Đại học Bách Khoa Tp. Hồ Chí Minh
Khoa Khoa học & Kỹ thuật Máy tính

THI CUỐI KỲ
Môn: **Khai phá dữ liệu** (CO3029)

Lớp: **20182** Nhóm: **LO2**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **07/06/2019**

Họ tên sinh viên: _____ **Mã số sinh viên.:** | | | | | | | | |

Điểm: _____ **Người ra đề:** _____ **Lê Hồng Trang**

Bằng chữ: _____ **Người coi thi:** _____

*Đề thi gồm **30** câu trắc nghiệm (**7 điểm**) và **01** câu tự luận (**3 điểm**). Tô đậm phương án được chọn trong phiếu trả lời và viết lời giải bài tự luận vào sau đề bài tương ứng.*

Câu 1 [L.O.3.2]. Mạng nơ-ron nhân tạo (ANN) là một mô hình tính toán:

(A) thường được dùng cho bài toán phân lớp hay nhận dạng.
(B) tất cả những đặc điểm này.
(C) mô phỏng cơ chế hoạt động của não người.
(D) số nút (node) đầu ra có thể là một hoặc nhiều.

Câu 2 [L.O.3.3]. Giải thuật k-means

(A) luôn dừng tại điểm tối toàn cục.
(B) thường sẽ kết thúc tại điểm tối ưu địa phương.
(C) không chắn chắn sẽ dừng.

Câu 3 [L.O.3.3]. Khoảng cách giữa các cụm dữ liệu C_i và C_j có thể được tính bởi

(A) Tất cả đều được.
(B) liên kết đơn (single link): $d(C_i, C_j) = \min\{d(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j\}$.
(C) liên kết đầy đủ (complete link): $d(C_i, C_j) = \max\{d(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j\}$.
(D) khoảng cách tâm (centroid): $d(C_i, C_j) = d(c_i, c_j)$, với c_i, c_j là tâm của C_i và C_j.

Câu 4 [L.O.1]. Tri thức có thể thu được từ quá trình khai phá dữ liệu là

(A) Mô hình phân loại.
(B) Mô hình phân cụm.
(C) Tập mẫu thường xuyên và tập luật.
(D) Tất cả những phương án còn lại.

Câu 5 [L.O.3.1]. Hồi quy logistic dùng để

(A) phân lớp dữ liệu.
(B) phân cụm dữ liệu.
(C) dự đoán.
(D) mô tả dữ liệu.

Câu 6 [L.O.3.3]. Một phương pháp phân cụm tốt cần đưa ra được các cụm mà

(A) tính tương tự trong cụm cao và tính tướng tự ngoài cụm cao.
(B) tính tương tự trong cụm cao và tính tướng tự ngoài cụm thấp.
(C) tính tương tự trong cụm thấp và tính tướng tự ngoài cụm thấp.
(D) tính tương tự trong cụm thấp và tính tướng tự ngoài cụm cao.

Câu 7 [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 $A \to B$.
(B) đo sự tương quan giữa hai sự kiện A và B.
(C) đánh giá luật kết hợp dạng $< A, B > \to A$.
(D) đánh giá luật kết hợp dạng $< A, B > \to B$.

Các câu hỏi 8 và 9 xét một mô hình phân lớp dùng hàm $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$ cho giả thuyết phân lớp.

Câu 8 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây sai?

(A) Đây là hàm hồi quy logistic.
(B) Đây là hàm sigmoid.
(C) X là tập dữ liệu mẫu.
(D) $h_\theta(X)$ là xác suất để $Y =$ "1", với Y là thuộc tính nhãn và "1" là nhãn đang được quan tâm.

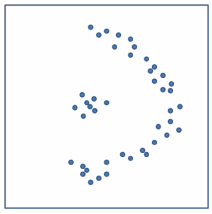
Câu 9 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây đúng?

(A) $h_\theta(X) \in [-1, 1]$.
(B) $h_\theta(X) \in [0, 1]$.
(C) $h_\theta(X) \in \mathbb{R}$.
(D) Không có phát biểu đúng.

Câu 10 [L.O.3.4]. Độ hỗ trợ của A, ký hiểu bởi $support(A)$, được định nghĩa là số giao dịch (transaction)

(A) không chứa A trên tổng số giao dịch.
(B) chứa A.
(C) không chứa A.
(D) chứa A trên tổng số giao dịch.

Câu 11 [L.O.3.3]. Giải thuật nào thích hợp nhất để phân cụm tập điểm dữ liệu dưới đây, nếu sử dụng hàm khoảng cách Ơclit (Euclidean)?



(A) DBSCAN.
(B) k-means.
(C) k-medoids.
(D) Các giải thuật này cho kết quả tương tự.

Câu 12 [L.O.3.2]. Hàm độ đo nào thường được dùng với dữ liệu nhị phân?

(A) Mahattan.
(B) Jaccard.
(C) Eiuclidean.
(D) Minkowski.

Câu hỏi 13 và 14 xét mô hình phân lớp M thực hiện phân loại dữ liệu có ba nhãn A, B và C. Kết qủa phân loại được cho bởi ma trận confusion sau đây

	A	B	C
A	116	13	10
B	14	11	20
C	11	10	122

Câu 13 [L.O.3.2]. Độ chính xác (precision) của việc phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập phân) là

(A) 0.832.
(B) 0.823.
(C) 0.825.
(D) 0.852.

Câu 14 [L.O.3.2]. Độ truy hồi (recall) của việc phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập phân) là

(A) 0.892.
(B) 0.289.
(C) 0.829.
(D) 0.298.

Câu 15 [L.O.4.4]. Để thu giảm dữ liệu, ta có thể sử dụng phương pháp

(A) Tất cả những phương án còn lại.
(B) Phân tích thành phần chính.
(C) Lấy mẫu dữ liệu.
(D) Kết hợp khối dữ liệu.

Câu 16 [L.O.3.3]. Có bao nhiêu cụm được sinh bởi giải thuật k-means?

(A) 2^k.
(B) e^k.
(C) Một bội số của k.
(D) k.

Các câu hỏi 17–21 xét danh sách giao dịch dưới đây
(1) $I_1, I_2, I_3, I_4, I_5, I_6$
(2) $I_7, I_2, I_3, I_4, I_5, I_6$
(3) I_1, I_8, I_4, I_5
(4) $I_1, I_9, I_{10}, I_4, I_6$
(5) $I_{10}, I_2, I_4, I_{11}, I_5$

Câu 17 [L.O.3.4]. Danh sách có

(A) 11 giao dịch.
(B) 6 giao dịch.
(C) 5 giao dịch.
(D) 9 giao dịch.

Câu 18 [L.O.3.4, L.O.5.1]. Với $support = 0.6$, danh sách các mẫu (itemsets) xuất hiện thường xuyên là

(A) gồm tất cả các mẫu trong các phương án còn lại.
(B) $< I_1 >, < I_2 >, < I_4 >, < I_5 >, < I_6 >$.
(C) $< I_1, I_4 >, < I_2, I_4 >, < I_2, I_5 >, < I_4, I_5 >, < I_4, I_6 >$.
(D) $< I_2, I_4, I_5 >$.

