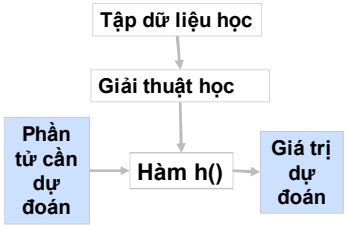
Phân loại học máy – học có giám sát

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

- • 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)



Quy ước

- Biến đầu vào (input variables)/đặc trưng (features), kí hiệu: $x^{(i)}$
- Biến đầu ra (output variable)/biến mục tiêu, kí hiệu $y^{(i)}$
- Mẫu huấn luyện (training example)
  kí hiệu ($x^{(i)}$, $y^{(i)}$)
- Tập huấn luyện X = {($x^{(i)}$, $y^{(i)}$)}, i = 1..m

Square meters	Bedrooms	Floors	Age of building (years)	Price in 1000€
x1	*x2*	*x3*	*x4*	*y*
200	5	1	45	460
131	3	2	40	232
142	3	2	30	315
756	2	1	36	178
…	…	…	…	…

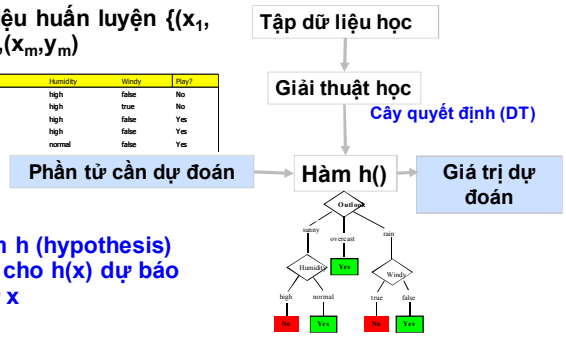
$$x^{(3)} = \begin{bmatrix} 142 \\ 3 \\ 2 \\ 30 \end{bmatrix}$$

$$x_1^{(4)} = 756$$

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

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

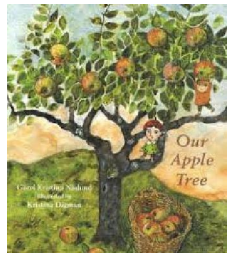
[See: Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997]

Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ

Phương pháp học cây quyết định
Decision Tree

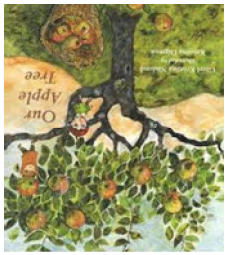
Nội dung

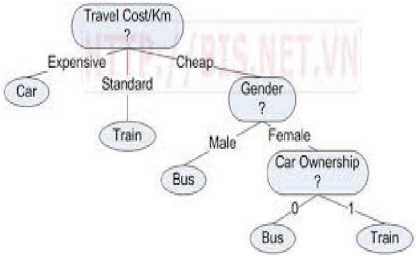
- Giới thiệu về cây quyết định
- Giải thuật học của cây quyết định
- Kết luận và hướng phát triển



Nội dung

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





Cây quyết định

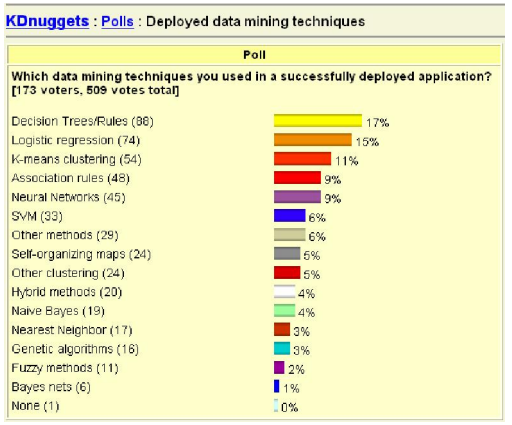
- lớp các giải thuật học
  - kết quả sinh ra dễ dịch **(if … then …)**
  - khá đơn giản, nhanh, hiệu quả được sử dụng nhiều
  - liên tục trong nhiều năm qua, cây quyết định được bình chọn là giải thuật được sử dụng nhiều nhất và thành công nhất
  - giải quyết các vấn đề của phân loại, hồi quy
  - làm 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

