

Chương 7. MÁY HỌC



TRÍ TUỆ NHÂN TẠO *Artificial Intelligence*

Đoàn Vũ Thịnh
Khoa Công nghệ Thông tin
Đại học Nha Trang
Email: thinhdv@ntu.edu.vn

Nha Trang, 06-2023

Chương 7. MÁY HỌC

Trong chương 6 đã thảo luận về biểu diễn và suy luận tri thức. Trong trường hợp này giả định đã có sẵn tri thức và có thể biểu diễn tường minh tri thức.

Tuy vậy, trong nhiều tình huống sẽ không có sẵn tri thức như:

- Kỹ sư tri thức cần thu nhận tri thức từ chuyên gia cùng lĩnh vực
- Cần biết các luật mô tả lĩnh vực cụ thể
- Bài toán không được biểu diễn tường minh theo luật, sự kiện hay các quan hệ.

Chương 7. MÁY HỌC

- Hệ thống được gọi là có khả năng học (có dáng vẻ học như con người) là hệ thống có khả năng tìm ra một sự khái quát hoặc mô hình cho các dữ liệu huấn luyện (dữ liệu có gán nhãn nhận diện hoặc phân loại).
- Đặc trưng khái quát hoặc mô hình đó có thể được sử dụng để nhận diện hoặc phân loại dữ liệu mới.
- Hai hướng tiếp cận cho hệ thống học:
 - Học từ ký hiệu
 - Học từ dữ liệu số

Chương 7. MÁY HỌC

Các hình thức học

- **Học vẹt:** Hệ tiếp nhận các khẳng định của các quyết định đúng. Khi hệ tạo ra một quyết định không đúng, hệ sẽ đưa ra các luật hay quan hệ đúng mà hệ đã sử dụng. Hình thức học vẹt nhằm cho phép chuyên gia cung cấp tri thức theo kiểu tương tác.
- **Học bằng cách chỉ dẫn:** Thay vì đưa ra 1 luật cụ thể cần áp dụng vào tình huống cho trước, hệ thống sẽ được cung cấp bằng các chỉ dẫn tổng quát. Ví dụ “Sinh viên đạt học bổng khi có trung bình học kỳ từ 7.0 trở lên”. Hệ thống phải tự mình đề ra các biến đổi trừu tượng đến các luật khả dụng.

Chương 7. MÁY HỌC

Các hình thức học

- **Học bằng quy nạp:** Hệ thống được cung cấp một tập các ví dụ và kết luận được rút ra từng ví dụ. Hệ liên tục lọc ra các luật và quan hệ nhằm xử lý từng ví dụ mới.
- **Học bằng tương tự:** Hệ thống được cung cấp đáp ứng đúng cho các tác vụ tương tự nhưng không giống nhau. Hệ thống cần làm thích ứng đáp ứng trước đó nhằm tạo ra một luật mới có khả năng áp dụng cho tình huống mới.
- **Học dựa trên giải thích:** Hệ thống phân tích tập các lời giải ví dụ nhằm ấn định khả năng đúng/sai và tạo ra các giải thích dùng để hướng dẫn cách giải bài toán trong tương lai.

Chương 7. MÁY HỌC

Các hình thức học

- **Học dựa trên tình huống:** Bất kỳ tình huống nào được hệ thống lập luận đều được lưu trữ cùng với kết quả cho dù đúng hay sai. Khi gặp tình huống mới, hệ thống sẽ làm thích nghi hành vi đã lưu trữ với tình huống mới.
- **Học khám phá hay học không giám sát:** Thay vì có mục tiêu tường minh, hệ khám phá liên tục tìm kiếm các mẫu và quan hệ trong dữ liệu nhấp. Các ví dụ về học không giám sát bao gồm gom cụm dữ liệu, học để nhận dạng các đặc tính cơ bản từ các điểm ảnh.

Chương 7. MÁY HỌC

Cây định danh

- Xây dựng cây định danh dựa trên sự phân hoạch của các thuộc tính. Trong đó, phân hoạch là:
 - Nút cha: là thuộc tính được phân hoạch
 - Các nút con: Các giá trị phân biệt ứng với thuộc tính được phân hoạch.
- Là công cụ phổ biến trong một số ứng dụng

Chương 7. MÁY HỌC

Cây định danh

- Xây dựng các quy luật để có thể kết luận 1 người như thế nào thì khi tắm biển sẽ bị cháy nắng.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh

- Ta gọi tính chất cháy nắng hay không cháy nắng là thuộc tính quan tâm (thuộc tính mục tiêu)
- Như vậy, trong trường hợp này tập R chỉ bao gồm 2 phần tử {nám, không}
- Ta gọi tập P là tất cả những người tham gia khảo sát ($n=8$)
- Chúng ta quan sát các hiện tượng cháy nắng dựa trên 4 thuộc tính: chiều cao (TB, cao, thấp), cân nặng (nhẹ, TB, nặng), Màu tóc (vàng, nâu, đỏ), dùng kem chống nắng (có, không). Ta gọi các thuộc tính này là dẫn xuất.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh

- Bước 1: Phân hoạch tập P ban đầu thành các tập P_i sao cho tất cả các phần tử trong tất cả các tập P_i có cùng thuộc tính mục tiêu.

$$P = P_1 \cup P_2 \cup \dots \cup P_n \text{ và}$$

$$\forall (i, j), i \neq j \text{ thì } (P_i \cap P_j) = \emptyset \text{ và}$$

$$\forall (i, k, l): P_k \in P_i \text{ và } P_l \in P_i \text{ thì } f(P_k) = f(P_l)$$

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh

- Bước 2: Sau khi đã phân hoạch xong tập P thành các tập P_i đặc trưng bởi thuộc tính đích r_i ($r_i \in R$), bước tiếp theo là ứng với mỗi phân hoạch P_i

ta xây dựng các luật $L_i: GT_i \rightarrow r_i$

Trong đó, GT_i là mệnh đề được hình thành bằng cách kết hợp các thuộc tính dẫn xuất.

Cụ thể, có 2 cách phân hoạch dễ thấy nhất.

- Cách 1: Cho mỗi người vào một danh sách phân hoạch:

$P_1 = \{\text{Sarah}\}, P_2 = \{\text{Dana}\}, \dots, P_8 = \{\text{Kartie}\}$

- Cách 2: Tập 1 = {Nám}, Tập 2 = {Không}

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh

- Thử cách khác:

Đầu tiên quan sát màu tóc, ta có

$$P_{\text{vàng}} = \{Sarah, Dana, Amie, Kartie\}$$

$$P_{\text{đỏ}} = \{Emilie\}$$

$$P_{\text{nâu}} = \{Alex, John, Peter\}$$

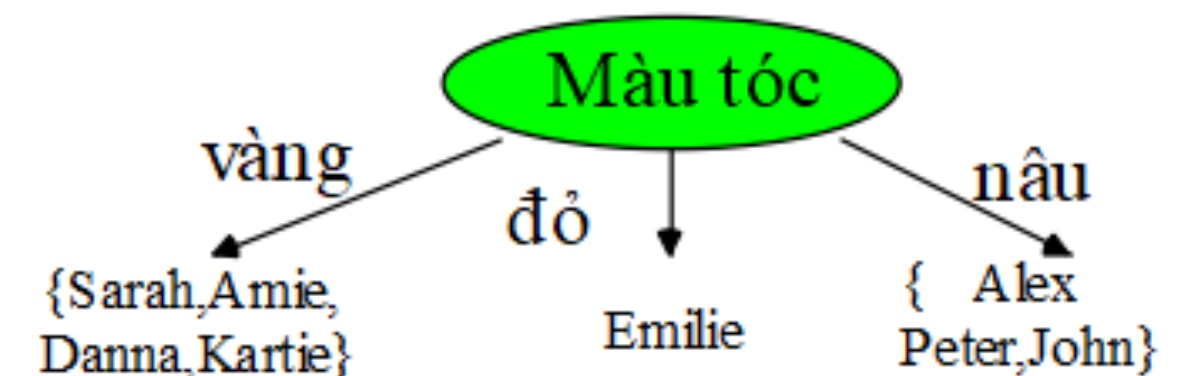
Nhìn vào bảng ta thấy:

$$P_{\text{nâu}} = \{\text{toàn người không bị cháy nắng}\}$$

$$P_{\text{đỏ}} = \{\text{toàn người bị cháy nắng}\}$$

tức là P_i có cùng chung thuộc tính mục tiêu. Còn lại $P_{\text{vàng}}$ có 2 trường hợp bị nám và không bị nám nên cần phải phân hoạch.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

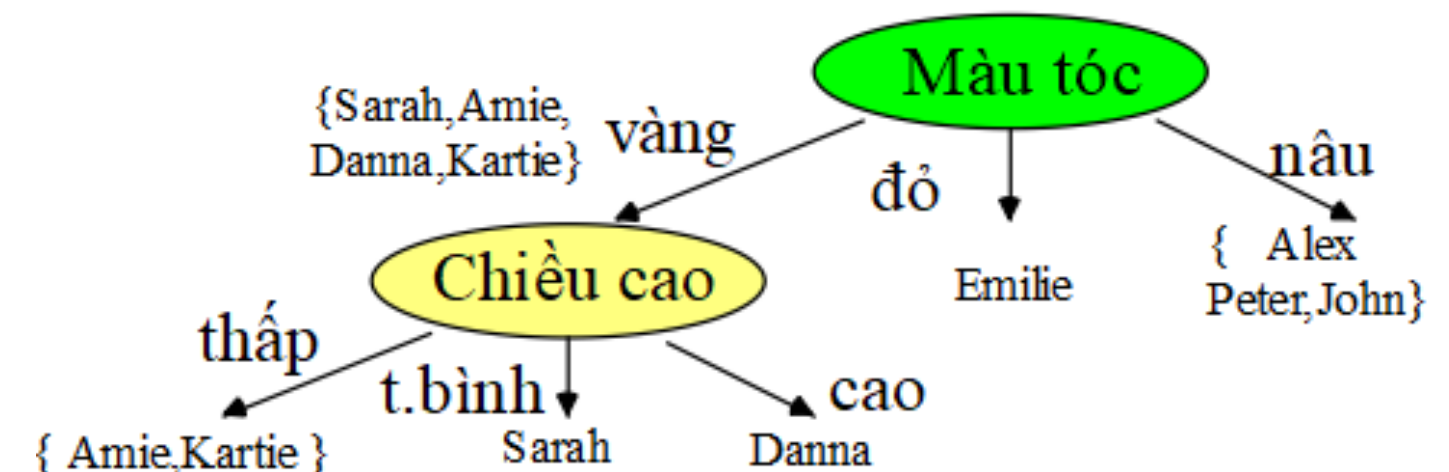


Chương 7. MÁY HỌC

Cây định danh

- Tiếp theo, từ tập $P_{\text{nâu}}$, ta quan sát thuộc tính chiều cao và phân hoạch dựa theo thuộc tính này:
- $P_{\text{vàng, thấp}} = \{Amie, Kartie\}$
- $P_{\text{vàng, cao}} = \{Dana\}$
- $P_{\text{vàng, TB}} = \{Sarah\}$
- Quá trình này cứ tiếp tục cho đến khi tất cả các nút lá của cây không còn lẫn lộn giữa nám và không bị nám.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

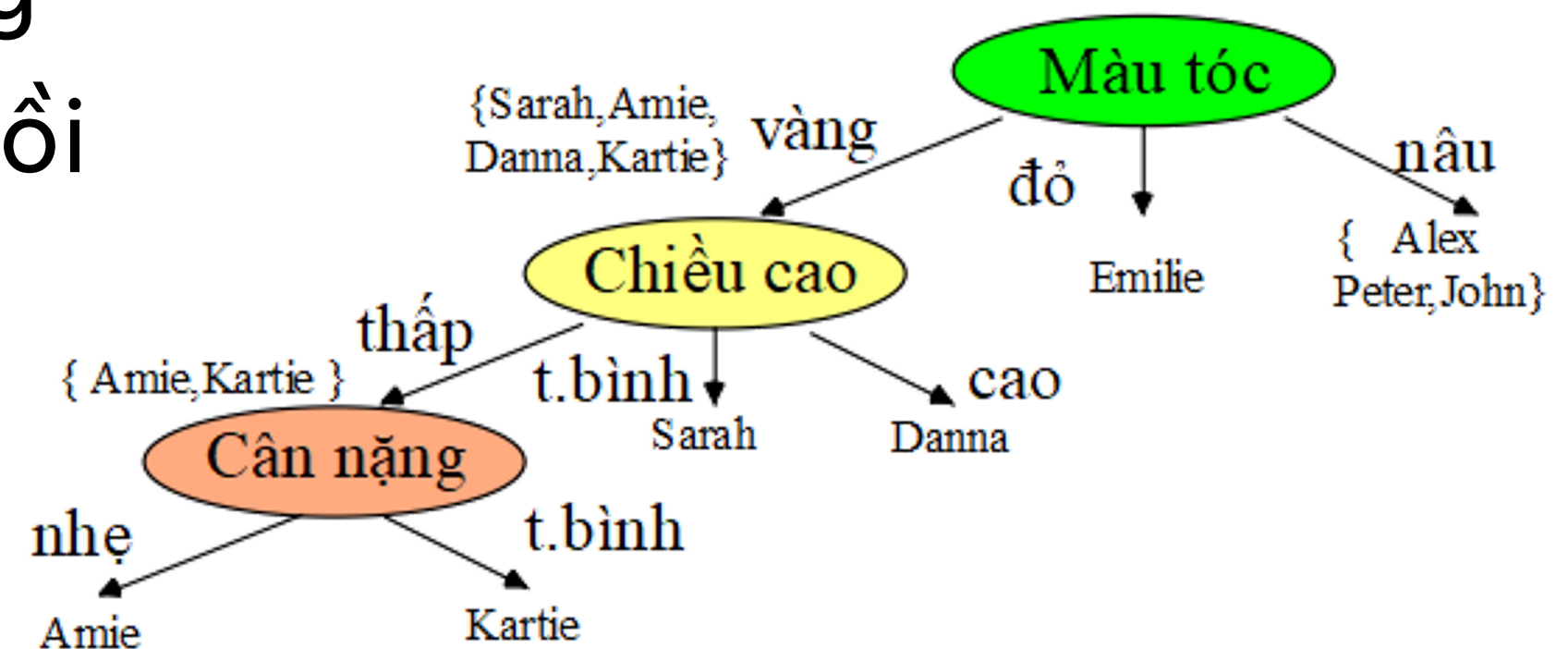


Chương 7. MÁY HỌC

Cây định danh

- Tiếp theo, từ tập $P_{\text{nâu, thấp}}$, ta quan sát thuộc tính cân nặng và phân hoạch dựa theo thuộc tính này:
- $P_{\text{vàng, thấp, nhẹ}} = \{Amie\}$
- $P_{\text{vàng, thấp, trung bình}} = \{Kartie\}$
- Có thể thấy rằng, qua mỗi bước phân hoạch thì cây phân hoạch ngày càng phình ra, quá trình này gọi là đâm chồi hay còn gọi là cây định danh.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không



