


<b>Giảng viên ra đề:</b> <b>Lê Hồng Trang</b> <small>(Chữ ký và Họ tên)</small>	<small>(Ngày ra đề)</small> <b>15/1/2021</b>	<b>Người phê duyệt:</b> <b>PGS. TS. Trần Minh Quang</b> <small>(Chữ ký, Chức vụ và Họ tên)</small>	<small>(Ngày duyệt đề)</small> <b>18/1/2021</b>
---	---	--	--

*(phần phía trên cần che đi khi in sao đề thi)*

 <b>TRƯỜNG ĐH BÁCH KHOA – ĐHQG-HCM</b> <b>KHOA KH &amp; KT MÁY TÍNH</b>	<b>THI CUỐI KỲ</b>		Học kỳ/năm học	1	2020-2021	
			Ngày thi	19/1/2021		
	Môn học	Khai phá Dữ liệu				
	Mã môn học	CO3029				
	Thời lượng	90 phút	Mã đề	201...		
<b>Ghi chú:</b> - Được sử dụng tài liệu - Nộp lại đề thi cùng với bài làm						

Đề thi gồm **25** câu trắc nghiệm (**6 điểm**) và **01** câu tự luận (**4 điểm**). Tô đậm phương án được chọn trong phiếu trả lời và viết lời giải bài tự luận vào sau đề bài tương ứng.

Bảng dưới đây là kết quả thống kê sau khi thực hiện phân cụm một tập 6000 điểm dữ liệu thành 3 cụm A, B, C.

		Actual			
		A	B	C	SUM
<b>Predicted</b>	A	600	400	200	1200
	B	1000	1200	200	2400
	C	400	400	1600	2400
	SUM	2000	2000	2000	

Các câu hỏi 1 và 4 xét với các số liệu cho trong bảng trên.

**Câu hỏi 1 [L.O.3.3, L.O.5.1].** Các chỉ số TP, TN, FP và FN được tính tương ứng là

- ☐ (A) 2200, 1200, 1200, 800.
☐ (B) 1200, 2200, 1200, 800.
☐ (C) 1200, 2200, 800, 1200.
☐ (D) 1200, 1200, 2200, 800.

**Câu hỏi 2 [L.O.3.3, L.O.5.1].** Chỉ số Precision là

- ☐ (A) 0.3.
☐ (B) 0.4.
☐ (C) 0.5.
☐ (D) 0.6.

**Câu hỏi 3 [L.O.3.3, L.O.5.1].** Chỉ số Recall là

- ☐ (A) 0.3.
☐ (B) 0.4.
☐ (C) 0.5.
☐ (D) 0.6.

**Câu hỏi 4 [L.O.3.3, L.O.5.1].** Chỉ số  $F_1$ -score là

- ☐ (A) 0.54.
☐ (B) 0.45.
☐ (C) 0.64.
☐ (D) 0.46.

**Câu hỏi 5 [L.O.3.3].** Giải thuật  $k$ -means có một số hạn chế. Một trong số đó là việc gán cứng một điểm vào một cụm (tức một điểm chỉ thuộc hoàn toàn vào một cụm hoặc không). Giải thuật nào sau đây được xem là sự cải tiến của  $k$ -means cho hạn chế này?

- ☐ (A) AGNES.
☐ (B) DIANA.
☐ (C) DBSCAN.
☐ (D) Fuzzy  $c$ -means.

Các câu hỏi 6 và 7 xét bài toán sau. Giả sử ta cần phân cụm 7 điểm dữ liệu thành 3 cụm ( $C_1, C_2, C_3$ ) sử dụng giải thuật  $k$ -means. Sau một lần lặp ta có các cụm như sau:

