

**Bài 1,**

Cho tập mẫu và kết quả đoán nhận của 1 mô hình học máy như sau

Mẫu	Phân lớp trong thực tế	Phân lớp sử dụng 1 mô hình học máy
X <sub>1</sub>	C <sub>1</sub>	C <sub>1</sub>
X <sub>2</sub>	C <sub>1</sub>	C <sub>1</sub>
X <sub>3</sub>	C <sub>2</sub>	C <sub>1</sub>
X <sub>4</sub>	C <sub>1</sub>	C <sub>2</sub>
X <sub>5</sub>	C <sub>2</sub>	C <sub>2</sub>
X <sub>6</sub>	C <sub>2</sub>	C <sub>2</sub>
X <sub>7</sub>	C <sub>2</sub>	C <sub>2</sub>
X <sub>8</sub>	C <sub>1</sub>	C <sub>2</sub>
X <sub>9</sub>	C <sub>2</sub>	C <sub>2</sub>
X <sub>10</sub>	C <sub>1</sub>	C <sub>2</sub>
X <sub>11</sub>	C <sub>1</sub>	C <sub>1</sub>
X <sub>12</sub>	C <sub>1</sub>	C <sub>1</sub>
X <sub>13</sub>	C <sub>2</sub>	C <sub>1</sub>
X <sub>14</sub>	C <sub>2</sub>	C <sub>2</sub>
X <sub>15</sub>	C <sub>1</sub>	C <sub>2</sub>

a, Hãy tính các giá trị Recall, Precision và F1 cho mỗi lớp đầu ra tương ứng

b, Hãy tính Recall, Precision và F1 cho mô hình phân lớp trên theo phương pháp Micro-average

c, Hãy tính Recall, Precision và F1 cho mô hình phân lớp trên theo phương pháp Macro-average

**Bài 2,**

Cho tập mẫu và kết quả đoán nhận của 1 mô hình học máy như sau

Mẫu	Phân lớp trong thực tế	Phân lớp sử dụng 1 mô hình học máy
D <sub>1</sub>	C <sub>1</sub>	C <sub>2</sub>
D <sub>2</sub>	C <sub>1</sub>	C <sub>1</sub>
D <sub>3</sub>	C <sub>2</sub>	C <sub>2</sub>
D <sub>4</sub>	C <sub>1</sub>	C <sub>2</sub>
D <sub>5</sub>	C <sub>2</sub>	C <sub>2</sub>
D <sub>6</sub>	C <sub>2</sub>	C <sub>1</sub>

D <sub>7</sub>	C <sub>2</sub>	C <sub>2</sub>
D <sub>8</sub>	C <sub>1</sub>	C <sub>2</sub>
D <sub>9</sub>	C <sub>2</sub>	C <sub>2</sub>
D <sub>10</sub>	C <sub>1</sub>	C <sub>2</sub>
D <sub>11</sub>	C <sub>1</sub>	C <sub>1</sub>
D <sub>12</sub>	C <sub>1</sub>	C <sub>1</sub>
D <sub>13</sub>	C <sub>2</sub>	C <sub>1</sub>
D <sub>14</sub>	C <sub>2</sub>	C <sub>2</sub>
D <sub>15</sub>	C <sub>1</sub>	C <sub>2</sub>
D <sub>16</sub>	C <sub>1</sub>	C <sub>2</sub>
D <sub>17</sub>	C <sub>2</sub>	C <sub>2</sub>
D <sub>18</sub>	C <sub>2</sub>	C <sub>1</sub>
D <sub>19</sub>	C <sub>2</sub>	C <sub>1</sub>
D <sub>20</sub>	C <sub>1</sub>	C <sub>1</sub>

a, Hãy tính các giá trị Recall, Precision và F1 cho mỗi lớp đầu ra tương ứng

b, Hãy tính Recall, Precision và F1 cho mô hình phân lớp trên theo phương pháp Micro-average

c, Hãy tính Recall, Precision và F1 cho mô hình phân lớp trên theo phương pháp Macro-average

