TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐỀ 1

KHOA HỆ THỐNG THÔNG TIN

ĐỀ THI CUỐI KỲ
HỌC KỲ II – NĂM HỌC 2022-2023
MÔN KHAI THÁC DỮ LIỆU
*Thời gian làm bài: **90** phút*

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)**

21522425 - Lê Thị Lan Nhi

Câu 2:

1. minsup = 25% = 2/5

C_1

itemset	sup
{Vietjet}	4
{Vietnam}	3
{Pacific}	2
{Bamboo}	1
{Đà Lạt}	3
{Phú Quốc}	2
{Hà Nội}	2
{Đà Nẵng}	3
{6.2023}	4
{5.2023}	4
{4.2023}	2
{17:25}	3
{5:40}	2
{9:30}	3
{10:10}	2
{Trễ giờ}	5
{Đúng giờ}	5

$\Rightarrow$

L_1

item	sup
{Vietjet}	4
{VietNam}	3
{Đà Lạt}	3
{Đà Nẵng}	3
{6.2023}	4
{5.2023}	4
{17:25}	3
{9:30}	3
{Trễ giờ}	5
{Đúng giờ}	5

C_2

item	sup
{Vietjet, Đà Lạt}	1
{Vietjet, Đà Nẵng}	1
{Vietjet, 6.2023}	2
{Vietjet, 5.2023}	1
{Vietjet, 17:25}	1
{Vietjet, 9:30}	3
{Vietjet, Trễ giờ}	4
{Vietjet, Đúng giờ}	0

$\Rightarrow$

L_2

item	sup
{Vietjet, 9:30}	3
{Vietjet, Trễ giờ}	4
{Đà Lạt, 6.2023}	3
{9:30, Trễ giờ}	3

{Việt Nam, Đà Lạt} 1

{Việt Nam, Đà Nẵng} 0

{ VietNam, 6.2023} 1

{VietNam, 5.2023} 2

{VietNam, 17:25} 1

{ VietNam, 9:30} 0

{ Việt Nam, Trễ giờ } 1

{Việt Nam, Đúng giờ} 2

{Đà Lạt, 6.2023} 3

{ Đà Lạt, 5.2023} 0

{ Đà Lạt, 17:25} 2

{ Đà Lạt, 9:30} 0

{ Đà Lạt, Trễ giờ } 1

{ Đà Lạt, Đúng giờ } 2

{Đà Nẵng, 6.2023} 0

{Đà Nẵng, 5.2023} 1

{Đà Nẵng, 17:25} 1

{Đà Nẵng, 9:30} 1

{Đà Nẵng, Trễ giờ} 1

{Đà Nẵng, Đúng giờ} 2

{6.2023, 17:25} 2

{6.2023, 9:30} 1

{6.2023, Trễ giờ} 2

{6.2023, Đúng giờ} 2

{5.2023, 17:25} 1
{5.2023, 9:30} 1
{5.2023, Trễ giờ} 2
{5.2023, Đúng giờ} 2
{17:25, Trễ giờ} 1
{17:25, Đúng giờ} 0
{9:30, Trễ giờ} 3
{9:30, Đúng giờ} 0

C_3 {Việt, et, 9:30, Trễ giờ} sup 3 ⇒ L_3 {Việt, et, 9:30, Trễ giờ} sup 3

Các tập phổ biến

$L = L_1 \cup L_2 \cup L_3$

= {Việt, et} {Việt Nam} {Đà Lạt}, {Đà Nẵng}, {6.2023}, {5.2023}

{17:25}, {9:30}, {Trễ giờ}, {Đúng giờ}, {Việt, et, 9:30}

{Việt, et, Trễ giờ}, {Đà Lạt, 6.2023}, {9:30, Trễ giờ}

, {Việt, et, 9:30, Trễ giờ}

Các tập phổ biến cuối cùng: {Việt, et, 9:30, Trễ giờ}, {Đà Lạt, 6.2023}, {Việt Nam}, {Đà Nẵng}, {5.2023}, {17:25}, {Đúng giờ}