- $C_1 = \{(2, 2), (4, 4), (6, 6)\}$
- $C_2 = \{(0, 4), (4, 0)\}$
- $C_3 = \{(5, 5), (9, 9)\}$

**Câu hỏi 6 [L.O.3.3].** Khi đó, tâm cụm được xác định cho bước lặp tiếp theo sẽ là

- |   |   |
|---|---|
| <input type="radio"/> (A) $C_1 : (4, 4), C_2 : (2, 2), C_3 : (7, 7).$ | <input type="radio"/> (B) $C_1 : (6, 6), C_2 : (4, 4), C_3 : (9, 9).$ |
| <input type="radio"/> (C) $C_1 : (2, 2), C_2 : (0, 0), C_3 : (5, 5).$ | <input type="radio"/> (D) Tất cả đều sai.                             |

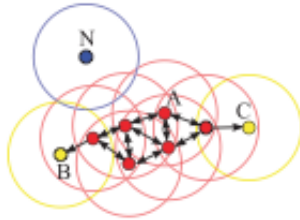
**Câu hỏi 7 [L.O.3.3].** Khoảng cách Mahattan giữa điểm  $(9, 9)$  đến tâm của cụm  $C_1$  trong bước lặp tiếp theo là

- |                               |                               |
|-------------------------------|-------------------------------|
| <input type="radio"/> (A) 8.  | <input type="radio"/> (B) 9.  |
| <input type="radio"/> (C) 10. | <input type="radio"/> (D) 11. |

**Câu hỏi 8 [L.O.3.3].** Giải thuật  $k$ -means sẽ cho kết quả không tốt với tập dữ liệu nào sau đây?

- |  |   |
|--|---|
| <input type="radio"/> (A) Có nhiều.                        | <input type="radio"/> (B) Tất cả trường hợp này.                  |
| <input type="radio"/> (C) Có các mật độ phân bố khác nhau. | <input type="radio"/> (D) Có các cụm có hình dáng kiểu không lồi. |

Các câu hỏi 9 và 10 xét hình ảnh dưới đây.



**Câu hỏi 9 [L.O.3.3, L.O.5.1].** Đây là hình ảnh minh họa cho giải thuật nào?

- |                                       |  |
|---------------------------------------|--|
| <input type="radio"/> (A) $k$ -means. | <input type="radio"/> (B) Agglomerative.   |
| <input type="radio"/> (C) DBSCAN.     | <input type="radio"/> (D) <i>Apriori</i> . |

**Câu hỏi 10 [L.O.3.3, L.O.5.1].** Điểm nào sẽ bị loại bỏ trong giải thuật phân cụm đúng được chọn ở câu 9?

- |                              |                              |
|------------------------------|------------------------------|
| <input type="radio"/> (A) A. | <input type="radio"/> (B) N. |
| <input type="radio"/> (C) B. | <input type="radio"/> (D) C. |

Các câu hỏi 11 và 13 Giả sử có một tập dữ liệu  $D_1$ . Xây dựng một mô hình hồi quy tuyến tính với đa thức bậc 3. Sau đó, nhận thấy rằng sai số huấn luyện (training error) và sai số thử nghiệm (testing error) là 0.

**Câu hỏi 11 [L.O.3.2].** Điều gì xảy ra nếu sử dụng đa thức bậc 4 để xây dựng một mô hình hồi quy khác cho tập dữ liệu trên?

- |  |
|--|
| <input type="radio"/> (A) Có thể mô hình mới sẽ underfit.        |
| <input type="radio"/> (B) Tất cả hiện tượng này đều xảy ra.      |
| <input type="radio"/> (C) Có thể mô hình mới sẽ overfit.         |
| <input type="radio"/> (D) Mô hình sẽ mới sẽ cho kết quả tốt hơn. |

**Câu hỏi 12 [L.O.3.2].** Điều gì xảy ra nếu sử dụng đa thức bậc 2 để xây dựng một mô hình hồi quy khác cho tập dữ liệu trên?

