

(Sinh viên được sử dụng 01 tờ A4 tài liệu, đề thi gồm 02 trang)

HỌ VÀ TÊN SV: MSSV: STT: PHÒNG THI:	<u>CÁN BỘ COI THI</u>
---	------------------------------

Câu 1 (2.0 điểm) Sinh viên chọn MỘT trong các câu sau:

1. Trình bày tóm tắt kiến trúc cơ bản của mạng neural nhân tạo. Nêu các ưu điểm và nhược điểm của mạng neural nhân tạo.
2. So sánh các điểm khác nhau giữa bài toán phân lớp và gom cụm. Cho ví dụ minh họa.
3. Lấy một ứng dụng khai thác dữ liệu trong lĩnh vực y tế hoặc quản lý nhân sự. 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 khởi hành các chuyến bay nội địa* của các hãng hàng không hoạt động tại Việt Nam xuất phát từ thành phố Hồ Chí Minh, chi tiết trong bảng sau.

	Hãng hàng không (HHK)	Điểm đến (DD)	Tháng bay (TB)	Giờ khởi hành (GKH)	Kết quả (KQ)
1	Vietjet	Đà Lạt	6.2023	17:25	Trễ giờ
2	Vietnam	Phú Quốc	5.2023	5:40	Đúng giờ
3	Vietjet	Hà Nội	5.2023	9:30	Trễ giờ
4	Pacific	Đà Nẵng	4.2023	10:10	Đúng giờ
5	Vietnam	Đà Lạt	6.2023	17:25	Đúng giờ
6	Vietjet	Phú Quốc	6.2023	9:30	Trễ giờ
7	Vietnam	Hà Nội	5.2023	10:10	Trễ giờ
8	Vietjet	Đà Nẵng	4.2023	9:30	Trễ giờ
9	Bamboo	Đà Lạt	6.2023	5:40	Đúng giờ
10	Pacific	Đà Nẵng	5.2023	17:25	Đúng giờ

Lưu ý:

- Thuộc tính *Kết quả (KQ)* là thuộc tính quyết định.
- Sinh viên có thể dùng từ viết tắt của thuộc tính trong khi làm bài.
- Làm tròn các số thập phân đến 04 chữ số thập phân.

1. Áp dụng thuật toán Apriori, tìm tập phổ biến thỏa ngưỡng $minsup = 25\%$. Chọn 1 tập phổ biến tối đại, liệt kê các luật kết hợp thỏa $minconf = 70\%$. **(1.75đ)**
2. Cho $B = \{\text{Điểm đến}, \text{Tháng bay}\}$, $X = \{1, 3, 6, 7, 8\}$ (tập các mẫu có giá trị $Kết\ quả = \text{“Trễ giờ”}$). Sử dụng tập thô tính: xấp xỉ trên, xấp xỉ dưới và hệ số xấp xỉ. **(1.0đ)**
3. Xác định nút gốc của cây quyết định, sử dụng *Chỉ số Gini*. **(1.75đ)**
4. Sử dụng công thức *Naïve Bayes* có làm tròn *Laplace* để phân lớp mẫu sau: **(1.5đ)**
 $X = \{\text{Hãng hàng không} = \text{“Vietjet”}, \text{Điểm đến} = \text{“Phủ Quốc”}, \text{Tháng bay} = \text{“6.2023”}, \text{Thời gian khởi hành} = \text{“9:30”}\}$

Câu 3 (2.0đ) Cho 7 điểm trong không gian 2 chiều như sau: $x_1 = \{3, 8\}$, $x_2 = \{2, 7.5\}$, $x_3 = \{3, 7\}$, $x_4 = \{4, 7\}$, $x_5 = \{8, 3\}$, $x_6 = \{7, 2.5\}$, $x_7 = \{8, 2\}$.

Với ma trận U_0 được khởi tạo như sau:

U_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
C1	1	0	0	0	0	0	0
C2	0	1	0	0	0	0	0
C3	0	0	1	1	1	1	1

Áp dụng thuật toán K-means và sử dụng độ đo Euclide để gom 7 điểm trên vào **3 cụm**.

Yêu cầu: Chỉ thực hiện các bước sau:

- *Bước 1: Tính trọng tâm cho các cụm*
- *Bước 2: So sánh khoảng cách điểm với trọng tâm từng cụm*
- *Bước 3: Xác định ma trận U_1*

Khoa/ Bộ môn duyệt đề

TM. Giảng viên ra đề

Bảng ma trận đáp ứng chuẩn đầu ra.

