Khoa Khoa học & Kỹ thuật Máy tính



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

 $\frac{\underline{\text{L\acute{o}p:}}}{\underline{\text{Th\acute{o}i gian:}}} \frac{20191}{\text{Ph\acute{o}i gian:}} \frac{\underline{\text{Nh\acute{o}m:}}}{\text{90 ph\acute{u}t}} \\ (\textit{\texttt{du\acute{o}c xem t\grave{a}i liệu giấy}})$

Ngày thi: 21/12/2019 Mã số sinh viên.: Họ tên sinh viên: Điểm: Người ra đề: Lê Hồng Trang Bằng chữ: Người coi thi: Đề thi gồm 30 câu trắc nghiêm (7 điểm) và 01 câu tư luân (3 điểm). Tô đâm phương án được chon trong phiếu trả lời và viết lời giải bài tư luân vào sau đề bài tương ứng. Câu 1 [L.O.3.2]. Mang nơ-ron nhân tao (ANN) là một mô hình tính toán: (B) tất cả những đặc điểm này. (A) thường được dùng cho bài toán phân lớp hay nhận dạng. (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 (\mathbf{A}) ngẫu nhiên. (B) chính là tập các đối tượng dữ liệu. (C) chính là các đối tương dữ liêu. (\mathbf{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 dang $A \to B$. (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 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) dư đoán. (D) mô tả dữ liêu. Câu 6 [L.O.3.2]. Hàm đô đo nào thường được dùng với dữ liệu nhi phân? (A) Mahattan. (B) Jaccard. C) Eiuclidean. 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

${f Câu}$ 9 [L.O.3.4]. Nếu giảm giá trị của $support$ xươ	ống, thì
A số mẫu (itemsets) xuất hiện thường xuyên vẫ B một số mẫu (itemsets) sẽ được đưa ra khỏi tặ C không xác định được tăng hay giảm số mẫu. D một số mẫu (itemsets) sẽ được thêm vào tập	ập xuất hiện thường xuyên hiện tại.
Câu 10 [L.O.3.4, L.O.5.1]. Các luật kết hợp có th ${\rm g\ddot{o}m}$	nể được khai phá với $support = 0.5$ và $confidence = 0.7$
	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$
${f Câu}$ 11 [L.O.3.4]. Nếu tăng giá trị của $confidence$	xuống, thì
 A một số luật kết hợp khác sẽ được thêm vào t B tập luật không thay đổi. C một số luật kết hợp khác sẽ bị đưa ra khỏi tậ D không thể xác định số lượng luật trong tập lu 	îp luật.
Câu 12 [L.O.3.4]. Một luật kết hợp được quan tâm	ı nếu nó thoả mãn
A điều kiện về $min_support$. B điều kiện về $min_confidence$. C đồng thời cả hai điều kiện về $min_support$ v Câu hỏi 13 và 14 xét mô hình phân lớp M thực hi loại được cho bởi ma trận confusion sau đây $ \frac{A B C}{A 116 13 10} $ $ B 14 11 20 $ $ C 11 10 122 $	r à $min_confidence$. iện phân $loại$ dữ liệu có ba nhãn $loại$ và $loace$. Kết qủa phân
Câu 13 [L.O.3.2]. Độ chính xác (precision) của việ thập phân) là	ệc phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số
(A) 0.832.(C) 0.825.	B 0.823.D 0.852.
Câu 14 [L.O.3.2]. Độ truy hồi (recall) của việc ph phân) là	ân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập
(A) 0.892. (C) 0.829.	B 0.289.D 0.298.
Trưởng bộ môn:	ã đề 1910 Trang 2

 $egin{array}{c} egin{array}{c} egin{array}{c} B & 4 & {
m giao} & {
m dich.} \\ \hline D & 7 & {
m giao} & {
m dich.} \\ \end{array}$

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

Câu 7 [L.O.3.4]. Danh sách có

A 5 giao dịch.
C 6 giao dịch.

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

$$\begin{array}{c|c} \textbf{A} & |N_{\epsilon}(p)| \leq MinPts. \\ \textbf{C} & |N_{\epsilon}(p)| \text{ tuỳ \acute{y}}. \end{array}$$

(B) $|N_{\epsilon}(p)| = MinPts$. (D) $|N_{\epsilon}(p)| \geq MinPts$.

$$(\mathbf{C}) |N_{\epsilon}(p)| \text{ tuỳ } \acute{\mathbf{v}}.$$

Câu 16 [L.O.3.4]. Đô hỗ trơ 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.
- (\mathbf{C}) không chứa A.
- (\mathbf{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 (Oclit)?



(A) DBSCAN.

(B) k-means.

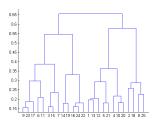
k-medoids.

