-out (những tập dữ liệu nhỏ)

Tóm tắt (1)

- **Phân lớp là một vấn đề nghiên cứu bao quát**
- **Phân lớn có khả năng là một trong những kỹ thuật khai phá dữ liệu được dùng rộng rãi nhất với rất nhiều mở rộng**

Tóm tắt (2)

- **Tính uyển chuyển vẫn đang là một vấn đề quan trọng của tất các ứng dụng cơ sở dữ liệu**
- **Các hướng nghiên cứu: phân lớp dữ liệu không-quan hệ, ví dụ như text, không gian và đa phương tiện**