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: **20191** Nhóm: **LO1**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **21/12/2019**

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]. 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 3 [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 $< A, B > \rightarrow A$.
(D) đánh giá luật kết hợp dạng $< A, B > \rightarrow B$.

Câu 4 [L.O.3.3]. Trường hợp nào sau đây mà k-means sẽ cho kết quả phân cụm không tốt

(A) Tập dữ liệu bao gồm điểm ngoại biên (outlier).
(B) Các điểm dữ liệu phân bố với nhiều mật độ khác nhau.
(C) Tập dữ liệu có hình dạng không lồi (non-convex).
(D) Tất cả các đặc điểm này.

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.3.2]. Hàm độ đo nào thường được dùng với dữ liệu nhị phân?

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

Các câu hỏi 7–11 xét danh sách giao dịch dưới đây
(1) I_1, I_5, I_4, I_2
(2) I_3, I_1, I_5, I_4
(3) I_5, I_6
(4) I_4, I_3, I_6, I_5
(5) I_4, I_6, I_1
(5) I_2, I_6

Câu 7 [L.O.3.4]. Danh sách có

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

Câu 8 [L.O.3.4, L.O.5.1]. Với $support = 0.5$, danh sách các mẫu (itemsets) xuất hiện thường xuyên là

(A) $< I_3 >, < I_6 >, < I_1, I_4 >, < I_5, I_4 >$.
(B) $< I_4 >, < I_6 >, < I_1, I_4 >, < I_5, I_4 >$.
(C) $< I_4 >, < I_2 >, < I_1, I_4 >, < I_5, I_4 >$.
(D) $< I_4 >, < I_6 >, < I_1, I_5 >, < I_6, I_4 >$.

Câu 9 [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 10 [L.O.3.4, L.O.5.1]. Các luật kết hợp có thể được khai phá với $support = 0.5$ và $confidence = 0.7$ gồm

(A) $I_1 \to I_5, I_5 \to I_1, I_5 \to I_4, I_4 \to I_5$.
(B) $I_1 \to I_4, I_4 \to I_1, I_5 \to I_4, I_4 \to I_5$.
(C) $I_1 \to I_4, I_4 \to I_1, I_5 \to I_1, I_1 \to I_5$.
(D) $I_1 \to I_6, I_6 \to I_1, I_5 \to I_6, I_6 \to I_5$.

Câu 11 [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 12 [L.O.3.4]. Một luật kết hợp được quan tâm nếu nó thoả mãn

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

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.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 16 **[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.

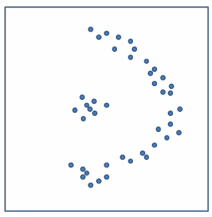
Câu 17 **[L.O.3.4].** Nguyên lý của giải thuật Apriori là

(A) Bất kỳ tập con của một tập tập mẫu xuất hiện thường xuyên thì không xuất hiện thường xuyên.
(B) Vét cạn để để đưa ra các mẫu xuất hiện thường xuyên.
(C) Bất kỳ tập con của một tập tập mẫu xuất hiện thường xuyên thì phải xuất hiện thường xuyên.

Câu 18 **[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 19 **[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 Euclidean (Óclit)?

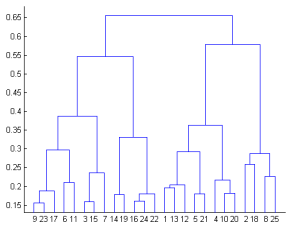


(A) DBSCAN.
(B) k-means.
(C) k-medoids.
(D) Các giải thuật này cho kết quả tương tự.

Câu 20 **[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) $\frac{support(A \cup B)}{support(B)}$.

Các câu hỏi 21 và 22 xét hình ảnh dưới đây.



Câu 21 [L.O.3.3, L.O.5.1]. Đây là hình ảnh minh hoạ cho phương pháp phân cụm nào?

(A) k-means.
(B) Phân cấp.
(C) DBSCAN.
(D) Apriori.