(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

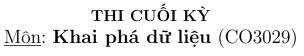
support(A) $support(A \cap B)$

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 h	oạ cho phương pháp phân cụm nào?
A k-means.C DBSCAN.	B Phân cấp. D Apriori.
Câu 22 [L.O.3.3, L.O.5.1]. Số cụm thích hợp nhất fo trong Câu 21là	or tập dữ liệu được biểu diễn bởi cây phả hệ (dendrogram)
(A) 2. (C) 6.	B 4. D 8.
Câu 23 [L.O.3.1]. Hàm $y = a \log(bx)$ là	
A một hàm hồi quy tuyến tính.C một hàm mất mát (loss function).	B một hàm sigmoid.D một hàm hồi quy phi tuyến.
$\it C\'{a}c$ $\it c\^{a}u$ $\it h\'{o}i$ 24 $\it v\`{a}$ 25 xét một mô hình phân lớp dì	ùng 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	
 A Đây là hàm hồi quy logistic. B Đây là hàm sigmoid. C X là tập dữ liệu mẫu. D h_θ(X) là xác suất để Y = "1", với Y là thuộc quan tâm. 	tính nhãn và "1" là nhãn đang được
Câu 25 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây	đúng?
$ \begin{array}{l} \textbf{A} h_{\theta}(X) \in [-1, 1]. \\ \textbf{C} h_{\theta}(X) \in \mathbb{R}. \end{array} $	$(\mathbf{B}) \ h_{\theta}(X) \in [0, 1].$ $(\mathbf{D}) \ \text{Không có phát biểu đúng.}$
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.(C) Lấy mẫu dữ liệu.	(B) Phân tích thành phần chính.(D) Kết hợp khối dữ liệu.
Câu 27 [L.O.3.3]. Khoảng cách giữa các cụm dữ liệu	ı C_i và C_j có thể được tính bởi
(A) Tất cả đều được. (B) liên kết đơn (single link): $d(C_i, C_j) = \min\{d(o_i)\}$ (C) liên kết đầy đủ (complete link): $d(C_i, C_j) = \min\{d(o_i)\}$ (D) khoảng cách tâm (centroid): $d(C_i, C_j) = d(c_i, c_j)$	$\operatorname{ax}\{d(o_{ip},o_{jq}):o_{ip}\in C_i,o_{jq}\in C_j\}.$
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 tượn sau t bước lặp thì thời gian tính toán là	ng, nếu giải thuật k -means kết thúc quá trình phân cụm
$ \begin{array}{c} \textbf{(A)} \ O(ktn). \\ \textbf{(C)} \ tO(kn). \end{array} $	$\begin{array}{c} \begin{array}{c} \begin{array}{c} \\ \\ \end{array} \begin{array}{c} \\ \end{array} B) \ kO(tn). \\ \\ \begin{array}{c} \\ \end{array} D) \ O(kt\log n). \end{array}$
Câu 30 [L.O.3.3]. Có bao nhiêu cụm được sinh bởi g	giải thuật k -means?
	$ \begin{array}{c} $

Khoa Khoa học & Kỹ thuật Máy tính





Lớp: **20191** $\underline{\text{Nh\'om}}$: LO1 Thời gian: 90 phút $(du \overline{\phi c} \ xem \ tài \ liệu \ giấy)$ Ngày thi: 21/12/2019

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

Khoa Khoa học & Kỹ thuật Máy tính



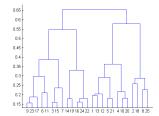
Trưởng bộ môn:....

THI CUỐI KỲ $\underline{\text{Môn}}$: Khai phá dữ liệu (CO3029)

Lớp: **20191** $\underline{\text{Nh\'om}}$: LO1 Thời gian: 90 phút $(du \overline{\phi c} \ xem \ tài \ liệu \ giấy)$

	$\underline{ ext{Ngà}}$	$\underline{\text{Ngày thi}}$: 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 điể phiếu trả lời và viết lời giải bài tự luận	êm) và 01 câu tự luận (3 điểm). Tô đậ n vào sau đề bài tương ứng.	ìm phương án được chọn trong	
Câu 1 [L.O.3.4]. Nguyên lý của giải th	nuật Apriori là		
thường xuyên.	mẫu xuất hiện thường xuyên thì không mẫu xuất hiện thường xuyên thì phải ất hiện thường xuyên.		
Câu 2 [L.O.3.4]. Một luật kết hợp đượ	ợc quan tâm nếu nó thoả mãn		
loại được cho bởi ma trận confusion sa	M thực hiện phân loại dữ liệu có ba n ${ m au}$ đây		
Câu 3 [L.O.3.2]. Độ chính xác (precis thập phân) là	ion) của việc phân loại dữ liệu thuộc l	lớp A (làm tròn đến 3 chữ số	
A) 0.852.C) 0.823.	B 0.832.D 0.825.		
Câu 4 [L.O.3.2]. Độ truy hồi (recall) c phân) là	của việc phân loại dữ liệu thuộc lớp A	(làm tròn đến 3 chữ số thập	
(A) 0.298.(C) 0.289.	B 0.892.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 ư 	u 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.
- (C) Phân cấp.

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

Câu 8 [L.O.3.2]. Mạng nơ-ron nhân tạo (ANN) là một mô hình tính toán:

- $m{(A)}$ số nút (node) đầu ra có thể là một hoặc
- (C) tất cả những đặc điểm này.

- (B) thường được dùng cho bài toán phân lớp hay nhận dạng.
- (D) mô phỏng cơ chế hoat đông của não người.

Câu 9 [L.O.3.1]. Hàm $y = a \log(bx)$ là

- (A) một hàm hồi quy phi tuyến.
- (C) một hàm sigmoid.

- B một hàm hồi quy tuyến tính.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.
- (\mathbf{B}) không chứa A trên tổng số giao dịch.
- chứa A.
- 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.
D 6 giao dịch.