Xét {Việt, et, 9:30, Trễ giờ}

Các tập con khác rỗng của tập phổ biến {Việt, et, 9:30}

, {Việt, et, Trễ giờ}, {9:30, Trễ giờ}, {Việt, et, 9:30}

{Trễ giờ}

Các luật kết hợp

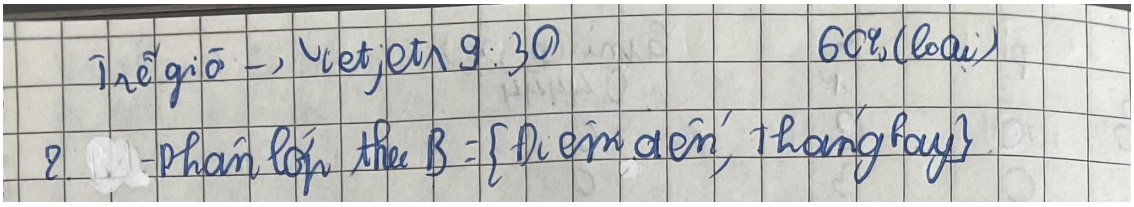
Việt, et ∧ 9:30 → Trễ giờ conf 100% (thỏa)

Việt, et ∧ Trễ giờ → 9:30 75% (thỏa)

9:30 ∧ Trễ giờ → Việt, et 100% (thỏa)

Việt, et → 9:30 ∧ Trễ giờ 75% (thỏa)

9:30 → Việt, et ∧ Trễ giờ 100% (thỏa)



2. Cho B={*Điểm đến, Tháng bay*}, X={1, 3, 6, 7, 8} (tập các mẫu có giá trị *Kết quả* = "*Trễ giờ*"). Sử dụng tập thô tính: xấp xỉ trên, xấp xỉ dưới và hệ số xấp xỉ. (1.0đ)

21520339 - Nguyễn Lê Ngọc Mai

- Phân lớp theo B = {Điểm đến, Tháng bay}

$U/B = \{\{1,5,9\},\{2\},\{3,7\},\{4,8\},\{6\},\{10\}\}$

- Xấp xỉ dưới

$\underline{B}X = \{x \mid [x]_B \subseteq X\} = \{3,7,6\}$

- Xấp xỉ trên

$\overline{B}X = \{x \mid [x]_B \cap X \neq \emptyset\} = \{1,5,9,3,7,4,8,6\}$

- Biên:

$B\_Biên = \overline{B}X - \underline{B}X = \{1,5,9,4,8\}$

$B\_Ngoài = U - \overline{B}X = \{2,10\}$

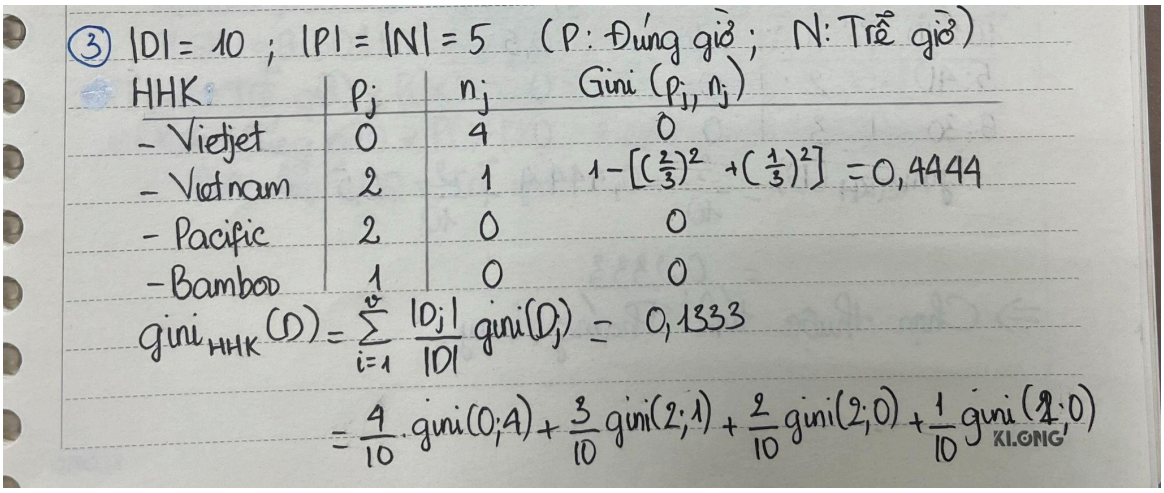
- Hệ số xấp xỉ

$$\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|} = \frac{|\{3,7,6\}|}{|\{1,5,9,3,7,4,8,6\}|} = \frac{3}{8} = 0.375$$

