



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 chắc 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 tương tự ngoài cụm cao. (B) tính tương tự trong cụm cao và tính tương tự ngoài cụm thấp.
- (C) tính tương tự trong cụm thấp và tính tương tự ngoài cụm thấp. (D) tính tương tự trong cụm thấp và tính tương 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 $A \rightarrow B$. (B) đo sự tương quan giữa hai sự kiện A và B .
- (C) đánh giá luật kết hợp dạng $\langle A, B \rangle \rightarrow A$. (D) đánh giá luật kết hợp dạng $\langle A, B \rangle \rightarrow B$.

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.

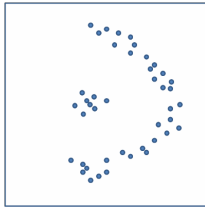
Câu 9 [L.O.3.2, L.O.5.1]. Phát biểu nào dưới đây đúng?

- ☐ (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 hiệu 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) Euclidean.
- ☐ (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 quả 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.
- ☒ (C) 0.829.
- ☐ (B) 0.289.
- ☐ (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) $\langle I_1 \rangle, \langle I_2 \rangle, \langle I_4 \rangle, \langle I_5 \rangle, \langle I_6 \rangle$.
☐ (C) $\langle I_1, I_4 \rangle, \langle I_2, I_4 \rangle, \langle I_2, I_5 \rangle, \langle I_4, I_5 \rangle, \langle I_4, I_6 \rangle$.
☐ (D) $\langle I_2, I_4, I_5 \rangle$.

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) $\langle I_2, I_4 \rangle \rightarrow I_1, \langle I_2, I_5 \rangle \rightarrow I_3$. ☒ (B) $\langle I_2, I_4 \rangle \rightarrow I_5, \langle I_2, I_5 \rangle \rightarrow I_4$.
☐ (C) $\langle I_2, I_4 \rangle \rightarrow I_5, \langle I_1, I_5 \rangle \rightarrow I_2$. ☐ (D) $\langle I_3, I_5 \rangle \rightarrow I_4, \langle I_3, I_4 \rangle \rightarrow I_5$.

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 hỗ 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 hỗ 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) $|N_\epsilon(p)|$ tùy ý.
 (D) $|N_\epsilon(p)| \geq MinPts$.

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

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

Câu 25 [L.O.3.4]. Một luật kết hợp được quan tâm nếu

- (A) nó thỏa mãn điều kiện về $min_support$.
 (B) nó thỏa mãn điều kiện về $min_confidence$.
 (C) nó thỏa 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 định

- (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.

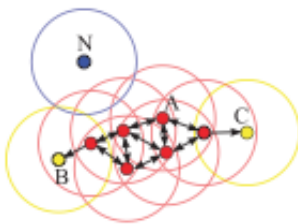
Câu 27 [L.O.3.3]. Với một tập dữ liệu có n đối tượng, nếu giải thuật k -means kết thúc quá trình phân cụm sau t bước lặp thì thời gian tính toán là

- (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 họa 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

- 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).
- 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



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



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 định

- (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 $\langle A, B \rangle \rightarrow A$. (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) dự đoán. (D) mô tả dữ liệu.

Câu 6 [L.O.3.2]. Hàm đo nào thường được dùng với dữ liệu nhị phân?

- (A) Mahattan. (B) Jaccard.
- (C) Euclidean. (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) $\langle I_3 \rangle, \langle I_6 \rangle, \langle I_1, I_4 \rangle, \langle I_5, I_4 \rangle$.
(B) $\langle I_4 \rangle, \langle I_6 \rangle, \langle I_1, I_4 \rangle, \langle I_5, I_4 \rangle$.
(C) $\langle I_4 \rangle, \langle I_2 \rangle, \langle I_1, I_4 \rangle, \langle I_5, I_4 \rangle$.
(D) $\langle I_4 \rangle, \langle I_6 \rangle, \langle I_1, I_5 \rangle, \langle I_6, I_4 \rangle$.

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

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 quả 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) $|N_\epsilon(p)|$ tùy ý.
 (D) $|N_\epsilon(p)| \geq MinPts$.

Câu 16 [L.O.3.4]. Độ hỗ trợ của A , ký hiệu 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.

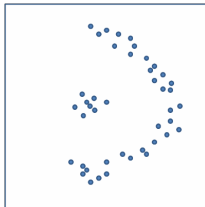
Câu 17 [L.O.3.4]. Nguyên lý của giải thuật Apriori là

- (A) Bất kỳ tập con của một tập tập mẫu xuất hiện thường xuyên thì không xuất hiện thường xuyên.
 (B) Vết cạn 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.
 (D) Tất cả những phương án còn lại.

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 (Ơclit)?

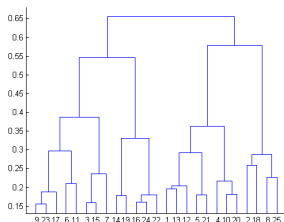


- (A) DBSCAN.
 (B) k -means.
 (C) k -medoids.
 (D) Các giải thuật này cho kết quả tương tự.

Câu 20 [L.O.3.4]. Độ tin cậy của $A \rightarrow B$, ký hiệu bởi $confidence(A \rightarrow B)$, được định nghĩa là

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

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 họa 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 21 là

- (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 chắc 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 .



