# I. Lý thuyết

## 1. Cây quyết định – thuật toán ID3

- Biểu diễn cây dưới dạng quy tắc logic:

(Trời = nắng ∧ Độ ẩm = bình_thường)
v (Trời = u_ám)
v (Trời = mưa ∧ Gió = yếu)

- Tiêu chuẩn chọn thuộc tính của ID3: **Thuộc tính có độ tăng thông tin lớn nhất.**
- **Entropy:**

▸ Trường hợp tập dữ liệu $S$ có 2 loại nhãn: đúng (+) hoặc sai (-)

$$Entropy(S) = -p_+ log_2 p_+ - p_- log_2 p_-$$

$p_+$: % số mẫu đúng, $p_-$: % số mẫu sai

▸ Trường hợp tổng quát: có $C$ loại nhãn

$$Entropy(S) = \sum_{i=1}^{C} -p_i log_2 p_i$$

$p_i$ : % ví dụ của $S$ thuộc loại $i$

▸ Ví dụ

$$Entropy([9^+, 5^-]) = -(9/14) log_2(9/14) - (5/14) log_2(5/14) = 0.94$$

- **Độ tăng thông tin IG:**

Với tập (con) mẫu $S$ và thuộc tính $A$

$$IG(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Trong đó:

*values* $(A)$: tập các giá trị của $A$

$S_v$ là tập con của $S$ bao gồm các mẫu có giá trị của $A$ bằng $v$

$|S|$ số phần tử của $S$

## 2. Xác định nhãn bằng phân loại Bayes

Tần xuất quan sát thấy nhãn $c_j$ trên tập dữ liệu D: $\frac{count(c_j)}{|D|}$

$$y = argmax_{c_j \in C} P(x_1, x_2, \dots, x_n | c_j) P(c_j)$$

Sử dụng giả thiết về tính độc lập (**Đơn giản!!!**)

$$P(x_1, x_2, \dots, x_n | c_j) = P(x_1 | c_j) P(x_2 | c_j) \dots P(x_n | c_j)$$

Số lần xuất hiện $x_i$ cùng với $c_j$ chia cho số lần xuất hiện $c_j$: $\frac{count(x_i, c_j)}{count(c_j)}$

## 3. Xác định nhãn bằng K-NN

- Giả sử mẫu $x$ có giá trị thuộc tính là $< a_1(x), a_2(x), \dots, an(x) >$, thuộc tính là số thực
- Khoảng cách giữa hai mẫu $x_i$ và $x_j$ là khoảng cách Euclidean

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{n} (a_l(x_i) - a_l(x_j))^2}$$

**Sử dụng K-NN cho bài toán hồi quy thì chỉ cần tính trung bình k giá trị có d nhỏ nhất là được, còn phân loại thì dựa vào số lượng nhãn đa số trong k nhãn lấy ra.**

## II. Bài tập

**1. Cho dữ liệu huấn luyện, trong đó Loại, Độ ồn, KL là thuộc tính và f là nhãn phân loại.**

| Loại | Độ ồn | KL | f |
|---|---|---|---|
| Trống | To | Nặng | - |
| Ghita | To | Nhẹ | + |
| Trống | Nhỏ | Nhẹ | - |
| Piano | Nhỏ | Nặng | - |
| Ghita | Nhỏ | Nặng | + |
| Piano | To | Nhẹ | + |
| Piano | Nhỏ | Nhẹ | - |
| Trống | Nhỏ | Nặng | - |

**a) Bằng phương pháp phân lớp Bayes đơn giản (chỉ rõ các xác suất điều kiện thành phần), xác định nhãn cho ví dụ: Loại = Piano, Độ ồn = To, KL = Nặng.**

**b) Xác định nút gốc cho cây quyết định sử dụng ID3**

**Chú ý TH các thuộc tính cùng độ ưu tiên thì chọn thuộc tính theo thứ tự từ trái sang phải, tức: Loại, Độ ồn, KL.**

BL

a) Ta có:

$y = \text{argmax}_{c_j \in C} P(\text{Loại} = \text{Piano}, \text{Độ ồn} = \text{To}, \text{KL} = \text{Nặng} \mid c_j) P(c_j)$

