Bài tập Cây Quyết Định

Trần Chí Vỹ

Ngày 30 tháng 10 năm 2025

Bài 1: Phân loại lá cây

a) Xây dựng cây dùng Gini Index

Chúng ta chọn thuộc tính có Gini Index trung bình (có điều kiện) thấp nhất.

Lần 1: Chọn nút gốc

$Gini(\text{Gốc}) = 1 - (6/10)^2 - (4/10)^2 = 0.48$

Gini(Hình dạng):

- Tròn (6 mẫu: 3 Có, 3 Không): $Gini = 1 - (3/6)^2 - (3/6)^2 = 0.5$
- Dài (4 mẫu: 3 Có, 1 Không): $Gini = 1 - (3/4)^2 - (1/4)^2 = 0.375$
- $Gini_{\text{TB}}(\text{Hình dạng}) = (6/10) \times 0.5 + (4/10) \times 0.375 = \mathbf{0.45}$

Gini(Dạng):

- Đơn (6 mẫu: 3 Có, 3 Không): $Gini = 0.5$
- Kép (4 mẫu: 3 Có, 1 Không): $Gini = 0.375$
- $Gini_{\text{TB}}(\text{Dạng}) = (6/10) \times 0.5 + (4/10) \times 0.375 = \mathbf{0.45}$

Gini(Màu):

- Đỏ (4 mẫu: 3 Có, 1 Không): $Gini = 0.375$
- Xanh (3 mẫu: 1 Có, 2 Không): $Gini = 1 - (1/3)^2 - (2/3)^2 \approx 0.444$
- Vàng (3 mẫu: 2 Có, 1 Không): $Gini = 1 - (2/3)^2 - (1/3)^2 \approx 0.444$
- $Gini_{\text{TB}}(\text{Màu}) = (4/10) \times 0.375 + (3/10) \times 0.444 + (3/10) \times 0.444 \approx \mathbf{0.417}$

Gini(Kích thước):

- Nhỏ (4 mẫu: 2 Có, 2 Không): $Gini = 0.5$
- To (2 mẫu: 0 Có, 2 Không): $Gini = 0$
- Vừa (4 mẫu: 4 Có, 0 Không): $Gini = 0$
- $Gini_{\text{TB}}(\text{Kích thước}) = (4/10) \times 0.5 + (2/10) \times 0 + (4/10) \times 0 = \mathbf{0.2}$

Kết luận Lần 1: Thuộc tính "Kích thước"có Gini Index thấp nhất (0.2), được chọn làm nút gốc.

- Nhánh Kích thước = To: (0 Có, 2 Không) → Nút lá, Quyết định = Không.
- Nhánh Kích thước = Vừa: (4 Có, 0 Không) → Nút lá, Quyết định = Có.
- Nhánh Kích thước = Nhỏ: (2 Có, 2 Không) → Cần phân nhánh tiếp.

Lần 2: Phân nhánh từ "Kích thước = Nhỏ"

Xét 4 mẫu con: {1, 2, 4, 5} (2 Có, 2 Không).

Gini(Dạng):

- Đơn (2 mẫu: 1K, 4K → 0 Có, 2 Không): $Gini = 0$
- Kép (2 mẫu: 2C, 5C → 2 Có, 0 Không): $Gini = 0$
- $Gini_{\text{TB}}(\text{Dạng}) = (2/4) \times 0 + (2/4) \times 0 = \mathbf{0.0}$

Gini(Hình dạng):

- Tròn (2 mẫu: 1K, 2C → 1 Có, 1 Không): $Gini = 0.5$
- Dài (2 mẫu: 4K, 5C → 1 Có, 1 Không): $Gini = 0.5$
- $Gini_{\text{TB}}(\text{Hình dạng}) = (2/4) \times 0.5 + (2/4) \times 0.5 = 0.5$

Gini(Màu):

- Đỏ (2 mẫu: 1K, 2C → 1 Có, 1 Không): $Gini = 0.5$
- Xanh (1 mẫu: 4K → 0 Có, 1 Không): $Gini = 0$
- Vàng (1 mẫu: 5C → 1 Có, 0 Không): $Gini = 0$
- $Gini_{\text{TB}}(\text{Màu}) = (2/4) \times 0.5 + (1/4) \times 0 + (1/4) \times 0 = 0.25$

Kết luận Lần 2: Thuộc tính "Dạng"có Gini Index thấp nhất (0.0), được chọn làm nút con.

- Nhánh Dạng = Đơn: (0 Có, 2 Không) → Nút lá, Quyết định = Không.
- Nhánh Dạng = Kép: (2 Có, 0 Không) → Nút lá, Quyết định = Có.

