CHAPTER 4.

AN INTRODUCTION TO MACHINE LEARNING

Nội dung

- Một số định nghĩa
  - Học máy
  - Cây quyết định
- Bài toán phân lớp dữ liệu và một số thuật toán
- Bài toán gom cụm dữ liệu và một số thuật toán

Một số định nghĩa

Học máy

- Học là một trong những khả năng thông minh cơ bản và khó tự động hóa nhất của con người nói riêng và động vật nói chung.
- Học máy là việc mô hình hóa môi trường xung quanh, hay khả năng một chương trình máy tính sinh ra một cấu trúc dữ liệu mới khác với cấu trúc dữ liệu hiện có, lưu trữ và khai thác tri thức phục vụ nhu cầu con người; chương trình máy tính đó gọi là chương trình học.

- Những bài toán quan trọng và có nhiều ứng dụng thường có kích thước dữ liệu lớn, nên các chương trình học chỉ có thể khảo sát một phần nhỏ trong không gian các mẫu có thể.
- Từ đó, chương trình học phải khái quát hóa được một cách đúng đắn những mẫu đã biết thông tin; để có thể đạt được hiệu quả khi giải quyết những vấn đề mới, chưa từng gặp trong lĩnh vực đó; đây chính là vấn đề trọng tâm, vấn đề khó khăn của việc học.

Cây quyết định

- Cây quyết định (tiếng Anh: decision tree) là một đồ thị của các quyết định và các hậu quả có thể của nó (bao gồm rủi ro và hao phí tài nguyên).
- Cây quyết định được sử dụng để xây dựng một kế hoạch nhằm đạt được mục tiêu mong muốn.
- Các cây quyết định được dùng để hỗ trợ quá trình ra quyết định. Cây quyết định là một dạng đặc biệt của cấu trúc cây.

Cây quyết định (…)

- Trong lĩnh vực học máy, cây quyết định là một kiểu mô hình dự báo (predictive model), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng.
- Mỗi một nút trong (internal node) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó.
- Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó.

Cây quyết định (…)

- Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định.
- Học bằng cây quyết định cũng là một phương pháp thông dụng trong khai phá dữ liệu. Khi đó, cây quyết định mô tả một cấu trúc cây, trong đó, các lá đại diện cho các phân loại còn cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân loại đó.

Cây quyết định (...)

- Một cây quyết định có thể được học bằng cách chia tập hợp nguồn thành các tập con dựa theo một kiểm tra giá trị thuộc tính. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất. Quá trình đệ quy hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất. Một bộ phân loại rừng ngẫu nhiên (random forest) sử dụng một số cây quyết định để có thể cải thiện tỉ lệ phân loại.
- Cây quyết định cũng là một phương tiện có tính mô tả dành cho việc tính toán các xác suất có điều kiện.

Bài toán phân lớp dữ liệu & một số thuật toán phân lớp dữ liệu

Bài toán phân lớp

- Bài toán phân lớp (classification) và bài toán gom cụm (cluster) là hai bài toán quan trọng trong lĩnh vực học máy.
- Bài toán phân lớp là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp đã cho trước nhờ một mô hình phân lớp.
- Mô hình này được xây dựng dựa trên một tập dữ liệu được xây dựng trước đó có gán nhãn (hay còn gọi là tập huấn luyện).
- Quá trình phân lớp là quá trình gán nhãn cho đối tượng dữ liệu.

Bài toán phân lớp…

- Như vậy, nhiệm vụ của bài toán phân lớp là cần tìm một mô hình phần lớp để khi có dữ liệu mới thì có thể xác định được dữ liệu đó thuộc vào phân lớp nào.
- Có nhiều bài toán phân lớp dữ liệu như phân lớp nhị phân (binary), phân lớp đa lớp (multiclass), phân lớp đa trị.
- Bài toán phân lớp nhị phân là bài toán gắn nhãn dữ liệu cho đối tượng vào một trong hai lớp khác nhau dựa vào việc dữ liệu đó có hay không có các đặc trưng (feature) của bộ phân lớp

Bài toán phân lớp…

- Bài toán phân lớp đa lớp là quá trình phân lớp dữ liệu với số lượng lớp lớn hơn hai. Như vậy với từng dữ liệu chúng ta phải xem xét và phân lớp chúng vào những lớp khác nhau chứ không phải là hai lớp như bài toán phân lớp nhị phân. Và thực chất bài toán phân lớp nhị phân là một bài toán đặt biệt của phân lớp đa lớp.
- Ứng dụng của bài toán này được sử dụng rất nhiều và rộng rãi trong thực tế ví dụ như bài toán nhận dạng khuôn mặt, nhận diện giọng nói, phát hiện email spam…

Mô hình phân lớp dựa trên cây quyết định

- Trong mô hình phân lớp, thuật toán phân lớp giữ vai trò trung tâm, quyết định tới sự thành công của mô hình phân lớp.
- Do vậy chìa khóa của vấn đề phân lớp dữ liệu là tìm ra được một thuật toán phân lớp nhanh, hiệu quả, có độ chính xác cao và có khả năng mở rộng được.

