KHOA CNTT & TRUYỀN THÔNG
BM KHOA HỌC MÁY TÍNH

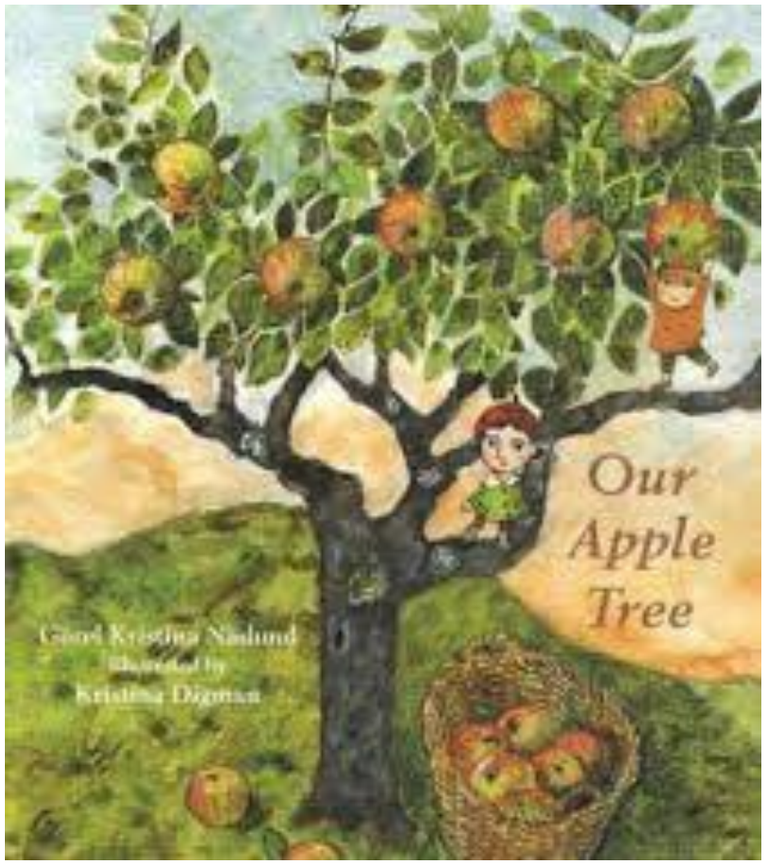
Phương pháp học cây quyết định

Decision Tree

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

Nội dung

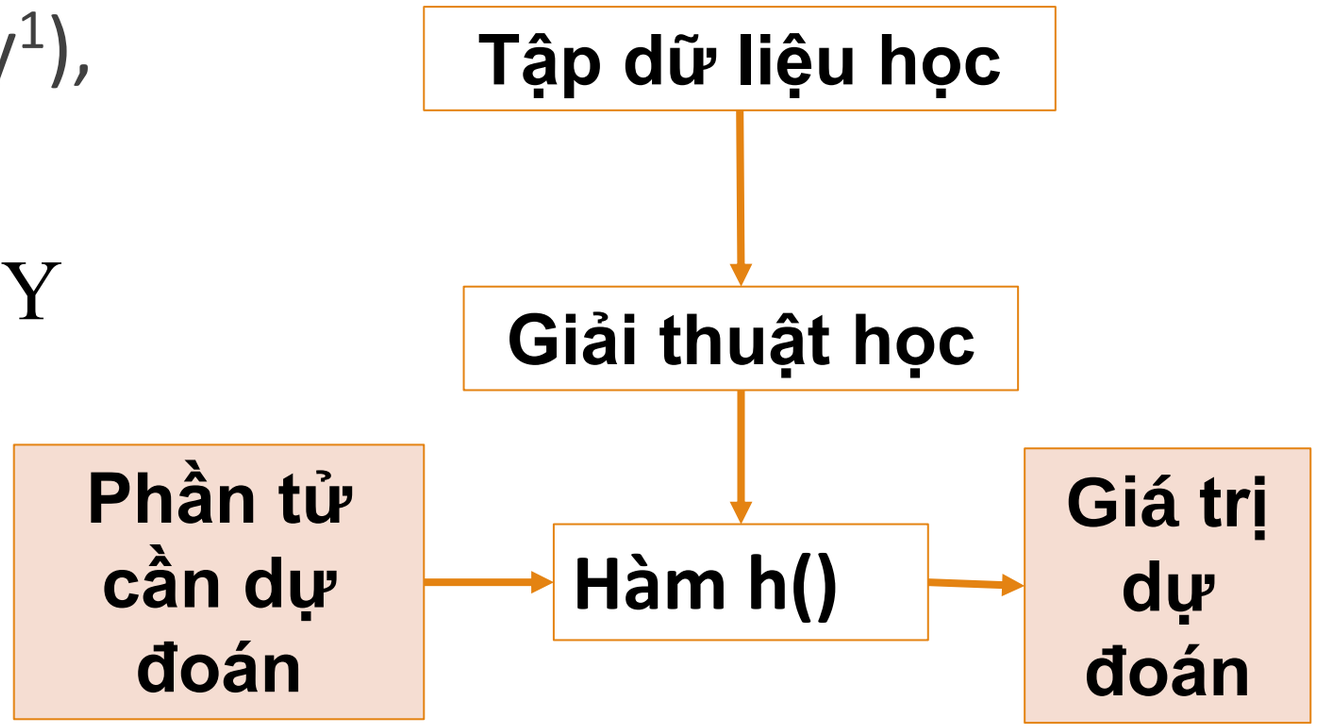
- **Học có giám sát**
- Giới thiệu về cây quyết định
- Giải thuật học của cây quyết định
- Kết luận và hướng phát triển



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



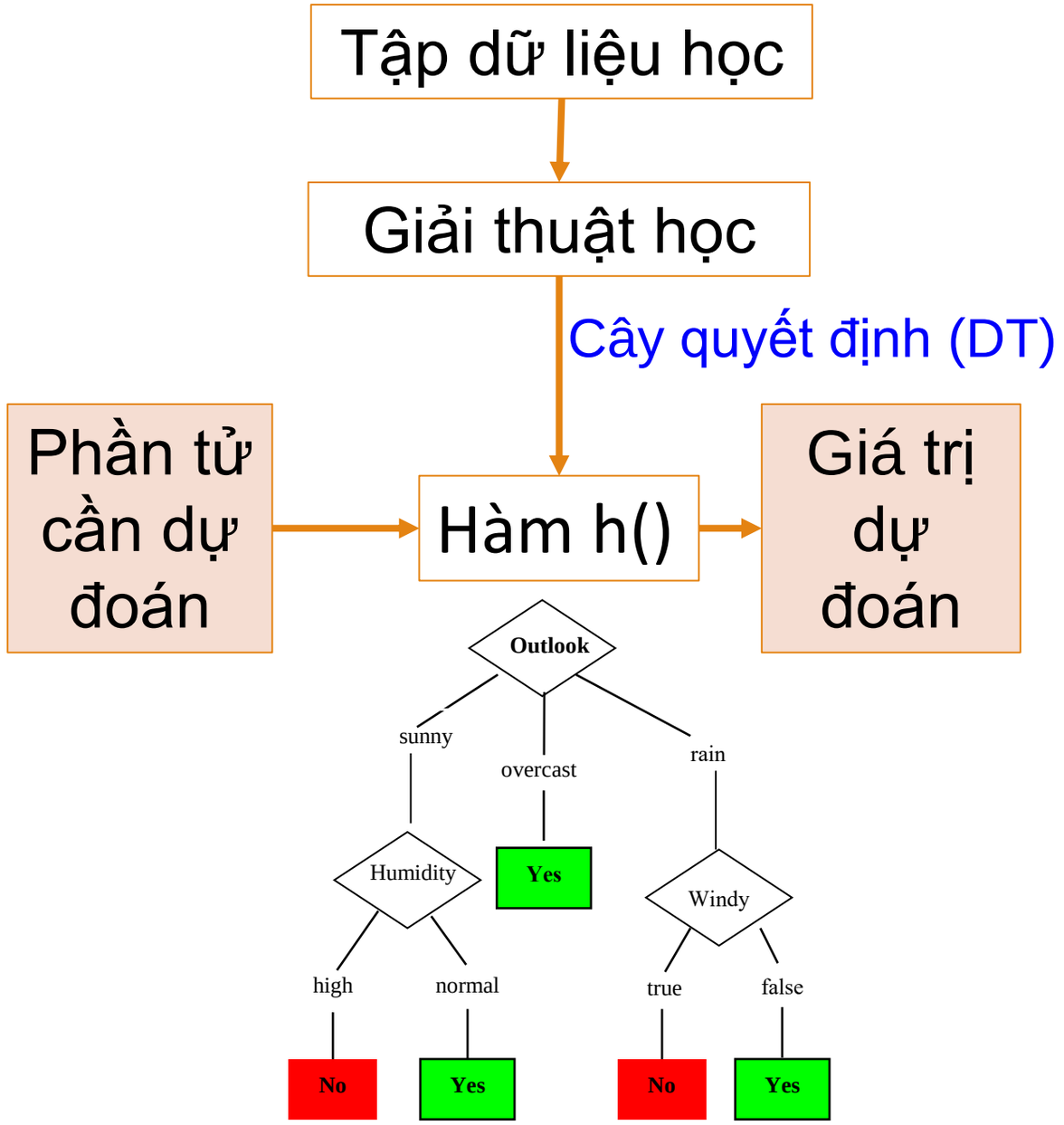
- **Y là giá trị liên tục**: sử dụng pp hồi quy (regression)
- **Y là giá trị rời rạc**: sử dụng pp phân lớp (classification)

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



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

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)

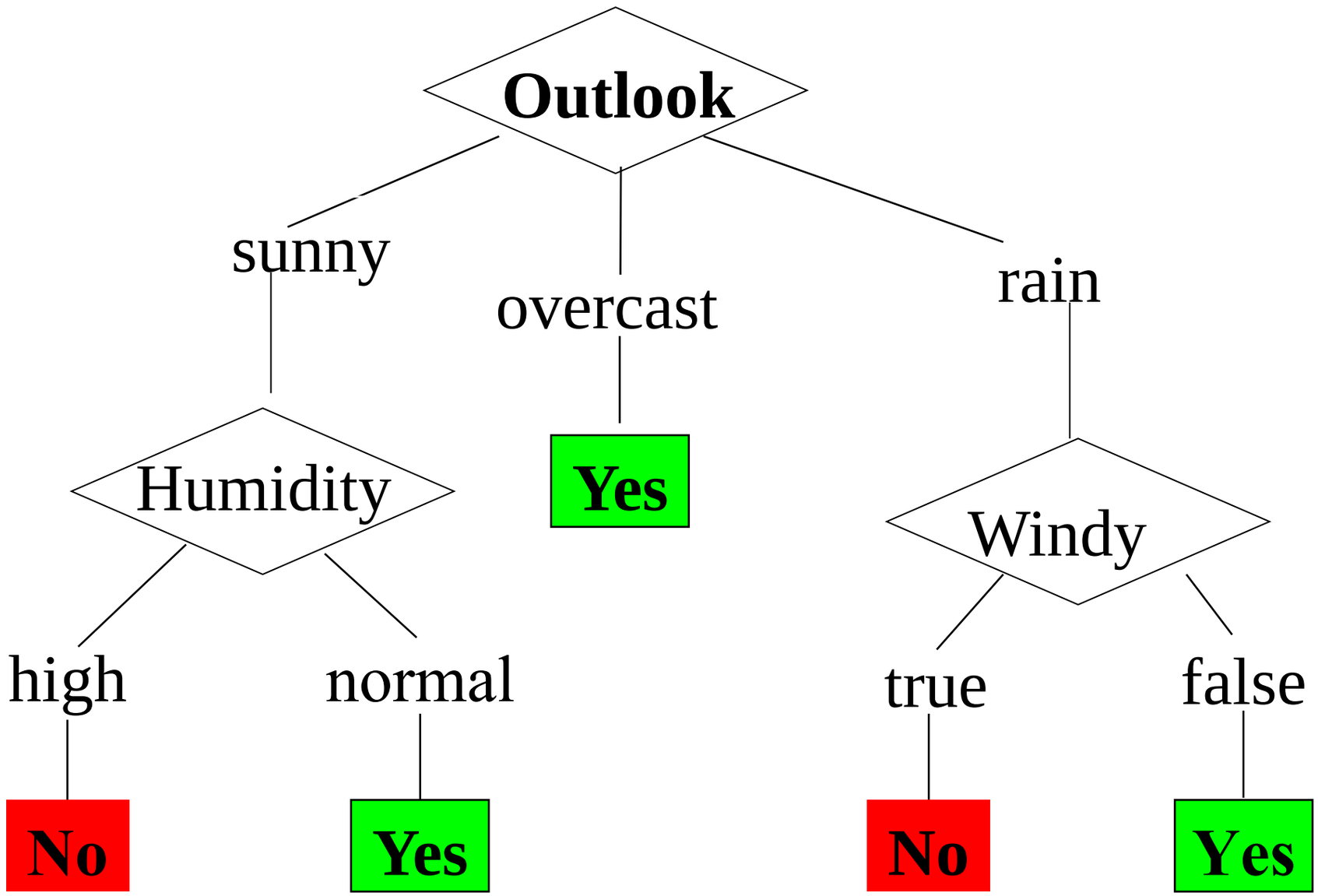
Kỹ thuật DM thành công trong ứng dụng thực

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

Ví dụ cây quyết định

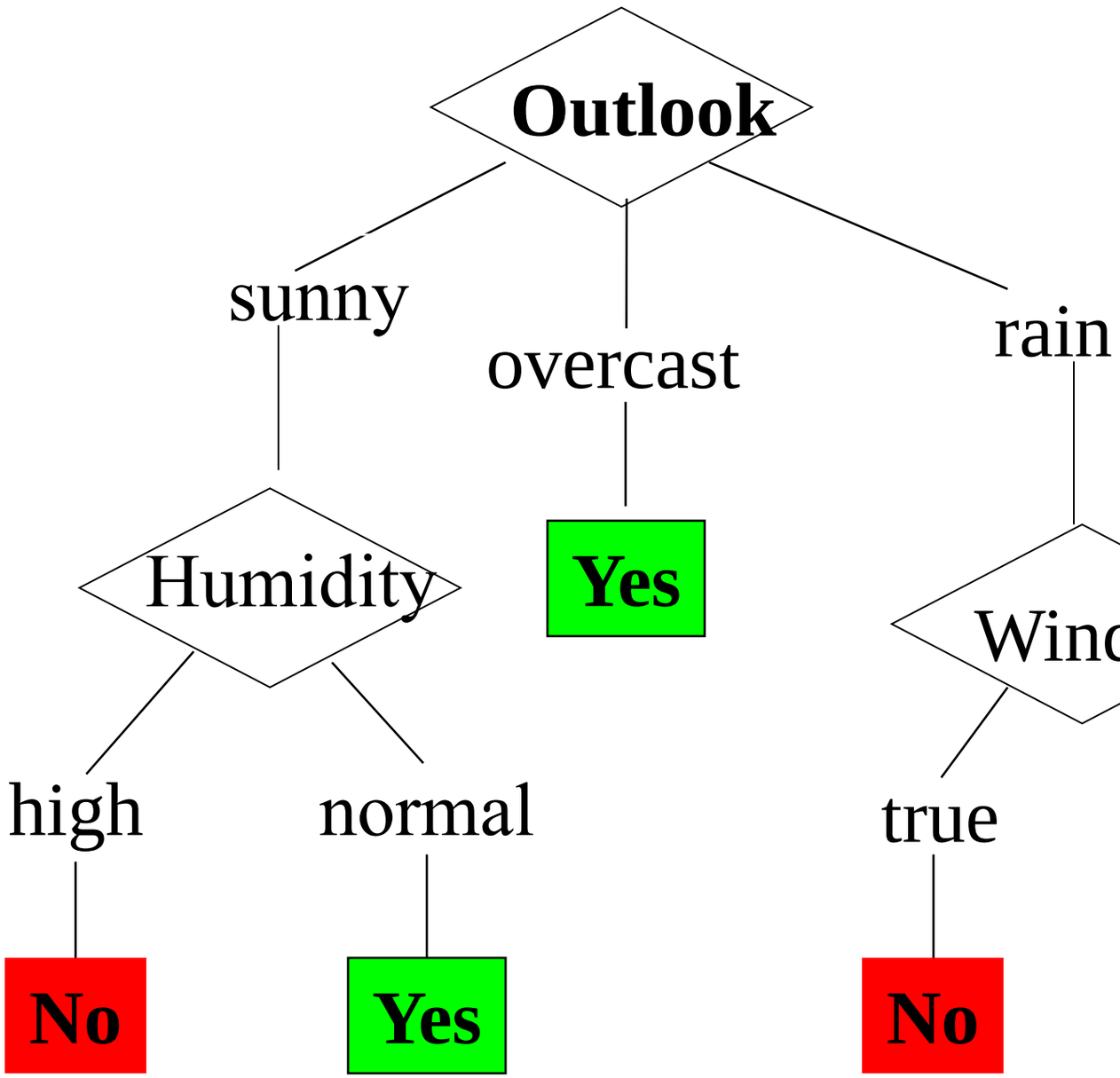


Cây quyết định

- 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

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.

