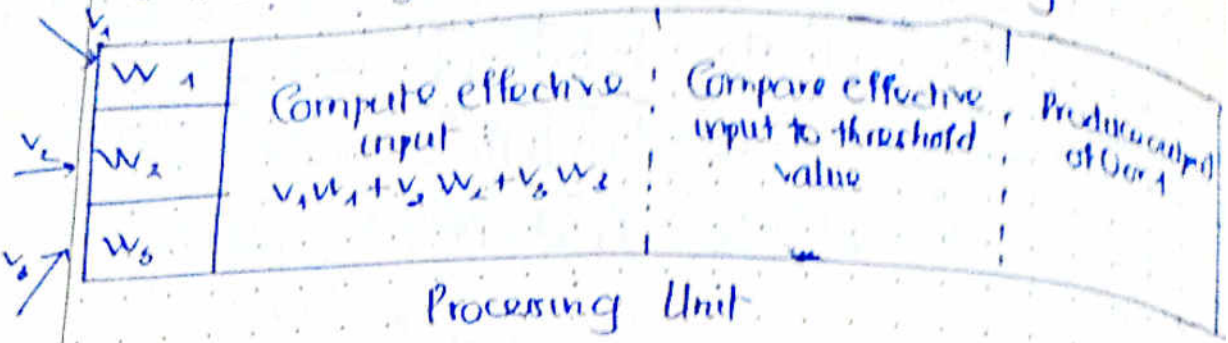
Key points

nhận dạng mẫu

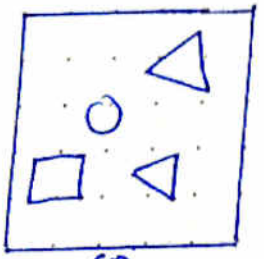
B. Hoạt động bên trong của một đơn vị xử lý

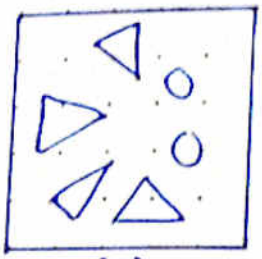


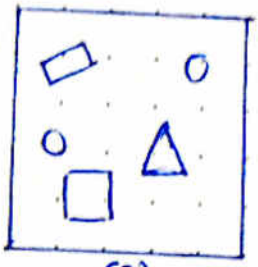
Processing Unit

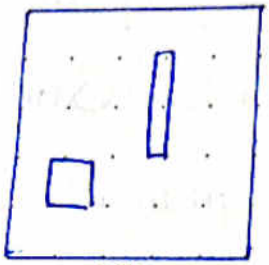
HỌC QUA LOGIC

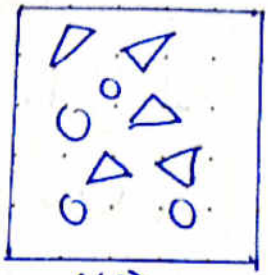
Lớp B


(6)


(7)


(8)


(9)


(10)

P1: 3 hình tam giác

P2: 3 hình tròn

....

Lớp	Hình	Tam giác (P1)	Hình tròn (P2)	Ovan (P3)	Chữ nhật (P4)	Đa giác (P5)
B	6	1	1	0	1	0
B	7	1	1	0	0	0
B	8	1	1	0	1	0
B	9	0	0	0	1	0
B	10	1	0	1	0	0

ary

$q_1 = \bar{P}_1 P_2 \bar{P}_3 P_4 \bar{P}_5$

$q_2 = P_1 P_2 \bar{P}_3 \bar{P}_4 \bar{P}_5$

$q_3 = P_1 P_2 \bar{P}_3 P_4 \bar{P}_5$

$q_4 = \bar{P}_1 \bar{P}_2 \bar{P}_3 P_4 \bar{P}_5$

$q_5 = P_1 \bar{P}_2 P_3 \bar{P}_4 \bar{P}_5$

$X \in D$: X thoả

$q = P_1 P_2 \bar{P}_3 P_4 \bar{P}_5 \vee P_1 P_2 \bar{P}_3 \bar{P}_4 \bar{P}_5 \vee P_1 P_2 \bar{P}_3 P_4 \bar{P}_5$
$\vee \bar{P}_1 \bar{P}_2 \bar{P}_3 P_4 \bar{P}_5 \vee P_1 \bar{P}_2 P_3 \bar{P}_4 \bar{P}_5$

$X \in D$: $X \in$ [

II. Học qua sự quan sát

1. Ý tưởng thuật toán Quinlan:

+ Cho một bảng quan sát là tập hợp các mẫu với các thuộc tính nhất định của các đối tượng nào đó.