Câu 22 [L.O.3.3, L.O.5.1]. Số cụm thích hợp nhất for tập dữ liệu được biểu diễn bởi cây phả hệ (dendrogram) trong Câu 21là

(A) 2.
(B) 4.
(C) 6.
(D) 8.

Câu 23 [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 24 và 25 xét một mô hình phân lớp dùng hàm $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$ cho giả thuyết phân lớp.

Câu 24 [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 25 [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 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) Phân tích thành phần chính.
(C) Lấy mẫu dữ liệu.
(D) Kết hợp khối dữ liệu.

Câu 27 [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 28 [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 về sự hội tụ.

Câu 29 [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 30 [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.

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: **20191** Nhóm: **LO1**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **21/12/2019**

Đáp án – Mã đề: 1820

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

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: **20191** Nhóm: **LO1**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **21/12/2019**

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]. Nguyên lý của giải thuật Apriori là

(A) Bất kỳ tập con của một tập tập mẫu xuất hiện thường xuyên thì không xuất hiện thường xuyên.
(B) Bất kỳ tập con của một tập tập mẫu xuất hiện thường xuyên thì phải xuất hiện thường xuyên.
(C) Vét cạn để để đưa ra các mẫu xuất hiện thường xuyên.

Câu 2 [L.O.3.4]. Một luật kết hợp được quan tâm nếu nó thoả mãn

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

Câu hỏi 3 và 4 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 3 [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.

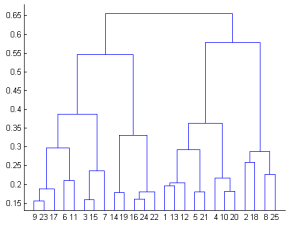
Câu 4 [L.O.3.2]. Độ truy hồi (recall) của việc phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập phân) là

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

Câu 5 [L.O.3.3]. Giải thuật k-means

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

Các câu hỏi 6 và 7 xét hình ảnh dưới đây.



Câu 6 [L.O.3.3, L.O.5.1]. Đây là hình ảnh minh hoạ cho phương pháp phân cụm nào?

(A) Apriori.
(B) k-means.
(C) Phân cấp.
(D) DBSCAN.

Câu 7 [L.O.3.3, L.O.5.1]. Số cụm thích hợp nhất for tập dữ liệu được biểu diễn bởi cây phả hệ (dendrogram) trong Câu 6là

(A) 8.
(B) 2.
(C) 4.
(D) 6.

Câu 8 [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 9 [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 10 [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.

Các câu hỏi 11–15 xét danh sách giao dịch dưới đây

(1) I_1, I_5, I_4, I_2
(2) I_3, I_1, I_5, I_4
(3) I_5, I_6
(4) I_4, I_3, I_6, I_5
(5) I_4, I_6, I_1
(5) I_2, I_6

Câu 11 [L.O.3.4]. Danh sách có

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

Câu 12 [L.O.3.4, L.O.5.1]. Với $support = 0.5$, danh sách các mẫu (itemsets) xuất hiện thường xuyên là

(A) $< I_4 >, < I_6 >, < I_1, I_5 >, < I_6, I_4 >$.
(B) $< I_3 >, < I_6 >, < I_1, I_4 >, < I_5, I_4 >$.
(C) $< I_4 >, < I_6 >, < I_1, I_4 >, < I_5, I_4 >$.
(D) $< I_4 >, < I_2 >, < I_1, I_4 >, < I_5, I_4 >$.

Câu 13 [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 14 [L.O.3.4, L.O.5.1]. Các luật kết hợp có thể được khai phá với $support = 0.5$ và $confidence = 0.7$ gồm

(A) $I_1 \to I_6, I_6 \to I_1, I_5 \to I_6, I_6 \to I_5$.
(B) $I_1 \to I_5, I_5 \to I_1, I_5 \to I_4, I_4 \to I_5$.
(C) $I_1 \to I_4, I_4 \to I_1, I_5 \to I_4, I_4 \to I_5$.
(D) $I_1 \to I_4, I_4 \to I_1, I_5 \to I_1, I_1 \to I_5$.

