Trường Đại Học Bách Khoa - Tp. HCM
Khoa Khoa Học & Kỹ Thuật Máy Tính
***

Mã đề: **1171**

Kiểm tra giữa kỳ

Môn: **Khai Phá Dữ Liệu (CO3029)** - Ngành: **Khoa Học Máy Tính**

Ngày: 6/11/2017, HK1 – 2017-2018

Thời gian làm bài: **45** phút

(*Bài kiểm tra gồm **6** câu hỏi trắc nghiệm và **2** câu hỏi tự luận. Sinh viên được **tham khảo** tài liệu.)*

Cho câu 1-6, sinh viên chọn 1 câu trả lời đúng nhất. Nếu chọn câu (e) thì sinh viên cần trình bày đáp án khác so với đáp án ở các câu (a), (b), (c), và (d) và/hoặc giải thích lựa chọn (e) của mình. Điểm đúng của mỗi câu hỏi từ câu 1 đến câu 4 là 1 điểm.

Câu 1. Trong các quá trình xử lý dữ liệu sau, quá trình nào là quá trình khai phá dữ liệu?

a. Xác định giá trị lợi nhuận dự kiến trong mục tiêu của kế hoạch kinh doanh 2017-2022 của công ty AA.

b. Xác định lượng mưa trung bình tại thành phố Hồ Chí Minh cho tháng 10 của năm 2018.

c. Xác định lưu lượng giao thông qua cầu Mỹ Thuận trong thời gian trước, trong, và sau Tết sắp tới.

d. Câu a, b, và c đều đúng.

e. Ý kiến khác.

Câu 2. Chọn phát biểu **ĐÚNG** về công tác làm sạch dữ liệu.

a. Công tác này xử lý các thuộc tính dư thừa trong tập dữ liệu bằng cách kiểm tra tương quan.

b. Công tác này xử lý các thuộc tính liên tục trong tập dữ liệu bằng kỹ thuật binning.

c. Công tác này xử lý các dữ liệu vượt biên miền trị của thuộc tính khi thực hiện chuẩn hóa.

d. Công tác này xử lý các dữ liệu ngoại biên bằng việc phát hiện và lọc nhiễu.

e. Ý kiến khác.

Câu 3. Thực hiện chuẩn hóa dữ liệu trong **Bảng 3** (ở Câu 8) ở thuộc tính X_2 về miền trị [-1, 1]. Giá trị X_2 mới của đối tượng có ID = 5 là bao nhiêu? Giả sử làm tròn chính xác với 2 số thập phân.

a. -0.71

b. -0.43

c. -0.14

d. 1.14

e. Ý kiến khác.

Câu 4. Chọn phát biểu **ĐÚNG** về đặc điểm có thể giúp phân biệt giữa 3 mô hình phân lớp : cây quyết định C4.5, mạng neuron nhân tạo nhiều tầng truyền thẳng, máy vec-tơ hỗ trợ ?

a. Cấu trúc mô hình phân lớp.

b. Cơ chế xây dựng mô hình phân lớp.

c. Mức độ khả hiểu của mô hình đối với quyết định phân lớp.

d. Câu a, b, và c đều đúng.

e. Ý kiến khác.

Câu 5. Cho cây quyết định CART trong **Hình 1**, luật phân lớp nào **KHÔNG** được dẫn ra từ cây quyết định này?