Câu 19 [L.O.3.4]. Nếu giảm giá trị của *support* xuống, thì

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

Câu 20 [L.O.3.4, L.O.5.1]. Các luật kết hợp có thể được khai phá với $support = 0.6$ và $confidence = 0.8$ là

(A) $< I_2, I_4 > \rightarrow I_1, < I_2, I_5 > \rightarrow I_3$.
(B) $< I_2, I_4 > \rightarrow I_5, < I_2, I_5 > \rightarrow I_4$.
(C) $< I_2, I_4 > \rightarrow I_5, < I_1, I_5 > \rightarrow I_2$.
(D) $< I_3, I_5 > \rightarrow I_4, < I_3, I_4 > \rightarrow I_5$.

Câu 21 [L.O.3.4]. Nếu tăng giá trị của *confidence* xuống, thì

(A) một số luật kết hợp khác sẽ được thêm vào tập luật.
(B) tập luật không thay đổi.
(C) một số luật kết hợp khác sẽ bị đưa ra khỏi tập luật.
(D) không thể xác định số lượng luật trong tập luật.

Câu 22 [L.O.3.4]. Giải thuật Apriori dùng để

(A) khám phá ra tất cả mẫu xuất hiện thường xuyên bằng việc cắt bỏ các luật có độ hộ trợ (support) lớn hơn.
(B) phân cụm các đối tượng dữ liệu.
(C) khám phá ra tất cả mẫu xuất hiện thường xuyên bằng việc cắt bỏ các luật có độ hộ trợ (support) nhỏ hơn.
(D) phân lớp các đối tượng dữ liệu.

Câu 23 [L.O.3.3, L.O.5.1]. Gọi ϵ là bán kính hình cầu lân cận của một điểm trong một tập dữ liệu $\mathcal{D}$ cho trước, ký hiệu $N_\epsilon(p) = \{q \in \mathcal{D} : d(p,q) \leq \epsilon\}$, trong đó $d(p,q)$ là khoảng cách giữa p và q. Gọi $MinPts$ là số điểm tối thiểu trong một lân cận của một điểm trong $\mathcal{D}$. Khi đó, nếu $p \in \mathcal{D}$ là một điểm nhân (core) thì

(A) $|N_\epsilon(p)| \leq MinPts$.
(B) $|N_\epsilon(p)| = MinPts$.
(C) $|N_\epsilon(p)|$ tuỳ ý.
(D) $|N_\epsilon(p)| \geq MinPts$.

Câu 24 [L.O.3.4]. Độ tin cậy của $A \to B$, ký hiệu bởi $confidence(A \to B)$, 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 25 [L.O.3.4]. Một luật kết hợp được quan tâm nếu

(A) nó thoả mãn điều kiện về $min\_support$.
(B) nó thoả mãn điều kiện về $min\_confidence$.
(C) nó thoả mãn đồng thời cả hai điều kiện về $min\_support$ và $min\_confidence$.

Câu 26 [L.O.3.3]. Trong giải thuật gom cụm trộn (agglomerative), các cụm ban đầu được xác đinh

(A) ngẫu nhiên.
(B) chính là tập các đối tượng dữ liệu.
(C) chính là các đối tượng dữ liệu.
(D) bởi k đối tượng dữ liệu ngẫu nhiên.

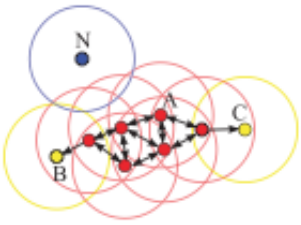
Câu 27 [L.O.3.3]. Với một tập dữ liệu có n đối tượng, nếu giải thuật k-means kết thúc quá trình phân cụm sau t bước lặp thì thời gian tính toán là

(A) $O(ktn)$.
(B) $kO(tn)$.
(C) $tO(kn)$.
(D) $O(kt \log n)$.

Câu 28 [L.O.3.1]. Hàm $y = a \log(bx)$ là

(A) một hàm hồi quy tuyến tính.
(B) một hàm sigmoid.
(C) một hàm mất mát (loss function).
(D) một hàm hồi quy phi tuyến.

Các câu hỏi 29 và 30 xét hình ảnh dưới đây.



Câu 29 [L.O.3.3, L.O.5.1]. Đây là hình ảnh minh hoạ cho giải thuật nào?

(A) k-means.
(B) Agglomerative.
(C) DBSCAN.
(D) *Apriori*.

Câu 30 [L.O.3.3, L.O.5.1]. Điểm nào sẽ bị loại bỏ trong giải thuật phân cụm đúng được chọn ở câu 29?

(A) A.
(B) N.
(C) B.
(D) C.

Câu 31 [L.O.3.3, L.O.5.1]. Tự luận – *Phân cụm dữ liệu theo theo tiếp cận phân cấp*

Xét tập dữ liệu gồm 6 điểm (với 2 thuộc tính) được cho trong bảng dưới đây.

Điểm	x**-toạ độ**	y**-toạ độ**
p_1	0.4005	0.5306
p_2	0.2148	0.3854
p_3	0.3457	0.3156
p_4	0.2652	0.1875
p_5	0.0789	0.4139
p_6	0.4548	0.3022

Yêu cầu

(a) Xây dựng ma trận sai khác (khoảng cách) cho tập dữ liệu cho bởi bảng trên, biết rằng độ đo về sự sai khác (dissimilarity) giữa các điểm là khoảng cách Ơclit (Euclidean distance).

(b) Bằng giải thuật AGNES (Agglomerative Nesting) với độ đo single-link cho khoảng cách giữa các cụm, hãy xây dựng cấu trúc phân cấp cụm cho tập dữ liệu (biểu diễn dưới dạng biểu đồ Venn) và cây phả hệ (dendrogram) tương ứng.

Lời giải

Đại học Bách Khoa Tp. Hồ Chí Minh
Khoa Khoa học & Kỹ thuật Máy tính

THI CUỐI KỲ
Môn: **Khai phá dữ liệu** (CO3029)

Lớp: **20182** Nhóm: **LO2**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **07/06/2019**

Đáp án – Mã đề: 1820

Câu 1 (B)
Câu 2 (B)
Câu 3 (A)
Câu 4 (D)
Câu 5 (A)
Câu 6 (B)
Câu 7 (B)
Câu 8 (C)
Câu 9 (B)
Câu 10 (D)
Câu 11 (A)
Câu 12 (B)
Câu 13 (B)
Câu 14 (C)
Câu 15 (A)
Câu 16 (D)
Câu 17 (C)
Câu 18 (A)
Câu 19 (D)
Câu 20 (B)
Câu 21 (C)
Câu 22 (C)
Câu 23 (D)
Câu 24 (A)
Câu 25 (C)
Câu 26 (C)
Câu 27 (A)
Câu 28 (D)
Câu 29 (C)
Câu 30 (B)

Đại học Bách Khoa Tp. Hồ Chí Minh
Khoa Khoa học & Kỹ thuật Máy tính

THI CUỐI KỲ
Môn: **Khai phá dữ liệu** (CO3029)

Lớp: **20182** Nhóm: **LO2**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **07/06/2019**

Họ tên sinh viên: _____ **Mã số sinh viên.:** | | | | | | | | |

Điểm: _____ **Người ra đề:** _____ **Lê Hồng Trang**

Bằng chữ: _____ **Người coi thi:** _____

*Đề thi gồm **30** câu trắc nghiệm (**7 điểm**) và **01** câu tự luận (**3 điểm**). Tô đậm phương án được chọn trong phiếu trả lời và viết lời giải bài tự luận vào sau đề bài tương ứng.*

Câu 1 [L.O.3.4]. Độ tin cậy của $A \to B$, ký hiệu bởi $confidence(A \to B)$, được định nghĩa là

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

Câu 2 [L.O.3.3]. Giải thuật k-means

(A) luôn dừng tại điểm tối toàn cục.
(B) không chắn chắn sẽ dừng.
(C) thường sẽ kết thúc tại điểm tối ưu địa phương.

