TRÍ TUỆ NHÂN TẠO

Bài 12: Học máy và Cây quyết định

Nội dung

1. Học máy là gì?
2. Một số khái niệm liên quan
3. Cây quyết định (decision tree)
4. Giải thuật đâm chồi
5. Thuật toán ID3
6. Xây dựng tập luật từ cây quyết định
7. Bài tập ứng dụng

Phần 1

Học máy là gì?

Học máy là gì?

- **Tiếng Anh:** Machine Learning
- **Khái niệm:** Nghiên cứu về các phương pháp xây dựng khả năng tự thu thập tri thức của máy tính (từ dữ liệu đã có hoặc từ môi trường)
  - Chỉ là một trong nhiều định nghĩa
- Các phương pháp cơ bản: rất nhiều, không kể hết được
  - Hồi quy
  - Cây quyết định (DC – Decision Tree)
  - Phân loại bayer đơn giản (NBC – Naive Bayes Classifier)
  - Máy vector hỗ trợ (SVM - Support Vector Machine)
  - Mạng thần kinh nhân tạo (ANN – Artificial Neural Network)
  - …

Học máy là gì?

- Học máy ≠ Học thuộc lòng:
  - Học thuộc lòng (học vẹt): tri thức đã có được nạp vào máy tính
  - Học máy = học hiểu: máy tính nhận thức được các tri thức nạp vào, tổng quát hóa và áp dụng vào các tình huống mới
- Học máy ≈ Cung cấp cho máy tính khả năng quyết định trong những trường hợp không chuẩn bị trước
- Học có giám sát (học có thầy):
  - Học có chỉ dẫn (learning by instruction)
  - Học bằng suy luận (learning by deduction)
  - Học bằng quy nạp (learning by induction)
- Học không giám sát (học không thầy):
  - Học qua quan sát (learning by observation)
  - Học qua khám phá (learning by discovery)

Các lớp bài toán cơ bản

- Học có giám sát (supervised learning): học cách tiên đoán đầu ra
  - Hồi quy (regression): đầu ra là số hoặc vector
  - Phân lớp (classification): đầu ra là xác suất dự báo
- Học tăng cường (reinforcement learning): hiệu chỉnh các siêu tham số (hyperparameter) để cực đại hóa lợi ích trong tương lai
  - “reinforcement learning is difficult” – Geoffrey Hilton
  - Chìa khóa để tạo ra “strong AI” – những cỗ máy có thể tự học và tự hoàn thiện
  - Hiện chưa có nhiều tiến bộ trong các mô hình
  - Nhưng có nhiều thành công khi kết hợp với các kĩ thuật mới (AlphaZero chẳng hạn)

Các lớp bài toán cơ bản

- Học không giám sát (unsupervised learning): tự khai phá các đặc trưng nội tại hợp lý của đầu vào
- Như thế nào là "hợp lý":
  - Biến đổi dữ liệu đầu vào có số chiều cao thành dữ liệu có số chiều thấp hơn (nhưng không mất thông tin hoặc mất không đáng kể)
  - Dữ liệu có số chiều cao nhưng các đặc trưng thành phần có tính "kinh tế" (economical) hơn
  - Gom cụm dữ liệu đầu vào

