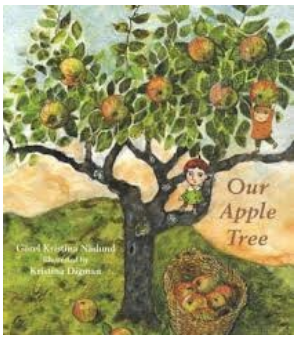
KHOA CNTT & TRUYỀN THÔNG
BM KHOA HỌC MÁY TÍNH

Phương pháp học cây quyết định
Decision Tree

PGS. TS. Đỗ Thanh Nghị
TS. Trần Nguyễn Minh Thư
tnmthu@ctu.edu.vn

Nội dung

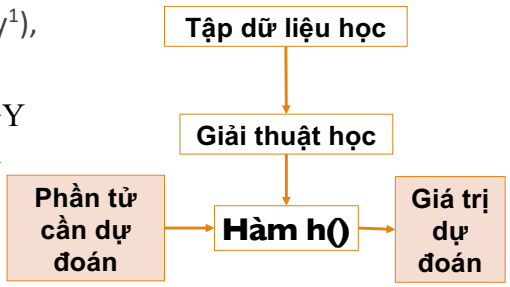
- **Học có giám sát**
- Giới thiệu về cây quyết định
- Giải thuật học của cây quyết định
- Kết luận và hướng phát triển



Học có giám sát

Từ tập dữ liệu huấn luyện $\{(X^1, y^1), (X^2,y^2),..,(X^m,y^m)$

- Tìm hàm h (hypothesis) X=>Y sao cho h(x) dự báo được y từ x



- **Y là giá trị liên tục**: sử dụng pp hồi quy (regression)
- **Y là giá trị rời rạc**: sử dụng pp phân lớp (classification)

5

5

Từ tập dữ liệu học/huấn luyện $\{ (x^1, y^1), (x^2,y^2),..,(x^m,y^m) \}$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	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

[See: Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997]

Chỉ ra thuộc tính? Nhãn/lớp của tập dữ liệu thời tiết trong bảng trên

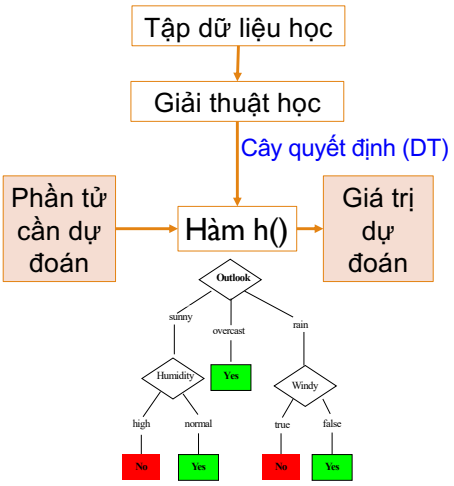
6

Phân loại học máy – học có giám sát

Từ tập dữ liệu huấn luyện { (x^1, y^1), $(x^2,y^2),..,(x^m,y^m)$ }

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes

- Tìm hàm h (hypothesis) X=>Y sao cho h(x) dự báo được y từ x

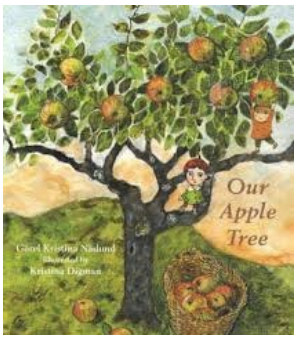


7

7

Nội dung

- **Học có giám sát**
- **Giới thiệu về cây quyết định**
- Giải thuật học của cây quyết định
- Kết luận và hướng phát triển



8

8

Cây quyết định

Cây quyết định là giải thuật học:

- kết quả sinh ra dễ diễn dịch **(if … then …)**
- khá đơn giản, nhanh, hiệu quả được sử dụng nhiều
- liên tục trong nhiều năm qua, cây quyết định được bình chọn là giải thuật được sử dụng nhiều nhất và thành công nhất
- giải quyết các vấn đề của phân loại, hồi quy
- làm việc cho **dữ liệu số và kiểu liệt kê**
- được ứng dụng thành công trong hầu hết các lãnh vực về phân tích dữ liệu, phân loại text, spam, phân loại gien, etc