Câu 3 [L.O.1]. Tri thức có thể thu được từ quá trình khai phá dữ liệu là

(A) Tất cả những phương án còn lại.
(B) Mô hình phân loại.
(C) Mô hình phân cụm.
(D) Tập mẫu thường xuyên và tập luật.

Các câu hỏi 4–8 xét danh sách giao dịch dưới đây
(1) $I_1, I_2, I_3, I_4, I_5, I_6$
(2) $I_7, I_2, I_3, I_4, I_5, I_6$
(3) I_1, I_8, I_4, I_5
(4) $I_1, I_9, I_{10}, I_4, I_6$
(5) $I_{10}, I_2, I_4, I_{11}, I_5$

Câu 4 [L.O.3.4]. Danh sách có

(A) 9 giao dịch.
(B) 11 giao dịch.
(C) 6 giao dịch.
(D) 5 giao dịch.

Câu 5 [L.O.3.4, L.O.5.1]. Với $support = 0.6$, danh sách các mẫu (itemsets) xuất hiện thường xuyên là

(A) $< I_2, I_4, I_5 >$.
(B) gồm tất cả các mẫu trong các phương án còn lại.
(C) $< I_1 >, < I_2 >, < I_4 >, < I_5 >, < I_6 >$.
(D) $< I_1, I_4 >, < I_2, I_4 >, < I_2, I_5 >, < I_4, I_5 >, < I_4, I_6 >$.

Câu 6 [L.O.3.4]. Nếu giảm giá trị của *support* xuống, thì

(A) một số mẫu (itemsets) sẽ được thêm vào tập xuất hiện thường xuyên hiện tại.
(B) số mẫu (itemsets) xuất hiện thường xuyên vẫn 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 7 [L.O.3.4, L.O.5.1]. Các luật kết hợp có thể được khai phá với $support = 0.6$ và $confidence = 0.8$ là

(A) $< I_3, I_5 > \rightarrow I_4, < I_3, I_4 > \rightarrow I_5$.
(B) $< I_2, I_4 > \rightarrow I_1, < I_2, I_5 > \rightarrow I_3$.
(C) $< I_2, I_4 > \rightarrow I_5, < I_2, I_5 > \rightarrow I_4$.
(D) $< I_2, I_4 > \rightarrow I_5, < I_1, I_5 > \rightarrow I_2$.

Câu 8 [L.O.3.4]. Nếu tăng giá trị của $confidence$ xuống, thì

(A) không thể xác định số lượng luật trong tập luật.
(B) một số luật kết hợp khác sẽ được thêm vào tập luật.
(C) tập luật không thay đổi.
(D) một số luật kết hợp khác sẽ bị đưa ra khỏi tập luật.

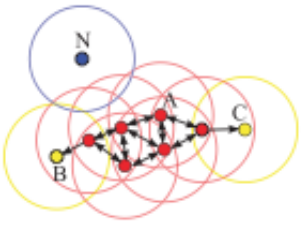
Câu 9 [L.O.3.2]. Hàm độ đo nào thường được dùng với dữ liệu nhị phân?

(A) Minkowski.
(B) Mahattan.
(C) Jaccard.
(D) Eiuclidean.

Câu 10 [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 $< A, B > \rightarrow B$.
(B) đánh giá luật kết hợp dạng $A \rightarrow B$.
(C) đo sự tương quan giữa hai sự kiện A và B.
(D) đánh giá luật kết hợp dạng $< A, B > \rightarrow A$.

Các câu hỏi 11 và 12 xét hình ảnh dưới đây.



Câu 11 [L.O.3.3, L.O.5.1]. Đây là hình ảnh minh hoạ cho giải thuật nào?

(A) *Apriori*.
(B) k-means.
(C) Agglomerative.
(D) DBSCAN.

Câu 12 [L.O.3.3, L.O.5.1]. Điểm nào sẽ bị loại bỏ trong giải thuật phân cụm đúng được chọn ở câu 11?

(A) C.
(B) A.
(C) N.
(D) B.

Câu 13 [L.O.3.4]. Giải thuật Apriori dùng để

(A) phân lớp các đối tượng dữ liệu.
(B) khám phá ra tất cả mẫu xuất hiện thường xuyên bằng việc cắt bỏ các luật có độ hộ trợ (support) lớn hơn.
(C) phân cụm các đối tượng dữ liệu.
(D) khám phá ra tất cả mẫu xuất hiện thường xuyên bằng việc cắt bỏ các luật có độ hộ trợ (support) nhỏ hơn.

Câu 14 [L.O.3.2]. Mạng nơ-ron nhân tạo (ANN) là một mô hình tính toán:

(A) số nút (node) đầu ra có thể là một hoặc nhiều.
(B) thường được dùng cho bài toán phân lớp hay nhận dạng.
(C) tất cả những đặc điểm này.
(D) mô phỏng cơ chế hoạt động của não người.

Câu hỏi 15 và 16 xét mô hình phân lớp M thực hiện phân loại dữ liệu có ba nhãn A, B và C. Kết qủa phân loại được cho bởi ma trận confusion sau đây

	A	B	C
A	116	13	10
B	14	11	20
C	11	10	122

Câu 15 [L.O.3.2]. Độ chính xác (precision) của việc phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập phân) là

(A) 0.852.
(B) 0.832.
(C) 0.823.
(D) 0.825.

Câu 16 [L.O.3.2]. Độ truy hồi (recall) của việc phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập phân) là

(A) 0.298.
(B) 0.892.
(C) 0.289.
(D) 0.829.

Câu 17 [L.O.3.3]. Có bao nhiêu cụm được sinh bởi giải thuật k-means?

(A) k.
(B) 2^k.
(C) e^k.
(D) Một bội số của k.

Câu 18 [L.O.3.4]. Độ hỗ trợ của A, ký hiểu bởi $support(A)$, được định nghĩa là số giao dịch (transaction)

(A) chứa A trên tổng số giao dịch.
(B) không chứa A trên tổng số giao dịch.
(C) chứa A.
(D) không chứa A.

Câu 19 [L.O.3.4]. Một luật kết hợp được quan tâm nếu

(A) nó thoả mãn điều kiện về $min\_support$.
(B) nó thoả mãn đồng thời cả hai điều kiện về $min\_support$ và $min\_confidence$.
(C) nó thoả mãn điều kiện về $min\_confidence$.