Phần 2

Một số khái niệm liên quan

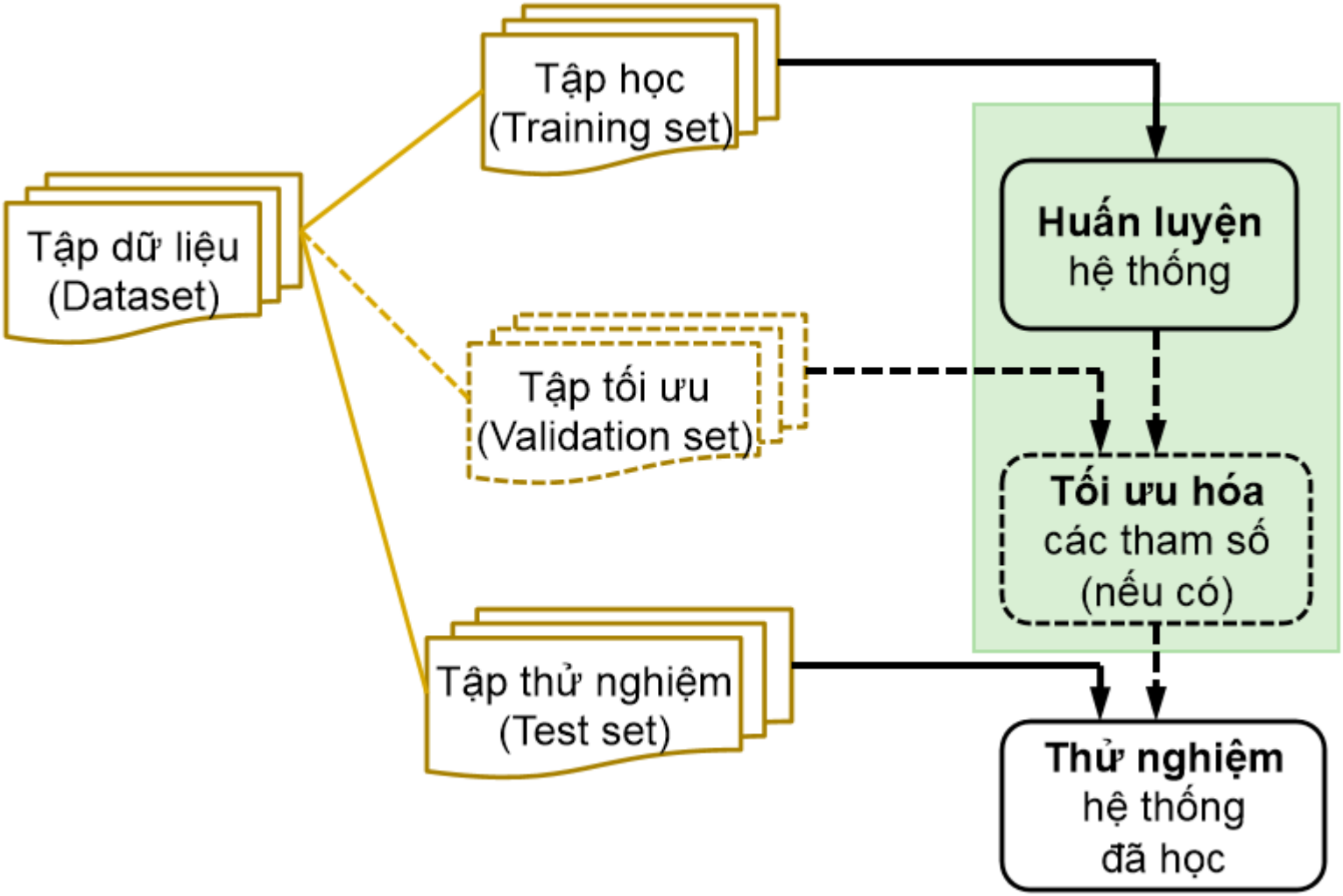
Một số khái niệm liên quan

- Tập dữ liệu huấn luyện (training dataset): tập dữ liệu sử dụng để dạy máy tính học
  - Dữ liệu thật được thu thập từ thực tế
  - Tập dữ liệu cần có tính phổ quát (đa dạng), không quá tập trung vào những trường hợp đặc thù
  - Chất lượng mẫu đủ tốt để học
  - Càng nhiều mẫu càng tốt (?)
- Một số phương pháp học máy tự tách tập dữ liệu này làm đôi (khi đang huấn luyện) để kiểm chứng quá trình học, kỹ thuật này gọi là k-fold cross-validation (xác thực chéo gấp k)

Một số khái niệm liên quan

- Tập dữ liệu kiểm tra (testing dataset): tập dữ liệu sử dụng để kiểm tra kết quả học của máy tính
  - Dữ liệu thật được thu thập từ thực tế, có tính phổ quát
  - Có những mẫu chất lượng không thật tốt để kiểm tra các trường hợp nhập nhằng
- Làm sao để đánh giá kết quả học của máy?
  - Cứ kiểm tra thử, máy trả lời đúng càng nhiều càng tốt! Vậy nếu kết quả trả lời là dạng số thì sao?
  - Có những bài toán trả lời đúng thì không sao, nhưng trả lời sai thì rất nghiêm trọng (chẳng hạn như chuẩn đoán bệnh), vậy nên đánh giá kết quả học thế nào?
  - Nói chung: rất nhiều kĩ thuật, tùy thuộc vào bài toán cụ thể

Một số khái niệm liên quan



Một số khái niệm liên quan

- Hiện tượng "quá kém" (underfitting): Máy thể hiện kết quả kém cả khi học và khi kiểm tra
- Hiện tượng "quá kém" thể hiện mô hình học không phù hợp ➔ máy không có khả năng học bài đạt yêu cầu
  - Khắc phục: điều chỉnh mô hình (quy mô hoặc tham số)
  - Đôi khi phải đổi cả phương pháp huấn luyện
- Hiện tượng "quá khớp" (overfitting): Máy thể hiện tốt khi huấn luyện nhưng lại cho kết quả kém khi kiểm tra
- Hiện tượng "quá khớp" thể hiện phương pháp học không hiệu quả ➔ khả năng tổng quát hóa của máy kém
  - Thường do mô hình quá mạnh, nên khả năng nhớ cao nhưng khả năng tổng quát hóa yếu

Phần 3

Cây quyết định

Cây quyết định: phân loại dựa trên thuộc tính

TT	Độ tuổi	Nghề nghiệp	Chuyên môn	Hiện trạng
1	Già	Bác sĩ	Đa khoa	***Nghỉ hưu***
2	Trung niên	Giảng viên	Toán	***Đi làm***
3	Thanh niên	Sinh viên	Toán	***Đi học***
4	Thanh niên	Làm nông	-	***Đi làm***
5	Già	Giảng viên	Tin học	***Nghỉ hưu***
6	Trung niên	Bác sĩ	Răng	***Đi làm***

Yêu cầu: cho một người A, độ tuổi Trung niên, nghề Bác sĩ, chuyên môn Răng, dự đoán xem hiện trạng của A là thế nào?