+ Sử dụng một độ đo để đánh giá và đề ra tiêu chuẩn nhằm chọn lựa một thuộc tính mang tính chất "phân loại" để phân bảng thành các bảng con nhỏ hơn, sao cho từ mỗi bảng con này dễ dàng phân tích hơn ra một quy luật chung

(Đệ quy)

+ Tạo độ thuật toán dựa "Cây quyết định" cho thấy thứ tự của các thuộc tính đang xét.

VD: Xác định là người châu Á hay châu Âu khi xem xét một nhóm người căn cứ trên: hình dáng, chiều cao và giới tính.

- Định nghĩa độ đo V:

+ $V(H.Dáng = To) = (A'_{To}, Âu_{To})$

. A'_{To} = Tổng số quan sát châu A' có (H.Dáng = To) / Tổng số quan sát có (H.Dáng = To)

. $Âu_{To}$ = Tổng số quan sát châu Âu có (H.Dáng = To) / Tổng số quan sát có (H.Dáng = To)

- Tiêu chuẩn phân loại: Chọn thuộc tính có Nhiều vectơ đơn vị nhất.

Giả sử có bảng quan sát sau:

STT	Hình dáng	Chiều cao	Giới tính	Quan sát
1	To	Trung bình	Nam	Châu Á
2	Nhỏ	Thấp	Nam	Châu Á
3	Nhỏ	Trung bình	Nam	Châu Á
4	To	Cao	Nam	Châu Âu
5	Nhỏ	Trung bình	Nữ	Châu Âu
6	Nhỏ	Cao	Nam	Châu Âu
7	Nhỏ	Cao	Nữ	Châu Âu
8	To	Trung bình	Nữ	Châu Âu

- Tính toán:

+ Lần 1:

• Hình dáng:

V(H.Dáng = To) = (113, 213)

V(H.Dáng = Nhỏ) = (215, 315)

• Chiều cao:

V(Chiều cao = TB) = (112, 112)

V(Chiều cao = Thấp) = (1, 0) *

V(Chiều cao = Cao) = (0, 1) *

• Giới tính:

V(Giới tính = Nam) = (315, 215)

V(Giới tính = Nữ) = (0, 1)

STT	Hình dáng	Giới tính	Quan sát
1	To	Nam	Châu Á
3	Nhỏ	Nam	Châu Á
5	Nhỏ	Nữ	Châu Âu
8	To	Nữ	Châu Âu

+ Lần 2:

• Hình dáng:

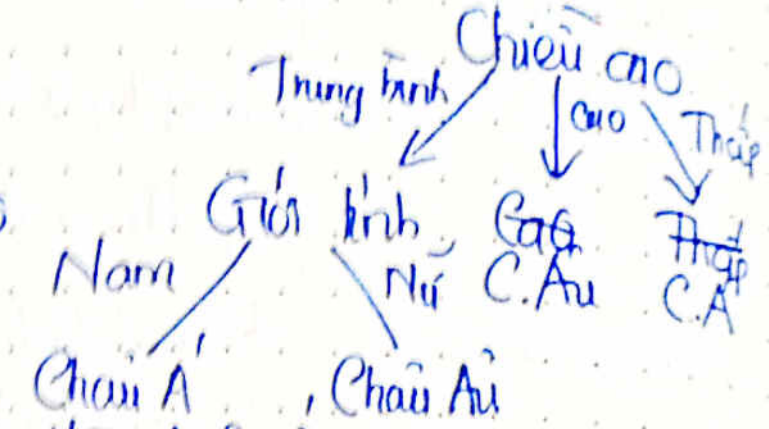
V(H.Dáng = To) = (112, 112)

V(H.Dáng = Nhỏ) = (112, 112)

. Giới hạn:

V(G.Tính = Nam) = (1,0)

V(G.Tính = Nữ) = (0,1)