 $\overline{\mathbf{C}}$ 4 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

- $(\mathbf{A}) < I_4>, < I_6>, < I_1, I_5>, < I_6, I_4>.$
- $\begin{array}{c} (A) > I_4 > , \\ (A) > I_6 > , \\ (A) > I_1, I_4 > , \\ (A) > I_6 > , \\ (A) > I_1, I_4 > , \\ (A) > I_6 > , \\ (A) > I_1, I_4 > , \\ (A) > I_6 > , \\ (A) > I_1, I_4 > , \\ (A) > I_6 > , \\ (A) > I_1, I_4 > , \\ (A) > I_6 > , \\ (A) > I_1, I_4 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_1, I_2 > , \\ (A) > I_2 > , \\ (A) > I_1, I_2 > , \\$

Câu 13 [L.O.3.4]. Nếu giảm giá trị của support xuống, thì (a) một số mẫn (itemsets) viất hiện thường xuyên viất hiện thường xuyên hiện tại. (b) số màn (itemsets) xuất hiện thường xuyên vớn giữ nguyên. (c) một số mẫn (tiemsets) số được đưa ta khởi tạ xuất hiện thường xuyên hiện tại. (d) không xác định được tăng hay giảm số mẫn. (a) không xác định được tăng hay giảm số mẫn. (b) không xác định được tăng hay giảm số mẫn. (c) hị - la, la - li, l ₅ - la, la - l ₅ . (a) li - la, la - li, l ₅ - la, la - l ₅ . (b) li - li, li		
 B số mẫu (itemsets) xuất hiện thường xuyên vẫn giữ nguyên. © một số mẫu (itemsets) sẽ được đươ ra khối tập xuất hiện thường xuyên hiện tại. B không xốc định được tăng hay giảm số mấu. 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 f₁→ f₀, f₀→ f₁, f₅→ f₀, f₀→ f₅. B f₁→ f₀, f₀→ f₁, f₅→ f₁, f₅→ f₀. I₁→ f₃, f₁→ f₁, f₂→ f₃, f₀→ f₅. I₁→ f₃, f₃→ f₁, f₂→ f₁, f₁→ f₅. I₁→ f₃, f₁→ f₁, f₂→ f₁, f₁→ f₅. I₁→ f₃, f₃→ f₁, f₂→ f₁, f₁→ f₅. Câu 15 [L.O.3.4]. Nêu 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. Câu 16 [L.O.3.3]. Trường hợp nào sau đây mà f₅-means sẽ cho kết quả phân cụm không tốt A 'Tất cả các đặc điểm này. B 'Tập dữ liệu bao gồm điểm ngoại biển (outlier). Câc điểm dữ liệu phān bố với nhiều mặt dộ 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 Tắt cả những phương án còn lại. Chai 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 (Oclit)? B DBSCAN. Câu 19 [L.O.3.1]. Hỗi quy logistic dùng để A mó tả đữ liệu. B phân lợp dữ liệu. B phân lợp dữ liệu. Cỳ 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 lạp thì thời gian tính toán là A O(kt log n). B O(ktn). Câu 21 [L.O.3.2]. Hàm độ do nào thưởng được dùng với dữ liệu nhị phân? A Minkowski. B Mahattan. 	Câu 13 [L.O.3.4]. Nếu giảm giá trị của $support$ xuốt	ng, thì
gồm (A) $I_1 \rightarrow I_6, I_6 \rightarrow I_1, I_5 \rightarrow I_6, I_6 \rightarrow I_5$. (B) $I_1 \rightarrow I_5, I_5 \rightarrow I_1, I_5 \rightarrow I_4, I_4 \rightarrow I_5$. (C) $I_1 \rightarrow I_4, I_4 \rightarrow I_1, I_5 \rightarrow I_4, I_4 \rightarrow I_6$. (D) $I_1 \rightarrow I_4, I_4 \rightarrow I_1, I_5 \rightarrow I_4, I_4 \rightarrow I_5$. (Câu 15 [L.O.3.4], Nếu tămg giá trị của cơn fidence xuống, thì (A) không thể xác định số hượng 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 được thêm vào tập luật. (Câu 16 [L.O.3.3], Tướng hợp nào sau đây mà k-means sẽ cho kết quả phân cụm không tốt (A) Tắt cả các đa điểm này. (B) Tập dữ liệu bao gồm điểm ngoại biên (outlier). (C) Các điểm dữ liệu phân bố với nhiều mặt độ khác nhau. (D) Tặp dữ liệu có hình dạng không lỗi (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) Tắt cả những phương ân còn lại. (C) Phân tích thành phân chính. (D) Lây mâu dữ liệu. (C) Phân tích thành phân chính. (D) Lây mâu dữ liệu. (E) Phân tích thành phân chính. (E) 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 (Oclit)? (E) A-means. (D) k-medoids. (Câu 19 [L.O.3.1], Hỗi quy logistic dùng để (A) mô tả dữ liệu. (E) phân cụm kết tập dữ liệu có n đối tượng, nếu giải thuật k-means kết thúc quá trình phân cụm sau t bước lập thì thời gian tính toàn là (A) O(kt log n). (E) kO(tm). (B) O(kt n). (Câu 21 [L.O.3.2], Hàm độ do nào thường được dùng với dữ liệu nhị phân? (A) Minkowski. (B) Mahattan.	B số mẫu (itemsets) xuất hiện thường xuyên vẫn C một số mẫu (itemsets) sẽ được đưa ra khỏi tập	n giữ nguyên.
Câu 15 [L.O.3.4]. Nếu tăng giá trị của confidence xuống, thì (A) không thể xác dịnh 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 khôi tập luật. (Câu 16 [L.O.3.3]. Trường hợp nào sau dây mà k-means sẽ cho kết quả phân cụm không tốt (A) Tất cả các đặc diễm nây. (B) Tập dữ liệu bào gồm điểm ngoại biên (outlier). (C) Các diễm dữ liệu phân bố với nhiều mặt độ khác nhau. (D) Tập dữ liệu có hình dạng không lỗi (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 phân chính. (D) Lấy mẫu dữ liệu dưới đây, nếu sử dụng hàm khoảng cách Euclidean (Oclit)? (A) Các giải thuật này cho kết quả tương tự. (B) DBSCAN. (C) k-meaus. (D) k-medoids. (Câu 19 [L.O.3.1]. Hồi quy logistic dùng để (A) mô tả đĩ 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 lập thì thời gian tính toán là (A) O(kt log n). (B) O(ktn). (Cau 21 [L.O.3.2]. Hàm độ đo nào thường được dùng với dữ liệu nhị phân? (A) Minkowski. (B) Mahattan.		ể được khai phá với $support = 0.5$ và $confidence = 0.7$
