Chương 2: Getting to know your data

|  |  |  |
| --- | --- | --- |

|  |  |  |
| --- | --- | --- |

|  |  |  |
| --- | --- | --- |

# Đề thi mẫu

| Câu 1. Cho tập dữ liệu thu thập được của 2 tập giá trị của X và Y như sau:   | X | 5.1 | 4.9 | 4.6 | 5.7 | 5.4 | 5.1 | 5.1 | 5.3 | 5 | 7 | 6.4 | 4.5 | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Y | 3.5 | 3 | 3.1 | 4.4 | 3.9 | 3.5 | 3.8 | 3.7 | 3.3 | 3.2 | 3.2 | 2.3 |  1. Hãy xác định các giá trị trung bình, trung vị, mode của X và Y. 2. Vẽ biểu đồ Boxplot của X và Y. 3. Sử dụng phương pháp chuẩn hóa min-max để chuẩn hóa dữ liệu quan sát của X và Y về đoạn [1, 5]. 4. Hãy làm trơn dữ liệu ban đầu của X và Y bằng phương pháp làm trơn trung bình (bin means), trong đó việc phân chia thùng theo chiều sâu (Equal-depth) với số bin là 3. Mô tả các bước thực hiện. 5. Xác định hệ số tương quan giữa X và Y. | Câu 2. Cho tập dữ liệu giao dịch sau:   | ID | Items | | --- | --- | | T1 | F, M, B, E | | T2 | O, F, E | | T3 | O, A, H, S | | T4 | D, B, F, E | | T5 | D, A, H, S | | T6 | O, M, E | | T7 | O, A, D, H, S |  1. Giả sử ngưỡng của độ hỗ trợ (minimum support) là 40%. Sử dụng thuật toán Apriori tìm tất cả các tập đối tượng thường xuyên. 2. Giả sử ngưỡng của độ hỗ trợ là 40%, ngưỡng của độ tin cậy (minimum confidence) là 75%. Từ tập dữ liệu trên hãy xác định các luật kết hợp mạnh. |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

# 

| **Câu 1**  a.  - Giá trị trung bình  - Giá trị trung vị  Sắp xếp 2 tập dữ liệu, ta được:   | 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 |   - Xác định Mode của 2 tập giá trị  + ModeX = 5.1 với 3 lần xuất hiện  + ModeY = 3.2, 3.5 với 2 lần xuất hiện  b.  Vẽ biểu đồ boxplot cho tập X:  Giá trị khoảng ngoại lai:  =  = [4.95 - 1.5\*0.6; 5.55 + 1.5\*0.6]  = [4.05; 6.45]  → có 1 giá trị ngoại lai là 7    Vẽ biểu đồ boxplot cho tập Y:  Giá trị khoảng ngoại lai:  =  = [3.15 - 1.5\*0.6; 3.75+ 1.5\*0.6]  = [2.25; 4.65]  → Không có giá trị ngoại lai | c.  Ta sử dụng công thức để chuẩn hóa bằng phương pháp min-max  Chuẩn hóa dữ liệu tập X  →   |  |  | | --- | --- |   Tương tự, ta có kết quả chuẩn hóa dữ liệu tập Y   |  |  | | --- | --- |   d. Với số bin là 3 ta chia được mỗi bin sẽ gồm 4 phần tử  Làm trơn dữ liệu cho tập X  + **bin 1**: 4.5, 4.6, 4.9, 5  + **bin 2**: 5.1, 5.1, 5.1, 5.3  + **bin 3**: 5.4, 5.7, 6.4, 7  Bằng phương pháp làm trơn trung bình, ta được:  + **bin 1**: 4.75, 4.75, 4.75, 4.75  + **bin 2**: 5.15, 5.15, 5.15, 5.15  + **bin 3**: 6.125, 6.125, 6.125, 6.125  Làm trơn dữ liệu cho tập Y  + **bin 1**: 2.3, 3, 3.1, 3.2  + **bin 2**: 3.2, 3.3, 3.5, 3.5  + **bin 3**: 3.7, 3.8, 3.9, 4.4  Bằng phương pháp làm trơn trung bình, ta được:  + **bin 1**: 2.9, 2.9, 2.9, 2.9  + **bin 2**: 3.375, 3.375, 3.375, 3.375  + **bin 3**: 3.95, 3.95, 3.95, 3.95  e. Tính hệ số tương quan giữa X và Y |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Câu 2**  a, Ta có:  min\_sup\_count = [40% \* 7] = [2.8] = 3   | C1 | | → | L1 | | | --- | --- | --- | --- | --- | | Item set | supcount | Item set | supcount | | A | 3 | A | 3 | | B | 2 | D | 3 | | D | 3 | E | 4 | | E | 4 | F | 3 | | F | 3 | H | 3 | | H | 3 | O | 4 | | M | 2 | S | 3 | | O | 4 |  |  |  | | S | 3 |  |  |  |  | C2 | | → | L2 | | | --- | --- | --- | --- | --- | | Item set | sup count | Item set | sup count | | {A, D} | 2 | {A, H} | 3 | | {A, E} | 0 | {A, S} | 3 | | {A, F} | 0 | {E, F} | 3 | | {A, H} | 3 | {H, S} | 3 | | {A, O} | 2 |  |  | | {A, S} | 3 |  |  | | {D, E} | 1 |  |  | | {D, F} | 1 |  |  | | {D, H} | 2 |  |  | | {D, O} | 1 |  |  | | {D, S} | 2 |  |  | | {E, F} | 3 |  |  | | {E, H} | 0 |  |  | | {E, O} | 2 |  |  | | {E, S} | 0 |  |  | | {F, H} | 0 |  |  | | {F, O} | 0 |  |  | | {F, S} | 0 |  |  | | {H, O} | 2 |  |  | | {H, S} | 3 |  |  | | {O, S} | 2 |  |  |  | C3 | |  | L3 | | | --- | --- | --- | --- | --- | | Item set | sup count | → | Item set | sup count | | {A, H, S} | 3 | {A, H, S} | 3 |   ⇒ Tập các đối tượng thường xuyên là  {A, D, E, F, H, O, S, AH, AS, EF, HS, AHS} | b,  min\_conf = 75%   | Tập phổ biến | Luật kết hợp | Độ tin cậy | | --- | --- | --- | | {E, F} | E → F | 3/4 | |  | F → E | 3/3 | | {H, S} | H → S | 3/3 | |  | S → H | 3/3 | | {A, H} | A → H | 3/3 | |  | H → A | 3/3 | | {A, S} | A → S | 3/3 | |  | S → A | 3/3 | | {A, H, S} | {A, H} → S | 3/3 | |  | {A, S} → H | 3/3 | |  | {H, S} → A | 3/3 | |  | S → {A, H} | 3/3 | |  | H -> {A, S} | 3/3 | |  | A → {H, S} | 3/3 | |

# 

|  |  |
| --- | --- |

# 

# 

|  |  |  |
| --- | --- | --- |

|  |  |  |
| --- | --- | --- |