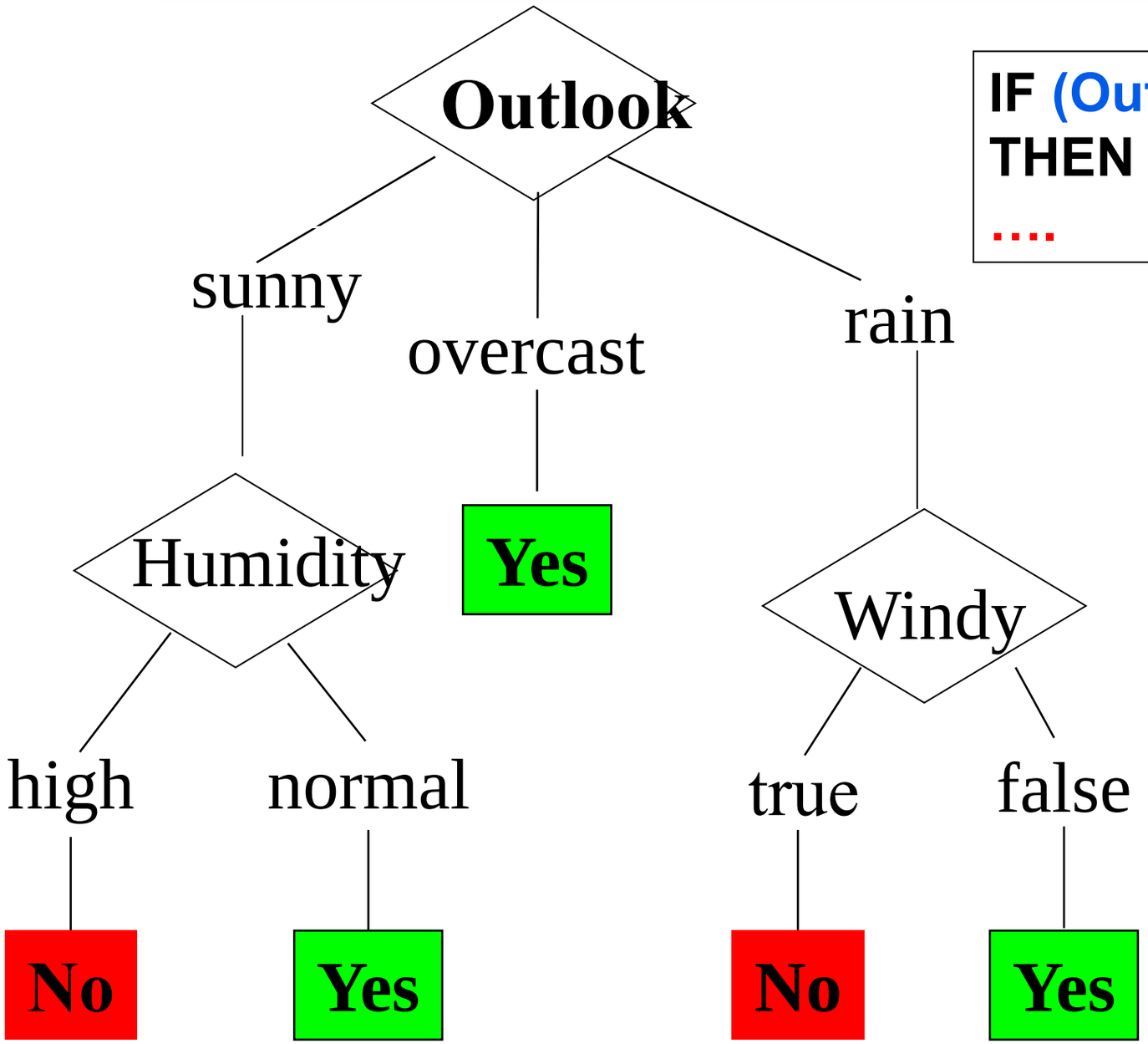


Mỗi nhánh tương ứng với một giá trị của thuộc tính

Mỗi nút mang một thuộc tính (biến độc lập)

Mỗi nút lá là một lớp (biến phụ thuộc)

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
….

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

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)

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)

*Claude Shannon

Born: 30 April 1916
Died: 23 February 2001

"Father of information theory"

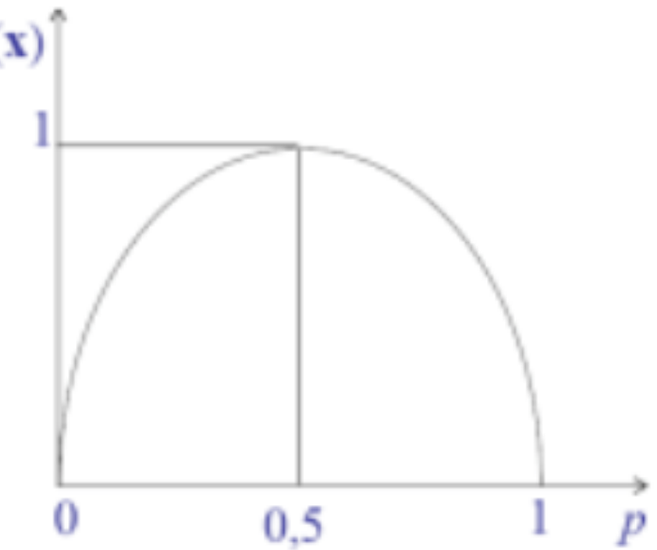
Entropy

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

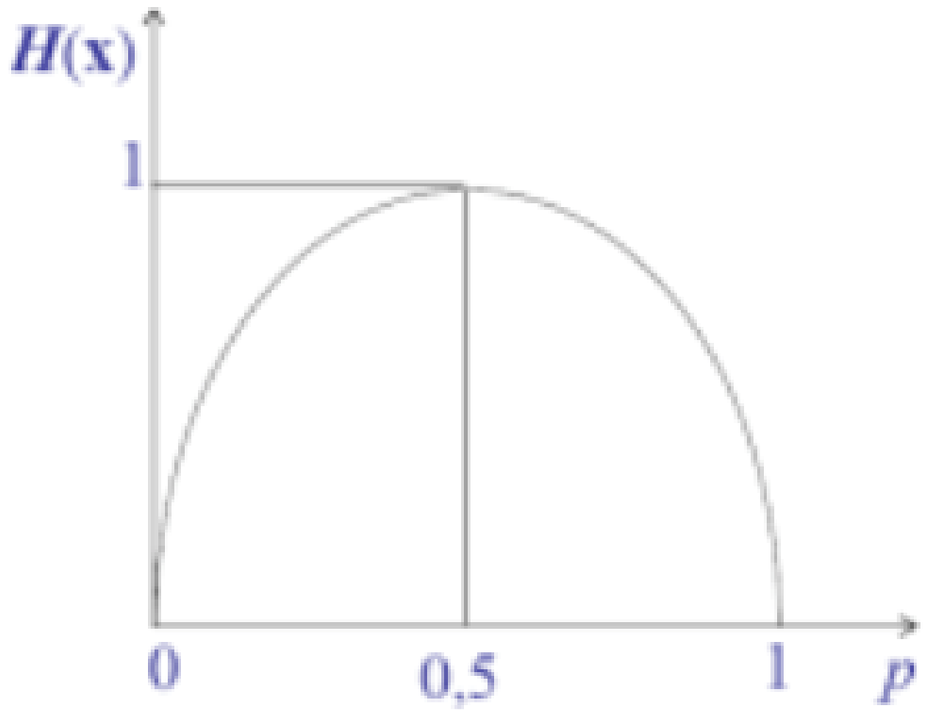
- Entropy = 0?
- Entropy = 1?



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

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

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

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

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

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

Độ 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

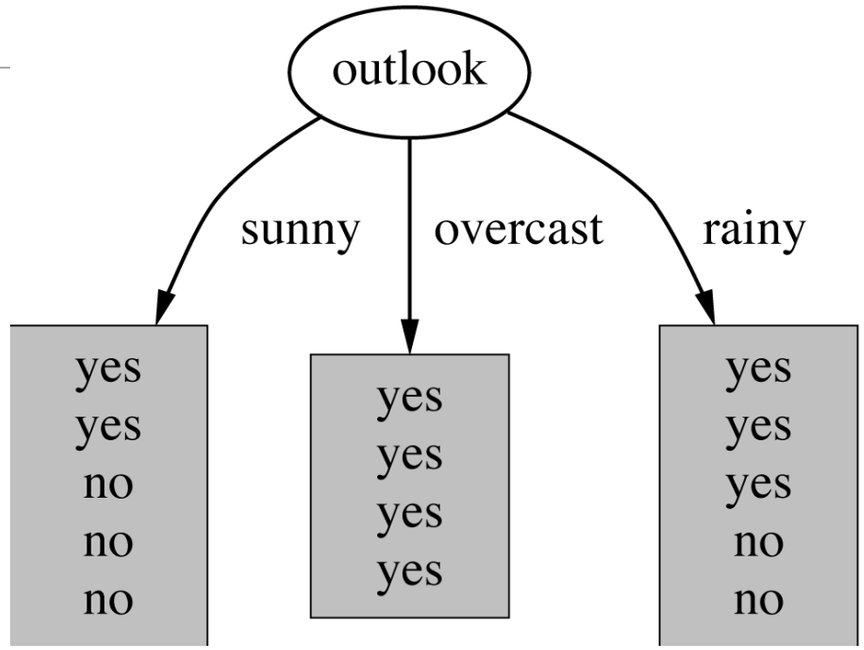
- Độ lợi thông tin khi chọn thuộc tính **A** phân hoạch dữ liệu **D** thành **v** phần

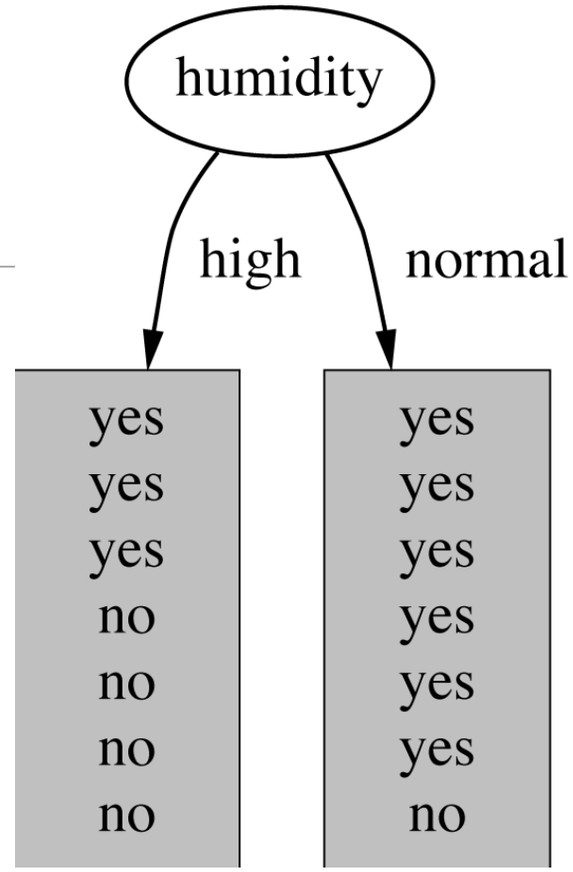
$$Gain(A) = Info(D) - Info_A(D)$$

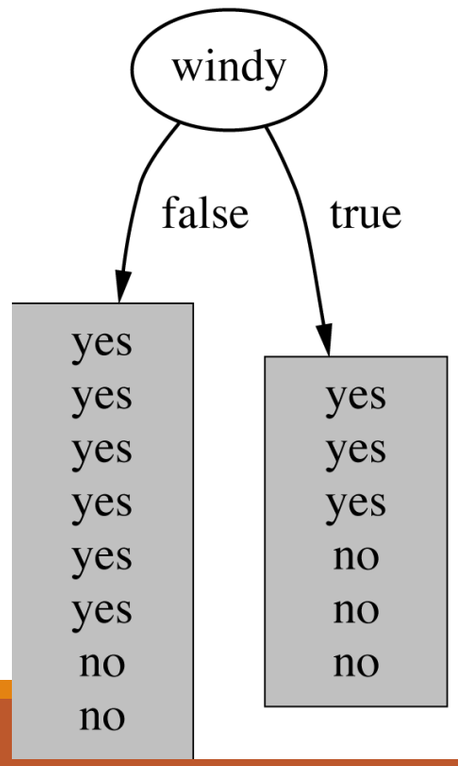
Giải thuật cây quyết định

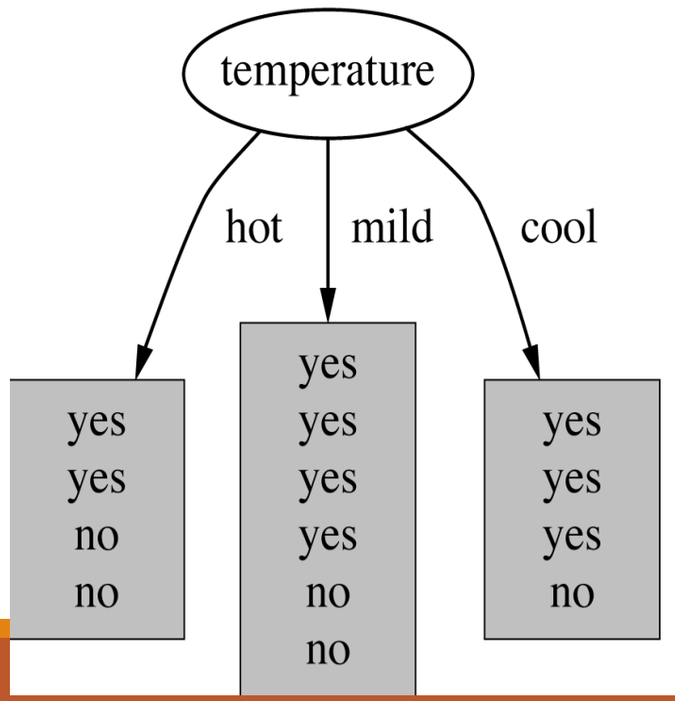
Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes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

Chọn thuộc tính phân hoạch ?