- Luật:

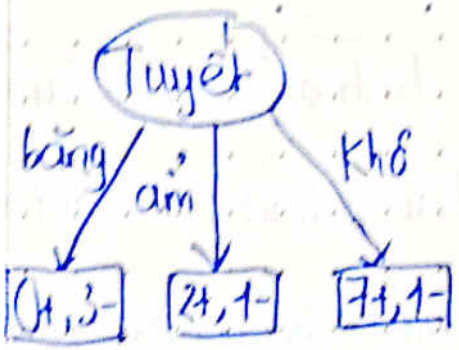
+ Nếu chiều cao = Cao thì châu Âu

+ Nếu chiều cao = Thấp thì châu Á

+ Nếu chiều cao = T.bình và G.Tính = Nam thì châu Á

+ Nếu chiều cao = T.bình và G.Tính = Nữ thì châu Âu

2. Chọn thuộc tính lần 1

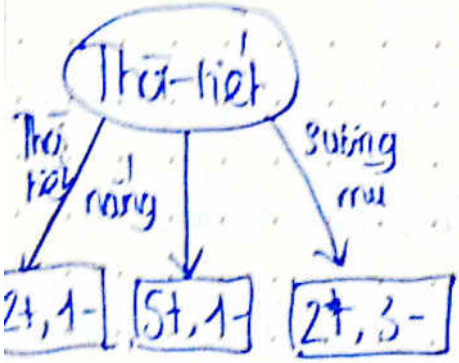


$H_{băng} = -0{,}3 * \log_2 0{,}3 - 3{,}3 * \log_2 3{,}3 = 0$

$H_{ấm} = -2{,}3 * \log_2 2{,}3 - 1{,}3 * \log_2 1{,}3 = 0{,}918$

$H_{khô} = -7{,}8 \log_2 2{,}3 - 1{,}8 * \log_2 1{,}8 = 0{,}54$

$AE = 3{,}14 * 0 + 3{,}14 * 0{,}918 + 8{,}14 * 0{,}544 = 0{,}508$



$H_{gió} = -2{,}3 * \log_2 2{,}3 - 1{,}3 * \log_2 1{,}3 = 0{,}918$

$H_{ráng} = -5{,}6 * \log_2 5{,}6 - 1{,}6 * \log_2 1{,}6 = 0{,}65$

$H_{sương\ mù} = -2{,}5 * \log_2 2{,}5 - 3{,}5 * \log_2 3{,}5 = 0{,}971$

$AE = 3{,}14 * 0{,}918 + 5{,}14 * 0{,}65 + 0{,}971 * 5{,}14 = 0{,}822$

Date: ...../...../.....

3. Học là gì?

- Học là quá trình tiếp nhận tri thức mới hoặc cập nhật tri thức về hành vi, kĩ năng, ... và biến quan niệm thông tin ≠ nhau.

+ Thay đổi để tốt hơn (Theo 1 điều kiện định trước) khi có tình huống tương tự xảy ra.

+ Học ko phải là "học thuộc lòng"

- Học là một trong những khả năng quan trọng của con người.

4. Các mức độ của học

- Học là 1 quá trình được chia làm nhiều mức độ:

1. Được truyền tải tri thức và ghi nhớ chúng.

2. Tiếp nhận tri thức thông qua quan sát và thăm dò các sự kiện.

3. Cải thiện tri thức thông qua luyện tập các kỹ năng vận động và nhận thức.

4. Tổ chức tri thức mới thành các biểu diễn hợp quát, hiệu quả.

Date: ...../...../.....

