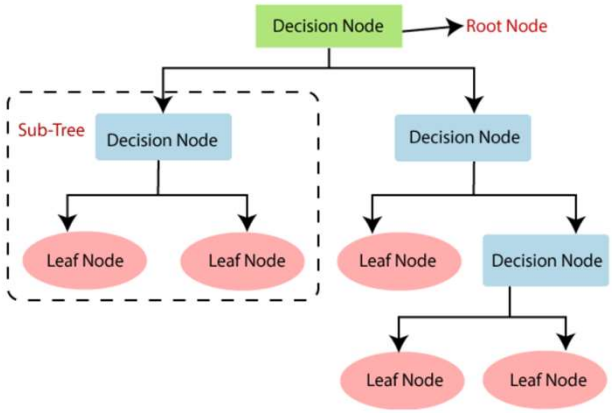
KHAI PHÁ DỮ LIỆU

Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giáo viên: TS.Nguyễn Khắc Cường

CHỦ ĐỀ 4

PHÂN LỚP
(Decision tree)

Decision tree



- Giới thiệu
  - Decision tree = Cây quyết định
  - Giải thuật phân lớp đơn giản, tốc độ thực hiện nhanh
  - Được ứng dụng trong nhiều lĩnh vực phân tích dữ liệu
- Idea:
  - Quá trình học:
    - Tách tập huấn luyện thành tập luật có dạng hình cây
    - Nút cha: chọn trong số các feature sao cho training data được tách "tốt nhất"
    - Nút con: chọn feature thỏa điều kiện tương tự nút cha
    - Nút lá: nhãn
  - Classification:
    - New data được phân loại theo đường dẫn từ nút gốc đến nút lá

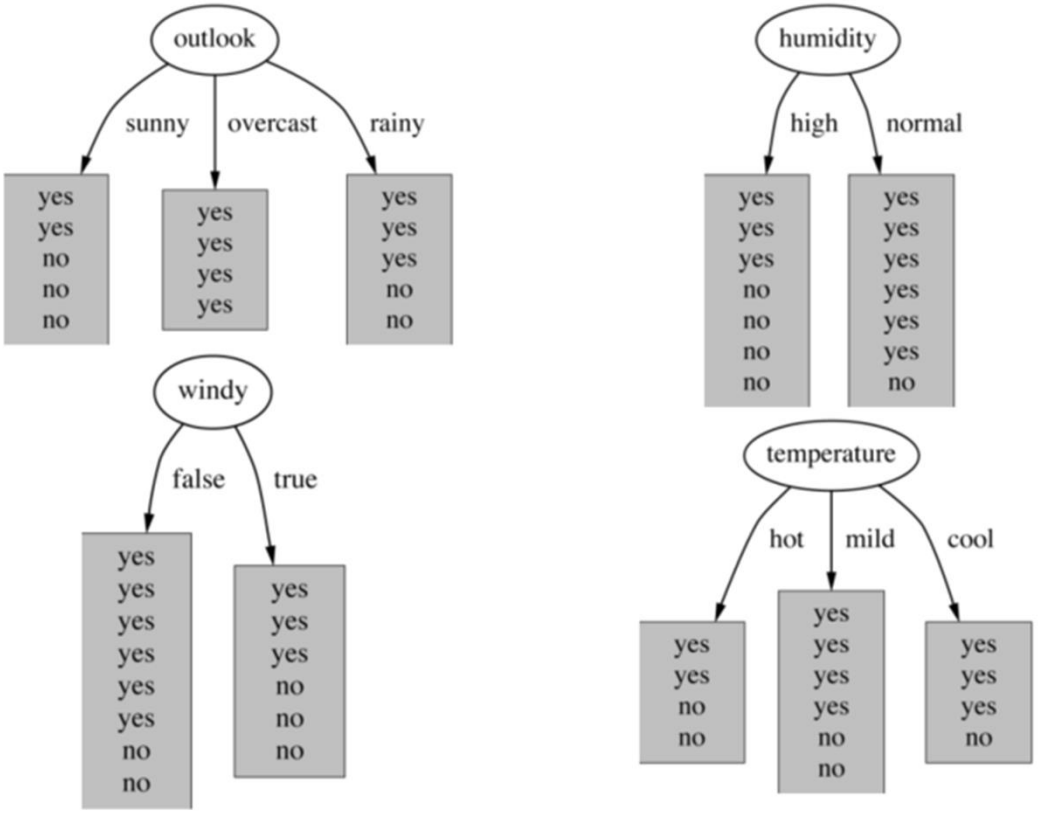
Decision tree

- Xét training data:
  - Features: Outlook, Temp, Humidity, Windy
  - Label: Play

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

Decision tree

- Các trường hợp có thể tách training data thành cây:



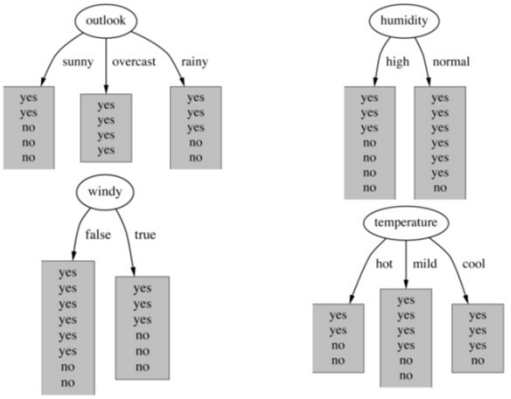
- Câu hỏi: chọn cây nào? = chọn nút gốc nào?

