PHÂN LỚP DỮ LIỆU

Giảng viên: Đặng Thị Thu Hiền, Nguyễn Tu Trung
BM HTTT, Khoa CNTT, Trường ĐH Thủy Lợi

Hà Nội, 2023

Nội dung

- Tổng quan về phân lớp dữ liệu
- Phân lớp dữ liệu với mạng Bayesian
- Phân lớp dữ liệu với kNN
- Phân lớp dữ liệu với cây quyết định
- Một số phương pháp phân lớp dữ liệu khác

Tổng quan về phân lớp dữ liệu

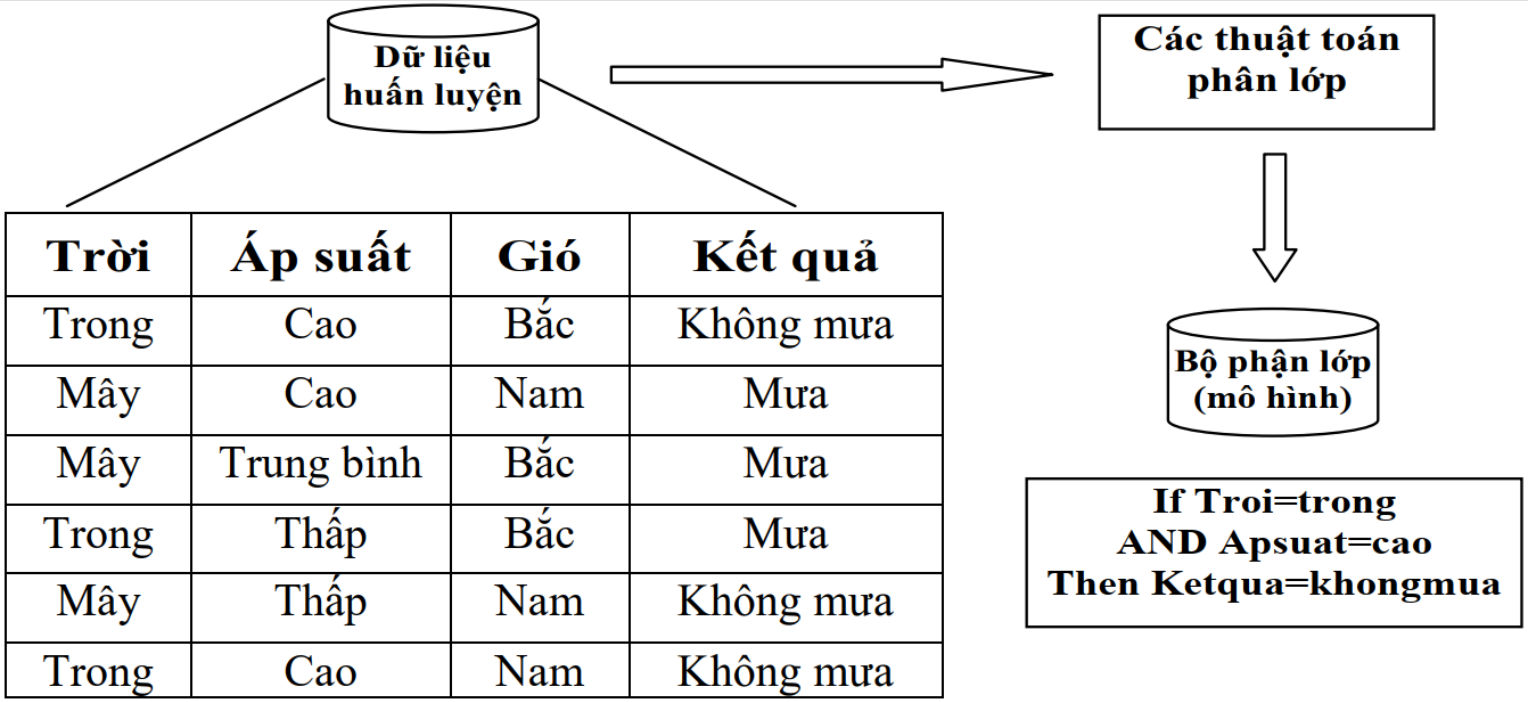
- Phân lớp và dự đoán là hai dạng của phân tích dữ liệu nhằm trích rút ra một mô hình mô tả các lớp dữ liệu quan trọng hay dự đoán xu hướng dữ liệu tương lai
  - Phân lớp dự đoán giá trị của những nhãn xác định hay những giá trị rời rạc, nghĩa là *phân lớp thao tác với những đối tượng dữ liệu mà có bộ giá trị là biết trước*
  - Dự đoán xây dựng mô hình với các hàm nhận giá trị liên tục
- Ví dụ:
  - Mô hình phân lớp dự báo thời tiết có thể cho biết ngày mai mưa hay nắng dựa vào những thông số về độ ẩm, sức gió, nhiệt độ,… của ngày hôm nay và các ngày trước đó