$\alpha_B(X) < 1 \Rightarrow$ Lớp quyết định KQ là thô

3. Xác định nút gốc của cây quyết định, sử dụng Chỉ số Gini. (1.75đ)

21520404 - Đặng Ánh Phước



③ $|D| = 10$; $|P| = |N| = 5$ (P: Đúng giờ ; N: Trễ giờ)

HHK	p_j	n_j	Gini (p_j, n_j)
- Vietjet	0	4	0
- Vietnam	2	1	$1-[(\frac{2}{3})^2+(\frac{1}{3})^2] = 0,4444$
- Pacific	2	0	0
- Bamboo	1	0	0

$gini_{HHK}(D) = \sum_{i=1}^{v} \frac{|D_j|}{|D|} gini(D_j) = 0,1333$

$= \frac{4}{10} \cdot gini(0;4) + \frac{3}{10} gini(2;1) + \frac{2}{10} gini(2;0) + \frac{1}{10} gini(1;0)$

DD	p_j	n_j	Gini (p_j, n_j)
Đà Lạt	2	1	$1-\left[\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right]=0,4444$
Phú Quốc	1	1	$1-\left[\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right]=0,5$
Hà Nội	0	2	~~$1-\left[\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right]=0,5$~~ 0
Đà Nẵng	2	1	$1-\left[\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2\right]=0,4444$

$$gini_{DD}(D)=\frac{3}{10}\cdot 0,4444+\frac{2}{10}\cdot 0,5+0+\frac{3}{10}\cdot 0,4444$$

$$=0,3667$$

TB	p_j	n_j	Gini (p_j, n_j)
6.2023	2	2	$1-\left[\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right]=0,5$
5.2023	2	2	0,5
4.2023	1	1	0,5

$$gini_{TB}(D)=\frac{4}{10}\cdot 0,5\cdot 2+\frac{2}{10}\cdot 0,5$$

$$=0,5$$

GKH	p_j	n_j	Gini (p_j, n_j)
17:25	2	1	0,4444
10:10	1	1	0,5
5:40	2	0	0
9:30	0	3	0

$$gini_{GKH}(D)=\frac{3}{10}\cdot 0,4444+\frac{2}{10}\cdot 0,5$$

$$=0,2333$$

→ Chọn thuộc tính Hãng hàng không

4. **Sử dụng công thức *Naïve Bayes* có làm trơn *Laplace* để phân lớp mẫu sau: (1.5đ)**

X={Hãng hàng không="Vietjet", Điểm đến="Phú Quốc", Tháng bay= "6.2023", Thời gian khởi hành = "9:30"}

21520129 - Bùi Thị Như Ý

- Ước lượng $P(C_i)$ với C_1= "Đúng giờ", C_2= "Trễ giờ"

P(KQ = "Đúng giờ") = 6/12

P(KQ = "Trễ giờ") = 6/12

- Với thuộc tính HHK

P(HHK = "Vietjet" | KQ = "Đúng giờ") = 1/9

P(HHK = "Vietjet" | KQ = "Trễ giờ") = 5/9

- Với thuộc tính DD

P(DD = "Phú Quốc" | KQ = "Đúng giờ") = 2/9

P(DD = "Phú Quốc" | KQ = "Trễ giờ") = 2/9

- Với thuộc tính TB

P(TB = "6.2023" | KQ = "Đúng giờ") = 3/8

P(TB = "6.2023" | KQ = "Trễ giờ") = 3/8

- Với thuộc tính GTH

P(GTH = "9:30" | KQ = "Đúng giờ") = 1/9

P(GTH = "9:30" | KQ = "Trễ giờ") = 4/9

- Xác định lớp cho X

Có:

- P(KQ = "Đúng giờ") * P(X | KQ = "Đúng giờ")

  = P(KQ = "Đúng giờ") * P(HHK = "Vietjet" | KQ = "Đúng giờ") * P(DD = "Phú Quốc" | KQ = "Đúng giờ") * P(TB = "6.2023" | KQ = "Đúng giờ") * P(GTH = "9:30" | KQ = "Đúng giờ")

  $= \frac{6}{12} * \frac{1}{9} * \frac{2}{9} * \frac{3}{8} * \frac{1}{9} = \frac{1}{1944} \approx 0.0005$