9

9

Cây quyết định

Có rất nhiều giải thuật sẵn dùng

- ID3 (Quinlan 79)
- CART – Classification and Regression Trees (Brieman et al. 84)
- Assistant (Cestnik et al. 87)
- C4.5 (Quinlan 93)
- See5 (Quinlan 97)
- ...
- Orange (Demšar, Zupan 98-03)

10

Kỹ thuật DM thành công trong ứng dụng thực (2011)

Which methods/algorithms did you use for data analysis in 2011? [311 voters]	
Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %
Neural Nets (84)	27.0 %
Boosting (73)	23.5 %
Bayesian (68)	21.9 %
Bagging (63)	20.3 %
Factor Analysis (58)	18.7 %
Anomaly/Deviation detection (51)	16.4 %
Social Network Analysis (44)	14.2 %
Survival Analysis (29)	9.32 %
Genetic algorithms (29)	9.32 %
Uplift modeling (15)	4.82 %

Top 10 DM algorithms (2015)

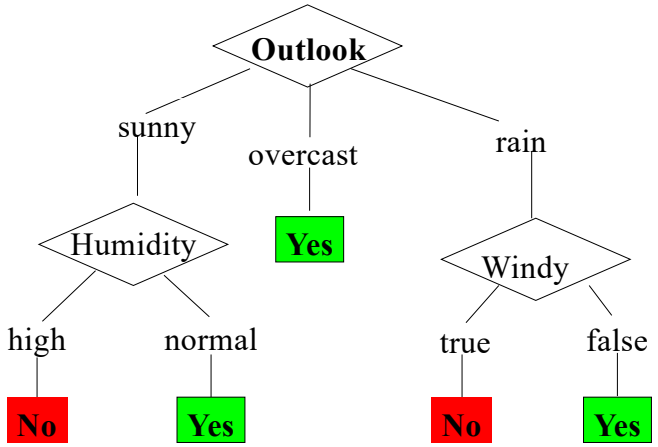
Here are the algorithms:

- 1. C4.5
- 2. k-means
- 3. Support vector machines
- 4. Apriori
- 5. EM
- 6. PageRank
- 7. AdaBoost
- 8. kNN
- 9. Naive Bayes
- 10. CART

11

11

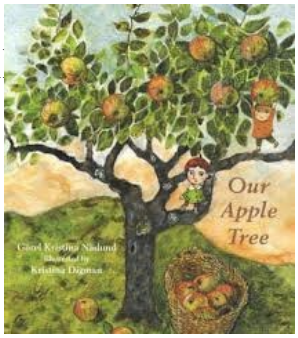
Ví dụ cây quyết định



12

Nội dung

- **Học có giám sát**
- **Giới thiệu về cây quyết định**
- **Giải thuật học của cây quyết định**
- Kết luận và hướng phát triển



13

13

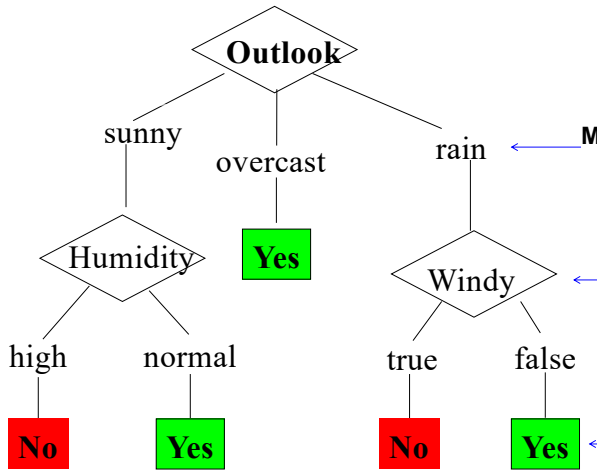
Cây quyết định

- Nút trong : được tích hợp với điều kiện để kiểm tra rẽ nhánh
- Nút lá : được gán nhãn tương ứng với lớp của dữ liệu
- 1 nhánh : trình bày cho dữ liệu thỏa mãn điều kiện kiểm tra, ví dụ : age < 25.
- ở mỗi nút, 1 thuộc tính được chọn để phân hoạch dữ liệu học sao cho tách rời các lớp tốt nhất có thể
- Một luật quyết định có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá.
- Dữ liệu mới đến được phân loại bằng cách duyệt từ nút gốc của cây cho đến khi đụng đến nút lá, từ đó rút ra lớp của đối tượng cần xét