Câu 20 [L.O.3.3, L.O.5.1]. Gọi ϵ là bán kính hình cầu lân cận của một điểm trong một tập dữ liệu $\mathcal{D}$ cho trước, ký hiệu $N_\epsilon(p) = \{q \in \mathcal{D} : d(p,q) \leq \epsilon\}$, trong đó $d(p,q)$ là khoảng cách giữa p và q. Gọi $MinPts$ là số điểm tối thiểu trong một lân cận của một điểm trong $\mathcal{D}$. Khi đó, nếu $p \in \mathcal{D}$ là một điểm nhân (core) thì

(A) $|N_\epsilon(p)| \geq MinPts$.
(B) $|N_\epsilon(p)| \leq MinPts$.
(C) $|N_\epsilon(p)| = MinPts$.
(D) $|N_\epsilon(p)|$ tuỳ ý.

Câu 21 [L.O.3.1]. Hàm $y = a\log(bx)$ là

(A) một hàm hồi quy phi tuyến.
(B) một hàm hồi quy tuyến tính.
(C) một hàm sigmoid.
(D) một hàm mất mát (loss function).

Câu 22 [L.O.3.1]. Hồi quy logistic dùng để

(A) mô tả dữ liệu.
(B) phân lớp dữ liệu.
(C) phân cụm dữ liệu.
(D) dự đoán.

Câu 23 [L.O.3.3]. Với một tập dữ liệu có n đối tượng, nếu giải thuật k-means kết thúc quá trình phân cụm sau t bước lặp thì thời gian tính toán là

(A) $O(kt\log n)$.
(B) $O(ktn)$.
(C) $kO(tn)$.
(D) $tO(kn)$.

Câu 24 [L.O.3.3]. Trong giải thuật gom cụm trộn (agglomerative), các cụm ban đầu được xác đinh

(A) bởi k đối tượng dữ liệu ngẫu nhiên.
(B) ngẫu nhiên.
(C) chính là tập các đối tượng dữ liệu.
(D) chính là các đối tượng dữ liệu.

Câu 25 [L.O.3.3]. Khoảng cách giữa các cụm dữ liệu C_i và C_j có thể được tính bởi

(A) khoảng cách tâm (centroid): $d(C_i, C_j) = d(c_i, c_j)$, với c_i, c_j là tâm của C_i và C_j.
(B) Tất cả đều được.
(C) liên kết đơn (single link): $d(C_i, C_j) = \min\{d(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j\}$.
(D) liên kết đầy đủ (complete link): $d(C_i, C_j) = \max\{d(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j\}$.

Các câu hỏi 26 và 27 xét một mô hình phân lớp dùng hàm $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$ cho giả thuyết phân lớp.

Câu 26 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây sai?

(A) $h_\theta(X)$ là xác suất để $Y =$ "1", với Y là thuộc tính nhãn và "1" là nhãn đang được quan tâm.
(B) Đây là hàm hồi quy logistic.
(C) Đây là hàm sigmoid.
(D) X là tập dữ liệu mẫu.

Câu 27 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây đúng?

(A) Không có phát biểu đúng.
(B) $h_\theta(X) \in [-1, 1]$.
(C) $h_\theta(X) \in [0, 1]$.
(D) $h_\theta(X) \in \mathbb{R}$.

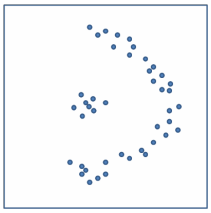
Câu 28 [L.O.4.4]. Để thu giảm dữ liệu, ta có thể sử dụng phương pháp

(A) Kết hợp khối dữ liệu.
(B) Tất cả những phương án còn lại.
(C) Phân tích thành phần chính.
(D) Lấy mẫu dữ liệu.

Câu 29 [L.O.3.3]. Một phương pháp phân cụm tốt cần đưa ra được các cụm mà

(A) tính tương tự trong cụm thấp và tính tướng tự ngoài cụm cao.
(B) tính tương tự trong cụm cao và tính tướng tự ngoài cụm cao.
(C) tính tương tự trong cụm cao và tính tướng tự ngoài cụm thấp.
(D) tính tương tự trong cụm thấp và tính tướng tự ngoài cụm thấp.

Câu 30 [L.O.3.3]. Giải thuật nào thích hợp nhất để phân cụm tập điểm dữ liệu dưới đây, nếu sử dụng hàm khoảng cách Ơclit (Euclidean)?



(A) Các giải thuật này cho kết quả tương tự.
(B) DBSCAN.
(C) k-means.
(D) k-medoids.

Câu 31 [L.O.3.3, L.O.5.1]. Tự luận – *Phân cụm dữ liệu theo theo tiếp cận phân cấp*

Xét tập dữ liệu gồm 6 điểm (với 2 thuộc tính) được cho trong bảng dưới đây.

Điểm	x**-toạ độ**	y**-toạ độ**
p_1	0.4005	0.5306
p_2	0.2148	0.3854
p_3	0.3457	0.3156
p_4	0.2652	0.1875
p_5	0.0789	0.4139
p_6	0.4548	0.3022

Yêu cầu

(a) Xây dựng ma trận sai khác (khoảng cách) cho tập dữ liệu cho bởi bảng trên, biết rằng độ đo về sự sai khác (dissimilarity) giữa các điểm là khoảng cách Ơclit (Euclidean distance).

(b) Bằng giải thuật AGNES (Agglomerative Nesting) với độ đo single-link cho khoảng cách giữa các cụm, hãy xây dựng cấu trúc phân cấp cụm cho tập dữ liệu (biểu diễn dưới dạng biểu đồ Venn) và cây phả hệ (dendrogram) tương ứng.

Lời giải

Đại học Bách Khoa Tp. Hồ Chí Minh
Khoa Khoa học & Kỹ thuật Máy tính

THI CUỐI KỲ
Môn: **Khai phá dữ liệu** (CO3029)

Lớp: **20182** Nhóm: **LO2**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **07/06/2019**

Đáp án – Mã đề: 1821

Câu 1 (B)
Câu 2 (C)
Câu 3 (A)
Câu 4 (D)
Câu 5 (B)
Câu 6 (A)
Câu 7 (C)
Câu 8 (D)
Câu 9 (C)
Câu 10 (C)
Câu 11 (D)
Câu 12 (C)
Câu 13 (D)
Câu 14 (C)
Câu 15 (C)
Câu 16 (D)
Câu 17 (A)
Câu 18 (A)
Câu 19 (B)
Câu 20 (A)
Câu 21 (A)
Câu 22 (B)
Câu 23 (B)
Câu 24 (D)
Câu 25 (B)
Câu 26 (D)
Câu 27 (C)
Câu 28 (B)
Câu 29 (C)
Câu 30 (B)

Đại học Bách Khoa Tp. Hồ Chí Minh
Khoa Khoa học & Kỹ thuật Máy tính

THI CUỐI KỲ
Môn: **Khai phá dữ liệu** (CO3029)

Lớp: **20182** Nhóm: **LO2**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **07/06/2019**

Họ tên sinh viên: _____ **Mã số sinh viên.:** | | | | | | | | |

Điểm: _____ **Người ra đề:** _____ **Lê Hồng Trang**

Bằng chữ: _____ **Người coi thi:** _____

*Đề thi gồm **30** câu trắc nghiệm (**7 điểm**) và **01** câu tự luận (**3 điểm**). Tô đậm phương án được chọn trong phiếu trả lời và viết lời giải bài tự luận vào sau đề bài tương ứng.*

Câu 1 [L.O.3.4]. Một luật kết hợp được quan tâm nếu

(A) nó thoả mãn điều kiện về $min\_confidence$.
(B) nó thoả mãn điều kiện về $min\_support$.
(C) nó thoả mãn đồng thời cả hai điều kiện về $min\_support$ và $min\_confidence$.