Tổng quan về phân lớp dữ liệu

- Ví dụ (tiếp):
  - Nhờ các luật về xu hướng mua hàng của khách hàng trong siêu thị, các nhân viên kinh doanh có thể ra những quyết sách đúng đắn về lượng mặt hàng cũng như chủng loại bày bán…
  - Một mô hình dự đoán có thể dự đoán lượng tiền tiêu dùng của các khách hàng tiềm năng dựa trên những thông tin về thu nhập và nghề nghiệp của họ
- Quá trình phân lớp dữ liệu gồm hai bước:
  - Xây dựng mô hình
  - Sử dụng mô hình

Xây dựng mô hình

- Mô tả một tập những lớp đã được định nghĩa trước, trong đó mỗi bộ hoặc mẫu dữ liệu được gán về một lớp đã xác trước bởi thuộc tính nhãn lớp
- Tập hợp những bộ và lớp được "học" bởi chương trình để xây dựng mô hình được gọi là tập dữ liệu huấn luyện
- Mô hình biểu diễn dưới dạng luật phân lớp, cây quyết định hoặc công thức toán học...



Trời	Áp suất	Gió	Kết quả
Trong	Cao	Bắc	Không mưa
Mây	Cao	Nam	Mưa
Mây	Trung bình	Bắc	Mưa
Trong	Thấp	Bắc	Mưa
Mây	Thấp	Nam	Không mưa
Trong	Cao	Nam	Không mưa

Sử dụng mô hình

- Nhằm mục đích xác định lớp của dữ liệu trong tương lai hoặc phân lớp những đối tượng chưa biết
- Trước khi sử dụng mô hình cần đánh giá độ chính xác của mô hình:
  - 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
  - Lưu ý: tập kiểm tra và tập huấn luyện là độc lập với nhau

Trời	Áp suất	Gió	Kết quả
Trong	Cao	Nam	Không mưa
Mây	Thấp	Nam	Không mưa
Mây	Trung bình	Bắc	Mưa
Trong	Cao	Bắc	Không mưa

Trong,Cao,Nam?

Kết quả

KhôngMưa

Phân lớp dữ liệu với mạng Bayesian

- Định lý Bayes
- Giới thiệu thuật toán Naïve Bayes
- Thuật toán phân lớp Bayes
- Ví dụ minh họa
- Lưu đồ thuật toán phân lớp Bayes
- Thực thi thuật toán với ví dụ
- Phân tích thuật toán Bayes

Định lý Bayes

- Định lý Bayes được phát biểu như sau:
  - $P(Y|X) = \frac{P(X|Y).P(Y)}{P(X)}$
- P(Y): Xác suất của sự kiện Y xảy ra
- P(X): Xác suất của sự kiện X xảy ra
- P(X|Y): Xác suất (có điều kiện) của sự kiện X xảy ra, nếu biết rằng sự kiện Y đã xảy ra
- P(Y|X): Xác suất (có điều kiện) của sự kiện Y xảy ra, nếu biết rằng sự kiện X đã xảy ra