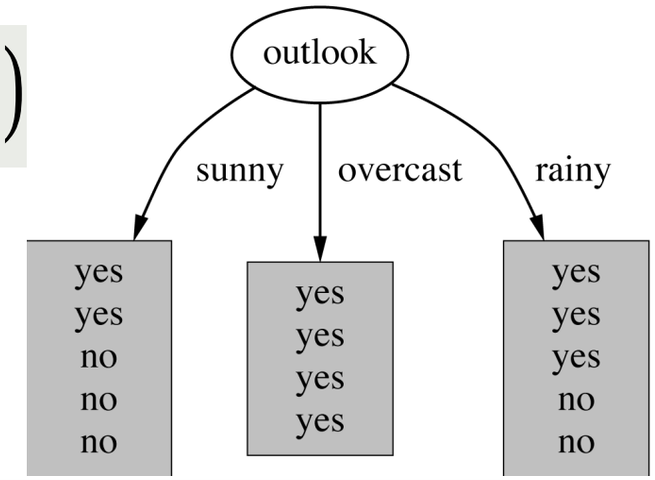








$$Info_A(D) = D_1/D * Info(D_1) + D_2/D * Info(D_2) \square + D_v/D * Info(D_v)$$



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":**

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

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

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

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

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

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

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

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

Ví dụ : thuộc tính outlook

- Độ hỗn loạn thông tin trước khi phân hoạch

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

- độ lợi thông tin của outlook

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

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

$$= 0.247 \text{ bits}$$

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

$$\text{info}([3,4],[6,1]) = (7/14)\times 0.985 + (7/14)\times 0.592 = 0.788\,\text{bits}$$

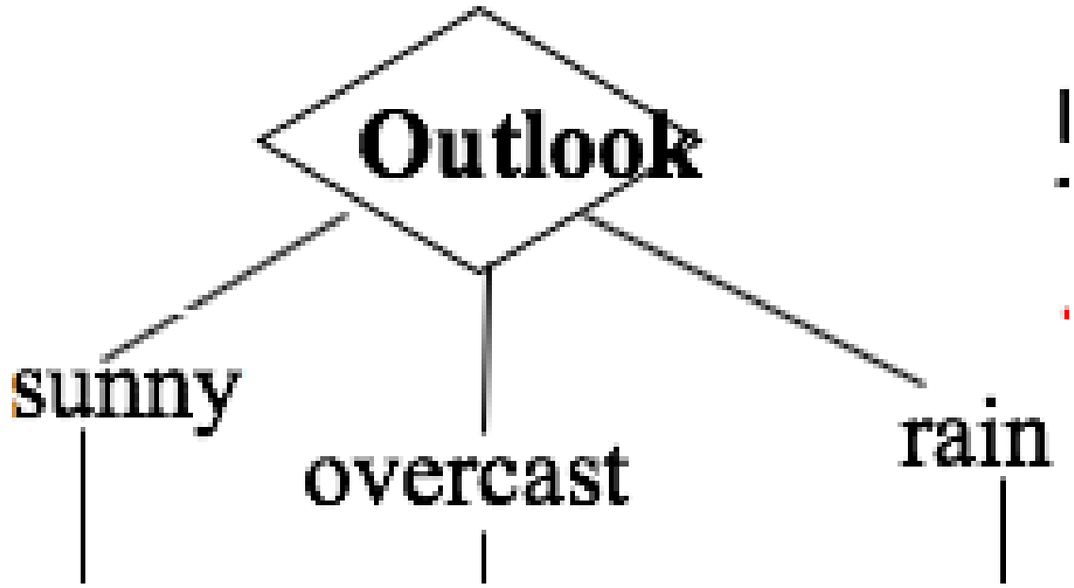
- **độ lợi thông tin của thuộc tính humidity**

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

Độ lợi thông tin

- độ lợi thông tin của các thuộc tính

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



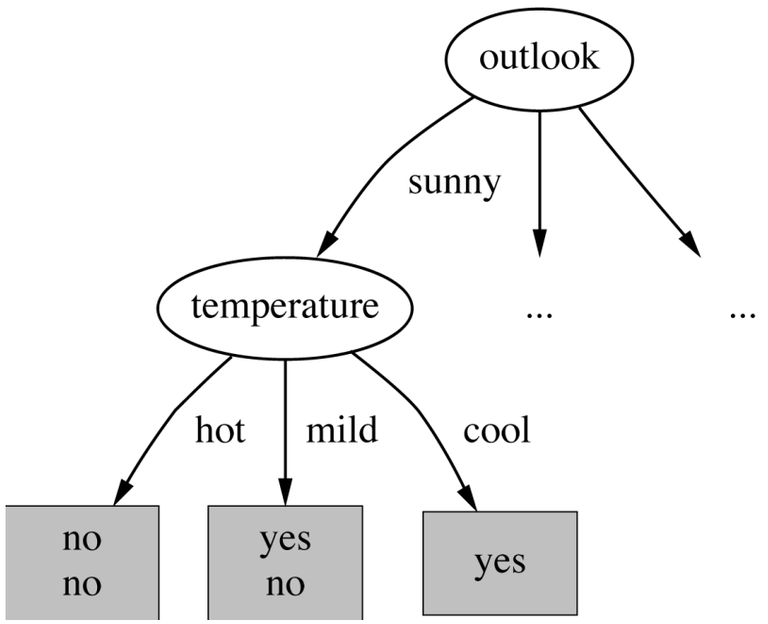
$$\text{gain("Outlook")} = 0.247 \text{ bits}$$

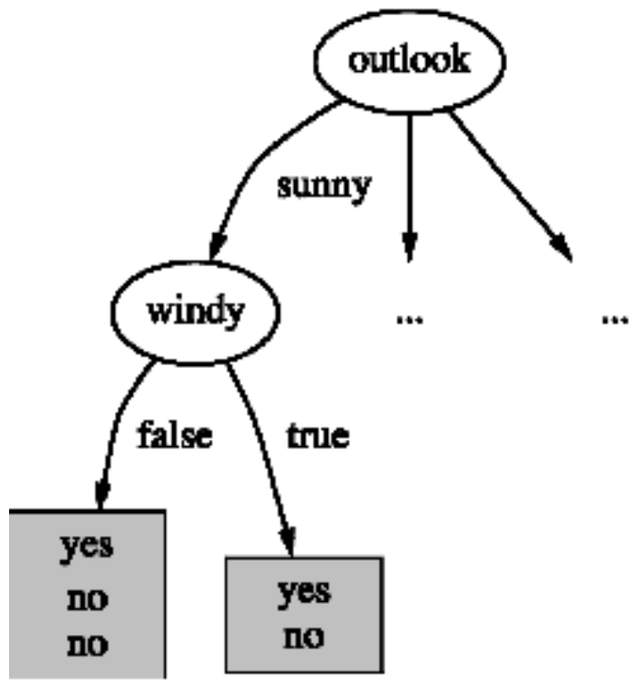
$$\text{gain("Humidity")} = 0.152 \text{ bits}$$

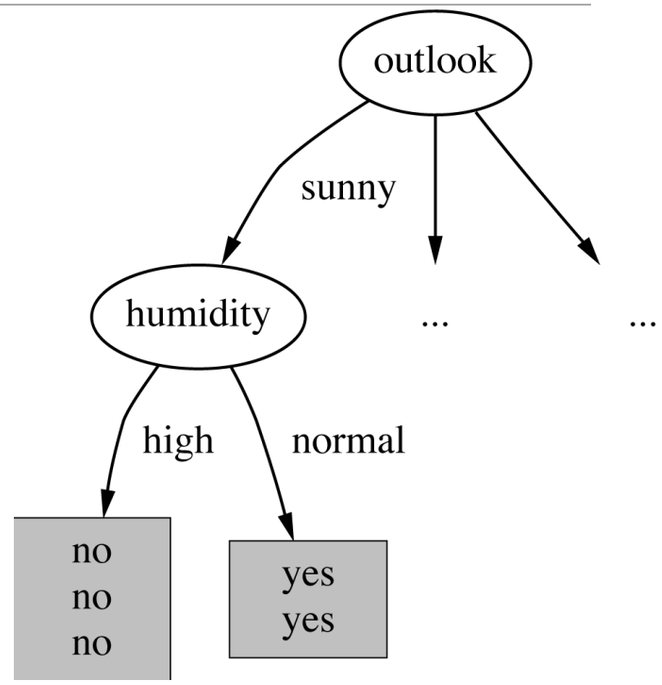
$$\text{gain("Temperature")} = 0.029 \text{ bits}$$

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

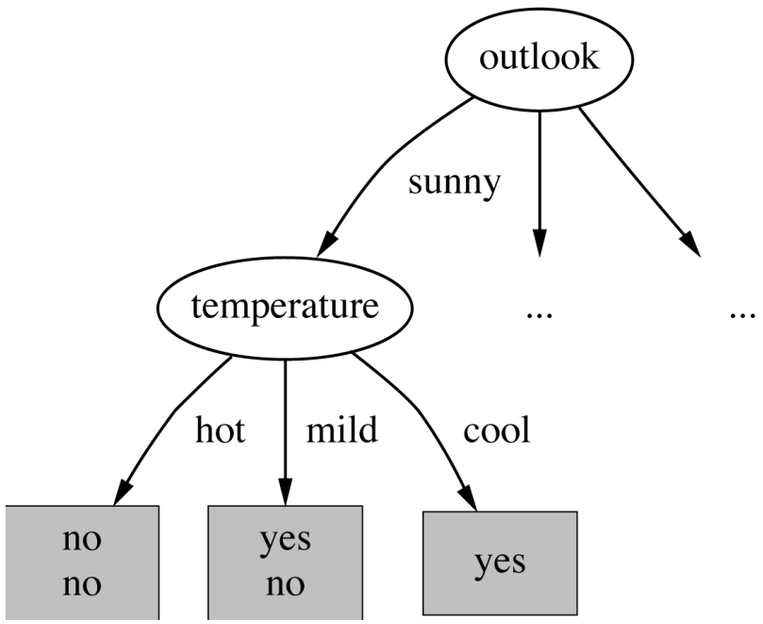
Tiếp tục phân hoạch dữ liệu

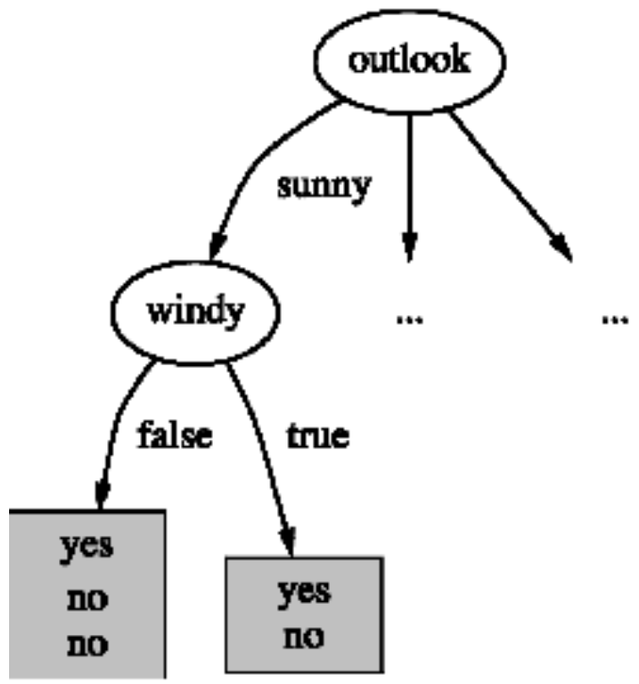


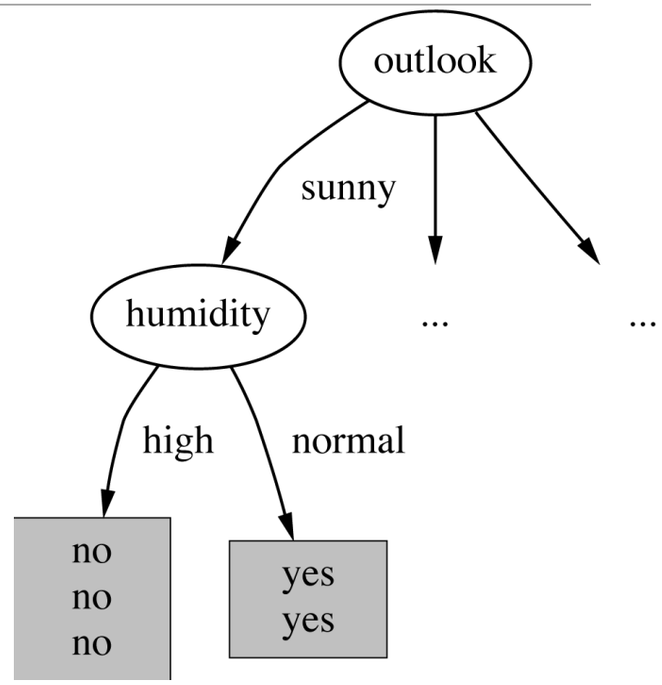




Tiếp tục phân hoạch dữ liệu





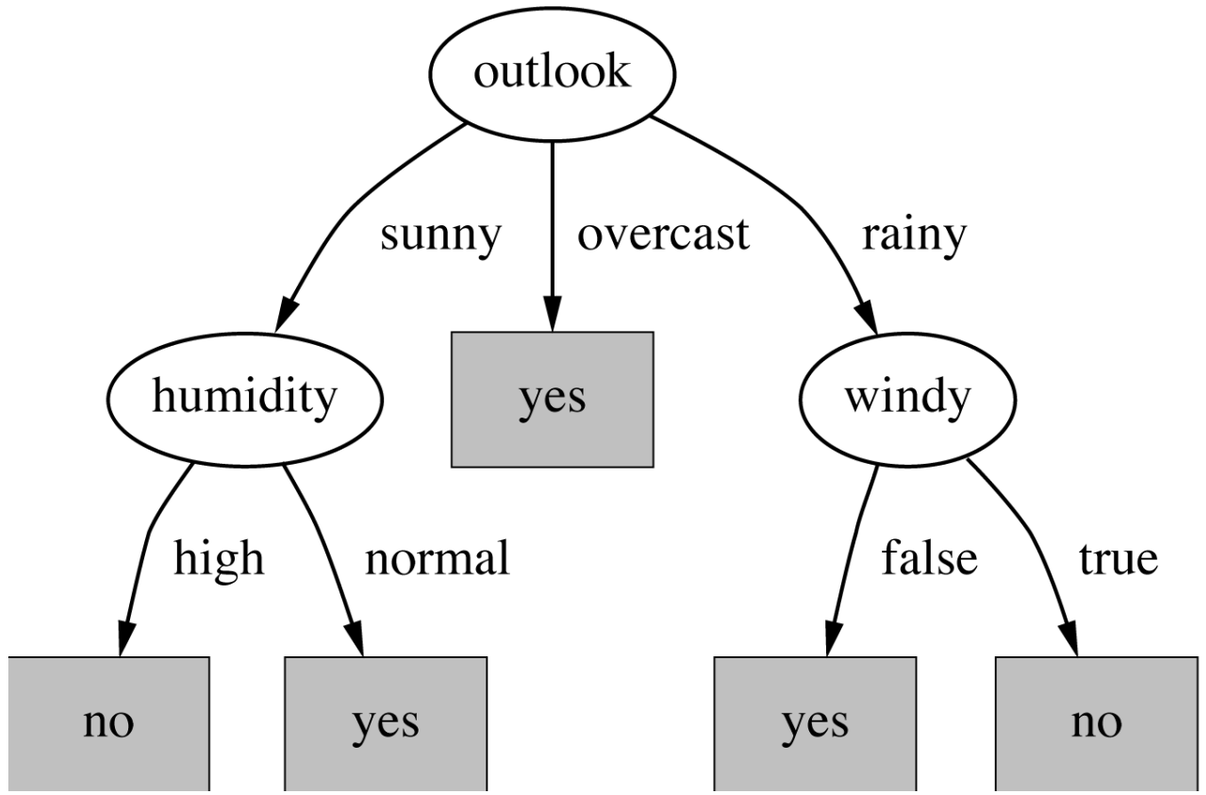


$$gain("Humidity") = 0.971\,bits$$

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

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

Kết quả



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

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

Chỉ số gini (CART)

