Bài tập

Câu hỏi 1:

Giả sử một kho dữ liệu bao gồm 3 chiều: thời gian, bác sĩ và bệnh nhân, và hai độ đo là số đếm và chi phí với chi phí là tổng số tiền mà một bệnh nhân phải trả cho một bác sĩ trong một lần khám bệnh.

(a) Hãy liệt kê ba loại lược đồ dữ liệu có thể được sử dụng để mô hình hóa kho dữ liệu và sử dụng một trong số chúng để vẽ lược đồ dữ liệu cho kho dữ liệu mô tả ở trên.

(b) Xuất phát từ khối cơ bản [ngày, bác sĩ, bệnh nhân], những thao tác OLAP nào cần được thực hiện để liệt kê tổng số tiền kiếm được của mỗi bác sỹ trong năm 2022?

(c) Viết các câu lệnh DMQL để định nghĩa lược đồ dữ liệu trong câu (a) và câu lệnh SQL để tính tổng tiền trong câu (b)

Câu hỏi 2:

Một cơ sở dữ liệu có 5 giao dịch như dưới đây. Giả sử minsup = 60% và minconf = 80%.

Mã giao dịch	**Mặt hàng mua**
T100	A, B, C, D, E
T200	B, E, A, C, H
`T300	P, A, G, E
T400	U, R, B, A, N, A
T500	G, A, U, D, I

a) Hãy tìm tất cả những tập mặt hàng thường xuyên sử dụng thuật toán Apriori và FP-growth. So sánh tính hiệu quả của hai phương pháp khai phá trên.

b) Hãy liệt kê tất cả những luật kết hợp (với độ hỗ trợ s và độ tin cậy c) tuân thủ theo siêu luật được mô tả dưới đây, trong đó *X* là một biến đại diện cho khách hàng và *mhi* thể hiện các biến đại điện cho các mặt hàng (ví dụ "A", "B", v.v...): cho tất cả x trong giao dịch; mua(X, mh1) → mua(X, mh2) [s, c]

Câu hỏi 3:

Giả sử một bệnh viện thực hiện việc lấy dữ liệu về tuổi và độ béo của 18 bệnh nhân với kết quả như sau

Tuổi	23	23	27	27	39	41	47	49	50
Độ béo	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
Tuổi	52	54	54	56	57	58	58	60	61
Độ béo	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

a) Hãy tính giá trị trung bình, giá trị trung vị và phương sai chuẩn của tuổi và độ béo;

b) Chuẩn hóa hai biến này dựa trên chuẩn hóa z-score;

Câu hỏi 4:

Bảng sau đây chứa các thuộc tính bao gồm: Tên, Giới tính, Xét nghiệm (XN) 1, Xét nghiệm 2, Xét nghiệm 3, Xét nghiệm 4 trong đó Tên là một định danh của một đối tượng, giới tính là một thuộc tính đối xứng, các thuộc tính còn lại đều là không đối xứng mô tả kết quả xét nghiệm của mỗi cá nhân,…Giả sử rằng một dịch vụ tồn tại để xác định xem cặp nào tương thích với nhau. Với giá trị của thuộc tính không đối, gán giá trị dương tính là 1 và giá trị âm tính là 0. Giả sử thêm rằng khoảng cách giữa hai đối tượng được tính chỉ dựa vào các thuộc tính không đối xứng.

Tên	Giới tính	XN-1	XN-2	XN-3	XN-4
Kiên	nam	-	+	+	-
Châu	nữ	-	+	+	+
Bình	nam	+	-	-	+

(a) Tính hệ số Jaccard cho mỗi cặp.

(b) Cặp nào là tương thích với nhau nhất, cặp nào ít phù hợp với nhau nhất?

Câu hỏi 5:

Bảng dưới đây chứa dữ liệu huấn luyện từ một cơ sở dữ liệu của nhân viên. Các dữ liệu được tổng quát hóa. Ví dụ “31 : : : 35” cho thuộc tính *tuổi* thể hiện phạm vi tuổi trong khoảng từ 31

đến 35. Với một hàng dữ liệu, *số bản ghi* thể hiện số lượng bộ dữ liệu có các giá trị của các thuộc tính *phòng, trạng thái, tuổi, lương* xác định của mỗi dòng.

Phòng	**Trạng thái**	**tuổi**	**lượng**	**số bản ghi**
Bán hàng	lâu năm	31. . .35	46K.. .50K	30
Bán hàng	mới vào	26. . .30	26K.. .30K	40
Bán hàng	mới vào	31. . .35	31K.. .35K	40
Hệ thống	mới vào	21. . .25	46K.. .50K	20
Hệ thống	lâu năm	31. . .35	66K.. .70K	5
Hệ thống	mới vào	26. . .30	46K.. .50K	3
Hệ thống	lâu năm	41. . .45	66K.. .70K	3
Quảng cáo	lâu năm	36. . .40	46K.. .50K	10
Quảng cáo	mới vào	31. . .35	41K.. .45K	4
Thư ký	lâu năm	46. . .50	36K.. .40K	4
Thư ký	mới vào	26. . .30	26K.. .30K	6

Đặt **Trạng thái** là thuộc tính nhãn phân lớp.

Sử dụng thuật toán xây dựng cây quyết định cho dữ liệu được cho ở trên.

Câu hỏi 6:

Cho một tập gồm 14 điểm như sau:

(0, 1), (1, 2), (2, 1), (2, 2), (2, 3), (3, 2), (3, 5), (4, 3), (4, 5), (5, 4), (5, 5), (6, 3), (6, 4), (6, 5)

Hãy liệt kê thuật toán k-means sẽ sinh ra những cụm nào cho tập điểm trên với trường hợp giá trị k = 2.