Các câu hỏi 2–6 xét danh sách giao dịch dưới đây
(1) $I_1, I_2, I_3, I_4, I_5, I_6$
(2) $I_7, I_2, I_3, I_4, I_5, I_6$
(3) I_1, I_8, I_4, I_5
(4) $I_1, I_9, I_{10}, I_4, I_6$
(5) $I_{10}, I_2, I_4, I_{11}, I_5$

Câu 2 [L.O.3.4]. Danh sách có

(A) 11 giao dịch.
(B) 9 giao dịch.
(C) 6 giao dịch.
(D) 5 giao dịch.

Câu 3 [L.O.3.4, L.O.5.1]. Với $support = 0.6$, danh sách các mẫu (itemsets) xuất hiện thường xuyên là

(A) gồm tất cả các mẫu trong các phương án còn lại.
(B) $< I_2, I_4, I_5 >$.
(C) $< I_1 >, < I_2 >, < I_4 >, < I_5 >, < I_6 >$.
(D) $< I_1, I_4 >, < I_2, I_4 >, < I_2, I_5 >, < I_4, I_5 >, < I_4, I_6 >$.

Câu 4 [L.O.3.4]. Nếu giảm giá trị của *support* xuống, thì

(A) số mẫu (itemsets) xuất hiện thường xuyên vẫn giữ nguyên.
(B) một số mẫu (itemsets) sẽ được thêm vào tập xuất hiện thường xuyên hiện tại.
(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 5 [L.O.3.4, L.O.5.1]. Các luật kết hợp có thể được khai phá với $support = 0.6$ và $confidence = 0.8$ là

(A) $< I_2, I_4 > \rightarrow I_1, < I_2, I_5 > \rightarrow I_3$.
(B) $< I_3, I_5 > \rightarrow I_4, < I_3, I_4 > \rightarrow I_5$.
(C) $< I_2, I_4 > \rightarrow I_5, < I_2, I_5 > \rightarrow I_4$.
(D) $< I_2, I_4 > \rightarrow I_5, < I_1, I_5 > \rightarrow I_2$.

Câu 6 [L.O.3.4]. Nếu tăng giá trị của $confidence$ xuống, thì

(A) một số luật kết hợp khác sẽ được thêm vào tập luật.
(B) không thể xác định số lượng luật trong tập luật.
(C) tập luật không thay đổi.
(D) một số luật kết hợp khác sẽ bị đưa ra khỏi tập luật.

Câu 7 [L.O.3.3]. Với một tập dữ liệu có n đối tượng, nếu giải thuật k-means kết thúc quá trình phân cụm sau t bước lặp thì thời gian tính toán là

(A) $O(ktn)$.
(B) $O(kt \log n)$.
(C) $kO(tn)$.
(D) $tO(kn)$.

Câu 8 [L.O.1]. Tri thức có thể thu được từ quá trình khai phá dữ liệu là

(A) Mô hình phân loại.
(B) Tất cả những phương án còn lại.
(C) Mô hình phân cụm.
(D) Tập mẫu thường xuyên và tập luật.

Câu 9 [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 $A \to B$.
(B) đánh giá luật kết hợp dạng $< A, B > \to B$.
(C) đo sự tương quan giữa hai sự kiện A và B.
(D) đánh giá luật kết hợp dạng $< A, B > \to A$.

Câu 10 [L.O.3.1]. Hồi quy logistic dùng để

(A) phân lớp dữ liệu.
(B) mô tả dữ liệu.
(C) phân cụm dữ liệu.
(D) dự đoán.

Câu 11 [L.O.3.4]. Giải thuật Apriori dùng để

(A) khám phá ra tất cả mẫu xuất hiện thường xuyên bằng việc cắt bỏ các luật có độ hộ trợ (support) lớn hơn.
(B) phân lớp các đối tượng dữ liệu.
(C) phân cụm các đối tượng dữ liệu.
(D) khám phá ra tất cả mẫu xuất hiện thường xuyên bằng việc cắt bỏ các luật có độ hộ trợ (support) nhỏ hơn.

Các câu hỏi 12 và 13 xét một mô hình phân lớp dùng hàm $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$ cho giả thuyết phân lớp.

Câu 12 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây sai?

(A) Đây là hàm hồi quy logistic.
(B) $h_\theta(X)$ là xác suất để $Y =$ "1", với Y là thuộc tính nhãn và "1" là nhãn đang được quan tâm.
(C) Đây là hàm sigmoid.
(D) X là tập dữ liệu mẫu.

Câu 13 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây đúng?

(A) $h_\theta(X) \in [-1, 1]$.
(B) Không có phát biểu đúng.
(C) $h_\theta(X) \in [0, 1]$.
(D) $h_\theta(X) \in \mathbb{R}$.

Câu 14 [L.O.3.3]. Giải thuật k-means

(A) thường sẽ kết thúc tại điểm tối ưu địa phương.
(B) luôn dừng tại điểm tối toàn cục.
(C) không chắn chắn sẽ dừng.

Câu hỏi 15 và 16 xét mô hình phân lớp M thực hiện phân loại dữ liệu có ba nhãn A, B và C. Kết qủa phân loại được cho bởi ma trận confusion sau đây

	A	B	C
A	116	13	10
B	14	11	20
C	11	10	122

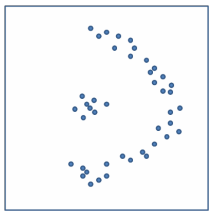
Câu 15 [L.O.3.2]. Độ chính xác (precision) của việc phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập phân) là

(A) 0.832.
(B) 0.852.
(C) 0.823.
(D) 0.825.

Câu 16 [L.O.3.2]. Độ truy hồi (recall) của việc phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập phân) là

(A) 0.892.
(B) 0.298.
(C) 0.289.
(D) 0.829.

Câu 17 [L.O.3.3]. Giải thuật nào thích hợp nhất để phân cụm tập điểm dữ liệu dưới đây, nếu sử dụng hàm khoảng cách Ơclit (Euclidean)?



(A) DBSCAN.
(B) Các giải thuật này cho kết quả tương tự.
(C) k-means.
(D) k-medoids.

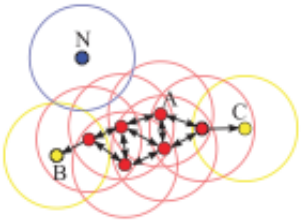
Câu 18 [L.O.3.3]. Có bao nhiêu cụm được sinh bởi giải thuật k-means?

(A) 2^k.
(B) k.
(C) e^k.
(D) Một bội số của k.

Câu 19 [L.O.3.2]. Hàm độ đo nào thường được dùng với dữ liệu nhị phân?

(A) Mahattan.
(B) Minkowski.
(C) Jaccard.
(D) Eiuclidean.