- nếu dữ liệu T có n lớp, chỉ số gini(T) được định nghĩa như sau :

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

p_j là xác suất của lớp j trong T

- gini(T) là nhỏ nhất nếu những lớp trong T bị lệch

Chỉ số gini (CART)

- nếu dữ liệu T có n lớp, chỉ số gini(T) được định nghĩa như sau :

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

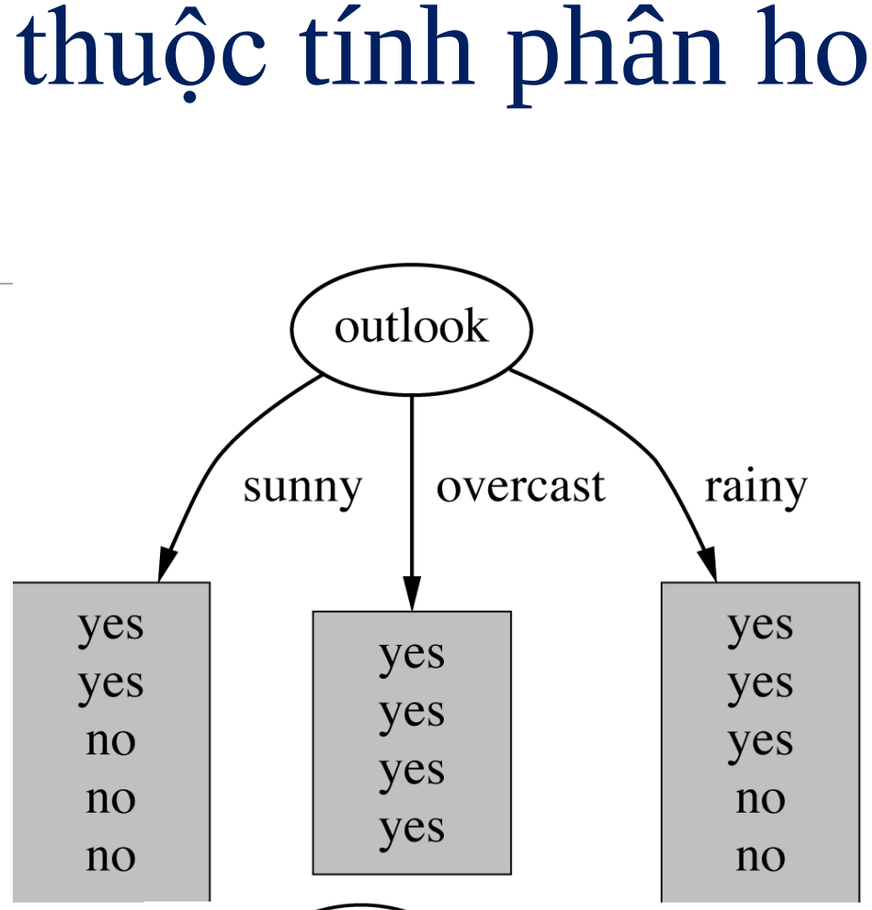
p_j là xác suất của lớp j trong T

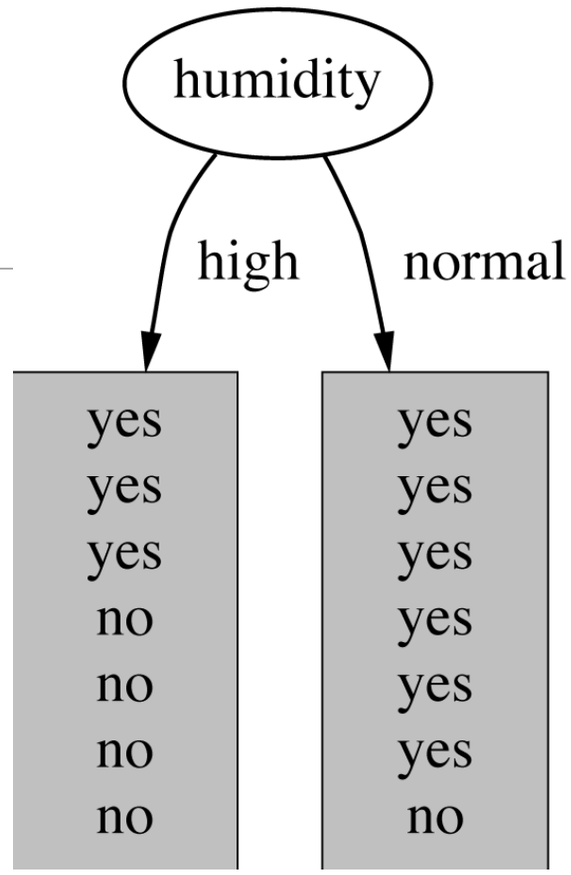
- gini(T) là nhỏ nhất nếu những lớp trong T bị lệch

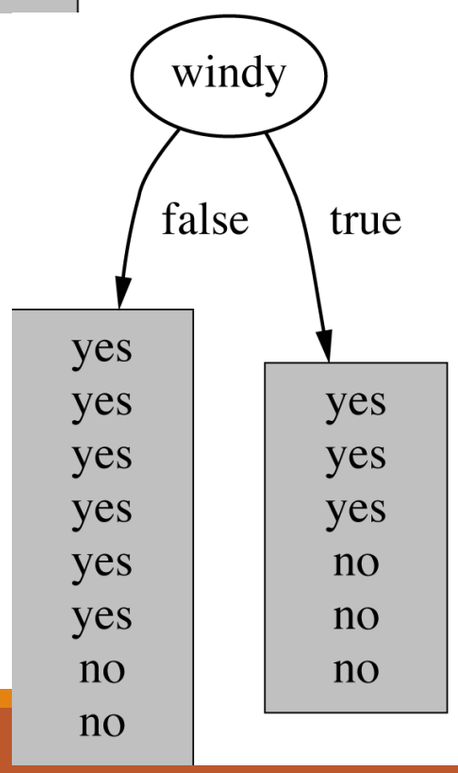
$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

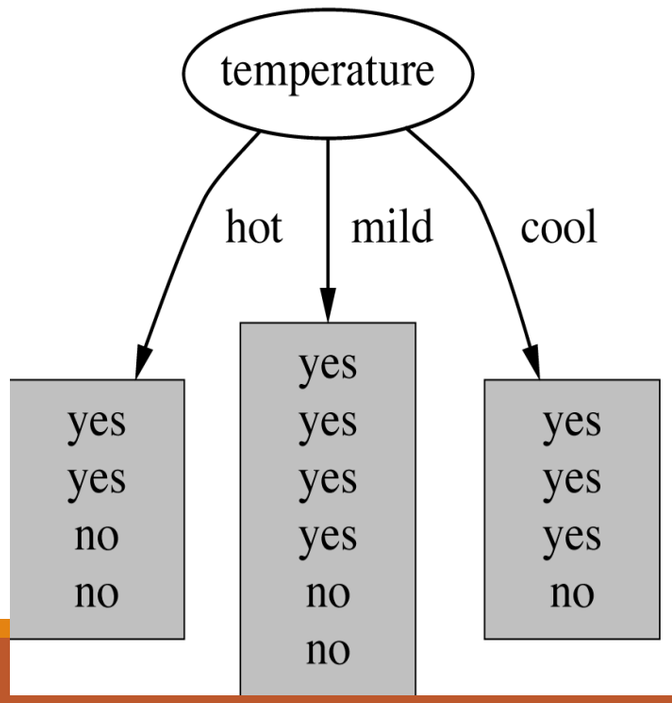
- sau khi phân hoạch T thành 2 tập con T1 & T2 với kích thước N1 & N2, chỉ số gini
- thuộc tính có **$gini_{split}$(T) nhỏ nhất** được chọn để phân hoạch

Chọn thuộc tính phân hoạch ?







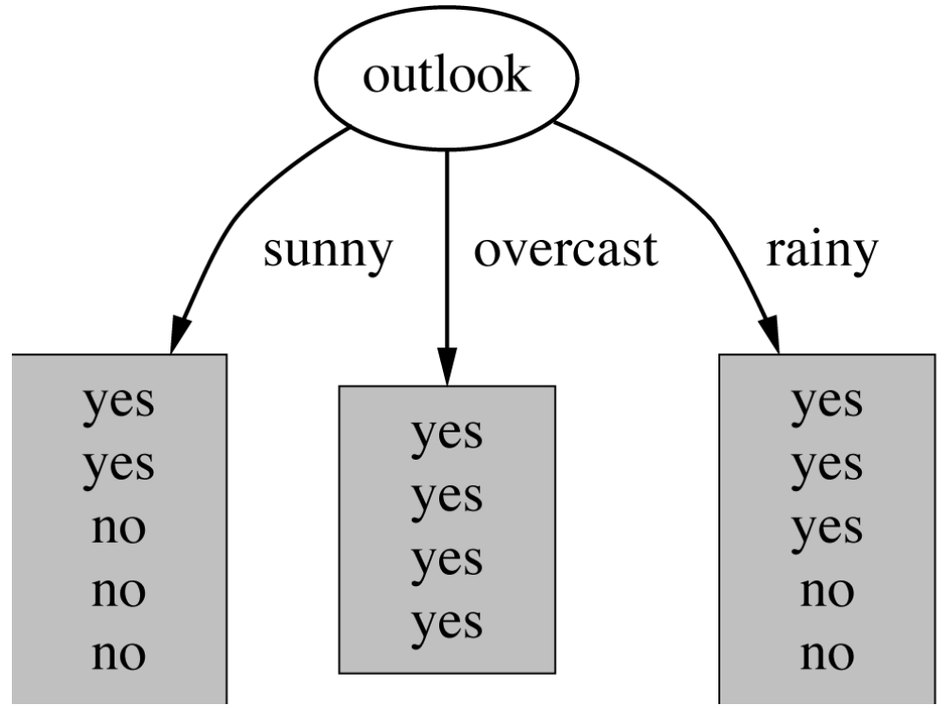


Xây dựng cây với chỉ số gini

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

Xây dựng cây với chỉ số gini

Tính Gini cho thuộc tính Outlook

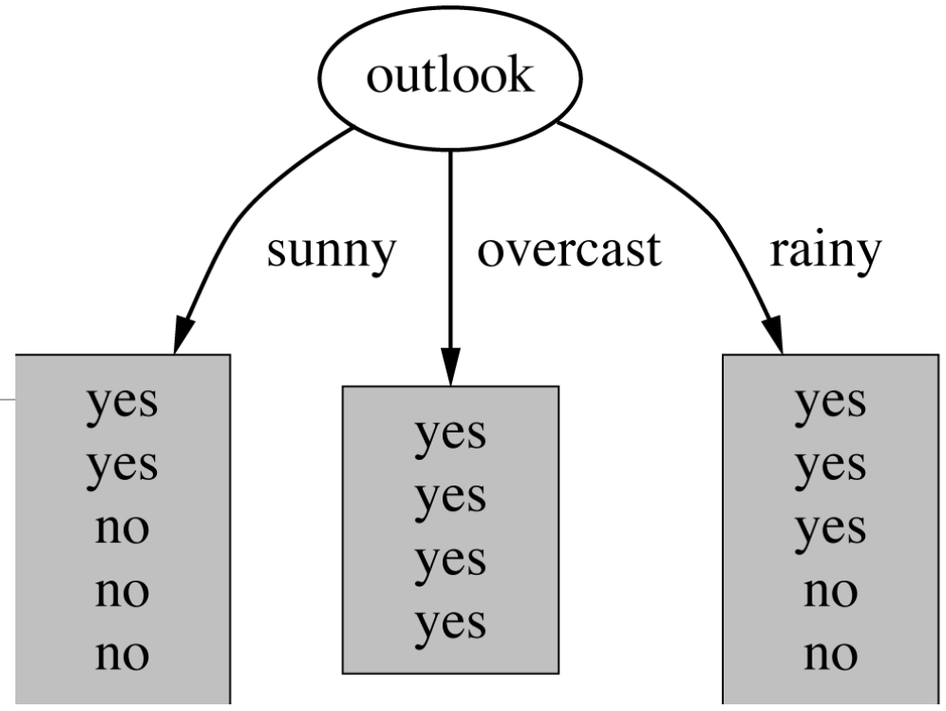


Gini(Outlook=Sunny) = $1 – [(2/5)^2 + (3/5)^2]= 1 – 0.16 – 0.36 = 0.48$

Gini(Outlook=Overcast) = $1 – [(4/4)^2 + (0/4)^2]= 0$

Gini(Outlook=Rain) = $1 – (3/5)^2 – (2/5)^2 = 1 – 0.36 – 0.16 = 0.48$

Xây dựng cây với chỉ số gini



Tính Gini cho thuộc tính Outlook

Gini(Outlook=Sunny) = $1 – [(2/5)^2 + (3/5)^2]= 1 – 0.16 – 0.36 = 0.48$

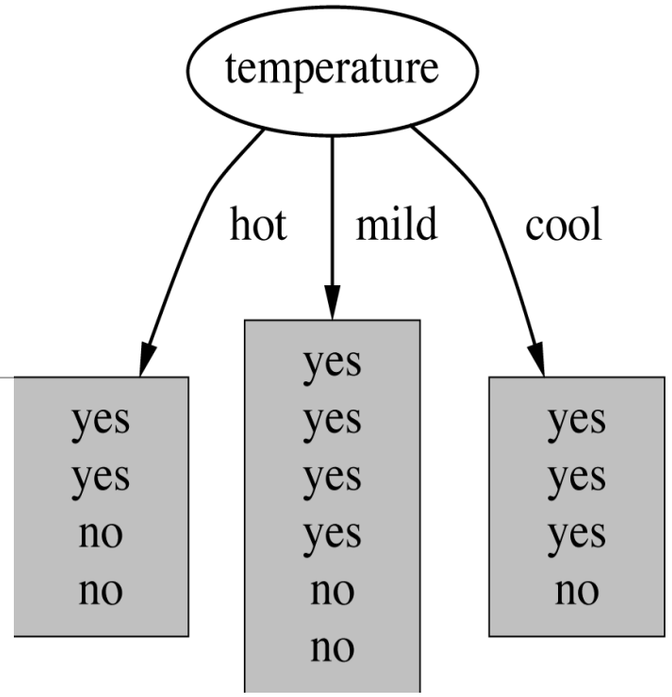
Gini(Outlook=Overcast) = $1 – [(4/4)^2 + (0/4)^2]= 0$

