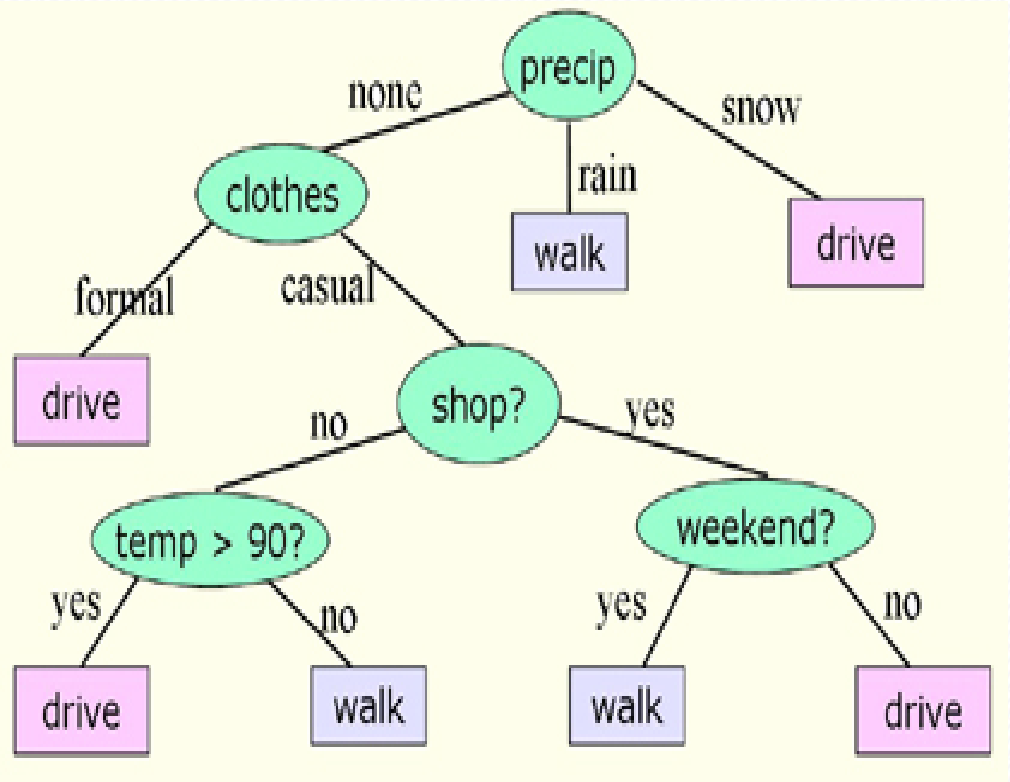
Cây quyết định (ID3) và Học Quy nạp (ILA)

Tô Hoài Việt
Khoa Công nghệ Thông tin
Đại học Khoa học Tự nhiên TPHCM
thviet@fit.hcmuns.edu.vn

Nội dung

- Cây quyết định
- Học cây quyết định – Thuật toán ID3
- Biểu diễn tri thức bằng luật
- Rút luật từ cây quyết định
- Thuật toán học quy nạp

Cây quyết định



Cây quyết định biểu diễn:

- Mỗi nút trong kiểm tra một thuộc tính
- Mỗi nhánh tương ứng với giá trị thuộc tính
- Mỗi nút lá được gán một phân lớp

Định luật Occam: những cây đơn giản là những cây quyết định tốt hơn

Thuật toán học ID3

Được phát triển đồng thời bởi Quinlan trong AI và Breiman, Friedman, Olsen và Stone trong thống kê