Chương 7. MÁY HỌC

Cây định danh

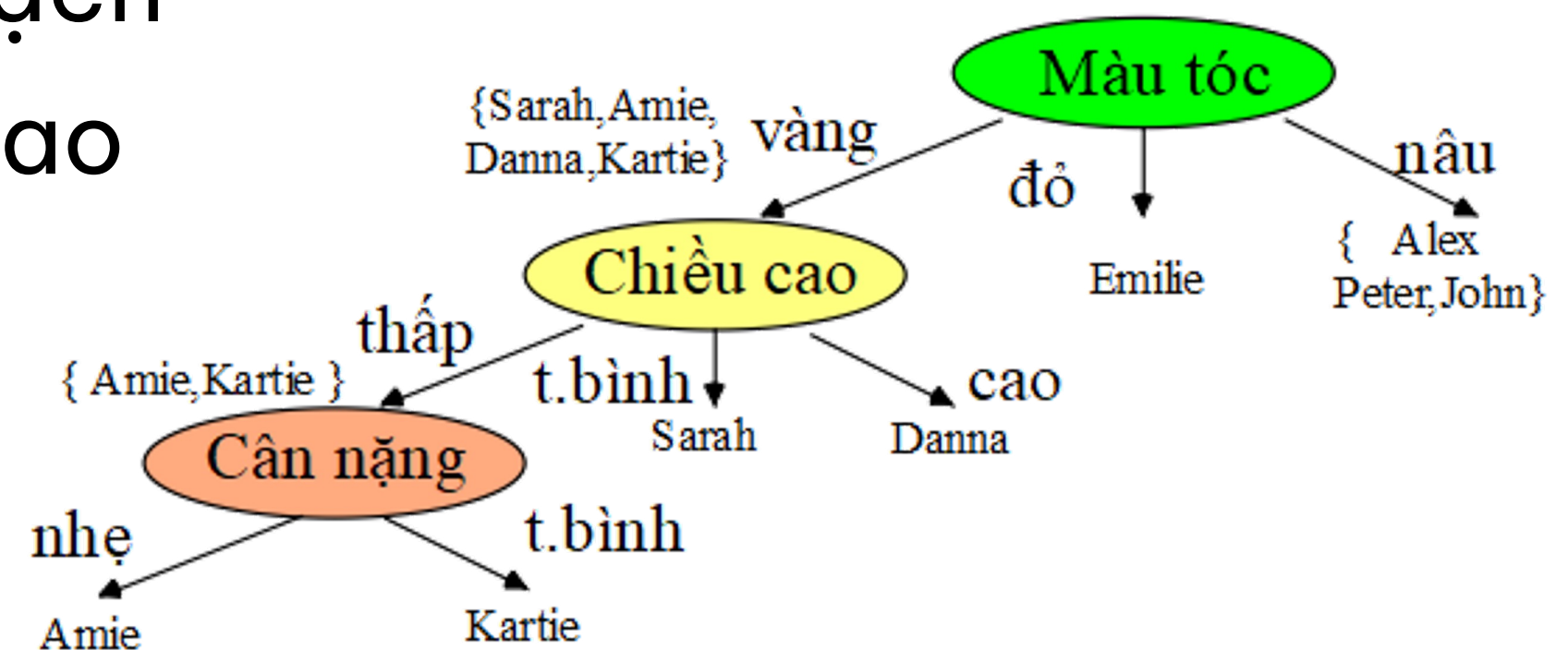
- Để chuyển từ tri thức học thành luật, đi đi từ nút gốc đến lá lấy nút gốc làm GT, nút lá làm KL

- $if(\text{Màu tóc} = \text{vàng}) \text{ and}$

$(\text{chiều cao} = \text{trung bình}) \text{ then}$

- Câu hỏi đặt ra: Nếu từ đầu ta không chọn thuộc tính màu tóc để phân hoạch mà chọn thuộc tính khác như chiều cao thì cách nào sẽ tốt hơn?

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không



Chương 7. MÁY HỌC

Cây định danh

- Ví dụ 2: Xây dựng cây định danh cho kết quả trời mưa với bảng dữ liệu sau

STT	Trời	Áp suất	Gió	Kết quả
1	Trong	Cao	Bắc	Không mưa
2	Mây	Cao	Nam	Mưa
3	Mây	Trung bình	Bắc	Mưa
4	Trong	Thấp	Bắc	Không mưa
5	Mây	Thấp	Bắc	Mưa
6	Mây	Cao	Bắc	Mưa
7	Mây	Thấp	Nam	Không mưa
8	Trong	Cao	Nam	Không mưa

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Giải thuật Quinland quyết định thuộc tính phân hoạch bằng cách xây dựng các vector đặc trưng cho mỗi giá trị của từng thuộc tính dẫn xuất và mục tiêu.
- Với mỗi thuộc tính dẫn xuất A có thể sử dụng để phân hoạch, tính
- $V_{A(j)} = (T(j, r_1), T(j, r_2), \dots, T(j, r_n))$, với
- A : thuộc tính dẫn xuất r_1, r_2, \dots, r_n : thuộc tính mục tiêu
- $T(j, r_j) = T_{Aij} / T_{Aj}$
 - T_{Aij} : tổng số phần tử trong phân hoạch có thuộc tính A là j và thuộc tính mục tiêu là r
 - T_{Aj} : Tổng số phần tử trong phân hoạch có thuộc tính A là j
- Lưu ý: $T(j, r_1) + T(j, r_2) + \dots + T(j, r_n) = 1$

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Như vậy nếu 1 thuộc tính A có thể nhận 1 trong 5 giá trị khác nhau thì nó sẽ có 5 vector đặc trưng.
- Một vector V_{Aj} được gọi là vector đơn vị nếu chỉ có duy nhất 1 thành phần có giá trị là 1, các thành phần khác có giá trị là 0.
- Thuộc tính được chọn để phân hoạch là thuộc tính có nhiều vector đơn vị nhất.

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Xây dựng các quy luật để có thể kết luận 1 người như thế nào thì khi tắm biển sẽ bị cháy nắng theo giải thuật Quinland.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Màu tóc có 3 giá trị (vàng, đỏ, nâu): có 3 vector đặc trưng

- $V_{tóc(vàng)} = T_{(vàng, nám)}, T_{(vàng, không)}$

- Số người tóc vàng: 4
- Số người tóc vàng, nám: 2
- Số người tóc vàng, không nám: 2
- Do đó, $V_{tóc(vàng)} = 0.5, 0.5$

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Màu tóc có 3 giá trị (vàng, đỏ, nâu): có 3 vector đặc trưng

- $V_{tóc(nâu)} = T_{(nâu, nám)}, T_{(nâu, không)}$

- Số người tóc nâu: 3
- Số người tóc nâu, nám: 0
- Số người tóc nâu, không nám: 3
- Do đó, $V_{tóc(nâu)} = 0, 1$ là một vector đơn vị

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Màu tóc có 3 giá trị (vàng, đỏ, nâu): có 3 vector đặc trưng

- $V_{tóc(đỏ)} = T_{(đỏ, nám)}, T_{(đỏ, không)}$

- Số người tóc đỏ: 1
 - Số người tóc đỏ, nám: 1
 - Số người tóc đỏ, không nám: 0
 - Do đó, $V_{tóc(đỏ)} = 1, 0$ là một vector đơn vị
- Tổng số vector đơn vị của màu tóc: 2

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Chiều cao có 3 giá trị (cao, t.bình, thấp): có 3 vector đặc trưng
 - $V_{chiều\ cao(cao)} = 0, 1$
 - $V_{chiều\ cao(trung\ bình)} = 0.67, 0.33$
 - $V_{chiều\ cao(thấp)} = 0.33, 0.67$
- Tổng số vector đơn vị của chiều cao: 1

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Cân nặng có 3 giá trị (nhẹ, t.bình, nặng): có 3 vector đặc trưng
 - $V_{cân\ nặng(nhẹ)} = 0.5, 0.5$
 - $V_{cân\ nặng(trung\ bình)} = 0.33, 0.67$
 - $V_{cân\ nặng(nặng)} = 0.33, 0.67$
- Tổng số vector đơn vị của cân nặng: 0

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Dùng kem có 2 giá trị (có, không): có 2 vector đặc trưng
 - $V_{dùng\ kem(có)} = 0, 1$
 - $V_{dùng\ kem(không)} = 0.6, 0.4$
- Tổng số vector đơn vị của dùng kem: 1

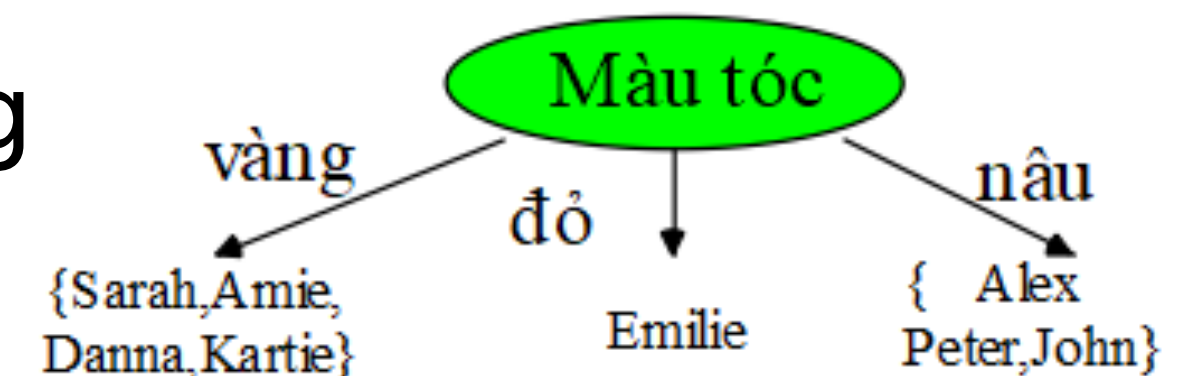
Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Ta thấy:
 - Tổng số vector đơn vị của màu tóc: 2
 - Tổng số vector đơn vị của chiều cao: 1
 - Tổng số vector đơn vị của cân nặng: 0
 - Tổng số vector đơn vị của dùng kem: 1
- Từ đó, ta chọn màu tóc để phân hoạch
- Sau khi phân hoạch ta thấy còn màu vàng chứa lẫn lộn bị nám và không bị nám.
- Tiếp tục phân hoạch cho tập này.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không



Chương 7. MÁY HỌC

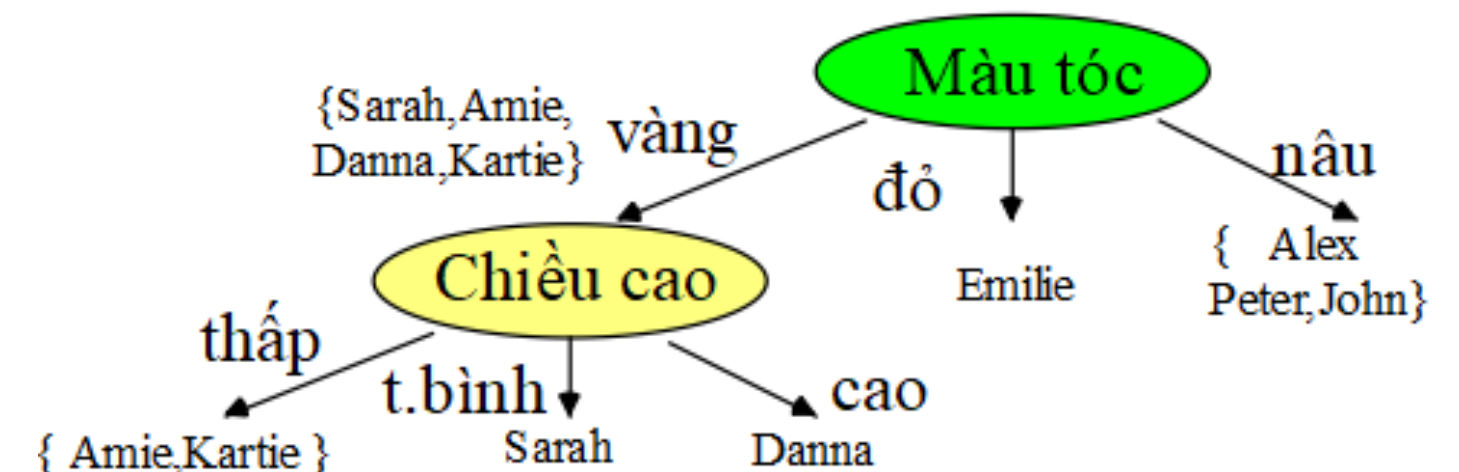
Cây định danh – Giải thuật Quinland

- Tiếp tục tính vector đặc trưng cho các thuộc tính: chiều cao, cân nặng, dùng kem cho màu tóc = vàng.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Kartie	Vàng	Thấp	Nhẹ	Có	Không

- $V_{chiều\ cao(cao)} = 0, 1$
- $V_{chiều\ cao(trung\ bình)} = 1, 0$
- $V_{chiều\ cao(thấp)} = 0.5, 0.5$
- $V_{cân\ nặng(nhẹ)} = 0.5, 0.5$
- $V_{cân\ nặng(trung\ bình)} = 0.5, 0.5$
- $V_{cân\ nặng(nặng)} = 0, 0$
- $V_{dùng\ kem(có)} = 0, 1$
- $V_{dùng\ kem(không)} = 1, 0$

- Tổng số vector đơn vị của dùng kem = chiều cao
- Chiều cao có 3 phân hoạch > dùng kem (2) nên sử dụng chiều cao cho đợt phân hoạch tiếp theo.

