

ĐỀ 1

Đề thi gồm 2 trang

(Sinh viên được sử dụng 1 tờ A4 tài liệu)

Câu 1 (2.0 điểm) Chọn MỘT trong các câu sau:

1. Phân biệt thuộc tính rời rạc và thuộc tính liên tục. Cho ví dụ minh họa.
2. Nêu ưu điểm và hạn chế của cây quyết định.
3. Lấy một ứng dụng khai thác dữ liệu trong lĩnh vực giáo dục hoặc giao thông công cộng. Dựa vào ví dụ vừa chọn, hãy nêu đặc trưng của tập dữ liệu, và đề xuất thuật toán khai thác dữ liệu nên áp dụng.

Câu 2 (6.0 điểm)

Cho CSDL về *Thông tin giao hàng trễ hạn* của các doanh nghiệp vận chuyển hàng hóa vừa và nhỏ, chi tiết trong bảng sau. Ghi chú:

- Thuộc tính *Kết quả giao hàng (KQ)* là thuộc tính quyết định.
- Sinh viên có thể dùng từ viết tắt của thuộc tính trong khi làm bài.

	Xếp loại Shipper (XL)	Độ quan trọng món hàng (QT)	Được giảm giá ship (GG)	Kết quả giao hàng (KQ)
1	B	Thấp	Có	Trễ hạn
2	C	Trung bình	Có	Đúng hạn
3	C	Cao	Không	Đúng hạn
4	C	Thấp	Có	Trễ hạn
5	C	Trung bình	Không	Trễ hạn
6	A	Cao	Không	Đúng hạn
7	B	Trung bình	Có	Trễ hạn
8	A	Cao	Có	Đúng hạn
9	A	Trung bình	Không	Trễ hạn
10	A	Thấp	Có	Đúng hạn

- a. Tìm tập phổ biến có ngưỡng $minsup = 30\%$. Chọn 1 tập phổ biến tối đại để phát sinh luật, từ đó liệt kê 3 luật kết hợp thỏa $minconf = 60\%$. (2.0đ)
- b. Cho $B = \{\text{Độ quan trọng món hàng}, \text{Được giảm giá ship}\}$, $X = \{1, 4, 5, 7, 9\}$ (tập các mẫu có giá trị *Kết quả giao hàng* = “Trễ hạn”). Sử dụng tập thô tính: xấp xỉ trên, xấp xỉ dưới và hệ số xấp xỉ. (1.0đ)
- c. Xác định nút gốc của cây quyết định, sử dụng *Chỉ số Gini (Gini index)*. (1.5đ)
- d. Sử dụng công thức *Naïve Bayes* có làm trơn *Laplace* để phân lớp mẫu sau: (1.5đ)

$$X = \{XL = "B", QT = "Thấp", GG = "Không"\}$$

Câu 3 (2.0đ)

Cho 8 điểm như sau: $A1=(2,10)$, $A2=(2,5)$, $A3=(8,4)$, $A4=(5,8)$, $A5=(7,5)$, $A6=(6,4)$, $A7=(1,2)$, $A8=(4,9)$.
Với ma trận $U0$ được khởi tạo như sau:

U0	A1	A2	A3	A4	A5	A6	A7	A8
C1	1	0	0	0	0	0	0	0
C2	0	1	0	0	0	0	0	0
C3	0	0	1	1	1	1	1	1

Áp dụng thuật toán K_means và sử dụng độ đo Euclide để gom 8 điểm trên vào 3 cụm.

Lưu ý: chỉ thực hiện các bước

- Bước 1: tính trọng tâm cho các cụm
- Bước 2: so sánh khoảng cách điểm với trọng tâm từng cụm
- Bước 3: cập nhật ma trận $U1$

Đề thi gồm 2 trang

(Sinh viên được sử dụng 1 tờ A4 tài liệu)

Câu 1 (2.0 điểm) Chọn **MỘT** trong các câu sau:

1. Nêu ưu điểm và hạn chế của Support Vector Machine (Máy vector hỗ trợ).
2. Phân biệt thuộc tính rời rạc và thuộc tính liên tục. Cho ví dụ minh họa.
3. Lấy một ứng dụng khai thác dữ liệu trong lĩnh vực giáo dục hoặc giao thông công cộng. Dựa vào ví dụ vừa chọn, hãy nêu đặc trưng của tập dữ liệu, và đề xuất thuật toán khai thác dữ liệu nên áp dụng.

Câu 2 (6.0 điểm)

Cho CSDL về *Thông tin giao hàng trễ hạn* của các doanh nghiệp vận chuyển hàng hóa vừa và nhỏ, chi tiết trong bảng sau. Ghi chú:

- Thuộc tính *Kết quả giao hàng (KQ)* là thuộc tính quyết định.
- Sinh viên có thể dùng từ viết tắt của thuộc tính trong khi làm bài.

	Xếp loại Shipper (XL)	Độ quan trọng món hàng (QT)	Được giảm giá ship (GG)	Kết quả giao hàng (KQ)
1	C	Thấp	Không	Đúng hạn
2	B	Trung bình	Không	Trễ hạn
3	A	Trung bình	Không	Đúng hạn
4	A	Cao	Có	Trễ hạn
5	C	Trung bình	Không	Đúng hạn
6	A	Thấp	Có	Trễ hạn
7	C	Cao	Không	Đúng hạn
8	B	Thấp	Có	Đúng hạn
9	B	Thấp	Không	Trễ hạn
10	B	Cao	Có	Trễ hạn

- a. Tìm tập phổ biến có ngưỡng $minsup = 30\%$ và liệt kê 3 luật kết hợp thỏa $minconf = 95\%$. (2.0đ)
- b. Cho $B = \{\text{Độ quan trọng món hàng, Được giảm giá ship}\}$, $X = \{1, 3, 5, 7, 8\}$ (tập các mẫu có giá trị *Kết quả giao hàng* = “Đúng hạn”). Sử dụng tập thô tính: xấp xỉ trên, xấp xỉ dưới và hệ số xấp xỉ. (1.0đ)
- c. Xác định nút gốc của cây quyết định, sử dụng *Chỉ số Gini (Gini index)*. (1.5đ)
- d. Sử dụng công thức *Naïve Bayes* có làm tròn *Laplace* để phân lớp mẫu sau: (1.5đ)

$$X = \{XL = "C", QT = "Trung bình", GG = "Có"\}$$

Câu 3 (2.0đ)

Cho 8 điểm như sau: $A1=(10,10)$, $A2=(17,5)$, $A3=(8,4)$, $A4=(5,8)$, $A5=(7,5)$, $A6=(6,4)$, $A7=(1,2)$, $A8=(4,9)$.
Với ma trận $U0$ được khởi tạo như sau:

$U0$	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$	$A8$
$C1$	1	0	0	0	0	0	0	0
$C2$	0	1	0	0	0	0	0	0
$C3$	0	0	1	1	1	1	1	1

Áp dụng thuật toán K-means và sử dụng độ đo Euclide để gom 8 điểm trên vào 3 cụm.

Lưu ý: chỉ thực hiện các bước

- Bước 1: tính trọng tâm cho các cụm
- Bước 2: so sánh khoảng cách điểm với trọng tâm từng cụm
- Bước 3: cập nhật ma trận $U1$

Khoa Hệ thống Thông tin

TM. Giảng viên ra đề

Cao Thị Nhàn