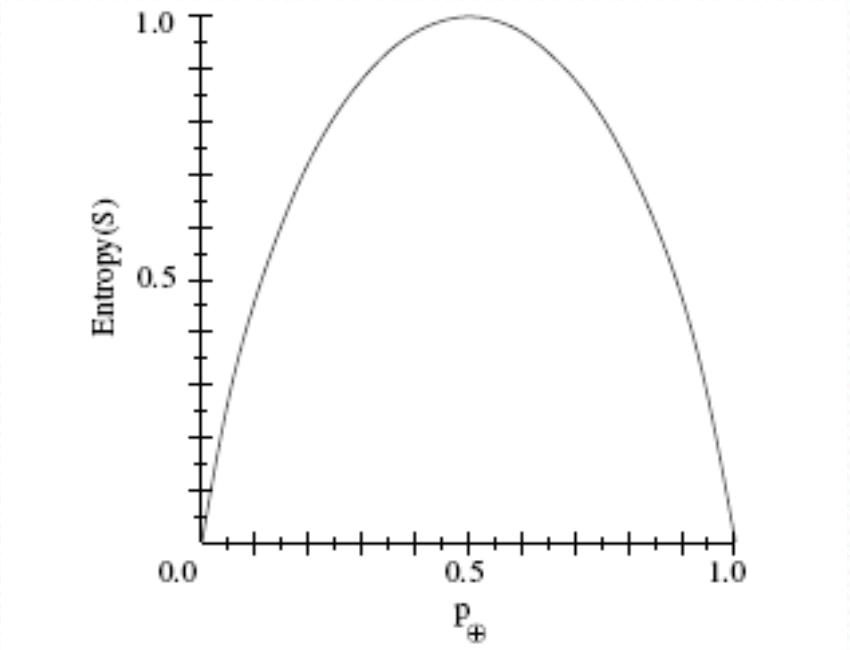
Lặp:

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.

Thuộc tính tốt nhất là gì?

Entropy

- S là tập các mẫu huấn luyện
- p là tỷ lệ các mẫu dương trong S
- $H \equiv - p.\log_2 p - (1 - p).\log_2(1 - p)$



Thuật toán học ID3

Được phát triển đồng thời bởi Quinlan trong AI và Breiman, Friedman, Olsen và Stone trong thống kê

Lặp:

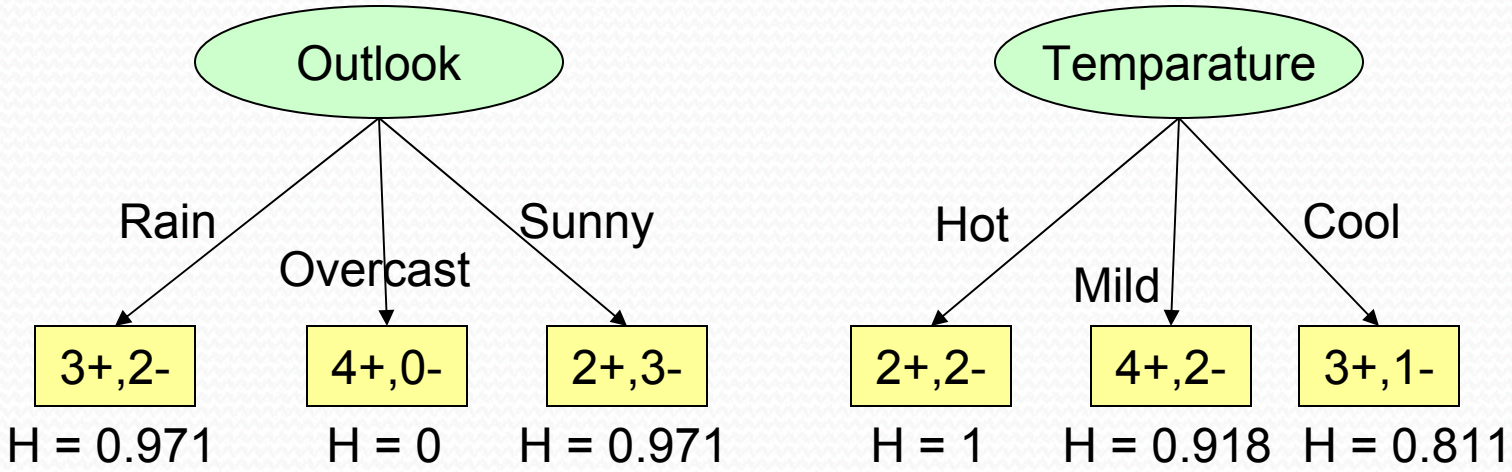
1. Chọn A ← thuộc tính quyết định "tốt nhất" cho *nút* kế tiếp
2. Gán A là thuộc tính quyết định cho *nút*
3. Với mỗi giá trị của A, tạo nhánh con mới của *nút*
4. Phân loại các mẫu huấn luyện cho các nút lá
5. Nếu các mẫu huấn luyện được phân loại hoàn toàn thì NGƯNG, Ngược lại, lặp với các nút lá mới.