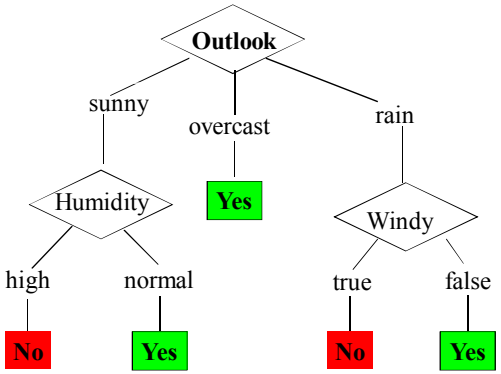
- ID3 (Quinlan 79)
- **CART – Classification and Regression Trees (Brieman et al. 84)**
- Assistant (Cestnik et al. 87)
- **C4.5 (Quinlan 93)**
- See5 (Quinlan 97)
- ...
- Orange (Demšar, Zupan 98-03)

Kỹ thuật DM thành công trong ứng dụng thực (2004)



10

Example Decision Tree



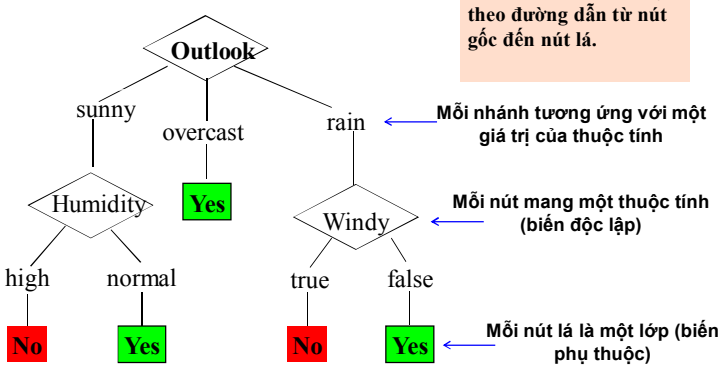
Cây quyết định

- 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

12

Ví dụ Decision Tree

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)

Nội dung

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

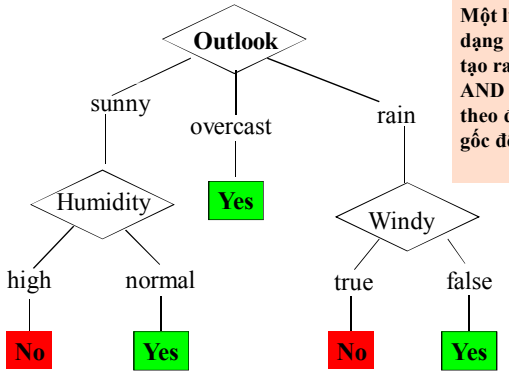
14

Dữ liệu weather, dựa trên các thuộc tính (Outlook, Temp, Humidity, Windy), quyết định (play/no)

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

15

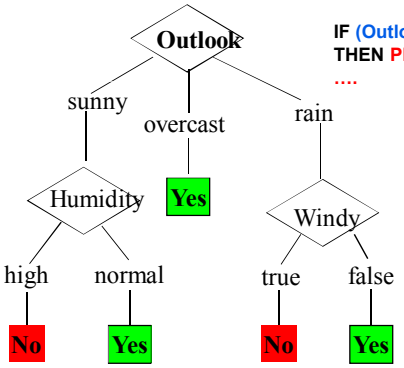
Cây quyết định cho tập dữ liệu weather, dựa trên các thuộc tính (Outlook, Temp, Humidity, Windy)



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.

16

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
....

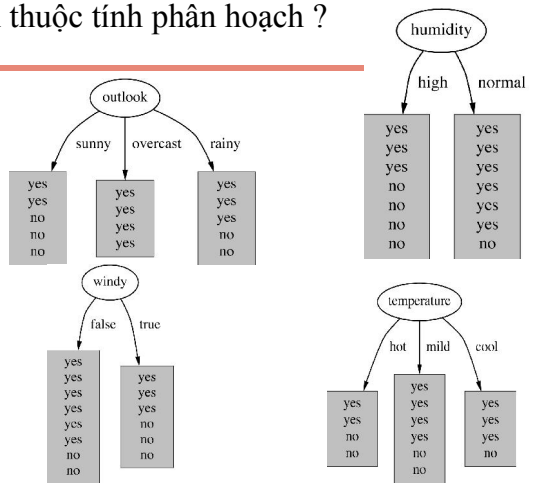
Giải thuật cây quyết định

- 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 trong 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
- cắt nhánh Bottom-up
  - cắt những cây con hoặc các nhánh từ dưới lên trên, để tránh học vẹt (overfitting, over learning)

Chọn thuộc tính phân hoạch

- ở mỗi nút, các thuộc tính được đánh giá dựa trên phân tách dữ liệu học **tốt nhất** có thể
- việc đánh giá 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)

Chọn thuộc tính phân hoạch ?