A không thế 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 khôi 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 Tắt cả các đặc điểm nây. D Tập dữ liệu bào gồm điểm ngoại biên (outlier). C Các điểm dữ liệu bào gồm điểm ngoại biên (outlier). C Các điểm dữ liệu phân bố với nhiều mặt độ khác nhau. D Tập dữ liệu có hình dạng không lỗi (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 Tắt cả những phương án còn lại. C Phân tích thành phân 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 đười đây, nếu sử dụng hàm khoảng cách Euclidean (Oclit)? A Các giải thuật này cho kết quả tương tự. B DBSCAN. C k-means. D k-medoids. Câu 19 [L.O.3.1]. Hỗi quy logistic dùng để A mó tả đữ liệu. B phân lớp 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 lạp thì thời gian tính toán là A O(kt log n). B O(ktn). C kO(tn). B Mahattan.		$\begin{array}{l} \textbf{(B)} \ \ I_1 \to I_5, I_5 \to I_1, I_5 \to I_4, I_4 \to I_5. \\ \textbf{(D)} \ \ I_1 \to I_4, I_4 \to I_1, I_5 \to I_1, I_1 \to I_5. \end{array}$
B một số luật kết hợp khác sẽ được thêm vào tặp luật. (C) tặp luật không thay đổi. (D) một số luật kết hợp khác sẽ bị đưa ra khỗi tập luật. (Cau 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 mặt độ khác nhau. (D) Tập dữ liệu có hình dạng không lỗi (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. (C) Phân tích thành phân chính. (D) Lấy mẫu dữ liệu. (C) Phân tích thành phân chính. (D) Lấy mẫu dữ liệu dưới đây, nếu sử dụng hàm khoảng cách Euclidean (Oclit)? (A) Các giải thuật này cho kết quả tương tự. (B) DBSCAN. (C) k-means. (D) k-medoids. (Cau 19 [L.O.3.1]. Hồi quy logistic dùng để (A) mô tả dữ liệu. (D) 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 lập thì thời gian tính toán là (A) O(kt log n). (B) O(kt n). (C) kO(tn). (B) Mahattan.	Câu 15 [L.O.3.4]. Nếu tăng giá trị của confidence	xuống, thì
A Tắt cả các đặc điểm này. B Tạp dữ liệu bao gồm điểm ngoại biên (outlier). C Các điểm dữ liệu phân bố với nhiều mặt độ khác nhau. D Tặp dữ liệu có hình dạng không lồi (non-convex). Câu 17 [L.O.4.4]. Dể thu giảm dữ liệu, ta có thể sử dụng phương pháp A Kết hợp khối dữ liệu. B Tắt 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 diễm dữ liệu dưới đây, nếu sử dụng hàm khoảng cách Euclidean (Oclit)? A) Các giải thuật này cho kết quả tương tự. C k-means. D k-medoids. Câu 19 [L.O.3.1]. Hỗi quy logistic dùng để A) mô tả dữ liệu. D phân lớp 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 lập thì thời gian tính toán là A) O(kt log n). B) O(ktn). C kO(tn). C lau 21 [L.O.3.2]. Hàm độ do nào thường được đùng với dữ liệu nhị phân? A) Minkowski. B) Mahattan.	B một số luật kết hợp khác sẽ được thêm vào tậj C tập luật không thay đổi.	p luật.
 Tặp dử liệu bao gồm diễm ngoại biên (outlier). Các điểm dữ liệu phân bố với nhiều mặt độ khác nhau. Tặp dữ liệu có hình dạng không lồi (non-convex). Câu 17 [L.O.4.4]. Dể 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 đãy, nếu sử dụng hàm khoảng cách Euclidean (Oclit)? A Các giải thuật này cho kết quả tương tự. B DBSCAN. C k-means. 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 doán. 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). Câu 21 [L.O.3.2]. Hàm độ đo nào thường được dùng với dữ liệu nhị phân? A Minkowski. B Mahattan. 	Câu 16 [L.O.3.3]. Trường hợp nào sau đây mà k -me	ans sẽ cho kết quả phân cụm không tốt
A Kết hợp khối dữ liệu. (C) Phân tích thành phần chính. (D) Lấy mẫu dữ liệu. (C) Phân tích thành phần chính. (D) Lấy mẫu dữ liệu. (C) Lấy mẫu dữ liệu dưới dây, nếu sử dụng hàm khoảng cách Euclidean (Oclit)? (A) Các giải thuật này cho kết quả tương tự. (B) DBSCAN. (C) k-means. (D) k-medoids. (C) k-means. (D) k-medoids. (E) Phân cụm dữ liệu. (E) phân cụm 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). (E) A Minkowski. (B) Mahattan.	B Tập dữ liệu bao gồm điểm ngoại biên (outlier) C Các điểm dữ liệu phân bố với nhiều mật độ kh	nác nhau.
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 (Oclit)? (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 (D) 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). (Câu 21 [L.O.3.2]. Hàm độ đo nào thường được dùng với dữ liệu nhị phân? (A) Minkowski. (B) Mahattan.	Câu 17 [L.O.4.4]. Để thu giảm dữ liệu, ta có thể sử	dụng phương pháp
khoảng cách Euclidean (Oclit)? 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.	=	\succeq
 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. 	· · · · · · · · · · · · · · · · · · ·	phân cụm tập điểm dữ liệu dưới đây, nếu sử dụng hàm
 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. 		
 (A) mô tả 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. 	\asymp	\supseteq
 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. 	Câu 19 [L.O.3.1]. Hồi quy logistic dùng để	
sau t bước lặp thì thời gian tính toán là	\simeq	
Câu 21 [L.O.3.2]. Hàm độ đo nào thường được dùng với dữ liệu nhị phân? (A) Minkowski. (B) Mahattan.	· · · · · ·	ng, nếu giải thuật k -means kết thúc quá trình phân cụm
(A) Minkowski. (B) Mahattan.		$\begin{array}{c} \textbf{(B)} \ O(ktn). \\ \textbf{(D)} \ tO(kn). \end{array}$
	Câu 21 [L.O.3.2]. Hàm độ đo nào thường được dùng	g với dữ liệu nhị phân?
	\sim	\simeq