Thuộc tính tốt nhất sẽ làm tối thiểu hoá entropy trung bình của dữ liệu trong các nút con

Ví dụ Huấn luyện

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Ví dụ (tt)



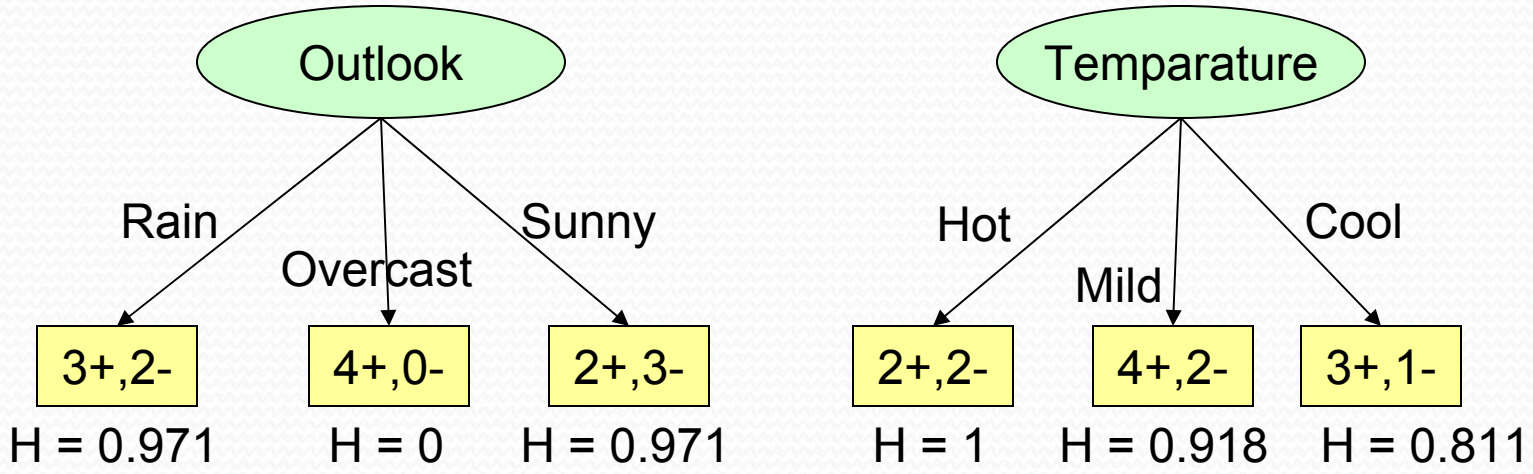
$H_{rain} = -3/5.\log_2 3/5 - 2/5.\log_2 2/5 = 0.442 + 0.529 = 0.971$

$H_{overcast} = -4/4.\log_2 4/4 - 0/4.\log_2 0/4 = 0 + 0 = 0$

$H_{sunny} = -2/5.\log_2 2/5 - 3/5.\log_2 3/5 = 0.529 + 0.442 = 0.971$

$$AE(ĐHLTB) = \sum_{v \in Value(A)} p_v H_{Av}$$

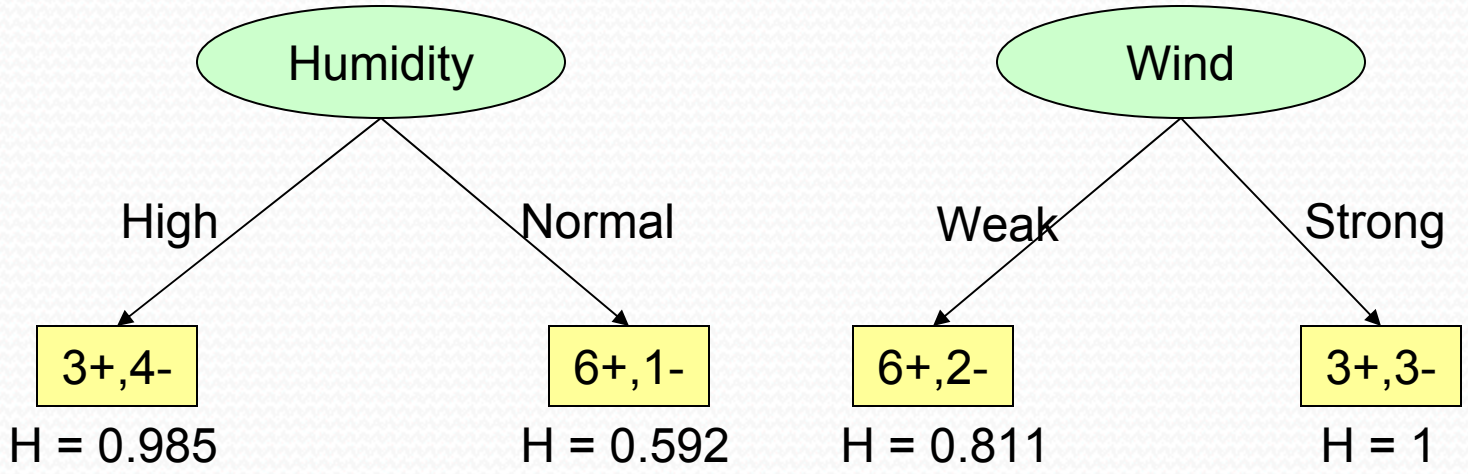
Ví dụ (tt)