- Các kỹ thuật phân lớp đã được sử dụng gồm:
  - Phân lớp cây quyết định
  - Bộ phân lớp Bayesian
  - Phân tích thống kê
  - Phương pháp tập thô
  - Thuật toán di truyền
  - Mạng nơron,…

Thuật toán Quinlan

- Thuật toán Quinlan: là thuật toán học theo quy nạp dùng luật, đa mục tiêu. Nó được phát triển bởi John Ross Quinlan – nhà khoa học máy tính – đưa ra năm 1979 với ý tưởng: chọn thuộc tính quan trọng nhất để tạo cây quyết định.

Ý tưởng thuật toán Quinlan

- Cho một bảng quan sát (cơ sở dữ liệu) là một tập hợp các mẫu với các thuộc tính nhất định của các đối tượng nào đó.
- Sử dụng một độ đo để định lượng và đề ra một tiêu chuẩn nhằm chọn lựa một thuộc tính mang tính chất phân loại để phân bảng này thành các bảng con nhỏ hơn. Từ các bảng con này dễ dàng phân tích tìm ra qui luật chung.
- Từ đó thiết lập cây quyết định cho thấy thứ tự của thuộc tính đang xét.
- Tìm cây quyết định, xây dựng bộ luật, và đưa ra quyết định một số mẫu cụ thể.

VD1

Cho cơ sở dữ liệu gồm các mẫu sau:

Mẫu	Thời tiết	Lá cây	Nhiệt độ	Quyết định (mùa)
1	Mưa	Rụng	Thấp	Đông
2	Nắng	Xanh	Trung bình	Xuân
3	Nắng	Vàng	Trung bình	Thu
4	Nắng	Xanh	Cao	Hè
5	Nắng	Rụng	Thấp	Đông
6	Tuyết	Vàng	Thấp	Đông
7	Mưa	Rụng	Trung bình	Thu
8	Mưa	Xanh	Cao	Hè
9	Tuyết	Xanh	Thấp	Đông
10	Tuyết	Rụng	Thấp	Đông
11	Mưa	Vàng	Trung bình	Thu
12	Mưa	Xanh	Trung bình	Xuân
x	Mưa	Vàng	Cao	?
y	Tuyết	Rụng	Trung bình	?
z	Tuyết	Vàng	Trung bình	?

a. Từ mẫu 1 đến mẫu 12 hãy rút ra bộ luật cho sự quyết định Mùa theo thuật toán Quinlan.

b. Áp dụng cho biết kết quả các mẫu x,y,z.

Hướng dẫn giải

Bước 0: Gọi vect ơ độ đo v=(Xuân, Hè, Thu, Đông)