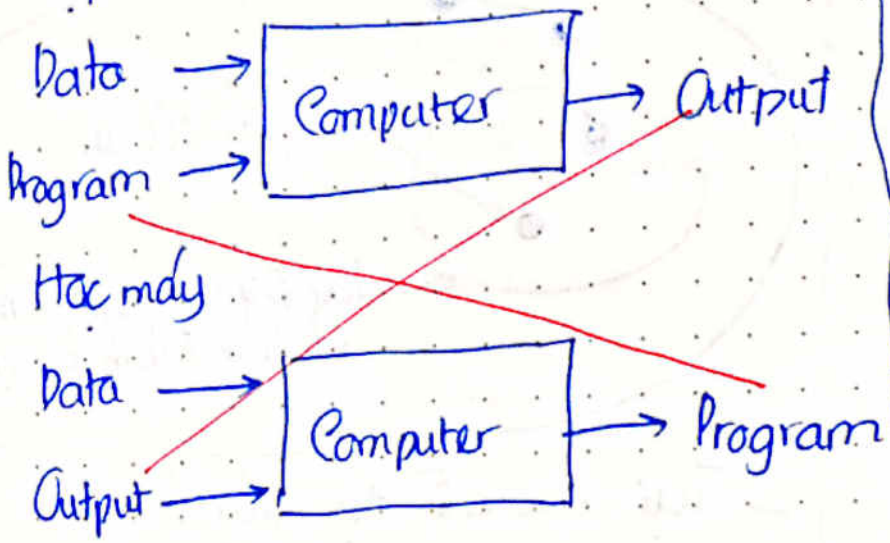
5. Học máy là gì?

- Học máy (Machine Learning) nghiên cứu các thay đổi mang tính thích nghi trong hệ thống, cho phép hệ thống thực thi những tác vụ tương tự hiệu quả hơn qua nhiều lần. (Herbert Simon)

Learning - Pattern Detection - Data - Self - Programming

- Nhiệm vụ của học máy là thiết kế các chương trình máy tính có thể học luật từ dữ liệu, thích nghi với thay đổi và cải thiện hiệu quả thông qua kinh nghiệm.

Lập trình hệ thống



Học máy

Algorithms = Hạt giống

Data = Chất dinh dưỡng

You = Kẻ làm vườn

Programs = Cây

6. Tại sao học lại khó?

- Cho trước một lượng hữu hạn dữ liệu huấn luyện. Ta cần suy ra một quan hệ trong không gian vô hạn.

- Trong thực tế, có vô số quan hệ như thế.

Date: ...../...../.....

7. Nguyên lý Occam Razor

- Nguyên lý: Giải pháp đơn giản là giải pháp tốt. (William of Occam)

- Khi có nhiều giải pháp cho cùng một vấn đề, ta nên chọn giải pháp đơn giản nhất.

- Ta quan niệm thế nào là đơn giản?

+ Sd tri thức hiển nó cj bài toán để định nghĩa giải pháp đơn giản. Vd: Smoothness prior.

8. Học máy và tìm kiếm

- Tại sao cần học máy?

+ Sự phát hiện cj học máy nhờ vào 1 số yếu tố sau:

. Hiển bổ mới về lý thuyết và thuật toán.

. Dữ liệu mức khuyến tăng mạnh như vũ bão.

. Sức mạnh tính toán đã sẵn sàng.

. Ngành công nghiệp phát triển nở rộ

9. Ứng dụng của học máy

- Ba lĩnh vực thích hợp cho học máy:

+ Khai thác dữ liệu: Sử dụng dữ liệu thu thập trong quá khứ để cải thiện quyết định.

+ Các phần mềm ứng dụng có thể lập trình bằng thủ công

+ Các chương trình hỉ điều chỉnh.

+ ...

- Xử lý thị giác:

+ Phát hiện / xác minh khuôn mặt

+ Nhận dạng chữ viết tay

- Xử lý hàng nói: Nhận dạng ân, vị, chữ / câu / đối tg X.

- Các ứng dụng khác:

+ Đánh chỉ mục: Google, khai thác văn bản, truy vấn thông tin.

+ Tài chính: Dự đoán khả năng vay nợ, quản lý rủi ro và đầu tư

+ Truyền thông: Dự đoán tình hình giao thông.

+ Trò chơi: Cờ vua, cờ vây

Date: ...../...../.....

+ Điều khiển: Huấn luyện Robot

... và còn rất nhiều ứng dụng.

- Học máy được ứng dụng trong những lĩnh vực:

+ Con k có thể tham gia vào quá trình.

+ Con k có thể giải thích về chuyên môn.

+ Các giải pháp thay đổi theo thời gian.

+ Giải pháp cần ~~giải thích~~ giải thích nghi vs ~~trường hợp~~ cụ thể.