Gini(Outlook=Rain) = $1 – (3/5)^2 – (2/5)^2 = 1 – 0.36 – 0.16 = 0.48$

Gini(Outlook) = (5/14) x 0.48 + (4/14) x 0 + (5/14) x 0.48

= 0.171 + 0 + 0.171 = 0.342

Xây dựng cây với chỉ số gini



Tính Gini cho thuộc tính Temperature

Gini(Temp=Hot) = $1 – (2/4)^2 – (2/4)^2 = 0.5$

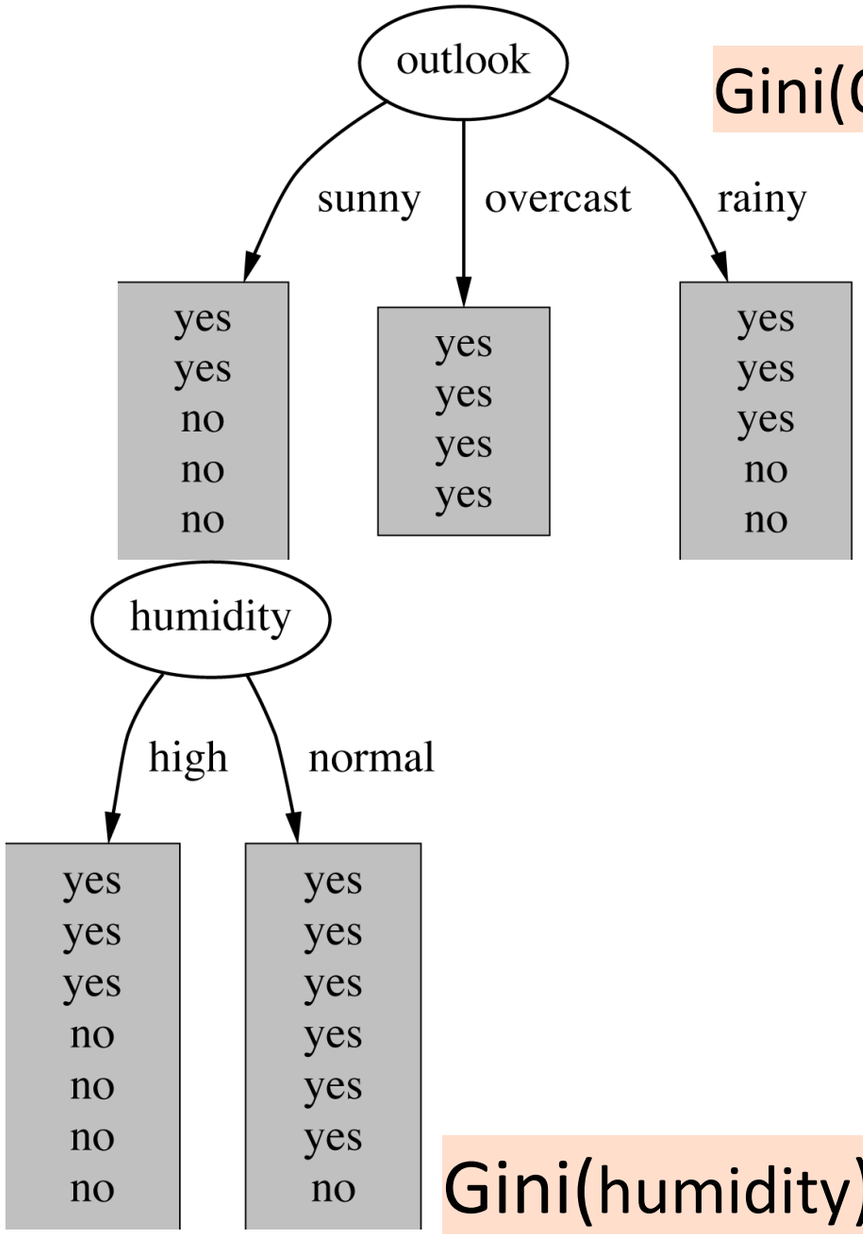
Gini(Temp=Cool) = $1 – (3/4)^2 – (1/4)^2 = 1 – 0.5625 – 0.0625 = 0.375$

Gini(Temp=Mild) = $1 – (4/6)^2 – (2/6)^2 = 1 – 0.444 – 0.111 = 0.445$

Gini(Temp) = (4/14) x 0.5 + (4/14) x 0.375 + (6/14) x 0.445

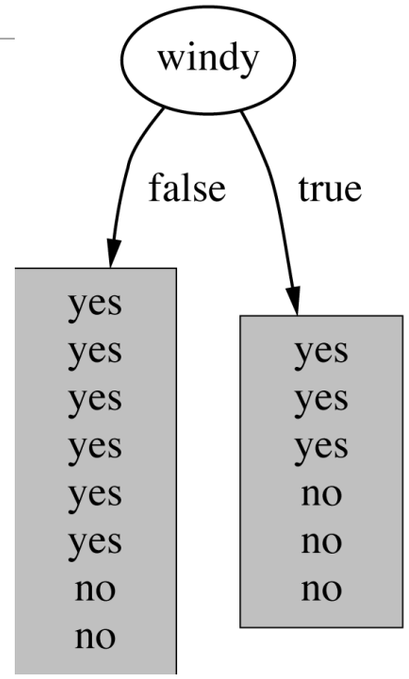
= 0.142 + 0.107 + 0.190 = 0.439

Tương tự tính Gini cho thuộc tính Humidity và Windy

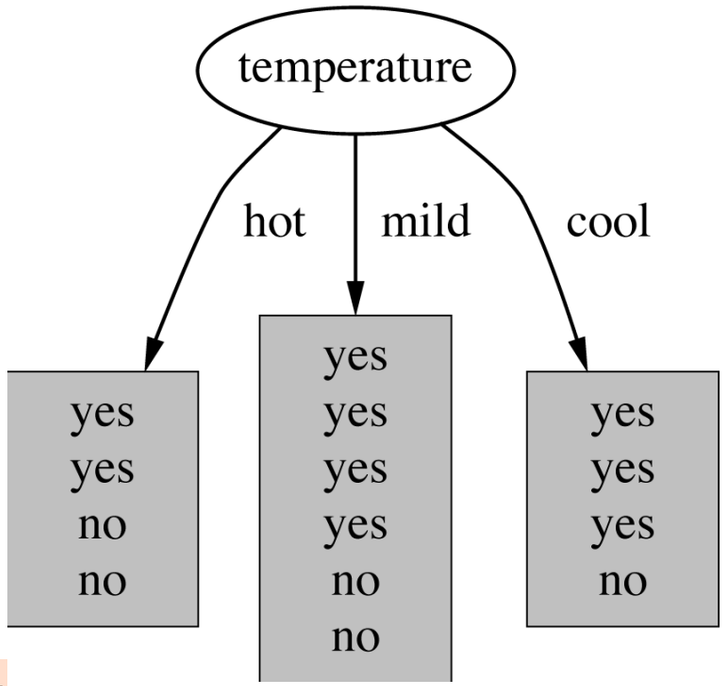


Gini(Outlook) = 0.343

Gini(humidity) = 0.367



Gini(Windy) = 0.429



Gini(temperature) = 0.44

Xây dựng cây với chỉ số gini

Tại nhánh Sunny, tính Gini cho Temperature, Humidity, Wind

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Xây dựng cây với chỉ số gini

Tại nhánh Sunny, tính Gini cho Temperature, Humidity, Wind

- Gini của Temperature đối với Outlook = Sunny

Temperature	Yes	No	Number of instances
Hot	0	2	2
Cool	1	0	1
Mild	1	1	2

Gini(Outlook=Sunny, Temp.=Hot) = $1 - (0/2)^2 - (2/2)^2 = 0$

Gini(Outlook=Sunny, Temp.=Cool) = $1 - (1/1)^2 - (0/1)^2 = 0$

Gini(Outlook=Sunny, Temp.=Mild) = $1 - (1/2)^2 - (1/2)^2$

$= 1 - 0.25 - 0.25 = 0.5$

Gini(Outlook=Sunny, Temp.) = (2/5)x0 + (1/5)x0 + (2/5)x0.5 = 0.2

Xây dựng cây với chỉ số gini

Tại nhánh Sunny, tính Gini cho Temperature, Humidity, Wind

- Gini của Humidity đối với Outlook = Sunny

Humidity	Yes	No	Number of instances
High	0	3	3
Normal	2	0	2

Gini(Outlook=Sunny, Humidity=High) = $1 – (0/3)^2 – (3/3)^2 = 0$

Gini(Outlook=Sunny, Humidity=Normal) = $1 – (2/2)^2 – (0/2)^2 = 0$

Gini(Outlook=Sunny, Humidity) = (3/5)x0 + (2/5)x0 = 0

Xây dựng cây với chỉ số gini

Khi Outlook = Sunny,

các giá trị Gini của các đặc trưng lần lượt:

Feature	Gini index
Temperature	0.2
Humidity	0
Wind	0.466

Giải thuật C4.5, dữ liệu kiểu số

- phân hoạch nhị phân
  - ví dụ : temp < 45
- không như dữ liệu liệt kê, dữ liệu kiểu số có nhiều nhánh phân hoạch
- phương pháp
  - tính độ lợi thông tin cho mọi giá trị phân nhánh của thuộc tính
  - chọn giá trị phân nhánh tốt nhất