Bước 1:Tính vectơ độ đo của các thuộc tính ở CSDL ban đầu

+Thuộc tính **thời tiết**

$V_{(\text{thời tiết=nắng})} = (1/4,\ 1/4,\ 1/4,\ 1/4);$

$V_{(\text{thời tiết=tuyết})} = (0,\ 0,\ 0,\ 1);$

$V_{(\text{thời tiết=mưa})} = (1/5,\ 1/5,\ 2/5,\ 1/5);$

+Thuộc tính lá cây

$V_{(\text{lá cây=Vàng})} = (0,\ 0,\ 2/3,\ 1/3);$

$V_{(\text{lá cây=Xanh})} = (2/5,\ 2/5,\ 0,\ 1/5);$

$V_{(\text{lá cây=Rụng})} = (0,\ 0,\ 1/4,\ 3/4);$

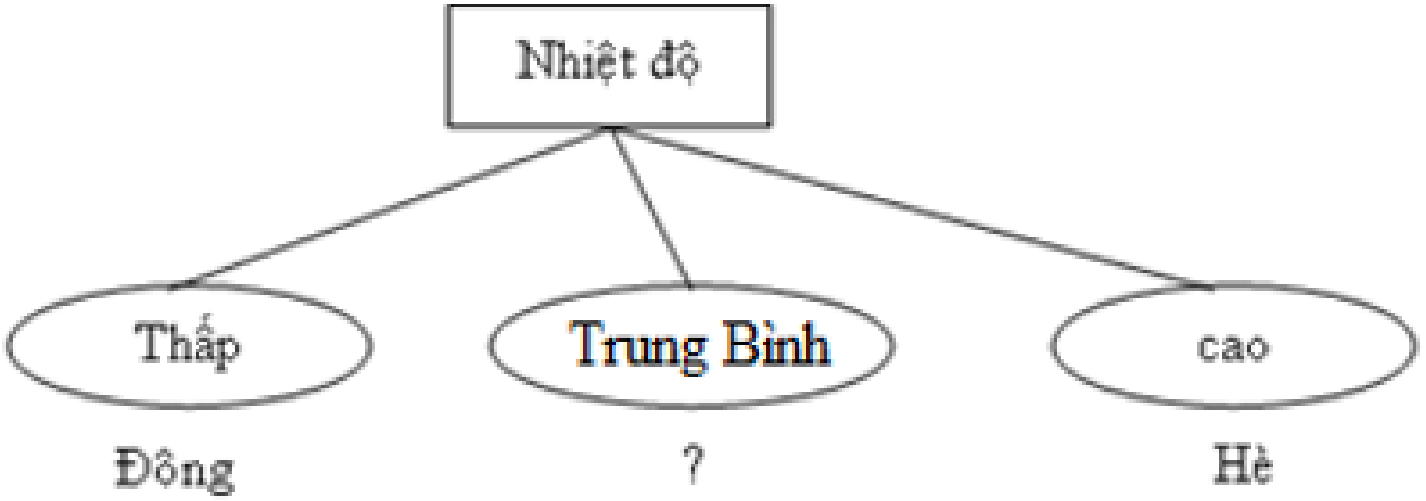
+Thuộc tính Nhiệt độ

$V_{(\text{nhiệt độ=Trung bình})} = (2/5, 0, 3/5, 0);$

$V_{(\text{nhiệt độ=thấp})} = (0, 0, 0, 1);$

$V_{(\text{nhiệt độ=cao})} = (0, 1, 0, 0);$

Chọn thuộc tính Nhiệt độ làm thuộc tính phân loại:



Bước 2: CSDL ứng với Nhiệt độ bằng Trung bình

#	Thời tiết	Lá cây	Quyết định (Mùa)
2	Nắng	Xanh	Xuân
3	Nắng	Vàng	Thu
7	Mưa	Rụng	Thu
11	Mưa	Vàng	Thu
12	Mưa	Xanh	Xuân

Tính vectơ độ đo của các thuộc tính

+Thuộc tính **thời tiết**

$V_{(\text{thời tiết=nắng})} = (1/2,\ 0,\ 1/2,\ 0);$

$V_{(\text{thời tiết=mưa})} = (1/3,\ 0,\ 2/3,\ 0);$

+Thuộc tính lá cây

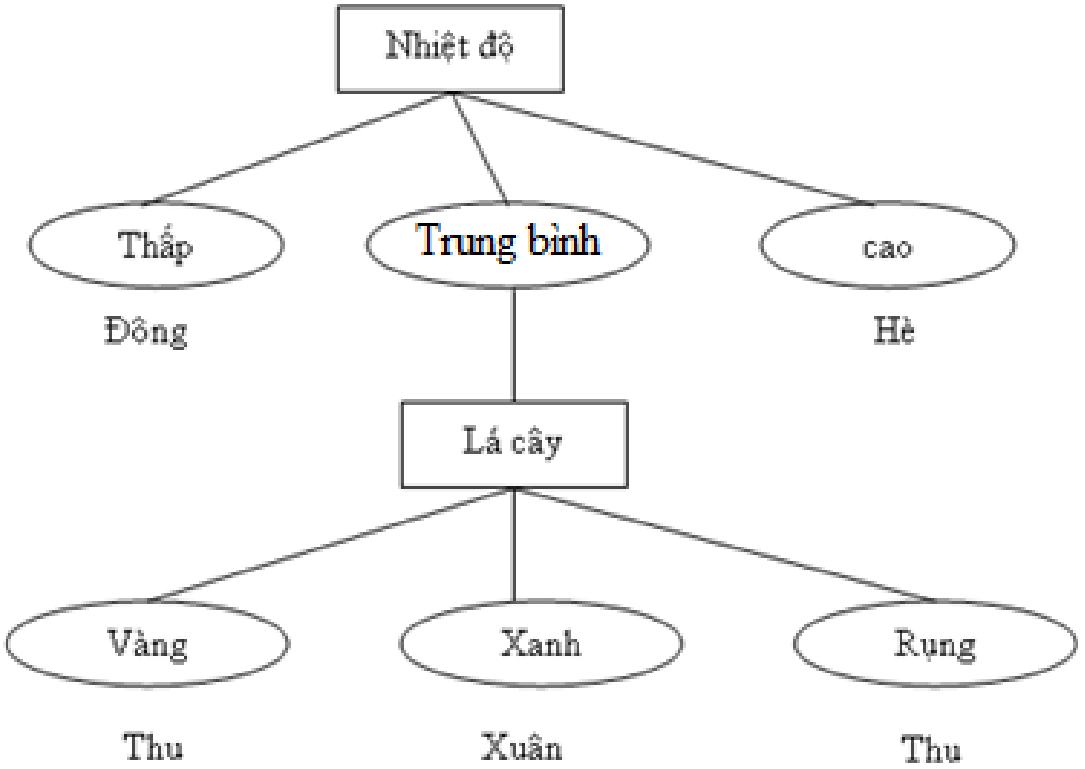
$V_{(\text{lá cây=Vàng})} = (0, 0, 1, 0);$

