

Câu 1: Khai thác dữ liệu có thể giúp cho doanh nghiệp?

- Thu thập dữ liệu hiệu quả hơn.
- Cải thiện hiệu năng hoạt động của cơ sở dữ liệu.
- Phát hiện và khắc phục điểm yếu của cơ sở dữ liệu.
- ✓ Cải thiện chiến lược kinh doanh.

Câu 2: Khai thác dữ liệu là bước phát triển tiếp theo của?

- Khoa học thực nghiệm
- ✓ Khoa học về dữ liệu
- Khoa học máy tính
- Công nghệ phần mềm

Câu 3: Tri thức khám phá được từ quy trình khai thác dữ liệu được sử dụng bởi ai?

- Chuyên viên phân tích số liệu
- Chuyên viên quản trị cơ sở dữ liệu
- ✓ Chuyên viên hoạch định chiến lược
- Chuyên viên quản trị mạng

Câu 4: Dữ liệu lịch sử (historical data) được hiểu là

- Dữ liệu về lịch sử hình thành của công ty.
- Dữ liệu không dùng đến nữa.
- Dữ liệu về hoạt động kinh doanh hiện tại.
- ✓ Dữ liệu hình thành trong toàn bộ quá trình kinh doanh.

Câu 5: Vấn đề nào sau đây KHÔNG phải là khó khăn (thách thức) trong khai phá dữ liệu?

- Có quá nhiều dữ liệu
- Phải tích hợp dữ liệu từ nhiều nguồn
- Dữ liệu có quá nhiều định dạng
- ✓ Có quá nhiều thuật toán

Câu 6: Phân tích giỏ hàng là bài toán

- ✓ Tìm luật kết hợp
- Phân lớp
- Phân cụm
- Hồi quy

Câu 7: Hệ thống nào sau đây sử dụng khai thác dữ liệu ?

- Hệ thống thông tin quản lý
- ✓ Hệ thống hỗ trợ quyết định
- Hệ thống thông tin địa lý

- Hệ thống thông tin phân tán

Câu 8: Lọc thư rác là ứng dụng của bài toán

- Phân cụm
- ✓ Phân lớp
- Hồi quy
- Phân tích luật kết hợp

Câu 9: Một siêu thị muốn sắp xếp vị trí của các món hàng trên kệ hàng, sao cho có thể tối đa hóa lợi nhuận. Vậy, siêu thị nên thực hiện bài toán gì trên lịch sử mua hàng của khách hàng?

- Phân cụm
- Phân lớp
- Hồi quy
- ✓ Phân tích luật kết hợp

Câu 10: Giai đoạn nào được thực hiện đầu tiên trong quy trình phát hiện tri thức?

- Khai thác dữ liệu
- Đánh giá tri thức
- ✓ Làm sạch dữ liệu
- Tích hợp dữ liệu

Câu 11: Hãy chọn phát biểu ĐÚNG trong các phát biểu sau:

- Khai thác dữ liệu được sử dụng trong các công ty tư nhân không gây ra tác động xã hội.
- Chỉ khai thác dữ liệu được sử dụng bởi các tổ chức nhà nước mới gây ra tác động xã hội.
- ✓ Khai thác dữ liệu cho dù được sử dụng bởi tổ chức nào cũng gây nên tác động xã hội.
- Khai thác dữ liệu cho dù được sử dụng bởi tổ chức nào cũng không gây nên tác động xã hội.

Câu 12: Tích hợp dữ liệu nhằm

- Gom những mẫu dữ liệu giống nhau vào một nhóm.
- Gom những mẫu dữ liệu cùng nằm vào một khối.
- ✓ Tập hợp dữ liệu từ nhiều nguồn thành một khối.
- Tập hợp dữ liệu cũ và dữ liệu mới thành một khối.

Câu 13: Thao tác nhóm các mẫu dữ liệu gần giống nhau vào những nhóm chưa xác định trước. Chúng ta sử dụng kỹ thuật

- Học có giám sát
- Học có ràng buộc
- ✓ Học không giám sát
- Học không ràng buộc

Câu 14: Một trong những kỹ thuật để phát hiện ngoại lệ (outlier detection) là sử dụng bài toán

- Phân cụm
- Phân lớp
- Phân tích chuỗi
- ✓ Khai thác luật kết hợp

Câu 15: Một công ty viễn thông muốn nhóm những khách hàng của họ vào những nhóm riêng biệt để triển khai chương trình khuyến mãi phù hợp cho từng nhóm. Đây là ví dụ của

- ✓ Học không giám sát
- Học có giám sát
- Phân tích tương quan
- Phân tích luật kết hợp

Câu 16: Cho bảng dữ liệu sau về kế hoạch vào đại học của các em học sinh Học sinh Điểm thi ĐH Gia đình khuyến khích Học bổng (ngàn đồng/tháng) Kế hoạch vào ĐH An 21 Có 800 Có Hòa 23 Không 1000 Có Thái 18 Có 0 Có Bình 28 Không ? Không Giả sử bạn là một mệnh thường quân muốn xác định mức học bổng hàng tháng để giúp các học sinh giỏi có hoàn cảnh khó khăn được học đại học. Bạn sử dụng bài toán gì để xác định số tiền học bổng cho em Bình để em thay đổi kế hoạch vào ĐH từ “không” thành “có”?

- Phân lớp
- Khai thác luật kết hợp
- Phân cụm
- ✓ Hồi quy

Câu 17: Hệ thống khuyến nghị (recommender systems) trong tiếp thị chéo (cross-marketing) thường sử dụng kết quả của

- ✓ Luật kết hợp
- Phân cụm
- Hồi quy
- Phân lớp

Câu 18: Khi khai thác dữ liệu được sử dụng phổ biến, phát biểu nào sau đây là KHÔNG ĐÚNG

- Hành vi người dùng trên internet có thể được khai thác mà người dùng không biết.
- ✓ Người dùng có thể kiểm soát được thông tin nào của mình có thể được khai thác.
- Người dùng có được lợi ích từ việc dễ dàng ra quyết định hơn.
- Người dùng được lợi ích do các công ty chăm sóc khách hàng tốt hơn.

Câu 19: Để tư vấn ngành học cho thí sinh A muốn xét tuyển vào đại học, dựa vào thông tin cá nhân của thí sinh A và một lượng lớn sinh viên đang theo học hoặc đã ra trường,

ta cần làm bài toán:

- ✓ Phân lớp
- Phân cụm
- Hồi quy
- Phân tích chuỗi thời gian

Câu 20: Câu hỏi “Khách hàng có thông tin là X có tiềm năng hay không?” có thể trả lời bằng:

- Phân tích luật kết hợp
- Hồi quy
- ✓ Phân lớp
- Phân cụm

Câu 21: Để xác định một khách hàng tiềm năng ta sử dụng bài toán

- ✓ Phân lớp
- Phân cụm
- Khai thác tập phổ biến
- Hồi quy

Câu 22: Câu hỏi “Một khách hàng là giáo viên thường mua gì sau khi mua máy tính?” được trả lời bởi bài toán nào?

- Khai thác luật kết hợp đơn mức
- Khai thác luật kết hợp đa mức
- Khai thác luật kết hợp đơn chiều
- ✓ Khai thác luật kết hợp đa chiều

Câu 23: Thói quen mua hàng của những nhóm khách hàng khác nhau trong siêu thị có thể được phát hiện nhờ phương pháp

- Phân lớp
- ✓ Phân cụm
- Hồi quy
- Suy diễn

Câu 24: Cho một cơ sở dữ liệu về lý lịch, điểm số về quá trình học tập của sinh viên. Bài toán phân lớp có thể:

- ✓ Dự đoán xếp loại tốt nghiệp của một sinh viên.
- Dự đoán điểm trung bình tốt nghiệp của một sinh viên.
- Xác định số tiền học bổng tối ưu cho một sinh viên.
- Phát hiện sự liên quan giữa thu nhập gia đình và xếp loại tốt nghiệp.

Câu 25: Với một công ty bán lẻ có khoảng một triệu khách hàng, công ty muốn thực hiện chương trình khuyến mãi sao cho phù hợp với nhu cầu của khách hàng, bước đầu tiên công ty nên làm là

- ✓ Phân cụm khách hàng dựa vào thông tin cá nhân và hành vi mua hàng.
- Phân lớp khách hàng dựa vào thông tin cá nhân và hành vi mua hàng.
- Khai thác tập phổ biến trên lịch sử mua hàng của khách hàng.
- Dự đoán giá trị của khách hàng trong tương lai.

Câu 26: Thao tác nào sau đây không thuộc giai đoạn tiền xử lý dữ liệu?

- Làm sạch dữ liệu
- ✓ Trục quan hóa dữ liệu
- Rút gọn dữ liệu
- Chuẩn hóa dữ liệu

Câu 27: Kích cỡ quần áo (size = {XS, S, M, L, XL, XXL,...}) là kiểu dữ liệu

- Định danh
- ✓ Thứ bậc
- Chia quãng
- Chia tỉ lệ

Câu 28: Vẽ biểu đồ cho dữ liệu ta gọi là

- ✓ Trục quan hóa dữ liệu
- Tiền xử lý dữ liệu
- Chuyển dạng dữ liệu
- Chuẩn hóa dữ liệu

Câu 29: _____ không phải là một yêu cầu của chất lượng dữ liệu?

- Tính đúng đắn của dữ liệu
- Tính hợp thời của dữ liệu
- Độ sạch sẽ của dữ liệu
- ✓ Dung lượng của dữ liệu

Câu 30: Cho một tập dữ liệu về độ tuổi bao gồm 10 phần tử như sau: Age = {15, 16, 18, 22, 24, 24, 24, 25, 26, 30} Trung vị của tập dữ liệu Age là?

- ✓ 24
- 25
- 20
- 22.5

Câu 31: Cho một tập dữ liệu về độ tuổi bao gồm 10 phần tử như sau: Age = {15, 16, 18, 22, 24, 24, 24, 25, 26, 30} Độ lệch chuẩn của tập dữ liệu Age là?

- ✓ 4.71
- 16.52
- 6

• 4.23

Câu 32: Histogram của phân bố chuẩn có dạng

- Đường parabol
- Đường thẳng
- ✓ Hình chuông
- Hình tròn

Câu 33: Việc thay thế số tuổi bằng giá trị đại diện như thiếu nhi, thiếu niên, thanh niên, trung niên, cao niên... gọi là

- Chuẩn hóa dữ liệu
- ✓ Rời rạc hóa dữ liệu
- Chuyển dạng dữ liệu
- Rút gọn dữ liệu

Câu 34: Dữ liệu rời rạc...

- chỉ chấp nhận giá trị kiểu số nguyên
- chỉ chấp nhận giá trị kiểu số thực
- ✓ chỉ chấp nhận một số giá trị hữu hạn
- chỉ chấp nhận giá trị kiểu chuỗi

Câu 35: Hàm cosine để đo độ tương đồng giữa

- Hai mẫu dữ liệu rời rạc
- Hai mẫu dữ liệu kiểu thứ bậc
- ✓ Hai vector kiểu số
- Hai ma trận kiểu số

Câu 36: [<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image003.jpg>] được gọi là công thức tính khoảng cách Euclidean khi h bằng

- 1
- ✓ 2
- 3
- ∞

Câu 37: [<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image004.png>] là công thức tính

- ✓ Độ tương đồng cosine
- Khoảng cách Euclidean
- Khoảng cách Minkowski
- Khoảng cách Manhattan

Câu 38: Hãy cho biết các chỉ số thể hiện trong biểu đồ boxplot là gì?

- ✓ Min, Max, Q1, Q3, Median
- Min, Max, Q1, Q3, Mean
- Q1, Q2, Q3, Q4, Q5, Mean
- Q1, Q2, Q3, Mean, Median

Câu 39: OLAP nghĩa là

- ✓ Phân tích dữ liệu trực tuyến
- Phân tích dữ liệu ngoại tuyến
- Phân tích dữ liệu trực quan
- Phân tích dữ liệu trực tiếp

Câu 40: Đây là chức năng chính của OLAP?

- Hỗ trợ xử lý giao dịch nghiệp vụ thường xuyên
- ✓ Hỗ trợ quyết định
- Lưu trữ dữ liệu hoạt động thường xuyên
- Lưu trữ dữ liệu lịch sử

Câu 41: _____ là quá trình lấy dữ liệu từ nhiều nguồn và ánh xạ từng trường vào một cấu trúc trong nhà kho dữ liệu (data warehouse).

- Chuyển dạng dữ liệu
- ✓ Tích hợp dữ liệu
- Làm sạch dữ liệu
- Chuẩn hóa dữ liệu

Câu 42: Kiểu Data Mart có thể lấy dữ liệu từ kho dữ liệu hoặc hệ thống vận hành.