- P(KQ = "Trễ giờ") * P(X | KQ = "Trễ giờ")

  = P(KQ = "Trễ giờ") * P(HHK = "Vietjet" | KQ = "Trễ giờ") * P(DD = "Phú Quốc" | KQ = "Trễ giờ") * P(TB = "6.2023" | KQ = "Trễ giờ") * P(GTH = "9:30" | KQ = "Trễ giờ")

  $= \frac{6}{12} * \frac{5}{9} * \frac{2}{9} * \frac{3}{8} * \frac{4}{9} = \frac{5}{486} \approx 0.0103$

Vì P(KQ = "Trễ giờ") * P(X | KQ = "Trễ giờ") > P(KQ = "Đúng giờ") * P(X | KQ = "Đúng giờ")

=> X thuộc lớp C_2(KQ = "Trễ giờ")

Câu 3 (2.0đ) Cho 7 điểm trong không gian 2 chiều như sau: x1 ={3, 8}, x2 ={2, 7.5}, x3 ={3, 7}, x4 ={4, 7}, x5 ={8, 3}, x6 ={7, 2.5}, x7 ={8, 2}.

Với ma trận U0 được khởi tạo như sau:

U0	x1	x2	x3	x4	x5	x6	x7
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**.j

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 U1*

21521197 Hoàng Ngô Thảo Nguyên

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

x_1 thuộc C_1

x_2 thuộc C_2

x_3, x_4, x_5, x_6, x_7 thuộc C_3

(*) Tính vecto trọng tâm

C_1 có trọng tâm $c_1(3; 8)$

C_2 có trọng tâm $c_2(2; 7,5)$

C_3 có trọng tâm $c_3(6; 4,3)$

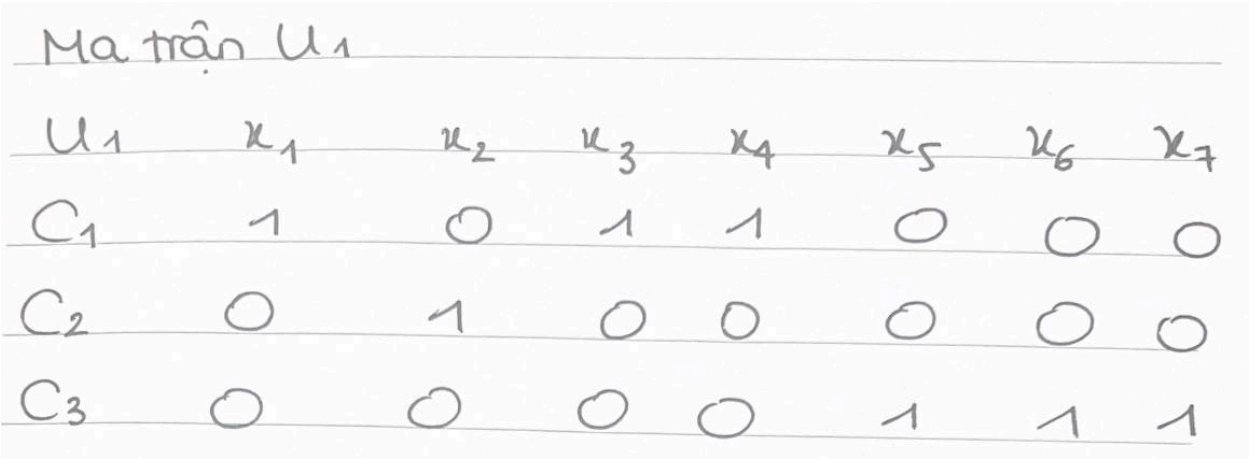
(*) Tính Khoảng cách

	$d(x, c_1)$	$d(x, c_2)$	$d(x, c_3)$
x_1	0	1,1180	4,7634
x_2	1,1180	0	5,1225
x_3	1	1,1180	4,0361
x_4	1,4142	2,0616	3,3601
x_5	7,0710	7,5	2,3854
x_6	6,8007	7,0711	2,0591
x_7	7,8102	8,1394	3,0479

Cụm $C_1 = \{x_1, x_3, x_4\}$

Cụm $C_2 = \{x_2\}$

Cụm $C_3 = \{x_5, x_6, x_7\}$



Ma trận U_1

U_1	x_1	x_2	x_3	x_4	x_5	x_6	x_7
C_1	1	0	1	1	0	0	0
C_2	0	1	0	0	0	0	0
C_3	0	0	0	0	1	1	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