- Các lĩnh vực liên quan: Electrical engineering, Statistics, Philosophy, Neuroscience, Bychology, CS,

+ Xác suất thống kê: Làm khớp mô hình với dữ liệu và kiểm tra kết quả.

+ Khai thác dữ liệu / Phân tích dữ liệu thăm dò: Phát hiện hình mẫu trong dữ liệu.

+ Lý thuyết điều khiển thích nghi: Học mức huyền các mô hình và dùng chúng để đạt tới một số mục tiêu.

+ Trí tuệ nhân tạo: Xây dựng máy thông minh.

10. Các dạng học máy

10.1. Học có giám sát (Supervised learning)

- Dựa vào các mẫu huấn luyện đã biết phân lớp
- Bộ học chịu hướng dẫn để nhận diện đặc trưng cho mỗi phân lớp.

10.2 Học không giám sát (Unsupervised learning)
- Dựa vào mẫu huấn luyện chưa biết phân lớp
- Bộ học cần phải hiểu hiểm thông tin trong hoàn cảnh nhiều sự hướng dẫn.

10.3 Học tăng cường (Reinforcement learning)
- Dựa vào mẫu huấn luyện chưa biết phân lớp
- Bộ học hành động hướng đến một hình thưởng có mức kết quả phản hồi.

→ Tự giải thích phản hồi thông qua việc tạo được tín hiệu hướng dẫn.

11. Các mức độ học máy
- Mức độ 1: Ghi nhớ: Dự đoán mẫu đã từng học trước đó
- Mức độ 2: Lấy trung bình để xử lý nhiễu
+ Dự đoán sai lệch do thiết bị.

12. Phân lớp là gì?
- Phân lớp là 1 bài toán học với một hàm mục tiêu f mà

Date: ...../...../.....

ánh xạ tập các thuộc tính x tới một nhãn lớp y.

- Một trong các thuộc tính là thuộc tính lớp.
- Hai nhãn lớp (hoặc là lớp): Yes (1), No (0)

Input → Mô hình → Output

Tập thuộc tính (X) → phân lớp → Nhãn lớp (y)

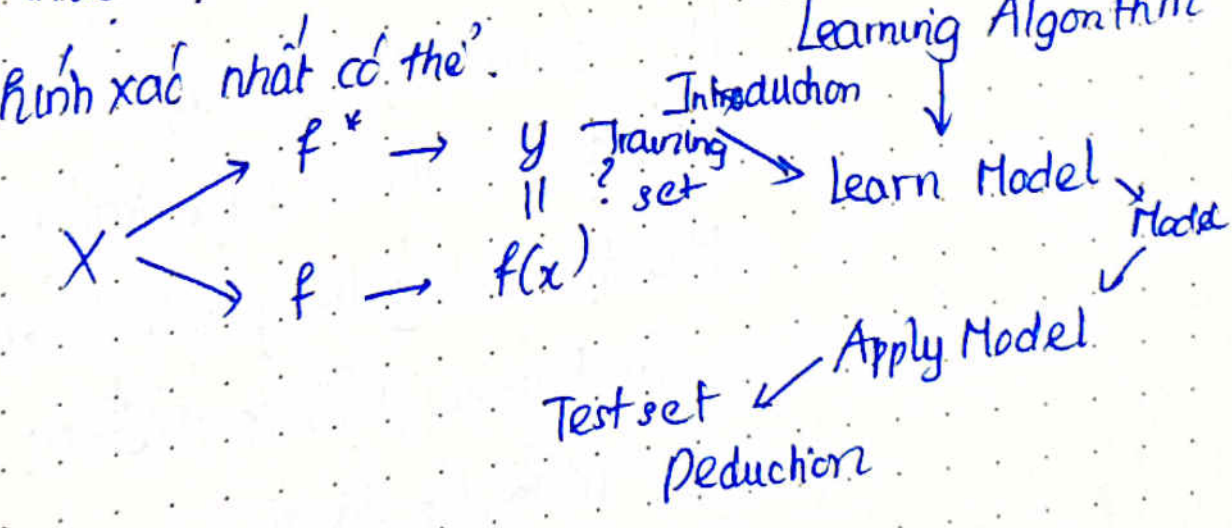
13. Học có giám sát

Chúng ta có các quan sát: $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ và 1 hàm f^* nào đó. Qua hàm f^* này, tạo thành các nhãn (Label): $y^{(1)} = f^*(x^{(1)})$ $y^{(2)} = f^*(x^{(2)})$

$\ldots$ $y^{(N)} = f^*(x^{(N)})$

Nhưng hàm f^* bị mất. Chúng ta chỉ biết: $(x^{(i)}, y^{(i)})$

Nhiệm vụ: Khôi phục lại f^* bằng mô hình f_w 1 cách chính xác nhất có thể.



