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

37

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

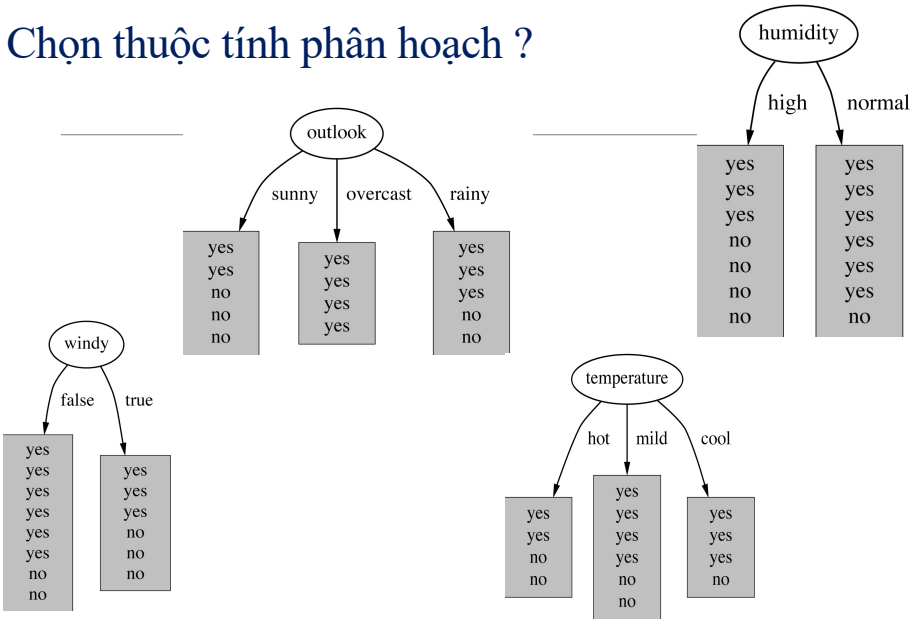
- gini(T) là nhỏ nhất nếu những lớp trong T bị lệch

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

- 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

38

Chọn thuộc tính phân hoạch ?



39

39

Xây dựng cây với chỉ số gini

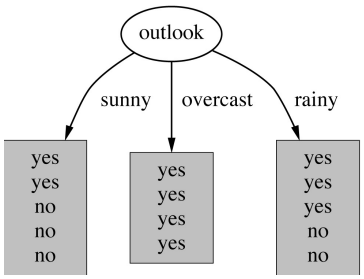
Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes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

40

40

Xây dựng cây với chỉ số gini

Tính Gini cho thuộc tính Outlook



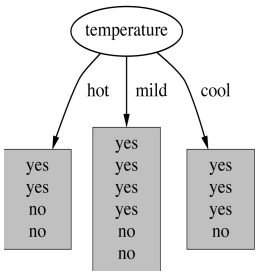
Gini(Outlook=Sunny) = $1 - [(2/5)^2 + (3/5)^2] = 1 - 0.16 - 0.36 = 0.48$

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

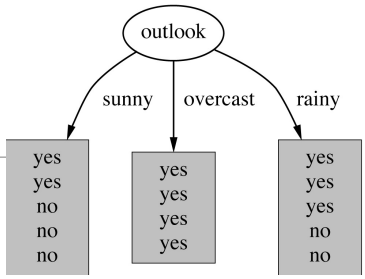
Gini(Outlook=Rain) = $1 - (3/5)^2 - (2/5)^2 = 1 - 0.36 - 0.16 = 0.48$

41

41



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

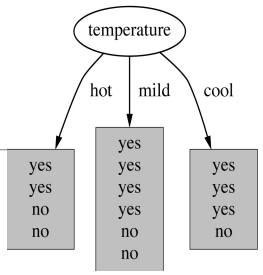
Gini(Outlook) = (5/14) x 0.48 + (4/14) x 0 + (5/14) x 0.48

= 0.171 + 0 + 0.171 = 0.342

42

42

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) = $1 - (3/4)^2 - (1/4)^2 = 1 - 0.5625 - 0.0625 = 0.375$