14

14

Ví dụ cây quyết định



Một luật quyết định có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá.

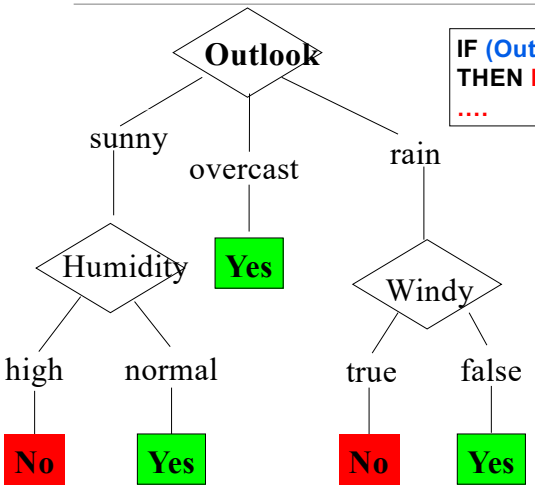
Mỗi nhánh tương ứng với một giá trị của thuộc tính

Mỗi nút mang một thuộc tính (biến độc lập)

Mỗi nút lá là một lớp (biến phụ thuộc)

15

Cây quyết định cho tập dữ liệu weather, dựa trên các thuộc tính (Outlook, Temp, Humidity, Windy)



IF (Outlook=sunny) and (Humidity= high)
THEN Play=No
....

Một luật quyết định có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá.

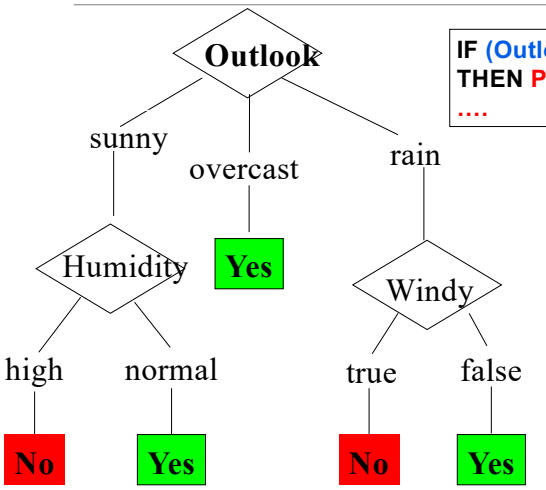
Dữ liệu mới đến được phân loại bằng cách duyệt từ nút gốc của cây cho đến khi đụng đến nút lá, từ đó rút ra lớp của đối tượng cần xét

16

16

A: Humidity=high, Windy=true, Outlook=overcast

B: Humidity=high, outlook=rain, windy=false



IF (Outlook=sunny) and (Humidity= high)
THEN Play=No
....

Một luật quyết định có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá.

Dữ liệu mới đến được phân loại bằng cách duyệt từ nút gốc của cây cho đến khi đụng đến nút lá, từ đó rút ra lớp của đối tượng cần xét

17

17

Giải thuật cây quyết định

- Xây dựng cây Top-down
  - bắt đầu nút gốc, tất cả các dữ liệu học ở nút gốc
  - Nếu dữ liệu tại 1 nút có cùng lớp -> nút lá (nhãn của nút chính là nhãn của các phần tử thuộc nút lá); Nếu dữ liệu ở nút chứa các phần tử có lớp rất khác nhau (không thuần nhất) thì phân hoạch dữ liệu một cách đệ quy bằng việc chọn 1 thuộc tính để thực hiện phân hoạch **tốt nhất** có thể => kết quả thu được cây nhỏ nhất