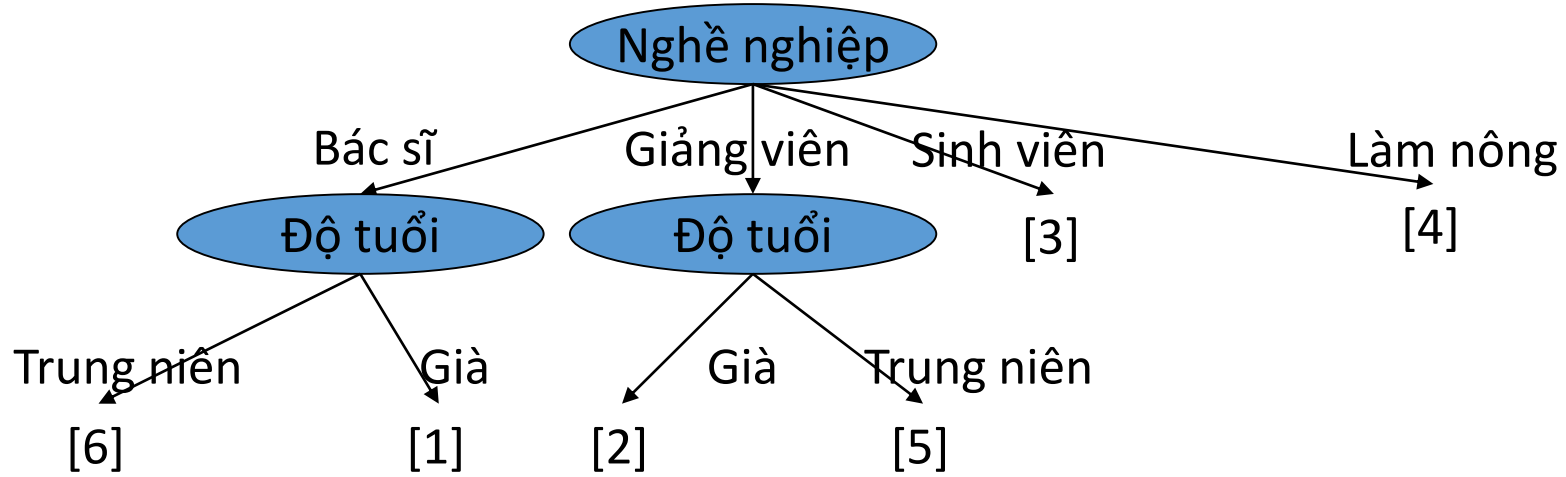
Cây quyết định: bài toán

- Bài toán phân loại mẫu (phân lớp):
  - Dataset: một tập các mẫu, mỗi mẫu gồm nhiều thuộc tính khác nhau và được chỉ định thuộc một phân loại nào đó
  - Huấn luyện: máy nhận các mẫu và tìm ra các đặc trưng để phân loại các mẫu
  - Thử nghiệm: máy nhận một mẫu mới và quyết định xem mẫu mới thuộc phân loại nào
- Mẫu: tập hợp nhiều thuộc tính
  - Có thể có thuộc tính dạng số (tuổi, cân nặng, chỉ số hóa sinh,…)
  - Có thể có thuộc tính phi số (phân loại, mô tả,…)
  - Có thể có thuộc tính thiếu khuyết (không có giá trị)

Cây quyết định: sinh cây từ gốc

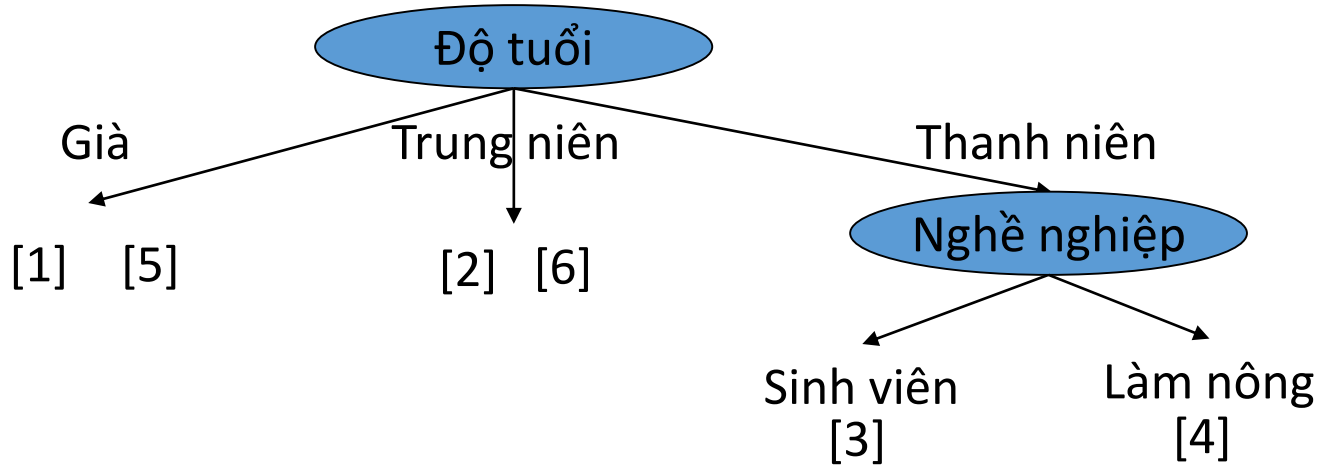
- Rất thích hợp cho bài toán phân hoạch theo **thuộc tính rời rạc**
- Từ một nút gốc chọn một thuộc tính nào đó để phân hoạch → Các mẫu ví dụ bị tách thành các tập nhỏ hơn
- Tiếp tục phân hoạch các tập con cho đến khi các mẫu là đồng nhất theo mục tiêu phân hoạch