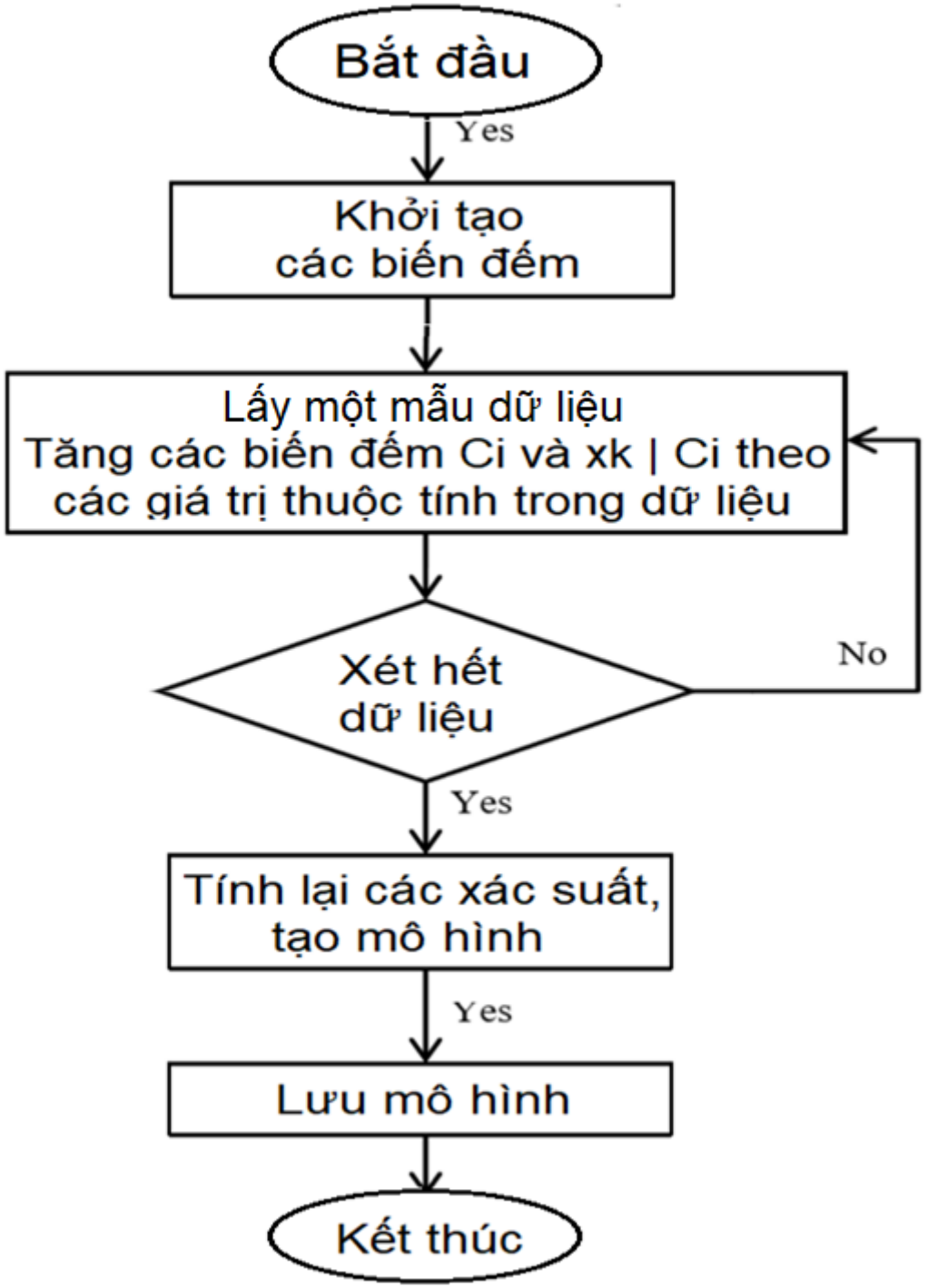
Giới thiệu thuật toán Naïve Bayes

- Là phương pháp phân loại dựa vào xác suất
- Sử dụng rộng rãi trong lĩnh vực học máy, phổ biến trong nhiều lĩnh vực như các công cụ tìm kiếm, các bộ lọc mail nói riêng và phân loại văn bản nói chung
- Ý tưởng cơ bản của cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại
- Điểm quan trọng của phương pháp này chính:
  - Giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau
  - Không khai thác sự phụ thuộc của nhiều từ vào trong một chủ đề cụ thể
- Được xem là thuật toán đơn giản nhất trong các phương pháp

Thuật toán phân lớp Bayes

- Dữ kiện cần có:
  - D: tập dữ liệu huấn luyện, được vector hoá dưới dạng $\vec{x} = (x_1, x_2, \dots, x_n)$
  - C_i : tập các điểm dữ liệu của D thuộc lớp C_i với i={1,2,3,…}
  - Các thuộc tính $x_1, x_2, \dots, x_n$ độc lập xác suất đôi một với nhau
- Thuật toán Naïve Bayes cơ bản:
  - Bước 1 : Huấn luyện Naïve Bayes (dựa vào tập dữ liệu)
    - Tính xác suất $P(C_i)$
    - Tính xác suất $P(x_k|C_i)$
  - Bước 2: Phân lớp X_{new}
    - Tính $F(X_{new}, C_i) = P(C_i) \prod_{k=1}^{n} P(x_k|C_i)$
    - X_{new} được gán vào lớp C_q sao cho
      - $F(X_{new}, C_q) = \max(F(X_{new}, C_i))$

Lưu đồ thuật toán phân lớp Bayes



Ví dụ minh họa thuật toán Bayes

- Yêu cầu: Dự đoán quyết định của người chơi có đi chơi Tennis hay không với các điều kiện về thời tiết đã được biết trước
- Bảng dữ liệu huấn luyện như sau:

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
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

Ví dụ minh họa

- Bảng dữ liệu huấn luyện như sau (tiếp):

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	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

Thực thi thuật toán Bayes với ví dụ

- Có 2 lớp dự báo:
  - C1 = “yes” => Có đi chơi Tennis
  - C2 = “no” => Không đi chơi Tennis
- B1: Huấn luyện Bayes
- B2: Phân lớp Bayes

B1: Huấn luyện Bayes

- Tính các xác suất $P(C_i)$
  - P(C1) = P("yes") = 9/14
  - P(C2) = P("no") = 5/14
- Tính các xác suất $P(x_k|C_i)$
  - Với thuộc tính Outlook
  - Với thuộc tính Temp
  - Với thuộc tính Humidity
  - Với thuộc tính Wind