$= \text{argmax}_{c_j \in C} P(\text{Loại} = \text{Piano} \mid c_j)\, P(\text{Độ ồn} = \text{To} \mid c_j)\, P(\text{KL} = \text{Nặng} \mid c_j) P(c_j)$

Với $c_1$ = +:

$P(\text{Loại} = \text{Piano} \mid +)\, P(\text{Độ ồn} = \text{To} \mid +)\, P(\text{KL} = \text{Nặng} \mid +)\, P(+)$

= 1/3 * 2/3 * 1/3 * 3/8 = 0.0278

Với $c_2$ = –:

$P(\text{Loại} = \text{Piano} \mid -)\, P(\text{Độ ồn} = \text{To} \mid -)\, P(\text{KL} = \text{Nặng} \mid -)\, P(-)$

= 2/5 * 1/5 * 3/5 * 5/8 = 0.03

Vì 0.0278 < 0.03 => y = – => Vậy nhãn cho ví dụ là –

b) Tập dữ liệu S có 2 nhãn là + và –

S = [3+, 5-] => $H(S) = -\frac{3}{8} log_2 \frac{3}{8} - -\frac{5}{8} log_2 \frac{5}{8} = 0.954$

- Tính IG(S, Loại)

  Values(Loại) = {Trống, Ghita, Piano}

  $S_{Trống}$ = [0+, 3-] => $H(S_{Trống}) = -\frac{0}{3} log_2 \frac{0}{3} - -\frac{3}{3} log_2 \frac{3}{3} = 0$

  $S_{Ghita}$ = [2+, 0-] => $H(S_{Ghita}) = -\frac{2}{2} log_2 \frac{2}{2} - -\frac{0}{2} log_2 \frac{0}{2} = 0$

$S_{Piano}$ = [1+, 2-] => H($S_{Piano}$) = $-\frac{1}{3}log_2\frac{1}{3} - -\frac{2}{3}log_2\frac{2}{3} = 0.918$

=> IG(S, Loại) = H(S) – (3/8) H($S_{Trống}$) – (2/8) H($S_{Ghita}$) – (3/8) H($S_{Piano}$) = 0.61

- Tính IG(S, Độ ồn)

  Values(Độ ồn) = {To, Nhỏ}

  $S_{To}$ = [2+, 1-] => H($S_{To}$) = 0.918

  $S_{Nhỏ}$ = [1+, 4-] => H($S_{nhỏ}$) = 0.722

=> IG(S, Độ ồn) = H(S) – (3/8) H($S_{To}$) – (5/8) H($S_{Nhỏ}$) = 0.159

- Tính IG(S, KL)

  Values(KL) = {Nặng, Nhẹ}

  $S_{Nặng}$ = [1+, 3-] => H($S_{Nặng}$) = 0.811

  $S_{Nhẹ}$ = [2+, 2-] => H($S_{Nhẹ}$) = 1

=> IG(S, KL) = H(S) – (4/8) H($S_{Nặng}$) – (4/8) H($S_{Nhẹ}$) = 0.0485

Vì thuộc tính "Loại" cho độ tăng thông tin lớn nhất nên "Loại" được chọn làm nút gốc cho cây quyết định sử dụng thuật toán ID3.

**2. Cho bảng dữ liệu như hình bên với Ai là các thuộc tính và f là nhãn phân loại.**

**Xác định nhãn cho ví dụ: A1 = 1, A2 = 0, A3 = 1 bằng Bayes.**

| A1 | A2 | A3 | f |
|---|---|---|---|
| 0 | 0 | 1 | + |
| 0 | 0 | 2 | + |
| 0 | 0 | 3 | + |
| 0 | 0 | 4 | + |
| 0 | 1 | 1 | - |
| 0 | 1 | 2 | - |
| 0 | 1 | 3 | - |
| 1 | 0 | 4 | - |
| 1 | 1 | 1 | + |
| 1 | 1 | 2 | + |

BL

Ta có:

$y = \text{argmax}_{c_j \in C} P(A1 = 1, A2 = 0, A3 = 1 \mid c_j) P(c_j)$

$= \text{argmax}_{c_j \in C} P(A1 = 1 \mid c_j)\ P(A2 = 0 \mid c_j)\ P(A3 = 1 \mid c_j) P(c_j)$

