TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ THI MÔN **KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**
LỚP CAO HỌC – HỆ THỐNG THÔNG TIN
THỜI GIAN LÀM BÀI: 150 phút
ĐƯỢC SỬ DỤNG TÀI LIỆU

Câu 1:

1.1 Cho trước danh sách các tập phổ biến (FIs) cùng với độ phổ biến (support) của chúng. Trình bày thuật toán tìm tất cả các luật kết hợp thỏa ngưỡng *minConf*. Anh/chị hãy cho biết độ phức tạp của thuật toán tương ứng theo |FIs|.

1.2 Cho CSDL giao tác dưới dạng nhị phân như sau:

Mã giao dịch	A	B	C	D	E	F
1	1	1	0	1	1	0
2	0	1	1	0	1	0
3	1	1	0	1	1	1
4	1	1	1	0	1	0
5	0	1	1	1	0	1
6	1	1	1	1	1	0

a) Tìm tất cả các tập phổ biến có trong CSDL với *minSup* = 50% theo phương pháp FP-Tree (hoặc IT-Tree)

b) Tìm tất cả các luật kết hợp với *minConf* = 80%

Câu 2:

2.1 Trình bày ngắn gọn thuật toán k-means.

2.2 Một xe đón khách về bến xe Miền Đông của công ty Mai Linh muốn đón n khách hàng. Do thời gian đón khách ít nên công ty muốn gom khách về k địa điểm để tiện việc đón. Giả sử n = 5 và k = 2. 5 khách hàng đang ở các tọa độ A(1,1), B(3,1), C(3,3), D(4,2), E(1,3). Anh/chị hãy cho biết nên hẹn khách nào tại địa điểm nào để việc đưa đón là thuận tiện nhất. Cho biết tọa độ của 2 địa điểm cần đón khách? Giả sử độ đo khoảng cách được sử dụng là độ đo Euclidean.

Câu 3:

Cho CSDL sau:

ID	Outlook	Tempurature	Humidity	Windy	Class
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rain	Mild	High	False	Yes
5	Rain	Cold	Normal	False	Yes
6	Rain	Cold	Normal	True	No
7	Overcast	Cold	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cold	Normal	False	Yes
10	Rain	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rain	Mild	High	True	No
15	Overcast	Mild	Normal	False	?
16	Rain	Hot	Normal	True	?

3.1. Sử dụng độ đo sau, tìm các luật phân lớp với cột quyết định là Class.

Độ đo Information Gain (IG):

$$Gain(S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Trong đó:

- Value(A) là tập tất cả các giá trị có thể có đối với thuộc tính A và S_v là tập con của S mà A có giá trị là v
- Với S bao gồm *c* lớp, thì Entropy của S được tính bằng công thức sau:

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Ở đây p_i là tỉ lệ của các mẫu thuộc lớp *i* trong tập S

Lưu ý: Chúng ta luôn chọn độ đo IG có giá trị lớn nhất

3.2. Cho biết lớp (Class) của mẫu 15, 16 dựa vào tập luật vừa tìm được.

3.3. Cho mẫu X = {Outlook = Rain, Tempurature = Hot, Humidity = Normal, Windy = False}. Dựa vào phương pháp Naïve Bayesian, tìm lớp của X.

- HẾT -