Đại Học Quốc Gia Tp. HCM
Trường Đại Học Bách Khoa
Khoa Khoa Học và Kỹ Thuật Máy Tính

Mã đề thi: 01

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 => B) là số lần xuất hiện đồng thời của A và B trong tập dữ liệu *D*
   c. support(A => B) luôn lớn hơn confidence(A=> 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 mứt* và *chứa đồng thời bánh mì và mứt* lần lượt là 20000, 30000 và 10000.

8. Độ hỗ trợ (support) của phát biểu *"ai mua mứt 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 mứt 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 mứt và bánh mì nêu ở trên không đổi thì phát biểu *"ai mua mứt 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 (sequetial 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 các 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 kiệu 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. $h_\theta(X) \in [0, 1]$
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.

Thực tế \ Phân lớp thành	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 dổi dữ liệu / các mẫu thường xuyên

e. Tất cả các câu trên đều đúng

31. Phép kiểm thống kê chi-square được dùng để:
    a. Tìm ra những điểm chia để rời rạc hóa dữ liệu
    b. Tạo ra các mức ý niệm để thực hiện việc tổng quát hóa dữ liệu
    c. Phân tích sự độc lập của các thuộc tính rời rạc
    d. Phân tích tương quan của các thuộc tính liên tục
32. Giải pháp nào được dùng để thu giảm dữ liệu:
    a. Phân tích nhân tố chính (Principal component analysis)
    b. Histogram, Data Sampling
    c. Kết hợp khối dữ liệu (data cube aggregation)
    d. Hai câu a và b đều đúng
    e. Ba câu a, b và c đều đúng
33. Chọn phát biểu ĐÚNG:
    a. Hàm Y = aX+b là hàm hồi qui phi tuyến (a, b là thông số)
    b. Hàm Y = aX1 + bX2 + cX3 + d là hàm hồi qui phi tuyến (a, b, c, d là thông số)
    c. Hàm Y = a.log(bX) là hàm hồi qui phi tuyến (a, b là thông số)
    d. Hàm Y = aXb là hàm hồi qui tuyến tính (a, b là thông số)
    e. Cả 4 câu trên đều sai

34. Các điểm ngoại biên (outlier) có thể phát hiện được nhờ phương pháp nào sau đây:
    a. Dùng trị trung bình và độ lệch chuẩn
    b. Dùng giá trị IQR (interquartile range), Q1 và Q3
    c. Dùng phương pháp gom cụm
    d. Cả ba phương pháp trên
35. Chọn phát biểu Đúng trong các câu sau:
    a. Giải thuật k-medoids giải quyết vấn đề nhiễu và điểm biên tốt hơn k-means
    b. 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
    c. Gom cụm bằng phân hoạch thường làm việc tốt với các cụm có dạng hình cầu
    d. 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:
    a. Thu giảm số chiều
    b. Chọn thuộc tính phân tách trong việc xây dựng bộ phân lớp dữ liệu
    c. Thu giảm lượng số dữ liệu
    d. 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:
    a. pass
    b. epoch
    c. stage
    d. iteration
38. Thành phần nào sau đây không là thành tố cơ bản để đặc tả tác vụ khai phá dữ liệu
    a. Dữ liệu cụ thể được khai phá
    b. Tri thức nền
    c. Các độ đo
    d. 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:
    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 dổi dữ liệu / các mẫu thường xuyên
    e. Tất cả các câu trên đều đúng

40. Mạng nơ-ron nhân tạo (ANN) là một mô hình tính toán
    a. mô phỏng cơ chế hoạt động của bộ não người
    b. 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
    c. thường được dùng trong việc phân lớp dữ liệu
    d. tất cả các câu trên đều đúng

Phần 2: Tự luận (3.0 điểm). Sinh viên làm bài trực tiếp trên đề thi
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 $P(C_i|X)$

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 bô môn

TS. Trần Minh Quang **ThS. Trương Quang Hải** **TS. Trần Minh Quang**