Trang 3

Trưởng bộ môn:....

$egin{aligned} \left(egin{aligned} \mathbf{A} \right) N_{\epsilon}(p) &\geq MinPts. \\ \hline \mathbf{C} N_{\epsilon}(p) &= MinPts. \end{aligned}$	$\begin{array}{c c} \textbf{(B)} & N_{\epsilon}(p) \leq MinPts. \\ \hline \textbf{(D)} & N_{\epsilon}(p) \text{ tuỳ \acute{y}.} \end{array}$
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): $d(C_i, C_j) = d(c_i, c_i)$ B Tất cả đều được. C liên kết đơn (single link): $d(C_i, C_j) = \min\{d(o_i)\}$ D liên kết đầy đủ (complete link): $d(C_i, C_j) = \max\{d(o_i)\}$	$(p,o_{jq}):o_{ip}\in C_i,o_{jq}\in C_j\}.$
${f Câu}$ 24 [L.O.3.4]. Đại lượng $lift$ được định nghĩa bở	$\operatorname{fi}\ lift = rac{P(A\cup B)}{p(A)p(B)},$ được dùng để
Câu 25 [L.O.1]. Tri thức có thể thu được từ quá trìn	h khai phá dữ liệu là
A Tất cả những phương án còn lại. C Mô hình phân cụm.	B Mô hình phân loại.D Tập mẫu thường xuyên và tập luật.
Câu 26 [L.O.3.3]. Trong giải thuật gom cụm trộn (a	gglomerative), các cụm ban đầu được xác đinh
$oldsymbol{A}$ bởi k đối tượng dữ liệu ngẫu nhiên. $oldsymbol{C}$ chính là tập các đối tượng dữ liệu.	B ngẫu nhiên. D chính là các đối tượng dữ liệu.
Câu 27 [L.O.3.4]. Độ tin cậy của $A \to B$, ký hiệu bỏ	si $confidence(A \rightarrow B)$, được định nghĩa là
	$ \begin{array}{c} & \underline{support(A \cap B)} \\ & \underline{support(A)} \\ & \\ & \underline{D} \\ & \underline{support(A \cap B)} \\ & \underline{support(B)} \\ \end{array} . $
Câu 28 [L.O.3.3]. Có bao nhiêu cụm được sinh bởi g	jiải thuật k -means?
$\begin{array}{c} (\mathbf{A}) & k. \\ (\mathbf{C}) & e^k. \end{array}$	
$\it C\'{a}\it c$ $\it c\^{a}\it u$ $\it h\'{o}\it i$ $\it 29$ $\it v\`{a}$ $\it 30$ xét một mô hình phân lớp dù	ng 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_θ(X) là xác suất để Y = "1", với Y là thuộc quan tâm. B Đây là hàm hồi quy logistic. C Đây là hàm sigmoid. D X là tập dữ liệu mẫu. 	tính nhãn và "1" là nhãn đang được
Câu 30 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây	đúng?
(A) Không có phát biểu đúng. (C) $h_{\theta}(X) \in [0, 1]$.	$ \begin{array}{ll} $

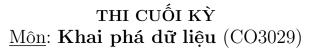
Câu 22 [L.O.3.3, L.O.5.1]. Gọi ϵ là bán kính hình cầu lân cận của một điểm trong một tập dữ liệu \mathcal{D} cho

(core) thì

Trưởng bộ môn:....

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

Khoa Khoa học & Kỹ thuật Máy tính





Lớp: **20191** $\underline{\text{Nh\'om}}$: $\mathbf{LO1}$ Thời gian: 90 phút $(du \overline{\not oc \ xem \ t} \grave{ai} \ li \hat{e}u \ gi \hat{a}y)$ Ngày thi: 21/12/2019

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

Khoa Khoa học & Kỹ thuật Máy tính



 $(2) I_3, I_1, I_5, I_4$

Trưởng bộ môn:....