18

18

Giải thuật cây quyết định

Chọn thuộc tính phân hoạch

- Tại mỗi nút, các thuộc tính được đánh giá dựa trên phân tách dữ liệu học **tốt nhất** có thể
- Thuộc tính nào tốt ?
  - cho ra kết quả là cây nhỏ nhất
  - Thường dựa trên giá trị heuristics để tìm được các thuộc tính sinh ra các nút "purest" (thuần khiết)

19

19

Giải thuật cây quyết định

Chọn thuộc tính phân hoạch

- Tại mỗi nút, các thuộc tính được đánh giá dựa trên phân tách dữ liệu học **tốt nhất** có thể
- Việc đánh giá tốt hay không dựa trên các heuristics
  - **độ lợi thông tin** (chọn thuộc tính có chỉ số lớn)- information gain (ID3/C4.5 - Quinlan)
  - Tỉ số độ lợi thông tin (information gain ratio)
  - **chỉ số gini** (chọn thuộc tính có chỉ số nhỏ)- gini index (CART - Breiman)

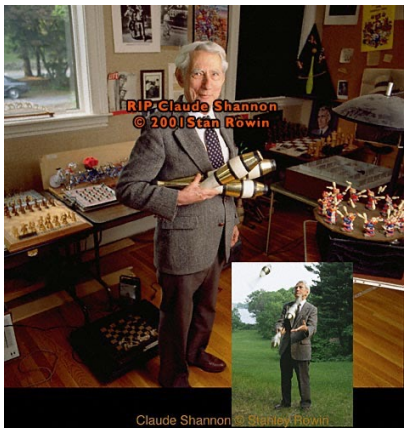
20

20

*Claude Shannon

Born: 30 April 1916
Died: 23 February 2001

"Father of information theory"



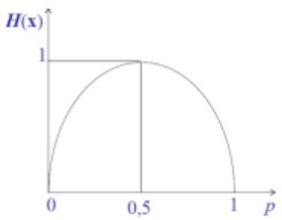
21

Entropy

Entropy là một đại lượng toán học dùng để đo lượng thông tin không chắc chắn (hay lượng ngẫu nhiên) của một sự kiện hay một phân phối ngẫu nhiên cho trước

Entropy – uncertainty measure

Entropy luôn >=0
- Entropy = 0?
- Entropy = 1?

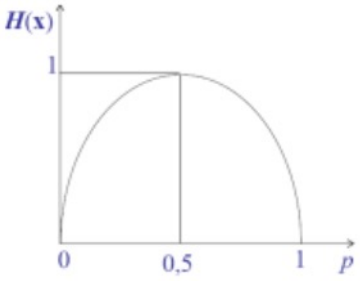


$Info(D) = entropy(p_1,p_2,..,p_n) = -p_1logp_1 - p_2logp_2 - p_nlogp_n$

p_i: xác suất mà phần tử trong dữ liệu D thuộc lớp C_i

22

Entropy



p: # phần tử có nhãn +

n: # phần tử có nhãn -

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

p = n = 6;

Entropy (0.5,0.5) = -0.5 $\log_2(0.5)$ –0.5 $\log_2(0.5)$ = 1

Entroypy = 1
(cực đại khi xác suất xuất hiên của các thành phần bằng nhau 50/50

23

23

Độ lợi thông tin

- Độ đo hỗn loạn trước khi phân hoạch trừ cho sau khi phân hoạch
- thông tin được đo lường bằng *bits*
  - cho 1 phân phối xác suất, thông tin cần thiết để dự đoán 1 sự kiện là *entropy*
- công thức tính entropy – độ hỗn loạn thông tin trước khi phân hoạch

  Info(D) = entropy(p_1,p_2,..,p_n) = -$p_1$logp_1 – $p_2$logp_2 - $p_n$logp_n
- p_i: xác suất mà phần tử trong dữ liệu D thuộc lớp C_i