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA HỆ THỐNG THÔNG TIN

ĐỀ THI CUỐI KỲ
HỌC KỲ II – NĂM HỌC 2022-2023
MÔN KHAI THÁC DỮ LIỆU

ĐỀ 2

Thời gian làm bài: ***90*** *phút*

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

HỌ VÀ TÊN SV: ..... **MSSV**:..... **STT**: ..... **PHÒNG THI:**.....	**CÁN BỘ COI THI**

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={*Điểm đến, Tháng bay*}, X={1, 3, 5, 7, 10} (tập các mẫu có giá trị *Kết quả* = "*Đúng giờ*"). Sử dụng tập thô tính: xấp xỉ trên, xấp xỉ dưới và hệ số xấp xỉ. (1.0đ)**

21520324_Lê Trần Thùy Linh

B={Điểm đến, Tháng bay}

X={1, 3, 5, 7, 10}

U / B = {{1}, {2, 4}, {3, 6, 10}, {5}, {7, 8}, {9}}

$\underline{B}X = \{1, 5\}$

$\bar{B}X = \{1, 3, 6, 10, 5, 7, 8\}$

Hệ số xấp xỉ: $\alpha_B(x) = \frac{2}{7} = 0.2857$

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={Hãng hàng không="Vietnam", Điểm đến="Đà Lạt", Tháng bay= "6.2023", Thời gian khởi hành = "17:25"}

21521495 - Nguyễn Kim Anh Thư

- Gọi C1, C2 lần lượt là các lớp có KQ là "Đúng giờ" và "Trễ giờ". Ước lượng P(Ci) với i = {1, 2}

P(KQ = "Đúng giờ") = 6/12

P(KQ = "Trễ giờ") = 6/12

- Với thuộc tính HHK

P(HHK = "Vietnam" | KQ = "Đúng giờ") = 3/9

P(HHK = "Vietnam" | KQ = "Trễ giờ") = 2/9

- Với thuộc tính DD

P(DD = "Đà Lạt" | KQ = "Đúng giờ") = 3/9

P(DD = "Đà Lạt" | KQ = "Trễ giờ") = 2/9

- Với thuộc tính TB

P(TB = "6.2023" | KQ = "Đúng giờ") = 3/8

P(TB = "6.2023" | KQ = "Trễ giờ") = 3/8

- Với thuộc tính GTH

P(GTH = "17:25" | KQ = "Đúng giờ") = 3/9

P(GTH = "17:25" | KQ = "Trễ giờ") = 2/9

- Xét

P(KQ = "Đúng giờ") * P(X | KQ = "Đúng giờ")

= P(KQ = "Đúng giờ") * P(HHK = "Vietnam" | KQ = "Đúng giờ") * P(DD = "Đà Lạt" | KQ = "Đúng giờ") * P(TB = "6.2023" | KQ = "Đúng giờ") * P(GTH = "17:25" | KQ = "Đúng giờ")

= 6/12 * 3/9 * 3/9 * 3/8 * 3/9 = 1/144 ≈ 0.0069

P(KQ = "Trễ giờ") * P(X | KQ = "Trễ giờ")

= P(KQ = "Trễ giờ") * P(HHK = "Vietnam" | KQ = "Trễ giờ") * P(DD = "Đà Lạt" | KQ = "Trễ giờ") * P(TB = "6.2023" | KQ = "Trễ giờ") * P(GTH = "17:25" | KQ = "Trễ giờ")

= 6/12 * 2/9 * 2/9 * 3/8 * 2/9= 1/488 ≈ 0.0021

Ta thấy 0.0069 > 0.0021 => X thuộc lớp C1(KQ = "Đúng giờ")

Câu 3 (2.0đ) Cho 7 điểm trong không gian 2 chiều như sau: x1 ={3, 8}, x2 ={3, 7.5}, x3 ={3, 7}, x4 ={4, 7}, x5 ={8, 3}, x6 ={9, 2.5}, x7 ={8, 2}.

Với ma trận U0 được khởi tạo như sau:

U0	x1	x2	x3	x4	x5	x6	x7
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**.