14. Hunt's Algorithm

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

(a) Defaulted - No
(7,3)

(b) Home Owner
Yes → Defaulted - Yes (3,0)
No → Defaulted - No (4,3)

(c) Home Owner
Yes → Defaulted - Yes (3,0)
No → Marital Status
Single, Divorced → Defaulted - Yes (1,3)
Married → Defaulted - No (3,0)

(d) Home Owner
Yes → Defaulted - Yes (3,0)
No → Marital Status
Single, Divorced → Annual Income
< 80k → Defaulted - No (1,0)
> 80k → Defaulted - Yes (0,3)
Married → Defaulted - No (3,0)

⊛ Thuật toán ID3

- Được phát triển dựa trên bởi Quinlan ở lĩnh vực Trí tuệ nhân tạo và Breiman, Friedman, Olsen và Stone ở thống kê.

- Lặp:

1. Chọn A ← E hình quyết định "tốt nhất" cho nút kế tiếp.
2. Gán A là thuộc tính quyết định cho nút.
3. Với mỗi giá trị của A, tạo nhánh con mới của nút.
4. Phân loại các mẫu huấn luyện cho các nút lá.
5. Nếu các mẫu huấn luyện được phân loại hoàn thì NGƯNG. Ngược lại, lặp với các nút lá mới.

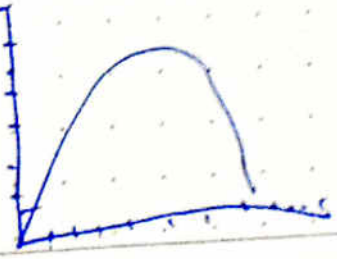
- Thuộc tính tốt nhất là gì?

⊛ Độ đo Entropy

- Gọi S là tập ngẫu nhiên mẫu huấn luyện, p⊕ tỷ lệ các mẫu dg ở S

- $H \equiv -p_{\oplus} \log_2 p_{\oplus} - (1-p_{\oplus}) \log_2 (1-p_{\oplus})$

- Nếu có n hơn 2 lớp?



Date: ..../..../....

- Thuộc tính tốt nhất tối thiểu hoá Entropy trung bình của dữ liệu trong các nút con.

$$\text{Average Entropy}(A) = \sum_{v \in \text{Value}(A)} P_v H_{A=v}$$

15. Cây quyết định

15.1 Áp dụng

Dữ liệu quan sát → Cây quyết định → Dự đoán = dữ liệu chưa biết

Luật quyết định → Dự đoán = dữ liệu chưa biết

15.2 Thuật toán

- Có n thuật toán xây dựng cây quyết định: Hunt's Algorithm, CART, ID3, C4.5, Cây định danh

15.3 Từ CSDL đến cây định danh

- Cơ sở dữ liệu:
  + Tập các mẫu quan sát được ở thực tế
  + Mỗi mẫu có 1 ∈ hình cầu cũng là ∈ tính dùng để gán lớp (dự đoán / định danh đối tg)
  + Sd để huấn luyện cây quyết định

VD: 3 màu tóc × 3 c cao × 3 cân nặng × 2 sở thích = 54 tổ hợp có thể.