Với thuộc tính Outlook

- Có các giá trị: sunny, overcast, rain
- P(sunny | yes) = 2/9
- P(sunny | no) = 3/5
- P(overcast | yes) = 4/9
- P(overcast | no) = 0/5
- P(rain | yes) = 3/9
- P(rain | no) = 2/5

Với thuộc tính Temp

- Có các giá trị: Hot, Cold, Mild
- P(hot | yes) = 2/9
- P(hot | no) = 2/5
- P(cold | yes) = 3/9
- P(cold | no) = 1/5
- P(mild | yes) = 4/9
- P(mild | no) = 2/5

Với thuộc tính Humidity

- Có các giá trị: Normal, High
- P(normal | yes) = 6/9
- P(normal | no) = 1/5
- P(high | yes) = 3/9
- P(high | no) = 4/5

Với thuộc tính Wind

- Có các giá trị: Weak, Strong
- P(weak | yes) = 6/9
- P(weak | no) = 2/5
- P(strong | yes) = 3/9
- P(strong | no) = 3/5

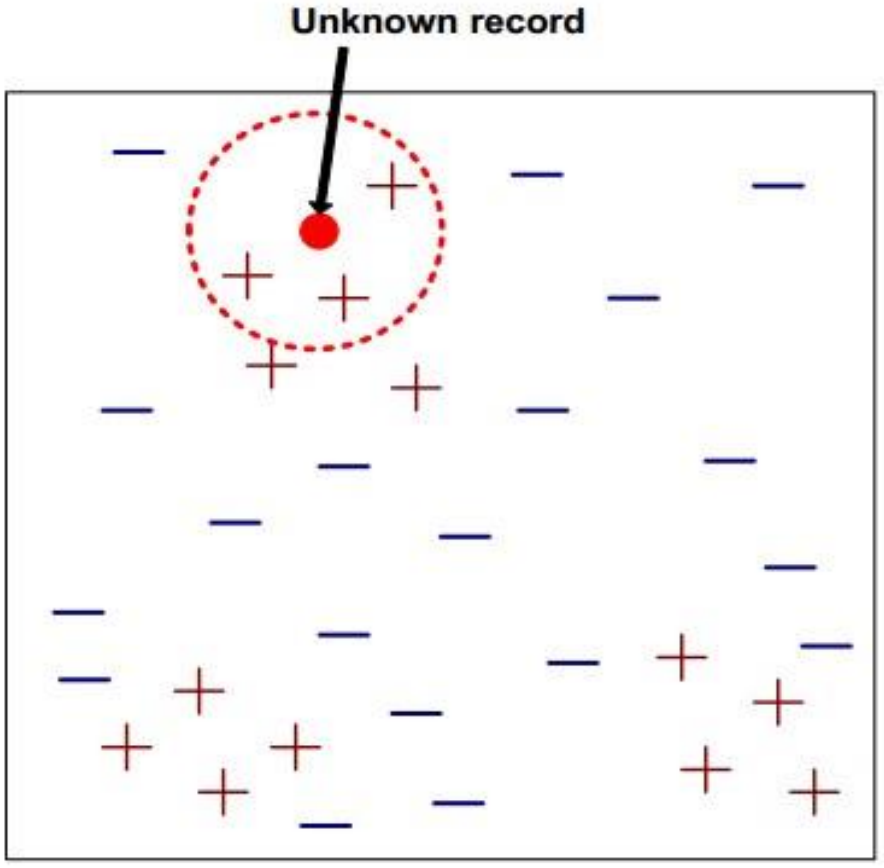
B2: Phân lớp Bayes

- X^{new} = {sunny, cool, high, strong}
- Tính các xác suất
  - $F(X^{new} | yes) = P(yes) * P(sunny | yes) * P(cool | yes) * P(high | yes) * P(strong | yes) = 9/14 * 2/9 * 3/9 * 3/9 * 3/9 = 0.0053$
  - $F(X^{new} | no) = P(no) * P(sunny | no) * P(cool | no) * P(high | no) * P(strong | no) = 5/14 * 3/5 * 1/5 * 4/5 * 3/5 = 0.0206$
- Kết luận: X^{new} thuộc vào lớp No

Phân lớp bằng KNN

Phân loại k-nn (k-nearest neighbor)

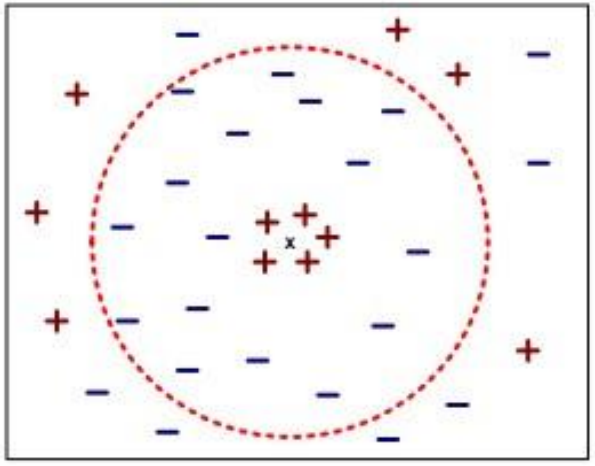
- Cho trước tập dữ liệu huấn luyện D với các lớp, phân loại record/object X vào các lớp dựa vào k phần tử tương tự với X nhất (dùng luật số đông: majority vote)



