**BUỔI 1**

## Chương 1: Định nghĩa học máy

* Định nghĩa học : Học là quá trình tìm hiểu, tiếp thu những thứ mình chưa biết
* Định nghĩa tri thức : Bao gồm những kiến thức, thông tin, sự hiểu biết, hay kỹ năng có được nhờ [trải nghiệm](https://vi.wikipedia.org/wiki/Tr%E1%BA%A3i_nghi%E1%BB%87m),thông qua [giáo dục](https://vi.wikipedia.org/wiki/Gi%C3%A1o_d%E1%BB%A5c) hay tự học hỏi.
* Độ to : Entropy : E = -p.log2p

P = m/n

## Chương 2: DECISION TREE (Cây quyết định)

**2.1 Problem Statement (Phát biểu bài toán)**

- Cho CSDL

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TT | Tên :X | Màu tóc A1 | Chiều cao A2 | Cân nặng A3 | SD thuốc A4 | Kết quả :y |
| 1 | Hoa | Đen | Tầm thước | Nhẹ | Không | Rám |
| 2 | Lan | Đen | Cao | Vừa phải | Có | Không |
| 3 | Xuân | Râm | Thấp | Vừa phải | Có | Không |
| 4 | Hạ | Đen | Thấp | Vừa phải | Không | Rám |
| 5 | Thu | Bạc | Tầm thước | Nặng | Không | Rám |
| 6 | Đông | Râm | Cao | Nặng | Không | Không |
| 7 | Mơ | Râm | Tầm thước | Nặng | Không | Không |
| 8 | Đào | Đen | Thấp | Nhẹ | Có | Không |

A = {A1, A2, A3, A4}

Phân tích :

* 8 bản ghi : 8 quan sát

+ Dữ liệu này gọi là dữ liệu học

+ Dữ liệu thử khi đã tìm được quy luật

**2.3 Phương pháp học:** **Chia để trị (Divide and Conquer)**

* Có 2 cách chia: chia theo đầu vào, chia theo đầu ra
* Chia theo đầu vào: (dung cho các thuật toán Quinlan, Độ lộn xộn, ID3, C4.5, C5.0; thuật toán Bayes…

+ **Bước 1**: Hoàn toàn tương tự như thuật toán Quinlan dung phương pháp “chia để trị (tức tạo cây)

**Cân nặng:A3**

|  |  |  |
| --- | --- | --- |
| *j=1 Nhẹ* | *j=2 Vừa* | *j=3 Nặng* |
| **Hoa** | **Lan** | **Thu** |
| **Đào** | **Xuân** | **Đông** |
|  | **Hạ** | **Mơ** |
|  |  |  |

**Chiều cao: A2**

|  |  |  |
| --- | --- | --- |
| *j=1 T.Thước* | *j=2 Cao* | *j=3 Thấp* |
| **Hoa** | **Lan** | **Xuân** |
| **Thu** | **Đông** | **Hạ** |
| **Mơ** |  | **Đào** |
|  |  |  |

**Màu tóc: A1**

|  |  |  |
| --- | --- | --- |
| *j=1 Đen* | *j=2 Bạc* | *j=3*  *Râm* |
| **Hoa** | **Thu** | **Đông** |
| **Lan** |  | **Xuân** |
| **Hạ** |  | **Mơ** |
| **Đào** |  |  |

**Dùng thuốc: A4**

|  |  |
| --- | --- |
| *j=1*  *Có* | *j=2*  *Không* |
| **Hoa** | **Lan** |
| **Thu** | **Xuân** |
| **Hạ** | **Đào** |
| **Đông** |  |
| **Mơ** |  |

**Tỷ lệ số nhóm có giá trị đồng nhất (hoặc đồng nhất rám hoặc không) trên tổng số nhóm**

Cách tính giá trị ở dưới :

Lộn xộn : Không tính đc

Ví dụ A1 : Đen : Lộn Xộn 🡺 Không tính đc

Bạc : Chắc chắn bị rám 🡺 TÍnh đc

Râm : Chắc chắn không bị rám 🡺 Tính đc

🡺Có 2/3 giá trị có thể tính đc

Các số khác trong bảng tương tự

|  |  |  |  |
| --- | --- | --- | --- |
| 2/3 | 1/3 | 0/3 | 1/2 |

+ **BƯỚC 2**: Tính toán theo độ lộn xộn dùng Entropy

trong đó:

*(3.2)*

*nj*: số mẫu nhánh có đặc trưng *Ai* (trong ví dụ đặc trưng *A1:màu tóc;…)*

*nt*: số tổng số mẫu quan sát *(nt=8* trong ví dụ của chúng ta*)*

*nic*: số mẫu nhánh cây thứ *i* với giá trị đầu ra *c* (*c=1:*Rám; *c=2:* Không cho ví dụ này)

a, = 0.5

|  |  |  |
| --- | --- | --- |
|  | ***Thuật toán độ lộn xộn*** | |
| *1* | *EA1: Màu tóc* | *0,50 (Min)* |
| 2 | *EA2*: Chiều cao | 0,69 |
| 3 | *EA3*: Cân nặng | 0,94 |
| 4 | *EA4*: Dùng thuốc | 0,64 |

*Xóa 3 gốc: Chiều cao, Cân năng, Dùng thuốc***.** *Giữ lại gốc “Màu tóc”*

**BƯỚC 3:** Chọn độ lộn xộn nhỏ nhất EA1 = 0.50

**BƯỚC 4:** Viết luật, dựa vào E=0 ở EA1

Luật 1 : If ‘ A1’ là bạc

Then ‘kết quả’ là rám

Luật 2 : if ‘A1’ là râm

Then ‘kết quả’ là không rám

**BƯỚC 5**: Tổ hợp chập 2 đặc trưng đầu vào, ta có 3 trường hợp sau:

Xóa nhánh A1(j=2), A1(j=3)

|  |  |
| --- | --- |
| J=1 đen | (a) (A1 là đen) AND (A2: chiều cao)  (b) (A1 là đen) AND (A3: cân nặng) |
| Hoa | (c) (A1 là đen) AND (A4:dùng thuốc) |
| Lan |  |
| Hạ |  |
| Đào |  |

***Bảng 2.1. Bảng dữ liệu thử nghiệm E « Rám Nắng » cò lại sau vòng 1***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TT** | **Tên**  **A :** | **Màu tóc**  **{A1** | **Chiều cao**  **A2** | **Cân năng**  **A3** | **Dùng thuốc**  **A4}** | **Kết quả**  (Label) B |
| 1 | Hoa | Đen | Tầm thước | Nhẹ | Không (N) | Bị rám (y) |
| 2 | Lan | Đen | Cao | Vừa phải | Có (Y) | Không (N) |
| 4 | Hạ | Đen | Thấp | Vừa phải | Không | Bị Rám |
| 8 | Đào | Đen | Thấp | Nhẹ | Có | Không |

và lặp lại các bước từ B2 đến B4.

Chọn E= 0

BTVN:

1. Cài PM WEKA
2. Trong thư mục data của WEKA có tệp dữ liệu Weather normal / Weather numeric

Chọn phân lớp

**BUỔI 3**

### 2.4 Remarks

2.4.1: Thuật toán có đầu ra mong muốn: Kết Quả

2.4.2: Loại học có 3 loại

* Học có giám sát (Học có tín hiệu chỉ đạo, hoặc học có giảng viên) : Supervised
* Học không giám sát: Unsupervised Learning
* Học củng cố:

### 2.5 Thuật toán ID3 ( Inductive Dichotomizer )

* ID3 là cải tiến của thuật toán độ lộn xộn
* Giống thuật toán độ lộn xộn
  + Phương pháp: Chia để trị
  + Các bước: Y hệt
  + Khác: Tiêu chí chọn gốc:

EToàn CSDL – EAi 🡪 MAX (Tiêu chí chọn gốc)

🡺

Trong đó: IG (Information Gain) là độ tăng (độ lợi) của thông tin

Với ví dụ trên:

*S = [3+, 5-]; H(S) =*

So sánh 2 thuật toán:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ***Thuật toán độ lộn xộn*** | | ***Thuật toán ID3*** | |
| *1* | *EA1: Màu tóc* | *0,50 (Min)* | *H(S, EA1): Màu tóc* | *0.95-0,50=0.45 (Max)* |
| 2 | *EA2*: Chiều cao | 0,69 | *H(S, EA2):* Chiều cao | 0.95-0,69=0,36 |
| 3 | *EA3*: Cân nặng | 0,94 | *H(S, EA3):* Cân nặng | 0.95-0,94=0.01 |
| 4 | *EA4*: Dùng thuốc | 0,64 | *H(S, E41)*: Dùng thuốc | 0.95-0,64=0,31 |

BTVN: Làm nốt bài này với thuật toán ID3

### 2.6 Phần mềm WEKA

2.7 Đánh giá **🡺** Sai số : error : Ký hiệu e = d – y

* MAE (Mean Absolute Error)
  + L1 = MAE =
* L2 = RMSE = alt
* Confusion matrix (Ma trận nhầm lẫn)

**BUỔI 4**

### 2.8 Các thuật toán khác của họ cây quyết định

* Thuật toán C4.5 (Trong WEKA là J4.8)
* Thuật toán C5.0
* Rừng ngẫu nhiên (RANDIM FOREST)
* Thuật toán QueenLan
* Thuật toán ILA

## Chương 4: Naïve Bayes A

4.1 Các khái niệm xác suất

4.2 Xác suất có điều kiện – Xác suất Bayes

4.3 Bài toán

***Input :*** chobảng dữ liệu học chơi Tennis

**Bảng 4.2. Bảng dữ liệu học chơi Tennis**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***A=*** | ***A1*** | ***A2*** | ***A3*** | ***A4*** | ***B*** |
| ***Day*** | ***Outlook (Sky)*** | ***Temperature*** | ***Humidity*** | ***Wind*** | ***PlayTennis*** |
| *Dl* | Sunny | Hot | High | False | No |
| *D2* | Sunny | Hot | High | True | No |
| *D3* | Overcast | Hot | High | False | Yes |
| *D4* | Rainy | Mild | High | False | Yes |
| *D5* | Rainy | Cool | Normal | False | Yes |
| *D6* | Rainy | Cool | Normal | True | No |
| *D7* | Overcast | Cool | Normal | True | Yes |
| *D8* | Sunny | Mild | High | False | No |
| *D9* | Sunny | Cool | Normal | False | Yes |
| *Dl0* | Rain | Mild | Normal | False | Yes |
| *Dl1* | Sunny | Mild | Normal | True | Yes |
| *Dl2* | Overcast | Mild | High | True | Yes |
| *Dl3* | Overcast | Hot | Normal | False | Yes |
| *Dl4* | Rain | Mild | High | True | No |

***Output:*** Dữ liệu của mẫu tin 1 cần dự báo Play= No? Play=Yes dùng thuật toán Bayes

|  |  |
| --- | --- |
|  | **Outlook Temp Humidity Windy Play: B** |
| *D15* | Sunny Hot High True ? |

**Bước 1**: *Chia (phân hoạch) theo đặc trưng đầu vào* hay còn gọi là lập mô hình

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GỐC: Outlook: A1 | | | Temp:A2 | | | Humidity:A3 | | | Windy: A4 | | | Play | |
| Yes | | No | Yes | | No | Yes | | No | Yes | | No | Yes | No |
| Sunny  Overcast  Rainy | 2  4  3 | 3  0  2 | Hot  Wild  Cool | 2  4  3 | 2  2  1 | High  Normal | 3  6 | 4  1 | False  True | 6  3 | 2  3 | 9 | 5 |



**Bước 2:** *Sử dụng công thức Bayes*

***B2.1.*** *Xác định biến hỏi (B(yes, no)=Play(yes, no)); Biến điều kiện*

*(A={A1, A2, A3, A4}={A1 A2A3A4}).* Từ đó, tính 2 công thức với:

***B2.2.*** (4)

(5)

Trong công thức (4), (5): =

=

Tính Để đơn giản, ta giả sử các đặc trưng đầu vào A1…A4 độc lập (không phụ thuộc) của xác suất có điều kiện. Khi đó ta có:

Từ đó, chúng ta có:

(6)

(7)

***B2.3.*** Lấy các giá trị yêu cầu và thay vào hai công thức cuối cho mỗi thuộc tính để dự báo:

|  |  |
| --- | --- |
|  | **Outlook: A1 Temp :A2 Humidity:A3 Windy:A4 Play:B** |
| *D15* | Sunny Hot High True ? |

= (8)

= (9)

***B3. So sánh (8) với (9)***

Hai công thức tính xác suất có điều kiện có mẫu số *P(A)* chung, để só sánh chúng, ta chỉ cần so sánh hai tử số của chúng. Dễ thấy (9)>(8), suy diễn, suy ra,

*Ghi chú: Kết luận mà ta có được bằng phép SUY DIỄN theo quy luật: SỐ LỚN (MAX), hay lối suy diễn này là “luật bầu cử, bỏ phiếu”*

**Buổi 5**

4.5. Many Bayes (Bayes Net)

4.5.1. Mục đích giải quyết trường hợp các biến không độc lập với nhau

4.5.2. Một số tính chất của xác suất có điều kiện

a) Luật tích: P(A|B) = P(B|A).P(A)/P(B) (Định lý Bayes)

P(A|B) = P(A˄B)/P(B) = P(A,B)/P(B)

🡺P(A,B) = P(A|B).P(B)

b) Luật tích mở rộng:

P(A,B,C) = P(A˄(B˄C)) = P(A|B˄C).P(B˄C) = P(A|B˄C).P(B|C).P(C)

= P(A | B,C).P(B,C)

Theo Bayes: P(A|B,C) = (P(B,C|A).P(A))/(P(B,C))

c) Xác suất của P(B,C|A). Theo đ/n x/s có điều kiện

Theo đ/n:

P(B,C|A) = P(B˄C˄A)/P(A) = P(A|B˄C).P(B|C).P(C)/P(A)

Theo tính chất giao hoán P(B˄C˄A) = P(A˄B˄C)

**4.5.4. Ví dụ về mạng Bayes**

Để tiện cho việc trình bày khái niệm mạng Bayes, xét một ví dụ sau

Một người đi làm về và muốn dự đoán xem ở nhà có người không thông qua một số dấu

hiệu có thể quan sát được. Cho biết một số dữ kiện sau:

- Nếu cả nhà đi vắng thì thường bật đèn ngoài sân. Tuy nhiên, đèn ngoài sân có thể

được cả trong một số trường hợp có người ở nhà, ví dụ khi có khách đến chơi.

- Nếu cả nhà đi vắng thì thường buộc chó ở sân sau.

Suy diễn xác suất

- Tuy nhiên chó có thể được buộc ở sân sau cả khi có người ở nhà nếu như chó bị

đau bụng.

- Nếu chó buộc ở ngoài thì có thể nghe tiếng sủa, tuy nhiên có thể nghe tiếng sủa

(của chó hàng xóm) cả khi chó không buộc ở ngoài.

Để thực hiện suy diễn xác suất cho bài toán trên, trước tiên cần xây dựng mô hình xác

suất. Ta sẽ sử dụng năm biến ngẫu nhiên sau để thể hiện các dữ kiện liên quan tới bài toán.

O: không ai ở nhà

L: đèn sáng

D: chó ở ngoài

B: chó bị ốm.

H: nghe thấy tiếng sủa.

Việc phân tích bài toán cho thấy:

- Nếu biết D thì H không phụ thuộc vào O, L, B.

- Nếu biết B thì D độc lập với O.

- O và B độc lập với nhau

Tiếp theo, ta xây dựng một đồ thị, trong đó mỗi biến ngẫu nhiên ở trên được biểu diễn

bởi một nút như trên hình vẽ dưới đây (hình 4.1). Các nút được nối với nhau bằng những cung

có hướng sao cho hai hai nút có quan hệ phụ thuộc được nối bởi một cung và hướng của cung

thể hiện chiều tác động của nút gốc tới nút đích. Với đồ thị có hướng, ta có thể xác định quan

hệ giữa các nút như sau: nếu tồn tại cung có hướng từ nút A tới nút B thì nút A được gọi là

nút cha (mẹ) và nút B là nút con.

Diagram, schematic

Description automatically generated

Hình 4.1: Một ví dụ mạng Bayes

Sau khi có đồ thị, ta thêm vào bảng xác suất điều kiện. Bảng xác suất điều kiện thể hiện

xác suất của biến khi biết giá trị cụ thể của các biến ở các nút cha mẹ. Trong trường hợp nút

không có cha mẹ, xác suất trở thành xác suất tiền nghiệm. Để thuận tiện, bảng xác suất điều

kiện được thể hiện ngay trên hình vẽ cùng với đồ thị.

Đồ thị vừa xây dựng cùng với các bảng xác suất điều kiện tạo thành mạng Bayes cho

bài toán trong ví dụ trên.

**Buổi 6**

## Chương 3: K-Láng Giềng Gần Nhất (K-NN)

## Chương 5: Neural Networks

### 5.1 Introduction

- Máy tính nơron

- Phỏng nơron : Neuron : Tế bào thần kinh

- Não người có 15 tỷ Neuron

### 5.2 Phỏng 1 neuron

5.2.1 Neuron sinh học

5.2.2 Neuron nhân tạo

5.2.3 Độ lệch hay ngưỡng

**Buổi 7**

### 5.3 Neuro Computing

5.3.1 Computing

- Digital Computing (Binary)

+ Cổng And (And Gate)

+ Cổng OR (OR Gate)

+ Cổng Not (Not Gate)

Phương pháp: Học Perceptron

**Buổi 8**

### 5.4 Mạng nơ ron nhân tạo

5.4.1 Mạng là gì, tại sao mạng

* Một nơ ron có thể được
  + And Gate
  + Or Gate
  + Not Gate

5.4.2 Mạng nơ ron 1 lớp – Mạng PERCETRAN