Cây quyết định: một cây kết quả ví dụ



TT	Độ tuổi	Nghề nghiệp	Chuyên môn	Hiện trạng
1	Già	Bác sĩ	Đa khoa	***Nghỉ hưu***
2	Trung niên	Giảng viên	Toán	***Đi làm***
3	Thanh niên	Sinh viên	Toán	***Đi học***
4	Thanh niên	Làm nông	-	***Đi làm***
5	Già	Giảng viên	Tin học	***Nghỉ hưu***
6	Trung niên	Bác sĩ	Răng	***Đi làm***

Cây quyết định: một cây kết quả tốt hơn



TT	Độ tuổi	Nghề nghiệp	Chuyên môn	Hiện trạng
1	Già	Bác sĩ	Đa khoa	***Nghỉ hưu***
2	Trung niên	Giảng viên	Toán	***Đi làm***
3	Thanh niên	Sinh viên	Toán	***Đi học***
4	Thanh niên	Làm nông	-	***Đi làm***
5	Già	Giảng viên	Tin học	***Nghỉ hưu***
6	Trung niên	Bác sĩ	Răng	***Đi làm***

Cây quyết định: chất lượng mẫu

- **Vấn đề chất lượng mẫu:** Nếu số mẫu ít hoặc không điển hình sẽ dẫn đến hiện tượng sinh cây quyết định không đúng
- Ví dụ: Nếu chọn thuộc tính "Chuyên môn" để phân lớp tiếp nhóm "Bác sĩ" hoặc "Giảng viên" sẽ dẫn đến kết luận: Bác sĩ + Đa khoa → Nghỉ hưu

Cây quyết định: chọn thuộc tính phân hoạch

- **Vấn đề chọn thuộc tính phân hoạch:** Chọn thuộc tính phân hoạch tùy tiện → Cây quyết định nhiều tầng → Tính tổng quát hóa thấp (overfitting)
- Vậy việc chọn thuộc tính để phân hoạch là vấn đề quan trọng nhất trong chiến lược xây dựng cây quyết định

Cây quyết định: thông tin thiếu khuyết

- **Vấn đề thông tin không rời rạc:** Tìm cách rời rạc hóa các dữ liệu thu thập được.
- **Vấn đề không đủ thông tin:** Đôi khi tập mẫu không có đủ thông tin để phân loại mẫu → Đưa ra kết luận dựa trên số đông các mẫu

Cây quyết định: thông tin nhiễu, không đủ

- Ví dụ: Có 300 mẫu học
  - Độ tuổi = “Già”
  - Các thông tin khác: Không có
  - Hiện trạng:
    - Nghỉ hưu (280 mẫu)
    - Đi làm (18 mẫu)
    - Đi học (2 mẫu)
  - Vậy kết luận: Già → Nghỉ hưu

Dataset: dữ liệu chưa được mã hóa

TT	Độ tuổi	Nghề nghiệp	Chuyên môn	Hiện trạng
1	Già	Bác sĩ	Đa khoa	***Nghỉ hưu***
2	Trung niên	Giảng viên	Toán	***Đi làm***
3	Thanh niên	Sinh viên	Toán	***Đi học***
4	Thanh niên	Làm nông	-	***Đi làm***
5	Già	Giảng viên	Tin học	***Nghỉ hưu***
6	Trung niên	Bác sĩ	Răng	***Đi làm***

Dữ liệu ở dạng thông tin đầy đủ giúp chúng ta hình dung được những mối liên hệ giữa các thuộc tính (do chúng ta có tri thức về ý nghĩa các thuộc tính)

Dataset: dữ liệu đã được mã hóa

TT	A	B	C	Hiện trạng
1	O	D	D	***Nghỉ hưu***
2	A	T	M	***Đi làm***
3	T	S	M	***Đi học***
4	T	A	-	***Đi làm***
5	O	T	I	***Nghỉ hưu***
6	A	D	C	***Đi làm***