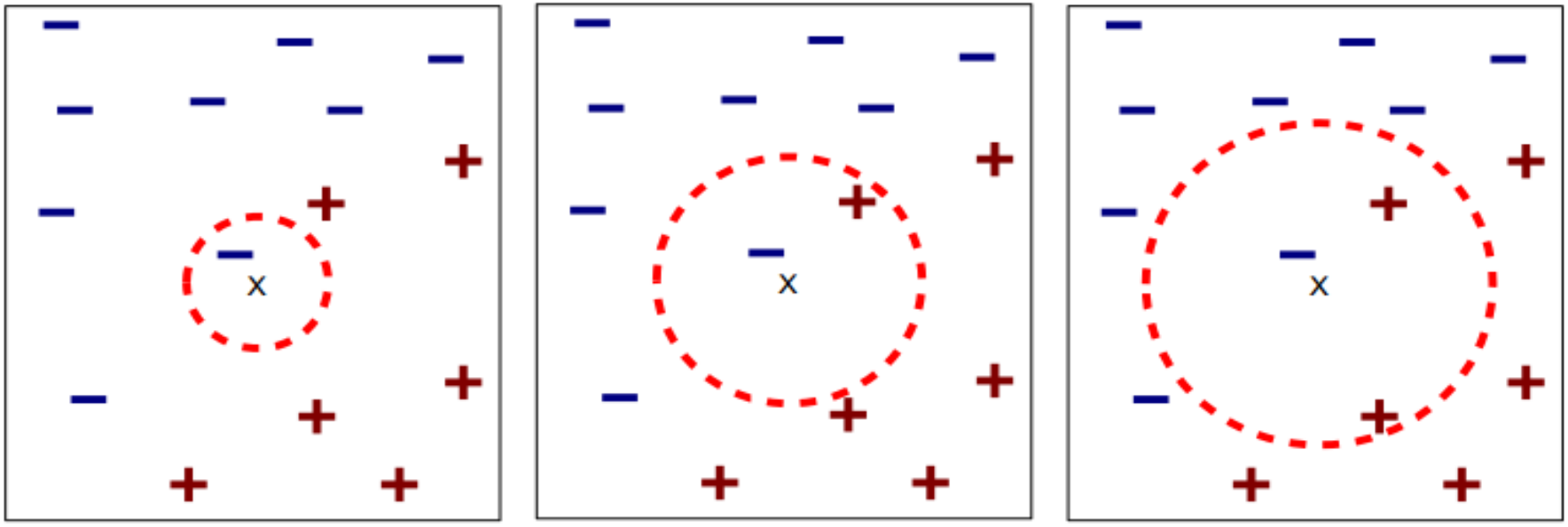
Phân lớp bằng KNN

- 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!



Phân lớp bằng KNN



(a) 1-nearest neighbor

X ∈ MINUS

(b) 2-nearest neighbor

X ∈ MINUS
hay
X ∈ PLUS ?

(c) 3-nearest neighbor

X ∈ PLUS

Phân lớp dữ liệu với cây quyết định

- Khái niệm cây quyết định
- Các bước thuật toán cây quyết định
- Độ lợi thông tin
- Thuật toán ID3
- Thuật toán C4.5
- Rút gọn cây quyết định và tập luật suy dẫn
- Phân lớp với chỉ số Gini

Khái niệm cây quyết định

- Là một flow-chart giống cấu trúc cây, gồm
  - Các nút biểu diễn thuộc tính
  - Các nhánh biểu diễn đầu ra của kiểm tra
  - Nút lá biểu diễn nhãn lớp
- Tạo theo hai giai đoạn là tạo cây và tỉa nhánh
  - Giai đoạn tạo cây:
    - Lúc bắt đầu tất cả các mẫu học đều nằm ở nút gốc
    - Tiếp đó, các mẫu học được chia một cách đệ quy dựa trên thuộc tính được chọn
  - Giai đoạn tỉa nhánh: Nhằm tìm và xóa những nhánh có phần tử không thể xếp vào lớp nào cả

Các bước thuật toán cây quyết định

- Các bước
  - B1: Cây được xây dựng đệ quy từ trên xuống và theo cách chia để trị
  - B2: Ban đầu tất cả mẫu học đều nằm ở gốc
  - B3: Thuộc tính được phân loại (nếu là giá trị liên tục được rời rạc hóa)
  - B4: Các mẫu học được chia đệ quy dựa trên thuộc tính chọn lựa
  - B5: Kiểm tra những thuộc tính được chọn dựa trên heuristic hoặc của một tiêu chuẩn thống kê
- 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