Chọn thuộc tính phân hoạch ?

- thuộc tính nào tốt ?
  - cho ra kết quả là cây nhỏ nhất
  - heuristics: chọn thuộc tính sinh ra các nút "purest" (thuần khiết)
- độ lợi thông tin
  - tăng với giá trị trung bình thuần khiết của các tập con của dữ liệu mà thuộc tính sinh ra
- chọn thuộc tính có độ lợi thông tin lớn nhất

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

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

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

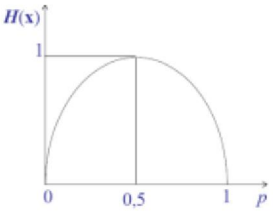
*Claude Shannon

Born: 30 April 1916
Died: 23 February 2001

"Father of information theory"

Entropy

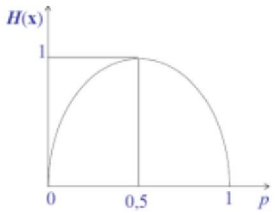
- Entropy là một đại lượng toán học dùng để đo lượng thông tin không chắc chắn (hay lượng ngẫu nhiên) của một sự kiện hay một phân phối ngẫu nhiên cho trước
- Entropy – uncertainty measure
- Entropy luôn >=0
  - **Entropy = 0?**
  - **Entropy = 1?**



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

- 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

25

Độ lợi thông tin

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

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

  p_i: xác suất mà phần tử trong dữ liệu D thuộc lớp C_i
- Độ hỗn loạn thông tin **sau** khi phân hoạch

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

  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

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

26

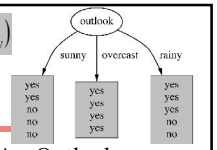
Ví dụ : thuộc tính outlook

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

27

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

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

Độ lợi thông tin

- 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)

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

Thuộc tính humidity

- "Humidity" = "High":

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

- "Humidity" = "Normal":

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

- thông tin của thuộc tính humidity

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

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

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

Độ lợi thông tin

- độ lợi thông tin của các thuộc tính

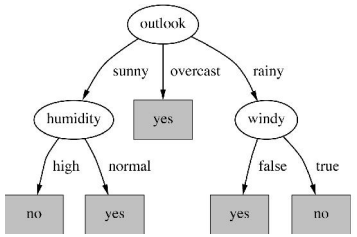
(trước khi phân hoạch) – (sau khi phân hoạch)

$$gain("Outlook") = 0.247 \text{ bits}$$
$$gain("Temperature") = 0.029 \text{ bits}$$
$$gain("Humidity") = 0.152 \text{ bits}$$
$$gain("Windy") = 0.048 \text{ bits}$$

Ví dụ : thuộc tính outlook

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

Kết quả

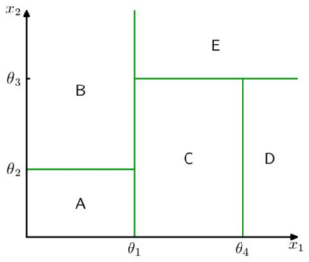


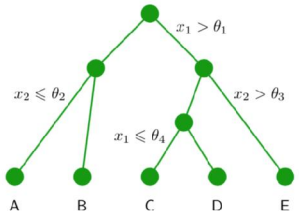
- chú ý : có thể có nút lá không thuần khiết

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

ID3 Algorithm

1. Create a root node for the tree
2. If all examples are positive, return leaf node ‘positive’
3. Else if all examples are negative, return leaf node ‘negative’
4. Calculate the entropy of current state E(S)
5. For each attribute, calculate the entropy with respect to the attribute ‘A’ denoted by E(S, A)
6. Select the attribute which has the maximum value of IG(S, A) and split the current (parent) node on the selected attribute
7. Remove the attribute that offers highest IG from the set of attributes
8. Repeat until we run out of all attributes, or the decision tree has all leaf nodes.