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 11 (C) | Câu 21 (B) |
| Câu 2 (C) | Câu 12 (C) | 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 15 (D) | Câu 25 (B) |
| Câu 6 (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) |

Đề thi môn Khai Phá Dữ Liệu
HK1/2018-2019 - Thời gian: 90 phút
MSMH: CO3029 - Ngày thi: 21/12/2018

(Đề thi gồm 6 trang. Sinh viên làm phần trắc nghiệm trên phiếu trả lời trắc nghiệm, phần tự luận ngay trên đề thi và nộp lại)

(Sinh viên được phép tham khảo tài liệu giấy)

Họ và Tên	
MSSV	

Phần 1. Trắc nghiệm (7.0 điểm): Chọn 1 câu trả lời đúng nhất và tô vào phiếu trả lời trắc nghiệm

1. Trong giải thuật *Apriori*

- a. $|C_k| \geq |L_k|$
- b. $|C_k| \geq |C_{k+1}|$
- c. tập dữ liệu D sẽ được quét m lần với m là chiều dài của tập thường xuyên xuất hiện (frequent itemset) dài nhất

d. câu a và c đều đúng

2. Để kiểm tra giải thuật *gradient descent* với mục tiêu là cực tiểu hóa hàm chi phí $J(\theta)$ có hội tụ hay không ta cần kiểm tra:

- a. $J(\theta)$ có giảm ở mỗi bước lặp
- b. $J(\theta)$ có tăng ở mỗi bước lặp
- c. $J(\theta) = 0$ sau 10,000 lần lặp
- d. hệ số học α có được thiết lập đủ lớn, ví dụ bằng 0.1

3. Khi phân loại dữ liệu dùng *cây quyết định*, độ đo nào sau đây giúp tránh tạo ra các phân hoạch có quá ít đối tượng

- a. Information Gain
- b. GainRatio
- c. GiniIndex
- d. tất cả các câu trên đều sai

4. Phương pháp gom cụm nào sau đây giúp phát hiện được các cụm có dạng hình ống (pipe) tốt nhất

- a. K-Means
- b. K-Medoids
- c. DBSCAN
- d. BIRCH

5. Trong kỹ thuật gom cụm dựa vào mật độ, phát biểu nào sau đây đúng:

- a. trong cụm chỉ có một core object, đó là trung tâm cụm
- b. mỗi phần tử trong một cụm có ít nhất $MinPts$ phần tử khác gần nó (trong phạm vi bán kính là ϵ)

c. khoảng cách từ một phần tử a đến một core object nào đó nhỏ hơn ϵ thì a thuộc về cụm

d. tất cả các câu trên đều sai

6. Phát biểu nào sau đây ĐÚNG trong khai phá luật kết hợp:

- a. support có ý nghĩa quan trọng hơn confidence
- b. $support_count(A \Rightarrow B)$ là số lần xuất hiện đồng thời của A và B trong tập dữ liệu D
- c. $support(A \Rightarrow B)$ luôn lớn hơn $confidence(A \Rightarrow B)$
- d. tất cả các câu trên đều sai

7. Giải thuật FP-Growth

- a. quét tập dữ liệu D (tập dữ liệu lớn) m lần với m là số dòng trong header table
- b. thường chạy chậm hơn giải thuật Apriori
- c. tập hợp các node trên một nhánh của FP-tree phải xuất hiện ít nhất k lần trong D , với k là số đếm (count) của node lá trong nhánh đang xét
- d. tất cả các câu trên đều sai

Dữ kiện dưới đây dùng cho 3 câu sau đây:

Cho T chứa 500,000 giao dịch trong đó số giao dịch chứa bánh mì, chứa mít và chứa đồng thời bánh mì và mít lần lượt là 20000, 30000 và 10000.

8. Độ hỗ trợ (support) của phát biểu "ai mua mít đều sẽ mua bánh mì" là:

- a. 2%
- b. 33.33%
- c. 50%
- d. Tất cả các câu trên đều sai

9. Độ tin cậy (confidence) của phát biểu "ai mua mít đều sẽ mua bánh mì" là:

- a. 66.66%
- b. 33.33%
- c. 45%
- d. 50%

10. Khi số lượng giao dịch trong T tăng lên 10,000,000 nhưng số lượng giao dịch mua mít và bánh mì nêu ở trên không đổi thì phát biểu "ai mua mít đều sẽ mua bánh mì" sẽ

- a. thay đổi độ hỗ trợ
- b. thay đổi độ tin cậy
- c. cả độ hỗ trợ và độ tin cậy đều thay đổi
- d. tất cả các câu trên đều sai

11. Sau khi chạy giải thuật FP-Growth trên tập dữ liệu D, trong tập kết quả có một số tập thường xuyên xuất hiện có chiều dài là 5. Giải thuật FG-Growth này đã quét (scan) qua D

- a. 1 lần
- b. 2 lần
- c. 5 lần
- d. ít nhất là 5 lần

12. Logistic regression là một phương pháp dùng để

- a. dự đoán (prediction)
- b. phân lớp (classification)
- c. mô tả dữ liệu (description)
- d. gom cụm dữ liệu (clustering)

13. Phát biểu nào sau đây SAI trong phân lớp dữ liệu

- a. dữ liệu huấn luyện luôn phải chứa nhãn (label)
- b. dữ liệu kiểm tra luôn phải chứa nhãn
- c. dữ liệu kiểm tra không cần phải chứa nhãn vì đây là tập được dùng để kiểm tra mô hình và nhãn sẽ được tạo ra từ mô hình
- d. dữ liệu huấn luyện và kiểm tra phải có cấu trúc giống nhau

14. Kỹ thuật gom cụm nào sau đây khởi động bằng cách xem mỗi đối tượng dữ liệu là một cụm

- a. K-Means
- b. phân hoạch (partition)
- c. trộn (agglomerative) dữ liệu dựa vào cây phân cấp

d. phân cụm dựa vào mật độ

15. Trong web mining, để hiểu được thứ tự các URL được truy cập, ta thường dùng phương pháp nào

- a. phân tích chuỗi tuần tự (sequential analysis)
- b. khai phá luật kết hợp (association rule)
- c. phân lớp (classification)
- d. phân tích tương quan (correlation analysis)

16. Các mẫu điều kiện cơ sở (conditional pattern base) được tạo ra

- a. cho mỗi frequent item trong header table
- b. bằng cách duyệt cây FP-Tree (từ dưới lên), xuất phát từ node đầu tiên trong danh sách node link của item đang xét và phải duyệt hết các node trong danh sách này

c. hai câu a và b đúng

d. tất cả các câu trên đều sai

17. Phát biểu nào sau đây về gom cụm dữ liệu là SAI

- a. khoảng cách giữa các phần tử trong cùng một cụm càng nhỏ càng tốt
- b. khoảng cách giữa các phần tử ở các cụm khác nhau càng nhỏ càng tốt
- c. mô hình gom cụm tốt khi nó phát hiện được các cụm có hình dạng bất kỳ
- d. giải thuật K-means thường cho kết quả là các cụm có dạng hình cầu và có kích thước gần giống nhau

18. Hồi qui tuyến tính có thể được dùng để

- a. xử lý dữ liệu bị nhiễu
- b. dự đoán giá trị dữ liệu số
- c. phân lớp dữ liệu có nhãn (classification)
- d. câu a và b đúng

Dữ liệu sau đây dùng cho **hai** câu sau:

Một mô hình phân lớp (classifier) dùng hàm sau

$$h_{\theta}(X) = \frac{1}{1 + e^{-\theta^T X}}$$

làm giả thuyết (hypothesis) cho việc phân lớp.

19. Phát biểu nào sau đây SAI

- a. X là tập dữ liệu mẫu
- b. đây là hàm hồi qui logistic
- 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 mà ta quan tâm)

20. Phát biểu nào sau đây ĐÚNG

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

- b. $h_0(X) \in [0, 1]$
 c. X là vector các thuộc tính đầu vào (input features) của tập dữ liệu mẫu (bao gồm $X_0=1$)
 d. **hai câu b và c đúng**

Cho bộ phân lớp M thực hiện việc phân loại dữ liệu có ba nhãn A, B và C. Kết quả phân loại được biểu diễn bởi ma trận sai biệt (confusion matrix) như sau. Hãy chọn câu trả lời đúng cho **hai** câu hỏi sau đây.

Phân lớp thành Thực tế	A	B	C
A	116	13	10
B	14	11	20
C	11	10	122

21. Độ chính xác (precision) của việc phân loại dữ liệu thuộc lớp A là (làm tròn đến 3 chữ số thập phân):

- a. **0.823**
 b. 0.835
 c. 0.803
 d. 0.745

22. Độ truy hồi (recall) của việc phân loại dữ liệu thuộc lớp A là (làm tròn đến 3 chữ số thập phân):

- a. 0.752
 b. 0.835
 c. 0.803
 d. **0.829**

23. Weka KHÔNG hỗ trợ chức năng nào sau đây?

- a. xây dựng (train) mô hình, lưu trữ mô hình và sử dụng lại mô hình đó để thực thi với dữ liệu mới
 b. lựa chọn các thuộc tính dựa vào tương quan giữa các thuộc tính độc lập với thuộc tính phụ thuộc (ví dụ thuộc tính phân lớp)
 c. đọc dữ liệu có định dạng file là ARFF
 d. **tất cả các câu trên đều sai**

24. Phát biểu nào sau đây SAI về mạng nơ-ron nhân tạo - Artificial Neural Network (ANN)

- a. hàm kích hoạt (activation function) thường được dùng là hàm sigmoid
 b. có thể có nhiều hơn một lớp ẩn (hidden layer)
 c. **việc tìm trọng số (weight) cho các liên kết được thực hiện dựa trên phương pháp feedforward**
 d. việc chọn hệ số học (learning rate) sẽ ảnh hưởng đến tốc độ cũng như khả năng hội tụ của giải thuật

25. Độ đo nào được dùng đối với các dữ liệu nhị phân

- a. Manhattan
 b. **Jaccard**
 c. Euclidean
 d. Minkowski

26. Gọi $R_{A,B}$ là sự tương quan giữa hai thuộc tính A và B trong tập dữ liệu D, phát biểu nào sau đây SAI

- a. $R_{A,B} \in [-1, 1]$
 b. $R_{A,B} = 1$ thì ta nên loại một trong hai thuộc tính trong quá trình khai phá dữ liệu
 c. **$R_{A,B} = -1$ thì ta nên loại một trong hai thuộc tính trong quá trình khai phá dữ liệu**
 d. $R_{A,B}$ cao thể hiện sự phụ thuộc lẫn nhau giữa A và B cao

27. Phát biểu nào dưới đây SAI về điều kiện dừng của giải thuật xây dựng cây quyết định:

- a. Tất cả những thể hiện trong phân hoạch D (tại nút N đang xét) thuộc về cùng một lớp
 b. Không còn thuộc tính nào nữa mà các thể hiện có thể được phân hoạch thêm
 c. **Việc tiếp tục lựa chọn các thuộc tính phân tách không làm tăng độ lợi thông tin**
 d. Không còn thể hiện nào nữa trên nhánh đang xét, tức là phân hoạch D bị rỗng

28. Trong số các phương pháp phân lớp dữ liệu, phương pháp nào có tính chất học tăng cường (incremental learning):

- a. Cây quyết định
 b. Naïve Bayes
 c. **Mạng nơ-ron**
 d. k-nearest neighbor

29. Các độ đo về sự phân tán của dữ liệu Q1, Q2, Q3, IQR có tác dụng trong việc:

- a. Phát hiện các phần tử nhiễu, các phần tử biên
 b. Cung cấp cái nhìn tổng quan về phân bố dữ liệu
 c. Chuẩn hóa dữ liệu, lựa chọn thuộc tính
 d. Phân lớp dữ liệu (classification)
 e. **Cả hai câu a và b đều đúng**

30. Tri thức có thể đạt được từ quá trình khai phá dữ liệu là:

- a. Mô hình phân loại / dự đoán
 b. Mô hình gom cụm / các mối quan hệ, luật kết hợp
 c. Các phần tử biên, ngoại lai
 d. Xu hướng biến đổi dữ liệu / các mẫu thường xuyên

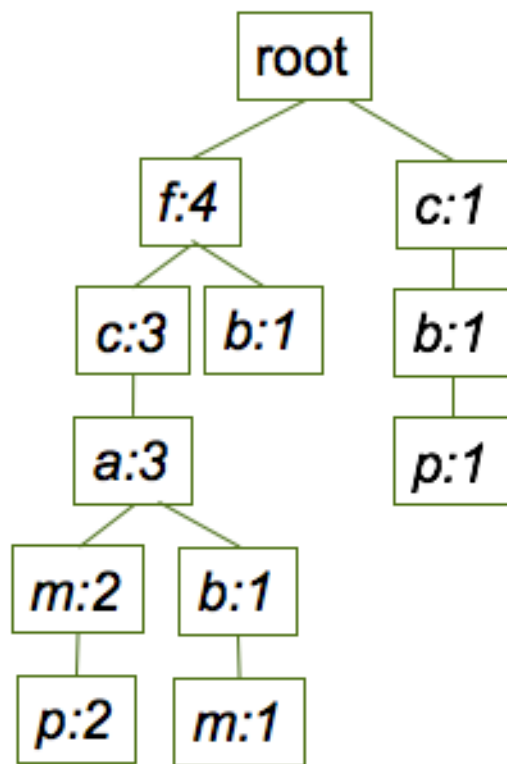
- e. Tất cả các câu trên đều đúng
31. Phép kiểm thống kê chi-square được dùng để:
- Tìm ra những điểm chia để rời rạc hóa dữ liệu
 - Tạo ra các mức ý niệm để thực hiện việc tổng quát hóa dữ liệu
 - Phân tích sự độc lập của các thuộc tính rời rạc
 - Phân tích tương quan của các thuộc tính liên tục
32. Giải pháp nào được dùng để thu giảm dữ liệu:
- Phân tích nhân tố chính (Principal component analysis)
 - Histogram, Data Sampling
 - Kết hợp khối dữ liệu (data cube aggregation)
 - Hai câu a và b đều đúng
 - Ba câu a, b và c đều đúng
33. Chọn phát biểu ĐÚNG:
- Hàm $Y = aX + b$ là hàm hồi qui phi tuyến (a, b là thông số)
 - Hàm $Y = aX_1 + bX_2 + cX_3 + d$ là hàm hồi qui phi tuyến (a, b, c, d là thông số)
 - Hàm $Y = a \cdot \log(bX)$ là hàm hồi qui phi tuyến (a, b là thông số)
 - Hàm $Y = aX^b$ là hàm hồi qui tuyến tính (a, b là thông số)
 - Cả 4 câu trên đều sai
34. Các điểm ngoại biên (outlier) có thể phát hiện được nhờ phương pháp nào sau đây:
- Dùng trị trung bình và độ lệch chuẩn
 - Dùng giá trị IQR (interquartile range), Q1 và Q3
 - Dùng phương pháp gom cụm
 - Cả ba phương pháp trên
35. Chọn phát biểu Đúng trong các câu sau:
- Giải thuật k-medoids giải quyết vấn đề nhiễu và điểm biên tốt hơn k-means
 - Cả 2 giải thuật gom cụm bằng phân hoạch (partition-based clustering) và gom cụm dựa vào cây phân cấp (hierarchical clustering) đều phải cho trước (input) số cụm
 - Gom cụm bằng phân hoạch thường làm việc tốt với các cụm có dạng hình cầu
 - Một điểm mạnh của gom cụm bằng phân hoạch so với gom cụm dựa vào cây phân cấp là nó có thể quay lại bước lặp trước đó

- e. Cả hai câu a và c đều đúng
36. Độ lợi thông tin (information gain) được dùng trong ngữ cảnh nào sau đây:
- Thu giảm số chiều
 - Chọn thuộc tính phân tách trong việc xây dựng bộ phân lớp dữ liệu
 - Thu giảm lượng số dữ liệu
 - Gộp khối dữ liệu
37. Trong giải thuật lan truyền ngược để huấn luyện mạng nơ-ron, mỗi lần lặp duyệt qua mọi phần tử trong tập huấn luyện được gọi bằng thuật ngữ tiếng Anh nào sau đây:
- pass
 - epoch
 - stage
 - iteration
38. Thành phần nào sau đây không là thành tố cơ bản để đặc tả tác vụ khai phá dữ liệu
- Dữ liệu cụ thể được khai phá
 - Tri thức nền
 - Các độ đo
 - Chuẩn áp dụng cho việc xây dựng ứng dụng khai phá dữ liệu.
39. Tri thức có thể đạt được từ quá trình khai phá dữ liệu là:
- Mô hình phân loại / dự đoán
 - Mô hình gom cụm / các mối quan hệ, luật kết hợp
 - Các phần tử biên, ngoại lai
 - Xu hướng biến đổi dữ liệu / các mẫu thường xuyên
 - Tất cả các câu trên đều đúng
40. Mạng nơ-ron nhân tạo (ANN) là một mô hình tính toán
- mô phỏng cơ chế hoạt động của bộ não người
 - số node đầu ra (output) có thể là một hoặc nhiều, phụ thuộc vào số lượng trạng thái của dữ liệu mà hệ thống cần khảo sát
 - thường được dùng trong việc phân lớp dữ liệu
 - tất cả các câu trên đều đúng

Câu 1 (1.0 điểm) Cho một bộ dữ liệu về giỏ mua hàng như sau:

TID	Giỏ hàng (items bought)
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f, h, j, o
4	b, c, k, s, p
5	a, f, c, e, l, p, m, n

Vẽ FP-tree từ bộ dữ liệu nêu trên, với min_sup = 3:



Câu 2 (1.0 điểm): Cho biết tuổi của các vận động viên tham gia môn cờ vua như sau: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

a) Hãy cho biết kết quả của các giá trị sau (0.5 điểm).

b) Cho biết các phân tử ngoại biên (outliers) dựa vào interquartile range (0.5 điểm).

Mean	
Median	
Mode	
Midrange	
Q1	
Q2	
Q3	

Câu 3 (1.0 điểm): Trong phân lớp dữ liệu dựa vào mạng Bayesian:

a) (0.5 điểm) Nêu ý nghĩa của $P(C_i|X)$ và biểu thức tổng quát tính $P(C_i|X)$

- Tập dữ liệu huấn luyện D với mô tả (nhãn) của các lớp $C_i, i=1..m$, quá trình phân loại một tuple/đối tượng $X = (x_1, x_2, \dots, x_n)$ với mạng Bayesian.....

.X được phân loại vào C_i nếu và chỉ nếu

. $P(C_i|X) > P(C_j|X)$ với $1 \leq j \leq m, j \neq i$

$$P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)}$$


b) (0.5 điểm) Nêu ý nghĩa của $P(X|C_i)$ và cách tính nó khi X chứa đồng thời thuộc tính rời rạc và liên tục

Giảng viên ra đề

Chủ nhiệm bộ môn

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

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

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

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

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

		Actual			
		A	B	C	SUM
Predicted	A	600	400	200	1200
	B	1000	1200	200	2400
	C	400	400	1600	2400
	SUM	2000	2000	2000	

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- ☒ (A) $C_1 : (4, 4), C_2 : (2, 2), C_3 : (7, 7).$
- ☐ (B) $C_1 : (6, 6), C_2 : (4, 4), C_3 : (9, 9).$
- ☐ (C) $C_1 : (2, 2), C_2 : (0, 0), C_3 : (5, 5).$
- ☐ (D) Tất cả đều sai.

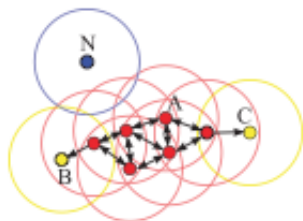
Câu hỏi 7 [L.O.3.3]. Khoảng cách Mahattan giữa điểm $(9, 9)$ đến tâm của cụm C_1 trong bước lặp tiếp theo là

- ☐ (A) 8.
- ☒ (C) 10.
- ☐ (B) 9.
- ☐ (D) 11.

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

- ☐ (A) Có nhiều.
- ☒ (B) Tất cả trường hợp này.
- ☐ (C) Có các mật độ phân bố khác nhau.
- ☐ (D) Có các cụm có hình dáng kiểu không lồi.

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



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

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

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

- ☐ (A) A.
- ☒ (B) N.
- ☐ (C) B.
- ☐ (D) C.

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

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

- ☐ (A) Có thể mô hình mới sẽ underfit.
- ☐ (B) Tất cả hiện tượng này đều xảy ra.
- ☒ (C) Có thể mô hình mới sẽ overfit.
- ☐ (D) Mô hình sẽ mới sẽ cho kết quả tốt hơn.

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

- ☒ (A) Có thể mô hình mới sẽ underfit.
- ☐ (B) Tất cả hiện tượng này đều xảy ra.
- ☐ (C) Có thể mô hình mới sẽ overfit.
- ☐ (D) Mô hình sẽ mới sẽ cho kết quả tốt hơn.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Câu hỏi 21 [L.O.3.4]. Danh sách có

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

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

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

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

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

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

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

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

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

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

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

Yêu cầu

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

Lời giải



Đáp án – Mã đề: 2010

Câu hỏi 1 (B)

Câu hỏi 2 (C)

Câu hỏi 3 (D)

Câu hỏi 4 (A)

Câu hỏi 5 (D)

Câu hỏi 6 (A)

Câu hỏi 7 (C)

Câu hỏi 8 (B)

Câu hỏi 9 (C)

Câu hỏi 10 (B)

Câu hỏi 11 (C)

Câu hỏi 12 (A)

Câu hỏi 13 (B)

Câu hỏi 14 (A)

Câu hỏi 15 (B)

Câu hỏi 16 (C)

Câu hỏi 17 (D)

Câu hỏi 18 (A)

Câu hỏi 19 (B)

Câu hỏi 20 (A)

Câu hỏi 21 (A)

Câu hỏi 22 (D)

Câu hỏi 23 (A)

Câu hỏi 24 (B)

Câu hỏi 25 (D)

Câu hỏi 26 Lời giải

[Trang chủ](#) / [Phòng thi của tôi](#) / [CO3029_1_DH_HK202](#) / [General](#) / [Khai phá dữ liệu - Thi cuối kỳ 2/2020-2021](#)

Đã bắt đầu vào lúc Thứ năm, 19 Tháng tám 2021, 9:55 AM

Tình trạng Đã hoàn thành

Hoàn thành vào lúc Thứ năm, 19 Tháng tám 2021, 10:45 AM

Thời gian thực hiện 50 phút

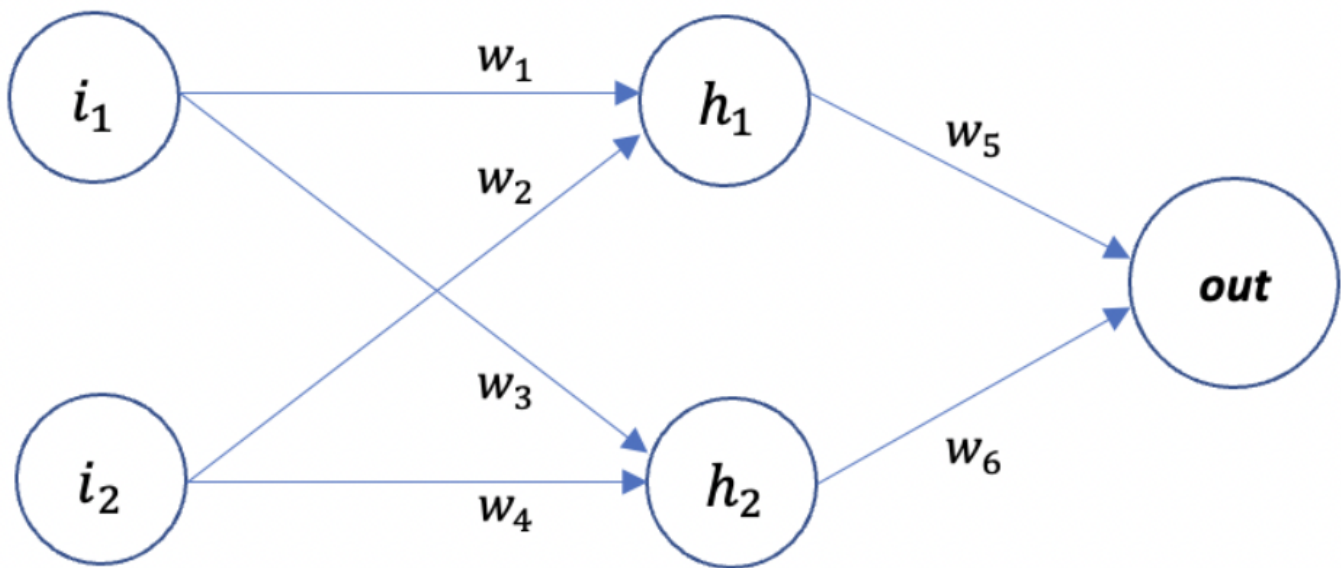
Câu hỏi 1

Hoàn thành

Chấm điểm của 0,40

Ta cần huấn luyện một mạng nơ-ron lan truyền ngược 3 tầng, được mô tả như dưới đây:

- Tầng vào gồm 2 nơ-ron, $i_1 = 2$ và $i_2 = 3$.
- Tầng ẩn: 2 nơ-ron, h_1 và h_2 .
- Tầng ra: 1 nơ-ron, $Out = 1$.
- Không sử dụng hàm kích hoạt tại tầng ẩn và tầng ra, tốc độ học (learning rate) được cho là 0.05.



Các trọng số được khởi tạo: $w_1 = 0.11$, $w_2 = 0.21$, $w_3 = 0.12$, $w_4 = 0.08$, $w_5 = 0.14$, $w_6 = 0.15$. Các câu hỏi 1 sau đây đến 6 xét lần huấn luyện đầu tiên. Trong giai đoạn truyền thẳng (forward pass), các giá trị tại các nơ-ron tầng ẩn, h_1 và h_2 lần lượt là

- ☐ a. 0.8 và 0.8.
- ☐ b. 0.84 và 0.85.
- ☒ c. 0.85 và 0.48.
- ☐ d. 0.58 và 0.84.

Câu trả lời của bạn là chính xác.

Câu hỏi **2**

Hoàn thành

Chấm điểm của 0,40

Giá trị của tầng ra là

- ☒ a. 0.919
- ☐ b. 0.191
- ☐ c. 1.091
- ☐ d. 0.119

Câu trả lời của bạn không chính xác.

Câu hỏi **3**

Hoàn thành

Chấm điểm của 0,40

Sử dụng hàm lỗi là một phần hai của bình phương khoảng cách Euclidean, giá trị lỗi là

- ☐ a. 0.327
- ☒ b. 0.237
- ☐ c. 0.723
- ☐ d. 0.372

Câu trả lời của bạn không chính xác.

Câu hỏi **4**

Hoàn thành

Chấm điểm của 0,40

Trong giai đoạn truyền ngược để cập nhật trong số, giá trị của w_5 và w_6 lần lượt là

- ☐ a. 0.71 và 0.71
- ☐ b. 0.17 và 0.71
- ☐ c. 0.71 và 0.17
- ☒ d. 0.17 và 0.17

Câu trả lời của bạn là chính xác.

Câu hỏi **5**

Hoàn thành

Chấm điểm của 0,40

Tiếp tục cập nhật, giá trị của w_1, w_2, w_3 và w_4 lần lượt là

- ☐ a. 0.12, 0.23, 0.1, 0.13
- ☒ b. 0.12, 0.13, 0.23, 0.1
- ☐ c. 0.12, 0.23, 0.13, 0.1
- ☐ d. 0.12, 0.13, 0.13, 0.1

Câu trả lời của bạn không chính xác.

Câu hỏi **6**

Hoàn thành

Chấm điểm của 0,40

Giá trị tại tầng ra trong lần huấn luyện tiếp theo sẽ là

- ☐ a. 0.62
- ☒ b. 0.26
- ☐ c. 0.22
- ☐ d. 0.66

Câu trả lời của bạn là chính xác.

Câu hỏi **7**

Hoàn thành

Chấm điểm của 0,40

Các câu hỏi 7 dưới đây đến 12 xét bài toán sau. Giả sử ta cần phân cụm 8 điểm dữ liệu $A_1 = (2, 10)$, $A_2 = (2, 5)$, $A_3 = (8, 4)$, $A_4 = (5, 8)$, $A_5 = (7, 5)$, $A_6 = (6, 4)$, $A_7 = (1, 2)$, $A_8 = (4, 9)$. thành 3 cụm C_1, C_2, C_3 sử dụng giải thuật k -means với khoảng cách Euclidean. Các tâm cụm được khởi tạo cần lượt là $C_1 = A_1$, $C_2 = A_4$ và $C_3 = A_7$.

Tại lần lặp đầu tiên, các cụm tìm được là

- ☒ a. $\{A_1\}, \{A_3, A_4, A_5, A_6, A_8\}, \{A_2, A_7\}$
- ☐ b. $\{A_1, A_8\}, \{A_3, A_4, A_5, A_6\}, \{A_2, A_7\}$
- ☐ c. $\{A_1, A_4, A_8\}, \{A_3, A_5, A_6\}, \{A_2, A_7\}$
- ☐ d. $\{A_1, A_3\}, \{A_4, A_5, A_6, A_8\}, \{A_2, A_7\}$

Câu trả lời của bạn là chính xác.

Câu hỏi 8

Hoàn thành

Chấm điểm của 0,40

Sau lần lặp đầu tiên, các tâm mới của cụm được cập nhật tương ứng là

- ☐ a. $C1 = (2,10), C2 = (6,6), C3 = (3.5,1.5)$
- ☐ b. $C1 = (10,2), C2 = (6,6), C3 = (1.5,3.5)$
- ☒ c. $C1 = (2,10), C2 = (6,6), C3 = (1.5,3.5)$
- ☐ d. $C1 = (6,6), C2 = (2,10), C3 = (1.5,3.5)$

Câu trả lời của bạn là chính xác.

Câu hỏi 9

Hoàn thành

Chấm điểm của 0,40

Tại lần lặp thứ 2, các cụm tìm được là

- ☐ a. $\{A1\}, \{A3, A4, A5, A6, A8\}, \{A2, A7\}$
- ☐ b. $\{A1, A4, A8\}, \{A3, A5, A6\}, \{A2, A7\}$
- ☒ c. $\{A1, A8\}, \{A3, A4, A5, A6\}, \{A2, A7\}$
- ☐ d. $\{A1.A3\}, \{A4, A5, A6, A8\}, \{A2, A7\}$

Câu trả lời của bạn là chính xác.

Câu hỏi **10**

Hoàn thành

Chấm điểm của 0,40

Sau lần lặp thứ 2, các tâm mới của cụm được cập nhật tương ứng là

- ☐ a. $C1 = (3, 9.5), C2 = (6.5, 5.25), C3 = (3.5, 1.5)$
- ☒ b. $C1 = (3, 9.5), C2 = (6.5, 5.25), C3 = (1.5, 3.5)$
- ☐ c. $C1 = (9.5, 3), C2 = (6.5, 5.25), C3 = (1.5, 3.5)$
- ☐ d. $C1 = (9.5, 3), C2 = (5.5, 5.25), C3 = (1.5, 3.5)$

Câu trả lời của bạn là chính xác.

Câu hỏi **11**

Hoàn thành

Chấm điểm của 0,40

Kết quả phân cụm cuối cùng là

- ☐ a. $\{A1\}, \{A3, A4, A5, A6, A8\}, \{A2, A7\}$
- ☐ b. $\{A4, A8\}, \{A3, A5, A6\}, \{A1, A2, A7\}$
- ☐ c. $\{A1, A8\}, \{A3, A4, A5, A6\}, \{A2, A7\}$
- ☒ d. $\{A1, A4, A8\}, \{A3, A5, A6\}, \{A2, A7\}$

Câu trả lời của bạn là chính xác.

Câu hỏi **12**

Hoàn thành

Chấm điểm của 0,40

Tâm của các cụm tìm được cuối cùng là

- ☒ a. $C1 = (3.66, 9)$, $C2 = (7, 4.33)$, $C3 = (1.5, 3.5)$
- ☐ b. $C1 = (3.66, 9)$, $C2 = (4.33, 7)$, $C3 = (1.5, 3.5)$
- ☐ c. $C1 = (9.5, 3)$, $C2 = (6.5, 5.25)$, $C3 = (1.5, 3.5)$
- ☐ d. $C1 = (9.5, 3)$, $C2 = (5.5, 5.25)$, $C3 = (1.5, 3.5)$

Câu trả lời của bạn là chính xác.

Câu hỏi **13**

Hoàn thành

Chấm điểm của 0,40

Các câu hỏi 13-17 xét danh sách giao dịch dưới đây:

- (1) trứng, thịt, khử khuẩn, khẩu trang
- (2) sữa, khẩu trang, trứng, mỳ gói, thịt
- (3) khẩu trang, sữa, trứng
- (4) mỳ gói, bia, sữa
- (5) khử khuẩn, mỳ gói, bia, sữa
- (6) thịt
- (7) sữa, bia, mỳ gói, khẩu trang, thịt
- (8) trứng, thịt, bia
- (9) sữa, mỳ gói
- (10) khẩu trang, sữa, mỳ gói

Danh sách có

- ☒ a. 10 giao dịch
- ☐ b. 9 giao dịch
- ☐ c. 8 giao dịch
- ☐ d. 7 giao dịch

Câu trả lời của bạn là chính xác.

Câu hỏi **14**

Hoàn thành

Chấm điểm của 0,40

Với $support = 0.4$, danh sách các mẫu (itemsets) xuất hiện thường xuyên là

- ☐ a. {trứng, mì gói}, {bia, khẩu trang}, {mì gói, khẩu trang}
- ☐ b. {trứng, mì gói}, {bia, thịt}, {sữa, mì gói, trứng}
- ☐ c. {mì gói, khẩu trang}, {khẩu trang, bia, thịt}, {sữa, mì gói, trứng}
- ☒ d. {trứng}, {thịt}, {khẩu trang}, {sữa}, {mì gói}, {bia}, {khẩu trang, sữa}, {sữa, mì gói}

Câu trả lời của bạn là chính xác.

Câu hỏi **15**

Hoàn thành

Chấm điểm của 0,40

Nếu giảm giá trị của $support$ xuống, thì

- ☐ a. số mẫu xuất hiện thường xuyên vẫn luôn giữ nguyên
- ☒ b. một số mẫu có thể được thêm vào tập xuất hiện thường xuyên hiện tại
- ☐ c. một số mẫu sẽ được đưa ra khỏi tập xuất hiện thường xuyên hiện tại
- ☐ d. các phương án đều đúng

Câu trả lời của bạn là chính xác.

Câu hỏi **16**

Hoàn thành

Chấm điểm của 0,40

Các luật kết hợp với $support = 0.4$ và $confidence = 0.7$ gồm

- ☐ a. $\{sữa\} \rightarrow \{bia\}$, $\{bia\} \rightarrow \{khẩu\ trang\}$
- ☐ b. $\{khẩu\ trang\} \rightarrow \{mỳ\ gói\}$, $\{mỳ\ gói\} \rightarrow \{khẩu\ trang\}$
- ☒ c. $\{khẩu\ trang\} \rightarrow \{sữa\}$, $\{sữa\} \rightarrow \{mỳ\ gói\}$, $\{mỳ\ gói\} \rightarrow \{sữa\}$
- ☐ d. $\{bia\} \rightarrow \{khẩu\ trang\}$, $\{khẩu\ trang\} \rightarrow \{bia\}$

Câu trả lời của bạn là chính xác.

Câu hỏi **17**

Hoàn thành

Chấm điểm của 0,40

Kết quả khai phá luật kết hợp thu được cho thấy

- ☐ a. sữa và bia thường sẽ được mua cùng nhau
- ☐ b. bia và khẩu trang thường sẽ được mua cùng nhau
- ☐ c. mỳ gói và khẩu trang thường sẽ được mua cùng nhau
- ☒ d. sữa và mỳ gói thường sẽ được mua cùng nhau

Câu trả lời của bạn là chính xác.

Câu hỏi **18**

Hoàn thành

Chấm điểm của 0,40

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

- ☐ a. AGNES
- ☐ b. DIANA
- ☐ c. DBSCAN
- ☒ d. Fuzzy c-means

Câu trả lời của bạn là chính xác.

Câu hỏi **19**

Hoàn thành

Chấm điểm của 0,40

Một mạng nơ-ron nhân tạo có n đầu vào x_1, x_2, \dots, x_n với các trọng số w_1, w_2, \dots, w_n . Giá trị tổng có trọng số sẽ được truyền tới hàm kích hoạt được tính là

- ☒ a. $x_1*w_1 + x_2*w_2 + \dots + x_n*w_n$
- ☐ b. $x_1 + x_2 + \dots + x_n$
- ☐ c. $w_1 + w_2 + \dots + w_n$
- ☐ d. $x_1 + w_1 + x_2 + w_2 + \dots + x_n + w_n$

Câu trả lời của bạn là chính xác.

Câu hỏi **20**

Hoàn thành

Chấm điểm của 0,40

Mạng nơ-ron nào sau đây dùng học có giám sát?

- ☐ a. Mạng Hopfield
- ☒ b. Mạng perceptron đa tầng
- ☐ c. Bản đồ đặc trưng tự tổ chức
- ☐ d. Tất cả các mạng này

Câu trả lời của bạn là chính xác.

Câu hỏi **21**

Hoàn thành

Chấm điểm của 0,40

Một itemset có giá trị hỗ trợ (support) lớn hơn hoặc bằng một ngưỡng cho trước gọi là

- ☐ a. xuất hiện không thường xuyên
- ☒ b. ngưỡng xuất hiện thường xuyên
- ☐ c. xuất hiện thường xuyên
- ☐ d. ngưỡng xuất hiện không thường xuyên

Câu trả lời của bạn không chính xác.

Câu hỏi **22**

Hoàn thành

Chấm điểm của 0,40

Kỹ thuật nào dưới đây giúp cải thiện giải thuật Apriori?

- ☐ a. Lấy mẫu
- ☐ b. Tăng số lượng giao dịch
- ☐ c. Giảm số lượng giao dịch
- ☒ d. Kỹ thuật băm

Câu trả lời của bạn là chính xác.

Câu hỏi **23**

Hoàn thành

Chấm điểm của 0,40

Độ tin cậy của $A \rightarrow B$, ký hiệu bởi $\text{confidence}(A \rightarrow B)$, được định nghĩa là

- ☐ a. $\text{support}(A \cap B) / \text{support}(A)$
- ☒ b. $\text{support}(A \cup B) / \text{support}(A)$
- ☐ c. $\text{support}(A \cap B) / \text{support}(B)$
- ☐ d. $\text{support}(A \cup B) / \text{support}(B)$

Câu trả lời của bạn là chính xác.

Câu hỏi **24**

Hoàn thành

Chấm điểm của 0,40

Đại lượng *lift* được định nghĩa bởi $lift = P(A \cup B) / (P(A) * P(B))$, được dùng để

- ☐ a. đánh giá luật kết hợp dạng $A \rightarrow B$
- ☒ b. đo sự tương quan giữa hai sự kiện A và B
- ☐ c. đánh giá luật kết hợp dạng $\langle A, B \rangle \rightarrow A$
- ☐ d. đánh giá luật kết hợp dạng $\langle A, B \rangle \rightarrow B$

Câu trả lời của bạn là chính xác.

Câu hỏi **25**

Hoàn thành

Chấm điểm của 0,40

Kỹ thuật nào dưới đây thích hợp nhất khi áp dụng để xác định một bài viết (trên mạng xã hội) được thích hay không?

- ☒ a. Phân lớp
- ☐ b. Phân cụm
- ☐ c. Hồi quy
- ☐ d. Khai phá luật kết hợp

Câu trả lời của bạn là chính xác.

[◀ For testing ...](#)

Chuyển tới...