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.

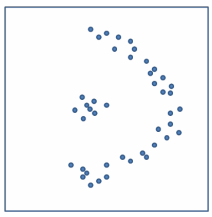
Câu 17 **[L.O.3.4].** Nguyên lý của giải thuật Apriori là

(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)?

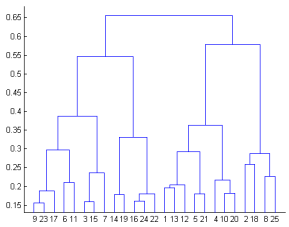


(A) DBSCAN.
(B) k-means.
(C) k-medoids.
(D) Các giải thuật này cho kết quả tương tự.

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.