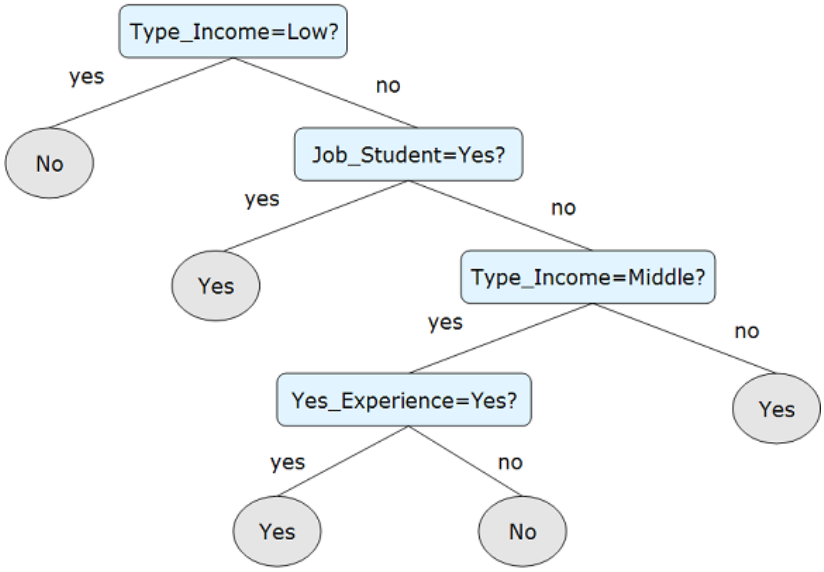
a. IF NOT(Type_Income=Low) AND NOT(Job_Student=Yes) AND NOT(Type_Income=Middle) THEN Class=Yes

b. IF Type_Income=Middle AND Job_Student=Yes AND NOT(Yes_Experience=Yes) THEN Class = No

c. IF NOT(Type_Income=Low) AND Job_Student=Yes THEN Class=Yes

d. IF Type_Income=Low THEN Class = No

e. Ý kiến khác.



Hình 1. Cây quyết định CART

Câu 6. Cho hai ma trận nhầm lẫn tương ứng cho 2 mô hình phân lớp M1 và M2 trong Bảng 1 và 2 như sau. Cho biết mô hình nào hiệu quả hơn cho việc phân lớp của các đối tượng thuộc lớp C3?

a. Mô hình M1.

b. Mô hình M2.

c. Cả hai mô hình M1 và M2 đều hiệu quả cho lớp C3.

d. Do thiếu thông tin về hai mô hình nên không thể xác định được mô hình nào hiệu quả hơn.

e. Ý kiến khác.

Bảng 1. Ma trận nhầm lẫn của M1

	Predicted C1	Predicted C2	Predicted C3	Predicted C4
C1	82	10	3	1
C2	2	65	1	0
C3	0	2	40	1
C4	2	1	1	10

Bảng 2. Ma trận nhầm lẫn của M2

	Predicted C1	Predicted C2	Predicted C3	Predicted C4
C1	82	8	2	4
C2	2	64	1	1
C3	2	1	38	2
C4	3	2	1	8

Câu 7. Phân biệt bài toán phân lớp với bài toán hồi quy. (1 *điểm*)

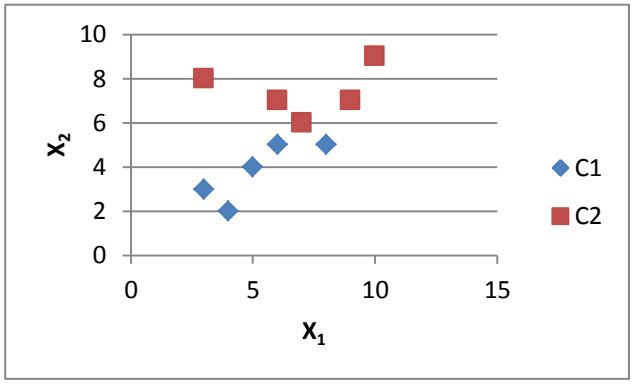
Câu 8. Cho các đối tượng dữ liệu trong **Bảng 3**.

a. Thực hiện rời rạc hóa dữ liệu dựa trên điểm phân tách là Q1 và Q3 ở mỗi chiều dữ liệu với 3 phân đoạn trị là [min, Q1), [Q1, Q3], và (Q3, max] ; trong đó, Q1 và Q3 là các điểm tứ phân vị tương ứng mức 25% và mức 75%. (1 *điểm*)

b. Xác định thuộc tính và kết quả phân tách dữ liệu tương ứng cho lần phân hoạch đầu tiên trong quá trình xây dựng cây quyết định C4.5 với dữ liệu huấn luyện trong Bảng 3. (2 *điểm*)

Bảng 3. Các đối tượng dữ liệu trong không gian R^2

OID	X_1	X_2	Lớp thật
O1	3	8	C2
O2	6	5	C1
O3	8	5	C1
O4	10	9	C2
O5	5	4	C1
O6	4	2	C1
O7	6	7	C2
O8	9	7	C2
O9	7	6	C2
O10	3	3	C1



Hình 1. Phân bố của các đối tượng trong R^2

Họ - Tên: …….