Với $c_1$ = +:

$P(A1 = 1 \mid +)\ P(A2 = 0 \mid +)\ P(A3 = 1 \mid +)\ P(+)$

= 2/6 * 4/6 * 2/6 * 6/10 = 0.044

Với $c_2$ = –:

$P(A1 = 1 \mid -)\ P(A2 = 0 \mid -)\ P(A3 = 1 \mid -)\ P(-)$

= 1/4 * 1/4 * 1/4 * 4/10 = 0.00625

Vì 0.00625 < 0.044 => y = + => Vậy nhãn cho ví dụ là +

3. **Cho bảng dữ liệu huấn luyện trong đó các dòng A, B, C là thuộc tính và D là nhãn phân loại. Sử dụng thuật toán k láng giềng (k = 3) tìm nhãn phân loại cho mẫu sau:**
**A = 2, B = 2, C = 1.**

| | | | | | | |
|---|---|---|---|---|---|---|
| A | 2 | 2 | 1 | 1 | 2 | 1 |
| B | 1 | 2 | 1 | 2 | 1 | 1 |
| C | 1 | 2 | 1 | 1 | 2 | 2 |
| D | + | + | + | + | - | - |

BL

Từ đề, ta có mẫu cần phân loại là x(2, 2, 1) và k = 3. Áp dụng công thức tính khoảng cách Euclid để tính khoảng cách giữa mẫu x và tất cả các mẫu xi khác trong bảng dữ liệu huấn luyện: $d(x, x_i) = \sqrt{\sum_{l=1}^{3} \left(a_l(x) - a_l(x_i)\right)^2}$

Có:

- x1(2, 1, 1) => d(x, x1) = 1
- x2(2, 2, 2) => d(x, x2) = 1
- x3(1, 1, 1) => d(x, x3) = 1.414
- x4(1, 2, 1) => d(x, x4) = 1
- x5(2, 1, 2) => d(x, x5) = 1.414
- x6(1, 1, 2) => d(x, x6) = 1.732

Vì k = 3 nên ta chọn 3 mẫu có d(x, xi) nhỏ nhất, đó là 3 mẫu x1, x2 và x4. Nhãn chiếm đa số trong 3 mẫu này là "+" => Nhãn phân loại cho mẫu x là "+"

4. **Cho bảng dữ liệu như hình bên với Ai là các thuộc tính và f là nhãn phân loại. Xây dựng cây quyết định sử dụng ID3.**
**Chú ý TH các thuộc tính cùng độ ưu tiên thì chọn thuộc tính theo thứ tự từ trái sang phải, tức: A1, A2, A3.**

| A1 | A2 | A3 | f |
|---|---|---|---|
| 0 | 0 | 1 | + |
| 0 | 0 | 2 | + |
| 0 | 0 | 3 | + |
| 0 | 0 | 4 | + |
| 0 | 1 | 1 | - |
| 0 | 1 | 2 | - |
| 0 | 1 | 3 | - |
| 1 | 0 | 4 | - |
| 1 | 1 | 1 | + |
| 1 | 1 | 2 | + |

BL

Tập dữ liệu S có 2 nhãn là + và –

S = [6+, 4-] => H(S) = 0.971

1. Đầu tiên ta sẽ chọn nút gốc cho cây:

- Tính IG(S, A1)

  Values(A1) = {0, 1}

  $S_0$ = [4+, 3-] => H($S_0$) = 0.985

  $S_1$ = [2+, 1-] => H($S_1$) = 0.918

=> IG(S, A1) = H(S) – (7/10) H($S_0$) – (3/10) H($S_1$) = 0.0061

- Tính IG(S, A2)

  Values(A2) = {0, 1}

  $S_0$ = [4+, 1-] => H($S_0$) = 0.722

  $S_1$ = [2+, 3-] => H($S_1$) = 0.971

=> IG(S, A2) = H(S) – (5/10) H($S_0$) – (5/10) H($S_1$) = 0.1245

- Tính IG(S, A3)