Các câu hỏi 20 và 21 xét hình ảnh dưới đây.



Câu 20 [L.O.3.3, L.O.5.1]. Đây là hình ảnh minh hoạ cho giải thuật nào?

(A) k-means.
(B) *Apriori*.
(C) Agglomerative.
(D) DBSCAN.

Câu 21 [L.O.3.3, L.O.5.1]. Điểm nào sẽ bị loại bỏ trong giải thuật phân cụm đúng được chọn ở câu 20?

(A) A.
(B) C.
(C) N.
(D) B.

Câu 22 [L.O.3.3]. Một phương pháp phân cụm tốt cần đưa ra được các cụm mà

(A) tính tương tự trong cụm cao và tính tướng tự ngoài cụm cao.
(B) tính tương tự trong cụm thấp và tính tướng tự ngoài cụm cao.
(C) tính tương tự trong cụm cao và tính tướng tự ngoài cụm thấp.
(D) tính tương tự trong cụm thấp và tính tướng tự ngoài cụm thấp.

Câu 23 [L.O.3.3]. Trong giải thuật gom cụm trộn (agglomerative), các cụm ban đầu được xác đinh

(A) ngẫu nhiên.
(B) bởi k đối tượng dữ liệu ngẫu nhiên.
(C) chính là tập các đối tượng dữ liệu.
(D) chính là các đối tượng dữ liệu.

Câu 24 [L.O.3.4]. Độ tin cậy của $A \to B$, ký hiệu bởi $confidence(A \to B)$, được định nghĩa là

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

Câu 25 [L.O.3.2]. Mạng nơ-ron nhân tạo (ANN) là một mô hình tính toán:

(A) thường được dùng cho bài toán phân lớp hay nhận dạng.
(B) số nút (node) đầu ra có thể là một hoặc nhiều.
(C) tất cả những đặc điểm này.
(D) mô phỏng cơ chế hoạt động của não người.

Câu 26 [L.O.4.4]. Để thu giảm dữ liệu, ta có thể sử dụng phương pháp

(A) Tất cả những phương án còn lại.
(B) Kết hợp khối dữ liệu.
(C) Phân tích thành phần chính.
(D) Lấy mẫu dữ liệu.

Câu 27 [L.O.3.4]. Độ hỗ trợ của A, ký hiểu bởi $support(A)$, được định nghĩa là số giao dịch (transaction)

(A) không chứa A trên tổng số giao dịch.
(B) chứa A trên tổng số giao dịch.
(C) chứa A.
(D) không chứa A.

Câu 28 [L.O.3.3]. Khoảng cách giữa các cụm dữ liệu C_i và C_j có thể được tính bởi

(A) Tất cả đều được.
(B) khoảng cách tâm (centroid): $d(C_i, C_j) = d(c_i, c_j)$, với c_i, c_j là tâm của C_i và C_j.
(C) liên kết đơn (single link): $d(C_i, C_j) = \min\{d(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j\}$.
(D) liên kết đầy đủ (complete link): $d(C_i, C_j) = \max\{d(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j\}$.

Câu 29 [L.O.3.3, L.O.5.1]. Gọi ϵ là bán kính hình cầu lân cận của một điểm trong một tập dữ liệu $\mathcal{D}$ cho trước, ký hiệu $N_\epsilon(p) = \{q \in \mathcal{D} : d(p, q) \leq \epsilon\}$, trong đó $d(p, q)$ là khoảng cách giữa p và q. Gọi $MinPts$ là số điểm tối thiểu trong một lân cận của một điểm trong $\mathcal{D}$. Khi đó, nếu $p \in \mathcal{D}$ là một điểm nhân (core) thì

(A) $|N_\epsilon(p)| \leq MinPts$.
(B) $|N_\epsilon(p)| \geq MinPts$.
(C) $|N_\epsilon(p)| = MinPts$.
(D) $|N_\epsilon(p)|$ tuỳ ý.

Câu 30 [L.O.3.1]. Hàm $y = a \log(bx)$ là

(A) một hàm hồi quy tuyến tính.
(B) một hàm hồi quy phi tuyến.
(C) một hàm sigmoid.
(D) một hàm mất mát (loss function).

Câu 31 [L.O.3.3, L.O.5.1]. Tự luận – *Phân cụm dữ liệu theo theo tiếp cận phân cấp*

Xét tập dữ liệu gồm 6 điểm (với 2 thuộc tính) được cho trong bảng dưới đây.

Điểm	x**-toạ độ**	y**-toạ độ**
p_1	0.4005	0.5306
p_2	0.2148	0.3854
p_3	0.3457	0.3156
p_4	0.2652	0.1875
p_5	0.0789	0.4139
p_6	0.4548	0.3022

Yêu cầu

(a) Xây dựng ma trận sai khác (khoảng cách) cho tập dữ liệu cho bởi bảng trên, biết rằng độ đo về sự sai khác (dissimilarity) giữa các điểm là khoảng cách Ơclit (Euclidean distance).

(b) Bằng giải thuật AGNES (Agglomerative Nesting) với độ đo single-link cho khoảng cách giữa các cụm, hãy xây dựng cấu trúc phân cấp cụm cho tập dữ liệu (biểu diễn dưới dạng biểu đồ Venn) và cây phả hệ (dendrogram) tương ứng.

Lời giải

Đại học Bách Khoa Tp. Hồ Chí Minh
Khoa Khoa học & Kỹ thuật Máy tính

THI CUỐI KỲ
Môn: **Khai phá dữ liệu** (CO3029)

Lớp: **20182** Nhóm: **LO2**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **07/06/2019**

Đáp án – Mã đề: 1822

Câu 1 (C)
Câu 2 (D)
Câu 3 (A)
Câu 4 (B)
Câu 5 (C)
Câu 6 (D)
Câu 7 (A)
Câu 8 (B)
Câu 9 (C)
Câu 10 (A)
Câu 11 (D)
Câu 12 (D)
Câu 13 (C)
Câu 14 (A)
Câu 15 (C)
Câu 16 (D)
Câu 17 (A)
Câu 18 (B)
Câu 19 (C)
Câu 20 (D)
Câu 21 (C)
Câu 22 (C)
Câu 23 (D)
Câu 24 (A)
Câu 25 (C)
Câu 26 (A)
Câu 27 (B)
Câu 28 (A)
Câu 29 (B)
Câu 30 (B)

Đại học Bách Khoa Tp. Hồ Chí Minh
Khoa Khoa học & Kỹ thuật Máy tính

THI CUỐI KỲ
Môn: **Khai phá dữ liệu** (CO3029)

Lớp: **20182** Nhóm: **LO2**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **07/06/2019**

Họ tên sinh viên: _____ **Mã số sinh viên.:** | | | | | | | | |

Điểm: _____ **Người ra đề:** _____ **Lê Hồng Trang**

Bằng chữ: _____ **Người coi thi:** _____

*Đề thi gồm **30** câu trắc nghiệm (**7 điểm**) và **01** câu tự luận (**3 điểm**). Tô đậm phương án được chọn trong phiếu trả lời và viết lời giải bài tự luận vào sau đề bài tương ứng.*

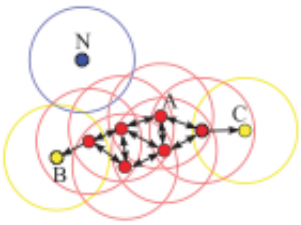
Câu 1 [L.O.3.4]. Giải thuật Apriori dùng để

(A) khám phá ra tất cả mẫu xuất hiện thường xuyên bằng việc cắt bỏ các luật có độ hộ trợ (support) lớn hơn.
(B) khám phá ra tất cả mẫu xuất hiện thường xuyên bằng việc cắt bỏ các luật có độ hộ trợ (support) nhỏ hơn.
(C) phân cụm các đối tượng dữ liệu.
(D) phân lớp các đối tượng dữ liệu.