Tập Weather, dữ liệu kiểu số

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
```

Tập Weather, dữ liệu kiểu số

- phân hoạch trên thuộc tính temperature

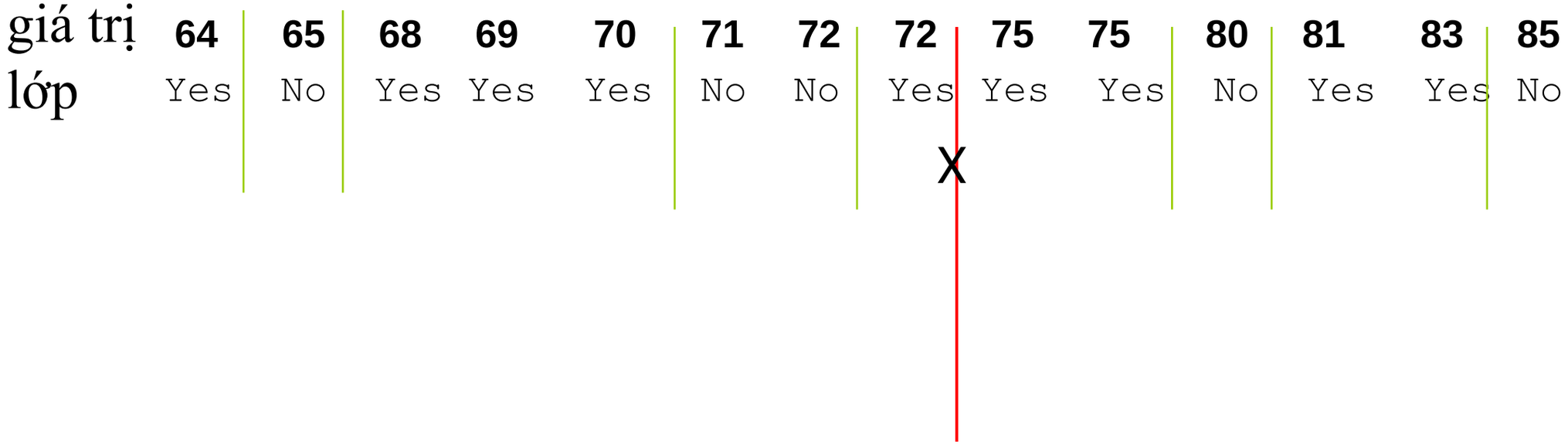
64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

  - ví dụ temperature < 71.5: yes/4, no/2
    temperature ≥ 71.5: yes/5, no/3

    - Info([4,2],[5,3]) = 6/14 info([4,2]) + 8/14 info([5,3])
      = 0.939 bits
- điểm phân hoạch : giữa
- có thể tính tất cả với 1 lần pass!
- cần sắp xếp dữ liệu

Cải tiến

chỉ cần tính entropy tại các điểm thay đổi lớp (Fayyad & Irani, 1992)



điểm giữa của cùng lớp không phải điểm tối ưu

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
14	Rain	Mild	High	Strong	30

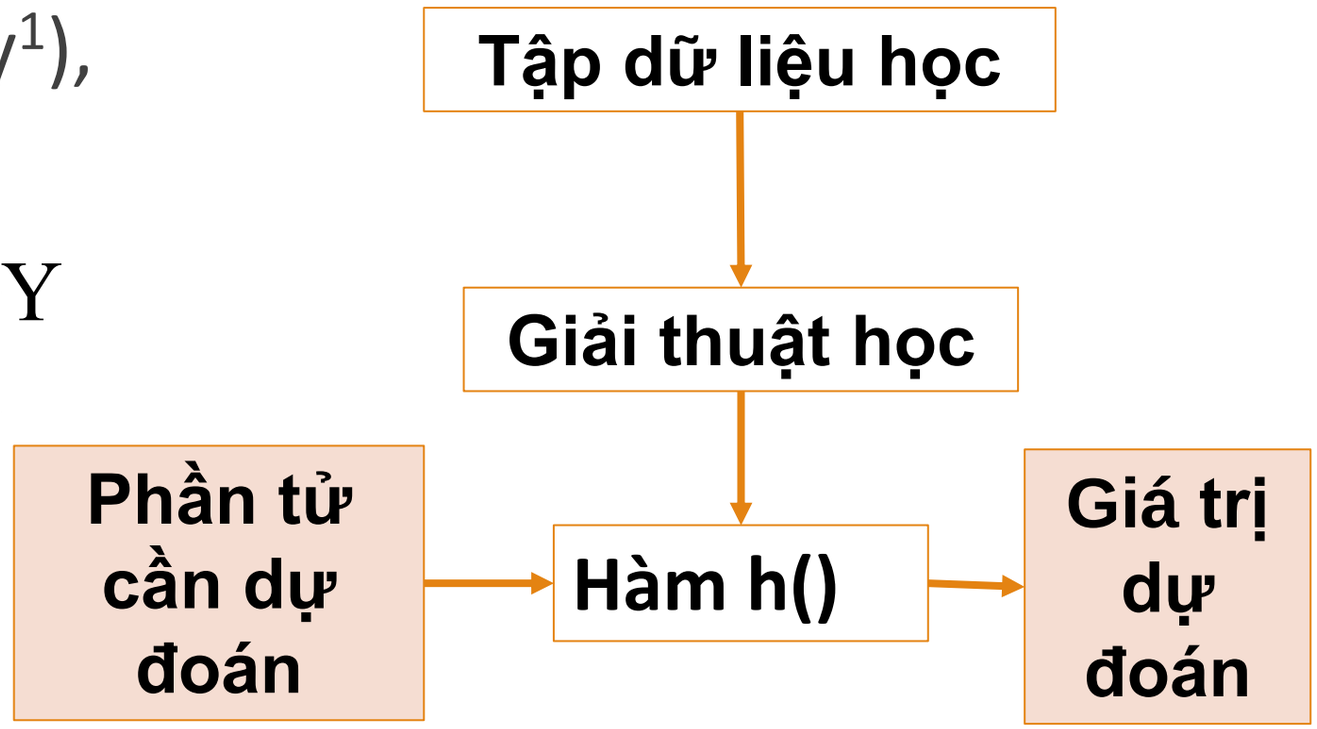
Cây quyết định cho bài toán hồi quy

Golf Players
25
30
46
45
52
23
43
35
38
46
48
52
44
30

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)

Chọn thuộc tính phân hoạch ?

- Bài toán phân lớp
  - độ lợi thông tin
  - *Chỉ số Gini*
- **Bài toán hồi quy**
  - Phương sai - Variance
  - Standard deviation (độ lệch chuẩn)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

  - The residual sum of squares

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

Cây quyết định cho bài toán hồi quy

CART - Regression Trees (Brieman et al. 84)

- Tính độ lệch chuẩn cho cột nhãn (Gold Playes)
- Tính độ lệch chuẩn của từng thuộc tính

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- Chọn thuộc tính có độ lệch chuẩn nhỏ nhất: có sự giảm độ lệch chuẩn nhiều nhất so với khi không phân hoạch

Cây quyết định cho bài toán hồi quy

Số lượng người chơi golf trung bình

μ = (25 + 30 + 46 + 45 + 52 + 23 + 43 + 35 + 38+ 46 + 48 + 52 + 44 + 30)/14

= 39.78

Độ lệch chuẩn (Standard deviation) số lượng người chơi (Toàn bộ tập dữ liệu)

σ= $\sqrt{[}$((25 – 39.78)^2 + (30 – 39.78)^2 + (46 – 39.78)^2 + … + (30 – 39.78)^2)/14]

= 9.32

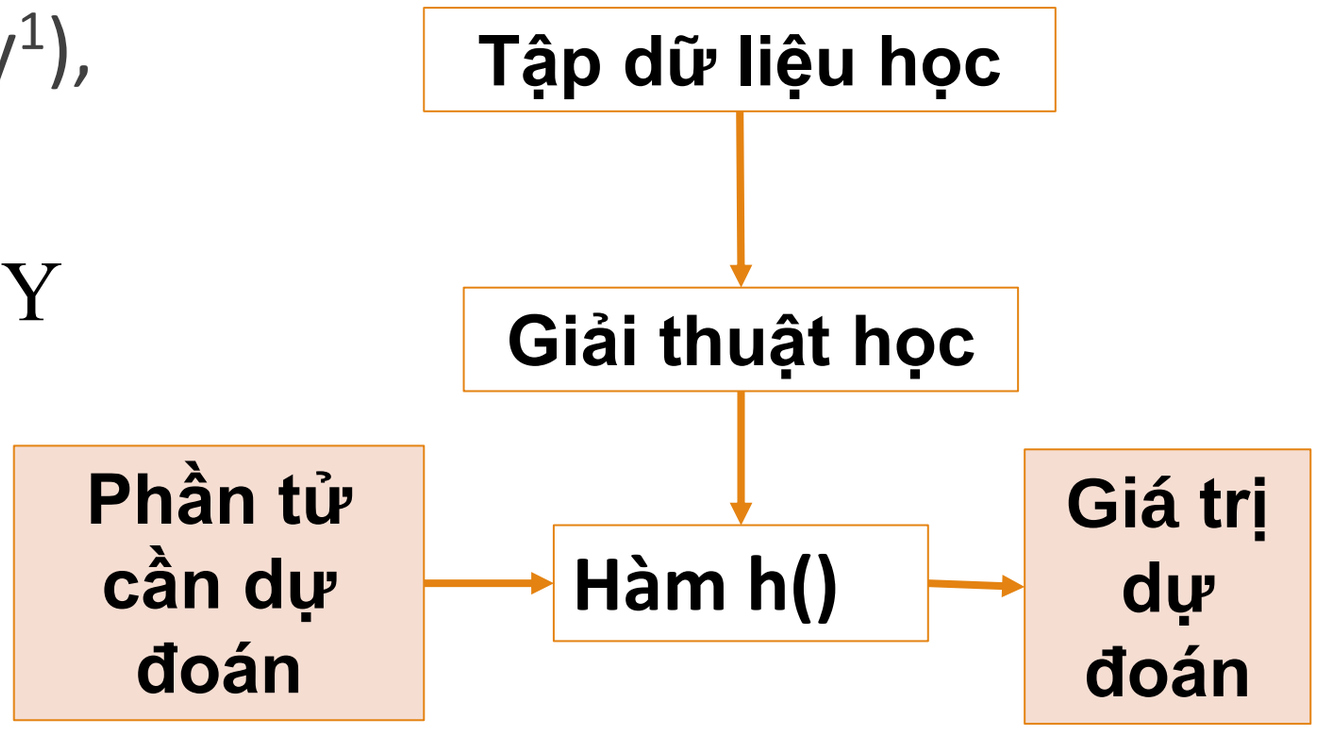
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Golf Players
25
30
46
45
52
23
43
35
38
46
48
52
44
30

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)

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Số lượng người chơi golf trung bình với Outlook = sunny

Độ lệch chuẩn (Standard deviation) số lượng người chơi

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	**25**
2	Sunny	Hot	High	Strong	**30**
8	Sunny	Mild	High	Weak	**35**
9	Sunny	Cool	Normal	Weak	**38**
11	Sunny	Mild	Normal	Strong	**48**

Số lượng người chơi golf trung bình với Outlook = sunny

$$\mu = (25 + 30 + 35 + 38 + 48)/5 = 35.2$$

Độ lệch chuẩn (Standard deviation) số lượng người chơi

$$\sigma = \sqrt{(((25 - 35.2)^2 + (30 - 35.2)^2 + (35 - 35.2)^2 + (38 - 35.2)^2 + (48 - 35.2)^2)/5)}$$

$$= 7.78$$

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Số lượng người chơi golf trung bình khi outlook = overcast

$\mu_{\text{outlook = overcast}}$

Độ lệch chuẩn (Standard deviation) số lượng người chơi

$\sigma_{\text{outlook = overcast}}$

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Số lượng người chơi golf trung bình khi outlook = overcast

$$\mu_{\text{outlook = overcast}} = (46 + 43 + 52 + 44)/4 = 46.25$$

Độ lệch chuẩn (Standard deviation) khi outlook = overcast

$$\sigma_{\text{outlook = overcast}} = \sqrt{(((46-46.25)^2+(43-46.25)^2+\ldots)/4} = 3.49$$

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Số lượng người chơi golf trung bình khi outlook = rain

$\mu_{\text{outlook = rain}}$

Độ lệch chuẩn (Standard deviation) khi outlook = rain

$\sigma_{\text{outlook = rain}}$

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Số lượng người chơi golf trung bình khi “outlook” =rain

= (45+52+23+46+30)/5 = 39.2

Độ lệch chuẩn (Standard deviation) khi “outlook” =rain

= $\sqrt{(((45 - 39.2)^2+(52 - 39.2)^2+\ldots)/5)}$=10.87

Cây quyết định cho bài toán hồi quy

Outlook	**Stdev of Golf Players**	**Instances**
Overcast	3.49	4
Rain	10.87	5
Sunny	7.78	5

$$S(T, X) = \sum_{c \in X} P(c) S(c)$$

Độ lệch chuẩn của thuộc tính Outlook