AE = 5/14*.971 + 4/14*0 + 5/14*.971
= 0.694

AE = 4/14*1 + 6/14*.918 + 4/14*.811
= 0.911

Ví dụ (tt)

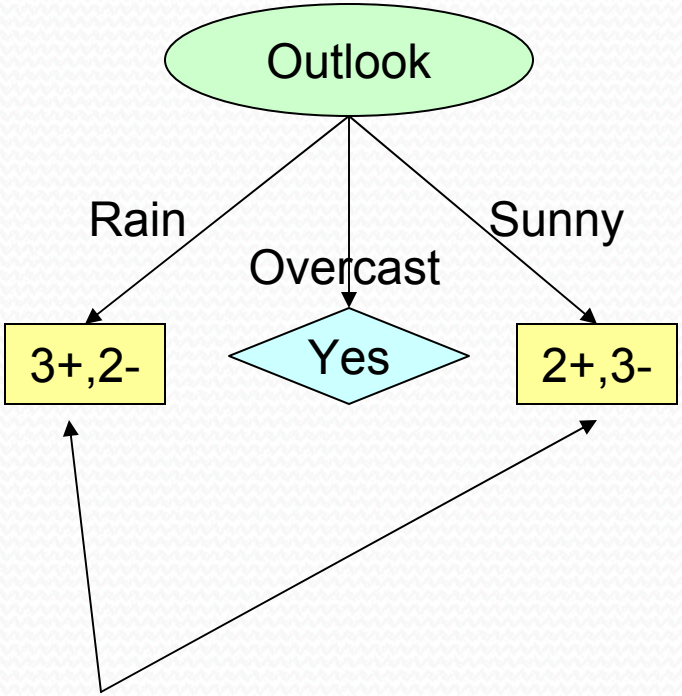


AE = 7/14*.985 + 7/14*.592
= 0.788

AE = 8/14*.811 + 6/14*1
= 0.892

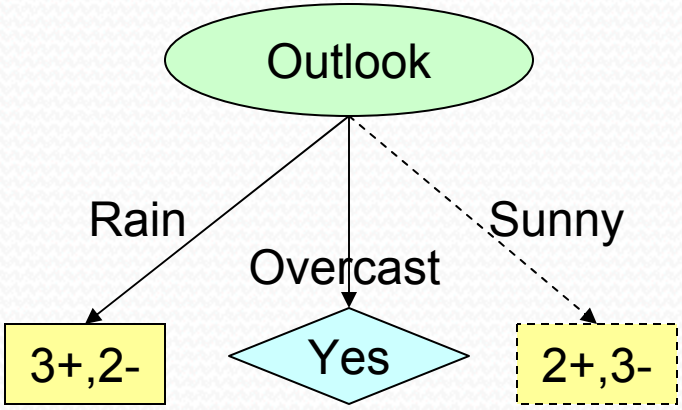
Chọn *Outlook* là thuộc tính quyết định

Ví dụ (tt)



Chọn thuộc tính gì tiếp theo?
Tiếp tục thực hiện việc phân chia

Ví dụ (tt)