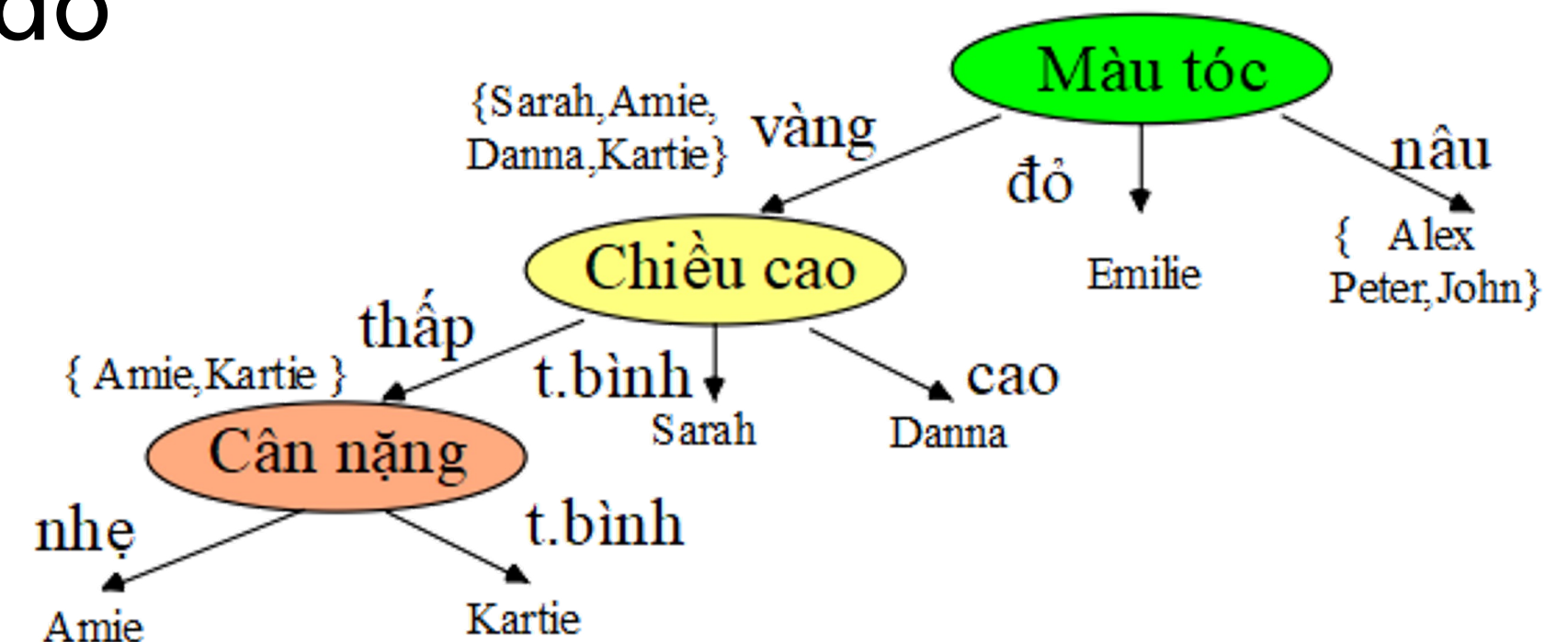


Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland

- Sau khi phân hoạch ta thấy thuộc tính chiều cao = thấp còn lẫn lộn kết quả bị nám và không bị nám.
- Tiếp tục sử dụng thuộc tính chiều cao có giá trị là thấp để phân hoạch.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Amie	Vàng	Thấp	TB	Không	Nám
Kartie	Vàng	Thấp	Nhẹ	Có	Không



- $V_{cân\ nặng(nhẹ)} = 0, 1$
- $V_{cân\ nặng(trung\ bình)} = 1, 0$
- $V_{cân\ nặng(nặng)} = 0, 0$
- $V_{dùng\ kem(có)} = 0, 1$
- $V_{dùng\ kem(không)} = 1, 0$

- Tổng số vector đơn vị của dùng kem = cân nặng
- Cân nặng có 3 phân hoạch > dùng kem (2) nên sử dụng cân nặng cho đợt phân hoạch tiếp theo.

Chương 7. MÁY HỌC

Cây định danh – Giải thuật Quinland (C4.5)

- Ví dụ 2: Xây dựng cây định danh bằng giải thuật Quinland cho kết quả trời mưa với bảng dữ liệu sau

STT	Trời	Áp suất	Gió	Kết quả
1	Trong	Cao	Bắc	Không mưa
2	Mây	Cao	Nam	Mưa
3	Mây	Trung bình	Bắc	Mưa
4	Trong	Thấp	Bắc	Không mưa
5	Mây	Thấp	Bắc	Mưa
6	Mây	Cao	Bắc	Mưa
7	Mây	Thấp	Nam	Không mưa
8	Trong	Cao	Nam	Không mưa

Chương 7. MÁY HỌC

Cây định danh – Entropy

- Thay vì phải xây dựng các vector đặc trưng như phương pháp của Quinland, ứng với mỗi thuộc tính dẫn xuất ta chỉ cần tính độ đo hỗn loạn và lựa chọn thuộc tính nào có độ đo hỗn loạn bé nhất để phân hoạch. Lặp lại cho đến khi hết các thuộc tính.
- Độ đo bất định cho thuộc tính X:

$$E(x) = \sum_b \left(\frac{n_b}{n_t} * \sum_c -\frac{n_{bc}}{n_b} * \log_a \frac{n_{bc}}{n_b} \right)$$

- Trong đó:
 - n_b : số mẫu nhánh b n_{bc} : tổng số mẫu trong nhánh b của lớp c
 - n_t : tổng số mẫu a : số lượng giá trị của thuộc tính mục tiêu

Chương 7. MÁY HỌC

Cây định danh – Entropy

- Xây dựng các quy luật để có thể kết luận 1 người như thế nào thì khi tắm biển sẽ bị cháy nắng theo độ bất định Entropy.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

- Số lượng giá trị thuộc tính mục tiêu: $a = 2$ (nám/không).

Chương 7. MÁY HỌC

Cây định danh – Entropy

- Xét thuộc tính màu tóc (vàng, nâu, đỏ)

- Vàng: 2/4 (nám) + 2/4 (không)
- Nâu: 0/3(nám) + 3/3 (không)
- Đỏ: 1/1 (nám) + 0/1 (không)

- Entropy của màu tóc:

$$\begin{aligned} E(\text{màu tóc}) &= \frac{4}{8} * \left(-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} \right) : \text{vàng} \\ &+ \frac{3}{8} * \left(-\frac{0}{3} \log_2 \frac{0}{3} - \frac{3}{3} \log_2 \frac{3}{3} \right) : \text{nâu} \\ &+ \frac{1}{8} * \left(-\frac{1}{1} \log_2 \frac{1}{1} - \frac{0}{1} \log_2 \frac{0}{1} \right) : \text{đỏ} \end{aligned}$$

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

$$E(\text{màu tóc}) = 0.5$$

Chương 7. MÁY HỌC

Cây định danh – Entropy

- Xét thuộc tính chiều cao (cao, trung bình, thấp)

- Cao: 0/2 (nám) + 2/2 (không)
- TB: 2/3(nám) + 1/3 (không)
- Thấp: 1/3 (nám) + 2/3 (không)

- Entropy của chiều cao:

- $E(\text{chiều cao}) = \frac{2}{8} * \left(-\frac{0}{2} \log_2 \frac{0}{2} - \frac{2}{2} \log_2 \frac{2}{2} \right) : \text{cao}$

$$+ \frac{3}{8} * \left(-\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \right) : \text{trung bình}$$

$$+ \frac{3}{8} * \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) : \text{thấp}$$

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

$$E(\text{chiều cao}) = 0.69$$

Chương 7. MÁY HỌC

Cây định danh – Entropy

- Xét thuộc tính cân nặng (nhẹ, trung bình, nặng)

- Nhẹ: 1/2 (nám) + 1/2 (không)
- TB: 1/3(nám) + 2/3 (không)
- Nặng: 1/3 (nám) + 2/3 (không)

- Entropy của cân nặng:

- $E(\text{cân nặng}) = \frac{2}{8} * \left(-\frac{1}{2} \log_2^{\frac{1}{2}} - \frac{1}{2} \log_2^{\frac{1}{2}} \right) : \text{nhẹ}$

$$+ \frac{3}{8} * \left(-\frac{1}{3} \log_2^{\frac{1}{3}} - \frac{2}{3} \log_2^{\frac{2}{3}} \right) : \text{trung bình}$$

$$+ \frac{3}{8} * \left(-\frac{1}{3} \log_2^{\frac{1}{3}} - \frac{2}{3} \log_2^{\frac{2}{3}} \right) : \text{nặng}$$

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

$$E(\text{cân nặng}) = 0.94$$

Chương 7. MÁY HỌC

Cây định danh – Entropy

- Xét thuộc tính dùng kem (có, không)
 - Có: 0/3 (nám) + 3/3 (không)
 - Không: 3/5 (nám) + 2/5 (không)

- Entropy của dùng kem:

$$\begin{aligned} E(\text{dùng kem}) = & \frac{3}{8} * \left(-\frac{0}{3} \log_2 \frac{0}{3} - \frac{3}{3} \log_2 \frac{3}{3} \right) : \text{có} \\ & + \frac{5}{8} * \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) : \text{không} \end{aligned}$$

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

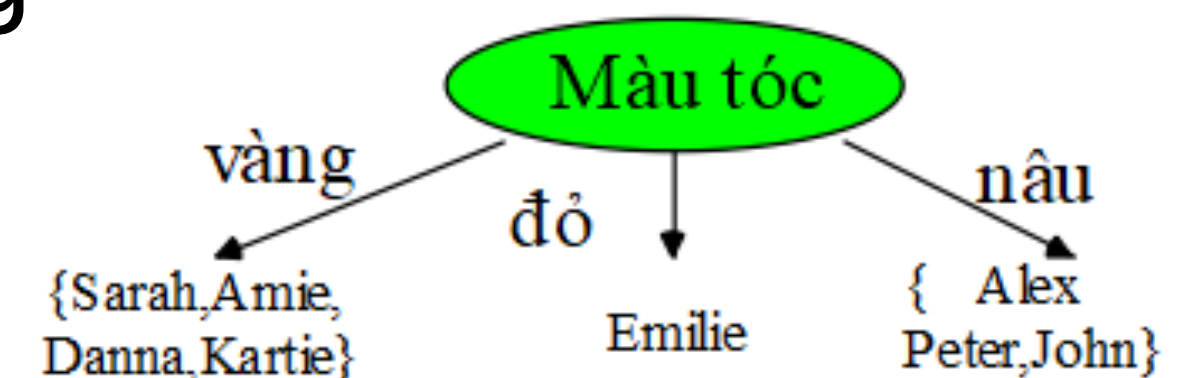
$$E(\text{dùng kem}) = 0.61$$

Chương 7. MÁY HỌC

Cây định danh – Entropy

- Ta thấy:
 - $E(\text{màu tóc}) = 0.5$
 - $E(\text{chiều cao}) = 0.69$
 - $E(\text{cân nặng}) = 0.94$
 - $E(\text{dùng kem}) = 0.61$
- Từ đó, ta chọn màu tóc để phân hoạch
- Sau khi phân hoạch ta thấy còn màu vàng chứa lẫn lộn bị nám và không bị nám.
- Tiếp tục phân hoạch cho tập này.

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không



Chương 7. MÁY HỌC

Cây định danh – Entropy

- Xét thuộc tính chiều cao (cao, trung bình, thấp)

- Cao: 0/1 (nám) + 1/1 (không)
- TB: 1/1 (nám) + 0/1 (không)
- Thấp: 1/2 (nám) + 1/2 (không)

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Kartie	Vàng	Thấp	Nhẹ	Có	Không

- Entropy của chiều cao:

- $E(\text{chiều cao}) = \frac{1}{4} * \left(-\frac{0}{1} \log_2^{\frac{0}{1}} - \frac{1}{1} \log_2^{\frac{1}{1}} \right) : \text{cao}$

$$+ \frac{1}{4} * \left(-\frac{1}{1} \log_2^{\frac{1}{1}} - \frac{0}{1} \log_2^{\frac{0}{1}} \right) : \text{trung bình}$$

$$+ \frac{2}{4} * \left(-\frac{1}{2} \log_2^{\frac{1}{2}} - \frac{1}{2} \log_2^{\frac{1}{2}} \right) : \text{thấp}$$

$$E(\text{chiều cao}) = 0.5$$

Chương 7. MÁY HỌC

Cây định danh – Entropy

- Xét thuộc tính cân nặng (nhẹ, trung bình, nặng)

- Nhẹ: 1/2 (nám) + 1/2 (không)
- TB: 1/2(nám) + 1/2 (không)

- Entropy của cân nặng:

- $$E(\text{cân nặng}) = \frac{2}{4} * \left(-\frac{1}{2} \log_2^{\frac{1}{2}} - \frac{1}{2} \log_2^{\frac{1}{2}} \right) : \text{nhẹ}$$
$$+ \frac{2}{4} * \left(-\frac{1}{2} \log_2^{\frac{1}{2}} - \frac{1}{2} \log_2^{\frac{1}{2}} \right) : \text{trung bình}$$

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Kartie	Vàng	Thấp	Nhẹ	Có	Không

$$E(\text{cân nặng}) = 1.0$$

Chương 7. MÁY HỌC

Cây định danh – Entropy

- Xét thuộc tính dùng kem (có, không)
 - Có: 0/2 (nám) + 2/2 (không)
 - Không: 2/2 (nám) + 0/2 (không)
- Entropy của dùng kem:

$$\begin{aligned} E(\text{dùng kem}) &= \frac{2}{4} * \left(-\frac{0}{2} \log_2 \frac{0}{2} - \frac{2}{2} \log_2 \frac{2}{2} \right) : \text{có} \\ &+ \frac{2}{4} * \left(-\frac{2}{2} \log_2 \frac{2}{2} - \frac{0}{2} \log_2 \frac{0}{2} \right) : \text{không} \end{aligned}$$

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Kartie	Vàng	Thấp	Nhẹ	Có	Không

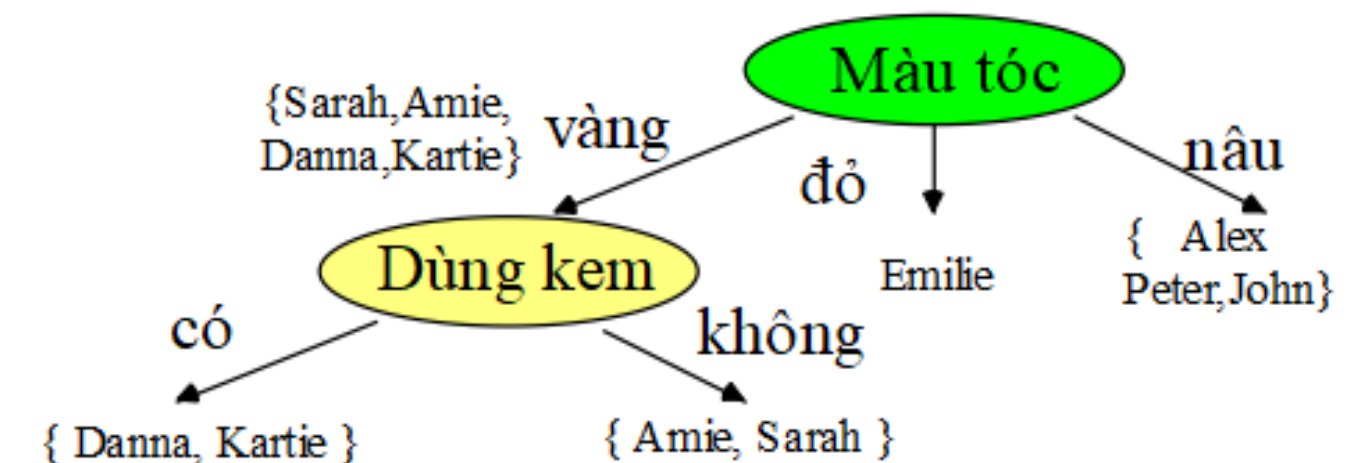
$$E(\text{dùng kem}) = 0$$

Chương 7. MÁY HỌC

Cây định danh – Entropy

- Ta thấy:
 - $E(\text{chiều cao}) = 0.5$
 - $E(\text{cân nặng}) = 1.0$
 - $E(\text{dùng kem}) = 0.0$
- Từ đó, ta chọn dùng kem để phân hoạch
- Sau khi phân hoạch ta không còn thuộc tính để phân hoạch nữa. KẾT THÚC

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Kartie	Vàng	Thấp	Nhẹ	Có	Không



Chương 7. MÁY HỌC

Cây định danh – Entropy (ID3)

- Ví dụ 2: Xây dựng cây định danh bằng giải thuật độ đo hỗn loạn cho kết quả trời mưa với bảng dữ liệu sau

STT	Trời	Áp suất	Gió	Kết quả
1	Trong	Cao	Bắc	Không mưa
2	Mây	Cao	Nam	Mưa
3	Mây	Trung bình	Bắc	Mưa
4	Trong	Thấp	Bắc	Không mưa
5	Mây	Thấp	Bắc	Mưa
6	Mây	Cao	Bắc	Mưa
7	Mây	Thấp	Nam	Không mưa
8	Trong	Cao	Nam	Không mưa

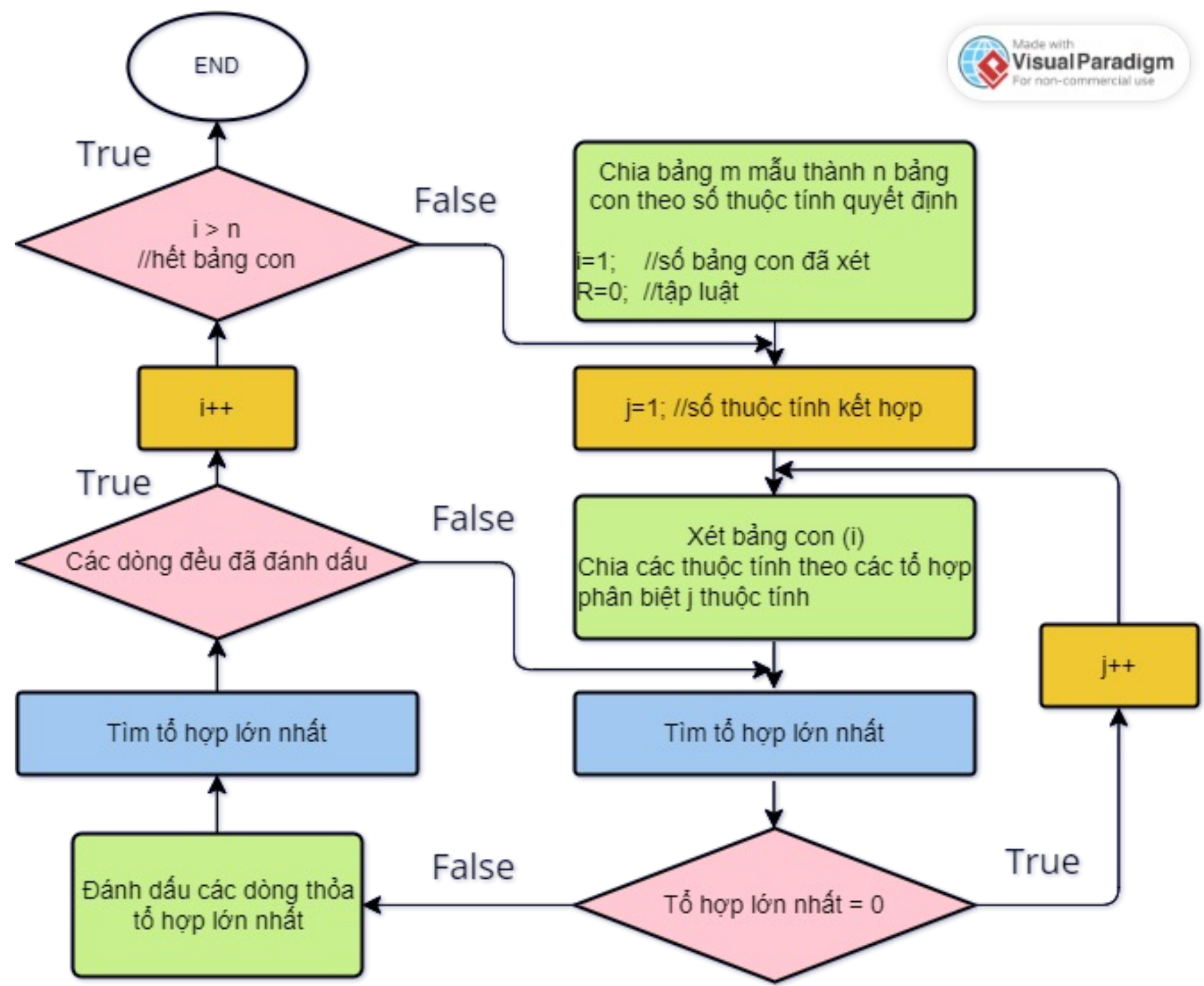
Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)

- Thuật giải ILA (Inductive Learning Algorithm) được dùng để xác định các luật phân loại cho tập hợp các mẫu học.
- Thuật giải này thực hiện theo cơ chế lặp, để tìm luật riêng đại diện cho tập mẫu của từng lớp.
- Sau khi xác định được luật, ILA loại bỏ các mẫu liên quan khỏi tập mẫu, đồng thời thêm luật mới này vào tập luật.

Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)



Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Xây dựng cây định danh bằng giải thuật ILA cho kết quả rám nắng với bảng dữ liệu sau

Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)

- Từ bảng dữ liệu có thể nhận ra số thuộc tính quyết định: $n=2$ nên tách thành 2 bảng con

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám

Bảng 1

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Bảng 2

- Thuộc tính kết quả (mục tiêu) có thể bỏ qua trong quá trình xét

Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)

▪ Xét bảng 1

▪ $j=1$ //xét TH có 1 thuộc tính

có 4 tổ hợp (màu tóc, c.cao, c.nặng, dùng kem)

• Xét thuộc tính: màu tóc

- vàng : 0 (cả bảng 1 và 2 đều có dữ liệu nên giá trị khác biệt = 0)
- đỏ:1 (bảng 1 có 1 dòng, bảng 2 không có dữ liệu)

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

{màu tóc}	{c.cao}	{c. nặng}	{dùng kem}
Vàng: 0 Đỏ: 1			

Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)

▪ Xét bảng 1

▪ $j=1$ // xét TH có 1 thuộc tính

có 4 tổ hợp (màu tóc, c.cao, c.nặng, dùng kem)

• Xét thuộc tính: chiều cao

- TB (bảng 1 và 2 đều có): 0
- thấp (bảng 1 và 2 đều có): 0

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

{màu tóc}	{c.cao}	{c. nặng}	{dùng kem}
Vàng: 0 Đỏ: 1	TB: 0 thấp: 0		

Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)

▪ Xét bảng 1

▪ $j=1$ //xét TH có 1 thuộc tính

có 4 tổ hợp (màu tóc, c.cao, c.nặng, dùng kem)

- Xét thuộc tính: cân nặng
 - nhẹ (bảng 1 và 2 đều có): 0
 - TB (bảng 1 và 2 đều có): 0
 - nặng (bảng 1 và 2 đều có): 0
- Xét thuộc tính: dùng kem
 - không (bảng 1 và 2 đều có): 0

IF màu tóc = đỏ THEN kết quả = bị nám

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

{màu tóc}	{c.cao}	{c. nặng}	{dùng kem}
Vàng: 0 Đỏ: 1	TB: 0 thấp: 0	TB: 0 nhẹ: 0 nặng: 0	không: 0

Tổ hợp max=1: Đỏ {màu tóc}

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám

~~Đánh dấu dòng màu tóc = đỏ~~

Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)

- **Xét bảng 1**

- **$j=1$ // xét TH có 1 thuộc tính**

có 4 tổ hợp (màu tóc, c.cao, c.nặng, dùng kem)

- Sau khi đánh dấu dòng đã xét, tính lại bảng tổ hợp cho các dòng chưa xét với 1 thuộc tính phân biệt

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

{màu tóc}	{c.cao}	{c. nặng}	{dùng kem}
Vàng: 0	TB: 0 thấp: 0	TB: 0 nhẹ: 0 nặng: 0	không: 0

Tăng giá trị của j (số thuộc tính cần xét = 2)



Tổ hợp max=0: không có luật mới

Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)

- **Xét bảng 1**

- **j=2 //xét TH có 2 thuộc tính**

có 6 tổ hợp: (m. tóc, c.cao), (m.tóc, c.nặng), (m.tóc, dùng kem),
(c.cao, c.nặng), (c.cao, dùng kem)
(c.nặng, dùng kem)

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

{màu tóc, c.cao}		{m.tóc, d.kem}
{vàng, TB}	{vàng, thấp}	{vàng, không}
1	0	2 (2 bản ghi # vs bảng 2)

{màu tóc, c.nặng}		{chiều cao, c.nặng}		{chiều cao, d.kem}		{c.nặng, d.kem}	
{vàng, TB}	{vàng, nhẹ}	{TB, nhẹ}	{thấp, TB}	{TB, không}	{thấp, không}	{TB, không}	{nhẹ, không}
0	0	1	0	1	1	1	1

Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)

- **Xét bảng 1**

- **$j=2$ // xét TH có 2 thuộc tính**

có 6 tổ hợp: (m. tóc, c.cao), (m.tóc, c.nặng), (m.tóc, dùng kem),
(c.cao, c.nặng), (c.cao, dùng kem)
(c.nặng, dùng kem)

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Thêm luật vào R:
IF màu tóc = vàng AND dùng kem = không
THEN kết quả = bị nám

Tổ hợp max=2: (màu tóc, dùng kem)
= {vàng, không}

~~Đánh dấu dòng màu tóc = vàng và dùng kem = không~~

Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)

- **Xét bảng 2 (sau khi hết dòng B1)**

- **j=1 //xét TH có 1 thuộc tính**

có 4 tổ hợp: {m. tóc}, {c.cao}, {c.nặng}, {dùng kem}

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám

{màu tóc}	{c.cao}	{c. nặng}	{dùng kem}
Vàng: 0 nâu: 3 (số dòng \notin B1)	TB: 0 thấp: 0 cao: 2	TB: 0 nhẹ: 0 nặng: 0	không: 0 có: 3

Thêm luật vào R:

IF màu tóc = nâu
THEN kết quả = không bị nám

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

Tổ hợp max=3: (màu tóc)= {nâu} và (dùng kem) = {có}. Chọn màu tóc

~~Đánh dấu dòng màu tóc = nâu~~

Chương 7. MÁY HỌC

Học quy nạp – Inductive learning Algorithm (ILA)

- **Xét bảng 2**
- **j=1 //xét TH có 1 thuộc tính**
có 4 tổ hợp (màu tóc, c.cao, c.nặng, dùng kem)
- Sau khi đánh dấu dòng đã xét, tính lại bảng tổ hợp cho các dòng chưa xét với 1 thuộc tính phân biệt

Thêm luật vào R:

IF dùng kem = có
THEN kết quả = không bị nám

Số dòng chưa đánh dấu bảng 2 = \emptyset : END

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Sarah	Vàng	TB	Nhẹ	Không	Nám
Amie	Vàng	Thấp	TB	Không	Nám
Emilie	Đỏ	TB	Nặng	Không	Nám

Tên	Màu tóc	C.cao	C.nặng	Dùng kem	Kết quả
Dana	Vàng	Cao	TB	Có	Không
Alex	Nâu	Thấp	TB	Có	Không
Peter	Nâu	Cao	Nặng	Không	Không
John	Nâu	TB	Nặng	Không	Không
Kartie	Vàng	Thấp	Nhẹ	Có	Không

{màu tóc}	{c.cao}	{c. nặng}	{dùng kem}
Vàng: 0	TB: 0 thấp: 0	TB: 0 nhẹ: 0	có: 2

Tổ hợp max=2: (dùng kem) = {có}

~~Đánh dấu dòng dùng kem = có~~

Chương 7. MÁY HỌC

Gini Index

- Xem xét ví dụ sau:
- Chúng ta sử dụng giải thuật **Entropy** để tính giá trị cho các thuộc tính.
- Xét thuộc tính Weather (Sunny:3, Windy: 4, Rainy:3)
 - $Sunny: \frac{1}{3} Cinema, \frac{2}{3} Tennis, \frac{0}{3} Stayin, \frac{0}{3} Shopping$
 - $Windy: \frac{3}{4} Cinema, \frac{0}{4} Tennis, \frac{0}{4} Stayin, \frac{1}{4} Shopping$
 - $Rainy: \frac{2}{3} Cinema, \frac{0}{3} Tennis, \frac{1}{3} Stayin, \frac{0}{3} Shopping$
 - $E(weather) = \frac{3}{10} \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} - 0 - 0 \right) + \frac{4}{10} \left(-\frac{3}{4} \log_2 \frac{3}{4} - 0 - 0 - \frac{1}{4} \log_2 \frac{1}{4} \right) + \frac{3}{10} \left(-\frac{2}{3} \log_2 \frac{2}{3} - 0 - \frac{1}{3} \log_2 \frac{1}{3} \right)$
 - $E(weather) = 0.8755$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Chương 7. MÁY HỌC