24

24

Độ lợi thông tin

- Độ hỗn loạn thông tin **trước** khi phân hoạch

$$Info(D) = entropy(p_1,p_2,..,p_n) = -p_1 log p_1 - p_2 log p_2 - p_n log p_n$$

p_i: xác suất mà phần tử trong dữ liệu **D** thuộc lớp **C_i**

- Độ hỗn loạn thông tin **sau** khi phân hoạch

$$Info_A(D) = D_1/D * Info(D_1) + D_2/D * Info(D_2) + \ldots + D_v/D * Info(D_v)$$

Thuộc tính **A** phân hoạch dữ liệu **D** thành **v** phần

- Độ lợi thông tin khi chọn thuộc tính **A** phân hoạch dữ liệu **D** thành **v** phần

$$Gain(A) = Info(D) - Info_A(D)$$

25

25

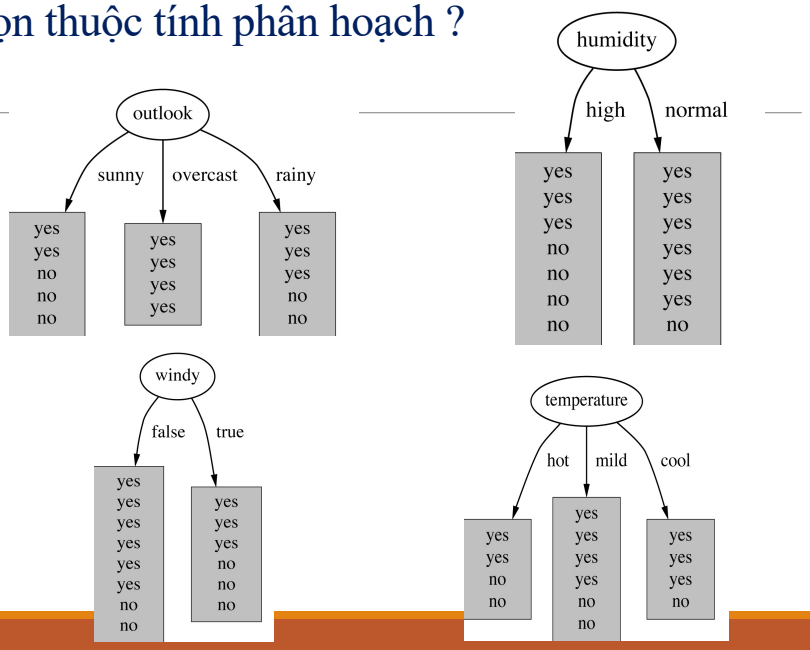
Giải thuật cây quyết định

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

26

26

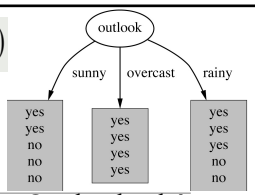
Chọn thuộc tính phân hoạch ?



27

27

$$Info_A(D) = D_1/D * Info(D_1) + D_2/D * Info(D_2) \ldots + D_v/D * Info(D_v)$$



Ví dụ : thuộc tính outlook

- Độ hỗn loạn thông tin sau khi chọn thuộc tính A= Outlook phân hoạch dữ liệu D thành v=3 phần
- **"Outlook" = "Sunny":**

$$info([2,3]) = entropy(2/5,3/5) = -2/5\log(2/5) - 3/5\log(3/5) = 0.971\,bits$$

- **"Outlook" = "Overcast":**

$$info([4,0]) = entropy(1,0) = -1\log(1) - 0\log(0) = 0\,bits$$

chú ý : log(0) không xác định nhưng 0*log(0) là 0

- **"Outlook" = "Rainy":**

$$info([3,2]) = entropy(3/5,2/5) = -3/5\log(3/5) - 2/5\log(2/5) = 0.971\,bits$$

- **thông tin của thuộc tính outlook:**

$$info([2,3],[4,0],[3,2]) = (5/14)\times 0.971 + (4/14)\times 0 + (5/14)\times 0.971 = 0.693\,bits$$

28

28

Ví dụ : thuộc tính outlook

- Độ hỗn loạn thông tin trước khi phân hoạch

$$info([9,5]) = entropy(9/14,5/14) = -9/14\log(9/14) - 5/14\log(5/14) = 0.940 \text{ bits}$$

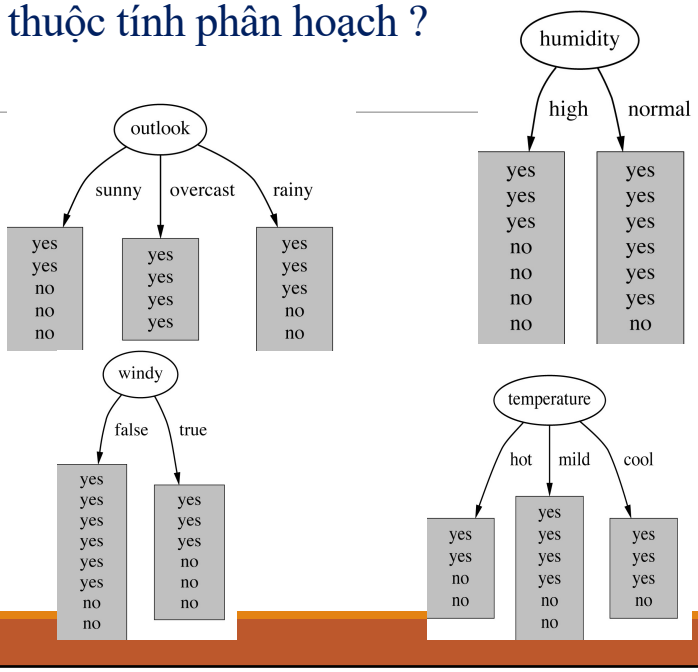
- độ lợi thông tin của outlook
(trước khi phân hoạch) – (sau khi phân hoạch)

$$gain("Outlook") = info([9,5]) - info([2,3],[4,0],[3,2]) = 0.940 - 0.693$$
$$= 0.247 \text{ bits}$$

29

29

Chọn thuộc tính phân hoạch ?



30

30

Thuộc tính humidity

- "Humidity" = "High":

$$\text{info}([3,4]) = \text{entropy}(3/7,4/7) = -3/7\log(3/7) - 4/7\log(4/7) = 0.985 \text{ bits}$$

- "Humidity" = "Normal":

$$\text{info}([6,1]) = \text{entropy}(6/7,1/7) = -6/7\log(6/7) - 1/7\log(1/7) = 0.592 \text{ bits}$$

- thông tin của thuộc tính humidity

$$= 0.788 \text{ bits}$$

$$\text{info}([3,4],[6,1]) = (7/14)\times 0.985 + (7/14)\times 0.592$$

- **độ lợi thông tin của thuộc tính humidity**

$$\text{info}([9,5]) - \text{info}([3,4],[6,1]) = 0.940 - 0.788 = 0.152$$

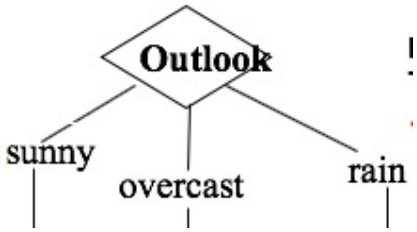
31

31

Độ lợi thông tin

- độ lợi thông tin của các thuộc tính

(trước khi phân hoạch) – (sau khi phân hoạch)