  Values(A3) = {1, 2, 3, 4}

  $S_1$ = [2+, 1-] => H($S_1$) = 0.918

  $S_2$ = [2+, 1-] => H($S_2$) = 0.918

  $S_3$ = [1+, 1-] => H($S_3$) = 1

  $S_4$ = [1+, 1-] => H($S_4$) = 1

=> IG(S, A3) = H(S) – (3/10) H($S_1$) – (3/10) H($S_2$) - (2/10) H($S_3$) – (2/10) H($S_4$) = 0.0202

Với thuộc tính A2 cho độ tăng thông tin lớn nhất, theo ID3 thì thuộc tính này sẽ được chọn làm nút gốc. Sau khi chọn nút gốc là A2, ta được các bộ dữ liệu con ở 2 nhánh tương ứng 2 giá trị của A2 là 0 và 1. Giả sử nhánh trái ta gọi là S’ và nhánh phải là S’’.

2. Chọn thuộc tính cho nút tiếp theo bên trái (hay nút gốc cho cây S’) (A2 = 0, tức chỉ nhìn các hàng A2 = 0):

S’ = [4+, 1-] => H(S) = 0.722

- Tính IG(S’, A1)

  Values(A1) = {0, 1}

$S_0'$ = [4+, 0-] => H($S_0'$) = 0

$S_1'$ = [0+, 1-] => H($S_1'$) = 0

=> IG(S', A1) = H(S') – (4/5) H($S_0'$) – (1/5) H($S_1'$) = 0.722

- Tính IG(S', A3)

  Values(A3) = {1, 2, 3, 4}

  $S_1'$ = [1+, 0-] => H($S_1'$) = 0

  $S_2'$ = [1+, 0-] => H($S_2'$) = 0

  $S_3'$ = [1+, 0-] => H($S_3'$) = 0

  $S_4'$ = [1+, 1-] => H($S_4'$) = 1

=> IG(S', A3) = H(S') – (1/5) H($S_1'$) – (1/5) H($S_2'$) - (1/5) H($S_3'$) – (2/5) H($S_4'$) = 0.322

Thuộc tính A1 có độ tăng thông tin lớn nhất nên được chọn cho nút này. Sau khi chọn nút là A1, ta được các bộ dữ liệu con ở 2 nhánh tương ứng 2 giá trị của A1 là 0 và 1. Đối với nhánh trái (A1 = 0, A2 = 0), toàn bộ mẫu có nhãn dương còn nhánh phải (A1 = 1, A2 = 0), toàn bộ mẫu có nhãn âm, do vậy quá trình học cây cho cả 2 nhánh này dừng lại, thuật toán tạo 2 nút lá nhãn "+" và "-".

3. Chọn thuộc tính cho nút tiếp theo bên phải (hay nút gốc cho cây S'') (A2 = 1):

S'' = [2+, 3-] => H(S) = 0.971

- Tính IG(S'', A1)

  Values(A1) = {0, 1}

  $S_0''$ = [0+, 3-] => H($S_0''$) = 0

  $S_1''$ = [2+, 0-] => H($S_1''$) = 0

=> IG(S'', A1) = H(S'') – (3/5) H($S_0''$) – (2/5) H($S_1''$) = 0.971

- Tính IG(S'', A3)

  Values(A3) = {1, 2, 3, 4}

  $S_1''$ = [1+, 1-] => H($S_1''$) = 1

  $S_2''$ = [1+, 1-] => H($S_2''$) = 1

  $S_3''$ = [0+, 1-] => H($S_3''$) = 0

  $S_4''$ = [0+, 0-] => H($S_4''$) = 0

=> IG(S'', A3) = H(S'') – (2/5) H($S_1''$) – (2/5) H($S_2''$) - (1/5) H($S_3''$) – (0/5) H($S_4''$) = 0.171

Thuộc tính A1 có độ tăng thông tin lớn nhất nên được chọn cho nút này. Sau khi chọn nút là A1, ta được các bộ dữ liệu con ở 2 nhánh tương ứng 2 giá trị của A1 là 0 và 1. Đối với nhánh trái (A1 = 0, A2 = 1), toàn bộ mẫu có nhãn âm còn nhánh phải (A1 = 1, A2 = 1), toàn bộ mẫu có nhãn dương, do vậy quá trình học cây cho cả 2 nhánh này dừng lại, thuật toán tạo 2 nút lá nhãn "+" và "-".

Cuối cùng chúng ta có cây quyết định theo ID3: