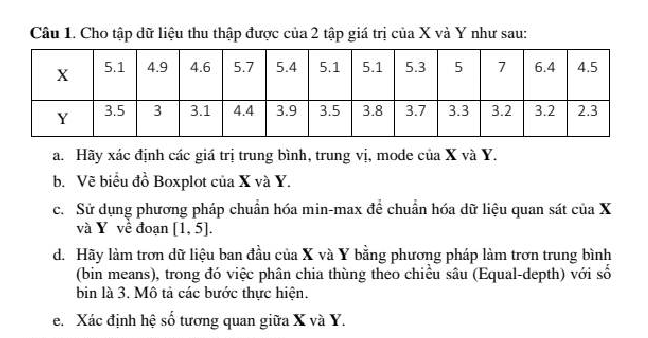
Đề cương Khai phá dữ liệu

Câu1: **(gồm các dạng y hệt đề thi mẫu như dưới)**



**Cách làm:**

1. Vì tập dữ liệu trong bài chưa được sắp xếp nên phải sắp xếp theo thứ tự từ bé đến lớn:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 4.5 | 4.6 | 4.9 | 5 | 5.1 | 5.1 | 5.1 | 5.3 | 5.4 | 5.7 | 6.4 | 7 |
| Y | 2.3 | 3 | 3.1 | 3.2 | 3.2 | 3.3 | 3.5 | 3.5 | 3.7 | 3.8 | 3.9 | 4.4 |

**\*Các giá trị trung bình (mean):**

**Công thức: mean =  =**

meanX = (4.5 + 4.6 + 4.9 + … + 7) / 12 = 5.34

meanY = (2.3 + 3 + 3.1 + … + 4.4) / 12 = 5.34

**\*Trung vị (median):**

Là điểm ở giữa của tập dữ liệu đã được sắp xếp từ bé đến lớn

Cách tính trung vị: xét trên tập có n phần tử

* **Nếu n lẻ: median = giá trị tại vị trí**
* **Nếu n chẵn : median = trung bình giá trị tại vị trí và**

Vị trí medianX và Y **=** 12/2 = 6 => Trung vị median là trung bình của giá trị tại vị trí 6 và 7.

* medianX = = 5.1
* medianY = = 5.1

**\*Mode: là giá trị xuất hiện nhiều nhất trong tập dữ liệu**

modeX = 5.1 và modeY = 3.2 và 3.5

**\*Ngoài ra có bài còn yêu cầu tính độ lệch chuẩn**

Độ lệch chuẩn () là căn bậc 2 của phương sai

Công thức phương sai: **2 =** **i2** ) - **2**

1. Để vẽ được biểu đồ Boxplot ta cần phải có 5 giá trị: min, max, Q1, Q2, Q3.

* Biểu đồ Boxplot của X:

\***Q1 (X) =** = = 3.25 (vị trí)

-> Q1 (X) = vị trí 3 + 0.25 \* (vị trí 4 – vị trí 3)

= 4.9 + 0.25 \* (5 – 4.9) = 4.925

\***Q3 (X) =3\*(** = 3\*3.25 = 9.75 (vị trí)

-> Q3 (X) = vị trí 9 + 0.75 \* (vị trí 10 – vị trí 9)

= 5.4 + 0.75 \* (5.7 – 5.4) = 5.625

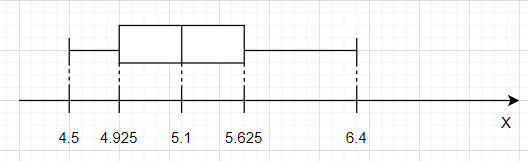
**\*IQR(X) = Q3 (X) – Q1 (X)** = 5.625 – 4.925 = 0.7

**\*[Q1 (X) – 1.5 \* IQR(X) ; Q3 (X) + 1.5 \* IQR(X) ]**

= [ 4.925 – 1.5\*0.7 ; 5.625 + 1.5\*0.7 ] = [3.875 ; 6.675] (\*\*)