Gini Index

- Xem xét ví dụ sau:
- Chúng ta sử dụng giải thuật **Entropy** để tính giá trị cho các thuộc tính.
- Xét thuộc tính Parent (Yes:5, No:5)

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Yes: $\frac{5}{5} Cinema, \frac{0}{5} Tennis, \frac{0}{5} Stayin, \frac{0}{5} Shopping$
- No: $\frac{2}{5} Cinema, \frac{2}{5} Tennis, \frac{1}{5} Stayin, \frac{1}{5} Shopping$
- $$E(parent) = \frac{5}{10} \left(-\frac{5}{5} \log_2 \frac{5}{5} - 0 - 0 - 0 \right) + \frac{5}{10} \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{1}{5} \log_2 \frac{1}{5} - \frac{1}{5} \log_2 \frac{1}{5} - \frac{1}{5} \log_2 \frac{1}{5} \right)$$
- $E(parent) = 0.961$

Chương 7. MÁY HỌC

Gini Index

- Xem xét ví dụ sau:
- Chúng ta sử dụng giải thuật **Entropy** để tính giá trị cho các thuộc tính.
- Xét thuộc tính Money (Rich:7, Poor:3)

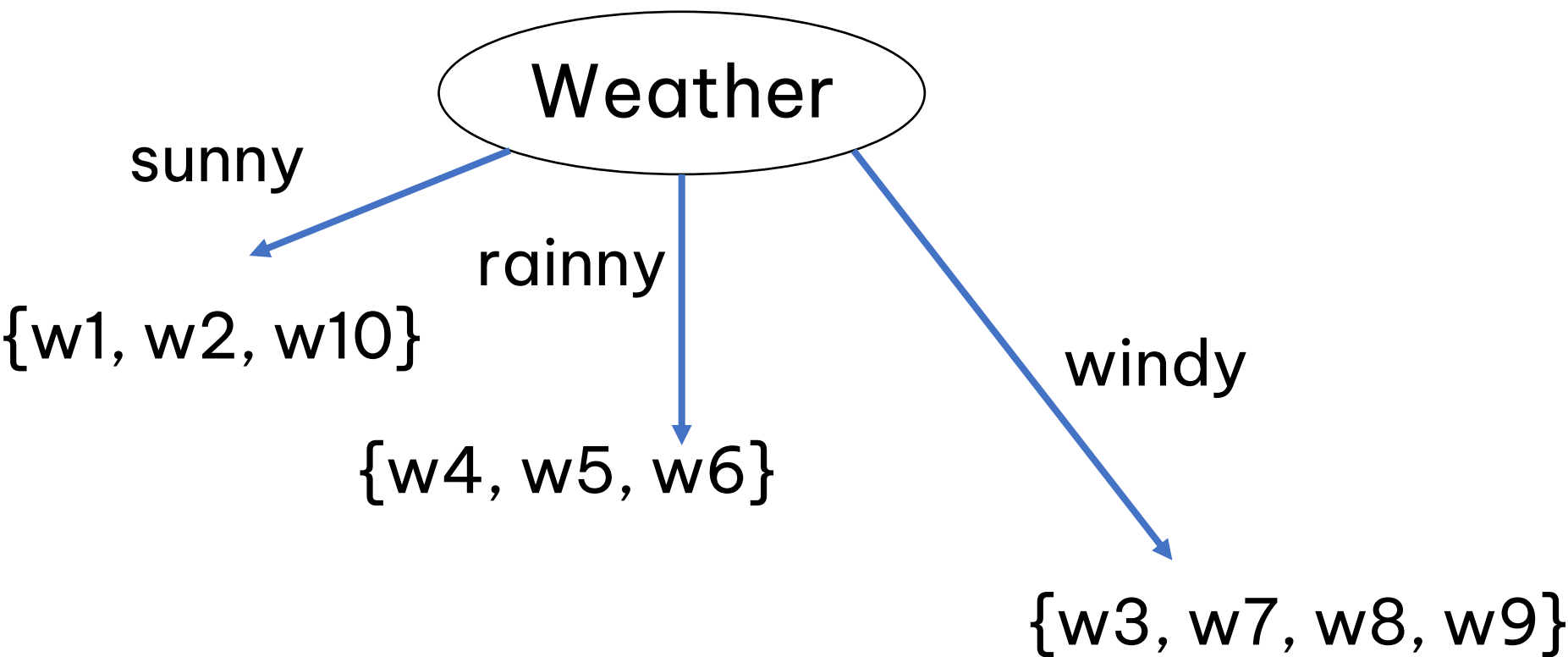
Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- $Rich: \frac{3}{7} Cinema, \frac{2}{7} Tennis, \frac{1}{7} Stayin, \frac{1}{7} Shopping$
- $Poor: \frac{3}{3} Cinema, \frac{0}{3} Tennis, \frac{0}{3} Stayin, \frac{0}{3} Shopping$
- $E(money) = \frac{7}{10} \left(-\frac{3}{7} \log_2^{\frac{3}{7}} - \frac{2}{7} \log_2^{\frac{2}{7}} - \frac{1}{7} \log_2^{\frac{1}{7}} - \frac{1}{7} \log_2^{\frac{1}{7}} \right) + \frac{3}{10} \left(-\frac{3}{3} \log_2^{\frac{3}{3}} - 0 - 0 - 0 \right)$
- $E(money) = 1.287$

Nhìn vào kết quả Entropy (thuộc tính) ta thấy giá trị của thuộc tính Weather là bé nhất = 0.8755 (<0.961, <1.287). Nên sử dụng thuộc tính Weather để phân hoạch

Chương 7. MÁY HỌC

Gini Index



Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Sự phân hoạch này là chưa tối ưu vì mỗi giá trị thuộc tính của weather đều chứa giá trị của Decision = {cinema, tennis, stayin, shopping}

Sẽ phù hợp hơn nếu chúng ta dùng giá trị của Parent để phân hoạch. Nếu Parent = Yes thì Decision = Cinema.

Chương 7. MÁY HỌC

Gini Index

- ❑ Gini Index (hoặc Gini Impurity) là một phương pháp được sử dụng trong việc xây dựng cây quyết định và trong bài toán phân loại trong lĩnh vực Machine Learning và Data Mining.
- ❑ Gini Index đo lường độ tinh khiết (impurity) của dữ liệu trong một nút của cây quyết định. Điểm số Gini Index càng thấp, thì dữ liệu càng tinh khiết và dễ dự đoán.
- ❑ Gini Index được tính bằng cách xem xét tỷ lệ các lớp dữ liệu khác nhau trong một nút.

Chương 7. MÁY HỌC

Gini Index

- Công thức tính Gini Index cho một nút dựa trên phân phối lớp là:

$$Gini(S) = 1 - \sum p_i^2$$

- Trong đó:
 - S là nút cần tính *Gini Index*
 - p_i là tỷ lệ của lớp i trong nút S
- Gini Index thường nằm trong khoảng từ 0 đến 0.5, và khi Gini Index càng thấp, tức là nút càng tinh khiết. Khi xây dựng cây quyết định, chúng ta chọn thuộc tính có Gini Index thấp nhất làm thuộc tính phân chia tại nút đó, vì nó giúp làm tăng độ tinh khiết của dữ liệu trong các nút con sau đó.

Chương 7. MÁY HỌC

Gini Index

- Sử dụng giải thuật Gini như sau:
- Xét thuộc tính Weather có 3 giá trị: Sunny (3), Windy(4), Rainny(3)
 - ✓ Weather=sunny:
 - + cinema = 1
 - + tennis = 2
 - + stayin = 0
 - + shopping = 0

$Gini(weather = sunny)$

$$= 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(\frac{0}{3}\right)^2 + \left(\frac{0}{3}\right)^2 \right] = 0.444$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Chương 7. MÁY HỌC

Gini Index

- Sử dụng giải thuật Gini như sau:
- Xét thuộc tính Weather có 3 giá trị: Sunny (3), Windy(4), Rainny(3)
 - ✓ Weather=windy:
 - + cinema = 3
 - + tennis = 0
 - + stayin = 0
 - + shopping = 1

Gini(weather = sunny)

$$= 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{0}{4}\right)^2 + \left(\frac{0}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right] = 0.375$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Chương 7. MÁY HỌC

Gini Index

- Sử dụng giải thuật Gini như sau:
- Xét thuộc tính Weather có 3 giá trị: Sunny (3), Windy(4), Rainny(3)
 - ✓ Weather=rainny:
 - + cinema = 2
 - + tennis = 0
 - + stayin = 1
 - + shopping = 0

$Gini(weather = sunny)$

$$= 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{0}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{0}{3}\right)^2 \right] = 0.444$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

$$\begin{aligned} Gini(weather) &= \frac{3}{10} * 0.444 \\ &+ \frac{4}{10} * 0.375 \\ &+ \frac{3}{10} * 0.444 \\ &= 0.416 \end{aligned}$$

Chương 7. MÁY HỌC

Gini Index

- Sử dụng giải thuật Gini như sau:
- Xét thuộc tính **Parent** có 2 giá trị: Yes (5), No(5)
- Parent=yes:
 - + cinema = 5
 - + tennis = 0
 - + stayin = 0
 - + shopping = 0

$Gini(parent = yes)$

$$= 1 - \left[\left(\frac{5}{5}\right)^2 + \left(\frac{0}{5}\right)^2 + \left(\frac{0}{5}\right)^2 + \left(\frac{0}{5}\right)^2 \right] = 0$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Chương 7. MÁY HỌC

Gini Index

- Sử dụng giải thuật Gini như sau:
- Xét thuộc tính **Parent** có 2 giá trị: Yes (5), No(5)
- Parent=no:
 - + cinema = 1
 - + tennis = 2
 - + stayin = 1
 - + shopping = 1

$Gini(parent = no)$

$$= 1 - \left[\left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2 \right] = 0.72$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

$$\begin{aligned} Gini(parent) &= \frac{5}{10} * 0 \\ &+ \frac{5}{10} * 0.72 \\ &= 0.36 \end{aligned}$$

Chương 7. MÁY HỌC

Gini Index

- Sử dụng giải thuật Gini như sau:
- Xét thuộc tính **Money** có 2 giá trị: Rich (7), Poor(3)
- Money=rich:
 - + cinema = 3
 - + tennis = 2
 - + stayin = 1
 - + shopping = 1

$Gini(money = rich)$

$$= 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{2}{7}\right)^2 + \left(\frac{1}{7}\right)^2 + \left(\frac{1}{7}\right)^2 \right] = 0.694$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Chương 7. MÁY HỌC

Gini Index

- Sử dụng giải thuật Gini như sau:
- Xét thuộc tính **Money** có 2 giá trị: Rich (7), Poor(3)
- Money=poor:
 - + cinema = 3
 - + tennis = 0
 - + stayin = 0
 - + shopping = 0

$Gini(money = rich)$

$$= 1 - \left[\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2 + \left(\frac{0}{3}\right)^2 + \left(\frac{0}{3}\right)^2 \right] = 0$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stayin
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

$$\begin{aligned} Gini(money) &= \frac{7}{10} * 0.694 \\ &\quad + \frac{3}{10} * 0 \\ &= 0.486 \end{aligned}$$

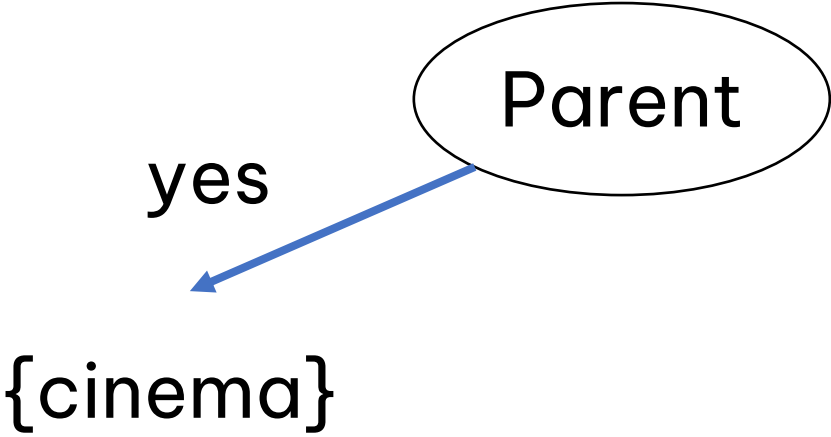
Chương 7. MÁY HỌC

Gini Index

Giá trị Gini trung bình của các thuộc tính	
Weather	0.416
Parent	0.36 ✓
Money	0.486

Ta chọn thuộc tính Parent để phân hoạch

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stayin
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis



Chương 7. MÁY HỌC

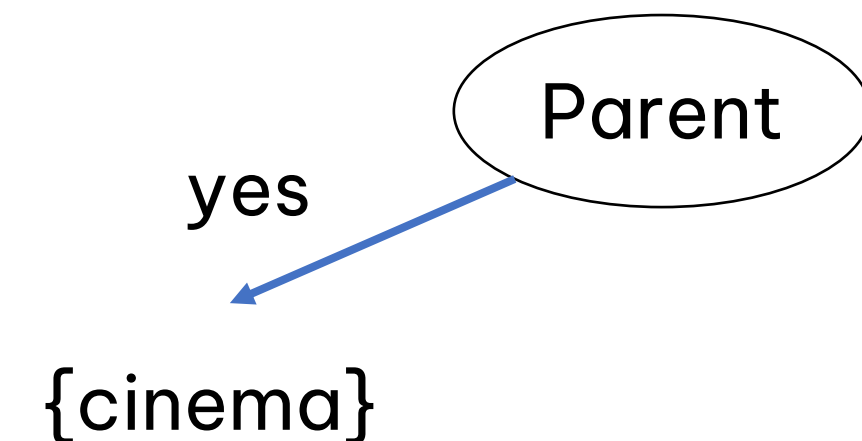
Gini Index

- Xét thuộc tính **Weather** có 3 giá trị: Sunny (2), Windy(2), Rainy(1)
- Weather=sunny:
 - + cinema = 0
 - + tennis = 2
 - + stayin = 0
 - + shopping = 0

$Gini(weather = sunny)$

$$= 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2 + \left(\frac{0}{2}\right)^2 \right] = 0$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stayin
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis



Chương 7. MÁY HỌC

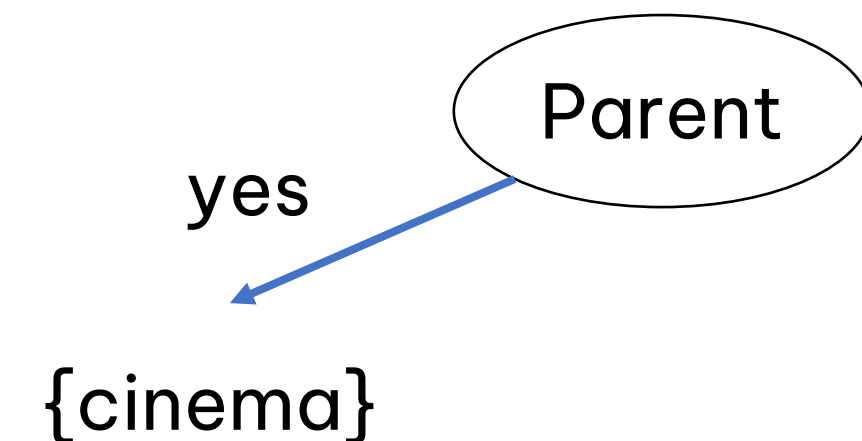
Gini Index

- Xét thuộc tính **Weather** có 3 giá trị: Sunny (2), Windy(2), Rainy(1)
- Weather=windy:
 - + cinema = 1
 - + tennis = 0
 - + stayin = 0
 - + shopping = 1

$Gini(weather = windy)$

$$= 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{0}{2}\right)^2 + \left(\frac{0}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right] = 0.5$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stayin
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis



Chương 7. MÁY HỌC

Gini Index

- Xét thuộc tính **Weather** có 3 giá trị: Sunny (2), Windy(2), Rainy(1)

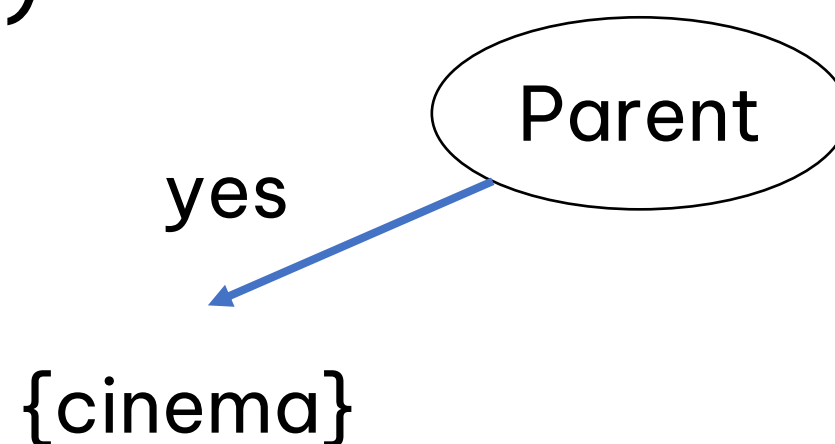
- Weather=rainy:

+ cinema = 0

+ tennis = 0

+ stayin = 1

+ shopping = 0



$Gini(weather = windy)$

$$= 1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2 \right] = 0$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stayin
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

$$\begin{aligned} Gini(weather) &= \frac{2}{5} * 0 \\ &+ \frac{2}{5} * 0.5 \\ &+ \frac{1}{5} * 0 \\ &= 0.2 \end{aligned}$$

Chương 7. MÁY HỌC

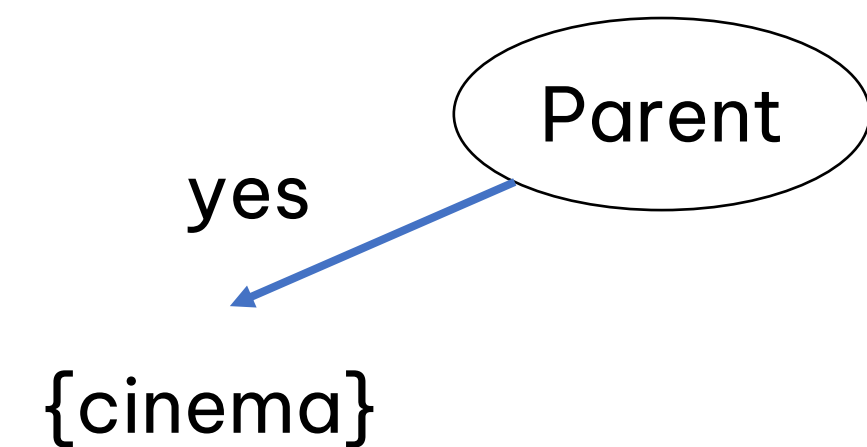
Gini Index

- Xét thuộc tính **Money** có 2 giá trị: Rich (4), Poor(1)
- Money=rich:
 - + cinema = 0
 - + tennis = 2
 - + stayin = 1
 - + shopping = 1

$Gini(money = rich)$

$$= 1 - \left[\left(\frac{0}{4}\right)^2 + \left(\frac{2}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right] = 0.625$$

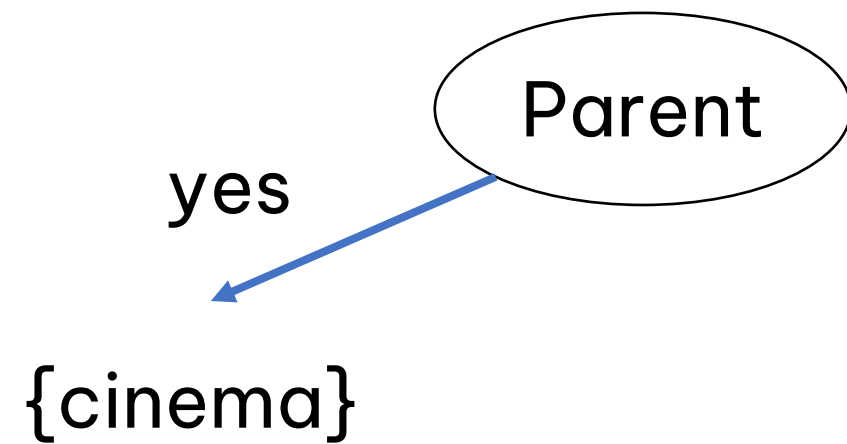
Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stayin
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis



Chương 7. MÁY HỌC

Gini Index

- Xét thuộc tính **Money** có 2 giá trị: Rich (4), Poor(1)
- Money=poor:
 - + cinema = 1
 - + tennis = 0
 - + stayin = 0
 - + shopping = 0



$Gini(money = rich)$


$$= 1 - \left[\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2 + \left(\frac{0}{1}\right)^2 + \left(\frac{0}{1}\right)^2 \right] = 0$$

Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stayin
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

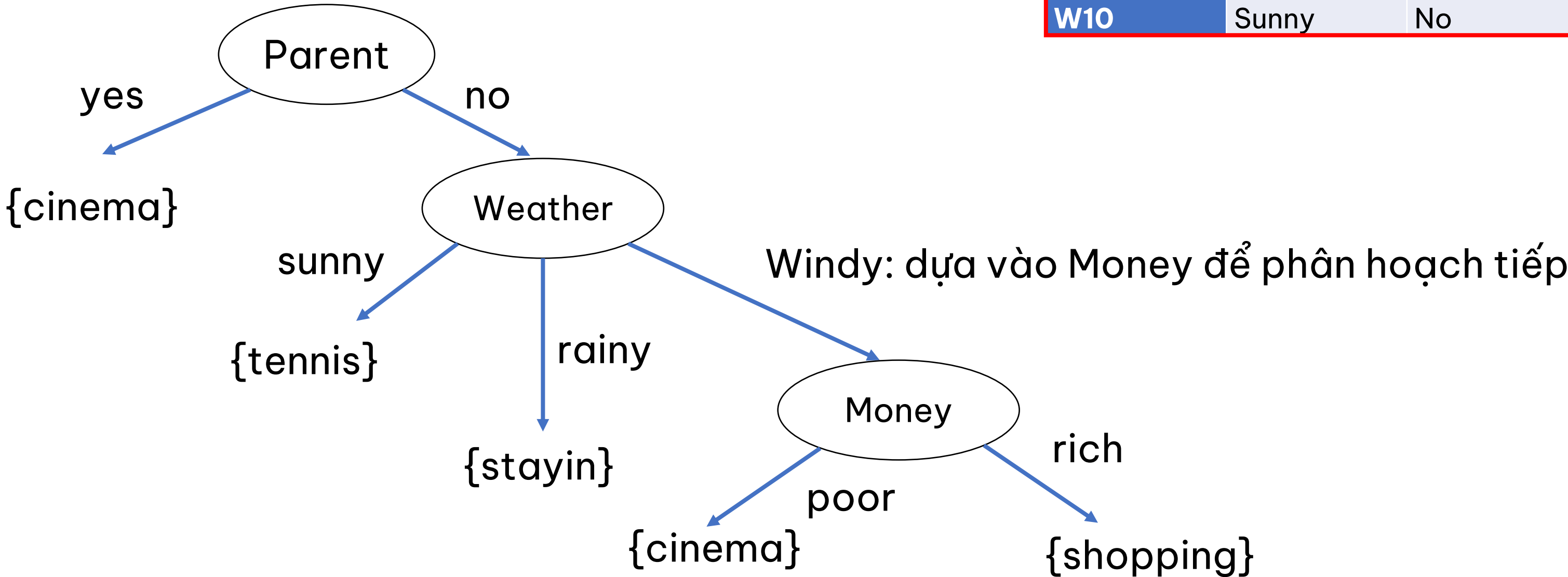
$$Gini(money) = \frac{4}{5} * 0.625 + \frac{1}{5} * 0 = 0.5$$

Chương 7. MÁY HỌC

Gini Index

Giá trị Gini trung bình của các thuộc tính	
Weather	0.2 
Money	0.5

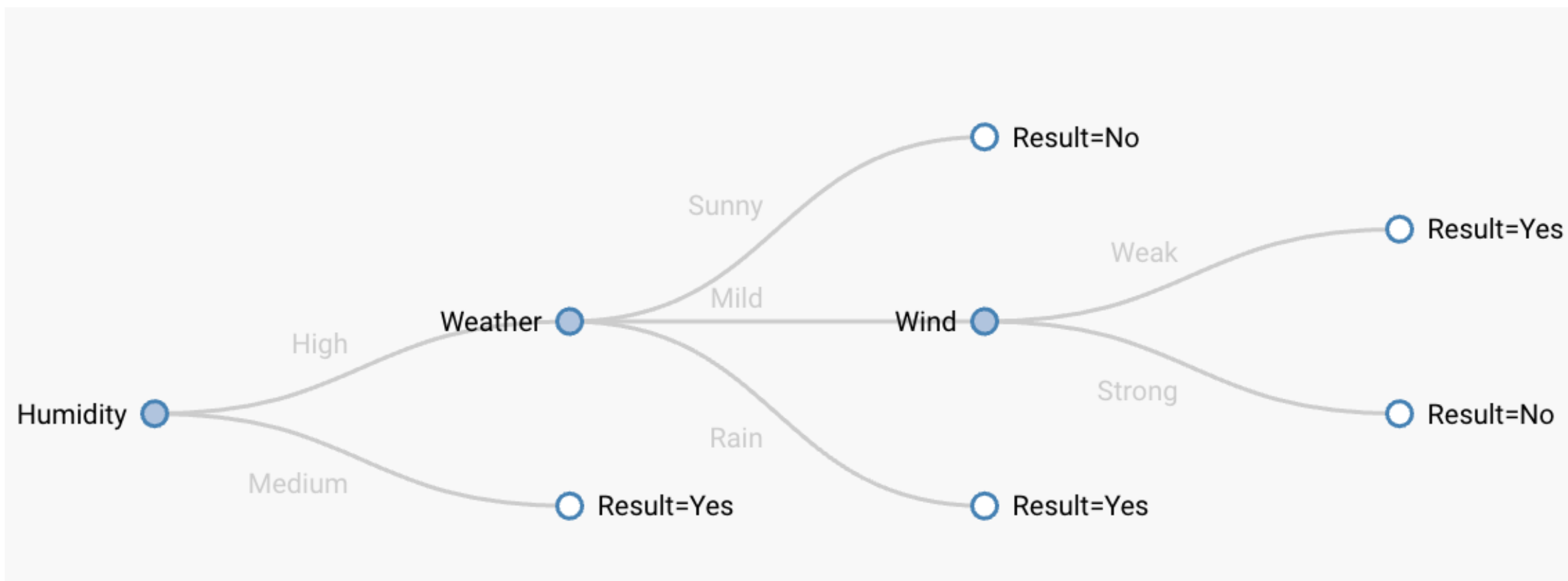
Ta chọn thuộc tính weather để phân hoạch tiếp theo



Weekend	Weather	Parent	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stayin
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Chương 7. MÁY HỌC