Decision tree

- Trả lời:
  - Chọn feature làm gốc nào để tạo ra cây
    - nhỏ nhất
    - nút gốc đó sinh ra các nút con là “purest”
- Giải pháp:
  - Cách đánh giá feature là “tốt nhất” để chọn làm nút gốc
    - Đánh giá information gain
      → chọn feature có information gain lớn nhất
    - Đánh giá chỉ số gini (xử lý dữ liệu số)
- Cách tính information gain:
  - Dùng thông tin của entropy

$$H(Y) = -\sum_{i=1}^{k} P(Y = y_i) \log_2 P(Y = y_i)$$

Decision tree



- Cách tính information gain:
  - Tính information gain của feature outlook

$$H(Y) = -\sum_{i=1}^{k} P(Y = y_i) \log_2 P(Y = y_i)$$

- "Outlook" = "Sunny":

$$\text{info}([2,3]) = \text{entropy}(2/5,3/5) = -2/5\log(2/5) - 3/5\log(3/5) = 0.971 \text{ bits}$$

- "Outlook" = "Overcast":

$$\text{info}([4,0]) = \text{entropy}(1,0) = -1\log(1) - 0\log(0) = 0 \text{ bits}$$

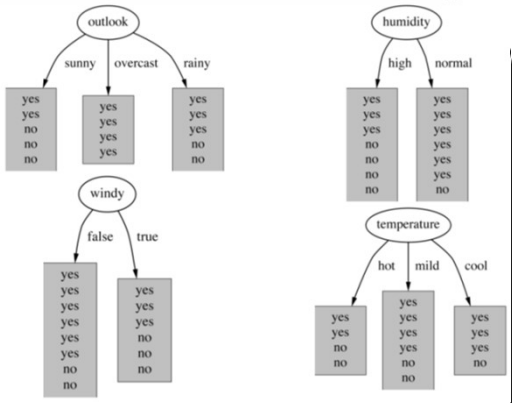
- "Outlook" = "Rainy":

$$\text{info}([3,2]) = \text{entropy}(3/5,2/5) = -3/5\log(3/5) - 2/5\log(2/5) = 0.971 \text{ bits}$$

- thông tin của thuộc tính outlook:

$$\text{info}([3,2],[4,0],[3,2]) = (5/14)\times 0.971 + (4/14)\times 0 + (5/14)\times 0.971$$
$$= 0.693 \text{ bits}$$

$$H(Y) = -\sum_{i=1}^{k} P(Y = y_i) \log_2 P(Y = y_i)$$



Decision tree

- Cách tính information gain:
  - Tính information gain (IG) của feature Outlook
    - "Outlook" = "Sunny":

$$\text{info}([2,3]) = \text{entropy}(2/5,3/5) = -2/5\log(2/5) - 3/5\log(3/5) = 0.971 \text{ bits}$$

    - "Outlook" = "Overcast":

$$\text{info}([4,0]) = \text{entropy}(1,0) = -1\log(1) - 0\log(0) = 0 \text{ bits}$$

    - "Outlook" = "Rainy":

$$\text{info}([3,2]) = \text{entropy}(3/5,2/5) = -3/5\log(3/5) - 2/5\log(2/5) = 0.971 \text{ bits}$$

    - thông tin của thuộc tính outlook:

$$\text{info}([3,2],[4,0],[3,2]) = (5/14)\times 0.971 + (4/14)\times 0 + (5/14)\times 0.971$$
$$= 0.693 \text{ bits}$$

  - IG(Outlook) = IG(trước khi tách) – IG(sau khi tách)

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

Decision tree

- Cách tính information gain:
  - Tính information gain (IG) của feature Humidity
    - "Humidity" = "High":

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

    - "Humidity" = "Normal":