Mã Số Sinh Viên: ……

Môn: **Khai phá dữ liệu (CO3029)** Ngày kiểm tra: **6/11/2017** Mã đề: **1171**

Phần trả lời

Câu hỏi	a	b	c	d	e (Ý kiến khác)
1				√	
2				√	
3		√			
4				√	
5		√			
6	√				

Câu 7. Phân biệt giữa bài toán phân lớp và bài toán hồi quy:

- Điểm giống nhau: Tập huấn luyện có thông tin của các biến cần được dự báo; Cơ chế học có giám sát/bán giám sát;

- Điểm khác nhau:

Hồi quy: Dự báo giá trị của các biến liên tục; Đánh giá mô hình dựa trên độ lệch giữa giá trị quan sát và giá trị tính được từ mô hình cho các biến cần được dự báo

Phân lớp: Dự báo giá trị của biến rời rạc; Đánh giá mô hình dựa trên sự giống nhau và khác nhau của giá trị lớp quan sát được và giá trị lớp dự báo được cho các biến cần được dự báo

Câu 8.

a. Sắp thứ tự cho mỗi thuộc tính:

X_1	X_2
3	2
3	3
4	4
5	5
6	5
6	6
7	7
8	7
9	8
10	9

Tìm Q1 và Q3 tương ứng cho mỗi thuộc tính:

- Cho thuộc tính X_1, Q1 = 3.5 và Q3 = 7.5
- Cho thuộc tính X_2, Q1 = 3.5 và Q3 = 7

Rời rạc hóa dữ liệu với L = [min, Q1), M = [Q1, Q3], và H = (Q3, max]:

OID	X_1	X_2	Lớp thật
O1	L	H	C2
O2	M	M	C1
O3	H	M	C1
O4	H	H	C2
O5	M	M	C1
O6	M	L	C1
O7	M	M	C2
O8	H	M	C2
O9	M	M	C2
O10	L	L	C1

b. Tính các giá trị Info(), Gain(), và GainRatio() tương ứng cho mỗi thuộc tính và sau đó chọn thuộc tính có giá trị GainRatio lớn nhất:

Info(D) = 1

Info(D, X1) = 0.960964

Info(D, X2) = 0.6

Gain(D, X1) = Info(D) – Info(D, X1) = 0.039036

Gain(D, X2) = Info(D) – Info(D, X2) = 0.4

SplitInfo(D, X1) = 1.485475

SplitInfo(D, X2) = 1.370951

GainRatio(D, X1) = Gain(D, X1)/SplitInfo(D, X1) = 0.026278

GainRatio(D, X2) = Gain(D, X2)/SplitInfo(D, X2) = 0.291768

Thuộc tính được chọn là X2.

Họ - Tên: ……

Mã Số Sinh Viên: ……

Môn: **Khai phá dữ liệu (CO3029)** Ngày kiểm tra: **6/11/2017** Mã đề: **1171**

Kết quả phân tách dữ liệu đối với thuộc tính được chọn:

D1 ứng với X2 = L:

OID	**X1**	**X2**	**Lớp thật**
O6	M	L	C1
O10	L	L	C1

D2 ứng với X2 = M:

OID	**X1**	**X2**	**Lớp thật**
O2	M	M	C1
O3	H	M	C1
O5	M	M	C1
O7	M	M	C2
O8	H	M	C2
O9	M	M	C2

D3 ứng với X2 = H:

OID	**X1**	**X2**	**Lớp thật**
O1	L	H	C2
O4	H	H	C2