### Bài 3,

Cho tập mẫu với 2 thuộc tính sau:

Mẫu	Thuộc tính A <sub>1</sub>	Thuộc tính A <sub>2</sub>	Lớp
D <sub>1</sub>	T	F	C <sub>2</sub>
D <sub>2</sub>	T	F	C <sub>1</sub>
D <sub>3</sub>	F	T	C <sub>2</sub>
D <sub>4</sub>	F	T	C <sub>2</sub>
D <sub>5</sub>	F	T	C <sub>2</sub>
D <sub>6</sub>	T	T	C <sub>1</sub>
D <sub>7</sub>	F	F	C <sub>2</sub>
D <sub>8</sub>	T	T	C <sub>2</sub>
D <sub>9</sub>	T	F	C <sub>2</sub>
D <sub>10</sub>	F	T	C <sub>2</sub>

a, Hãy tìm Entropy của tập dữ liệu theo phân lớp trên

b, Tính độ thu hoạch thông tin (Information Gain) của thuộc tính  $A_1$

c, Tính độ thu hoạch thông tin (Information Gain) của thuộc tính  $A_2$

#### Bài 4,

Cho tập dữ liệu huấn luyện với 3 thuộc tính  $A_1, A_2, A_3$  như trong bảng sau:

Mẫu	$A_1$	$A_2$	$A_3$	Lớp
$D_1$	T	H	P	Y
$D_2$	T	H	P	N
$D_3$	F	N	P	Y
$D_4$	T	L	N	N
$D_5$	F	H	N	Y
$D_6$	T	H	P	Y
$D_7$	F	L	N	Y
$D_8$	F	N	P	N
$D_9$	F	N	N	N
$D_{10}$	T	N	N	Y
$D_{11}$	T	L	N	N
$D_{12}$	T	H	P	Y
$D_{13}$	T	L	N	Y
$D_{14}$	T	H	N	N
$D_{15}$	F	L	P	Y

a, Hãy xây dựng cây quyết định tương ứng với bộ dữ liệu trên sử dụng thuật toán ID3

b, Giả sử  $D_5$  bị thiếu giá trị của thuộc tính  $A_2$ , hãy điền giá trị còn thiếu đó theo 2 cách

#### Bài 5,

Với bộ dữ liệu của bài 4, hãy xây dựng cây quyết định sử dụng thuật toán C4.5 (dùng Gain Ratio thay cho Information Gain)

#### Bài 6,

Cho tập dữ liệu huấn luyện với 4 thuộc tính như trong bảng sau:

Day	Outlook	Temperature	Humidity	Wind	Play tennis
$D_1$	Sunny	Hot	High	Weak	No

D <sub>2</sub>	Sunny	Hot	High	Strong	No
D <sub>3</sub>	Overcast	Hot	High	Strong	Yes
D <sub>4</sub>	Rain	Mild	Normal	Weak	Yes
D <sub>5</sub>	Rain	Cool	Normal	Weak	Yes
D <sub>6</sub>	Rain	Cool	Normal	Strong	No
D <sub>7</sub>	Overcast	Cool	Normal	Strong	No
D <sub>8</sub>	Sunny	Mild	High	Weak	No
D <sub>9</sub>	Sunny	Cool	Normal	Strong	Yes
D <sub>10</sub>	Rain	Mild	High	Weak	Yes
D <sub>11</sub>	Sunny	Mild	Normal	Strong	Yes
D <sub>12</sub>	Overcast	Mild	Normal	Strong	Yes
D <sub>13</sub>	Overcast	Hot	Normal	Weak	Yes
D <sub>14</sub>	Rain	Mild	High	Strong	No
D <sub>15</sub>	Sunny	Mild	Normal	Weak	Yes
D <sub>16</sub>	Overcast	Cool	High	Strong	No

a, Hãy xây dựng cây quyết định T1 tương ứng với bộ dữ liệu trên sử dụng thuật toán ID3.