= (4/14)x3.49 + (5/14)x10.87 + (5/14)x7.78 = **7.66**

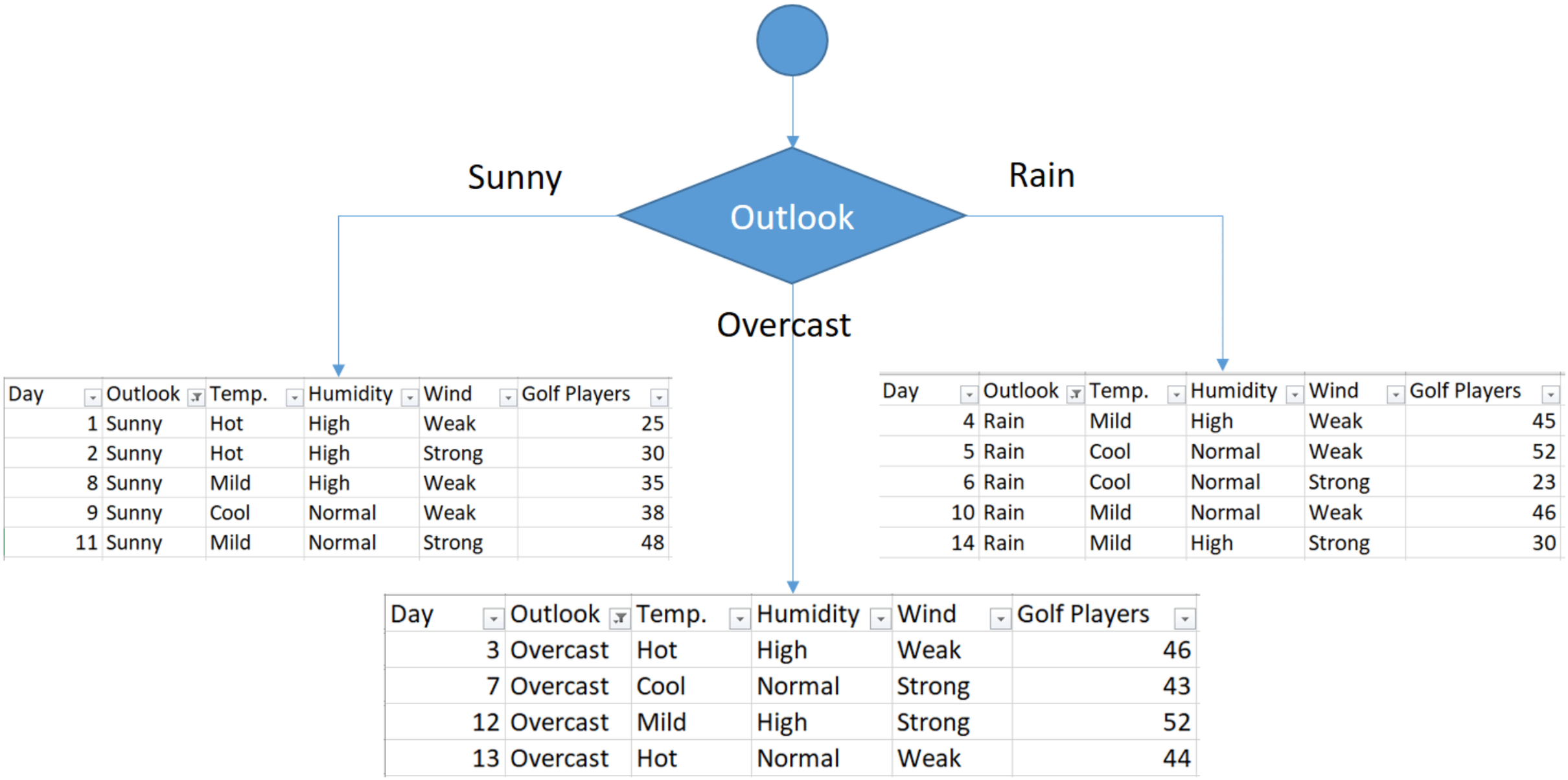
Độ chênh lệch giữa độ lệch chuẩn của toàn bộ dữ liệu và độ lệch chuẩn của thuộc tính outlook

Standard Deviation Reduction$_{\textbf{Outlook}}$ = 9.32 – 7.66 = **1.66**

Cây quyết định cho bài toán hồi quy

	Standard Deviation Reduction
Outlook	1.66
Temperature	0.47
Humidity	0.27
Wind	0.29

Cây quyết định cho bài toán hồi quy



Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Số người chơi golf khi outlook= sunny = {25, 30, 35, 38, 48}

Độ lệch chuẩn khi Outlook=Sunny: 7.78

Sử dung độ lệch chuẩn này như là độ lệch chuẩn cho toàn bộ dữ liệu của bước trước đó.

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
~~**8**~~	~~Sunny~~	~~Mild~~	~~High~~	~~Weak~~	~~35~~
~~**9**~~	~~Sunny~~	~~Cool~~	~~Normal~~	~~Weak~~	~~38~~
~~**11**~~	~~Sunny~~	~~Mild~~	~~Normal~~	~~Strong~~	~~48~~

Độ lệch chuẩn khi Outlook = sunny và temp. = hot

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30

Độ lệch chuẩn khi Outlook = sunny và temp. = hot

Số lượng người chơi golf trung bình khi outlook = sunny và temp.=hot

$$\mu_{\text{outlook = sunny và temp.=hot}} = (25+30)/2 = 27.5$$

Độ lệch chuẩn (Standard deviation) khi outlook = sunny

$$\sigma_{\text{ outlook =sunny và temp.=hot}} = \sqrt{((25-27.5)^2+(30-27.5)^2)/2} = 2.5$$

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
9	Sunny	Cool	Normal	Weak	38

Độ lệch chuẩn khi Outlook = sunny và temp. = cool

Số lượng người chơi golf trung bình khi outlook = sunny và temp.=cool

$$\mu_{\text{outlook = sunny và temp.=hot}} = 38$$

Độ lệch chuẩn (Standard deviation) khi outlook = sunny

$$\sigma_{\text{ outlook =sunny và temp.=hot}} = \sqrt{((38-38)^2} = 0$$

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
8	Sunny	Mild	High	Weak	35
11	Sunny	Mild	Normal	Strong	48

Độ lệch chuẩn khi Outlook = sunny và temp. = mild

Số lượng người chơi golf trung bình khi outlook = sunny và temp.=mild

$$\mu_{\text{outlook = sunny và temp.=}mild} = (35+48)/2 = 41.5$$

Độ lệch chuẩn (Standard deviation) khi outlook = sunny

$$\sigma_{\text{ outlook = sunny và temp.=}mild} = \sqrt{((35-41.5)^2+(48-41.5)^2)/2} = 6.5$$

Cây quyết định cho bài toán hồi quy

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Temperature	Stdev for Golf Players	Instances
Hot	2.5	2
Cool	0	1
Mild	6.5	2

Cây quyết định cho bài toán hồi quy

Temperature	Stdev for Golf Players	Instances
Hot	2.5	2
Cool	0	1
Mild	6.5	2

Độ lệch chuẩn khi outlook=sunny và xét thuộc tính temp.
= (2/5)x2.5 + (1/5)x0 + (2/5)x6.5 = 3.6
Độ chênh lệch của độ lệch chuẩn khi outlook=sunny và outlook =sunny + thuộc tính temp. = 7.78 – 3.6 = 4.18

Cây quyết định cho bài toán hồi quy

Outlook= sunny và humidity = high

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35

Độ lệch chuẩn khi outlook=sunny và humidity = high: 4.08

Cây quyết định cho bài toán hồi quy

Outlook= sunny và humidity = normal

Day	Outlook	Temp.	Humidity	Wind	Golf Players
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Độ lệch chuẩn khi outlook=sunny và thuộc tính humidity = normal: 5

Cây quyết định cho bài toán hồi quy

Độ lệch chuẩn khi outlook=sunny và xét thuộc tính humidity

Humidity	Stdev for Golf Players	Instances
High	4.08	3
Normal	5.00	2

Độ lệch chuẩn khi outlook=sunny và xét thuộc tính humidity
= (3/5)x4.08 + (2/5)x5 = 4.45
Độ chênh lệch của độ lệch chuẩn khi outlook=sunny và outlook =sunny + thuộc tính humidity = 7.78 – 4.45 = 3.33

Cây quyết định cho bài toán hồi quy

Outlook= sunny và windy = weak

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
~~**2**~~	~~Sunny~~	~~Hot~~	~~High~~	~~Strong~~	~~30~~
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
~~**11**~~	~~Sunny~~	~~Mild~~	~~Normal~~	~~Strong~~	~~48~~

Độ lệch chuẩn khi outlook=sunny và thuộc tính windy = weak: 5.56

Cây quyết định cho bài toán hồi quy

Outlook= sunny và windy = strong

Day	Outlook	Temp.	Humidity	Wind	Golf Players
~~**1**~~	~~Sunny~~	~~Hot~~	~~High~~	~~Weak~~	~~25~~
2	Sunny	Hot	High	Strong	30
~~**8**~~	~~Sunny~~	~~Mild~~	~~High~~	~~Weak~~	~~35~~