* Trong tập dữ liệu của X có 1 phần tử ngoại lai là 7 vì không thuộc (\*\*)

Bộ 5 số của X: min = 4.5 ; max = 6.4 ; Q1 (X) = 4.925; Q2 (X) = median(x) = 5.1; Q3 (X) = 5.625



* Biểu đồ Boxplot của Y:

\***Q1 (Y) =** = = 3.25 (vị trí)

-> Q1 (Y) = vị trí 3 + 0.25 \* (vị trí 4 – vị trí 3)

= 3.1 + 0.25 \* (3.2 – 3.1) = 3.125

\***Q3 (Y) =3\*(** = 3\*3.25 = 9.75 (vị trí)

-> Q3 (Y) = vị trí 9 + 0.75 \* (vị trí 10 – vị trí 9)

= 3.7 + 0.75 \* (3.8 – 3.7) = 3.775

**\*IQR(Y) = Q3 (Y) – Q1 (Y)** = 3.775 – 3.125 = 0.65

**\*[Q1 (Y) – 1.5 \* IQR(Y) ; Q3 (Y) + 1.5 \* IQR(Y) ]**

= [ 3.125 – 1.5\*0.65 ; 3.775 + 1.5\*0.65 ] = [2.15 ; 4.75]

* Trong tập dữ liệu của Y không có phần tử ngoại lai nào.

Bộ 5 số của Y: min = 2.3 ; max = 4.4 ; Q1 (Y) = 3.125; Q2 (Y) = median(Y) = 3.4; Q3 (Y) = 3.775

\*Công thức chuẩn hoá min-max:

**vi’ = \* (new\_maxA – new\_minA) + new\_minA**

* Chuẩn hoá min-max của X về đoạn [1; 5]

xi’ = \* (new\_maxX – new\_minX) + new\_minX

* x1’ = \* (5-1) + 1 = 1
* Tương tự tính được x2’ = 1.16; x3’ = 1.64; x4’ = 1.8; x5’ = 1.96; x6’ = 1.96; x7’ = 1.96; x8’ = 2.28; x9’ = 2.44; x10’ = 2.92; x11’ = 4.04; x12’ = 5
* Chuẩn hoá min-max của Y về đoạn [1; 5]

yi’ = \* (new\_maxY – new\_miny) + new\_minY

* y1’ = \* (5-1) + 1 = 1
* Tương tự tính được y2’ = 2.3; y3’ = 2.5; y4’ = 2.7; y5’ = 2.7; y6’ = 2.9; y7’ = 3.3; y8’ = 3.3; y9’ = 3.7; y10’ = 3.9; y11’ = 4.05; y12’ = 5

\*Ngoài ra, có bài yêu cầu tính theo chuẩn Z-Score, tỷ lệ thập phân.

**\*Công thức chuẩn hoá Z-Score:**

**vi’ =**

**SA = \* => vi’ =**

**\*Tỷ lệ thập phân:**

**vi’ =**

**max ( < 1; j là số nguyên nhỏ nhất**

* **max () < 1**
* **max < 10j**

d)

\*Phương pháp làm mịn dữ liệu = Binning

B1: Sắp xếp dữ liệu

B2: Chia các phần tử vào các bin

* Độ sâu của bin => Số bin
* Từ số bin => số phần tử được phân vào mỗi bin

B3: Tính trung bình/trung vị/biên của mỗi bin

B4: Thay thế các giá trị của bin = trung bình/trung vị/biên.

* Bin của X:

Bin1: 4.5; 4.6; 4.9; 5

Bin2: 5.1; 5.1; 5.1; 5.3

Bin3: 5.4; 5.7; 6.4; 7

* Bin means
  + Bin1: 4.75; 4.75; 4.75; 4.75
  + Bin2: 5.15; 5.15; 5.15; 5.15
  + Bin3: 6.125; 6.125; 6.125; 6.125
* Bin của Y:

Bin1: 2.3; 3; 3.1; 3.2

Bin2: 3.2; 3.3; 3.5; 3.5

Bin3: 3.7; 3.8; 3.9; 4.4

* Bin means
  + Bin1: 2.9; 2.9; 2.9; 2.9
  + Bin2: 3.375; 3.375; 3.375; 3.375
  + Bin3: 3.95; 3.95; 3.95; 3.95

e)