Chỉ số gini (CART)

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

  p_j là xác suất của lớp j trong T

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

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

Chỉ số gini (CART)

- sau khi phân hoạch T thành 2 tập con T1 & T2 với kích thước N1 & N2, chỉ số gini

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

- thuộc tính có **$gini_{split}$(T) nhỏ nhất** được chọn để phân hoạch

Ví dụ: chỉ số gini

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Ví dụ: chỉ số gini

- Tính Gini Outlook

Outlook	Yes	No	Number of instances
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

- 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) = $1 - (3/5)^2 - (2/5)^2 = 1 - 0.36 - 0.16 = 0.48$

Ví dụ: chỉ số gini

- Tính Gini Outlook

Outlook	Yes	No	Number of instances
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

- Gini(Outlook) = (5/14) x 0.48 + (4/14) x 0 + (5/14) x 0.48 = 0.171 + 0 + 0.171 = 0.342

Ví dụ: chỉ số gini

- Tính Gini Temperature

Temperature	Yes	No	Number of instances
Hot	2	2	4
Cool	3	1	4
Mild	4	2	6

- 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) = $1 - (4/6)^2 - (2/6)^2 = 1 - 0.444 - 0.111 = 0.445$

Ví dụ: chỉ số gini

- Tính Gini Temperature

Temperature	Yes	No	Number of instances
Hot	2	2	4
Cool	3	1	4
Mild	4	2	6

- 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

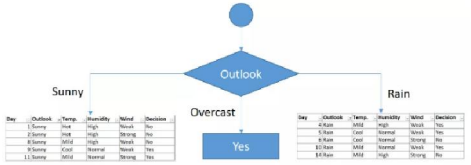
Ví dụ: chỉ số gini

- Tính Gini Humidity, Windy?

Ví dụ: chỉ số gini

- Tổng hợp các giá trị Gini

Feature	Gini index
Outlook	0.342
Temperature	0.439
Humidity	0.367
Wind	0.428



Ví dụ: chỉ số gini

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

Day	Outlook	Temp.	Humidity	Wind	Decision
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

Ví dụ: chỉ số gini

- 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

Ví dụ: chỉ số gini

- 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

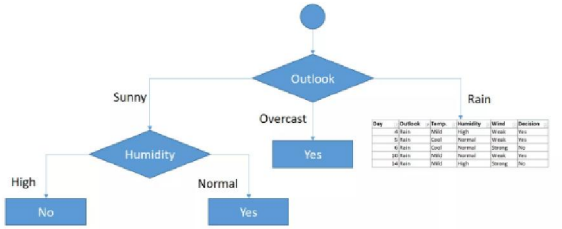
Ví dụ: chỉ số gini

- Khi Outlook = Sunny, các giá trị Gini của các đặc trưng lần lượt:

Feature	Gini index
Temperature	0.2
Humidity	0
Wind	0.466

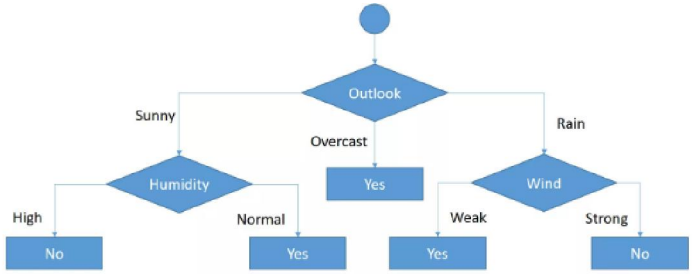
Ví dụ: chỉ số gini

- Cây quyết định tương ứng:



Ví dụ: chỉ số gini

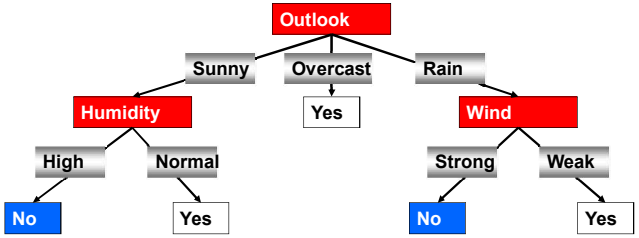
- Cách tính tương tự cho Rainy



Biến đổi cây quyết định thành luật

- **Biểu diễn tri thức dưới dạng luật IF-THEN**
- **Mỗi luật tạo ra từ mỗi đường dẫn từ gốc đến lá**
- **Mỗi cặp giá trị thuộc tính dọc theo đường dẫn tạo nên phép kết (phép AND – và)**
- **Các nút lá mang tên của lớp**