Câu 15 [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.

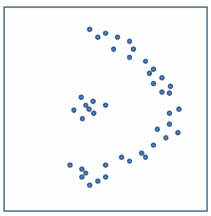
Câu 16 [L.O.3.3]. Trường hợp nào sau đây mà k-means sẽ cho kết quả phân cụm không tốt

(A) Tất cả các đặc điểm này.
(B) Tập dữ liệu bao gồm điểm ngoại biên (outlier).
(C) Các điểm dữ liệu phân bố với nhiều mật độ khác nhau.
(D) Tập dữ liệu có hình dạng không lồi (non-convex).

Câu 17 [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 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 Euclidean (Óclit)?



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

Câu 19 [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 20 [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 21 [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 22 [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 23 [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): $d(C_i, C_j) = \max\{d(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j\}$.

Câu 24 [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 $< A, B > \rightarrow A$.

Câu 25 [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.

Câu 26 [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 27 [L.O.3.4]. Độ tin cậy của $A \rightarrow B$, ký hiệu bởi $confidence(A \rightarrow B)$, được định nghĩa là

(A) $\frac{support(A \cup B)}{support(B)}$.
(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 28 [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.

Các câu hỏi 29 và 30 xét một mô hình phân lớp dùng hàm $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$ cho giả thuyết phân lớp.

Câu 29 [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 30 [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) $h_\theta(X) \in \mathbb{R}$.

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: **20191** Nhóm: **LO1**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **21/12/2019**

Đáp án – Mã đề: 1821

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

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: **20191** Nhóm: **LO1**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **21/12/2019**

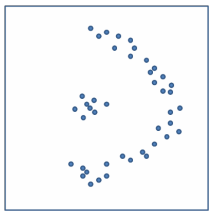
Họ tên sinh viên: _____ **Mã số sinh viên.:** | | | | | | | | |

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



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

Câu 2 [L.O.3.2]. Hàm độ đo nào thường được dùng với dữ liệu nhị phân?

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

Câu 3 [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 4 [L.O.3.3]. Trường hợp nào sau đây mà k-means sẽ cho kết quả phân cụm không tốt

(A) Tập dữ liệu bao gồm điểm ngoại biên (outlier).
(B) Tất cả các đặc điểm này.
(C) Các điểm dữ liệu phân bố với nhiều mật độ khác nhau.
(D) Tập dữ liệu có hình dạng không lồi (non-convex).

Câu 5 [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) $tO(kn)$.

Các câu hỏi 6–10 xét danh sách giao dịch dưới đây
(1) I_1, I_5, I_4, I_2
(2) I_3, I_1, I_5, I_4

(3) I_5, I_6
(4) I_4, I_3, I_6, I_5
(5) I_4, I_6, I_1
(5) I_2, I_6

Câu 6 [L.O.3.4]. Danh sách có

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

Câu 7 [L.O.3.4, L.O.5.1]. Với $support = 0.5$, danh sách các mẫu (itemsets) xuất hiện thường xuyên là

(A) $< I_3 >, < I_6 >, < I_1, I_4 >, < I_5, I_4 >$.
(B) $< I_4 >, < I_6 >, < I_1, I_5 >, < I_6, I_4 >$.
(C) $< I_4 >, < I_6 >, < I_1, I_4 >, < I_5, I_4 >$.
(D) $< I_4 >, < I_2 >, < I_1, I_4 >, < I_5, I_4 >$.

Câu 8 [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 9 [L.O.3.4, L.O.5.1]. Các luật kết hợp có thể được khai phá với $support = 0.5$ và $confidence = 0.7$ gồm

(A) $I_1 \to I_5, I_5 \to I_1, I_5 \to I_4, I_4 \to I_5$.
(B) $I_1 \to I_6, I_6 \to I_1, I_5 \to I_6, I_6 \to I_5$.
(C) $I_1 \to I_4, I_4 \to I_1, I_5 \to I_4, I_4 \to I_5$.
(D) $I_1 \to I_4, I_4 \to I_1, I_5 \to I_1, I_1 \to I_5$.

Câu 10 [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.

Các câu hỏi 11 và 12 xét một mô hình phân lớp dùng hàm $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$ cho giả thuyết phân lớp.

Câu 11 [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 12 [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 13 [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.

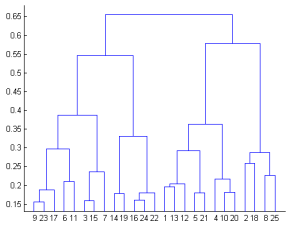
Câu 14 [L.O.4.4]. Để thu giảm dữ liệu, ta có thể sử dụng phương pháp

(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 15 [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): $d(C_i, C_j) = \max\{d(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j\}$.

Các câu hỏi 16 và 17 xét hình ảnh dưới đây.



Câu 16 [L.O.3.3, L.O.5.1]. Đây là hình ảnh minh hoạ cho phương pháp phân cụm nào?

(A) k-means.
(B) Apriori.
(C) Phân cấp.
(D) DBSCAN.

Câu 17 [L.O.3.3, L.O.5.1]. Số cụm thích hợp nhất for tập dữ liệu được biểu diễn bởi cây phả hệ (dendrogram) trong Câu 16là

(A) 2.
(B) 8.
(C) 4.
(D) 6.

Câu 18 [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 19 [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 20 [L.O.3.4]. Một luật kết hợp được quan tâm nếu nó thoả mãn

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

Câu 21 [L.O.3.4]. Nguyên lý của giải thuật Apriori là

(A) Vét cạn để để đưa ra các mẫu xuất hiện thường xuyên.
(B) Bất kỳ tập con của một tập tập mẫu xuất hiện thường xuyên thì không xuất hiện thường xuyên.
(C) Bất kỳ tập con của một tập tập mẫu xuất hiện thường xuyên thì phải xuất hiện thường xuyên.

Câu 22 [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 23 [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 về sự hội tụ.

Câu 24 [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 25 [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 26 [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 27 [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 hỏi 28 và 29 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 28 [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 29 [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 30 [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 $< A, B > \to A$.

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: **20191** Nhóm: **LO1**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **21/12/2019**

Đáp án – Mã đề: 1822

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

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: **20191** Nhóm: **LO1**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **21/12/2019**

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.3.3]. 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 2 [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.

Câu 3 [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 4 [L.O.3.3]. Giải thuật k-means

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

Câu 5 [L.O.3.4]. Nguyên lý của giải thuật Apriori là

(A) Vét cạn để để đưa ra các mẫu xuất hiện thường xuyên.
(B) Bất kỳ tập con của một tập tập mẫu xuất hiện thường xuyên thì phải xuất hiện thường xuyên.
(C) Bất kỳ tập con của một tập tập mẫu xuất hiện thường xuyên thì không xuất hiện thường xuyên.

Câu 6 [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 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) đánh giá luật kết hợp dạng $< A, B > \to A$.
(C) đo sự tương quan giữa hai sự kiện A và B.
(D) đánh giá luật kết hợp dạng $< A, B > \to B$.

Các câu hỏi 8–12 xét danh sách giao dịch dưới đây

(1) I_1, I_5, I_4, I_2
(2) I_3, I_1, I_5, I_4
(3) I_5, I_6
(4) I_4, I_3, I_6, I_5
(5) I_4, I_6, I_1
(5) I_2, I_6

Câu 8 [L.O.3.4]. Danh sách có

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

Câu 9 [L.O.3.4, L.O.5.1]. Với $support = 0.5$, danh sách các mẫu (itemsets) xuất hiện thường xuyên là

(A) $< I_3 >, < I_6 >, < I_1, I_4 >, < I_5, I_4 >$.
(B) $< I_4 >, < I_2 >, < I_1, I_4 >, < I_5, I_4 >$.
(C) $< I_4 >, < I_6 >, < I_1, I_4 >, < I_5, I_4 >$.
(D) $< I_4 >, < I_6 >, < I_1, I_5 >, < I_6, I_4 >$.

Câu 10 [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 11 [L.O.3.4, L.O.5.1]. Các luật kết hợp có thể được khai phá với $support = 0.5$ và $confidence = 0.7$ gồm

(A) $I_1 \to I_5, I_5 \to I_1, I_5 \to I_4, I_4 \to I_5$.
(B) $I_1 \to I_4, I_4 \to I_1, I_5 \to I_1, I_1 \to I_5$.
(C) $I_1 \to I_4, I_4 \to I_1, I_5 \to I_4, I_4 \to I_5$.
(D) $I_1 \to I_6, I_6 \to I_1, I_5 \to I_6, I_6 \to I_5$.

Câu 12 [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 13 [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 14 [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 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) Lấy mẫu dữ liệu.
(C) Phân tích thành phần chính.
(D) Kết hợp khối dữ liệu.

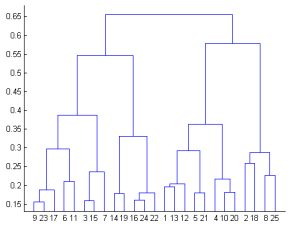
Câu 16 [L.O.3.3, L.O.5.1]. Gọi ϵ là bán kính hình cầu lân cận của một điểm trong một tập dữ liệu $\mathcal{D}$ cho trước, ký hiệu $N_\epsilon(p) = \{q \in \mathcal{D} : d(p,q) \leq \epsilon\}$, trong đó $d(p,q)$ là khoảng cách giữa p và q. Gọi $MinPts$ là số điểm tối thiểu trong một lân cận của một điểm 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 17 [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.

Các câu hỏi 18 và 19 xét hình ảnh dưới đây.



Câu 18 [L.O.3.3, L.O.5.1]. Đây là hình ảnh minh hoạ cho phương pháp phân cụm nào?

(A) k-means.
(B) DBSCAN.
(C) Phân cấp.
(D) Apriori.

Câu 19 [L.O.3.3, L.O.5.1]. Số cụm thích hợp nhất for tập dữ liệu được biểu diễn bởi cây phả hệ (dendrogram) trong Câu 18là

(A) 2.
(B) 6.
(C) 4.
(D) 8.

Các câu hỏi 20 và 21 xét một mô hình phân lớp dùng hàm $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$ cho giả thuyết phân lớp.

Câu 20 [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 21 [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 22 [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 23 [L.O.3.4]. Một luật kết hợp được quan tâm nếu nó thoả mãn

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

Câu 24 [L.O.3.3]. Trường hợp nào sau đây mà k-means sẽ cho kết quả phân cụm không tốt

(A) Tập dữ liệu bao gồm điểm ngoại biên (outlier).
(B) Tập dữ liệu có hình dạng không lồi (non-convex).
(C) Các điểm dữ liệu phân bố với nhiều mật độ khác nhau.
(D) Tất cả các đặc điểm này.

Câu hỏi 25 và 26 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 25 [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 26 [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.

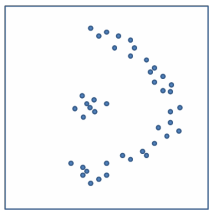
Câu 27 [L.O.3.4]. Độ tin cậy của $A \to B$, ký hiệu bởi $confidence(A \to B)$, được định nghĩa là

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

Câu 28 [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 29 [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 Euclidean (Ơclit)?



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

Câu 30 [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) $O(kt \log n)$.

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: **20191** Nhóm: **LO1**
Thời gian: **90 phút**
(***được xem tài liệu giấy***)
Ngày thi: **21/12/2019**

Đáp án – Mã đề: 1823

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