**Hệ số tương quan:**

**rA,B = ; rA,B [-1, 1]**

* 2 (X) = \* (4.52 + 4.62 + 4.92 +52 +…+ 72) – 5.342 = 0.4969

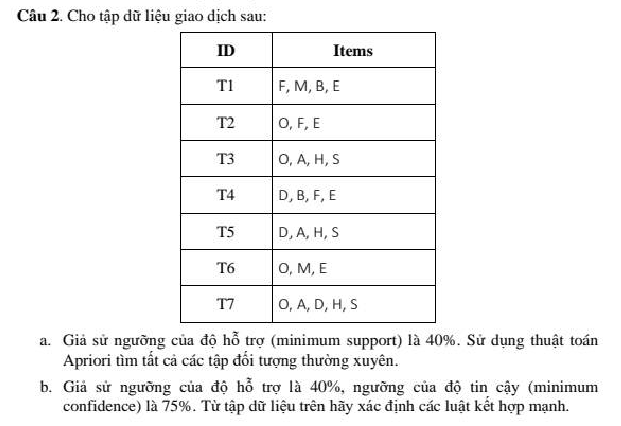
(X) = 0.7

* 2 (Y) = \* (2.32 + 32 + 3.12 + 3.22 +…+ 4.42) – 3.42 = 0.3125

(Y) = 0.56

* rX,Y = = 0.94 [-1, 1]

Câu2: (có 4 dạng phải ôn là Apriori, K-means, thuật toán ID3, tính khoảng cách. Câu2 trong đề thi mẫu này mới có 1 dạng là Apriori)



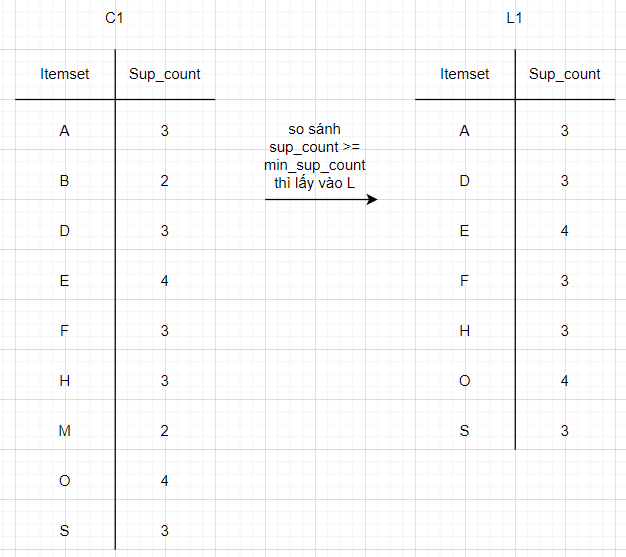
**Cách làm:**

1. yêu cầu tìm tất cả các tập thường xuyên (phổ biến) chính là các tập L

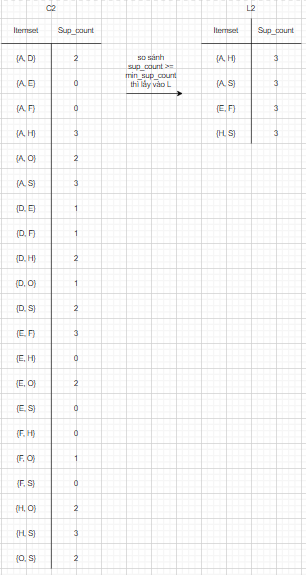
**Công thức: min\_sup\_count =**  (chú ý kq làm tròn lên)

* min\_sup\_count = = 3

B1: Đếm các tập có 1 phần tử là tập là tập C1( tập ứng viên). Sau đó so sánh sup\_count >= min\_sup\_count thì lấy vào tập L (tập phổ biến)