b, Hãy xây dựng cây quyết định T2 tương ứng với bộ dữ liệu trên sử dụng thuật toán C4.5 (dùng Gain Ratio thay cho Information Gain)

c, Cho mẫu quan sát mới D=<Sunny, Mild, High, Strong>, tìm kết quả “Chơi” hay “Không chơi” tennis tương ứng với cây T1 và T2 vừa xây dựng được

## Bài 7,

Mỗi đối tượng có hai thuộc tính đặc trưng, đặc trưng thứ nhất nhận hai giá trị 0 hoặc 1, đặc trưng thứ hai nhận giá trị trong 5 chữ cái. Dữ liệu quan sát được gồm hai lớp cho bởi bảng

c <sub>1</sub>	1A	0E	0C	1D
c <sub>2</sub>	0A	1B	1C	1D

a, , Hãy xây dựng cây quyết định T1 tương ứng với bộ dữ liệu trên sử dụng thuật toán ID3.

b, Hãy xây dựng cây quyết định T2 tương ứng với bộ dữ liệu trên sử dụng thuật toán C4.5 (dùng Gain Ratio thay cho Information Gain)

**Bài 8,**

Cho tập dữ liệu huấn luyện với 2 thuộc tính  $a_1$ ,  $a_2$  như sau:

Mẫu	$a_1$	$a_2$	Lớp
$x_1$	0,15	0,84	$\omega_1$
$x_2$	0,8	0,54	$\omega_1$
$x_3$	0,28	0,36	$\omega_1$
$x_4$	0,38	0,70	$\omega_1$
$x_5$	0,52	0,48	$\omega_1$
$x_6$	0,58	0,74	$\omega_1$
$x_7$	0,74	0,76	$\omega_1$
$x_8$	0,08	0,14	$\omega_2$
$x_9$	0,24	0,16	$\omega_2$
$x_{10}$	0,70	0,20	$\omega_2$
$x_{11}$	0,90	0,28	$\omega_2$
$x_{12}$	0,64	0,90	$\omega_2$
$x_{13}$	0,74	0,36	$\omega_2$
$x_{14}$	0,10	0,30	$\omega_2$

Hãy sử dụng kỹ thuật xử lý thuộc tính liên tục của Quinlan trong C4.5 để đưa giá trị của 2 thuộc tính  $a_1$  và  $a_2$  ở trên thành rời rạc (nhị phân).

**Bài 9,**

Cho bộ dữ liệu huấn luyện sau

Day	Outlook	Temperature	Humidity	Wind	Play tennis
D <sub>1</sub>	Sunny	Hot	High	Weak	No
D <sub>2</sub>	Sunny	Hot	High	Strong	No
D <sub>3</sub>	Overcast	Hot	High	Strong	Yes
D <sub>4</sub>	Rain	Mild	Normal	Weak	Yes
D <sub>5</sub>	Rain	Cool	Normal	Weak	Yes
D <sub>6</sub>	Rain	Cool	Normal	Strong	No
D <sub>7</sub>	Overcast	Cool	Normal	Strong	No
D <sub>8</sub>	Sunny	Mild	High	Weak	No
D <sub>9</sub>	Sunny	Cool	Normal	Strong	Yes
D <sub>10</sub>	Rain	Mild	High	Weak	Yes
D <sub>11</sub>	Sunny	Mild	Normal	Strong	Yes
D <sub>12</sub>	Overcast	Mild	Normal	Strong	Yes

D <sub>13</sub>	Overcast	Hot	Normal	Weak	Yes
D <sub>14</sub>	Rain	Mild	High	Strong	No
D <sub>15</sub>	Sunny	Mild	Normal	Weak	Yes
D <sub>16</sub>	Overcast	Cool	High	Strong	No

Sử dụng thuật toán Naïve Bayes xác định việc đi chơi hoặc không đối với các mẫu sau:

a, D<sub>17</sub> = (Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

b, D<sub>18</sub> = (Outlook=Overcast, Temperature=Hot, Humidity=Normal, Wind=Weak)

### Bài 10,

Một địa phương có 0,8% dân số bị bệnh lao. Xét nghiệm sinh hóa T cho thấy 98% người mắc bệnh lao cho kết quả dương tính (+) và 97% người không mắc bệnh lao cho kết quả âm tính (-). Một người xét nghiệm sinh hóa T cho kết quả dương tính, dùng tiêu chuẩn MAP của thuật toán Naïve Bayes để kết luận được người này bị lao hay không?

### Bài 11,

Giả sử có một trang trại phôi giống gà nuôi chung một giống gà nước ngoài GN (1200 con) và hai giống gà trong nước G1 (900 con) và G2 (900 con). Sau 3 tháng tuổi ta lọc các con lớn nuôi riêng. Biết rằng gà ta có trọng lượng 400g- 500g là lớn nhưng gà ngoại nhập thì là bé và các xác suất này trong mỗi loài đều bằng 0,5. Một con gà ngẫu nhiên có trọng lượng thuộc khoảng này, chỉ dựa vào trọng lượng để xác định loại gà lớn hay bé

### Bài 12,

Cho tập dữ liệu quan sát trong R2 gồm 3 lớp:

$$S1 = \{(0,0), (1,0), (0,1)\}$$

$$S2 = \{(1,1), (2,1)\}$$

$$S3 = \{(-1,0), (-1,-1), (1,-2), (1,-1)\}$$

$$\text{Tỉ lệ học nuy}=1$$

a, Áp dụng thuật toán học perceptron để phân biệt S1 và S2 với khởi tạo là:  $\mathbf{w} = (1,1)^T$ ,  $w_0 = 0$  và hàm kích hoạt là hàm hard limiter 1 (hl1)

b, Áp dụng thuật toán học perceptron để phân biệt S1 và S3 với khởi tạo là:  $\mathbf{w} = (1,1)^T$ ,  $w_0 = 0$  và hàm kích hoạt là hàm hard limiter 1 (hl1)

c, Áp dụng thuật toán học perceptron để phân biệt S2 và S3 với khởi tạo là:  $\mathbf{w} = (1,1)^T$ ,  $w_0 = 0$  và hàm kích hoạt là hàm hard limiter 1 (hl1)

### Bài 13,

Với tập dữ liệu ở bài 12, hãy:

a, Áp dụng thuật toán học perceptron để phân biệt S1 và S2 với khởi tạo là:  $\mathbf{w} = (1,1)^T$ ,  $w_0 = 0$  và hàm kích hoạt là hàm hard limiter 2 (hàm sign)

b, Áp dụng thuật toán học perceptron để phân biệt S1 và S3 với khởi tạo là:  $\mathbf{w} = (1,1)^T$ ,  $w_0 = 0$  và hàm kích hoạt là hàm hard limiter 2 (hàm sign)

c, Áp dụng thuật toán học perceptron để phân biệt S2 và S3 với khởi tạo là:  $\mathbf{w} = (1,1)^T$ ,  $w_0 = 0$  và hàm kích hoạt là hàm hard limiter 2 (hàm sign)

### Bài 14,

Sử dụng thuật toán GA giải phương trình sau với các tham số tương ứng:

$$x^2 = 81$$

- Hàm thích nghi:  $F(x) = 1000 - x^2 + 81$
- Ngưỡng thích nghi:  $\theta = 900$
- Số lượng cá thể của quần thể  $n = 6$
- Tỷ lệ lai ghép  $r_{co} = 0,5$
- Tỷ lệ đột biến  $r_{mu} = 0,5$
- Số vòng lặp tối đa  $lm = 3$
- Khởi tạo các lời giải bởi chuỗi nhị phân độ dài 4
- Phép lai ghép 2 chuỗi x và y: đổi chỗ 2 bit cuối và 2 bit đầu của chuỗi x và y cho nhau
- Phép đột biến: biến đổi bit thứ 3 trong chuỗi