Bộ luật cây quyết định (từ Gini Index):

- **Luật 1:** Nếu Kích thước = To **THÌ** Độc = Không
- **Luật 2:** Nếu Kích thước = Vừa **THÌ** Độc = Có
- **Luật 3:** Nếu Kích thước = Nhỏ **VÀ** Dạng = Đơn **THÌ** Độc = Không
- **Luật 4:** Nếu Kích thước = Nhỏ **VÀ** Dạng = Kép **THÌ** Độc = Có

b) Áp dụng bộ luật

- **Mẫu 11: (Dài, Kép, Đỏ, Vừa)**
  - Kiểm tra: Kích thước = Vừa.
  - Áp dụng Luật 2.
  - **Kết quả: Độc = Có.**
- **Mẫu 12: (Tròn, Đơn, Xanh, Nhỏ)**
  - Kiểm tra: Kích thước = Nhỏ.
  - Kiểm tra: Dạng = Đơn.
  - Áp dụng Luật 3.
  - **Kết quả: Độc = Không.**

c) Tính theo Entropy và Classification Error

1. Entropy (ID3 - Information Gain)

Lần 1 (Chọn gốc):

- $Entropy_{\text{TB}}$(Hình dạng) ≈ 0.924
- $Entropy_{\text{TB}}$(Dạng) ≈ 0.924
- $Entropy_{\text{TB}}$(Màu) ≈ 0.875
- $Entropy_{\text{TB}}$(Kích thước) = $\mathbf{0.4}$
- $\rightarrow$ Chọn "Kích thước"làm nút gốc.

Lần 2 (Nhánh "Nhỏ"):

- $Entropy_{\text{TB}}$(Hình dạng) = 1.0
- $Entropy_{\text{TB}}$(Dạng) = $\mathbf{0.0}$
- $Entropy_{\text{TB}}$(Màu) = 0.5
- $\rightarrow$ Chọn "Dạng"làm nút con.

2. Classification Error (Lỗi phân loại)

Lần 1 (Chọn gốc):

- $Error_{\text{TB}}(\text{Hình dạng}) = 0.4$
- $Error_{\text{TB}}(\text{Dạng}) = 0.4$
- $Error_{\text{TB}}(\text{Màu}) \approx 0.3$
- $Error_{\text{TB}}(\text{Kích thước}) = \mathbf{0.2}$
- $\rightarrow$ Chọn "Kích thước"làm nút gốc.

Lần 2 (Nhánh "Nhỏ"):

- $Error_{\text{TB}}(\text{Hình dạng}) = 0.5$
- $Error_{\text{TB}}(\text{Dạng}) = \mathbf{0.0}$
- $Error_{\text{TB}}(\text{Màu}) = 0.25$
- $\rightarrow$ Chọn "Dạng"làm nút con.

Kết luận: Đối với bộ dữ liệu này, kết quả tính toán cho cả Entropy (ID3) và Classification Error đều cho ra cây quyết định hoàn toàn giống như Gini Index. Do đó, bộ luật và kết quả dự đoán cho Mẫu 11 và 12 không thay đổi.

Bài 2: Phân loại quyết định

Dữ liệu gốc (Mẫu 1-10) có 10 mẫu, trong đó:

- Lớp "Quyết định- Có: 6 mẫu (2, 3, 4, 6, 7, 10)
- Lớp "Quyết định- Không: 4 mẫu (1, 5, 8, 9)

a) Xây dựng cây dùng Classification Error

Chúng ta chọn thuộc tính có Lỗi Phân Loại (Error) trung bình thấp nhất.

Lần 1: Chọn nút gốc

$Error(\text{Gốc}) = 1 - \max(6/10) = 0.4$

Error(Phái):

- Nam (6 mẫu: 3 Có, 3 Không): $Error = 1 - \max(3/6) = 0.5$
- Nữ (4 mẫu: 3 Có, 1 Không): $Error = 1 - \max(3/4) = 0.25$
- $Error_{\text{TB}}(\text{Phái}) = (6/10) \times 0.5 + (4/10) \times 0.25 = \mathbf{0.4}$

Error(Công việc):

- Chân tay (6 mẫu: 3 Có, 3 Không): $Error = 0.5$
- Trí óc (4 mẫu: 3 Có, 1 Không): $Error = 0.25$
- $Error_{\text{TB}}(\text{Công việc}) = (6/10) \times 0.5 + (4/10) \times 0.25 = \mathbf{0.4}$

Error(Học vấn):

- Cao đẳng (4 mẫu: 3 Có, 1 Không): $Error = 0.25$
- Đại học (3 mẫu: 2 Có, 1 Không): $Error = 1 - \max(2/3) \approx 0.333$
- Phổ thông (3 mẫu: 1 Có, 2 Không): $Error = 1 - \max(2/3) \approx 0.333$
- $Error_{\text{TB}}(\text{Học vấn}) = (4/10) \times 0.25 + (3/10) \times 0.333 + (3/10) \times 0.333 \approx \mathbf{0.3}$

Error(Độ tuổi):

- Trung niên (4 mẫu: 2 Có, 2 Không): $Error = 1 - \max(2/4) = 0.5$
- Già (4 mẫu: 4 Có, 0 Không): $Error = 0$
- Thanh niên (2 mẫu: 0 Có, 2 Không): $Error = 0$
- $Error_{\text{TB}}(\text{Độ tuổi}) = (4/10) \times 0.5 + (4/10) \times 0 + (2/10) \times 0 = \mathbf{0.2}$

Kết luận Lần 1: Thuộc tính "Độ tuổi"có Error thấp nhất (0.2), được chọn làm nút gốc.

- Nhánh Độ tuổi = Già: (4 Có, 0 Không) → Nút lá, Quyết định = Có.
- Nhánh Độ tuổi = Thanh niên: (0 Có, 2 Không) → Nút lá, Quyết định = Không.
- Nhánh Độ tuổi = Trung niên: (2 Có, 2 Không) → Cần phân nhánh tiếp.

Lần 2: Phân nhánh từ "Độ tuổi = Trung niên"

Xét 4 mẫu con: {1, 2, 4, 8} (2 Có, 2 Không).

Error(Công việc):

- Chân tay (2 mẫu: 1K, 8K → 0 Có, 2 Không): $Error = 0$
- Trí óc (2 mẫu: 2C, 4C → 2 Có, 0 Không): $Error = 0$
- $Error_{\text{TB}}(\text{Công việc}) = (2/4) \times 0 + (2/4) \times 0 = \mathbf{0.0}$

Error(Phái):

- Nam (2 mẫu: 1K, 4C → 1 Có, 1 Không): $Error = 0.5$
- Nữ (2 mẫu: 2C, 8K → 1 Có, 1 Không): $Error = 0.5$
- $Error_{\text{TB}}(\text{Phái}) = (2/4) \times 0.5 + (2/4) \times 0.5 = 0.5$

Error(Học vấn):

- Cao đẳng (2 mẫu: 1K, 4C → 1 Có, 1 Không): $Error = 0.5$
- Đại học (1 mẫu: 2C → 1 Có, 0 Không): $Error = 0$
- Phổ thông (1 mẫu: 8K → 0 Có, 1 Không): $Error = 0$
- $Error_{\text{TB}}(\text{Học vấn}) = (2/4) \times 0.5 + (1/4) \times 0 + (1/4) \times 0 = 0.25$

Kết luận Lần 2: Thuộc tính "Công việc"có Error thấp nhất (0.0), được chọn làm nút con.

- Nhánh Công việc = Chân tay: (0 Có, 2 Không) → Nút lá, Quyết định = Không.
- Nhánh Công việc = Trí óc: (2 Có, 0 Không) → Nút lá, Quyết định = Có.

Bộ luật cây quyết định (từ Classification Error):

- **Luật 1:** Nếu Độ tuổi = Già **THÌ** Quyết định = Có
- **Luật 2:** Nếu Độ tuổi = Thanh niên **THÌ** Quyết định = Không
- **Luật 3:** Nếu Độ tuổi = Trung niên **VÀ** Công việc = Chân tay **THÌ** Quyết định = Không
- **Luật 4:** Nếu Độ tuổi = Trung niên **VÀ** Công việc = Trí óc **THÌ** Quyết định = Có

b) Áp dụng bộ luật

- **Mẫu A: (Nữ, Trí óc, Cao đẳng, Già)**
  - Kiểm tra: Độ tuổi = Già.
  - Áp dụng Luật 1.
  - **Kết quả: Quyết định = Có.**
- **Mẫu B: (Nam, Chân tay, Phổ thông, Trung niên)**
  - Kiểm tra: Độ tuổi = Trung niên.
  - Kiểm tra: Công việc = Chân tay.
  - Áp dụng Luật 3.
  - **Kết quả: Quyết định = Không.**

c) Làm tương tự cho Gini Index và Entropy

Đối với bộ dữ liệu này, kết quả tính toán cho cả Gini Index và Entropy (ID3) đều cho ra cây quyết định hoàn toàn giống như Classification Error.

- Cả ba phương pháp đều chọn "Độ tuổi"làm nút gốc (vì có Gini/Entropy/Error trung bình thấp nhất).
- Cả ba phương pháp đều chọn "Công việc"làm nút con cho nhánh "Độ tuổi = Trung niên".

Do đó, bộ luật và kết quả dự đoán cho Mẫu A và B không thay đổi.