Biến đổi cây quyết định thành luật



R_1: If (Outlook=Sunny) ∧ (Humidity=High) Then Play=No
R_2: If (Outlook=Sunny) ∧ (Humidity=Normal) Then Play=Yes
R_3: If (Outlook=Overcast) Then Play=Yes
R_4: If (Outlook=Rain) ∧ (Wind=Strong) Then Play=No
R_5: If (Outlook=Rain) ∧ (Wind=Weak) Then Play=Yes

Giải thuật

- giải thuật ID3/C4.5 (Quinlan, 1993)
  - sử dụng Gain ratio
  - xử lý dữ liệu số, liệt kê, nhiễu
- CART (Breiman et al., 1984)
  - sử dụng chỉ số Gini
  - xử lý dữ liệu số, liệt kê, nhiễu

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

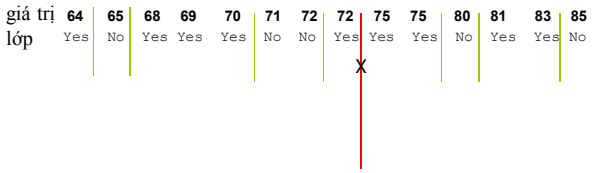
- phân hoạch trên thuộc tính temperature

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

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

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

https://www.mathsisfun.com/data/standard-deviation.html

Cây quyết định cho bài toán hồi quy

- Số lượng người chơi golf trung bình
  = (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)
  = $\sqrt{[(25 - 39.78)^2 + (30 - 39.78)^2 + (46 - 39.78)^2 + \ldots + (30 - 39.78)^2)/14]}$
  = 9.32

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

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

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

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
  = (25 + 30 + 35 + 38 + 48)/5 = 35.2
- Độ lệch chuẩn (Standard deviation) số lượng người chơi
  = $\sqrt{(((25 - 35.2)^2 + (30 - 35.2)^2 + (35 - 35.2)^2 + (48 - 35.2)^2)/5)}$ = 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
- Độ lệch chuẩn (Standard deviation) số lượng người chơi

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
  = (46 + 43 + 52 + 44)/4 = 46.25
- Độ lệch chuẩn (Standard deviation) số lượng người chơi
  = $\sqrt{(((46-46.25)^2+(43-46.25)^2+\ldots)}$= 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
- Độ lệch chuẩn (Standard deviation) số lượng người chơi

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
  = (45+52+23+46+30)/5 = 39.2
- Độ lệch chuẩn (Standard deviation) số lượng người chơi
  = $\sqrt{(((45 - 39.2)^2+(52 - 39.2)^2+\ldots)/5)}$=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

Độ lệch chuẩn của thuộc tính Outlook
= (4/14)x3.49 + (5/14)x10.87 + (5/14)x7.78 = 7.66

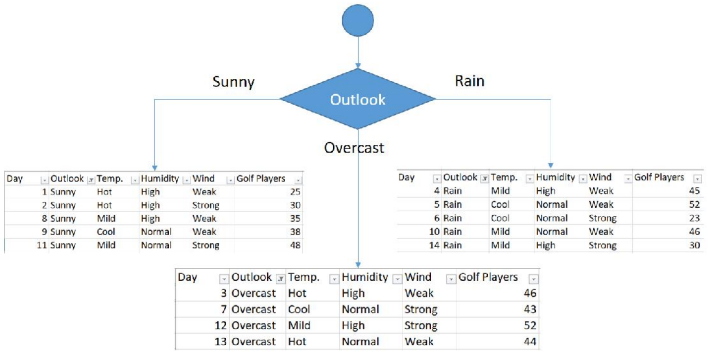
$$S(T,X) = \sum_{c \in X} P(c)S(c)$$

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

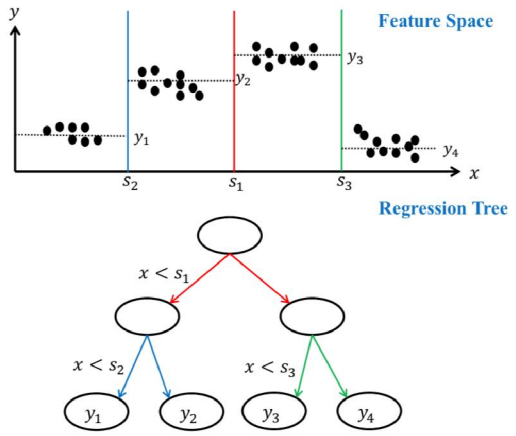
Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

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

- Golf players for sunny outlook = {25, 30, 35, 38, 48}
- Standard deviation for sunny outlook = 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

- Golf players for sunny outlook = {25, 30, 35, 38, 48}
- Standard deviation for sunny outlook = 7.78
- Notice that we will use this standard deviation value as global standard deviation for this sub data set.