Nhưng đa số các tập dữ liệu được công bố hiện nay đều ở dạng mã hóa, vì nhiều lý do, trong đó lý do lớn nhất là an toàn thông tin cá nhân hoặc doanh nghiệp

Phần 4

Giải thuật đâm chồi

Giải thuật đâm chồi (1/3)

- Giải thuật đâm chồi là giải thuật cơ bản để xây dựng cây quyết định
  - R: Nút gốc (chính xác là nút đang xét)
  - S: Tập các mẫu S = (s_1, s_2, …, s_n)
  - T: Tập kết luận của E, T = (t_1, t_2, …, t_m)
  - A: Tập thuộc tính chưa được chọn
- Thuật giải tìm kết luận gắn với R hoặc thuộc tính tiến hành phân hoạch tiếp

Giải thuật đâm chồi (2/3)

Generate (R, S, T, A)

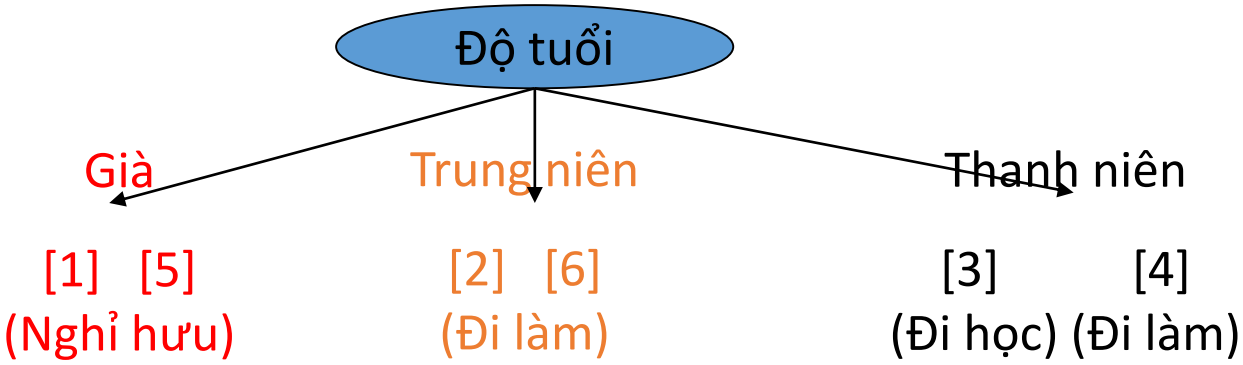
- Nếu T chỉ có 1 kiểu giá trị: kết luận tại R là t_1 và kết thúc nhánh này
- Nếu A rỗng: kết luận tại R là kết luận chiếm đa số đối với E và kết thúc nhánh này
- Chọn thuộc tính A_x để phân hoạch:
  - Nguyên tắc chọn là mở (*)
  - Ghi thuộc tính A_x vào R
  - Xây dựng tất cả các nhánh con từ R, mỗi nhánh là một giá trị có thể của A_x

(*) Chọn một thuộc tính trong tập A để đâm chồi: Việc chọn thế nào tùy vào từng thuật toán

Giải thuật đâm chồi (3/3)

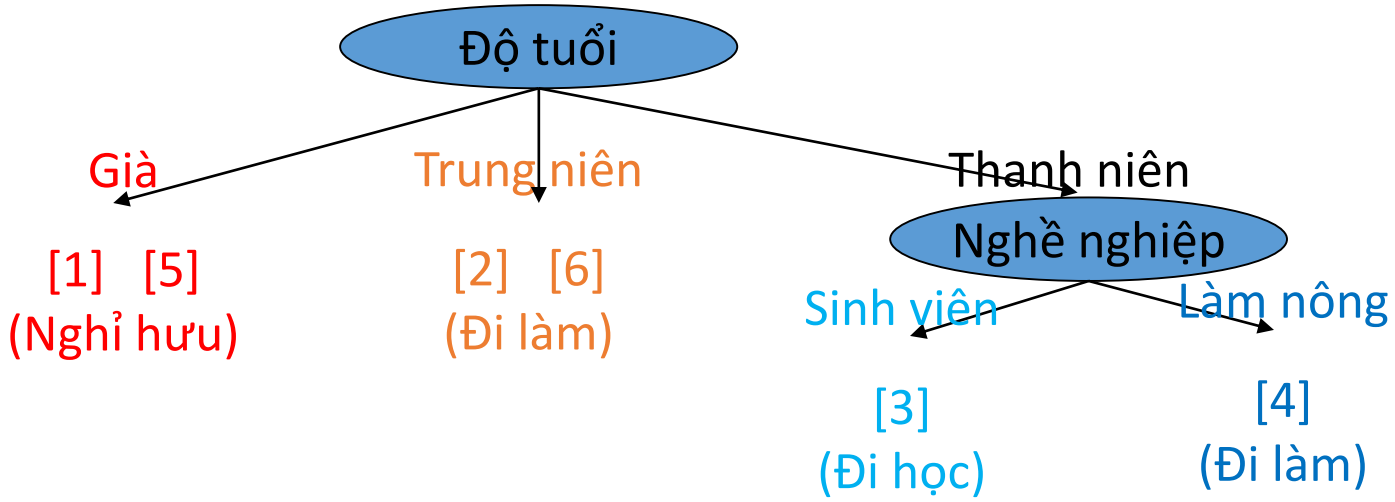
- Với mỗi nhánh giá trị V của A_x:
  - Tạo nút R_V
  - Xây dựng tập S' = {s_i | $s_i \in S$ & thuộc tính A_x của s_i là V}
  - Xây dựng tập T' là tập các kết luận của S'
  - Nếu S' rỗng thì kết luận tại R_V là kết luận chiếm đa số đối với S
  - Ngược lại: Generate (R_V, S', T', A\{A_X})