Xác suất 1 mẫu mới trùng với mẫu quan sát

Thực tế: Số thuộc tính và giá trị thuộc tính rất lớn 85% 15%

→ K° thể làm = cách so khớp tra bảng

+ Hướng giải quyết

• Đề xuất 1 thủ tục phân lớp chính xác từng mẫu

• Thủ tục đúng = số lượng chỉ tiêu mẫu → có thể đúng = các mẫu mà nhãn lớp là chưa biết

VD:

Name	Hair color	Height	Weight	Lotion	Result
Sarah	blonde	average	light	no	sunburned
Dana	blonde	tall	average	yes	none
Alex	brown	short	average	yes	none
Annie	blonde	short	average	no	sunburned
Emily	red	short	average	no	sunburned
Pete	brown	average	heavy	no	none
John	brown	average	heavy	no	none
Katie	blonde	short	light	yes	none

_ Cây định danh: là g mỗi nhánh chỉ cho ra 1 kết quả

_ Occam's Razor:

+ Thế giới vốn dĩ đơn giản

+ Cây định danh nhỏ nhất mà nhất quán với các mẫu

→ cây có khả năng nhất Ọ vc phân lớp chính xác các đtg chưa biết. Cây định danh (1) tốt hơn (2)

- Độ hỗn loạn trung bình : Average disorder

$$= \sum_b \left(\frac{n_b}{n_t}\right) \times \left(\sum_c - \frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b}\right)$$

Trong đó :

+ n_b là số mẫu nhánh b
+ n_t là tổng số mẫu g tất cả các nhánh,
+ n_{bc} là tổng số mẫu g nhánh b ∈ lớp c

- Độ hỗn loạn :

$$\text{Disorder} = \sum_c - \frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b}$$

+ Mô tả sự "hỗn loạn" ⊙ 1 tập dữ liệu :
  . ĐHL = 0 nếu tập dữ liệu là đồng nhất (chỉ gồm 1 lớp)
  . ĐHL = 1 nếu tập DL chứa tất cả các lớp.

- Xây dựng cây định danh :

+ Chỉ có thể xây dựng cây "nhỏ", mặc dù ko đảm bảo là "nhỏ nhất"

+ Thủ tục : Chia để trị
  . Tìm test cho nút gốc : Chia CSDL thành các tập con sao cho càng nhiều mẫu ∈ cùng 1 lớp càng tốt.
  . Với mỗi tập chứa ≥ hơn 1 lớp, chọn 1 test ≠ để chia tiếp

Key points

Date: ...../...../.....

Không đồng nhất = các tập con đồng nhất.

+ Điều kiện dừng:

. Mỗi nút lá chỉ gồm các mẫu đồng nhất.

. Ko còn ∈ tính nào có thể phân chia nữa.

⇒ Thuật ngữ cụ thể hoá độ hỗn loạn dữ liệu.

Hair color

blonde	red	brown
✓ Sarah	✓ Emily	Alex
Dana		Pete
✓ Annie		John
Katie		

Height

Short	tall	average
Alex ✗	Dana ✓	Sarah
✓ Annie		Pete ✗
✓ Emily ✗		John ✗
Katie		

Weight

light	average	heavy
✓ Sarah	Dana	Pete ✗
Katie	Alex ✗	John ✗
	✓ Annie	
	✓ Emily ✗	

Lotion Used

No	yes
✓ Sarah	Dana
✓ Annie	Alex ✗
✓ Emily ✗	Katie
Pete ✗	
John ✗	

Tính độ hỗn loạn trung bình

$$L1: AD_{\text{Hair color}} = \frac{4}{8} \times \left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{4}{8} \times \left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) + \frac{3}{8}\left(-\frac{3}{3}\log\frac{3}{23} - \frac{0}{3}\log\frac{0}{23}\right.$$

$$= 0,5 \quad (\text{Chọn})$$

nary

$$AD_{Height} = \frac{4}{8}\left(-\frac{4}{4}\log_2\frac{2}{4}\right) + \frac{1}{8}\cdot 0 + \frac{3}{8}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)$$
$$\approx 0,84436$$

$$AD_{Weight} = \frac{2}{8}\left(-\frac{2}{2}\log_2\frac{1}{2}\right) + \frac{4}{8}\left(-\frac{4}{4}\log_2\frac{2}{4}\right) + \frac{2}{8}\cdot 0$$
$$= 0,75$$