Câu 2 [L.O.3.3]. Một phương pháp phân cụm tốt cần đưa ra được các cụm mà

(A) tính tương tự trong cụm cao và tính tướng tự ngoài cụm cao.
(B) tính tương tự trong cụm thấp và tính tướng tự ngoài cụm thấp.
(C) tính tương tự trong cụm cao và tính tướng tự ngoài cụm thấp.
(D) tính tương tự trong cụm thấp và tính tướng tự ngoài cụm cao.

Các câu hỏi 3 và 4 xét hình ảnh dưới đây.



Câu 3 [L.O.3.3, L.O.5.1]. Đây là hình ảnh minh hoạ cho giải thuật nào?

(A) k-means.
(B) DBSCAN.
(C) Agglomerative.
(D) *Apriori*.

Câu 4 [L.O.3.3, L.O.5.1]. Điểm nào sẽ bị loại bỏ trong giải thuật phân cụm đúng được chọn ở câu 3?

(A) A.
(B) B.
(C) N.
(D) C.

Câu 5 [L.O.3.3]. Giải thuật k-means

(A) thường sẽ kết thúc tại điểm tối ưu địa phương.
(B) không chắn chắn sẽ dừng.
(C) luôn dừng tại điểm tối toàn cục.

Câu 6 [L.O.4.4]. Để thu giảm dữ liệu, ta có thể sử dụng phương pháp

(A) Tất cả những phương án còn lại.
(B) Lấy mẫu dữ liệu.
(C) Phân tích thành phần chính.
(D) Kết hợp khối dữ liệu.

Câu 7 [L.O.3.3]. Với một tập dữ liệu có n đối tượng, nếu giải thuật k-means kết thúc quá trình phân cụm sau t bước lặp thì thời gian tính toán là

(A) $O(ktn)$.
(B) $tO(kn)$.
(C) $kO(tn)$.
(D) $O(kt \log n)$.

Câu 8 [L.O.3.1]. Hồi quy logistic dùng để

(A) phân lớp dữ liệu.
(B) dự đoán.
(C) phân cụm dữ liệu.
(D) mô tả dữ liệu.

Câu 9 [L.O.3.2]. Mạng nơ-ron nhân tạo (ANN) là một mô hình tính toán:

(A) thường được dùng cho bài toán phân lớp hay nhận dạng.
(B) mô phỏng cơ chế hoạt động của não người.
(C) tất cả những đặc điểm này.
(D) số nút (node) đầu ra có thể là một hoặc nhiều.

Các câu hỏi 10 và 11 xét một mô hình phân lớp dùng hàm $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$ cho giả thuyết phân lớp.

Câu 10 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây sai?

(A) Đây là hàm hồi quy logistic.
(B) X là tập dữ liệu mẫu.
(C) Đây là hàm sigmoid.
(D) $h_\theta(X)$ là xác suất để $Y =$ "1", với Y là thuộc tính nhãn và "1" là nhãn đang được quan tâm.

Câu 11 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây đúng?

(A) $h_\theta(X) \in [-1, 1]$.
(B) $h_\theta(X) \in \mathbb{R}$.
(C) $h_\theta(X) \in [0, 1]$.
(D) Không có phát biểu đúng.

Câu 12 [L.O.3.4]. Độ tin cậy của $A \to B$, ký hiệu bởi $confidence(A \to B)$, được định nghĩa là

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

Các câu hỏi 13–17 xét danh sách giao dịch dưới đây
(1) $I_1, I_2, I_3, I_4, I_5, I_6$
(2) $I_7, I_2, I_3, I_4, I_5, I_6$
(3) I_1, I_8, I_4, I_5
(4) $I_1, I_9, I_{10}, I_4, I_6$
(5) $I_{10}, I_2, I_4, I_{11}, I_5$

Câu 13 [L.O.3.4]. Danh sách có

(A) 11 giao dịch.
(B) 5 giao dịch.
(C) 6 giao dịch.
(D) 9 giao dịch.

Câu 14 [L.O.3.4, L.O.5.1]. Với $support = 0.6$, danh sách các mẫu (itemsets) xuất hiện thường xuyên là

(A) gồm tất cả các mẫu trong các phương án còn lại.
(B) $< I_1, I_4 >, < I_2, I_4 >, < I_2, I_5 >, < I_4, I_5 >, < I_4, I_6 >$.
(C) $< I_1 >, < I_2 >, < I_4 >, < I_5 >, < I_6 >$.
(D) $< I_2, I_4, I_5 >$.

Câu 15 [L.O.3.4]. Nếu giảm giá trị của *support* xuống, thì

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

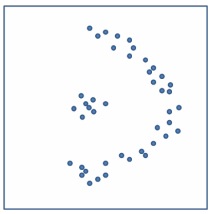
Câu 16 [L.O.3.4, L.O.5.1]. Các luật kết hợp có thể được khai phá với $support = 0.6$ và $confidence = 0.8$ là

(A) $< I_2, I_4 > \rightarrow I_1, < I_2, I_5 > \rightarrow I_3$.
(B) $< I_2, I_4 > \rightarrow I_5, < I_1, I_5 > \rightarrow I_2$.
(C) $< I_2, I_4 > \rightarrow I_5, < I_2, I_5 > \rightarrow I_4$.
(D) $< I_3, I_5 > \rightarrow I_4, < I_3, I_4 > \rightarrow I_5$.

Câu 17 [L.O.3.4]. Nếu tăng giá trị của $confidence$ xuống, thì

(A) một số luật kết hợp khác sẽ được thêm vào tập luật.
(B) một số luật kết hợp khác sẽ bị đưa ra khỏi tập luật.
(C) tập luật không thay đổi.
(D) không thể xác định số lượng luật trong tập luật.

Câu 18 [L.O.3.3]. Giải thuật nào thích hợp nhất để phân cụm tập điểm dữ liệu dưới đây, nếu sử dụng hàm khoảng cách Ơclit (Euclidean)?



(A) DBSCAN.
(B) k-medoids.
(C) k-means.
(D) Các giải thuật này cho kết quả tương tự.

Câu 19 [L.O.3.2]. Hàm độ đo nào thường được dùng với dữ liệu nhị phân?

(A) Mahattan.
(B) Eiuclidean.
(C) Jaccard.
(D) Minkowski.

Câu 20 [L.O.3.1]. Hàm $y = a\log(bx)$ là

(A) một hàm hồi quy tuyến tính.
(B) một hàm mất mát (loss function).
(C) một hàm sigmoid.
(D) một hàm hồi quy phi tuyến.

Câu 21 [L.O.3.3]. Trong giải thuật gom cụm trộn (agglomerative), các cụm ban đầu được xác đinh

(A) ngẫu nhiên.
(B) chính là các đối tượng dữ liệu.
(C) chính là tập các đối tượng dữ liệu.
(D) bởi k đối tượng dữ liệu ngẫu nhiên.

Câu 22 [L.O.3.4]. Một luật kết hợp được quan tâm nếu

(A) nó thoả mãn điều kiện về $min\_confidence$.
(B) nó thoả mãn đồng thời cả hai điều kiện về $min\_support$ và $min\_confidence$.
(C) nó thoả mãn điều kiện về $min\_support$.

Câu 23 [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) đánh giá luật kết hợp dạng $< A, B > \rightarrow A$.
(C) đo sự tương quan giữa hai sự kiện A và B.
(D) đánh giá luật kết hợp dạng $< A, B > \rightarrow B$.

Câu hỏi 24 và 25 xét mô hình phân lớp M thực hiện phân loại dữ liệu có ba nhãn A, B và C. Kết qủa phân loại được cho bởi ma trận confusion sau đây

	A	B	C
A	116	13	10
B	14	11	20
C	11	10	122

Câu 24 **[L.O.3.2].** Độ chính xác (precision) của việc phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập phân) là

(A) 0.832.
(B) 0.825.
(C) 0.823.
(D) 0.852.

Câu 25 **[L.O.3.2].** Độ truy hồi (recall) của việc phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập phân) là

(A) 0.892.
(B) 0.829.
(C) 0.289.
(D) 0.298.

Câu 26 **[L.O.3.3].** Có bao nhiêu cụm được sinh bởi giải thuật k-means?

(A) 2^k.
(B) Một bội số của k.
(C) e^k.
(D) k.

Câu 27 **[L.O.3.4].** Độ hỗ trợ của A, ký hiểu bởi $support(A)$, được định nghĩa là số giao dịch (transaction)

(A) không chứa A trên tổng số giao dịch.
(B) không chứa A.
(C) chứa A.
(D) chứa A trên tổng số giao dịch.

Câu 28 **[L.O.1].** Tri thức có thể thu được từ quá trình khai phá dữ liệu là

(A) Mô hình phân loại.
(B) Tập mẫu thường xuyên và tập luật.
(C) Mô hình phân cụm.
(D) Tất cả những phương án còn lại.

Câu 29 **[L.O.3.3].** Khoảng cách giữa các cụm dữ liệu C_i và C_j có thể được tính bởi

(A) Tất cả đều được.
(B) liên kết đầy đủ (complete link): $d(C_i, C_j) = \max\{d(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j\}$.
(C) liên kết đơn (single link): $d(C_i, C_j) = \min\{d(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j\}$.
(D) khoảng cách tâm (centroid): $d(C_i, C_j) = d(c_i, c_j)$, với c_i, c_j là tâm của C_i và C_j.

Câu 30 **[L.O.3.3, L.O.5.1].** Gọi ϵ là bán kính hình cầu lân cận của một điểm trong một tập dữ liệu $\mathcal{D}$ cho trước, ký hiệu $N_\epsilon(p) = \{q \in \mathcal{D} : d(p, q) \leq \epsilon\}$, trong đó $d(p, q)$ là khoảng cách giữa p và q. Gọi $MinPts$ là số điểm tối thiểu trong một lân cận của một điểm trong $\mathcal{D}$. Khi đó, nếu $p \in \mathcal{D}$ là một điểm nhân (core) thì

(A) $|N_\epsilon(p)| \leq MinPts$.
(B) $|N_\epsilon(p)|$ tuỳ ý.
(C) $|N_\epsilon(p)| = MinPts$.
(D) $|N_\epsilon(p)| \geq MinPts$.

Câu 31 [L.O.3.3, L.O.5.1]. Tự luận – *Phân cụm dữ liệu theo theo tiếp cận phân cấp*

Xét tập dữ liệu gồm 6 điểm (với 2 thuộc tính) được cho trong bảng dưới đây.

Điểm	x**-toạ độ**	y**-toạ độ**
p_1	0.4005	0.5306
p_2	0.2148	0.3854
p_3	0.3457	0.3156
p_4	0.2652	0.1875
p_5	0.0789	0.4139
p_6	0.4548	0.3022

Yêu cầu

(a) Xây dựng ma trận sai khác (khoảng cách) cho tập dữ liệu cho bởi bảng trên, biết rằng độ đo về sự sai khác (dissimilarity) giữa các điểm là khoảng cách Ơclit (Euclidean distance).

(b) Bằng giải thuật AGNES (Agglomerative Nesting) với độ đo single-link cho khoảng cách giữa các cụm, hãy xây dựng cấu trúc phân cấp cụm cho tập dữ liệu (biểu diễn dưới dạng biểu đồ Venn) và cây phả hệ (dendrogram) tương ứng.

Lời giải

Đại học Bách Khoa Tp. Hồ Chí Minh
Khoa Khoa học & Kỹ thuật Máy tính

THI CUỐI KỲ
Môn: **Khai phá dữ liệu** (CO3029)

Lớp: **20182** Nhóm: **LO2**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **07/06/2019**

Đáp án – Mã đề: 1823

Câu 1 (B)
Câu 2 (C)
Câu 3 (B)
Câu 4 (C)
Câu 5 (A)
Câu 6 (A)
Câu 7 (A)
Câu 8 (A)
Câu 9 (C)
Câu 10 (B)
Câu 11 (C)
Câu 12 (A)
Câu 13 (B)
Câu 14 (A)
Câu 15 (D)
Câu 16 (C)
Câu 17 (B)
Câu 18 (A)
Câu 19 (C)
Câu 20 (D)
Câu 21 (B)
Câu 22 (B)
Câu 23 (C)
Câu 24 (C)
Câu 25 (B)
Câu 26 (D)
Câu 27 (D)
Câu 28 (D)
Câu 29 (A)
Câu 30 (D)

Câu 31 [L.O.3.3, L.O.5.1]. Tự luận – *Phân cụm dữ liệu theo theo tiếp cận phân cấp*

Xét tập dữ liệu gồm 6 điểm (với 2 thuộc tính) được cho trong bảng dưới đây.

Điểm	x-**toạ độ**	y-**toạ độ**
p_1	0.4005	0.5306
p_2	0.2148	0.3854
p_3	0.3457	0.3156
p_4	0.2652	0.1875
p_5	0.0789	0.4139
p_6	0.4548	0.3022

Yêu cầu

(a) Xây dựng ma trận sai khác (khoảng cách) cho tập dữ liệu cho bởi bảng trên, biết rằng độ đo về sự sai khác (dissimilarity) giữa các điểm là khoảng cách Ơclit (Euclidean distance).

(b) Bằng giải thuật AGNES (Agglomerative Nesting) với độ đo single-link cho khoảng cách giữa các cụm, hãy xây dựng cấu trúc phân cấp cụm cho tập dữ liệu (biểu diễn dưới dạng biểu đồ Venn) và cây phả hệ (dendrogram) tương ứng.

Lời giải

(a) **Ma trận sai khác với khoảng cách Euclidean:**

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

(b) **Cấu trúc phân cấp cụm và cây phả hệ**