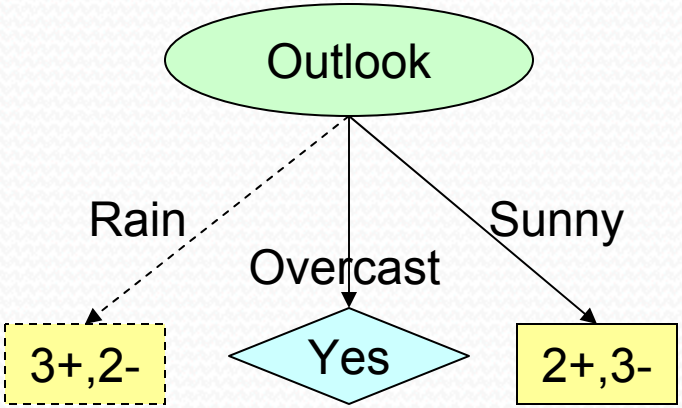


AE (Rain, Temperature) = 2/5*1 + 3/5*.918 = 0.951

AE (Rain, Humidity) = 2/5*1 + 3/5*.918 = 0.951

AE (Rain, Wind) = 2/5*0 + 3/5*0 = 0

Ví dụ (tt)

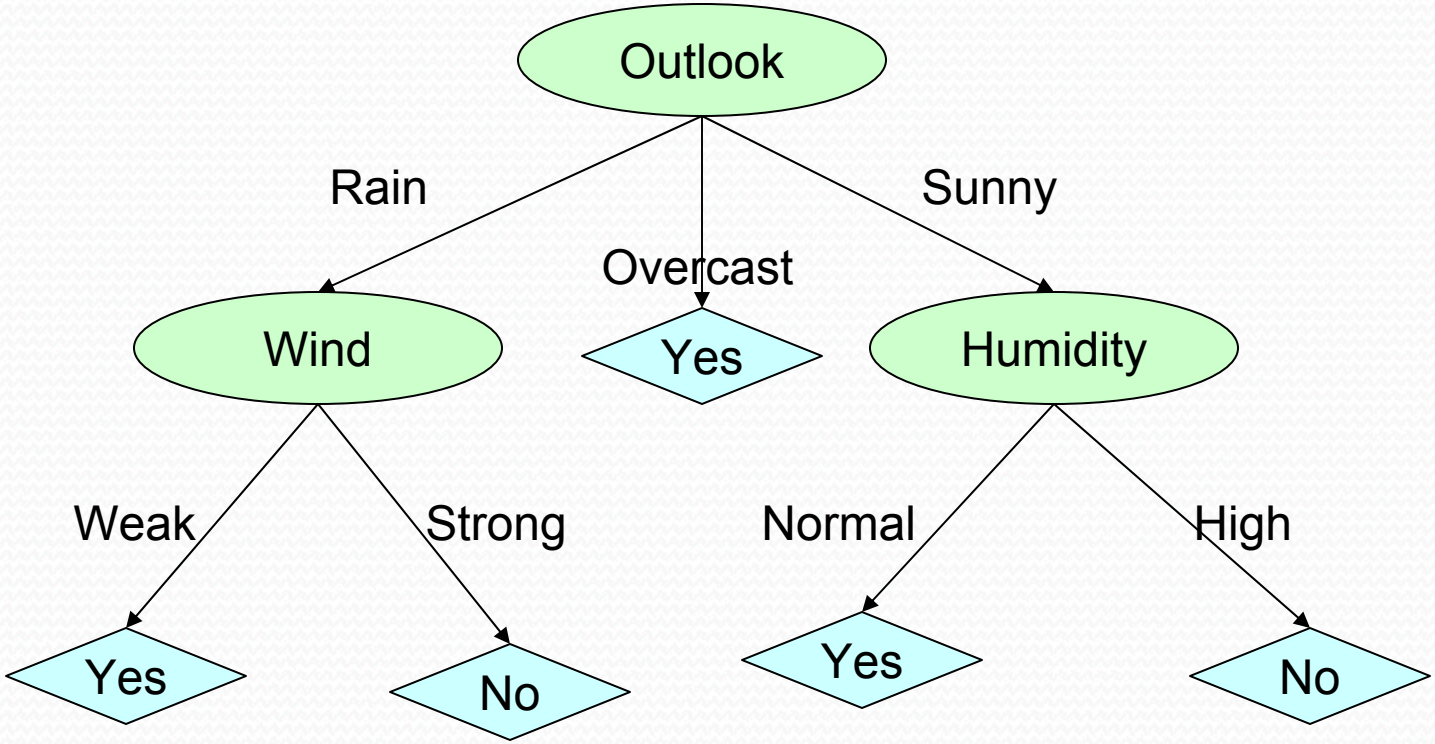


AE (Sunny, Temperature) = 2/5*0 + 2/5*1 + 1/5*0= 0.4

AE (Sunny, Humidity) = 2/5*0 + 3/5*0 = 0

AE (Sunny, Wind) = 2/5*1 + 3/5*.918 = 0.951

Ví dụ (tt)


Outlook
Rain
Overcast
Sunny
Wind
Yes
Humidity
Weak
Strong
Normal
High
Yes
No
Yes
No

Tri thức dạng luật

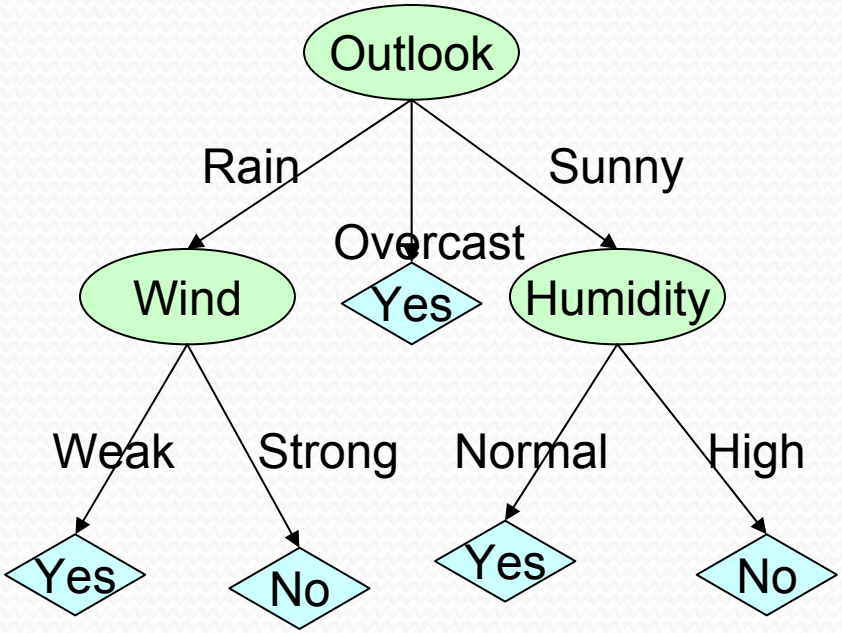
- Tri thức được biểu diễn dưới dạng luật:

  IF Điều kiện 1 ^ Điều kiện 2… THEN Kết luận

- Dễ hiểu với con người, được sử dụng chủ yếu trong các hệ chuyên gia

- Rút luật từ cây quyết định: đi từ nút gốc đến nút lá, lấy các phép thử làm tiền đề và phân loại của nút lá làm kết quả

Rút luật từ cây quyết định



- IF Outlook = Overcast THEN Yes
- IF Outlook = Rain AND Wind=Weak THEN Yes
- IF Outlook = Rain AND Wind=Strong THEN No
- IF Outlook = Sunny AND Humidity=Normal THEN Yes
- IF Outlook = Sunny AND Humidity=High THEN No

Thuật giải Học Quy nạp (ILA)

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

1. Chia tập mẫu thành các tập con ứng với thuộc tính quyết định
2. Với mỗi bảng con
3. Với mỗi tổ hợp thuộc tính có thể bắt (bắt đầu với số lượng = 1)
4. Tìm các giá trị chỉ xuất hiện ở bảng con này mà không xuất hiện ở các bảng con khác
5. (Nếu có nhiều tổ hợp thì chọn tổ hợp có số lượng mẫu tin nhiều nhất)
6. Sử dụng tổ hợp thuộc tính, giá trị vừa tìm được để tạo luật
7. Đánh dấu các dòng đã xét
8. Nếu còn dòng chưa xét, lặp lại bước 3
9. Lặp lại bước 2 với các bảng con

