**Bài tập về tiền xử lý dữ liệu**

Bài 1. Tiền xử lý tập dữ liệu sau sử dụng kỹ thuật binning:  
[16 17 20 21 22 23 23 25 28 28 33 34 36 37]

B1: sắp xếp dữ liệu:

[16 17 20 21 22 23 23 25 28 28 33 34 36 37]

B2: chia về các Bin:

bin1 = [16, 17, 20, 21]  
bin2 = [22, 23, 23, 25]  
bin3 = [28, 28, 33, 34, 36, 37]

1. Giảm nhiễu dữ liệu theo giá trị max.

B3: xác định giá trị max

bin1 = 21  
bin2 = 25  
bin3 = 37

B4:

bin1 = [21, 21, 21, 21]  
bin2 = [25, 25, 25, 25]  
bin3 = [37, 37, 37, 37, 37, 37]

1. Giảm nhiễu dữ liệu theo phần nguyên của giá trị trung bình.

B3: xác định giá trị phần nguyên của giá trị trung bình.

bin1 = 18  
bin2 = 23  
bin3 = 32

B4:

bin1 = [18, 18, 18, 18]  
bin2 = [23, 23, 23, 23]  
bin3 = [32, 32, 32, 32, 32, 32]

1. Giảm nhiễu dữ liệu theo giá trị biên.

Bài 2. Nếu một thuộc tính trong nguồn dữ liệu Điểm-Sinh viên có các giá trị A, B, C, D, F  
thì kiểu dữ liệu dự kiến của thuộc tính đó trong quá trình tiền xử lý là gì?

* kiểu dữ liệu dự kiến của thuộc tính đó trong quá trình tiền xử lý là kiểu dữ liệu rời rạc (categorical data). Điểm số của sinh viên được đánh giá theo hệ thống chữ cái, không phải ở dạng số liên tục, do đó không thể xử lý trực tiếp bằng các phép tính toán số học như cộng, trừ, nhân hay chia. Thay vào đó, ta có thể sử dụng các kỹ thuật xử lý dữ liệu rời rạc như One-Hot Encoding hoặc Label Encoding để biến đổi dữ liệu thành các giá trị số hoặc vector đại diện cho các giá trị rời rạc đó, và từ đó sử dụng trong các bước tiền xử lý và mô hình hóa dữ liệu.

Bài 3. Cho mảng một chiều X = {-5.0, 23.0, 17.6, 7.23, 1.11}, hãy chuẩn hóa mảng sử dụng.  
a. Decimal scaling: trong khoảng [-1, 1].

|  |  |
| --- | --- |
| x’ =  j là số nguyên nhỏ nhất sao cho: min(|x’|) >= -1 và max(|x’|) <= 1 | J = 2  x’ = |
| x | x’ |
| -5.0 | -0.05 |
| 23.0 | 0.23 |
| 17.6 | 0.176 |
| 7.23 | 0.0723 |
| 1.11 | 0.0111 |

b. Min-max: trong khoảng [0, 1].

|  |  |
| --- | --- |
| x ∈ [minX, maxX], x’ ∈ [new\_minX, new\_maxX]  x’ = | x ∈ [-5,23], x’ ∈ [0,1]  x’ = |
| x | x’ |
| -5.0 | = 0 |
| 23.0 | = 1 |
| 17.6 | = 0.807143 |
| 7.23 | = 0.436786 |
| 1.11 | = 0.218214 |

c. Min-max: trong khoảng [-1, 1].

|  |  |
| --- | --- |
| x ∈ [minX, maxX], x’ ∈ [new\_minX, new\_maxX]  x’ = | x ∈ [-5,23], x’ ∈ [-1,1]  x’ = |
| x | x’ |
| -5.0 | = -1 |
| 23.0 | = 1 |
| 17.6 | = 0.614286 |
| 7.23 | = - 0.12643 |
| 1.11 | = - 0.56357 |

d. Phương pháp độ lệch. (Z-Score Normalization)

|  |  |
| --- | --- |
| = (mean)  x’ = | n = 5  = = 8.788  10.3065  x’ = |
| x | x’ |
| -5.0 | -1.3378 |
| 23.0 | 1.378936 |
| 17.6 | 0.854994 |
| 7.23 | - 0.15117 |
| 1.11 | - 0.74497 |

e. So sánh kết quả của các dạng chuẩn trên và cho nhận xét về ưu nhược điểm  
của các phương pháp?

|  |  |  |
| --- | --- | --- |
|  | Ưu điểm | Nhược điểm |
| Phương pháp decimal scaling | giá trị chuẩn hóa thuộc vào khoảng [-1, 1], dễ hiểu và áp dụng | nếu giá trị lớn nhất hoặc nhỏ nhất trong mảng X quá lớn hoặc quá nhỏ, các giá trị chuẩn hóa có thể bị mất độ chính xác. |
| Phương pháp min-max | giá trị chuẩn hóa thuộc vào khoảng [0, 1], dễ hiểu và áp dụng | nếu giá trị lớn nhất hoặc nhỏ nhất trong mảng X quá lớn hoặc quá nhỏ, các giá trị chuẩn hóa có thể bị mất độ chính xác. |
| Phương pháp độ lệch | giá trị chuẩn hóa không bị giới hạn trong khoảng cố định, giá trị chuẩn hóa phản ánh được mức độ chênh lệch giữa các phần tử trong mảng X | phương pháp này yêu cầu tính toán giá trị trung bình và độ lệch chuẩn của mảng X, do đó cần nhiều thao tác tính toán hơn so với hai phương pháp trước đó. |

**Bài 4.** Làm mịn dữ liệu sử dụng kỹ thuật làm tròn cho tập sau:

Y = {1.17, 2.59, 3.38, 4.23, 2.67, 1.73, 2.53, 3.28, 3.44}  
Sau đó biểu diễn tập thu được với các độ chính xác:  
a. 0.1

{1.2, 2.6, 3.4, 4.2, 2.7, 1.7, 2.5, 3.3, 3.4}  
b. 1.0

{1, 3, 3, 4, 3, 2, 3, 3, 3}

**Bài 5.** Cho tập mẫu với các giá trị bị thiếu  
o X1 = {0, 1, 1, 2}  
o X2 = {2, 1, -, 1}  
o X3 = {1, -, -, 0}  
o X4 = {-, 2, 1, -}  
Nếu miền xác định của tất cả các thuộc tính là [0, 1, 2], hãy xác định các giá trị bị thiếu  
biết rằng các giá trị đó có thể là một trong số các xác trị của miền xác định? Hãy giải thích  
những cái được và mất nếu rút gọn chiều của kho dữ liệu lớn trong quá trình tiền xử lý dữ  
liệu?