THI CUỐI KỲ $\underline{\text{Môn}}\text{: }\mathbf{Khai}\ \mathbf{phá}\ \mathbf{dữ}\ \mathbf{liệu}\ (\text{CO3029})$

<u>Lớp</u>: **20191** <u>Nhóm</u>: **LO1**<u>Thời gian</u>: **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 đi ệ phiếu trả lời và viết lời giải bài tự luậ	ểm) và 01 câu tự luận (3 điểm). Tô đậ n vào sau đề bài tương ứng.	m phương án được chọn trong
Câu 1 [L.O.3.3]. Giải thuật nào thích khoảng cách Euclidean (Oclit)?	hợp nhất để phân cụm tập điểm dữ liệ	u dưới đây, nếu sử dụng hàm
A DBSCAN. C k-means.	$egin{array}{c} egin{array}{c} egin{array}$	cho kết quả tương tự.
Câu 2 [L.O.3.2]. Hàm độ đo nào thườ	ng được dùng với dữ liệu nhị phân?	
(A) Mahattan.(C) Jaccard.	(B) Minkowski.(D) Eiuclidean.	
Câu 3 [L.O.3.4]. Độ tin cậy của $A \rightarrow$	B , ký hiệu bởi $confidence(A \rightarrow B)$, đươ	ợc định nghĩa là
	$ \begin{array}{c} \textbf{B} \frac{support(A \cup B)}{support(B)}. \\ \textbf{D} \frac{support(A \cap B)}{support(B)}. \end{array} $	
Câu 4 [L.O.3.3]. Trường hợp nào sau	đây mà k -means sẽ cho kết quả phân cự	ım không tốt
 A Tập dữ liệu bao gồm điểm ngoại B Tất cả các đặc điểm này. C Các điểm dữ liệu phân bố với nh D Tập dữ liệu có hình dạng không l 	iều mật độ khác nhau.	
Câu 5 [L.O.3.3]. Với một tập dữ liệu sau t bước lặp thì thời gian tính	=: = =	kết thúc quá trình phân cụm
$ \begin{array}{c} \textbf{A} & O(ktn). \\ \textbf{C} & kO(tn). \end{array} $	$ \begin{array}{c} \textbf{B} \ O(kt \log n). \\ \textbf{D} \ tO(kn). \end{array} $	
Các câu hỏi 6 – 10 xét danh sách giao (1) I_1, I_5, I_4, I_2	dịch dưới đây	

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

Câu 6 [L.O.3.4]. Danh sách có

(A) 5 giao dịch.

B 7 giao dịch.
D 6 giao dịch.

C 4 giao dich.

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

- $\begin{array}{c} (\textbf{B}) < I_4>, < I_6>, < I_1, I_5>, < I_6, I_4>. \\ (\textbf{C}) < I_4>, < I_6>, < I_1, I_4>, < I_5, I_4>. \\ (\textbf{D}) < I_4>, < I_2>, < I_1, I_4>, < I_5, I_4>. \end{array}$

Câu 8 [L.O.3.4]. Nếu giảm giá tri 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.
- $(\widetilde{\mathbf{C}})$ một số mẫu (itemsets) sẽ được đưa ra khỏi tập xuất hiện thường xuyên hiện tại.
- (D) không xác định được tặng 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 = 0.7gồm

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.
- $(\overline{\mathbf{D}})$ 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.
- $(\mathbf{B}) h_{\theta}(X)$ 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.
- $(\mathbf{D})~X$ là tập dữ liệu mẫu.

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

(A) $h_{\theta}(X) \in [-1, 1].$ (C) $h_{\theta}(X) \in [0, 1].$

 $egin{aligned} egin{aligned} egin{aligned} egin{aligned} egin{aligned} egin{aligned} egin{aligned} egin{aligned} egin{aligned} h_{ heta}(X) \in \mathbb{R}. \end{aligned}$

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.
- (\mathbf{B}) chứa A trên tổng số giao dịch.
- \mathbf{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. C Phân tích thành phần chính.	B Kết hợp khối dữ liệu.D Lấy mẫu dữ liệu.
Câu 15 [L.O.3.3]. Khoảng cách giữa các cụm dữ l	liệu C_i và C_j có thể được tính bởi
A Tất cả đều được. B khoảng cách tâm (centroid): $d(C_i, C_j) = d(C_i, C_j)$ C liên kết đơn (single link): $d(C_i, C_j) = \min\{d(C_i, C_j) = d(C_i, C_j)$	$l(o_{ip}, o_{jq}) : o_{ip} \in C_i, o_{jq} \in C_j \}.$
Các câu hỏi 16 và 17 xét hình ảnh dưới đây.	
0.66 0.65 0.65 0.45 0.46 0.47 0.35 0.35 0.35 0.35 0.35 0.35 0.35 0.35	
Câu 16 [L.O.3.3, L.O.5.1]. Đây là hình ảnh minl	h hoạ cho phương pháp phân cụm nào?
	B Apriori. D DBSCAN.
Câu 17 [L.O.3.3, L.O.5.1]. Số cụm thích hợp nhất trong Câu 16là	t for tập dữ liệu được biểu diễn bởi cây phả hệ (dendrogram
(A) 2. (C) 4.	B 8.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. C chính là tập các đối tượng dữ liệu.	$egin{aligned} egin{aligned} egin{aligned\\ egin{aligned} egi$
Câu 19 [L.O.3.1]. Hàm $y = a \log(bx)$ là	
A một hàm hồi quy tuyến tính. C một hàm sigmoid.	B một hàm hồi quy phi tuyến.D một hàm mất mát (loss function).
Câu 20 [L.O.3.4]. Một luật kết hợp được quan tâ	m 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 	
Câu 21 [L.O.3.4]. Nguyên lý của giải thuật Aprio	ri là
 A Vét cạn để để đưa ra các mẫu xuất hiện thư B Bất kỳ tập con của một tập tập mẫu xuất thường xuyên. C Bất kỳ tập con của một tập tập mẫu xuất 	hiện thường xuyên thì không xuất hiện

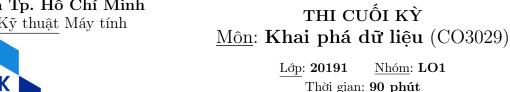
thường xuyên.

(A) Mô hình phân loại. (C) Mô hình phân cụm.	(B) Tất cả những phương án còn lại.(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 n	một mô hình tính toán:
 A thường được dùng cho bài toán phân lớp hay nhận dạng. C tất cả những đặc điểm này. 	 B) số nút (node) đầu ra có thể là một hoặc nhiều. 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.C phân cụm dữ liệu.	B mô tả dữ liệu. D dự đoán.
Câu 26 [L.O.3.3]. Có bao nhiêu cụm được sinh bởi g	giải thuật k -means?
$\begin{array}{c} \textbf{(A)} \ 2^k. \\ \textbf{(C)} \ e^k. \end{array}$	
trước, ký hiệu $N_{\epsilon}(p) = \{q \in \mathcal{D} : d(p,q) \leq \epsilon\}$, t	cầu lân cận của một điểm trong một tập dữ liệu \mathcal{D} cho trong đó $d(p,q)$ là khoảng cách giữa p và q . Gọi $MinPts$ lột điểm trong \mathcal{D} . Khi đó, nếu $p \in \mathcal{D}$ là một điểm nhân
$egin{aligned} igoldsymbol{igwedge} igl(oldsymbol{A} N_{\epsilon}(p) \leq MinPts. \ igl(oldsymbol{C} N_{\epsilon}(p) = MinPts. \end{aligned}$	$\begin{array}{c c} \hline \mathbf{B} & N_{\epsilon}(p) \geq MinPts. \\ \hline \mathbf{D} & N_{\epsilon}(p) \text{ tuỳ \acute{y}}. \end{array}$
$C\hat{a}u\ hỏi\ 28\ và\ 29$ xét mô hình phân lớp M thực hiệ loại được cho bởi ma trận confusion sau đây	n phân loại dữ liệu có ba nhãn A,B và $C.$ Kết qủa phân
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	
Câu 28 [L.O.3.2]. Độ chính xác (precision) của việc thập phân) là	phân loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số
(A) 0.832.(C) 0.823.	B 0.852.D 0.825.
Câu 29 [L.O.3.2]. Độ truy hồi (recall) của việc phân phân) là	n loại dữ liệu thuộc lớp A (làm tròn đến 3 chữ số thập
(A) 0.892.(C) 0.289.	B 0.298.D 0.829.
Câu 30 [L.O.3.4]. Đại lượng $lift$ được định nghĩa bở	$\mathrm{Si}\; lift=rac{P(A\cup B)}{p(A)p(B)},\; \mathrm{d}$ ược dùng để

 ${f Câu}$ 22 [L.O.1]. Tri thức có thể thu được từ quá trình khai phá dữ liệu là

Trưởng bộ môn:....

Khoa Khoa học & Kỹ thuật Máy tính



 $\underline{\text{Nh\'om}}$: LO1 Thời gian: 90 phút $(du \overline{\phi c} \ xem \ tài \ liệu \ giấy)$ Ngày thi: 21/12/2019

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

Khoa Khoa học & Kỹ thuật Máy tính



Trưởng bộ môn:....

THI CUỐI KỲ $\underline{\text{Môn}}\text{: }\mathbf{Khai}\ \mathbf{phá}\ \mathbf{dữ}\ \mathbf{liệu}\ (\text{CO3029})$

 $\frac{\text{L\acute{o}p: 20191}}{\text{Th\acute{o}i gian: 90 ph\acute{u}t}} \frac{\text{Nh\acute{o}m: LO1}}{\text{du\acute{o}c xem t\grave{a}i liệu giấy}} \\ \text{Ngày thi: 21/12/2019}$

Họ tên sinh viên:	Mã số sinh viên.:			
Điểm:	Người ra đề: Lê Hồng Trang			
Bằng chữ:	Người coi thi:	Người coi thi:		
Đề thi gồm 30 câu trắc nghiệm (7 điểm) v phiếu trả lời và viết lời giải bài tự luận vào		đậm phương án được chọn tron		
Câu 1 [L.O.3.3]. Trong giải thuật gom cụm	trộn (agglomerative), các cụm	ban đầu được xác đinh		
A) ngẫu nhiên.C) chính là tập các đối tượng dữ liệu.	\simeq	$egin{aligned} \mathbf{B} \end{aligned}$ chính là các đối tượng dữ liệu. $egin{aligned} \mathbf{D} \end{aligned}$ bởi k đối tượng dữ liệu ngẫu nhiên.		
Câu 2 [L.O.3.2]. Mạng nơ-ron nhân tạo (Al	NN) là một mô hình tính toán:			
A thường được dùng cho bài toán phân l	ớp B mô phỏng cơ chế	B mô phỏng cơ chế hoạt động của não người.		
hay nhận dạng. C tất cả những đặc điểm này.	D số nút (node) đầ nhiều.	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.C một hàm sigmoid.	\simeq	B) một hàm mất mát (loss function).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 B) không chắn chắn về sự hội tụ. C) luôn dừng tại điểm tối toàn cục. 	phương.			
Câu 5 [L.O.3.4]. Nguyên lý của giải thuật A	Apriori là			
 A) Vét cạn để để đưa ra các mẫu xuất hiệ B) Bất kỳ tập con của một tập tập mẫu thường xuyên. C) Bất kỳ tập con của một tập tập mẫu x thường xuyên. 	xuất hiện thường xuyên thì p			
Câu 6 [L.O.1]. Tri thức có thể thu được từ	quá trình khai phá dữ liệu là			
(A) Mô hình phân loại.(C) Mô hình phân cụm.	(B) Tập mẫu thường(D) Tất cả những pho	xuyên và tập luật. ương án còn lại.		
${f Câu}$ 7 [L.O.3.4]. Đại lượng $lift$ được định n	nghĩa bởi $lift = \frac{P(A \cup B)}{p(A)p(B)}$, được	dùng để		
(A) đánh giá luật kết hợp dạng $A \to B$. (C) đo sự tương quan giữa hai sự kiện A và	\cong	hợp dạng $< A, B > \rightarrow A$. hợp dạng $< A, B > \rightarrow B$.		

