

BÀI TẬP VỀ CÂY QUYẾT ĐỊNH

(1) Cho cơ sở dữ liệu sau:

Mã u	Thời tiết	Lá cây	Nhiệt độ	Quyết định (Mùa)
1T	Nắng	Vàng	Trung bình	Thu
2Đ	Tuyết	Xanh	Thấp	Đông
3Đ	Tuyết	Vàng	Thấp	Đông
4T	Mưa	Vàng	Trung bình	Thu
5Đ	Mưa	Rụng	Thấp	Đông
6Đ	Tuyết	Rụng	Thấp	Đông
7Đ	Nắng	Rụng	Thấp	Đông
8X	Nắng	Xanh	Trung bình	Xuân
9H	Nắng	Xanh	Cao	Hè
10T	Mưa	Rụng	Trung bình	Thu
11H	Mưa	Xanh	Cao	Hè
12X	Mưa	Xanh	Trung bình	Xuân
X	Mưa	Vàng	Thấp	?
Y	Tuyết	Rụng	Trung bình	?

- a) Từ mẫu 1 đến mẫu 12, hãy sử dụng thuật toán Quinlan hoặc thuật toán ID3 để tìm ra bộ luật cây quyết định.
- b) Áp dụng bộ luật vừa tìm được, cho biết kết quả của các mẫu X và Y.
- c) Áp dụng tương tự, tìm cây quyết định dùng Gini Index, và Classification Error

Hướng dẫn: Các bạn chỉnh sửa lại bài giải để quyết định tách nhánh theo độ đo **Information Gain**

+ **Xác định bài toán:** $X = (\text{Thời tiết}, \text{Lá cây}, \text{Nhiệt độ}) \rightarrow Y (\text{Quyết Định}) = \text{Thu} / \text{Đông} / \text{Xuân} / \text{Hè}$

+ **Tính lần 1,** xét lần lượt các thuộc tính Thời tiết, Lá cây, Nhiệt độ:

* Thuộc tính **Thời tiết:**

Nắng: 1T, 7D, 8X, 9H (Thu, Đông, Xuân, Hè) = (1/4, 1/4, 1/4, 1/4)

Tuyết: 2D, 3D, 6D (0/3, 3/3, 0/3, 0/3)

Mưa: 4T, 5D, 10T, 11H, 12X (2/5, 1/5, 1/5, 1/5)

$$\text{Entropy}(\text{Thời tiết}) = 4/12 * (-1/4 * \log_2(1,4) - 1/4 * \log_2(1,4) - 1/4 * \log_2(1,4) - 1/4 * \log_2(1,4)) + 3/12 * (-3/3 * \log_2(3,3)) + 5/12 * (-2/5 * \log_2(2,5) - 1/5 * \log_2(1,5) - 1/5 * \log_2(1,5) - 1/5 * \log_2(1,5)) = 1.4675$$

* Thuộc tính **Lá Cây:**

Vàng: 1T, 3D, 4T (2/3, 1/3, 0/3, 0/3)

Xanh: 2D, 8X, 9H, 11H, 12X (0/5, 1/5, 2/5, 2/5)

Rụng: 5D, 6D, 7D, 10T (1/4, 3/4, 0/4, 0/4)

BÀI TẬP VỀ CÂY QUYẾT ĐỊNH

$$\text{Entropy}(\text{Lá Cây}) = 3/12 * (-2/3 * \log_2(2,3) - 1/3 * \log_2(1,3)) + 5/12 * (-1/5 * \log_2(1,5) - 2/5 * \log_2(2,5)) - 2/5 * \log_2(2,5)) + 4/12 * (-1/4 * \log_2(1,4) - 3/4 * \log_2(3,4)) = 1.1338$$

* Thuộc tính **Nhiệt Độ**:

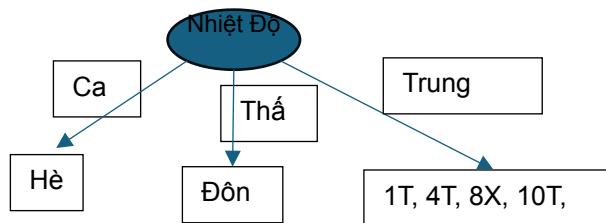
Trung Bình: 1T, 4T, 8X, 10T, 12X (3/5, 0/5, 2/5, 0/5)