B2: Đếm tiếp các tập có 2 phần tử , trong đó các tập 2 phần tử được tạo ra từ tập L1. Sau đó cũng so sánh sup\_count >= min\_sup\_count thì lấy vào tập L2 (tập phổ biến)



B3: cũng làm tương tự như trên

C3 -> L3

{A, H, S} (sup\_count=3)

1. min\_conf = 75%

|  |  |  |
| --- | --- | --- |
| Tập phổ biến | Luật kết hợp | Độ tin cậy |
| {A, H} | A -> H | 3/3 |
|  | H -> A | 3/3 |
| {A, S} | A-> S | 3/3 |
|  | S -> A | 3/3 |
| {E, F} | E -> F | 3/4 |
|  | F -> E | 3/3 |
| {H, S} | H -> S | 3/3 |
|  | S -> H | 3/3 |
| {A, H, S} | A, H -> S | 3/3 |
|  | A, S -> H | 3/3 |
|  | H, S -> A | 3/3 |
|  | A -> H, S | 3/3 |
|  | H -> A, S | 3/3 |
|  | S -> A, H | 3/3 |

**Để tính độ tin cậy như trên ta dùng công thức sau:**

**Confidence (A->B) =**

**Sau đó, so sánh độ tin cậy >= min\_conf thì lấy, nếu < thì loại.**

**Các dạng còn lại của câu 2**

**\*Thuật toán K-Means**

VD: Cho tập dữ liệu sau:

|  |  |  |
| --- | --- | --- |
|  | X1 | X2 |
| A | 1 | 2 |
| B | 2 | 2 |
| C | 2 | 3 |
| D | 3 | 3 |
| E | 3 | 4 |
| F | 2 | 4 |

Tiến hành phân cụm tập dữ liệu trên với k=2 với tâm khởi tạo cho 2 cụm là A và F.

**Cách làm:**

**B1:** Ta tính khoảng cách từng tâm đến tất cả các điểm có trong tập dữ liệu cho sẵn. Sau đó, phân các điểm của tập dữ liệu vào các cụm. Có nghĩa là điểm gần với tâm nào hơn thì thuộc cụm của tâm ấy.

Công thức tính khoảng cách:

**d(i,j) =**

Sau đó, cập nhập các điểm nằm trong các các cụm và tâm mới của các cụm sẽ có tạo độ bằng trung bình toạ độ các điểm nằm trong cụm.

**B2:** Ta lại làm tương tự như trên. Đến khi nào mà ta thấy sau 2 lần tính mà các điểm nằm trong các cụm vẫn giống như lần tính tính trước, có nghĩa là các điểm vẫn nằm trong các cụm đó không có sự thay đổi sang cụm khác, thì ta dừng.

**Giải:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | T1(1,2) | T2(2,4) | Cụm |
| A | 0 | 2.236 | 1 |
| B | 1 | 2 | 1 |
| C | 1.414 | 1 | 2 |
| D | 2.236 | 1.414 | 2 |
| E | 2.828 | 1 | 2 |
| F | 2.236 | 0 | 2 |

* Cụm 1 có các điểm là A(1,2) và B(2,2)

Tâm mới T1 = () = (1.5; 2)

* Cụm 2 có các điểm là C(2,3) , D(3,3) , E(3,4) , F(2, 4)

Tâm mới T2 = () = (2.5; 3.5)

|  |  |  |  |
| --- | --- | --- | --- |
|  | T1(1.5, 2) | T2( 2.5, 3.5) | Cụm |
| A | 0.5 | 2.121 | 1 |
| B | 0.5 | 1.581 | 1 |
| C | 1.12 | 0.707 | 2 |
| D | 1.803 | 0.707 | 2 |
| E | 2.5 | 0.707 | 2 |
| F | 2.062 | 0.707 | 2 |

(Ta thấy các điểm vẫn nằm ở cụm cũ nên đến đây ta dừng)

* Cụm 1 có các điểm là A(1,2) và B(2,2)

Tâm mới T1 = (1.5; 2)