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

    - thông tin của thuộc tính humidity

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

    - độ lợi thông tin của thuộc tính humidity

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

Decision tree

- Cách tính information gain:
  - Các feature còn lại tính tương tự, thu được

$$\text{gain}(\text{"Outlook"}) = 0.247 \text{ bits}$$

$$\text{gain}(\text{"Temperature"}) = 0.029 \text{ bits}$$

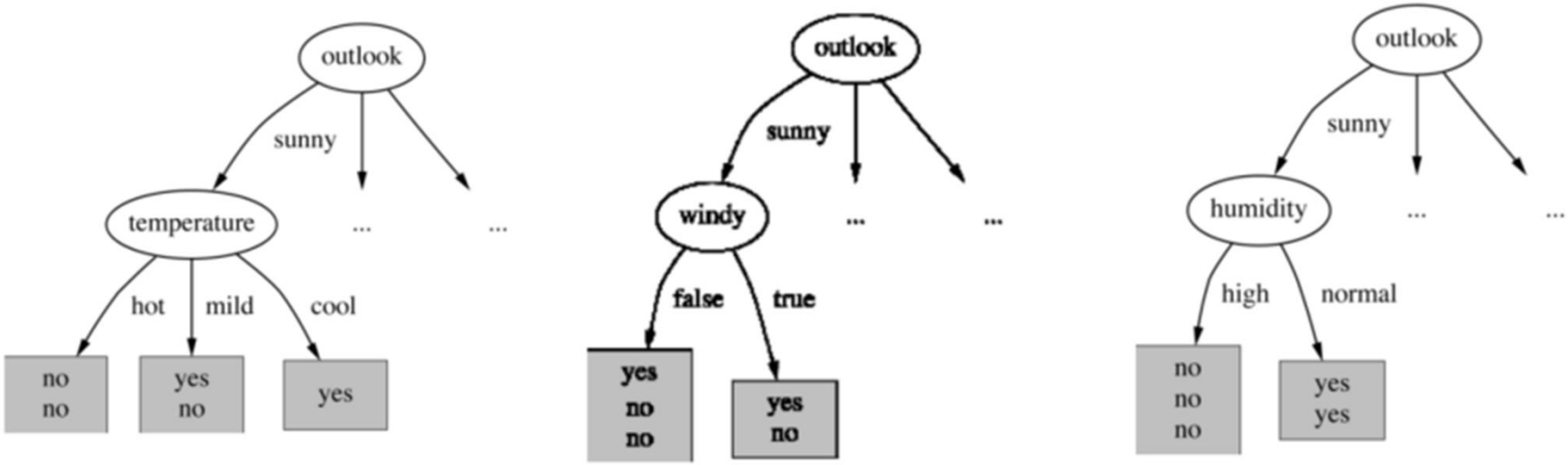
$$\text{gain}(\text{"Humidity"}) = 0.152 \text{ bits}$$

$$\text{gain}(\text{"Windy"}) = 0.048 \text{ bits}$$

→ Chọn feature Outlook làm nút gốc

Decision tree

- Tạo thành các cây có nút gốc là Outlook



- Tương tự, tính gain của các feature để chọn nút con (nút gốc tiếp theo)

Decision tree

- Tương tự, tính gain của các feature để chọn nút con (nút gốc tiếp theo)

$$\text{gain}(\text{"Windy"}) = 0.020 \text{ bits}$$

$$\text{gain}(\text{"Temperature"}) = 0.571 \text{ bits}$$

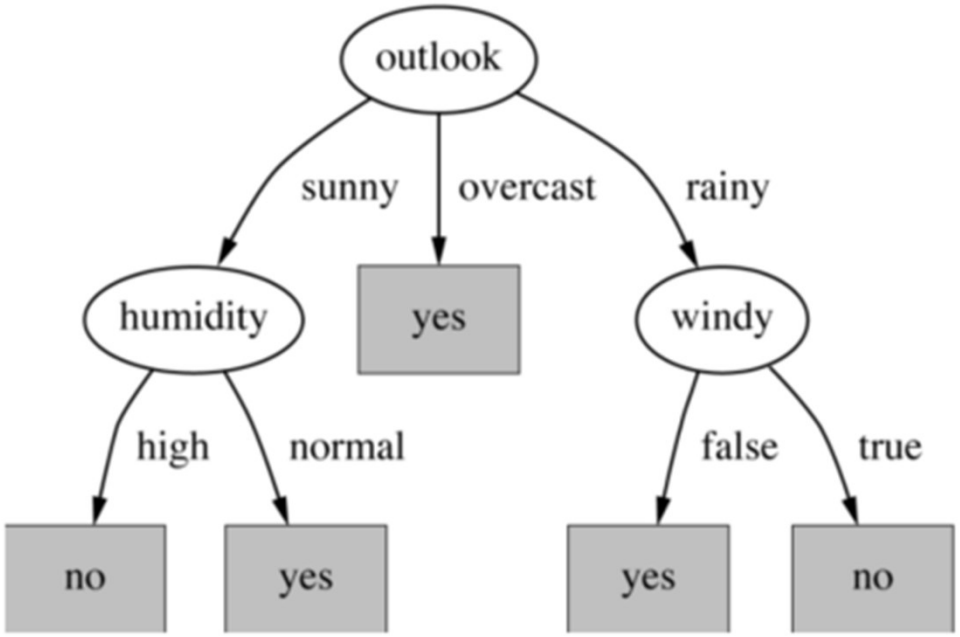
$$\text{gain}(\text{"Humidity"}) = 0.971 \text{ bits}$$

- Thao tác tách có thể dừng khi feature không thể tách được nữa
- Nhãn của nút con ở layer thấp nhất được gán cho lớp lớn nhất chứa trong lớp con đó

Decision tree

- Kết quả:

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



Q / A