$$\text{gain}(\text{"Temperature"}) = 0.029 \text{ bits}$$

$$\text{gain}(\text{"Windy"}) = 0.048 \text{ bits}$$

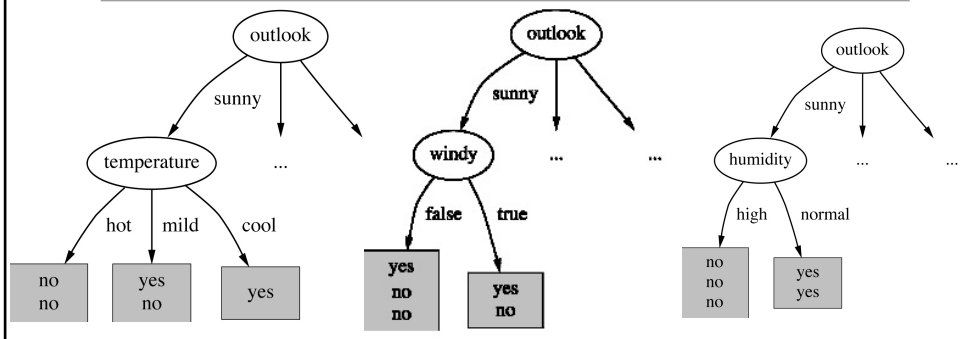
$$\text{gain}(\text{"Outlook"}) = 0.247 \text{ bits}$$

$$\text{gain}(\text{"Humidity"}) = 0.152 \text{ bits}$$

32

32

Tiếp tục phân hoạch dữ liệu

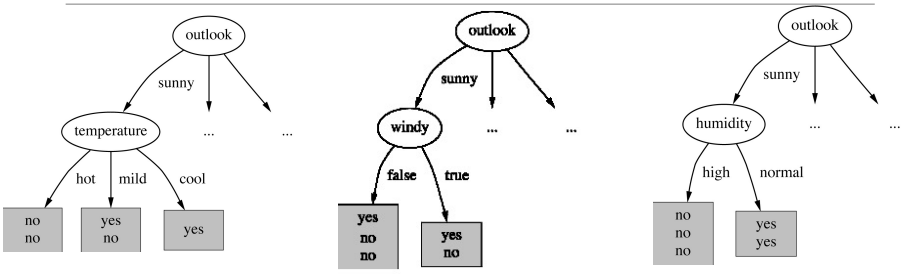


$\textbf{Info}_{\text{TEMP KHI OUTLOOK=SUNNY}}([0,2],[1,1],[1,0]) = \frac{2}{5}\left(-\frac{0}{2}\log\frac{0}{2} - \frac{2}{2}\log\frac{2}{2}\right)$
$+ \frac{2}{5}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) + \frac{1}{5}\left(-\frac{1}{1}\log\frac{1}{1} - \frac{0}{1}\log\frac{0}{1}\right) = 0+0.4+0$

33

33

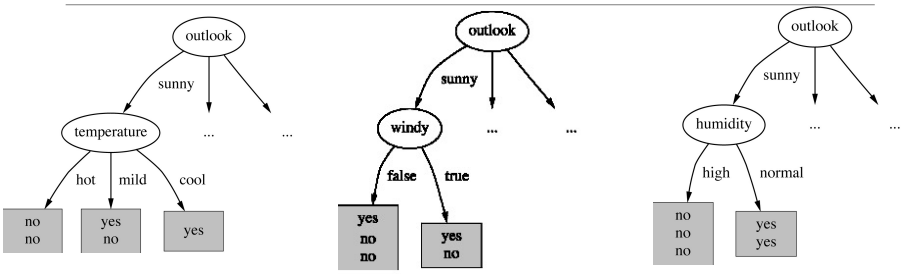
Tiếp tục phân hoạch dữ liệu



34

34

Tiếp tục phân hoạch dữ liệu



$$gain("Humidity") = 0.971\,bits$$

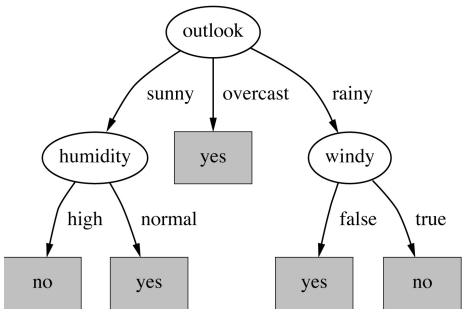
$$gain("Temperature") = 0.571\,bits$$

$$gain("Windy") = 0.020\,bits$$

35

35

Kết quả



- chú ý : có thể có nút lá không thuần khiết

  ⇒ phân hoạch dừng khi dữ liệu không thể phân hoạch, nhãn được gán cho lớp lớn nhất chứa trong nút lá

36

36