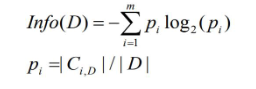
* Cụm 2 có các điểm là C(2,3) , D(3,3) , E(3,4) , F(2, 4)

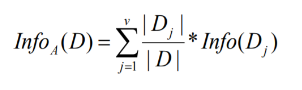
Tâm mới T2 = (2.5; 3.5)

**\*Thuật toán ID3**

**Áp dụng 3 công thức sau để tính cho các thuộc tính có trong bài:**

* **Entropy (Độ hỗn tạp của 1 tập):**

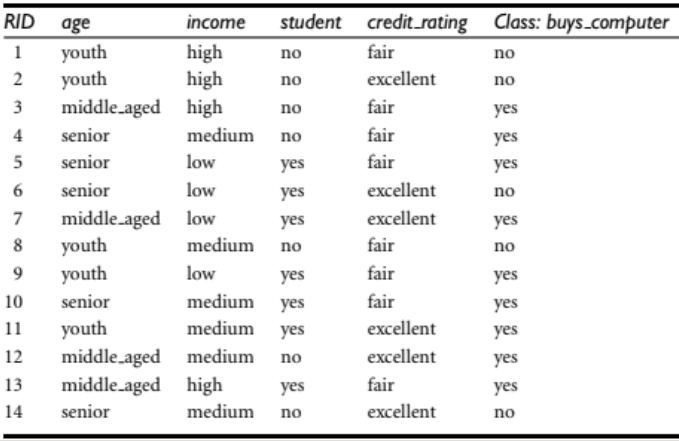


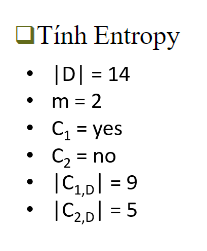


* **Information gain**

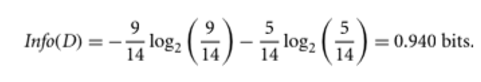


VD:

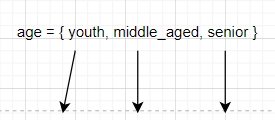




Giải:



\***Đối với thuộc tính “age”:**



=5 =4 =5 (có đc như này là ta đếm trong bảng trên)

Info(D1) = \* \* = 0.97 (bits)

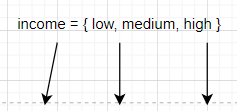
Info(D2) = \* - 0 = 0 (bits)

Info(D3) = \* \* = 0.97 (bits)

Infoage(D) = \*0.97 + \*0 + \*0.97 = 0.69 (bits)

Gain(age) = Info(D) – Infoage(D) = 0.94 – 0.69 = 0.25 (bits)

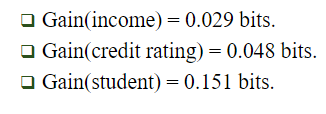
\*Đối với thuộc tính income:



=4 =6 =4

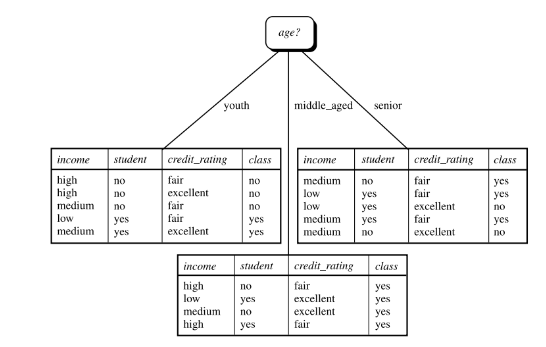
Cũng tính tương tự cái age kia, phải tính info D1 D2 D3, D và gain….Lười quá.

Phải tính hết cho tất cả các thuộc tính có trong bài. Ơ bài này có 4 thuộc tính: age, income, student, credit\_rating. Phải tính tất!



Sau khi tính xong hết. vẫn chưa xong đâu. Phải vẽ cái cây quyết định nứa =))

B1: từ cái bảng đề bài ấy, mình phân nó ra theo thuộc tính đầu là age này.



B2: tối giản cái bảng lại để ra cái cây. \*-\*