- |   |  |
|---|--|
| <input type="radio"/> (A) Có thể mô hình mới sẽ underfit. | <input type="radio"/> (B) Tất cả hiện tượng này đều xảy ra.      |
| <input type="radio"/> (C) Có thể mô hình mới sẽ overfit.  | <input type="radio"/> (D) Mô hình sẽ mới sẽ cho kết quả tốt hơn. |

**Câu hỏi 13 [L.O.3.2].** Nếu sử dụng đa thức bậc 2 để xây dựng một mô hình hồi quy khác cho tập dữ liệu trên, đặc trưng bias và variance của mô hình này sẽ

- (A) bias cao, variance cao. (B) bias cao, variance thấp.  
(C) bias thấp, variance thấp. (D) bias thấp, variance cao.

**Câu hỏi 14 [L.O.3.2].** Một mạng nơ-ron nhân tạo có  $n$  đầu vào  $x_1, x_2, \dots, x_n$  với các trọng số  $w_1, w_2, \dots, w_n$ . Giá trị tổng có trọng số sẽ được truyền tới hàm kích hoạt được tính là

- (A)  $\sum_{i=1}^n x_i w_i$ . (B)  $\sum_{i=1}^n x_i$ .  
(C)  $\sum_{i=1}^n w_i$ . (D)  $\sum_{i=1}^n x_i + \sum_{i=1}^n w_i$ .

**Câu hỏi 15 [L.O.3.2].** Mạng nơ-ron nào sau đây dùng học có giám sát?

- (A) Mạng Hopfield. (B) Mạng perceptron đa tầng.  
(C) Bản đồ đặc trưng tự tổ chức. (D) Tất cả các mạng này.

**Câu hỏi 16 [L.O.3.4].** Một itemset có giá trị hỗ trợ (support) lớn hơn hoặc bằng một ngưỡng cho trước gọi là

- (A) xuất hiện không thường xuyên.  
(B) ngưỡng xuất hiện thường xuyên.  
(C) xuất hiện thường xuyên.  
(D) ngưỡng xuất hiện không thường xuyên.

**Câu hỏi 17 [L.O.3.4].** Kỹ thuật nào dưới đây giúp cải thiện giải thuật Apriori?

- (A) Lấy mẫu. (B) Tăng số lượng giao dịch.  
(C) Giảm số lượng giao dịch. (D) Kỹ thuật băm (hash).

**Câu hỏi 18 [L.O.3.4].** Độ tin cậy của  $A \rightarrow B$ , ký hiệu bởi  $confidence(A \rightarrow B)$ , được định nghĩa là

- (A)  $\frac{support(A \cap B)}{support(A)}$ . (B)  $\frac{support(A \cup B)}{support(A)}$ .  
(C)  $\frac{support(A \cap B)}{support(B)}$ . (D)  $\frac{support(A \cup B)}{support(B)}$ .

**Câu hỏi 19 [L.O.3.4].** Đại lượng  $lift$  được định nghĩa bởi  $lift = \frac{P(A \cup B)}{p(A)p(B)}$ , được dùng để

- (A) đánh giá luật kết hợp dạng  $A \rightarrow B$ . (B) đo sự tương quan giữa hai sự kiện  $A$  và  $B$ .  
(C) đánh giá luật kết hợp dạng  $\langle A, B \rangle \rightarrow A$ . (D) đánh giá luật kết hợp dạng  $\langle A, B \rangle \rightarrow B$ .

**Câu hỏi 20 [L.O.3.4].** Kỹ thuật nào dưới đây thích hợp nhất khi áp dụng để xác định một bài viết (trên mạng xã hội) được thích hay không?

- (A) Phân lớp. (B) Phân cụm.  
(C) Hồi quy. (D) Khai phá luật kết hợp.

Các câu hỏi 21–25 xét danh sách giao dịch dưới đây

- (1) pointer, mouse, laptop, headphone, flash-disk  
(2) hard-disk, cleaner, pointer, laptop  
(3) pointer, mouse  
(4) laptop, cleaner, flash-disk  
(5) laptop, hard-disk, cleaner

**Câu hỏi 21 [L.O.3.4].** Danh sách có

- (A) 5 giao dịch. (B) 4 giao dịch.  
(C) 6 giao dịch. (D) 7 giao dịch.

**Câu hỏi 22 [L.O.3.4, L.O.5.1].** Với  $support = 0.5$ , danh sách các mẫu (itemsets) xuất hiện thường xuyên là

- (A) {laptop}, {mouse}.
- (B) {headphone}, {bag}.
- (C) {laptop, mouse}, {mouse, headphone}, {laptop, bag}.
- (D) {pointer}, {laptop}, {cleaner}, {laptop, cleaner}.

**Câu hỏi 23 [L.O.3.4].** Nếu giảm giá trị của  $support$  xuống, thì

- (A) một số mẫu (itemsets) có thể được thêm vào tập xuất hiện thường xuyên hiện tại.
- (B) số mẫu (itemsets) xuất hiện thường xuyên vẫn luôn giữ nguyên.
- (C) một số mẫu (itemsets) sẽ được đưa ra khỏi tập xuất hiện thường xuyên hiện tại.
- (D) không xác định được tăng hay giảm số mẫu.

**Câu hỏi 24 [L.O.3.4, L.O.5.1].** Các luật kết hợp với  $support = 0.5$  và  $confidence = 0.7$  gồm

- (A) {mouse}  $\rightarrow$  {headphone}, {mouse}  $\rightarrow$  {laptop}.
- (B) {laptop}  $\rightarrow$  {cleaner}, {cleaner}  $\rightarrow$  {laptop}.
- (C) {laptop}  $\rightarrow$  {mouse}, {mouse}  $\rightarrow$  {laptop}.
- (D) {bag}  $\rightarrow$  {mouse}, {mouse}  $\rightarrow$  {headphone}.

**Câu hỏi 25 [L.O.3.4,L.O.5.1].** Kết quả khai phá luật kết hợp thu được cho thấy

- (A) laptop và mouse thường sẽ được mua cùng nhau.
- (B) laptop và headphone thường sẽ được mua cùng nhau.
- (C) laptop và bag thường sẽ được mua cùng nhau.
- (D) laptop và cleaner thường sẽ được mua cùng nhau.

**Câu hỏi 26 [L.O.3.3, L.O.5.1]. Tự luận – Phân cụm dữ liệu**

Xét tập dữ liệu gồm 8 điểm  $A_1 = (2, 10), A_2 = (2, 5), A_3 = (8, 4), A_4 = (5, 8), A_5 = (7, 5), A_6 = (6, 4), A_7 = (1, 2), A_8 = (4, 9)$ . Thực hiện phân cụm tập dữ liệu với tập trên sử dụng phương pháp phân cấp agglomerative với các yêu cầu cụ thể dưới đây.

**Yêu cầu**

- (a) Xây dựng ma trận khoảng cách cho tập dữ liệu, với khoảng cách Euclidean. (1 điểm)
- (b) Thực hiện phân cụm cho hai trường hợp dùng độ đo khoảng cách *single-link* và *complete-link*. Với mỗi trường hợp, lập bảng cho các bước lặp, vẽ biểu đồ dendrogram và kết quả phân cụm thu được. (3 điểm)

**Lời giải**



**Đáp án – Mã đề: 2010**

Câu hỏi 1 (B)

Câu hỏi 12 (A)

Câu hỏi 25 (D)

Câu hỏi 2 (C)

Câu hỏi 13 (B)

Câu hỏi 26 Lời giải

Câu hỏi 3 (D)

Câu hỏi 14 (A)

Câu hỏi 4 (A)

Câu hỏi 15 (B)

Câu hỏi 5 (D)

Câu hỏi 16 (C)

Câu hỏi 6 (A)

Câu hỏi 17 (D)

Câu hỏi 7 (C)

Câu hỏi 18 (A)

Câu hỏi 8 (B)

Câu hỏi 19 (B)

Câu hỏi 20 (A)

Câu hỏi 9 (C)

Câu hỏi 21 (A)

Câu hỏi 10 (B)

Câu hỏi 22 (D)

Câu hỏi 23 (A)

Câu hỏi 11 (C)

Câu hỏi 24 (B)