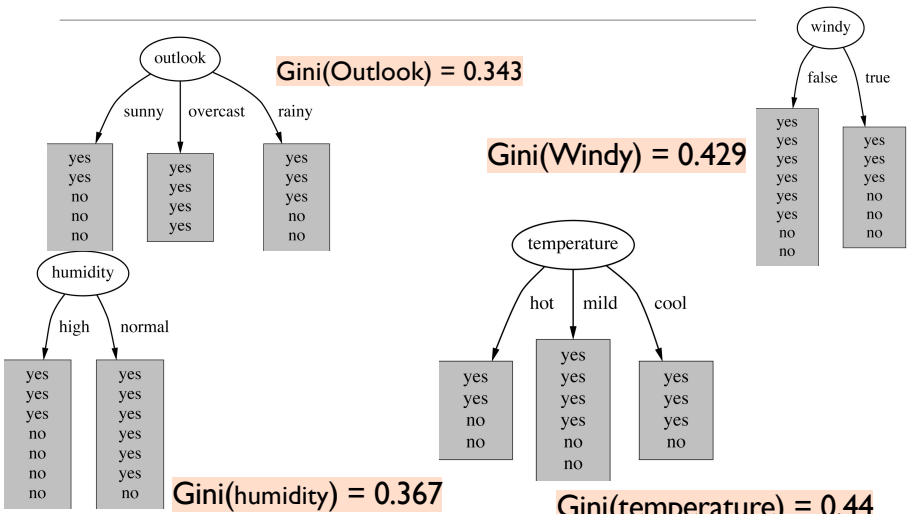
Gini(Temp=Mild) = $1 - (4/6)^2 - (2/6)^2 = 1 - 0.444 - 0.111 = 0.445$