Xây dựng cây quyết định để đưa ra luật phù hợp



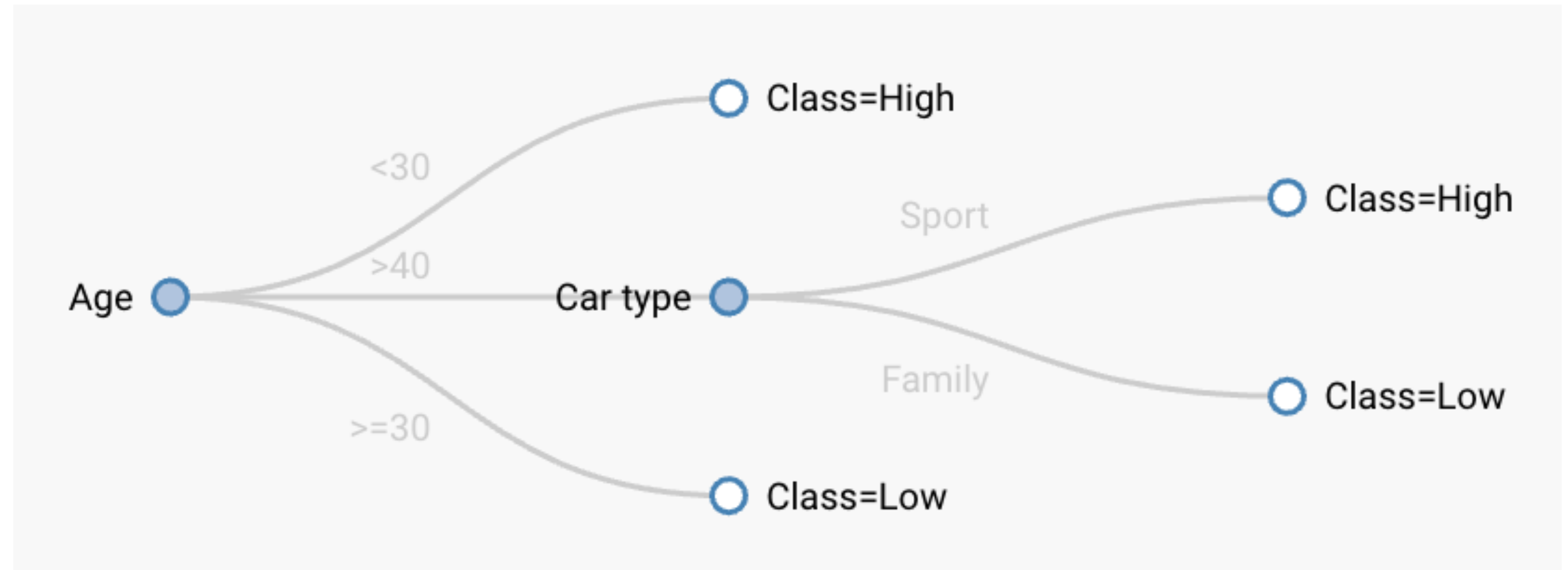
#	Day	Weather	Humidity	Wind	Result
1	D1	Sunny	High	Weak	No
2	D2	Sunny	High	Strong	No
3	D3	Mild	High	Weak	Yes
4	D4	Rain	High	Weak	Yes
5	D5	Rain	Medium	Strong	Yes
6	D6	Mild	High	Strong	No
7	D7	Mild	Medium	Weak	Yes
8	D8	Sunny	High	Weak	No
9	D9	Sunny	Medium	Weak	Yes
10	D10	Rain	Medium	Weak	Yes

Chương 7. MÁY HỌC

Xây dựng cây quyết định để đưa ra luật phù hợp

Age	Car type	Class
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High
AGE có thể phân thành 3 nhóm:		
<30	>=30	>=40

Age	Car type	Class
<30	Family	High
<30	Sport	High
>40	Sport	High
>40	Family	Low
>=30	Truck	Low
<30	Family	High



Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Thuật toán Naïve Bayes là một thuật toán học có giám sát, dựa trên định lý Bayes và được sử dụng để giải các bài toán phân loại.
- Được sử dụng chủ yếu trong phân loại văn bản.
- Naïve Bayes Classifier là một trong những thuật toán Phân loại đơn giản và hiệu quả nhất giúp xây dựng các mô hình học máy nhanh có thể đưa ra dự đoán nhanh.
 - Naïve: Giả định sự xuất hiện của một đặc điểm nào đó là độc lập với sự xuất hiện của các đặc điểm khác (độc lập ngẫu nhiên).
 - Bayes: Định lý Bayes

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Định lý Bayes: tìm xác suất của một sự kiện xảy ra với xác suất của một sự kiện khác đã xảy ra. Định lý Bayes được phát biểu về mặt toán học dưới dạng phương trình sau:

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

Trong đó A và B là các sự kiện và $P(B) \neq 0$.

- Giả định Naïve: nếu hai sự kiện A và B bất kỳ là độc lập, thì $P(A, B) = P(A).P(B)$

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Do vậy công thức Naïve Bayes được phát biểu như sau:

$$P(C_i/X) = \frac{P(X/C_i) * P(C_i)}{P(X)} = \frac{P(X/C_i) * P(C_i)}{P(x_1) * P(x_2) * \dots * P(x_n)}$$

- Ví dụ: Tìm xác suất của nhãn i = mưa hoặc không mưa khi biết điều kiện X = ít mây, áp suất thấp và gió thổi từ hướng nam.*
- $P(C_i/X)$: xác suất xảy ra nhãn i khi biết X
- $P(X/C_i)$: xác suất X khi biết nhãn i
- $P(C_i)$: xác suất xảy ra nhãn i
- $P(X)$: xác suất xảy ra nhãn X
- x_i : điều kiện thứ i

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Vì mẫu số không đổi đối với một đầu vào nhất định, chúng ta có thể loại bỏ phần mẫu:

$$P(C_i/X) \propto \prod_{i=1}^n P(X/C_i) * P(C_i)$$

- Ví dụ: Tìm xác suất của nhãn i = mưa hoặc không mưa khi biết điều kiện X = ít mây, áp suất thấp và gió thổi từ hướng nam.
- $P(C_i/X)$: xác suất xảy ra nhãn i khi biết X
- $P(X/C_i)$: xác suất X khi biết nhãn i
- $P(C_i)$: xác suất xảy ra nhãn i

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi $X_{(color=Red, Type=SUV, Origin=Domestic)}$?
- Bước 1: tính $P(C_i)$
- Bước 2: tính $P(X/c_i)$
- Bước 3: tính $P(c_i/x)$

Bước 1: tính $P(C_i)$

$$P(Yes) = \frac{5}{10} = 0.5$$

$$P(No) = \frac{5}{10} = 0.5$$

#	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi $X_{(color=Red, Type=SUV, Origin=Domestic)}$?

Bước 2: tính $P(X/c_i)$

Color	Yes	No
Red	3/5	2/5
Yellow	2/5	3/5

Type	Yes	No
Sports	4/5	2/5
SUV	1/5	3/5

Origin	Yes	No
Domestic	2/5	3/5
Imported	3/5	2/5

#	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi $X_{(color=Red, Type=SUV, Origin=Domestic)}$?
- Bước 3: tính

$$P(C_i/X) = \prod_{i=1}^n P(X/C_i) * P(C_i)$$

$$P(Yes/X) = 0.5 * \frac{3}{5} * \frac{1}{5} * \frac{2}{5} = 0.024$$

$$P(No/X) = 0.5 * \frac{2}{5} * \frac{3}{5} * \frac{1}{5} = 0.072$$

$$Ta\ có: P(Yes/X) + P(No/X) = 1$$

$$P(Yes/X) = \frac{0.024}{0.024 + 0.072} = 0.25$$

$$P(No/X) = 1 - 0.25 = 0.75$$

Khả năng mất cắp (stolen)
với những loại xe thoả điều
kiện X là 25%

#	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Origin	Yes	No
Domestic	2/5	3/5
Imported	3/5	2/5

Type	Yes	No	Color	Yes	No
Sports	4/5	2/5	Red	3/5	2/5
SUV	1/5	3/5	Yellow	2/5	3/5

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi $X_{(Outlook=Rainy, Temp=Hot, Humidity=High, Windy=False)}$?
- Bước 1: tính $P(C_i)$
- Bước 2: tính $P(X/C_i)$
- Bước 3: tính $P(C_i/X)$

Bước 1: tính $P(C_i)$

$$P(Yes) = \frac{9}{14}$$

$$P(No) = \frac{5}{14}$$

#	Outlook	Temperature	Humidity	Windy	Playgolf
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi $X_{(Outlook=Rainy, Temp=Hot, Humidity=High, Windy=False)}$?

Bước 2: tính $P(X/c_i)$

Outlook	Yes	No
Rainy	2/9	3/5
Overcast	4/9	0/5
Sunny	3/9	2/5

Temperature	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Yes	No
High	3/9	4/5
Normal	6/9	1/5

#	Outlook	Temperature	Humidity	Windy	Playgolf
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

Windy	Yes	No
True	6/9	2/5
False	3/9	3/5

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi $X_{(Outlook=Rainy, Temp=Hot, Humidity=High, Windy=False)}$?

- Bước 3: tính

$$P(C_i/X) = \prod_{i=1}^n P(X/C_i) * P(C_i)$$

$$P(Yes/X) = \frac{9}{14} * \frac{2}{9} * \frac{2}{9} * \frac{3}{5} * \frac{3}{9} = 0.0064$$

$$P(No/X) = \frac{5}{14} * \frac{3}{5} * \frac{2}{5} * \frac{4}{5} * \frac{3}{5} = 0.041$$

$$Ta\ có: P(Yes/X) + P(No/X) = 1$$

$$P(Yes/X) = \frac{0.0064}{0.0064 + 0.041} = 0.135$$

$$P(No/X) = 1 - 0.135 = 0.865$$

#	Outlook	Temperature	Humidity	Windy	Playgolf
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

Outlook	Yes	No	Temperature	Yes	No	Humidity	Yes	No	Windy	Yes	No
Rainy	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	True	6/9	2/5
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	False	3/9	3/5
Sunny	3/9	2/5	Cool	3/9	1/5						

Khả năng Playgolf = No chiếm tỷ lệ 86,5%

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi $X_{(Mây=ít, \text{Áp suất}=thấp, \text{Gió}=nam)}$?
- Bước 1: tính $P(C_i)$
- Bước 2: tính $P(X/c_i)$
- Bước 3: tính $P(c_i/x)$

Bước 1: tính $P(C_i)$

$$P(\text{Không mưa}) = \frac{4}{8}$$

$$P(\text{Mưa}) = \frac{4}{8}$$

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	Không mưa
2	nhiều	cao	Bắc	Mưa
3	ít	thấp	Bắc	Không mưa
4	nhiều	thấp	Bắc	Mưa
5	nhiều	trung bình	Bắc	Mưa
6	ít	cao	Nam	Không mưa
7	nhiều	cao	Nam	Mưa
8	nhiều	thấp	Nam	Không mưa

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi $X_{(Mây=ít, \text{Áp suất}=thấp, \text{Gió}=nam)}$?

Bước 2: tính $P(X/c_i)$

Mây	Không mưa	Mưa
ít	3/4	0/4
nhiều	1/4	4/4

Áp suất	Không mưa	Mưa
cao	2/4	2/4
thấp	2/4	1/4
trung bình	0/4	1/4

Gió	Không mưa	Mưa
nam	2/4	1/4
bắc	2/4	3/4

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	Không mưa
2	nhiều	cao	Bắc	Mưa
3	ít	thấp	Bắc	Không mưa
4	nhiều	thấp	Bắc	Mưa
5	nhiều	trung bình	Bắc	Mưa
6	ít	cao	Nam	Không mưa
7	nhiều	cao	Nam	Mưa
8	nhiều	thấp	Nam	Không mưa

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi $X_{(Mây=ít, \text{Áp suất}=thấp, \text{Gió}=nam)}$?

- Bước 3: tính

$$P(C_i/X) = \prod_{i=1}^n P(X/C_i) * P(C_i)$$

$$P(Không \text{ Mưa}/X) = \frac{4}{8} * \frac{3}{4} * \frac{2}{4} * \frac{2}{4} = 0.094$$

$$P(Mưa/X) = \frac{4}{8} * \frac{0}{4} * \frac{1}{4} * \frac{1}{4} = 0$$

Để tránh trường hợp cho xác suất = 0

ta sử dụng biến đổi Laplace

$$P_{Lap,k}(x|y) = \frac{C(x,y) + k}{c(y) + k * X}$$

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	Không mưa
2	nhiều	cao	Bắc	Mưa
3	ít	thấp	Bắc	Không mưa
4	nhiều	thấp	Bắc	Mưa
5	nhiều	trung bình	Bắc	Mưa
6	ít	cao	Nam	Không mưa
7	nhiều	cao	Nam	Mưa
8	nhiều	thấp	Nam	Không mưa

Mây	Không mưa	Mưa	Áp suất	Không mưa	Mưa
ít	3/4	0/4	cao	2/4	2/4
nhiều	1/4	4/4	thấp	2/4	1/4
Gió	Không mưa	Mưa	trung bình	0/4	1/4
nam	2/4	1/4			
bắc	2/4	3/4			

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi

$X_{(Mây=ít, \text{Áp suất}=thấp, \text{Gió}=nam)}$?

- Tính lại bước 2:

- $P(X/c_i) = P(X/mưa), \text{ với } X = \{ít, thấp, nam\}$

- $P(it/mưa) = \frac{C(it=mưa)+k}{C(mưa)+k*(\text{số chiều của thuộc tính } X=mây)}$

- $P(mây=ít/mưa) = \frac{0+1}{4+1*3} = \frac{1}{7}$

- $P(a.suất=thấp/mưa) = \frac{1+1}{4+1*2} = \frac{2}{6}$

- $P(gió=nam/mưa) = \frac{1+1}{4+1*2} = \frac{2}{6}$

- trong đó: $k=1$, C : count

$$P_{Lap,k}(x|y) = \frac{C(x,y) + k}{C(y) + k * X}$$

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	Không mưa
2	nhiều	cao	Bắc	Mưa
3	ít	thấp	Bắc	Không mưa
4	nhiều	thấp	Bắc	Mưa
5	nhiều	trung bình	Bắc	Mưa
6	ít	cao	Nam	Không mưa
7	nhiều	cao	Nam	Mưa
8	nhiều	thấp	Nam	Không mưa

Mây	Không mưa	Mưa	Áp suất	Không mưa	Mưa
ít	3/4	0/4	cao	2/4	2/4
nhiều	1/4	4/4	thấp	2/4	1/4
Gió	Không mưa	Mưa	trung bình	0/4	1/4
nam	2/4	1/4			
bắc	2/4	3/4			

số chiều $X_{mây} = 3$
số chiều $X_{áp suất} = 2$
số chiều $X_{gió} = 2$

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi trời ít mây, áp suất thấp và gió thổi từ hướng Nam?

- Bước 2:

- $P(X/c_i) = P(X/mưa), \text{ với } X = \{\text{ít}, \text{thấp}, \text{nam}\}$

- $$P(X/mưa) = P\left(\frac{\text{ít}}{mưa}\right) \cdot P\left(\frac{\text{thấp}}{mưa}\right) \cdot P\left(\frac{\text{nam}}{mưa}\right) =$$
$$= \frac{1}{7} * \frac{2}{6} * \frac{2}{6} = \frac{1}{63}$$

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	Không mưa
2	nhiều	cao	Bắc	Mưa
3	ít	thấp	Bắc	Không mưa
4	nhiều	thấp	Bắc	Mưa
5	nhiều	trung bình	Bắc	Mưa
6	ít	cao	Nam	Không mưa
7	nhiều	cao	Nam	Mưa
8	nhiều	thấp	Nam	Không mưa

$$P_{Lap,k}(x|y) = \frac{C(x,y) + k}{C(y) + k * X}$$

số chiều $X_{mây} = 3$

số chiều $X_{áp suất} = 2$

số chiều $X_{gió} = 2$

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi trời ít mây, áp suất thấp và gió thổi từ hướng Nam?