Độ lợi thông tin (information gain)

- Là đại lượng dùng để chọn thuộc tính để phân chia tập học
- Thuộc tính được chọn:
  - Là thuộc tính cho độ đo tốt nhất, có lợi nhất trong quá trình phân lớp
- D: tập huấn luyện
- $C_{i,D}$: tập các mẫu của D thuộc lớp C_i với i = {1,2,…,m}
- $|C_{i,D}|, |D|$: lực lượng của tập $C_{i,D}$ và D
- p_i (p_i # 0): xác suất để một mẫu bất kì của D thuộc về lớp C_i và tính như sau: $p_i = \frac{|C_{i,D}|}{|D|}$

Độ lợi thông tin (information gain) TH1

- Thông tin mong đợi để phân lớp một mẫu trong D theo nhãn lớp: $Entropy(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$
- Thuộc tính A chứa v giá trị $a_1, a_2, \ldots, a_v$ trong D
- Dùng A để chia tập huấn luyện D thành v tập con $D_1, D_2, \ldots, D_v$
- Thông tin cần thiết để phân chia D theo thuộc tính A:
  - $Entropy_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Entropy\ (D_j)$
- Độ lợi thông tin của sự phân chia dựa trên thuộc tính A:
  - $Gain(A) = Entropy(D) - Entropy_A\ (D)$

Thuật toán ID3

- Giới thiệu thuật toán ID3
- Thủ tục ID3(D,C,A)
- Ví dụ minh họa thuật toán ID3
- Rút luật từ cây quyết định
- Hạn chế của thuật toán ID3
- Mở rộng thuật toán ID3

Giới thiệu thuật toán ID3

- Là một thuật toán học các mẫu để tạo cây quyết định
- Ý tưởng: tạo cây quyết định bằng việc tìm kiếm cơ bản từ trên xuống trên tập huấn luyện
- Độ lợi thông tin được sử dụng để chọn thuộc tính có khả năng phân loại tốt nhất
- Input:
  - Tập huấn luyện D
  - Tập thuộc tính A
  - Thuộc tính quyết định C
- Ouput:
  - Cây quyết định
- Thủ tục thuật toán: ID3(D,C,A)

Thủ tục: ID3(D,C,A)

- B1: Tạo "nút_gốc" cho cây quyết định
- B2: Kiểm tra trường hợp đặc biệt
  - B2.1: If tất cả mẫu huấn luyện của D đều có trị của C là P, return cây có một nút duy nhất là nút_gốc với nhãn P
  - B2.2: If tất cả mẫu huấn luyện của D đều có trị của C là N, return cây có một nút duy nhất là nút_gốc với nhãn N
  - B2.2: If A là rỗng return cây có nút duy nhất là nút_gốc với nhãn là trị phổ biến nhất của C trong tập mẫu
- B3: Else begin
  - B3.1: Gọi X là thuộc tính của A phân lớp D tốt nhất //*Xác định X bằng cách tính Gain information*
  - B3.2: Gán nhãn nút_gốc với tên thuộc tính X
  - B3.3: For each giá trị v của X

Thủ tục: ID3(D,C,A)

- B3: Else begin (tiếp)
  - B3.3: For each giá trị v của X (tiếp)
    - B3.3.1: Thêm một nhánh cây mới dưới nút_gốc ứng với X=v
    - B3.3.2: Xác định tập con D_v ứng với X=v
    - B3.3.3: If Dv là rỗng
      - Thêm dưới nhánh mới này một nút lá có nhãn là trị phổ biến nhất của thuộc tính quyết định trong D
    - B3.3.4: Else
      - Thêm cây con vào dưới nhánh mới này bằng cách gọi đệ quy ID3(D_v, C, A-{X})
  - End
- Return nút_gốc

Ví dụ minh họa thuật toán ID3

- Tập dữ liệu học D *"chơi tennis"*
- Tập thuộc tính A = {Thời tiết, Nhiệt độ, Độ ẩm, Gió} => khác ∅
- B1: Tạo nút gốc cho cây quyết định
- Thuộc tính quyết định C: Lớp có có miền giá trị {P,N} (m=2) => sang B3
- B3.1: Xác định X bằng cách tính độ lợi của từng thuộc tính trong A

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
Náng	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

Ví dụ minh họa thuật toán ID3

- Số thuộc tính mang nhãn P là 9 trong tổng số 14 bộ trong D

$$Entropy(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) = -\frac{9}{14} log_2\left(\frac{9}{14}\right) - \frac{5}{14} log_2\left(\frac{5}{14}\right) = 0.94$$