Câu hỏi	CĐRMH
1	G1
2	G2
3	G2

(Sinh viên được sử dụng 01 tờ A4 tài liệu, đề thi gồm 02 trang)

HỌ VÀ TÊN SV: MSSV: STT: PHÒNG THI:	<u>CÁN BỘ COI THI</u>
---	------------------------------

Câu 1 (2.0 điểm) Sinh viên chọn MỘT trong các câu sau:

1. Trình bày tóm tắt kiến trúc cơ bản của mạng neural nhân tạo. Nêu các ưu điểm và nhược điểm của mạng neural nhân tạo.
2. So sánh các điểm khác nhau giữa bài toán phân lớp và gom cụm. Cho ví dụ minh họa.
3. Lấy một ứng dụng khai thác dữ liệu trong lĩnh vực y tế hoặc quản lý nhân sự. 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 khởi hành các chuyến bay nội địa* của các hãng hàng không hoạt động tại Việt Nam xuất phát từ thành phố Hồ Chí Minh, chi tiết trong bảng sau.

	Hãng hàng không (HHK)	Điểm đến (DD)	Tháng bay (TB)	Giờ khởi hành (GKH)	Kết quả (KQ)
1	Vietnam	Phú Quốc	5.2023	5:40	Đúng giờ
2	Vietjet	Hà Nội	5.2023	9:30	Trễ giờ
3	Vietnam	Đà Lạt	6.2023	17:25	Đúng giờ
4	Vietnam	Hà Nội	5.2023	10:10	Trễ giờ
5	Pacific	Đà Nẵng	5.2023	17:25	Đúng giờ
6	Vietjet	Đà Lạt	6.2023	17:25	Trễ giờ
7	Pacific	Đà Nẵng	4.2023	10:10	Đúng giờ
8	Vietjet	Đà Nẵng	4.2023	9:30	Trễ giờ
9	Vietjet	Phú Quốc	6.2023	9:30	Trễ giờ
10	Bamboo	Đà Lạt	6.2023	5:40	Đúng giờ

Lưu ý:

- Thuộc tính *Kết quả (KQ)* là thuộc tính quyết định.
- Sinh viên có thể dùng từ viết tắt của thuộc tính trong khi làm bài.
- Làm tròn các số thập phân đến 04 chữ số thập phân.

1. Áp dụng thuật toán Apriori, tìm tập phổ biến thỏa ngưỡng $minsup = 25\%$. Chọn 1 tập phổ biến tối đại, liệt kê các luật kết hợp thỏa $minconf = 75\%$. **(1.75đ)**
2. Cho $B = \{\text{Điểm đến}, \text{Tháng bay}\}$, $X = \{1, 3, 5, 7, 10\}$ (tập các mẫu có giá trị $Kết\ quả = \text{“Đúng giờ”}$). Sử dụng tập thô tính: xấp xỉ trên, xấp xỉ dưới và hệ số xấp xỉ. **(1.0đ)**
3. Xác định nút gốc của cây quyết định, sử dụng *Chỉ số Gini*. **(1.75đ)**
4. Sử dụng công thức *Naïve Bayes* có làm trơn *Laplace* để phân lớp mẫu sau: **(1.5đ)**
 $X = \{\text{Hãng hàng không} = \text{“Vietnam”}, \text{Điểm đến} = \text{“Đà Lạt”}, \text{Tháng bay} = \text{“6.2023”}, \text{Thời gian khởi hành} = \text{“17:25”}\}$

Câu 3 (2.0đ) Cho 7 điểm trong không gian 2 chiều như sau: $x_1 = \{3, 8\}$, $x_2 = \{3, 7.5\}$, $x_3 = \{3, 7\}$, $x_4 = \{4, 7\}$, $x_5 = \{8, 3\}$, $x_6 = \{9, 2.5\}$, $x_7 = \{8, 2\}$.

Với ma trận U_0 được khởi tạo như sau:

U_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
C1	1	0	0	0	0	0	0
C2	0	1	0	0	0	0	0
C3	0	0	1	1	1	1	1

Áp dụng thuật toán K-means và sử dụng độ đo Euclide để gom 7 điểm trên vào **3 cụm**.

Yêu cầu: Chỉ thực hiện các bước sau:

- Bước 1: Tính trọng tâm cho các cụm
- Bước 2: So sánh khoảng cách điểm với trọng tâm từng cụm
- Bước 3: Xác định ma trận U_1

Khoa/ Bộ môn duyệt đề

TM. Giảng viên ra đề

Bảng ma trận đáp ứng chuẩn đầu ra.

Câu hỏi	CĐRMH
1	G1
2	G2
3	G2