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

- Golf

Sunny outlook and Hot Temperature

Standard deviation for sunny outlook and hot temperature = 2.5

Cây quyết định cho bài toán hồi quy

Summary of standard deviations for temperature feature when outlook is sunny

Weighted standard deviation for sunny outlook and temperature = (2/5)x2.5 + (1/5)x0 + (2/5)x6.5 = 3.6

Standard deviation reduction for sunny outlook and temperature = 7.78 – 3.6 = 4.18

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

- Golf

Sunny outlook and Hot Temperature

Standard deviation for sunny outlook and hot temperature = 2.5

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

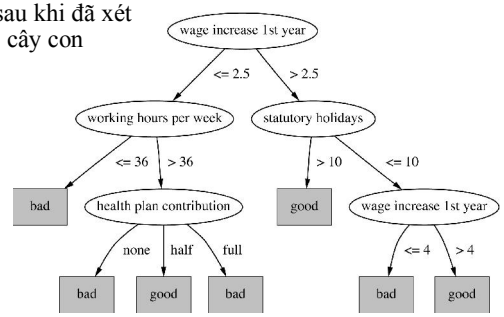
76

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

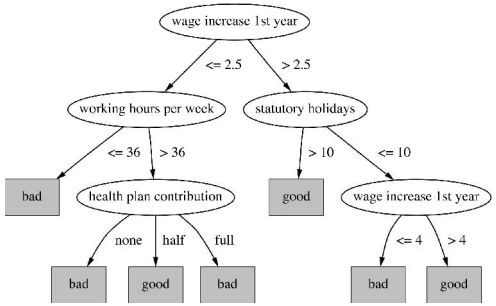
Thay thế cây con

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

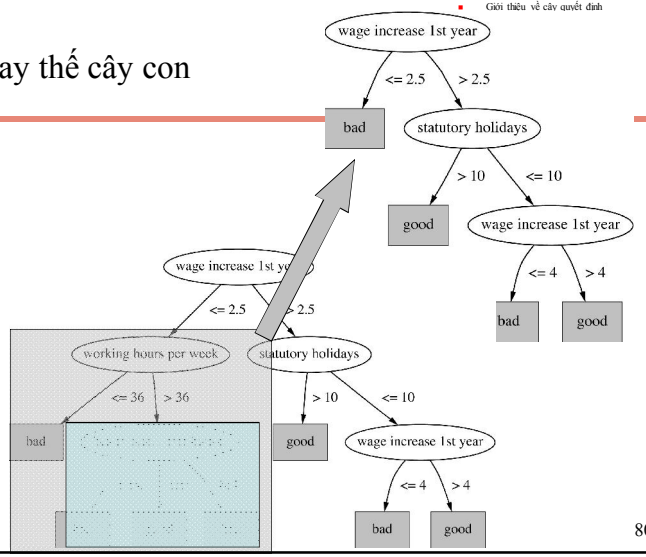


Thay thế cây con

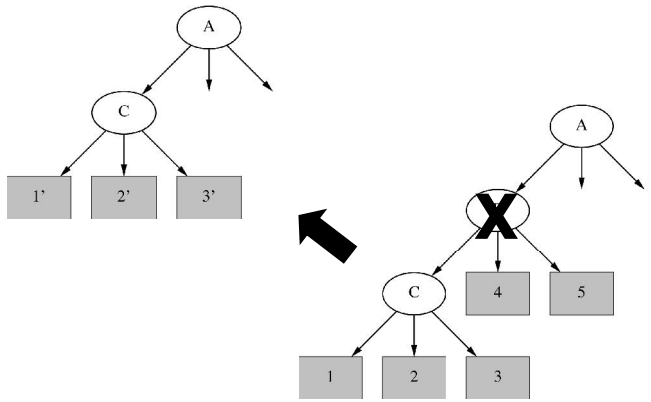
- thay thế cây con nào?



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

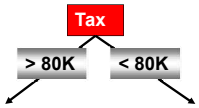
Hướng phát triển

- 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

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 số gini
  - Chọn giá trị phân chia có chỉ số gini thấp nhất

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 mean squared error)

88