Ví dụ về quá trình đâm chồi (1)



TT	Độ tuổi	Nghề nghiệp	Chuyên môn	Hiện trạng
1	Già	Bác sĩ	Đa khoa	***Nghỉ hưu***
2	Trung niên	Giảng viên	Toán	***Đi làm***
3	Thanh niên	Sinh viên	Toán	***Đi học***
4	Thanh niên	Làm nông	-	***Đi làm***
5	Già	Giảng viên	Tin học	***Nghỉ hưu***
6	Trung niên	Bác sĩ	Răng	***Đi làm***

Ví dụ về quá trình đâm chồi (2)



TT	Độ tuổi	Nghề nghiệp	Chuyên môn	Hiện trạng
1	Già	Bác sĩ	Đa khoa	***Nghỉ hưu***
2	Trung niên	Giảng viên	Toán	***Đi làm***
3	Thanh niên	Sinh viên	Toán	***Đi học***
4	Thanh niên	Làm nông	-	***Đi làm***
5	Già	Giảng viên	Tin học	***Nghỉ hưu***
6	Trung niên	Bác sĩ	Răng	***Đi làm***

Phần 5

Thuật toán ID3

Thế nào là cây quyết định tốt?

- Giải thuật đâm chồi có thể sinh nhiều cây quyết định khác nhau, tùy thuộc vào việc chọn thuộc tính đâm chồi
- Vậy trong những cây đó cây nào là tốt?
- Một trong những tiêu chuẩn của các thuật toán học máy "tốt" là khả năng tổng quát hóa cao
- Khả năng tổng quát hóa tốt ≈ ít nhánh
  - Đây chỉ là khả năng cao mà thôi, chẳng hạn như cây ít nhánh mà quá mất cân bằng thì cũng không tốt
- Ý tưởng: greedy (tham lam), chọn thuộc tính đem lại cho ta nhiều thông tin nhất
- Vấn đề: Thế nào là "đem lại nhiều thông tin nhất"?

Hàm đo entropy

- P là một tập n loại giá trị khác nhau
- Gọi p_i là xác suất xuất hiện của giá trị thứ i trong tập P
- Hàm đo Entropy của tập P được định nghĩa như sau:

$$E(P) = -\sum_{i=1}^{n} p_i \, log_2 \, p_i$$

- Trong trường hợp P gồm 2 loại giá trị:
  - E(P) = 0 nếu trong tập P tất cả đều thuộc một loại
  - E(P) = 1 nếu các mẫu phân bổ đều (mỗi loại chiếm một nửa)
  - 0 < E(P) < 1 trong các trường hợp còn lại
- Ví dụ: P = {1, 1, 2, 2, 2, 2}
  - Như vậy p_1=1/3, p_2=2/3
  - E(P) = -(1/3 x $\log_2$ 1/3 + 2/3 x $\log_2$ 2/3) = 0.918296

Thuật toán ID3

- Thuật toán ID3 mong muốn chọn ra thuộc tính phân loại tốt nhất với mỗi nút theo nghĩa cách chọn thuộc tính đó sẽ đem lại nhiều entropy nhất cho cây quyết định
- ID3 lập luận như sau:
  - Khi chọn thuộc tính A_X để phân hoạch: Tập S chia thành các tập (S_1, S_2, …, S_w) ứng với w giá trị của thuộc tính A_X
  - E(S) là lượng entropy ban đầu của S
  - E(S_i) là lượng entropy của tập con S_i
  - Vậy lượng entropy thu được qua phân hoạch A_X là:

$$E(S, A_X) = E(S) - \sum_{i=1}^{w} \frac{|S_i|}{|S|} E(S_i)$$

- ID3 = xét các thuộc tính A_i và chọn A_X có E(S, A_X) lớn nhất

Hãy thử thuật toán ID3 với dataset ví dụ

TT	Độ tuổi	Nghề nghiệp	Chuyên môn	Hiện trạng
1	Già	Bác sĩ	Đa khoa	***Nghỉ hưu***
2	Trung niên	Giảng viên	Toán	***Đi làm***
3	Thanh niên	Sinh viên	Toán	***Đi học***
4	Thanh niên	Làm nông	-	***Đi làm***
5	Già	Giảng viên	Tin học	***Nghỉ hưu***
6	Trung niên	Bác sĩ	Răng	***Đi làm***

(yêu cầu sinh viên tự thực hiện)

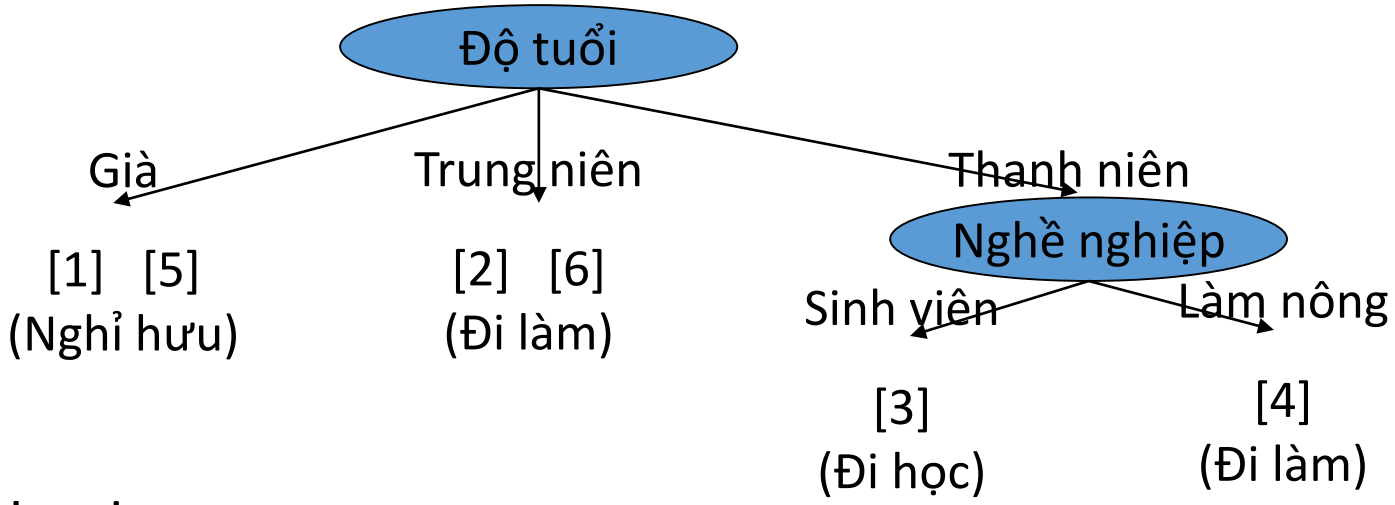
Phần 6

Xây dựng tập luật từ cây quyết định

Xây dựng tập luật từ cây quyết định

- Từ cây quyết định, có thể xây dựng tập luật suy dẫn bằng cách hình thành các luật lấy vế trái là các thuộc tính trên đường đi từ gốc, vế phải là thuộc tính kết luận
- Không thể làm ngược lại trong một số trường hợp (chuyển từ tập luật về cây quyết định)

Xây dựng tập luật từ cây quyết định



Tập luật thu được:

- Nếu "Độ tuổi" là "Già" thì "Nghỉ hưu"
- Nếu "Độ tuổi" là "Trung niên" thì "Đi làm"
- Nếu "Độ tuổi" là "Thanh niên" và "Nghề nghiệp" là "Sinh viên" thì "Đi học"
- Nếu "Độ tuổi" là "Thanh niên" và "Nghề nghiệp" là "Làm nông" thì "Đi làm"

Xét về khía cạnh nào đó thì tập luật này có thể xem như là quy luật của dữ liệu, bản thân con người cũng thường xuyên rút ra nhận xét như vậy khi quan sát thực tế.

Đặc điểm của cây quyết định

- Ưu điểm:
  - Dễ hiểu, đơn giản
  - Không cần chuẩn hóa dữ liệu
  - Xử lý được dữ liệu số và phi số
  - Trong suốt:
    - Có thể quan sát quá trình phát triển cây (khám phá dữ liệu)
    - Có thể quan sát quá trình ra quyết định (phân loại)
  - Có thể chuyển đổi thành luật
- Nhược điểm:
  - Không phù hợp với dữ liệu liên tục, phụ thuộc thời gian
  - Không tốt khi dữ liệu có quá nhiều phân lớp (và số lượng mẫu không đủ lớn và tốt)
  - Chi phí tính toán tương đối cao

Phần 7

Bài tập ứng dụng

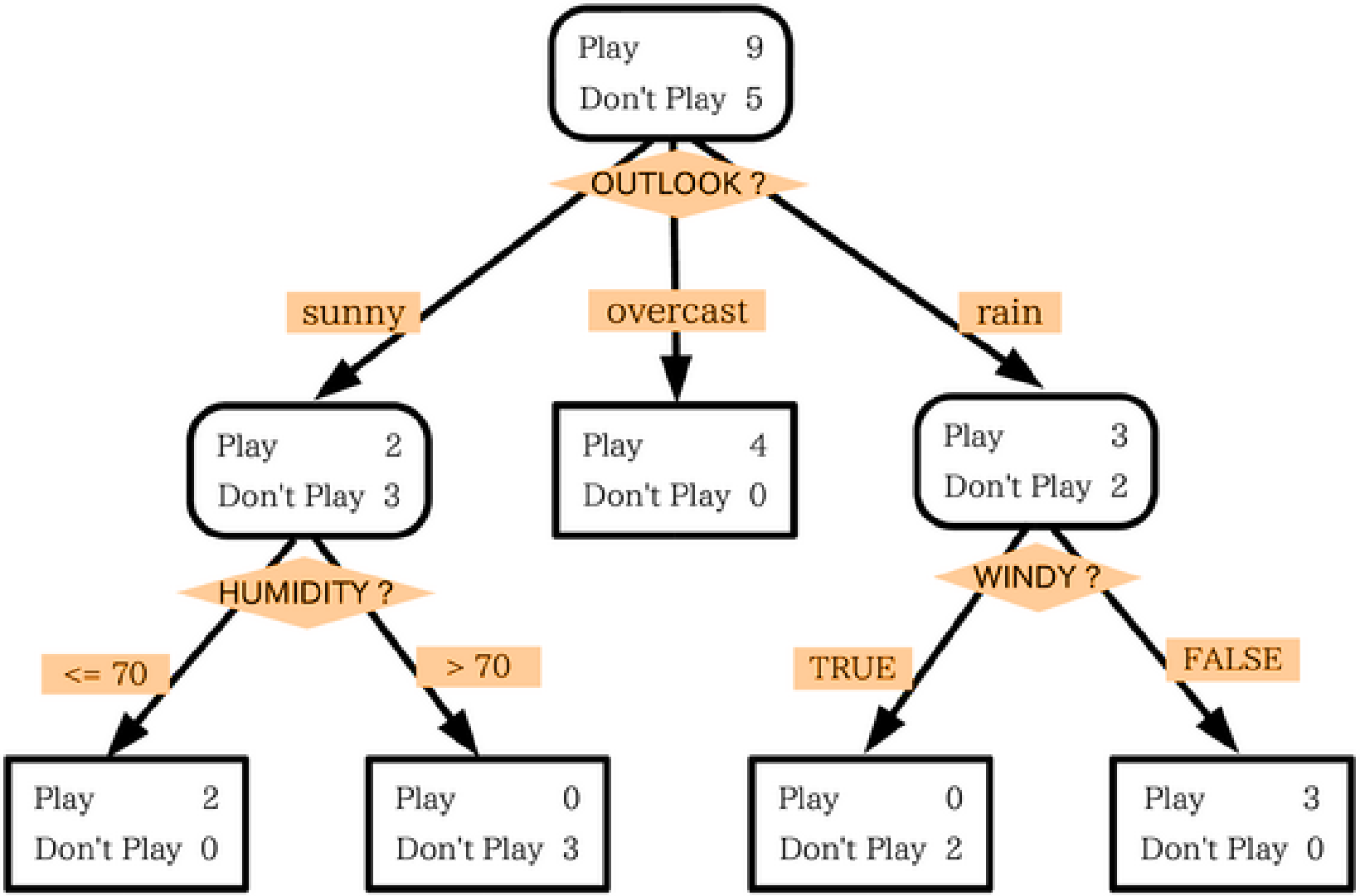
Quản lý sân golf: bài toán

- Một nhà quản lý sân golf thường phải ra quyết định xem cần bao nhiêu người phục vụ sân golf vào ngày hôm nay, số người phục vụ phụ thuộc vào số người đến chơi golf
- Nhà quản lý quan sát những người chơi golf và các thông số về thời tiết vào ghi chép vào sổ, các tham số sau:
  - Bầu trời (outlook): nắng (sunny) / mây (overcast) / mưa (rain)
  - Nhiệt độ (temperature): Độ F
  - Độ ẩm (humidity): số %, dưới 70% là khô
  - Gió mạnh (windy): có / không
  - Tình trạng có đến chơi golf hay không
- Dưa vào ghi chép của nhà quản lý hãy tìm quy luật đi chơi golf của khách hàng

Quản lý sân golf: dữ liệu

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

Quản lý sân golf: cây quyết định



Quản lý sân golf: quy luật và ứng dụng

- Như vậy có thể tạm rút kết luận (bộ luật):
  - Trời nhiều mây: Mọi người đều chơi golf
  - Trời nắng: Chỉ chơi nếu trời khô (ẩm <= 70%)
  - Trời mưa: Chỉ chơi nếu không có gió
- Ứng dụng: quản trị nhân lực
  - Trời nhiều mây: Thuê thêm phục vụ sân golf
  - Trời nắng + ẩm: Cho nghỉ bớt
  - Trời mưa + gió: Cho nghỉ bớt