Thấp: 2D, 3D, 5D, 6D, 7D (0/5, 5/5, 0/5, 0/5)

Cao: 9H, 11H (0/2, 0/2, 0/2, 2/2)

$$\text{Entropy}(\text{Nhiệt Độ}) = 5/12 * (-3/5 * \log_2(3,5) - 2/5 * \log_2(2,5)) + 5/12 * (-5/5 * \log_2(5,5)) + 2/12 * (-2/2 * \log_2(2,2)) = 0.4046$$

→ Chọn thuộc tính **Nhiệt Độ** có độ hỗn loạn trung bình thấp nhất là **0.4046** để phân lớp.



+ **Tính lần 3 với nhánh Nhiệt Độ = Trung Bình**, xét lần lượt các thuộc tính Thời tiết, Lá cây:

* Thuộc tính **Thời tiết**:

Nắng: 1T, 8X (Thu, Đôn, Xuân, Hè) = (1/2, 0/2, 1/2, 0/2)

Tuyết: x (x)

Mưa: 4T, 10T, 12X (2/3, 0/3, 1/3, 0/3)

$$\text{Entropy}(\text{Thời tiết}) = 2/5 * (-1/2 * \log_2(1,2) - 1/2 * \log_2(1,2)) + 3/5 * (-2/3 * \log_2(2,3) - 1/3 * \log_2(1,3)) = 0.9508$$

* Thuộc tính **Lá Cây**:

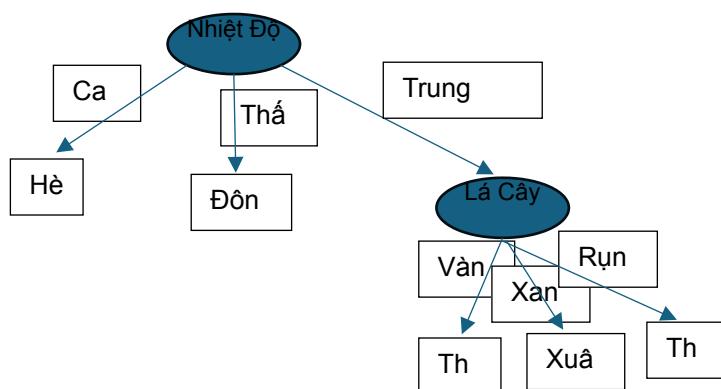
Vàng: 1T, 4T (2/2, 0/2, 0/2, 0/2)

Xanh: 8X, 12X (0/2, 0/2, 2/2, 0/2)

Rụng: 10T (1/1, 0/1, 0/1, 0/1)

$$\text{Entropy}(\text{Lá Cây}) = 2/5 * (-2/2 * \log_2(2,2)) + 2/5 * (-2/2 * \log_2(2,2)) + 1/5 * (-1/1 * \log_2(1,1)) = 0.0$$

→ Chọn thuộc tính **Lá Cây** có độ hỗn loạn trung bình thấp nhất là **0.0** để phân lớp.



Các bộ luật cây quyết định:

Luật 1) Nếu Nhiệt Độ = Cao thì Quyết định = Hè

Luật 2) Nếu Nhiệt Độ = Thấp thì Quyết định = Đông

Luật 3) Nếu Nhiệt Độ = Trung Bình và Lá Cây = Xanh thì Quyết định = Xuân

Luật 4) Nếu Nhiệt Độ = Trung Bình và Lá Cây = Vàng hay Rụng thì Quyết định = Thu

Câu b)

BÀI TẬP VỀ CÂY QUYẾT ĐỊNH

X = (Mưa, Vàng, Thấp) → Quyết định = Đóng (áp dụng luật 2)

Y = (Tuyết, Rụng, Trung bình) → Quyết định = Thu (áp dụng luật 4)

---o0o---

(Hết)