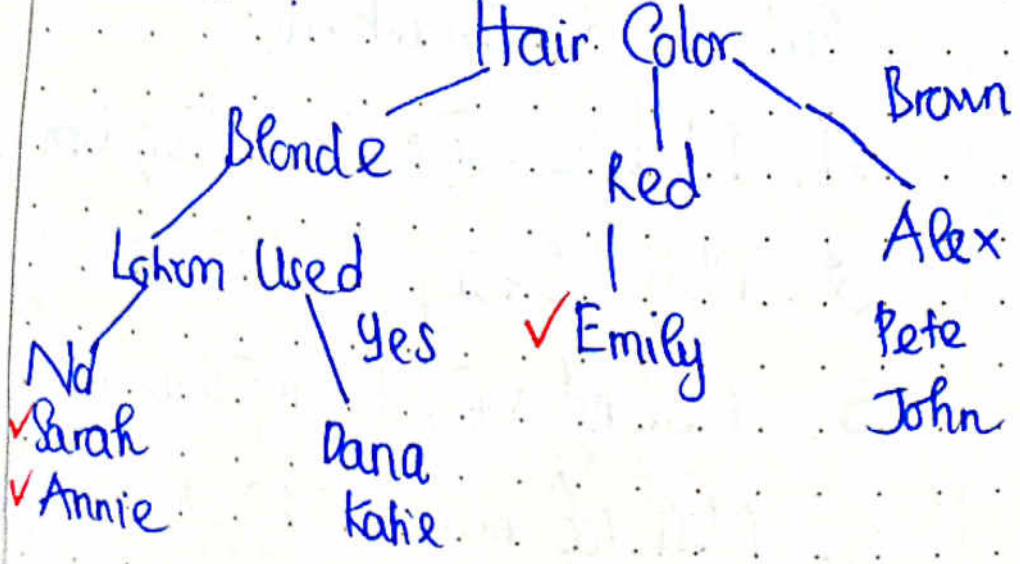
$$AD_{Lotion\ used} = \frac{5}{8}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) + \frac{3}{8}\cdot 0$$
$$\approx 0,6068$$

L2:
$$AD_{Height} = \frac{2}{4}\left(-\frac{2}{2}\log_2\frac{1}{2}\right) + \frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 0 = 0.5$$

$$AD_{Weight} = \frac{2}{4}\left(-\frac{2}{2}\log_2\frac{1}{2}\right) + \frac{2}{4}\left(-\frac{2}{2}\log_2\frac{1}{2}\right) + \frac{0}{4}\cdot 0$$
$$= 1$$

$$AD_{Lotion\ Used} = \frac{2}{4}\cdot 0 + \frac{2}{4}\cdot 0 = 0 \text{ (Chọn)}$$

Cây thì, dành cần hm:



- Thứ học đám chồi - SPROUTER:

+ Lặp đến khi: Mỗi nút lá chỉ gồm các mẫu đồng nhất hoặc ko đồng nhất hoặc ko còn ∈ hình nào có thể phân chia nữa.

• Chọn 1 nút lá bất kỳ có tập các mẫu ko đồng nhất.

• Thay thế nút lá đó bằng 1 nút test mà nó sẽ chia tập mẫu ko đồng nhất thành các tập ko đồng nhất ở mức độ tối thiểu dựa vào độ đo tính hỗn loạn.

15.4 Từ cây đến luật: Rút luật từ cây, Loại bỏ các nhánh để luật ko cần thiết, Loại bỏ luật thừa, Thứ học hỏ canh (Pruner)

- Rút luật:

+ Theo dấu mỗi đg dẫn từ gốc đến lá

+ Lấy các phép thử ở nhánh đó

+ Các nút lá làm kết luận

1. Nếu ~~nó vàng~~ và ko dừng kem thì cháy nắng

2. Nếu ~~ko vàng~~ và dừng kem thì ko cháy nắng

3. Nếu nó vàng đi thì cháy nắng

4. Nếu nó nâu thì ko cháy nắng

points

Date: ..../..../....

- Thử thực hiện cảnh:

+ Tạo 1 luật cho mỗi dãy đi từ gốc đến lá 0 cây quyết định.

+ Đơn giản hoá mỗi luật bằng cách loại bỏ ~ nên để k° ảnh hg KL mà cây có đc

+ Thay thế ~ luật có chung KL bằng 1 luật mặc định mà luật này sẽ đc kích hoạt khi k° có luật nào ± đc kích hoạt.

→ Các tiêu chuẩn

1. Nếu k° dùng kem hoặc tốc độ thì chạy nặng

2. Nếu tốc rằng s dùng kem hoặc tốc nâu thì chạy nặng