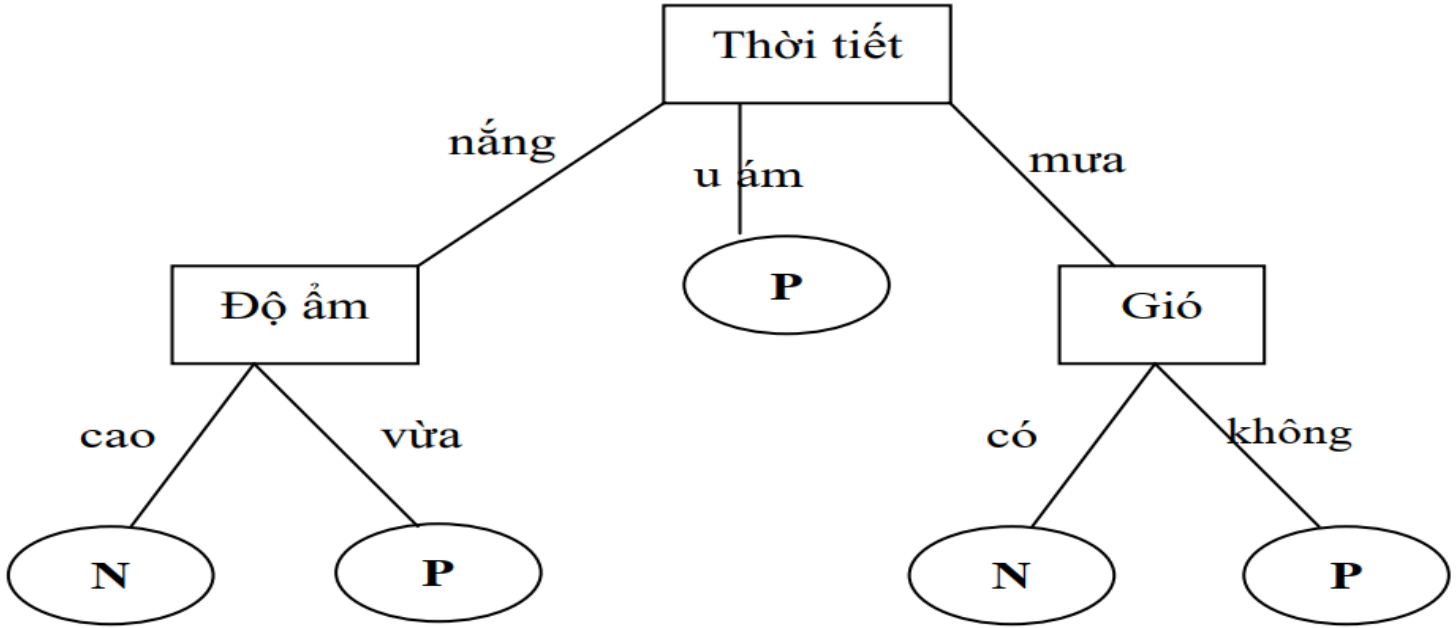
- Xét thuộc tính "Thời tiết":
  - Miền giá trị {nắng, u_ám, mưa} => v = 3
  - => Chia D thành 3 tập con: D_1 (nắng), D_2 (u_ám), D_2 (mưa) với số lượng mẫu như bảng bên

Thời tiết \ Lớp	P	N	Tổng
Nắng	2	3	5
U ám	4	0	4
mưa	3	2	5

  - $Entropy(D_1) = -\frac{2}{5} log_2\left(\frac{2}{5}\right) - \frac{3}{5} log_2\left(\frac{3}{5}\right) = 0.971$
  - $Entropy(D_2) = -\frac{4}{4} log_2\left(\frac{4}{4}\right) - \frac{0}{4} log_2\left(\frac{0}{4}\right) = 0$
  - $Entropy(D_3) = -\frac{3}{5} log_2\left(\frac{3}{5}\right) - \frac{2}{5} log_2\left(\frac{2}{5}\right) = 0.971$
  - $Entropy_{Thời\ tiết}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Entropy\ (D_j) = \frac{5}{14} * 0.971 + \frac{4}{14} * 0 + \frac{5}{14} * 0.971 = 0.694$

Ví dụ minh họa thuật toán IC3

- Xét thuộc tính "Thời tiết" (tiếp):
  - $Gain(Thời\ tiết) = Entropy(D) - Entropy_{Thời\ tiết}(D) = 0.94 - 0.694 = 0.246$
- Tương tự ta tính được
  - $Gain(Nhiệt\ độ) = 0.029; Gain(\text{Độ ẩm}) = 0.151; Gain(Gió) = 0.003$
- Chọn thuộc tính có độ lợi thông tin lớn nhất là thuộc tính "Thời tiết"
- Áp dụng ID3 cho mỗi nút con của nút gốc cho đến khi đạt đến nút lá hoặc nút có entropy = 0