21520417 - Huỳnh Ngọc Quí

Tính trọng tâm các cụm C1, C2, C3: c1(c11, c12), c2(c21, c22), c3(c31, c33)

$$c11 = \frac{3}{1} = 3$$

$$c12 = \frac{8}{1} = 8$$

=> C1 có trọng tâm là: c1(3, 8)

$$c21 = \frac{3}{1} = 3$$

$$c22 = \frac{7.5}{1} = 7.5$$

=> C2 có trọng tâm là: c2(3, 7.5)

$$c31 = \frac{3 + 4 + 8 + 9 + 8}{5} = 6.4$$

$$c32 = \frac{7 + 7 + 3 + 2.5 + 2}{5} = 4.3$$

=> C3 có trọng tâm là: c3(6.4, 4.3)

- **So sánh khoảng cách điểm với trọng tâm từng cụm**

d(xi, Ci)	C1	C2	C3	Gần trọng tâm cụm nhất
x1	0	0.5	5.0249	C1
x2	0.5	0	4.669	C2
x3	1	0.5	4.3417	C2
x4	1.4142	1.118	3.6125	C2
x5	7.0711	6.7268	2.0616	C3

x6	8.1394	7.8102	3.1623	C3
x7	7.8102	7.433	2.8018	C3

- **Xác định ma trận U1**

U1	X1	X2	X3	X4	X5	X6	X7
C1	1	0	0	0	0	0	0
C2	0	1	1	1	0	0	0
C3	0	0	0	0	1	1	1

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 U1*

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

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
Khoa Hệ thống Thông tin

ĐỀ THI CUỐI KỲ
HỌC KỲ II, NĂM HỌC 2021-2022
Môn: Khai thác dữ liệu
Thời gian: 90 phút

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.

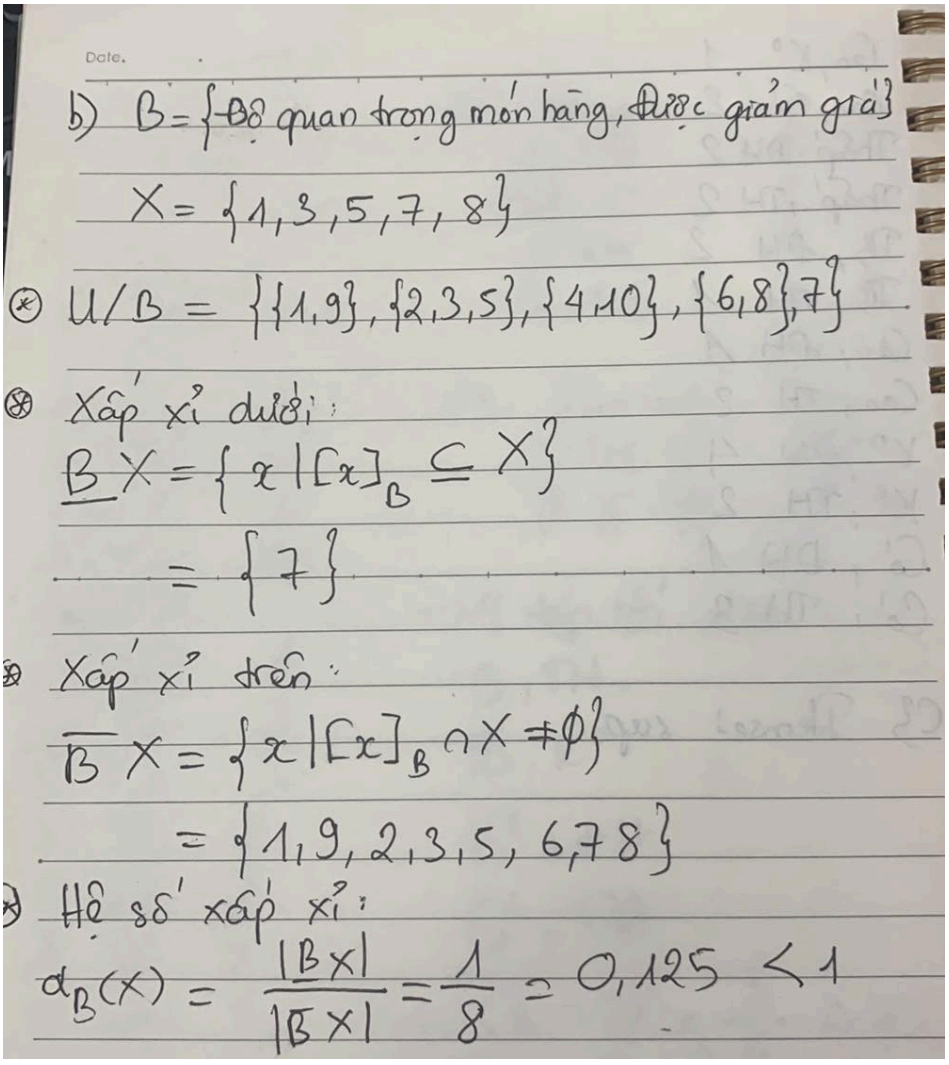
	Xếp loại Shipper (XL)	**Độ quan trọng món hàng (QT)**	**Được giảm giá ship (GG)**	**Kết quả giao hàng (KQ)**
1	B	Thấp	Có	Trễ hạn
2	C	Trung bình	Có	Đúng hạn
3	C	Cao	Không	Đúng hạn
4	C	Thấp	Có	Trễ hạn
5	C	Trung bình	Không	Trễ hạn
6	A	Cao	Không	Đúng hạn
7	B	Trung bình	Có	Trễ hạn

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={*Độ quan trọng món hàng, 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)

21522814 - Phan Quốc Vỹ



c. Xác định nút gốc của cây quyết định, sử dụng *Chỉ số Gini (Gini index)*. (1.5đ)

Phạm Duy Khánh - 21522211

Link: bt cuoi ki 2022 - de 1- cau 1 c

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"}

21521252 - Nguyễn Trọng Ninh

Ước lượng $P(C_i)$ với **C1**= "Đúng hạn", **C2**= "Trễ hạn" và $P(X|C_i)$

P(KQ = "Đúng hạn") = 6/12 = 0.5

P(KQ = "Trễ hạn") = 6/12 = 0.5

- Với thuộc tính XL
- Với thuộc tính QT
- Với thuộc tính GG

XL=B:

P (XL=B | KQ = "Đúng hạn") = $\frac{1}{8}$

P (XL=B | KQ = "Trễ hạn") = $\frac{3}{8}$

QT = "Thấp"

P (QT = "Thấp" | KQ = "Đúng hạn") = $\frac{2}{8}$

P (QT = "Thấp" | KQ = "Trễ hạn") = $\frac{3}{8}$

GG = "Không"

P (GG = "Không" | KQ = "Đúng hạn") = $\frac{3}{7}$

P (GG = "Không" | KQ = "Trễ hạn") = $\frac{3}{7}$

P (KQ = "Đúng hạn") * P (X| KQ = "Đúng hạn")

$= \frac{1}{2} * \frac{1}{8} * \frac{2}{8} * \frac{3}{7} = 0.0067$

P (KQ = "Trễ hạn") * P (X| KQ = "Trễ hạn")

$= \frac{1}{2} * \frac{3}{8} * \frac{3}{8} * \frac{3}{7} = 0.03$

P (KQ = "Trễ hạn") * P (X| KQ = "Trễ hạn") > P (KQ = "Đúng hạn") * P (X| KQ = "Đúng hạn") => X thuộc lớp C2 (KQ = "Trễ hạn")

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.
NHI - 21521230

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*

Trọng tâm cụm	X	y						
C1	2	10						
C2	2	5						
C3	5.17	5.33						
Ma trận khoảng cách								
	C1	C2	C3	**Cụm**				
A1	0	25	31.81	**C1**				
A2	25	0	10.14	**C2**				
A3	72	37	9.81	**C3**				
A4	13	18	7.14	**C3**				
A5	50	25	3.47	**C3**				
A6	52	17	2.47	**C3**				
A7	65	10	28.47	**C2**				
A8	5	20	14.81	**C1**				
Cập nhật U1								
U1	A1	A2	A3	A4	A5	A6	A7	A8
C1	1	0	0	0	0	0	0	1
C2	0	1	0	0	0	0	1	0
C3	0	0	1	1	1	1	0	0

Khoa Hệ thống Thông tin TM. Giảng viên ra đề

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
Khoa Hệ thống Thông tin

ĐỀ THI CUỐI KỲ
HỌC KỲ II, NĂM HỌC 2021-2022
Môn: Khai thác dữ liệu
Thời gian: 90 phút

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={*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 trơn *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.

<u>Lưu ý: chỉ thực hiện các bước</u>

- *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*