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

43

43

Tương tự tính Gini cho thuộc tính Humidity và Windy



44

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

47

47

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

48

48

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
```

49

49

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
  - Sắp xếp dữ liệu từ thấp đến cao hoặc ngược lại
    - Chọn giá trị chính giữa để phân hoạch
    - tính độ lợi thông tin cho mọi giá trị phân nhánh của thuộc tính và chọn giá trị phân nhánh tốt nhất

50

50

Giải thuật C4.5, dữ liệu kiểu số => phân hoạch nhị phân
ví dụ : temp < 45

- 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
- cần sắp xếp dữ liệu

51

51

Cải tiến

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

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

X

điểm giữa của cùng lớp không phải điểm tối ưu

52

52

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

53

Cây quyết định cho bài toán hồi quy

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

54

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 55

55

Cây quyết định cho bài toán hồi quy

CART - Regression Trees (Brieman et al. 84)

- Tính độ lệch chuẩn cho cột nhãn (Gold Players)
- 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

56

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)

$\sigma = \sqrt{[((25 - 39.78)^2 + (30 - 39.78)^2 + (46 - 39.78)^2 + \ldots + (30 - 39.78)^2)/14]}$

= 9.32

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

57

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

μ = (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

58

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

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

Độ lệch chuẩn (Standard deviation) khi outlook = overcast

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

59

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

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

60

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$_{\text{Outlook}}$ = 9.32 – 7.66 = **1.66**

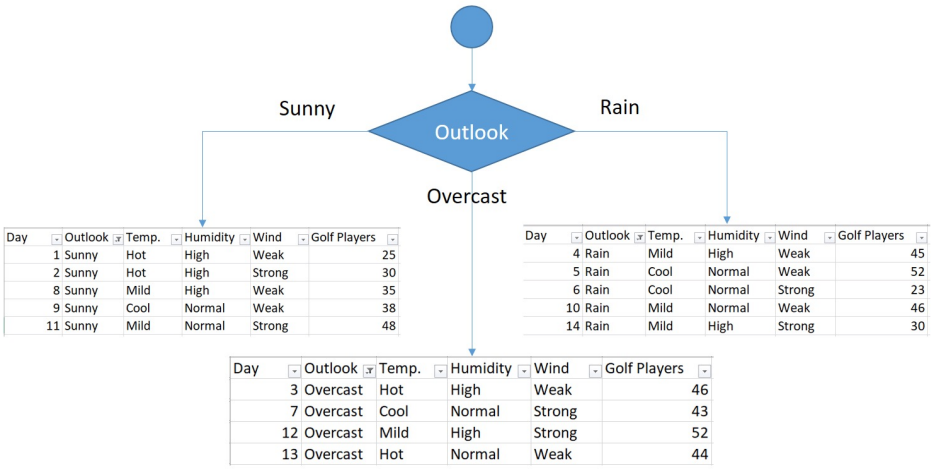
61

Cây quyết định cho bài toán hồi quy

	Standard Deviation Reduction
Outlook	1.66
Temperature	0.47