- Bước 2:

- $P(X/c_i) = P(X/\text{không mưa}), \text{ với } X = \{\text{ít}, \text{thấp}, \text{nam}\}$

- $P(X/k.\text{mưa}) = P\left(\frac{\text{ít}}{k.\text{mưa}}\right) \cdot P\left(\frac{\text{thấp}}{k.\text{mưa}}\right) \cdot P\left(\frac{\text{nam}}{k.\text{mưa}}\right) =$

- $P(X/k.\text{mưa}) = \frac{3+1}{4+1*3} * \frac{2+1}{4+1*2} * \frac{2+1}{4+1*2} = \frac{9}{63}$

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	Không mưa
2	nhiều	cao	Bắc	Mưa
3	ít	thấp	Bắc	Không mưa
4	nhiều	thấp	Bắc	Mưa
5	nhiều	trung bình	Bắc	Mưa
6	ít	cao	Nam	Không mưa
7	nhiều	cao	Nam	Mưa
8	nhiều	thấp	Nam	Không mưa

$$P_{Lap,k}(x|y) = \frac{C(x,y) + k}{C(y) + k * X}$$

số chiều $X_{\text{mây}} = 3$

số chiều $X_{\text{áp suất}} = 2$

số chiều $X_{\text{gió}} = 2$

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- Khả năng nào xảy ra khi trời ít mây, áp suất thấp và gió thổi từ hướng Nam?
- Bước 3: $P(C_i/X)$, với $X = \{\text{ít, thấp, nam}\}$

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	Không mưa
2	nhiều	cao	Bắc	Mưa
3	ít	thấp	Bắc	Không mưa
4	nhiều	thấp	Bắc	Mưa
5	nhiều	trung bình	Bắc	Mưa
6	ít	cao	Nam	Không mưa
7	nhiều	cao	Nam	Mưa
8	nhiều	thấp	Nam	Không mưa

$$P(\text{Không Mưa}/X) = \frac{9}{63} * \frac{4}{8} = 0.071$$

$$P(\text{Mưa}/X) = \frac{1}{63} * \frac{4}{8} = 0.008$$

$$\text{Ta có: } P(\text{Không mưa}/X) + P(\text{Mưa}/X) = 1$$

$$P(\text{Không mưa}/X) = \frac{0.071}{0.008 + 0.071} = 0.898$$

$$P(\text{Mưa}/X) = 1 - 0.898 = 0.102$$

$$P(C_i/X) = \prod_{i=1}^n P(X/C_i) * P(C_i)$$

Khả năng không mưa chiếm tỷ lệ 89,8%

Chương 7. MÁY HỌC

Phân loại Naive Bayes (Naive Bayes classification)

- today = (Overcast, Cool, High, False), khả năng nào của việc chơi golf sẽ xảy ra?
- Bước 1: tính $P(C_i)$
- Bước 2: tính $P(X/C_i)$
- Bước 3: tính $P(C_i/X)$

	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	FALSE	No
1	Rainy	Hot	High	TRUE	No
2	Overcast	Hot	High	FALSE	Yes
3	Sunny	Mild	High	FALSE	Yes
4	Sunny	Cool	Normal	FALSE	Yes
5	Sunny	Cool	Normal	TRUE	No
6	Overcast	Cool	Normal	TRUE	Yes
7	Rainy	Mild	High	FALSE	No
8	Rainy	Cool	Normal	FALSE	Yes
9	Sunny	Mild	Normal	FALSE	Yes
10	Rainy	Mild	Normal	TRUE	Yes
11	Overcast	Mild	High	TRUE	Yes
12	Overcast	Hot	Normal	FALSE	Yes
13	Sunny	Mild	High	TRUE	No

Chương 7. MÁY HỌC

Phân cụm dữ liệu K trung tâm (K-means clustering)

- Thuật toán k-means là thuật toán phân cụm từ n đối tượng ban đầu vào k cụm phân biệt, với $k < n$ được giới thiệu năm 1957 bởi Lloyd K-means.
- Thuật toán sử dụng độ đo tương tự giữa quan sát.
- Khoảng cách Euclidean là phương pháp phổ biến nhất được dùng để đánh giá khoảng cách của các quan sát.
- Cho 2 quan sát $u = \{u_1, u_2, \dots, u_q\}$ và $v = \{v_1, v_2, \dots, v_q\}$, mỗi quan sát bao gồm q biến.
- $d_{u,v} = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \dots + (u_q - v_q)^2}$: khoảng cách giữa vectors

Chương 7. MÁY HỌC

Phân cụm dữ liệu K trung tâm (K-means clustering)

- Nếu 2 vector thẳng hàng thì khoảng cách u, v là:
- $d_{u,v} = |u_1 - v_1| + |u_2 - v_2| + \dots + |u_q - v_q|$
- Phân cụm cho 8 điểm có tọa độ như sau thành 3 cụm:
 $A_1(2,10), A_2(2,5), A_3(8,4), A_4(5,8), A_5(7,5), A_6(6,4), A_7(1,2), A_8(4,9)$
- Có thể nhận thấy số cụm là 3: $k=3 < n=8$
- Bước 1: chọn ngẫu nhiên 3 điểm làm giá trị trung tâm:
 $A_1(2,10), A_4(5,8), A_7(1,2)$
- Áp dụng công thức tính khoảng cách giữa 2 điểm $a(x_1, y_1)$ và $b(x_2, y_2)$:
 $d(a,b) = |x_2 - x_1| + |y_2 - y_1|$

Chương 7. MÁY HỌC

Phân cụm dữ liệu K trung tâm (K-means clustering)

- Khoảng cách của các điểm tới k-means
- Cluster: tìm MIN của mỗi Point trong mỗi dist. mean_i và gán vào cluster.
- Ví dụ: A1(2,10): $\min = 0 \in \text{dist.mean 1}$
→ cluster = 1
 - cluster1: (2,10)
 - cluster2: (8,4), (5,8), (7,5), (6,4), (4,9)
 - cluster3: (2,5), (1,2)

		(2,10)	(5,8)	(1,2)	
	Point	Dist. mean1	Dist. mean2	Dist. mean3	Cluster
A1	(2,10)	0	5	9	1
A2	(2,5)	5	6	4	3
A3	(8,4)	12	7	9	2
A4	(5,8)	5	0	10	2
A5	(7,5)	10	5	9	2
A6	(6,4)	10	5	7	2
A7	(1,2)	9	10	0	3
A8	(4,9)	3	2	10	2

Chương 7. MÁY HỌC

Phân cụm dữ liệu K trung tâm (K-means clustering)

- Cập nhật lại giá trị của các giá trị trung tâm của cluster.
- cluster1: có 1 điểm $A1(2,10)$ nên $\text{dist.means1} = (2,10)$
- cluster2: $\text{dist.means2} = [(8+5+7+6+4)/5, (4+8+5+4+9)/5] = (6,6)$
- cluster3: $\text{dist.mean3} = [(2+1)/2, (5+2)/2] = (1.5,3.5)$

Chương 7. MÁY HỌC

Phân cụm dữ liệu K trung tâm (K-means clustering)

- Tính lại giá trị khoảng cách các means:
 - cluster1: (2,10), (4,9)
 - cluster2: (8,4), (5,8), (7,5), (6,4)
 - cluster3: (2,5), (1,2)

		(2,10)	(5,5)	(1.5,3.5)	
	Point	Dist. mean1	Dist. mean2	Dist. mean3	Cluster
A1	(2,10)	0	8	7	1
A2	(2,5)	5	3	2	3
A3	(8,4)	12	4	7	2
A4	(5,8)	5	3	8	2
A5	(7,5)	10	2	7	2
A6	(6,4)	10	2	5	2
A7	(1,2)	9	7	2	3
A8	(4,9)	3	5	8	1

Chương 7. MÁY HỌC

Phân cụm dữ liệu K trung tâm (K-means clustering)

- Cập nhật lại giá trị của các giá trị trung tâm của cluster.
- cluster1: $\text{dist.mean1} = [(2+4)/2, (10+9)/2] = (3, 9.5)$
- cluster2: $\text{dist.mean2} = [(8+5+7+6)/4, (4+8+5+4)/4] = (6.5, 5.25)$
- cluster3: $\text{dist.mean3} = [(2+1)/2, (5+2)/2] = (1.5, 3.5)$

Chương 7. MÁY HỌC

Phân cụm dữ liệu K trung tâm (K-means clustering)

- Tính lại giá trị khoảng cách các means:
 - cluster1: (2,10), (4,9), (5,8)
 - cluster2: (8,4), (7,5), (6,4)
 - cluster3: (2,5), (1,2)

		(3,9.5)	(6.5,5.25)	(1.5,3.5)	
	Point	Dist. mean1	Dist. mean2	Dist. mean3	Cluster
A1	(2,10)	1,5	9,25	7	1
A2	(2,5)	5,5	4,75	2	3
A3	(8,4)	10,5	2,75	7	2
A4	(5,8)	3,5	4,25	8	1
A5	(7,5)	8,5	0,75	7	2
A6	(6,4)	8,5	1,75	5	2
A7	(1,2)	9,5	8,75	2	3
A8	(4,9)	1,5	6,25	8	1

Chương 7. MÁY HỌC

Phân cụm dữ liệu K trung tâm (K-means clustering)

- Cập nhật lại giá trị của các giá trị trung tâm của cluster.
- cluster1: $\text{dist.mean1} = [(2+4+5)/3, (10+9+8)/3] = (3.67, 9)$
- cluster2: $\text{dist.means2} = [(8+7+6)/4, (4+5+4)/4] = (7, 4.3)$
- cluster3: $\text{dist.mean3} = [(2+1)/2, (5+2)/2] = (1.5, 3.5)$

Chương 7. MÁY HỌC

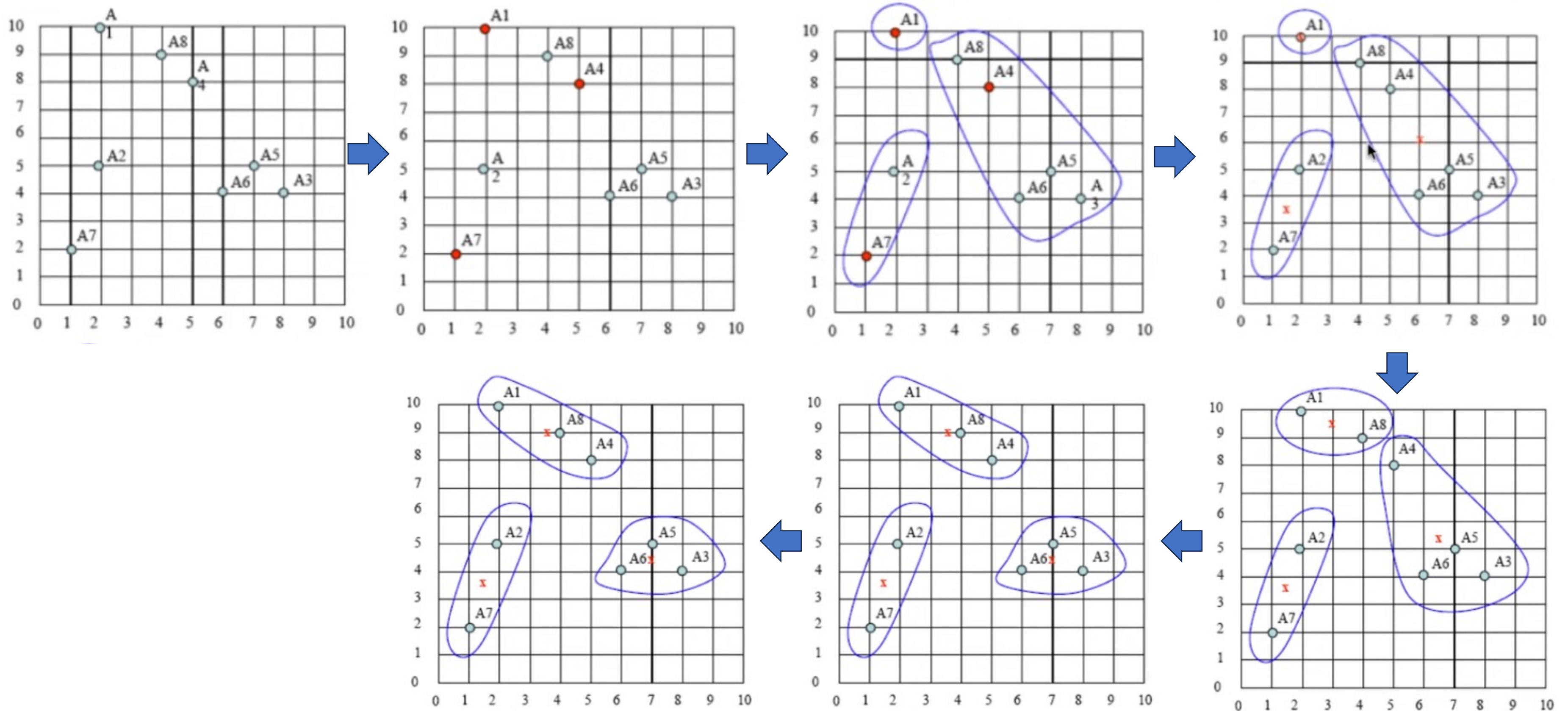
Phân cụm dữ liệu K trung tâm (K-means clustering)

- Tính lại giá trị khoảng cách các means:
 - cluster1: (2,10), (4,9), (5,8)
 - cluster2: (8,4), (7,5), (6,4)
 - cluster3: (2,5), (1,2)
- Nhận xét: số cụm không đổi so với lần lặp trước. TẠM DỪNG tính toán means

		(3.67,9)	(7,4.3)	(1.5,3.5)	
	Point	Dist. mean1	Dist. mean2	Dist. mean3	Cluster
A1	(2,10)	2,67	10,7	7	1
A2	(2,5)	5,67	5,7	2	3
A3	(8,4)	9,33	1,3	7	2
A4	(5,8)	2,33	5,7	8	1
A5	(7,5)	7,33	0,7	7	2
A6	(6,4)	7,33	1,3	5	2
A7	(1,2)	9,67	8,3	2	3
A8	(4,9)	0,33	7,7	8	1

Chương 7. MÁY HỌC

Phân cụm dữ liệu K trung tâm (K-means clustering)



Chương 7. MÁY HỌC

Phân cụm dữ liệu K trung tâm (K-means clustering)

- Phân cụm cho 6 điểm có tọa độ như sau thành 2 cụm:
 $A_1(1,1), A_2(2,1), A_3(2,3), A_4(3,2), A_5(4,3), A_6(5,5)$

HẾT CHƯƠNG 7