Rút luật từ cây quyết định

- Mỗi một đường dẫn từ gốc đến lá trong cây tạo thành một **luật**
  - Ví dụ: IF Thời tiết = nắng AND Độ ẩm = vừa THEN Chơi tennis
- Mỗi cặp giá trị thuộc tính trên một đường dẫn tạo nên một liên kết
- Nút lá giữ quyết định phân lớp dự đoán
- Các luật tạo được dễ hiểu hơn các cây

Hạn chế của thuật toán ID3

- ID3 có thể 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

Mở rộng thuật toán ID3

- ID3 được mở rộng cho trường hợp tập mẫu có thuộc tính liên tục
  - => cần phân tích thuộc tính liên tục thành một tập rời rạc các khoảng
- 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
- Trong VD chơi tennis, nếu có một bản ghi chưa biết giá trị của thuộc tính Độ ẩm, nhưng biết giá trị của thuộc tính Thời tiết là Nắng ta xử lý như sau:
  - Di chuyển từ nút gốc Thời tiết đến nút Độ ẩm theo cạnh được đánh nhãn là Nắng
  - Thuộc tính Độ ẩm có giá trị Cao thì có 2 bản ghi, nếu thuộc tính Độ ẩm có trị lớn hơn Vừa thì có 3 bản ghi => Có thể đưa ra câu trả lời cho xác suất xảy ra khả năng là 0.4 cho chơi tennis và 0.6 cho không chơi tennis

Thuật toán C4.5

- Với việc rời rạc hóa thuộc tính liên tục:
  - Giả sử A là thuộc tính liên tục => chuyển sang kiểu logic mới có tên là Ac, có giá trị Đúng nếu A<c và Sai nếu A>=c
  - Vấn đề là chọn ngưỡng c?
    - Xét các giá trị của thuộc tính A trong tập học: Giả sử được sắp theo thứ tự tăng dần là A1, A2,…, Am
    - Với từng giá trị Ai, i=1,1,…,m ta chia các bản ghi thành hai phần, một phần chứa các mẫu có giá trị A<c và phần chứa các mẫu có A>=c
    - Với những lần phân hoạch này, tính lại độ lợi thông tin của phép phân hoạch và tìm phân hoạch có độ lợi lớn nhất
- Giả sử Độ ẩm trong VD chơi tennis là thuộc tính liên tục
  - Cần xác định độ lợi thông tin cho mỗi lần phân hoạch theo thuộc tính Độ ẩm và tìm được sự phân hoạch tốt nhất tại ngưỡng c=75
  - => giá trị thuộc tính Độ ẩm biến đổi thành thuộc tính Độ ẩm_logic với giá trị Sai, Đúng cho các mẫu có giá trị < 75 và >=75

Rút gọn cây quyết định và tập luật suy dẫn

- Việc xây dựng cây quyết định đều dựa vào tập học
- Trong thực tế cây quyết định có thể phát sinh các đường đi dài và không đều
- Việc rút gọn cây quyết định được thực hiện bằng cách biến cây con thành nút lá: thực hiện tại nơi nếu lỗi phân lớp do cây con sinh ra lớn hơn nút lá
- Winston dùng phép thử Fisher để xác định thuộc tính phân loại có thực sự phụ thuộc vào các thuộc tính khác hay không?
  - Nếu điều này không xảy ra thì thuộc tính đó không cần phải xuất hiện trong đường đi hiện tại của cây
- Quinlan và Breiman đề xuất heuristic để rút gọn cây:
  - Từ mỗi đường đi từ gốc đến lá, ta tạo ra vế trái của luật phân lớp dựa trên nhãn của các nút và nhãn của các cung

Phân lớp với chỉ số Gini

- Tương tự như độ lợi ở trên, IBM trong phần mềm IBM Intelligent Miner đưa ra đại lượng cho việc phân lớp là chỉ số Gini như sau:
  - Nếu một tập dữ liệu T chứa những mẫu từ n lớp, chỉ số Gini, Gini(T) được định nghĩa: $Gini(T) = 1 - \sum_{j=1}^{n} p_j^2$
    - Với p_j là tần số liên quan của lớp j trong T
  - Nếu một tập hợp dữ liệu T được chia thành hai tập con T_1, T_2 với kích thước tương ứng là N_1 và N_2
    - Chỉ số Gini của dữ liệu chia cắt được định nghĩa như sau: $Gini_{split}(T) = \frac{N_1}{N} Gini(T_1) + \frac{N_2}{N} Gini(T_2)$
    - Thuộc tính có giá trị $Gini_{split}(T)$ nhỏ nhất được chọn để phân chia nút

Một số phương pháp phân lớp dữ liệu khác

- Phân lớp dựa trên luật kết hợp
- Thuật giải di truyền
- Tiếp cận tập thô
- Phân lớp dựa trên mạng Neural, SVM


THE END