Humidity	0.27
Wind	0.29

62

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

63

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.

64

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

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

65

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

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

66

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

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

Độ lệch chuẩn (Standard deviation) khi outlook = sunny

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

67

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

68

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

69

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

70

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

71

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

72

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

73

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

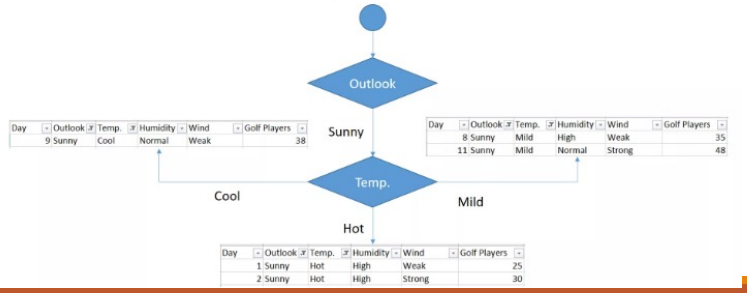
Độ lệch chuẩn khi outlook=sunny và xét thuộc tính windy
= (2/5)*9 + (3/5)x5.56 = 6.93
Độ chênh lệch của độ lệch chuẩn khi outlook=sunny và outlook =sunny + thuộc tính windy = 7.78 – 6.93 = 0.85

74

Cây quyết định cho bài toán hồi quy

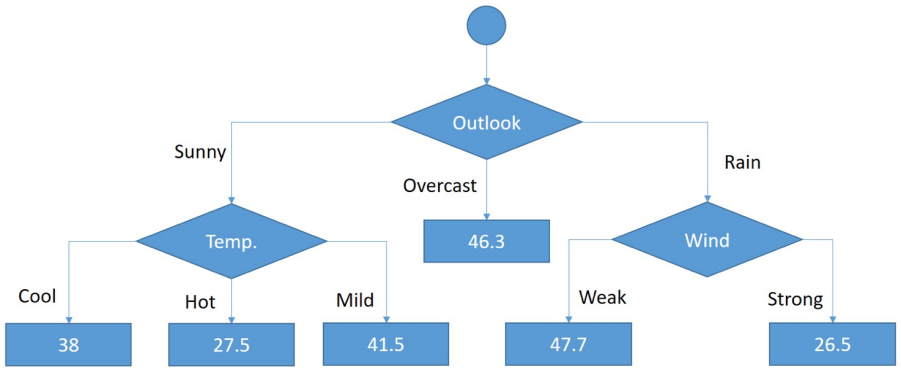
Cây quyết định được xây dựng:

Feature	Standard Deviation Reduction
Temperature	4.18
Humidity	3.33
Wind	0.85



75

Cây quyết định cho bài toán hồi quy



76

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

77

77

Postpruning

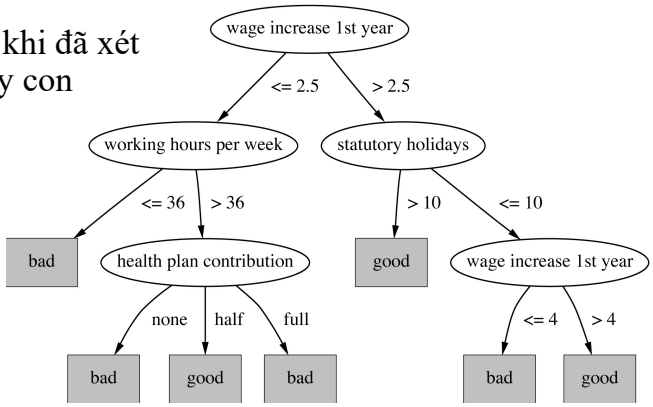
- xây dựng cây đầy đủ
- cắt nhánh
  - *thay thế cây con*
  - *đưa cây con lên trên*
- có nhiều chiến lược
  - ước lượng lỗi
  - significance test

78

78

Thay thế cây con

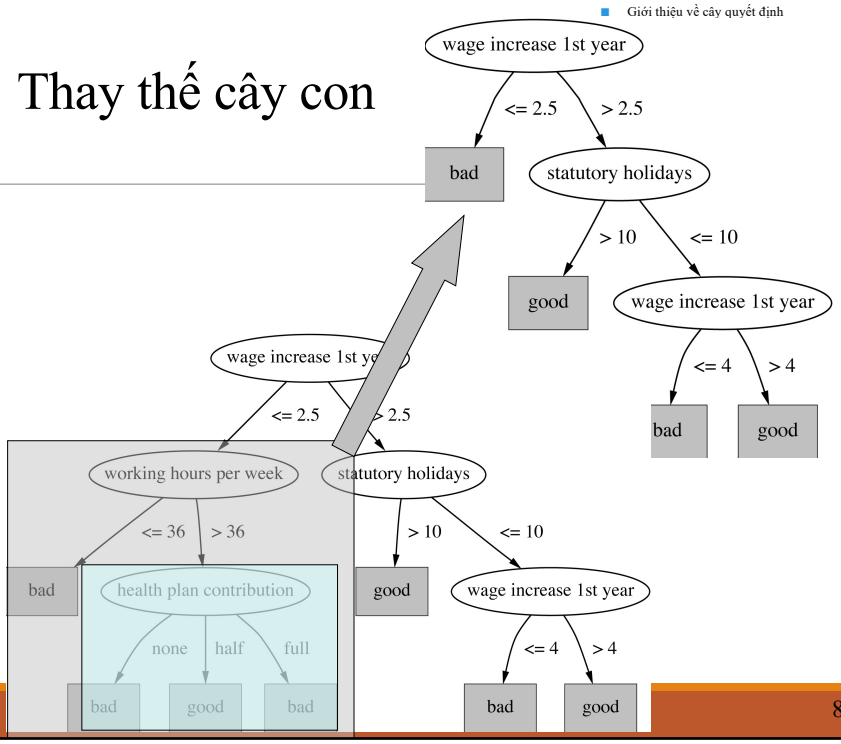
- *Bottom-up*
- thay thế sau khi đã xét tất cả các cây con



79

79

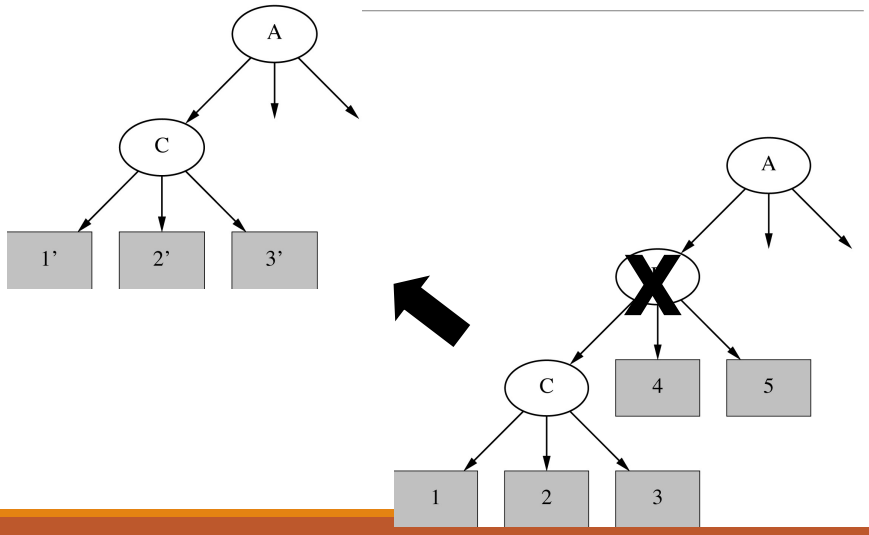
Thay thế cây con



80

80

Đưa cây con lên trên



81

81

Nội dung

Giới thiệu về cây quyết định

Giải thuật học của cây quyết định

Kết luận và hướng phát triển

82

82

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

- tăng độ chính xác
- xử lý dữ liệu không cân bằng
- dữ liệu phức tạp có số chiều lớn