Ví dụ ILA

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
1	Vừa	Xanh dương	Hộp	Mua
2	Nhỏ	Đỏ	Nón	Không mua
3	Nhỏ	Đỏ	Cầu	Mua
4	Lớn	Đỏ	Nón	Không mua
5	Lớn	Xanh lá	Trụ	Mua
6	Lớn	Đỏ	Trụ	Không mua
7	Lớn	Xanh lá	Cầu	Mua

Ví dụ ILA (tt)

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
1	Vừa	Xanh dương	Hộp	Mua
3	Nhỏ	Đỏ	Cầu	Mua
5	Lớn	Xanh lá	Trụ	Mua
7	Lớn	Xanh lá	Cầu	Mua

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
2	Nhỏ	Đỏ	Nón	Không mua
4	Lớn	Đỏ	Nón	Không mua
6	Lớn	Đỏ	Trụ	Không mua

Ví dụ ILA (tt)

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
1	Vừa	Xanh dương	Hộp	Mua
3	Nhỏ	Đỏ	Cầu	Mua
5	Lớn	Xanh lá	Trụ	Mua
7	Lớn	Xanh lá	Cầu	Mua

STT	Kích cỡ			Quyết định
2	Nhỏ			Không mua
4	Lớn			Không mua
6	Lớn		Trụ	Không mua

Chọn thuộc tính Màu sắc
với giá trị Xanh lá

Ví dụ ILA (tt)

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
1	Vừa	Xanh dương	Hộp	Mua
3	Nhỏ	Đỏ	Cầu	Mua

IF Màu sắc = Xanh lá THEN Quyết định = Mua

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
2	Nhỏ	Đỏ	Nón	Không mua
4	Lớn	Đỏ	Nón	Không mua
6	Lớn	Đỏ	Trụ	Không mua

Ví dụ ILA (tt)

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
3	Nhỏ	Đỏ	Cầu	Mua

IF Màu sắc = Xanh lá THEN Quyết định = Mua

IF Kích cỡ = Vừa THEN Quyết định = Mua

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
2	Nhỏ	Đỏ	Nón	Không mua
4	Lớn	Đỏ	Nón	Không mua
6	Lớn	Đỏ	Trụ	Không mua

Ví dụ ILA (tt)

IF Màu sắc = Xanh lá THEN Quyết định = Mua

IF Kích cỡ = Vừa THEN Quyết định = Mua

IF Hình dáng= Cầu THEN Quyết định = Mua

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
2	Nhỏ	Đỏ	Nón	Không mua
4	Lớn	Đỏ	Nón	Không mua
6	Lớn	Đỏ	Trụ	Không mua

Ví dụ ILA (tt)

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
1	Vừa	Xanh dương	Hộp	Mua
3	Nhỏ	Đỏ	Cầu	Mua
5	Lớn	Xanh lá	Trụ	Mua
7	Lớn	Xanh lá	Cầu	Mua

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
2	Nhỏ	Đỏ	Nón	Không mua
4	Lớn	Đỏ	Nón	Không mua
6	Lớn	Đỏ	Trụ	Không mua

IF Hình dáng = Nón THEN Quyết định = Không mua

Ví dụ ILA (tt)

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
1	Vừa	Xanh dương	Hộp	Mua
3	Nhỏ	Đỏ	Cầu	Mua
5	Lớn	Xanh lá	Trụ	Mua
7	Lớn	Xanh lá	Cầu	Mua

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
6	Lớn	Đỏ	Trụ	Không mua

IF Hình dáng = Nón THEN Quyết định = Không mua

Ví dụ ILA (tt)

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
1	Vừa	Xanh dương	Hộp	Mua
3	Nhỏ	Đỏ	Cầu	Mua
5	Lớn	Xanh lá	Trụ	Mua
7	Lớn	Xanh lá	Cầu	Mua

STT	Kích cỡ	Màu sắc	Hình dáng	Quyết định
6	Lớn	Đỏ	Trụ	Không mua

IF Hình dáng = Nón THEN Quyết định = Không mua

IF Kích cỡ = Lớn AND Màu sắc = Đỏ THEN Quyết định = Không mua

Điều cần nắm

- Nắm được khái niệm cây quyết định
- Hiểu và vận dụng thuật toán ID3
- Hiểu và vận dụng thuật toán học quy nạp