$V_{(\text{lá cây=Xanh})} = (1, 0, 0, 0);$

$V_{(\text{lá cây=Rụng})} = (0, 0, 1, 0);$

Chọn thuộc tính Lá cây làm thuộc tính phân loại

Từ các bước trên, ta có cây quyết định sau:



Bước 3: Tập luật **Từ cây quyết định trên, ta có tập luật sau:**

Luật 1: Nếu nhiệt độ thấp thì mùa Đông

Luật 2: Nếu nhiệt độ cao thì mùa Hè

Luật 3: Nếu nhiệt độ trung bình và lá cây Xanh thì mùa Xuân

Luật 4: Nếu nhiệt độ trung bình và lá cây màu Vàng hoặc lá rụng thì mùa Thu

#	Thời tiết	Lá cây	Nhiệt độ	Quyết định (Mùa)	Luật
x	Mưa	Vàng	Cao	Hè	Luật 2
y	Tuyết	Rụng	Trung bình	Thu	Luật 4
z	Tuyết	Vàng	Trung bình	Thu	Luật 4

Thuật toán ID3 (Iterative Dichotomiser 3)

- ID3 là một thuật toán được phát minh bởi John Ross Quinlan được sử dụng để tạo cây quyết định từ bộ dữ liệu.
- ID3 là tiền thân của thuật toán C4.5 và thường được sử dụng trong các lĩnh vực xử lý ngôn ngữ tự nhiên và học máy.

Sơ đồ thuật toán ID3

- Lặp:

1. Chọn A <= thuộc tính 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 toàn thì NGƯNG.

Ngược lại, lặp với các nút lá mới.

Trong đó thuộc tính tốt nhất ở đây là thuộc tính có entropy trung bình thấp nhất

Độ hỗn loạn trung bình: công thức xác định độ hỗn loạn trung bình (Entropy Average)

$$\text{Độ hỗn loạn TB} = \sum_b [\frac{n_b}{n_t} * \sum_c (\frac{-n_{bc}}{n_b} * \log_2 \frac{n_{bc}}{n_b})]$$

Trong đó:

b là số giá trị có trong mỗi thuộc tính đang xét.

c là số giá trị có trong thuộc tính quyết định ứng với mỗi giá trị của thuộc tinh đang xét.

$\mathbf{n_t}$: tong so mau tren cay

$\mathbf{n_b}$: tổng số mẫu thuộc về nhánh b

$\mathbf{n_{bc}}$:Tổng số mẫu trên nhánh b thuộc về lớp c

VD2

Ví dụ

cho bảng quan sát sau:

TT	Tên	Màu tóc	Chiều cao	Cân nặng	Dùng kem	Kết quả
1	Sarah	Vàng	Trung bình	Nhe	Không	Cháy nắng
2	Dana	Vàng	Cao	Trung bình	Có	Không cháy nắng
3	Alex	Nâu	Thấp	Trung bình	Có	Không cháy nắng
4	Annie	Vàng	Thấp	Trung bình	Không	Cháy nắng
5	Emily	Đỏ	Trung bình	Năng	Không	Cháy nắng
6	Peter	Nâu	Cao	Năng	Không	Không cháy nắng
7	John	Nâu	Trung bình	Năng	Không	Không cháy nắng
8	Katie	Vàng	Thấp	Nhe	Có	Không cháy nắng

Tính lần 1, xét lần lượt các thuộc tính:

Thuộc tính màu tóc: vang/nau/do

$$\textbf{ĐHL}_{\textbf{TB}} = 4/8[-2/4*\log_2(2/4)–2/4*\log_2(2/4)]+3/8[–3/3*\log_2(3/3)] +1/8\ [-1/1*\log_2(1/1)]$$

$$= ½ + 0 + 0 = \mathbf{0.5}$$

Thuộc tính chiều cao: THAP/TB/CAO

$$\textbf{ĐHL}_{\textbf{TB}} = 3/8\ [-2/3 * \log_2(2/3) – 1/3 * \log_2(1/3)]+3/8\ [–2/3 * \log_2(2/3) –1/3*\log_2(1/3)]$$

$$+ 2/8\ [- 2/2* \log_2 (2/2)]$$

$$= 1/4 * \log2\ (27/4) + 0 = \mathbf{0.69}$$

Thuộc tính cân nặng:

$$\textbf{ĐHL}_{\textbf{TB}} = 3/8[-1/3 * \log_2(1/3) - 2/3 * \log_2(2/3)] + 3/8[-1/3 * \log_2(1/3) - 2/3 * \log_2(2/3)]$$
$$+ 2/8\,[\,-1/2 * \log2\,(1/2) - 1/2 * \log2\,(1/2)]$$
$$= \mathbf{0.94}$$