Các câu hỏi $8-12$ xét danh sách giao dịch c (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	lưới đây
Câu 8 [L.O.3.4]. Danh sách có	
A 5 giao dịch.C 4 giao dịch.	B 6 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
${f C\hat{a}u~10~[L.O.3.4]}$. Nếu giảm giá trị của $supp$	ort xuống, thì
Câu 11 [L.O.3.4, L.O.5.1]. Các luật kết hợ gồm	p có thể được khai phá với $support = 0.5$ và $confidence = 0.7$
$ \begin{array}{c} \textbf{(A)} \ I_1 \to I_5, I_5 \to I_1, I_5 \to I_4, I_4 \to I_5. \\ \textbf{(C)} \ I_1 \to I_4, I_4 \to I_1, I_5 \to I_4, I_4 \to I_5. \end{array} $	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$
${f Câu~12~[L.O.3.4]}.$ Nếu tăng giá trị của $conf$	idence xuống, thì
 A một số luật kết hợp khác sẽ được thêm B một số luật kết hợp khác sẽ bị đưa ra C tập luật không thay đổi. D không thể xác định số lượng luật trong 	khỏi tập luật.
Câu 13 [L.O.3.4]. Độ hỗ trợ của A , ký hiểu l	oởi $support(A)$, 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 cụn	n dữ liệu C_i và C_j có thể được tính bởi
(A) Tất cả đều được. (B) liên kết đầy đủ (complete link): $d(C_i, C_j) = 1$ (C) liên kết đơn (single link): $d(C_i, C_j) = 1$ (D) khoảng cách tâm (centroid): $d(C_i, C_j)$	
$\mathbf{C\hat{a}u}$ 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.C Phân tích thành phần chính.	B Lấy mẫu dữ liệu.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			
$egin{aligned} \mathbf{A} & N_{\epsilon}(p) \leq MinPts. \ \mathbf{C} & N_{\epsilon}(p) = MinPts. \end{aligned}$	$\begin{array}{c c} \hline \mathbf{B} & N_{\epsilon}(p) \text{ tuỳ \acute{y}}. \\ \hline \mathbf{D} & N_{\epsilon}(p) \geq MinPts. \end{array}$		
Câu 17 [L.O.3.3]. Có bao nhiêu cụm được sinh bởi giải thuật k -means?			
$\begin{array}{c} \textbf{(A)} \ 2^k. \\ \textbf{(C)} \ e^k. \end{array}$			
$C\acute{a}c$ $c\^{a}u$ $h\acute{o}i$ 18 $v\grave{a}$ 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.

(C) Phân cấp.

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

B 6. D 8.

 $C\acute{a}c$ $c\^{a}u$ hỏi 20 $v\grave{a}$ 21 xét một mô hình phân lớp dùng hàm $h_{\theta}(X)=\frac{1}{1+e^{-\theta^TX}}$ 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.
- \bigcirc X là tập dữ liệu mẫu.
- $\overline{\mathbf{C}}$ Đây là hàm sigmoid.
- $\stackrel{\circ}{\mathbb{D}}$ $h_{ heta}(X)$ 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) $h_{\theta}(X) \in [-1, 1].$ (C) $h_{\theta}(X) \in [0, 1].$

 $egin{aligned} \mathbf{B} & h_{ heta}(X) \in \mathbb{R}. \\ \mathbf{D} & \text{Không có phát biểu đúng.} \end{aligned}$

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

 (\mathbf{A}) phân lớp dữ liệu.

(C) phân cụm dữ liệu.

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

- (\mathbf{A}) điều kiện về $min\ confidence$.
- (\mathbf{B}) đồng thời cả hai điều kiện về $min_support$ và $min_confidence$.
- (\mathbf{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
\overline{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. (C) 0.289. (B) 0.829. (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

Câu 28 [L.O.3.2]. Hàm đô đo nào thường được dùng với dữ liêu nhi phân?

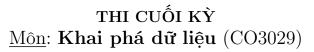
(A) Mahattan.(B) Eiuclidean.(C) Jaccard.(D) Minkowski.

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 (Oclit)?



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

Khoa Khoa học & Kỹ thuật Máy tính





 $\frac{\text{L\acute{o}p: 20191}}{\text{Th\acute{o}i gian: 90 ph\acute{u}t}} \frac{\text{Nh\acute{o}m: LO1}}{\text{d}\textit{u\acute{o}c xem t\grave{a}i liệu giấy}}$ Ngày thi: 21/12/2019

Câu 1 B	Câu 12 B	Câu 22 (A)
Câu 2 C	Câu 13 (D)	Câu 23 B
Câu 3 D	Câu 14 (A)	Câu 24 (D)
Câu 4 (A)	Câu 15 (A)	<u> </u>
Câu 5 B	Câu 16 (D)	Câu 2 5 🕜
Câu 6 D	Câu 17 (D)	Câu 26 (B)
Câu 7 C		
_	Câu 18 (C)	Câu 27 (A)
Câu 8 B	Câu 19 🕜	G2 23 G
Câu 9 (A)		Câu 28 (C)
Câu 10 (D)	Câu 20 (B)	Câu 29 🛕
	Câu 21 \bigcirc	G2 22 A
Câu 11 (C)		Câu 30 (A)