- ✓ Kiểu Data Mart phụ thuộc
- Kiểu Data Mart độc lập
- Kiểu Data Mart kết hợp
- Kiểu Data Mart riêng tư

Câu 43: Việc tạo cơ sở dữ liệu vật lý và các cấu trúc logic nằm trong giai đoạn nào

- Thiết kế
- ✓ Xây dựng
- Truy cập
- Quản lý

Câu 44: _____ là một tập hợp dữ liệu hướng chủ đề, toàn vẹn, không bị rò rỉ mất mát và có giá trị lịch sử phục vụ cho công tác quản lý, cung cấp thông tin một cách kịp thời, chính xác, đồng thời là nền tảng cho việc xây dựng các ứng dụng phân tích dữ liệu, hỗ trợ ra quyết định của tổ chức.

- Data structure - Cấu trúc dữ liệu
- ✓ Data warehouse - Kho dữ liệu

- Data base - Cơ sở dữ liệu
- Data analytics - Phân tích dữ liệu

Câu 45: OLTP là tên viết tắt của kĩ thuật nào?

- ✓ Online Transaction Processing
- Online Transaction Program
- Oracle Transaction Processing
- Online Terminal Processing

Câu 46: Trong khai phá dữ liệu, thao tác cắt khối (dice) giúp người sử dụng phân tích dữ liệu trên mấy chiều?

- Một chiều.
- Đúng hai chiều.
- ✓ Hai chiều trở lên.
- Không có đáp án đúng.

Câu 47: _____ dự đoán xu hướng và hành vi trong tương lai, hỗ trợ các nhà quản lý đưa ra quyết định tác động đến hoạt động kinh doanh của doanh nghiệp.

- Data warehouse - Kho dữ liệu
- ✓ Data mining - Khai phá dữ liệu
- Datamart - Kho dữ liệu cục bộ
- Metadata – Siêu dữ liệu

Câu 48: Đối với thuật toán Apriori, để tìm tất cả các luật kết hợp, ta cần phải làm gì trước đó?

- ✓ Tìm tất cả các tập phổ biến
- Tìm tất cả các tập không phổ biến
- Tính độ hỗ trợ của các mục đơn
- Tính độ tin cậy của các mục đơn

Câu 49: Duyệt cơ sở dữ liệu nhiều lần là yếu điểm của thuật toán

- FP-Growth
- ✓ Apriori
- TID-Apriori
- CLOSET

Câu 50: Bài toán khai thác tập phổ biến và luật kết hợp KHÔNG được áp dụng cho câu hỏi nào sau đây?

- Những món hàng nào thường hay được mua chung với nhau?
- Những loại DNA nào thì nhạy cảm với loại thuốc mới?
- ✓ Giá cổ phiếu X lên hay giảm trong 5 tháng tới?
- Món gì thường hay được mua sau khi mua laptop một tuần?

Câu 51: Ứng dụng nào sau đây KHÔNG áp dụng bài toán khai thác tập phổ biến?

- Phân tích giỏ hàng
- Phân tích lịch sử duyệt web (web logs)
- Tiếp thị chéo (cross marketing)
- ✓ Xác định hạn mức tín dụng

Câu 52: Thuật toán FP-Growth cần quét cơ sở dữ liệu bao nhiêu lần?

- 1 lần
- ✓ 2 lần
- 3 lần
- Không xác định

Câu 53: Việc sắp xếp vị trí các món hàng trong một catalogue có thể được cải tiến nhờ vào bài toán ?

- ✓ Khai thác luật kết hợp
- Phân lớp
- Phân cụm
- Phân tích dữ liệu chuỗi thời gian

Câu 54: Theo cơ sở dữ liệu giao tác trong bảng sau, độ tin cậy của luật kết hợp AàB là:
TID Itemset 100 A B C D 200 A C D 300 A B D 400 B C D E 500 A B C E

- ✓ 75%
- 60%
- 65%
- 80%

Câu 55: Theo cơ sở dữ liệu giao tác trong bảng sau, độ hỗ trợ và độ tin cậy của luật kết hợp AC àD lần lượt là: TID Itemset 100 A B C D 200 A C D 300 A B D 400 B C D E 500 A B C E

- ✓ 40%; 66.7%
- 66.7%; 40%
- 40%; 75%
- 75%; 40%

Câu 56: Chọn công thức đúng để tính độ tin cậy của luật

[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image005.png>]

- ✓ [<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image006.png>]
- [<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image007.png>]
- [<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image008.png>]
- [<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image009.png>]

Câu 57: Hãy cho biết trong các đặc điểm sau, đặc điểm nào KHÔNG phải của thuật toán Apriori?

- ✓ Cần xây dựng cây.
- Duyệt cơ sở dữ liệu nhiều lần.
- Sinh ra nhiều ứng viên.
- Sử dụng tính chất downward closure.

Câu 58: Độ hỗ trợ (support) của một luật kết hợp là

- Mức độ đáng tin của luật
- ✓ Mức độ phổ biến của luật
- Mức độ đúng đắn của luật
- Mức độ hỗ trợ người dùng của luật

Câu 59: Cho s là độ hỗ trợ, c là độ tin cậy của một luật kết hợp. Hãy chọn phát biểu ĐÚNG

- $s(A \Rightarrow B) = s(B \Rightarrow A)$ và $c(A \Rightarrow B) = c(B \Rightarrow A)$
- $s(A \Rightarrow B) \neq s(B \Rightarrow A)$ và $c(A \Rightarrow B) = c(B \Rightarrow A)$
- $s(A \Rightarrow B) \neq s(B \Rightarrow A)$ và $c(A \Rightarrow B) \neq c(B \Rightarrow A)$
- ✓ $s(A \Rightarrow B) = s(B \Rightarrow A)$ và $c(A \Rightarrow B) \neq c(B \Rightarrow A)$

Câu 60: Hãy chọn phát biểu ĐÚNG trong các phát biểu sau đây?

- ✓ Độ hỗ trợ tối thiểu (min_support) càng cao thì số tập phổ biến được phát hiện càng ít.
- Độ tin cậy tối thiểu (min_confidence) càng thấp thì số luật được phát hiện càng ít.
- Độ hỗ trợ tối thiểu (min_support) càng cao thì số tập phổ biến được phát hiện càng nhiều.
- Độ tin cậy tối thiểu (min_confidence) càng cao thì số luật được phát hiện càng nhiều.

Câu 61: Câu nào sau đây KHÔNG đúng về Apriori khi so sánh Apriori và FP-Growth?

- Apriori dễ cài đặt hơn FP-Growth.
- Apriori quét cơ sở dữ liệu nhiều lần.
- ✓ Apriori gặp khó khăn khi duyệt cây đệ quy.
- Apriori gặp khó khăn khi đếm độ hỗ trợ của các tập mục.

Câu 62: Một trong những cách để hạn chế yếu điểm của thuật toán Apriori là?

- ✓ Giảm số lần quét cơ sở dữ liệu
- Khử đệ quy trong thuật toán
- Tăng bộ nhớ hệ thống
- Tăng số lượng tập ứng viên

Câu 63: Bước đầu tiên của thuật toán FP-Growth là?

- Đưa các giao tác vào cây FP.
- ✓ Sắp xếp các mục trong giao tác theo thứ tự giảm dần của độ hỗ trợ.

- Xây dựng cơ sở mẫu điều kiện của các mục.
- Xây dựng cây FP điều kiện của các mục.

Câu 64: Trong hình sau, hãy lựa chọn tập cơ sở mẫu điều kiện đúng của mục
[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image011.png>]

- ✓ {fca:1}, {f:1}, {c:1}
- {fca:3}, {f:4}, {c:1}
- {fca:1}, {f:2}, {c:1}
- {fca:3}, {f:1}, {c:1}

Câu 65: Một luật có dạng: age(X, "19-25") $\cup$ occupation(X, "student") $\rightarrow$ buys(X, "coke") được gọi là luật gì?

- ✓ Luật kết hợp đa chiều
- Luật kết hợp đa mức
- Luật kết hợp hỗn hợp
- Luật kết hợp đa trị

Câu 66: Để rút trích số lượng tập phổ biến mà không mất mát thông tin, thay vì tìm tất cả các tập phổ biến, ta đi tìm

- ✓ Tất cả các tập phổ biến đóng (closed patterns).
- Tất cả các tập phổ biến cực đại (maximal patterns).
- Tất cả các tập ứng viên (candidate sets).
- Tất cả các tập phổ biến có độ dài cực đại.

Câu 67: Cho X là một tập mục, X là tập phổ biến đóng (closed-pattern) khi và chỉ khi X là phổ biến và

- ✓ Không tồn tại một tập mục Y sao cho
[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image012.png>] mà Y có cùng độ hỗ trợ với X.
- Không tồn tại một tập mục Y sao cho
[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image013.png>] , mà Y có cùng độ hỗ trợ với X.
- Không tồn tại một tập mục Y sao cho
[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image013.png>] , mà Y là phổ biến.
- Không tồn tại một tập mục Y sao cho
[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image014.png>] mà Y là phổ biến.

Câu 68: Thực thi thuật toán Apriori, nếu kết quả cuối cùng ta tìm được các tập phổ biến có độ dài bằng n, ta phải duyệt CSDL tối đa bao nhiêu lần?

- n
- ✓ n+1
- n-1

- $n+2$

Câu 69: Áp dụng Apriori, với $\text{minsup}=3$, nếu lần lặp thứ hai ta đếm độ hỗ trợ của các tập mục được $\{AB:3\}$, $\{AC:2\}$, $\{BC:3\}$, $\{BD:4\}$. Vậy tập ứng viên ở lần lặp thứ 3 là?

- $C_3 = \{ABC, ABD, ACD, BCD\}$
- $C_3 = \{ABC, ABD, BCD\}$
- ✓ $C_3 = \{ABD, BCD\}$
- $C_3 = \{BCD\}$

Câu 70: Các tập phổ biến chứa m khi duyệt cây FP ở hình sau là
[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image011.png>]

- ✓ $\{m:3\}$, $\{fm:3\}$, $\{cm:3\}$, $\{fcm:3\}$, $\{fam:3\}$, $\{fcam:3\}$
- $\{m:3\}$, $\{cm:3\}$, $\{fcm:3\}$, $\{fam:3\}$, $\{fcam:3\}$
- $\{m:3\}$, $\{fm:3\}$, $\{cm:3\}$, $\{fcm:3\}$, $\{fam:3\}$
- $\{m:3\}$, $\{fm:3\}$, $\{cm:3\}$

Câu 71: Từ nào sau đây KHÔNG liên quan đến thuật toán cây quyết định?

- Độ quy
- Chia để trị
- Tham lam
- ✓ Quy hoạch động

Câu 72: Câu nói “Hãy cho tôi biết bạn chơi với ai, tôi sẽ nói cho bạn biết bạn như thế nào” có thể dùng khi nói về thuật toán nào?

- SVM
- C4.5
- ID3
- ✓ KNN

Câu 73: Thuật toán nào sau đây tốn thời gian tính khoảng cách giữa các đối tượng dữ liệu?

- ID3
- CART
- ✓ KNN
- Naïve Bayes

Câu 74: Naïve Bayes là thuật toán phân lớp dựa trên gì?

- ✓ Thống kê
- Láng giếng
- Phân bố
- Luật

Câu 75: SVM là viết tắt của

- ✓ Support Vector Machine
- Space Vector Model
- Support Vector Model
- Space Vector Machine

Câu 76: Trong bài toán dự đoán độ tin cậy của khách hàng vay tiền ngân hàng. Trường hợp khách hàng là đáng tin cậy nhưng kết quả dự đoán là không tin cậy. Vậy kết quả dự đoán được gọi là

- True Positive
- False Positive
- True Negative
- ✓ False Negative

Câu 77: Dựa vào bảng sau, giá trị của xác suất $P(\text{Nghề nghiệp} = \text{"CNV"} | \text{An toàn} = \text{"Có"})$ là: Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có

- ✓ 2/5
- 2/3
- 3/5
- 1

Câu 78: Thuật toán phân lớp dựa vào láng giềng gần nhất thích hợp nhất để xử lý dữ liệu kiểu gì?

- ✓ Số liên tục
- Nhị phân
- Thứ bậc
- Định danh

Câu 79: Thuật toán nào sau đây sử dụng Entropy làm độ đo thông tin?

- ✓ ID3
- KNN
- Naïve Bayes
- SVM

Câu 80: Trong các độ đo để lựa chọn thuộc tính rẽ nhánh khi dựng cây quyết định, độ đo nào dễ bị nghiêng về thuộc tính có nhiều giá trị?

- ✓ Information Gain
- Split Info
- Gain Ratio
- Entropy

Câu 81: Nếu thuật toán dựng cây quyết định bị overfitting thì điều gì sau đây không xảy ra?

- Cây sẽ có nhiều nhánh
- Hiệu năng phân lớp sẽ kém
- Bị ảnh hưởng bởi nhiễu
- ✓ Thuật toán dựng cây bị rơi vào vòng lặp vô hạn

Câu 82: Thuật toán phân lớp nào sau đây chỉ thực hiện trên dữ liệu kiểu số

- ✓ SVM
- Naïve Bayes
- ID3
- C4.5

Câu 83: Độ bao phủ (recall) của phép tiên đoán trong bảng sau là Actual\Prediction
Cancer Not Cancer Total Cancer 1300 1200 2500 Not Cancer 2700 4800 7500 Total
4000 6000 10000

- 32.5%
- 13%
- ✓ 52%
- 36%

Câu 84: Độ tin cậy (precision) của phép tiên đoán trong bảng sau là Actual\Prediction
Cancer Not Cancer Total Cancer 1300 1200 2500 Not Cancer 2700 4800 7500 Total
4000 6000 10000

- ✓ 32.5%
- 13%
- 52%
- 36%

Câu 85: Thuật toán Naïve Bayes có thể cho độ chính xác không cao do điều gì?

- ✓ Cần giả định độc lập về mặt thống kê của các thuộc tính.
- Kết quả dựa trên quá nhiều tính toán xác suất.
- Thuật toán dễ bị ảnh hưởng bởi nhiễu.
- Thuật toán dễ bị overfitting.

Câu 86: Công thức sau được sử dụng trong thuật toán nào?

[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image020.png>]

- ✓ ID3
- KNN
- Naïve Bayes
- SVM

- Boosting
- ✓ Bagging
- Rừng ngẫu nhiên
- Bootstrap

- Precision
- ✓ Recall
- Accuracy
- Sensitivity

- Khi Precision tăng thì Recall tăng
- Khi Precision giảm thì Recall giảm
- ✓ Khi Recall tăng thì Precision giảm
- Precision và Recall tăng giảm độc lập không phụ thuộc nhau

- $3/5$
- ✓ $2/5$
- $2/4$
- $4/5$

- Info(D)
- Info Thu nhập (D)
- Info Tuổi (D)
- ✓ Info An toàn (D)

Câu 92: Dựa vào bảng sau, theo thuật toán Naïve Bayes, khi cần xét độ an toàn cho khách hàng {Thu nhập = “Cao”, Sở hữu nhà = “Không”, Tuổi = “>45”, Nghề nghiệp =

"CNV"}, ta KHÔNG cần tính xác suất nào? Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có

✓ $P(\text{An toàn} = \text{"Có"} | \text{Nghề nghiệp} = \text{"CNV"})$

- $P(\text{Tuổi} = \text{">45"} | \text{An toàn} = \text{"Có"})$
- $P(\text{An toàn} = \text{"Không"})$
- $P(\text{An toàn} = \text{"Có"})$

Câu 93: Cho các xác suất sau: $P(\text{An toàn} = \text{"Có"} | \text{Nghề nghiệp} = \text{"CNV"})$ 1) $P(\text{Tuổi} = \text{"40-45"} | \text{An toàn} = \text{"Có"})$ 2) $P(\text{Tuổi} = \text{"40-45"} | \text{An toàn} = \text{"Không"})$ 3) $P(\text{An toàn} = \text{"Không"} | \text{Nghề nghiệp} = \text{"CNV"})$ 4) $P(\text{An toàn} = \text{"Có"} | \text{Nghề nghiệp} = \text{"CNV"})$ 5) $P(\text{Thu nhập} = \text{"Cao"} | \text{An toàn} = \text{"Có"})$ 6) $P(\text{Thu nhập} = \text{"Cao"} | \text{An toàn} = \text{"Không"})$ 7) $P(\text{An toàn} = \text{"Có"} | \text{Thu nhập} = \text{"Cao"})$ 8) $P(\text{An toàn} = \text{"Không"} | \text{Thu nhập} = \text{"Cao"})$ 9) $P(\text{An toàn} = \text{"Có"})$ 10) $P(\text{An toàn} = \text{"Không"})$ Dựa vào bảng sau, theo thuật toán Naïve Bayes, hãy chọn các xác suất được sử dụng khi cần xét độ an toàn cho khách hàng {Thu nhập = "Cao", Tuổi = ">45", Nghề nghiệp = "CNV"}. Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có

✓ 1, 2, 5, 6, 9, 10

- 1, 2, 3, 4, 7, 8, 9, 10
- 1, 2, 3, 4, 5, 6, 7, 8
- 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Câu 94: Đường mũi tên biểu thị khoảng cách gì giữa 2 cụm

[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image022.jpg>]

✓ Single - link

- Complete - link
- Average - link
- Full - link

Câu 95: Chọn phát biểu ĐÚNG theo thuật toán DBSCAN

[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image023.png>]

✓ Điểm p là đến được trực tiếp theo mật độ từ q.

- Điểm p là đến được theo mật độ từ q.
- Điểm p là liên thông theo mật độ từ q.
- Điểm p là đến được từ q.

Câu 96: Thuật toán nào thích hợp để phân cụm tập dữ liệu như hình bên dưới?

[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image024.png>]

- K-means
- ✓ DBSCAN
- AGNES

- DIANA

Câu 97: Lựa chọn nào SAI khi nói về yếu điểm của thuật toán K-means?

- Không gom được những cụm có hình dạng không lồi
- Dễ bị ảnh hưởng bởi nhiễu
- ✓ Không chắc tìm được đủ k cụm
- Tốn thời gian tính lại khoảng cách giữa các điểm dữ liệu

Câu 98: Tiêu chí của phân cụm là

- Cực đại hóa độ tương đồng nội cụm và cực đại hóa độ tương đồng liên cụm.
- Cực tiểu hóa độ tương đồng nội cụm và cực tiểu hóa độ bất tương đồng liên cụm.
- Cực tiểu hóa độ tương đồng nội cụm và cực đại hóa độ bất tương đồng liên cụm.
- ✓ Cực đại hóa độ tương đồng nội cụm và cực tiểu hóa độ tương đồng liên cụm.

Câu 99: Kết quả của thao tác phân cụm KHÔNG phụ thuộc vào

- Đặc trưng được lựa chọn để phân cụm.
- Hàm đo khoảng cách giữa các điểm dữ liệu.
- Thuật toán phân cụm.
- ✓ Thời gian thực hiện thuật toán phân cụm.

Câu 100: Hình thành cụm bằng cách lan truyền theo mật độ là đặc điểm của thuật toán nào?

- AGNES
- DIANA
- ✓ DBSCAN
- K-means

Câu 101: Chọn phát biểu ĐÚNG khi so sánh thuật toán PageRank và thuật toán HITS

- ✓ Thứ hạng trang web trong thuật toán PageRank được tính KHÔNG phụ thuộc vào câu truy vấn, còn thứ hạng trang web trong thuật toán HITS thì phụ thuộc vào câu truy vấn.
- Thứ hạng trang web trong thuật toán PageRank được tính phụ thuộc vào câu truy vấn, còn thứ hạng trang web trong thuật toán HITS thì KHÔNG phụ thuộc vào câu truy vấn.
- Trong cả hai thuật toán PageRank và HITS, thứ hạng trang web được tính KHÔNG phụ thuộc vào câu truy vấn.
- Trong cả hai thuật toán PageRank và HITS, thứ hạng trang web được tính phụ thuộc vào câu truy vấn.

Câu 102: Chức năng tư vấn người dùng mua hàng trên web thương mại điện tử là kết quả của khai thác

- Nội dung web
- ✓ Lịch sử sử dụng web
- Cấu trúc web

- Thứ hạng web

Câu 103: Thao tác tính hạng trang web trong máy tìm kiếm là kết quả của khai thác

- Số lượt truy cập web
- ✓ Cấu trúc web
- Lịch sử sử dụng web
- Nội dung web

Câu 104: Hai bước chính của máy tìm kiếm thực hiện là gì?

- ✓ Truy hồi thông tin và tính hạng trang web.
- Lập chỉ mục thông tin và tính hạng trang web.
- Rút trích thông tin và truy hồi thông tin.
- Rút trích thông tin và tính hạng trang web.

Câu 105: Ký hiệu $a(i)$ và $h(i)$ lần lượt là authority và hub của trang i . Hãy chọn công thức đúng theo thuật toán HITS để tính hạng cho trang 1 theo lược đồ bên dưới
[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image026.jpg>]

- ✓ $h(1) = a(5) + a(6) + a(7); a(1) = h(2) + h(3) + h(4)$
- $h(1) = h(2) + h(3) + h(4); a(1) = a(5) + a(6) + a(7)$
- $h(1) = a(2) + a(3) + a(4); a(1) = h(5) + h(6) + h(7)$
- $h(1) = h(5) + h(6) + h(7); a(1) = a(1) + a(2) + a(3)$

Câu 106: Nếu sử dụng thuật toán PageRank để phân hạng các trang web liên kết như đồ thị bên dưới, giả sử khởi đầu 4 nút đều có điểm pagerank là $1/4$. Sau 3 lần lặp, nút nào sẽ có điểm pagerank cao nhất?

[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image028.jpg>]

- a
- b
- c
- ✓ d

Câu 107: Vì sao một tổ chức, công ty cần phải khai thác dữ liệu?

- Vì tổ chức đó có quá nhiều dữ liệu.
- Vì tổ chức đó thiếu lao động có tri thức.
- Vì cơ sở dữ liệu của tổ chức đó không được thiết kế tốt.
- ✓ Vì tổ chức đó muốn rút trích tri thức từ nguồn dữ liệu có sẵn.

Câu 108: Trong quy trình khám phá tri thức, khai thác dữ liệu là bước nằm ngay sau thao tác

- Thu thập dữ liệu
- ✓ Tiền xử lý dữ liệu
- Trực quan hóa dữ liệu
- Tích hợp dữ liệu

Câu 109: Khái niệm nào sau đây không có liên quan đến khai thác dữ liệu?

- Phát hiện tri thức (Knowledge discovery)
- Rút trích tri thức (knowledge extraction)
- Nghiệp vụ thông minh (business intelligence)
- ✓ Phân tích nghiệp vụ (business analysis)

Câu 110: Hãy chọn phát biểu ĐÚNG trong các phát biểu sau:

- Khai thác dữ liệu là một bước tiến của khoa học máy tính.
- ✓ Khai thác dữ liệu là một bước tiến của khoa học dữ liệu.
- Khai thác dữ liệu là một bước tiến của khoa học thực nghiệm.
- Khai thác dữ liệu là một bước tiến của khoa học lý thuyết.

Câu 111: Một siêu thị muốn triển khai một số chính sách khuyến mãi cho khách hàng. Siêu thị muốn chính sách khuyến mãi của mình phù hợp nhất với nhu cầu của khách hàng. Vậy, siêu thị nên thực hiện bài toán gì trên thông tin và lịch sử mua hàng của khách hàng?

- ✓ Phân cụm
- Phân lớp
- Hồi quy
- Phân tích luật kết hợp

Câu 112: Trong một công ty, ai là người sử dụng kết quả của khai thác dữ liệu?

- Bộ phận quản trị cơ sở dữ liệu
- Bộ phận phân tích dữ liệu
- ✓ Bộ phận ra quyết định nghiệp vụ
- Bộ phận quản trị hệ thống thông tin

Câu 113: Bạn phân tích dữ liệu về dân số của một địa phương, sau đó bạn muốn dự đoán tỷ lệ sinh của địa phương đó trong năm tới. Bạn dùng bài toán

- Phân cụm
- Phân lớp
- ✓ Hồi quy
- Phân tích tương quan

Câu 114: Việc khai thác mối quan hệ bạn bè trên mạng xã hội sử dụng bài toán

- ✓ Khai thác đồ thị
- Khai thác hình ảnh
- Khai thác văn bản
- Khai thác đa phương tiện

Câu 115: Bùng nổ dữ liệu không liên quan đến

- Năng lực của máy tính
- Định luật Moore
- Giá thành phần cứng
- ✓ Giá thành phần mềm

Câu 116: Tại sao nói khai thác dữ liệu được xem là là một hợp lưu của nhiều lĩnh vực?

- Vì phải có nhiều tổ chức cùng hợp lại để thực hiện khai thác dữ liệu.
- ✓ Vì khai thác dữ liệu sử dụng kiến thức, kỹ thuật của nhiều lĩnh vực khác nhau.
- Vì khai thác dữ liệu được sử dụng trong nhiều lĩnh vực khác nhau
- Vì nhiều lĩnh vực cùng sử dụng những bài toán khai thác dữ liệu giống nhau

Câu 117: Câu hỏi “Ta nên duyệt cho khách hàng có thông tin là X vay bao nhiêu tiền?” có thể trả lời bằng:

- Phân tích luật kết hợp
- ✓ Hồi quy
- Phân lớp
- Phân cụm

Câu 118: Câu hỏi “Liệu doanh thu của công ty tăng hay giảm trong 3 tháng kế tiếp?” có thể được trả lời bằng:

- Phân tích luật kết hợp
- ✓ Phân tích dữ liệu chuỗi thời gian
- Phân lớp
- Phân cụm

Câu 119: Để xác định đặc trưng của các đối tượng khách hàng ta sử dụng bài toán

- Phân lớp
- ✓ Phân cụm
- Khai thác tập phổ biến
- Hồi quy

Câu 120: Trung vị (median) của một tập dữ liệu là

- Giá trị trung bình của tập dữ liệu đó
- ✓ Giá trị chính giữa của tập dữ liệu đó
- Giá trị cao nhất của tập dữ liệu đó
- Giá trị xuất hiện nhiều nhất của trong tập dữ liệu

Câu 121: Chiều cao, cân nặng là dữ liệu kiểu

- Số rời rạc
- ✓ Số liên tục
- Định danh

- Thứ bậc

Câu 122: Hãy cho biết hình bên dưới được gọi là gì?

[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image001.png>]

- ✓ Boxplot
- Scatter plot
- Histogram
- Bar chart

Câu 123: Để đánh giá mức độ đáng tin của giá trị trung bình (mean) của một tập dữ liệu, ta cần xem xét thêm giá trị

- Trung vị
- Yếu vị (mode)
- ✓ Độ lệch chuẩn
- Cực đại

Câu 124: Dữ liệu nào là dữ liệu nhị phân đối xứng?

- Kết quả xét nghiệm HIV (+/-)
- Độ an toàn của khách hàng vay tiền (+/-)
- ✓ Giới tính (+/-)
- Đánh giá khách hàng tiềm năng (+/-)

Câu 125: [<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image002.jpg>] là công thức tính khoảng cách

- Manhattan
- ✓ Minkowski
- Euclidean
- Hamming

Câu 126: Hai kiểu lược đồ thường sử dụng trong data warehouse là

- Lược đồ hình sao và lược đồ hình vòng.
- Lược đồ hình bông tuyết và lược đồ hình vòng.
- ✓ Lược đồ hình sao và lược đồ hình bông tuyết.
- Lược đồ hình bông tuyết và lược đồ hình trục.

Câu 127: Hãy lựa chọn đặc trưng đúng của OLAP khi so sánh với OLTP

- OLAP thường có dung lượng nhỏ hơn OLTP.
- ✓ OLAP thường có truy vấn phức tạp hơn OLTP.
- OLAP thường có nhiều người dùng cùng lúc hơn OLTP.
- OLAP được cập nhật thường xuyên hơn OLTP.

Câu 128: Trình tự các bước triển khai Data Mart

- Xây dựng, Thiết kế, Cơ sở dữ liệu, Truy cập, Quản lý.
- ✓ Thiết kế, Xây dựng, Cơ sở dữ liệu, Truy cập, Quản lý.
- Quản lý, Thiết kế, Xây dựng, Cơ sở dữ liệu, Truy cập.
- Truy cập, Thiết kế, Xây dựng, Cơ sở dữ liệu, Quản lý.

Câu 129: OLAP là tên viết tắt của kỹ thuật nào?

- Online Advanced Program
- Online Analytical Program
- Online Advanced Processing
- ✓ Online Analytical Processing

Câu 130: Hãy chọn phát biểu ĐÚNG khi so sánh thuật toán Apriori và thuật toán FP-Growth

- FP-Growth có cần số lần quét cơ sở dữ liệu nhiều hơn Apriori.
- ✓ FP-Growth không sinh tập ứng viên như Apriori.
- FP-Growth cần thời gian để đếm độ hỗ trợ của các tập ứng viên ít hơn Apriori.
- FP-Growth dễ cài đặt hơn Apriori.

Câu 131: Độ đo về tính dễ bắt gặp của luật kết hợp được gọi là

- Độ tin cậy (confidence)
- ✓ Độ hỗ trợ (support)
- Độ nâng (lift)
- Độ nhạy (sensitivity)

Câu 132: Trong các khuyết điểm sau đây, khuyết điểm nào KHÔNG phải của thuật toán Apriori?

- Duyệt cơ sở dữ liệu nhiều lần
- ✓ Không sinh tập ứng viên
- Khó khăn trong đếm độ hỗ trợ
- Tốn thời gian thực thi

Câu 133: Theo cơ sở dữ liệu giao tác trong bảng sau, độ hỗ trợ của tập AB là: TID Itemset 100 A B C D 200 A C D 300 A B D 400 B C D E 500 A B C E

- ✓ 60%
- 75%
- 80%
- 65%

Câu 134: Theo cơ sở dữ liệu giao tác trong bảng sau, độ hỗ trợ và độ tin cậy của luật kết hợp BC → D lần lượt là: TID Itemset 100 A B C D 200 A C D 300 A B D 400 B C D E 500 A B C E

- ✓ 40%; 66.7%

- 66.7%; 40%
- 40%; 75%
- 75%; 40%

Câu 135: Hãy cho biết phát biểu nào sau đây ĐÚNG với tính chất downward closure?

- Mọi tập con của một tập phổ biến thì không phổ biến.
- Mọi tập con của một tập không phổ biến thì không phổ biến.
- ✓ Mọi tập bao của một tập không phổ biến thì không phổ biến.
- Mọi tập bao của một tập phổ biến thì phổ biến.

Câu 136: Ưu điểm của thuật toán FP-Growth là

- Duyệt đệ quy cây FP
- Nén cơ sở dữ liệu vào cây FP
- ✓ Không sinh tập ứng viên
- Tốn ít bộ nhớ để duy trì cây FP

Câu 137: Thuật toán FP-Growth cần nhiều bộ nhớ để thực thi do

- ✓ Phải nén toàn bộ cơ sở dữ liệu giao tác vào cấu trúc cây FP.
- Phải sinh ra nhiều tập ứng viên.
- Phải quét cơ sở dữ liệu nhiều lần.
- Phải thực thi duyệt cây đệ quy.

Câu 138: Hãy chọn cụm từ nào sau đây có liên quan đến thuật toán FP-Growth?

- Chia để trị
- Tham lam
- Quy hoạch động
- ✓ Đệ quy

Câu 139: Để xác định khách hàng thường mua gì sau khi mua siêu xe Rolls Royce, ta làm bài toán khai thác

- Luật kết hợp phổ biến
- ✓ Luật kết hợp hiếm
- Luật kết hợp phủ định
- Luật kết hợp đa chiều

Câu 140: Cho X là một tập mục, X là tập phổ biến cực đại (max-pattern) khi và chỉ khi X là phổ biến và

- Không tồn tại một tập mục Y sao cho [\[https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image014.png\]](https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image014.png) mà Y có cùng độ hỗ trợ với X
- Không tồn tại một tập mục Y sao cho [\[https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image013.png\]](https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image013.png) , mà Y có cùng độ hỗ trợ với X

✓ Không tồn tại một tập mục Y sao cho

[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image013.png>] , mà Y là phổ biến

• Không tồn tại một tập mục Y sao cho

[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image014.png>] mà Y là phổ biến

Câu 141: Trong các ứng dụng sau, ứng dụng nào KHÔNG PHẢI là ứng dụng của phân lớp?

- Nhận dạng khuôn mặt
- Nhận dạng chữ viết
- Phát hiện thư rác
- ✓ Phân tích giỏ hàng

Câu 142: Công thức sau được ứng dụng trong thuật toán nào?

[<https://s3.cloud.cmctelecom.vn/tvu/QuizImg/IT207-22110228/image017.png>]

- ID3
- CART
- KNN
- ✓ Naïve Bayes

Câu 143: Độ đo Information Gain được sử dụng để

- ✓ Lựa chọn thuộc tính để rẽ nhánh
- Làm điều kiện dừng dựng cây
- Đo độ chính xác của mô hình
- Đo độ bao phủ của mô hình

Câu 144: Thuật toán nào KHÔNG áp dụng được cho dữ liệu trong bảng sau nếu không xử lý gì thêm? Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có

- C4.5
- CART
- ✓ KNN
- ID3

Câu 145: Thuật toán nào sau đây cần giả định các thuộc tính của dữ liệu là độc lập về mặt thống kê?

- KNN
- ✓ Naïve Bayes
- SVM
- Cây quyết định

Câu 146: Overfitting là hiện tượng xảy ra khi

- Sử dụng quá ít dữ liệu để huấn luyện mô hình.
- ✓ Sử dụng quá nhiều dữ liệu để huấn luyện mô hình.
- Sử dụng dữ liệu không chính xác để huấn luyện mô hình.
- Sử dụng dữ liệu không đầy đủ để huấn luyện mô hình.

Câu 147: Độ chính xác (accuracy) của phép tiên đoán trong bảng 2 là: Actual\Prediction
Cancer Not Cancer Total Cancer 1300 1200 2500 Not Cancer 2700 4800 7500 Total
4000 6000 10000

- 75%
- ✓ 61%
- 60%
- 40%

Câu 148: Tại sao nói phân lớp là phương pháp học có giám sát?

- Vì cần theo dõi từng lần lặp của thuật toán.
- Vì cần tri thức của chuyên gia để phân lớp.
- ✓ Vì có thể kiểm định tính chính xác của mô hình và huấn luyện lại.
- Vì có thể giám sát hoạt động của thuật toán bằng một công cụ nào đó.

Câu 149: Dựa vào bảng sau, theo thuật toán Naïve Bayes, khi cần xét độ an toàn cho khách hàng {Thu nhập = "Thấp", Sở hữu nhà = "Không", Tuổi = "30-45", Nghề nghiệp = "CNV"}, ta KHÔNG cần tính xác suất nào? Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có

- ✓ $P(\text{An toàn} = \text{"Có"} | \text{Nghề nghiệp} = \text{"CNV"})$
- $P(\text{Tuổi} = \text{"40-45"} | \text{An toàn} = \text{"Có"})$
- $P(\text{An toàn} = \text{"Không"})$
- $P(\text{An toàn} = \text{"Có"})$

Câu 150: Trong bài toán dự đoán khách hàng tiềm năng để tiếp thị. Một khách hàng là không tiềm năng, nhưng mô hình dự đoán là tiềm năng. Khi đó doanh nghiệp sẽ...

- ✓ Mất chi phí
- Mất lợi nhuận
- Mất cả chi phí lẫn lợi nhuận
- Không mất gì cả

Câu 151: Dựa vào bảng sau, giá trị của xác suất $P(\text{Nghề nghiệp} = \text{"CNV"} | \text{An toàn} = \text{"Không"})$ là: Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có

- ✓ 1/4
- 1/2
- 1/3
- 1

Câu 152: K-means là phương pháp phân cụm dựa trên

- ✓ Phân hoạch (partitioning)
- Mật độ (density based)
- Phân cấp (hierarchical)
- Lưới (grid based)

Câu 153: Trong thuật toán phân cụm dựa trên phân cấp top-down (thuật toán DIANA), với tập dữ liệu có N điểm, kết quả cuối cùng sẽ là:

- ✓ N cụm
- 1 cụm
- K cụm (K là một số cho trước)
- Không xác định được số cụm

Câu 154: Cho các bước sau: B1. Gán các điểm dữ liệu vào các cụm theo trọng tâm gần nhất B2. Chọn k trọng tâm B3. Tính lại trọng tâm mới B4. Tính khoảng cách từ các điểm dữ liệu đến k trọng tâm Thứ tự đúng của các bước theo thuật toán k-means lần lượt là:

- B1 – B2 – B3 – B4
- B2 – B3 – B4 – B1
- ✓ B2 – B4 – B1 – B3
- B2 – B1 – B4 – B3

Câu 155: Ứng dụng nào sau đây KHÔNG phải là ứng dụng của phân cụm?

- Phát hiện ngoại lệ
- Phát hiện thể loại
- ✓ Phát hiện thư rác
- Tiền xử lý cho phân lớp

Câu 156: Đối tượng của khai thác dữ liệu web bao gồm:

- ✓ Cấu trúc web, nội dung web và lịch sử sử dụng web.
- Cấu trúc web, thứ hạng web và lịch sử sử dụng web.
- Nội dung web, thứ hạng web và lịch sử sử dụng web.
- Cấu trúc web, nội dung web và thứ hạng web.

Câu 157: Mục đích chính của khai thác cấu trúc web là tìm ra những mối quan hệ chưa biết giữa

- ✓ Các trang web
- Các siêu liên kết

- Dữ liệu web
- Nội dung web

Câu 158: Chọn phát biểu ĐÚNG trong thuật toán HITS

- Authority của trang p cao khi p được nhiều trang có authority cao trỏ đến
- Hub của trang p cao khi p được nhiều trang có authority cao trỏ đến
- Hub của trang p cao khi p trỏ đến nhiều trang có hub cao
- ✓ Authority của trang p cao khi p được nhiều trang có hub cao trỏ đến