Thuộc tính dùng kem:

$$\text{ĐHL}_{\text{TB}} = 5/8\ [\ -3/5 * \log2\ (3/5) - 2/5 * \log2\ (2/5)] + 3/8\ [\ 3/3 * \log2\ (3/3)]$$

$$= \mathbf{0.61}$$

Test	Độ hỗn loạn
Màu tóc	**0.5 *****
Chiều cao	**0.69**
Cân nặng	**0.94**
Dùng kem	**0.61**

➔ Chọn thuộc tính **màu tóc** là thuộc tính có độ hỗn loạn TB thấp nhất là thuộc tính quyết định.

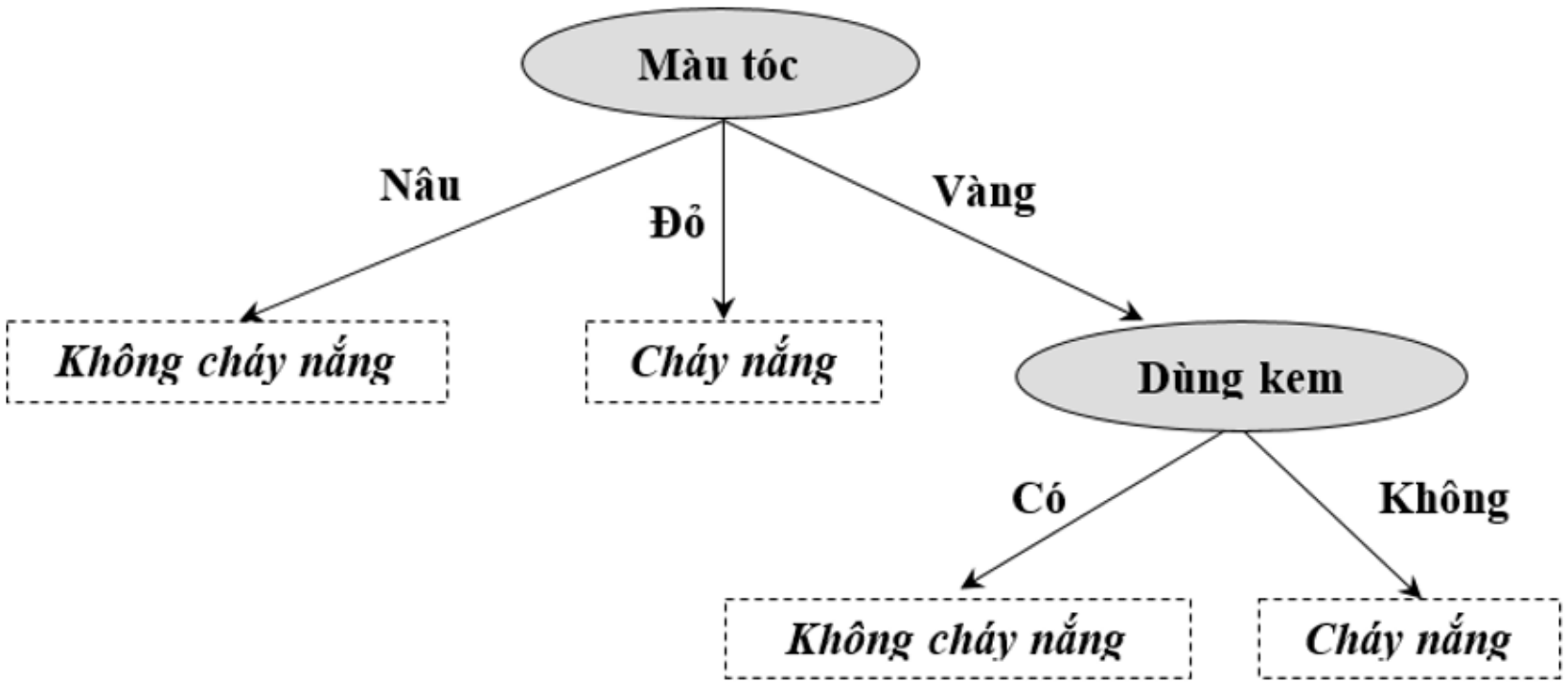
➔ Ta loại ra 3 mẫu ổn định và có bảng CSDL nhỏ hơn:

Tên	Chiều cao	Cân nặng	Dùng kem	Kết quả
Sarah	Trung bình	Nhẹ	Không	Cháy nắng
Dana	Cao	Trung bình	Có	Không cháy nắng
Annie	Thấp	Trung bình	Không	Cháy nắng
Katie	Thấp	Nhẹ	Có	Không cháy nắng

➔ Từ bảng CSDL nhỏ này, ta tính lần 2

Test	Độ hỗn loạn
Chiều cao	**0.5**
Cân nặng	**1**
Dùng kem	**0 *****

➔ Chọn thuộc tính **dùng kem** cho lần quyết định thứ 2, ta có cây định danh:



Bộ luật

Nếu màu tóc là nâu thì không cháy nắng.

Nếu màu tóc là đỏ thì cháy nắng.

Nếu màu tóc là vàng và có dùng kem thì không cháy nắng.

Nếu màu tóc là vàng và có không dùng kem thì cháy nắng.

➔ Được rút gọn thành bộ luật:

L1.Nếu tóc nâu thì không cháy nắng.

L2.Nếu tóc vàng và có dùng kem thì không cháy nắng

L3.Không luật nào thỏa thì cháy nắng.

Thuật toán ILA
(ILA INDUCTIVE learning algorithm)

Dùng để rút các luật phân lớp từ tập mẫu dữ liệu:

- 1.Chia bảng con có chứa m mẫu thành n bảng con. Một bảng con ứng với một giá trị của thuộc tính phân lớp (lặp lại từ bước 2 đến bước 8 cho mỗi bảng con).
- 2.Khởi tạo số lượng thuộc tính kết hợp với j=1;
- 3.Với mỗi bảng con đang xét, phân chia các thuộc tính của nó thành một danh sách các thuộc tính kết hợp, mỗi thành phần của danh sách có j thuộc tính phân biệt.

- 4.Với mỗi kết hợp các thuộc tính trong danh sách trên, đếm số lần xuất hiện các giá trị cho các thuộc tính trong kết hợp đó ở các dòng chưa bị khóa của bảng đang xét nhưng nó không được xuất hiện cùng giá trị ở những bảng con khác. Chọn ra một kết hợp trong danh sách sao cho nó có giá trị tương ứng xuất hiện nhiều nhất và được gọi là max_combination
- 5.Nếu max_combination=0 thì j=j+1 quay lại bước 3.
- 6.khóa các dòng ở bảng con đang xét mà tại đó giá trị bằng với giá trị tạo ra max_combination.

- 7.Thêm vào R luật mới với giả thiết là các giá trị tạo ra max_combination kết nối các bộ này bằng phép AND, kết luận là giá trị của thuộc tính quyết định trong bảng con đang xét.
- 8.Nếu tất cả các dòng đều khóa
    - Nếu còn bảng con thì qua bảng con tiếp theo và quay lại bước 2.
    - Ngược lại chấm dứt thuật toán
  - Ngược lại thì quay lại bước 4.

VD3

Thuật toán ILA (Inductive Learning Algorithm)

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi
Mưa	Ấm áp	Cao	Nhẹ	Có
Mưa	Mát	Trung bình	Nhẹ	Có
Nắng	Mát	Trung bình	Nhẹ	Có
Mưa	Ấm áp	Trung bình	Nhẹ	Có
Nắng	Ấm áp	Trung bình	Mạnh	Có
Âm u	Ấm áp	Cao	Mạnh	Có
Âm u	Nóng	Trung bình	Nhẹ	Có
Âm u	Nóng	Cao	Nhẹ	Có
Âm u	Mát	Trung bình	Mạnh	Có

Số lượng thuộc tính kết hợp j = 1

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi
Nắng	Nóng	Cao	Nhẹ	Không
Nắng	Nóng	Cao	Mạnh	Không
Mưa	Mát	Trung bình	Mạnh	Không
Nắng	Ấm áp	Cao	Nhẹ	Không
Mưa	Ấm áp	Cao	Mạnh	Không

Thuật toán ILA (Inductive Learning Algorithm)

IF Quang cảnh="Âm u" then Chơi Tennis="Có"

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi
Mưa	Ấm áp	Cao	Nhẹ	Có
Mưa	Mát	Trung bình	Nhẹ	Có
Mưa	Ấm áp	Trung bình	Nhẹ	Có
Nắng	Ấm áp	Trung bình	Mạnh	Có
Nắng	Mát	Trung bình	Nhẹ	Có

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi
Nắng	Nóng	Cao	Nhẹ	Không
Nắng	Nóng	Cao	Mạnh	Không
Mưa	Mát	Trung bình	Mạnh	Không
Nắng	Ấm áp	Cao	Nhẹ	Không
Mưa	Ấm áp	Cao	Mạnh	Không

Số lượng thuộc tính kết hợp j = 2

Thuật toán ILA (Inductive Learning Algorithm)

IF Quang cảnh="Âm u" then Chơi Tennis="Có"

IF Quang cảnh="Mưa" and Gió="Nhẹ" then Chơi Tennis="Có"

IF Quang cảnh="Nắng" and Độ ẩm="Trung bình" then Chơi Tennis="Có"

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi
Nắng	Nóng	Cao	Nhẹ	Không
Nắng	Nóng	Cao	Mạnh	Không
Mưa	Mát	Trung bình	Mạnh	Không
Nắng	Ấm áp	Cao	Nhẹ	Không
Mưa	Ấm áp	Cao	Mạnh	Không

Số lượng thuộc tính kết hợp j = 3 ⇔ Không còn bảng con

Thuật toán ILA (Inductive Learning Algorithm)

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi
Nắng	Nóng	Cao	Nhẹ	Không
Nắng	Nóng	Cao	Mạnh	Không
Mưa	Mát	Trung bình	Mạnh	Không
Nắng	Ấm áp	Cao	Nhẹ	Không
Mưa	Ấm áp	Cao	Mạnh	Không

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi
Mưa	Ấm áp	Cao	Nhẹ	Có
Mưa	Mát	Trung bình	Nhẹ	Có
Nắng	Mát	Trung bình	Nhẹ	Có
Mưa	Ấm áp	Trung bình	Nhẹ	Có
Nắng	Ấm áp	Trung bình	Mạnh	Có
Âm u	Ấm áp	Cao	Mạnh	Có
Âm u	Nóng	Trung bình	Nhẹ	Có
Âm u	Nóng	Cao	Nhẹ	Có
Âm u	Mát	Trung bình	Mạnh	Có

Số lượng thuộc tính kết hợp j = 1

⇔

max_combination =0

Thuật toán ILA (Inductive Learning Algorithm)

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi
Nắng	Nóng	Cao	Nhẹ	Không
Nắng	Nóng	Cao	Mạnh	Không
Mưa	Mát	Trung bình	Mạnh	Không
Nắng	Ấm áp	Cao	Nhẹ	Không
Mưa	Ấm áp	Cao	Mạnh	Không

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi
Mưa	Ấm áp	Cao	Nhẹ	Có
Mưa	Mát	Trung bình	Nhẹ	Có
Nắng	Mát	Trung bình	Nhẹ	Có
Mưa	Ấm áp	Trung bình	Nhẹ	Có
Nắng	Ấm áp	Trung bình	Mạnh	Có
Âm u	Ấm áp	Cao	Mạnh	Có
Âm u	Nóng	Trung bình	Nhẹ	Có
Âm u	Nóng	Cao	Nhẹ	Có
Âm u	Mát	Trung bình	Mạnh	Có

Số lượng thuộc tính kết hợp j = 2

Thuật toán ILA (Inductive Learning Algorithm)

IF Quang cảnh="Âm u" then Chơi Tennis="Có"

IF Quang cảnh="Mưa" and Gió="Nhẹ" then Chơi Tennis="Có"

IF Quang cảnh="Nắng" and Độ ẩm="Trung bình" then Chơi Tennis="Có"

IF Quang cảnh="Nắng" and Độ ẩm="Cao" then Chơi Tennis="Không"

IF Quang cảnh="Mưa" and Gió="Mạnh" then Chơi Tennis="Không"

Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi
Mưa	Ấm áp	Cao	Nhẹ	Có
Mưa	Mát	Trung bình	Nhẹ	Có
Nắng	Mát	Trung bình	Nhẹ	Có
Mưa	Ấm áp	Trung bình	Nhẹ	Có
Nắng	Ấm áp	Trung bình	Mạnh	Có
Âm u	Ấm áp	Cao	Mạnh	Có
Âm u	Nóng	Trung bình	Nhẹ	Có
Âm u	Nóng	Cao	Nhẹ	Có
Âm u	Mát	Trung bình	Mạnh	Có

Số lượng thuộc tính kết hợp j = 3

⇔ Không còn bảng

Bài toán gom cụm dữ liệu & thuật toán gom cụm dữ liệu

Nội dung

- Bài toán phân cụm
- Giới thiệu các độ đo
- Giới thiệu các thuật toán phân cụm - thuật toán k-means

Bài toán phân cụm

- Quá trình phân nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm
- Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở các cụm khác.

Các yêu cầu tiêu biểu về việc phân cụm dữ liệu

- Khả năng co giãn về tập dữ liệu (scalability)
- Khả năng xử lý nhiều kiểu thuộc tính khác nhau (different types of attributes)
- Khả năng khám phá các cụm với hình dạng tùy ý (clusters with arbitrary shape)
- Tối thiểu hóa yêu cầu về tri thức miền trong việc xác định các thông số nhập (domain knowledge for input parameters)
- Khả năng xử lý dữ liệu có nhiễu (noisy data)

- Khả năng phân cụm tăng dần và độc lập với thứ tự của dữ liệu nhập (incremental clustering and insensitivity to the order of input records)
- Khả năng xử lý dữ liệu đa chiều (high dimensionality)
- Khả năng gom cụm dựa trên ràng buộc (constraint-based clustering)
- Khả diễn và khả dụng (interpretability and usability)

Giới thiệu các độ đo

Two dimensions

In the Euclidean plane, let point p have Cartesian coordinates (p_1, p_2) and let point q have coordinates (q_1, q_2). Then the distance between p and q is given by:[2]

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

Manhattan distance in 2D space

In a 2 dimensional space, a point is represented as (x, y).

Consider two points P1 and P2:

P1: (x1, y1)

P2: (x2, y2)

Then, the manhattan distance between P1 and P2 is given as:
|x1-x2| + |y1-y2|

Giới thiệu các thuật toán phân cụm - thuật toán k-means

- Phân hoạch (partitioning): các phân hoạch được tạo ra và đánh giá theo một tiêu chí nào đó.
- Phân cấp (hierarchical): phân rã tập dữ liệu/đối tượng có thứ tự phân cấp theo một tiêu chí nào đó.
- Dựa trên mật độ (density-based): dựa trên connectivity and density functions.
- Dựa trên lưới (grid-based): dựa trên a multiple-level granularity structure.
- Dựa trên mô hình (model-based): một mô hình giả thuyết được đưa ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng nhất.

Các phương pháp đánh giá việc phân cụm dữ liệu

Đánh giá ngoại (external validation)
Đánh giá nội (internal validation)
Đánh giá tương đối (relative validation)
*Tiêu chí cho việc đánh giá và chọn kết quả phân cụm tối ưu:
- Độ nén (compactness): các đối tượng trong cụm nên gần nhau.
- Độ phân tách (separation): các cụm nên xa nhau.

Thuật toán K-means

- Chọn ngẫu nhiên k tâm cho k cụm (hoặc cho trước tâm của mỗi cụm; tâm mỗi cụm không nhất thiết trùng với một trong các đối tượng trong cụm đó)
- Lặp lại bước sau cho đến khi không có sự thay đổi nhóm của các đối tượng:
  - ✓ Tính khoảng cách giữa các đối tượng đến tâm của mỗi cụm
  - ✓ Cập nhật các đối tượng ở các nhóm bằng cách: Gom các đối tượng vào nhóm gần nhất.
  - ✓ Xác định lại tâm mới cho mỗi nhóm.

Ví dụ 4:

Trong mặt phẳng tọa độ *OXY*, cho 5 điểm *A*(2;3); *B*(7;5); *C*(4;2); *D*(8;4); *E*(2;0). Số cụm *k*=2; tâm cụm 1 là C1 (2;3), C2(4;2).

Sử dụng thuật toán *k*-means giải bài toán phân cụm dữ liệu trên theo độ đo khoảng cách euclid.

Hỏi sau khi kết thúc thuật toán thì cụm 1 và cụm 2 gồm các đối tượng nào ?

Giải:

A(2;3) B(7;5) C(4;2) D(8;4) E(2;0)

c1=(2;3) c2=(4;2)

Bước lặp 1: Tìm khoảng cách từ các điểm đến tâm các cụm

	A	B	C	D	E
c1(2;3)	0	5.39	2.24	6.08	3
c2(4;2)	2.24	4.24	0	4.47	2.83

Cụm 1 gồm các điểm: A

Cụm 2 gồm các điểm: B, C, D, E

Tâm cụm 1 là c1(2;3)

Tâm cụm 2 là c2(5.25;2.75)

Bước lặp 2: Tìm khoảng cách từ các điểm đến tâm các cụm

	A	B	C	D	E
c1(2;3)	0	5.39	2.24	6.08	3
c2(5.25;2.75)	3.26	2.85	1.46	3.02	4.26

Cụm 1 gồm: A,E
Cụm 2 gồm: B,C,D
Tâm cụm 1 là c1(2;1.5)
Tâm cụm 2 là c2(6.33;3.67)

Bước lặp 3: Tìm khoảng các từ các điểm đến tâm các cụm

	A	B	C	D	E
c1(2;1,5)	1.5	6.10	2.06	6.5	1.5
c2(6.33;3.67)	4.38	1.49	2.87	1.7	5.68

Cụm 1 gồm: A, C, E
Cụm 2 gồm: B, D
Tâm cụm 1 là c1(2.67;1.67)
Tâm cụm 2 là c2 (7.5;4.5)

Bước lặp 4: Tìm khoảng các từ các điểm đến tâm các cụm

	A	B	C	D	E
c1(2.67;1.67)	1.49	5.46	1.37	5.82	1.8
c2 (7.5;4.5)	5.7	0.71	4.3	0.71	7.11

Cụm 1 gồm: A, C, E

Cụm 2 gồm: B, D

Không có sự thay đổi các đối tượng trong các cụm → Dừng thuật toán

Kết quả gom cụm các đối tượng là

Cụm 1: A, C, E

Cụm 2: B, D