~~**9**~~	~~Sunny~~	~~Cool~~	~~Normal~~	~~Weak~~	~~38~~
11	Sunny	Mild	Normal	Strong	48

Độ lệch chuẩn khi outlook=sunny và thuộc tính windy = strong: 9

Cây quyết định cho bài toán hồi quy

Độ lệch chuẩn khi outlook=sunny và xét thuộc tính windy

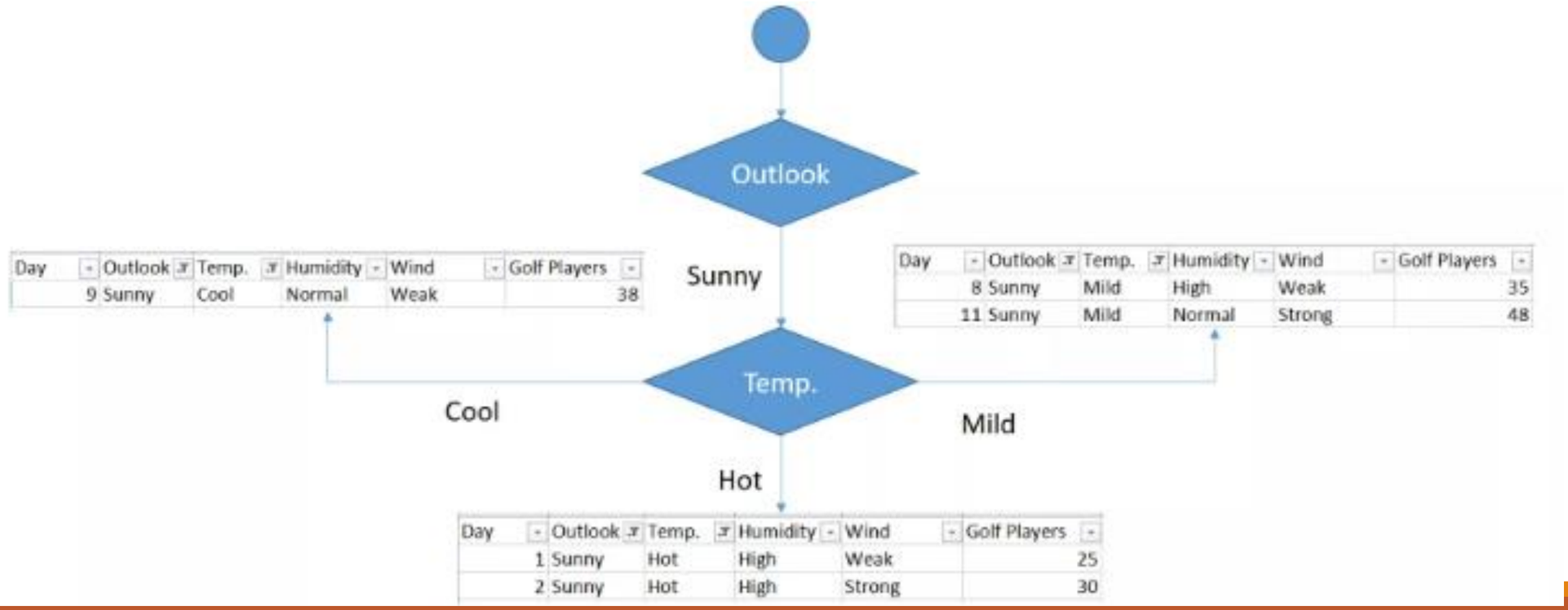
Wind	Stdev for Golf Players	Instances
Strong	9	2
Weak	5.56	3

Độ lệch chuẩn khi outlook=sunny và xét thuộc tính windy
= (2/5)*9 + (3/5)x5.56 = 6.93
Độ chênh lệch của độ lệch chuẩn khi outlook=sunny và outlook =sunny + thuộc tính windy = 7.78 – 6.93 = 0.85

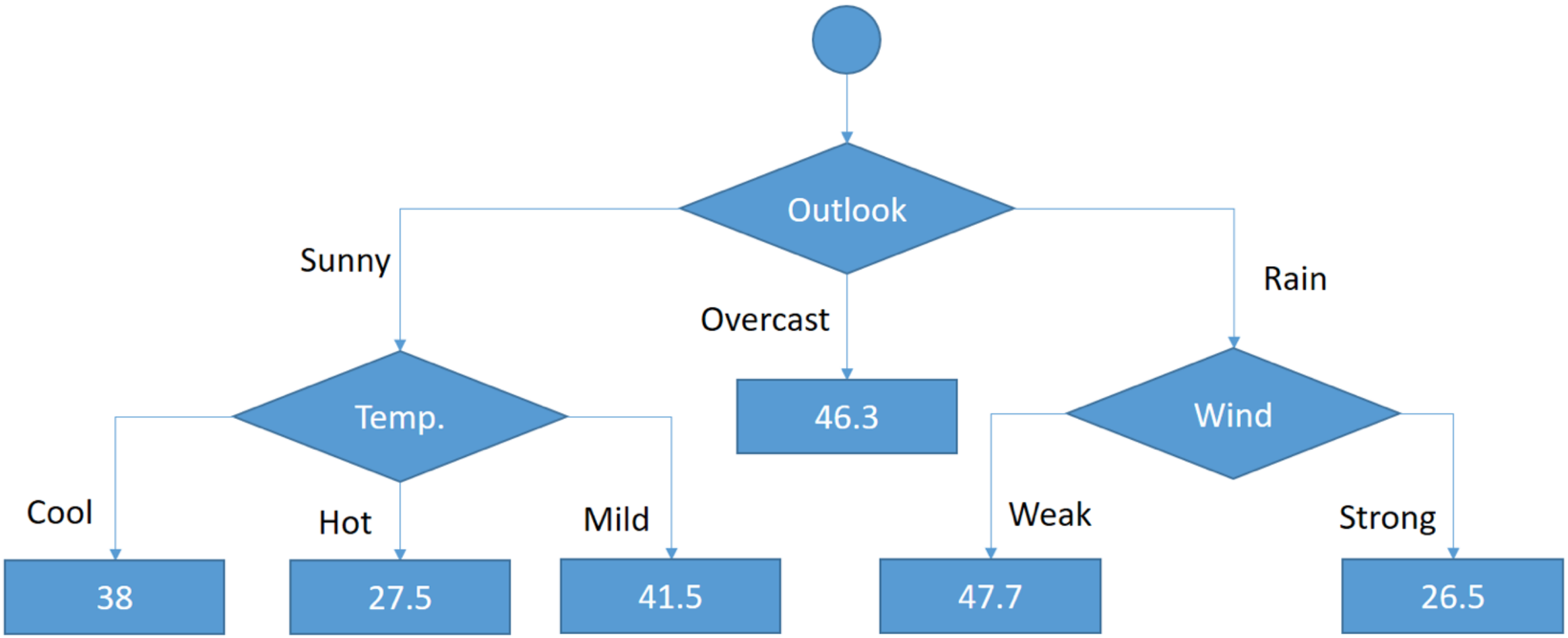
Cây quyết định cho bài toán hồi quy

Cây quyết định được xây dựng:

Feature	Standard Deviation Reduction
Temperature	4.18
Humidity	3.33
Wind	0.85



Cây quyết định cho bài toán hồi quy



Cắt nhánh

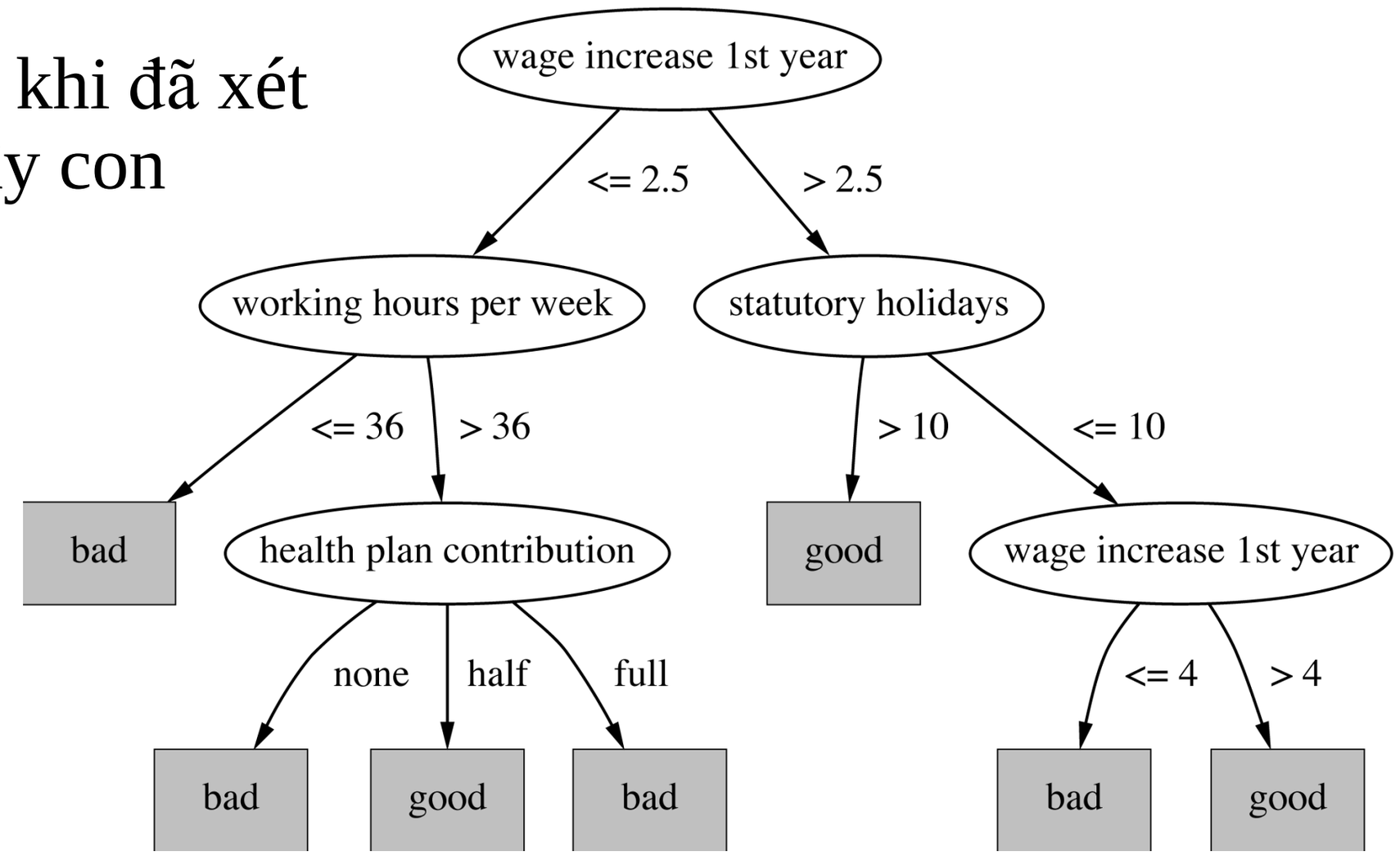
- mục tiêu : tránh học vẹt (overfitting), chịu đựng nhiễu, tăng độ chính xác khi phân loại tập test
- có 2 pha
  - *postpruning* – cắt nhánh cây sao cho tăng khả năng phân loại của cây
    - xây dựng cây đầy đủ
    - cắt nhánh
      - *thay thế cây con*
      - *đưa cây con lên trên*
  - *prepruning* – dừng sớm quá trình phân nhánh
- trong thực tế, postpruning được sử dụng nhiều hơn prepruning

Postpruning

- xây dựng cây đầy đủ
- cắt nhánh
  - *thay thế cây con*
  - *đưa cây con lên trên*
- có nhiều chiến lược
  - ước lượng lỗi
  - significance test

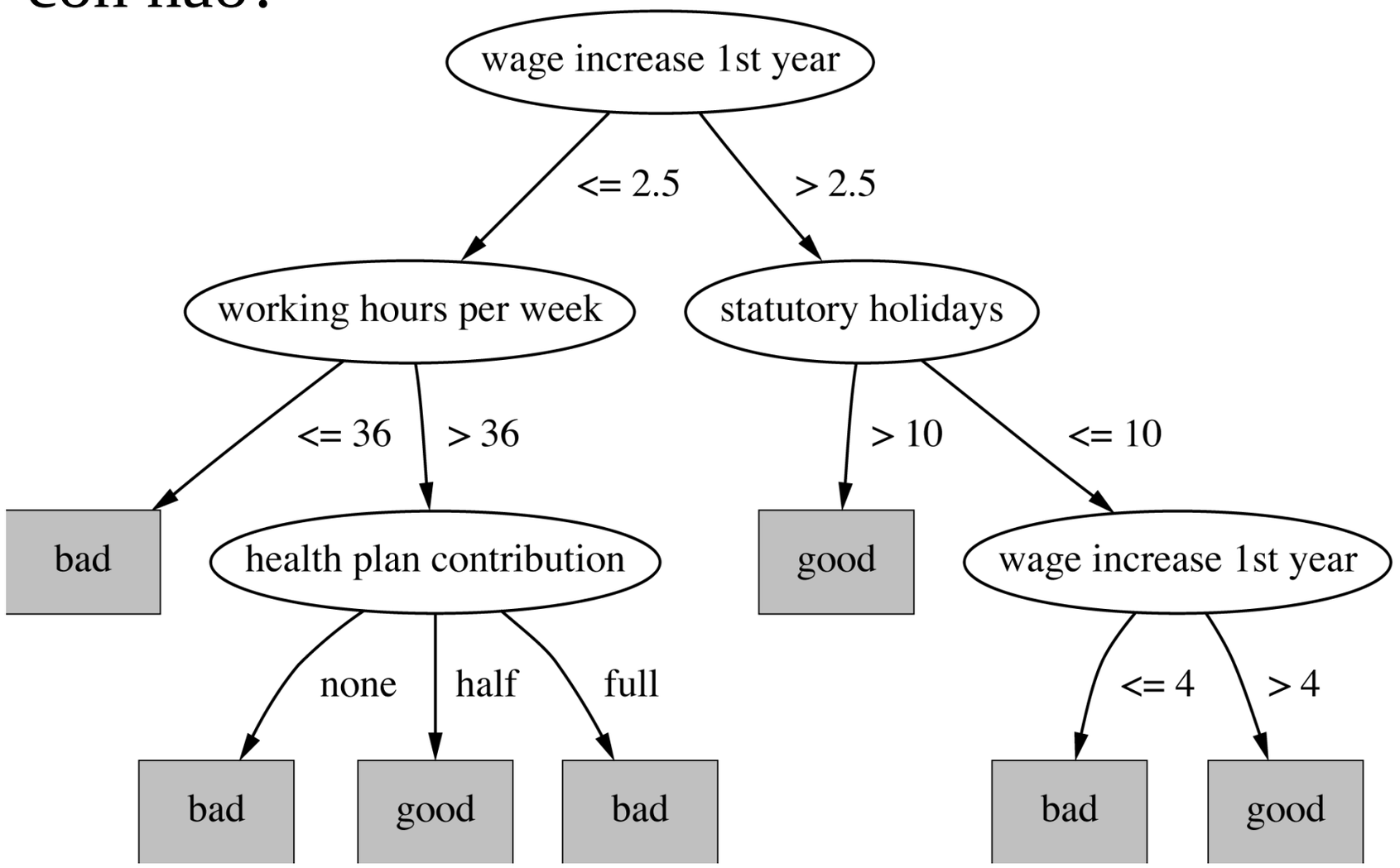
Thay thế cây con

- *Bottom-up*
- thay thế sau khi đã xét tất cả các cây con

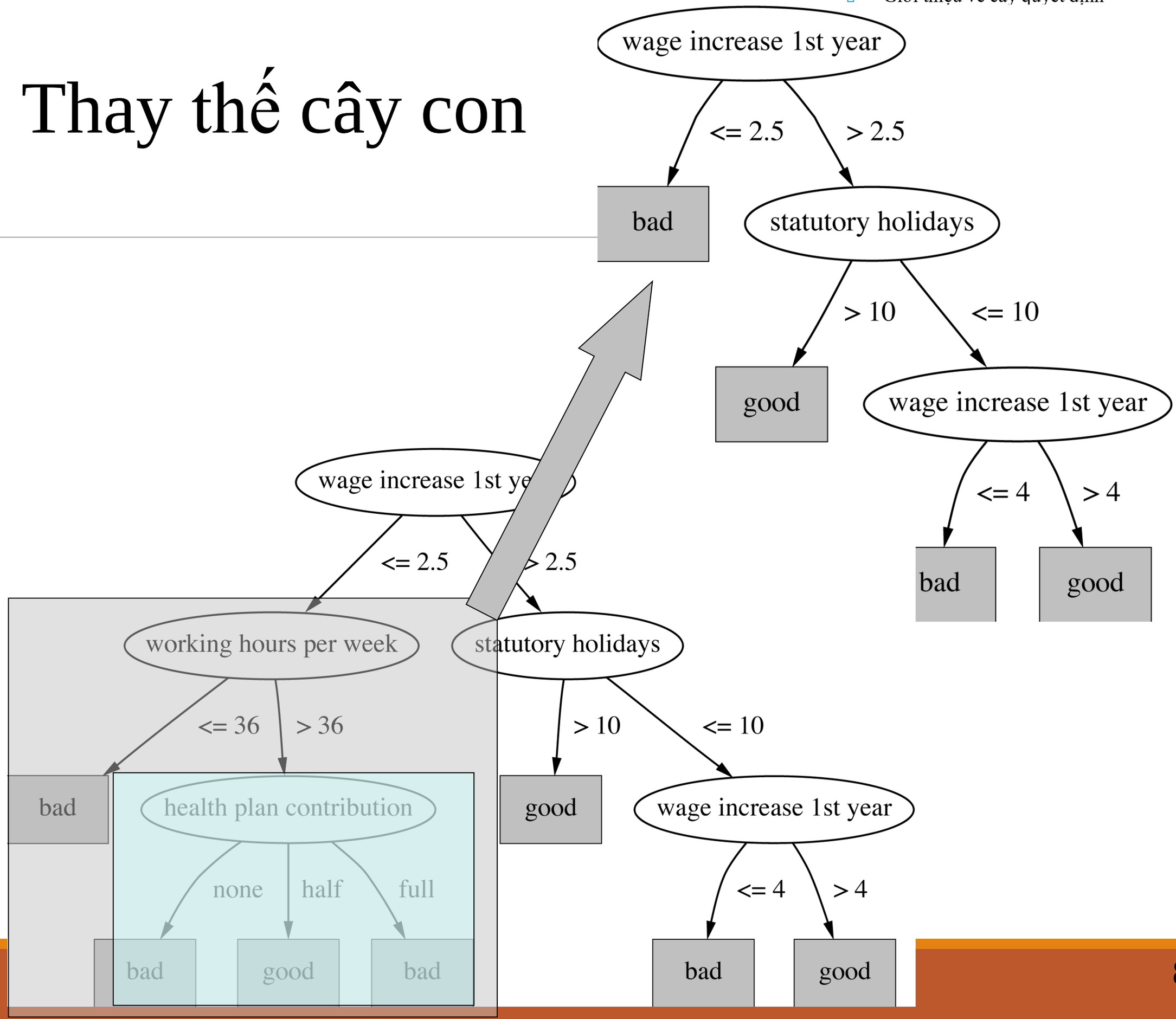


Thay thế cây con

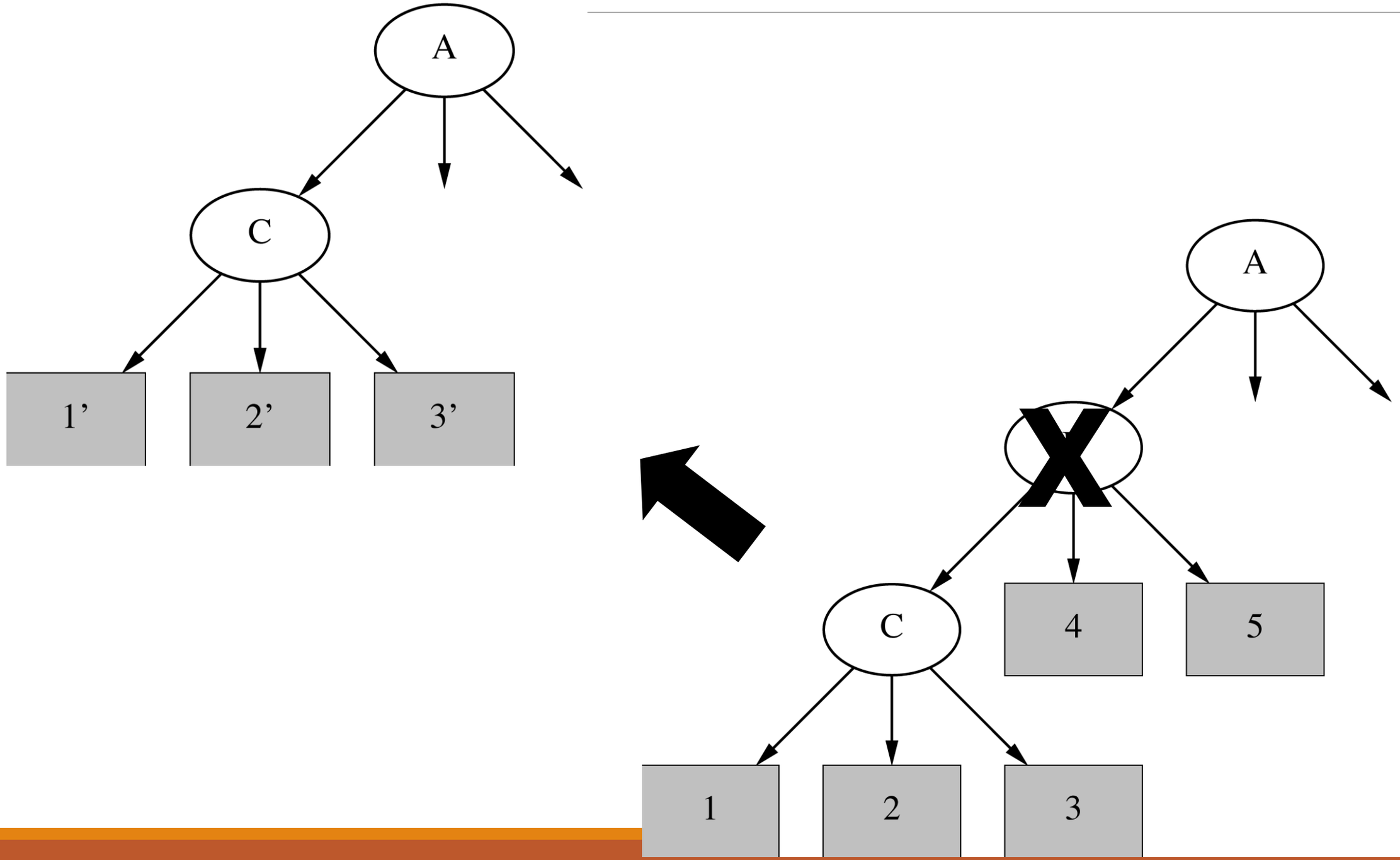
- thay thế cây con nào?



Thay thế cây con



Đưa cây con lên trên



Nội dung

Giới thiệu về cây quyết định

Giải thuật học của cây quyết định

Kết luận và hướng phát triển

Kết luận

- cây quyết định
  - xây dựng top-down
    - chọn thuộc tính để phân hoạch (độ lợi thông tin, entropy, chỉ số Gini, etc)
    - cắt nhánh bottom-up
  - dễ cài đặt, học nhanh, kết quả dễ hiểu
  - được sử dụng nhiều và thành công nhất trong các ứng dụng thực

Hướng phát triển

- phát triển
  - tăng độ chính xác
  - xử lý dữ liệu không cân bằng
  - dữ liệu phức tạp có số chiều lớn
  - cây oblique
  - tìm kiếm thông tin (ranking)
  - clustering

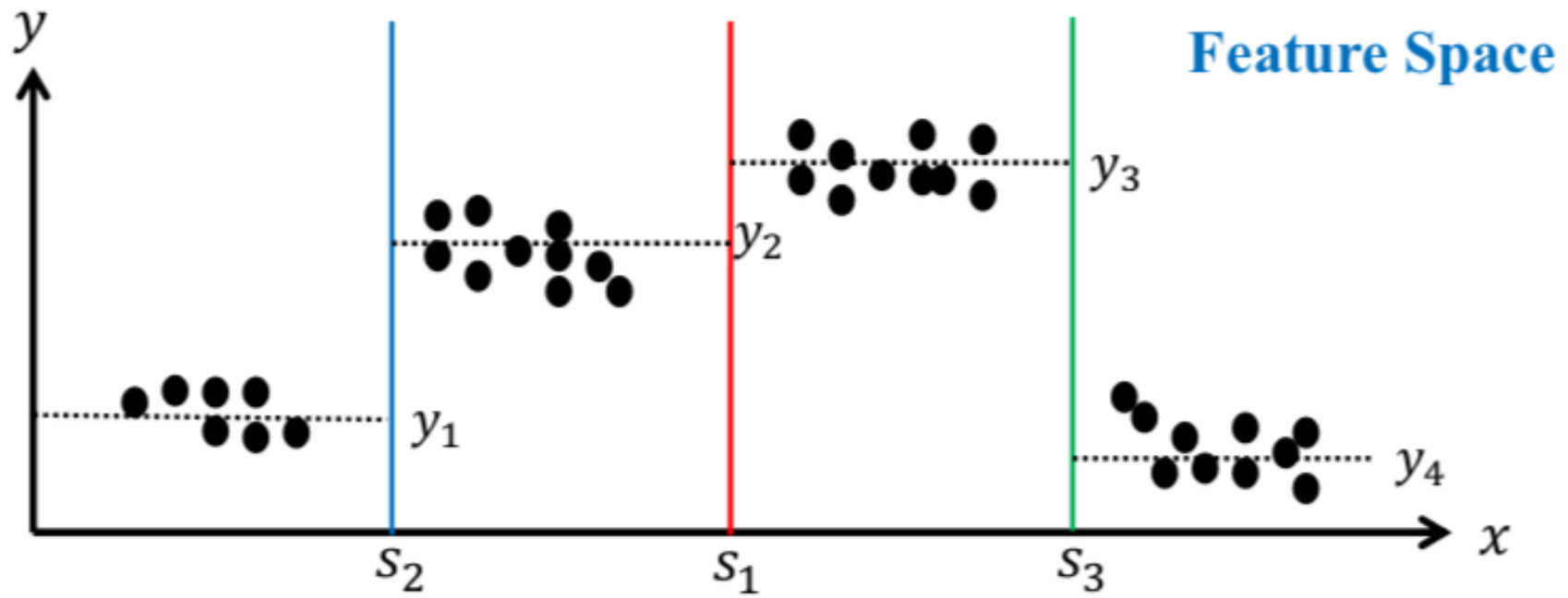
Cám ơn !

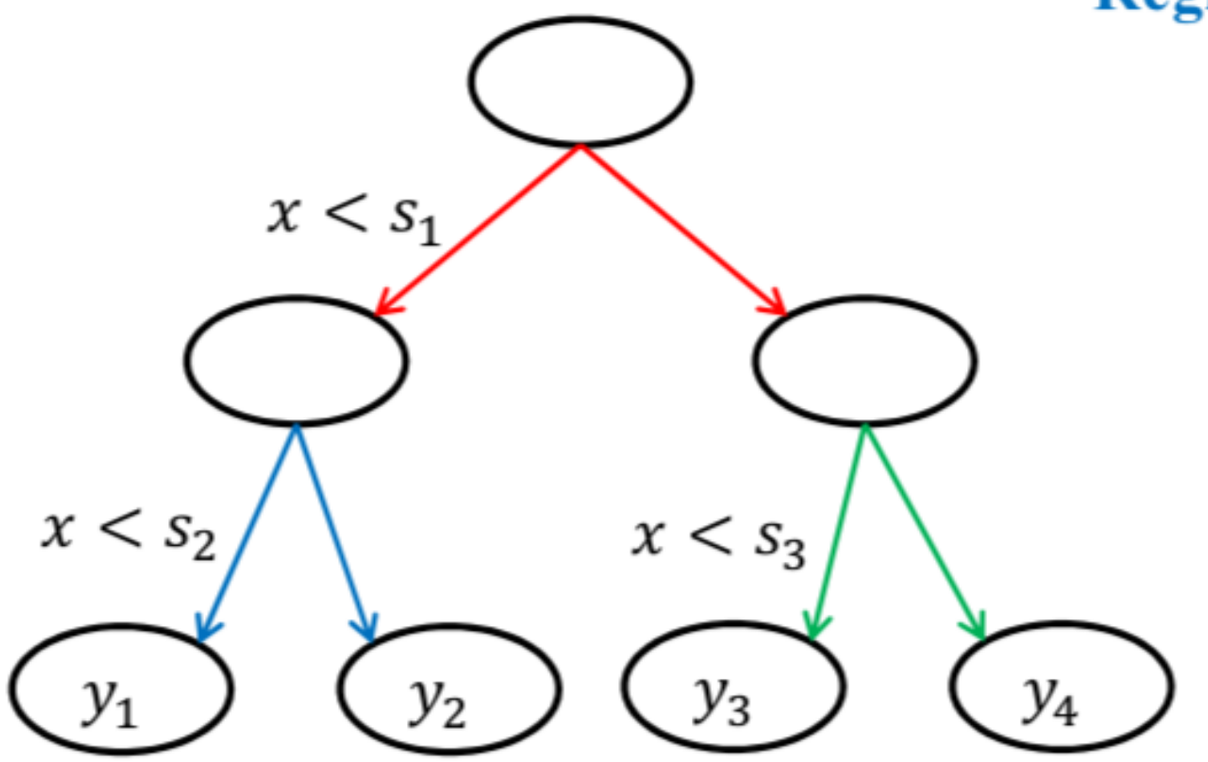
The **Population** Standard Deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

The **Sample** Standard Deviation:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$


Feature Space
y
x
y_1
y_2
y_3
y_4
s_2
s_1
s_3


Regression Tree
$x < s_1$
$x < s_2$
$x < s_3$
y_1
y_2
y_3
y_4

Phân chia thuộc tính có giá trị liên tục

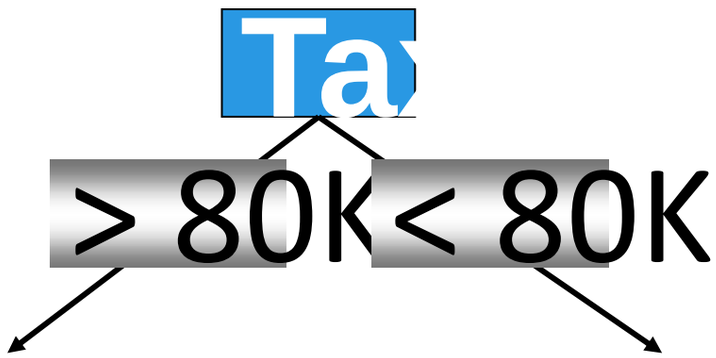
Dựa trên một giá trị nếu muốn phân chia nhị phân

Dựa trên vài giá trị nếu muốn có nhiều nhánh

Với mỗi giá trị tính các mẫu thuộc một lớp theo dạng A<*v* và A>*v*

Cách chọn giá trị *v* đơn giản: với mỗi giá trị *v* trong CSDL đều tính Gini của nó và lấy giá trị có Gini nhỏ nhất → kém hiệu quả

TID	Refund	Marital	Tax	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**



Phân chia thuộc tính có giá trị liên tục

Cách chọn giá trị *v* hiệu quả:

- Sắp xếp các giá trị tăng dần
- Chọn giá trị trung bình của từng giá trị của thuộc tính để phân chia và tính chỉ sô gini

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Taxable Income										
Sorted Values →	60	70	75	85	90	95	100	120	125	220
Split Positions →	55	65	72	80	87	92	97	110	122	172	230
	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >	<= >
Yes	0 3	0 3	0 3	0 3	1 2	2 1	3 0	3 0	3 0	3 0	3 0
No	0 7	1 6	2 5	3 4	3 4	3 4	3 4	4 3	5 2	6 1	7 0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	***0.300***	0.343	0.375	0.400	0.420

Differences from CT

Prediction is computed as the **average** of numerical target variable in the rectangle (in CT it is majority vote)

Impurity measured by **sum of squared deviations** from leaf mean

The residual sum of squares

Performance measured by RMSE (root

Model Selection in Trees: