KHAI PHÁ DỮ LIỆU – CHƯƠNG 1
TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU (DATA MINING)

Tổng hợp kiến thức, câu hỏi tự luận và trắc nghiệm

1. GIỚI THIỆU VÀ TÌNH HUỐNG THỰC TIỄN

1.1. Thực tế thúc đẩy Khai Phá Dữ Liệu

Trong thời đại số, chúng ta đang đối mặt với nghịch lý: **giàu dữ liệu nhưng nghèo thông tin** (*"We are data rich, but information poor"*). Lượng dữ liệu khổng lồ được thu thập mỗi ngày nhưng chỉ một phần nhỏ được khai thác để tạo ra giá trị.

Các tình huống điển hình:

- **Phát hiện gian lận tín dụng:** Nhận biết liệu người dùng thẻ tín dụng có phải chủ thẻ thật sự hay không.
- **Dự đoán giá cổ phiếu:** Xây dựng mô hình dự báo biến động giá thị trường.
- **Phát hiện tấn công mạng:** Phân tích lưu lượng mạng để nhận diện các cuộc tấn công dựa trên mẫu lưu lượng.

2. KHÁM PHÁ TRI THỨC TỪ CƠ SỞ DỮ LIỆU (KDD)

2.1. Định nghĩa KDD

- **Frawley et al. (1991):** "KDD là quá trình phi tầm thường (nontrivial) xác định các mẫu hợp lệ, mới lạ, có khả năng hữu ích và cuối cùng là có thể hiểu được."
- **Fayyad et al. (1996):** KDD là quá trình sử dụng cơ sở dữ liệu cùng với các bước lựa chọn, tiền xử lý, biến đổi để áp dụng các phương pháp khai phá dữ liệu nhằm liệt kê các mẫu và đánh giá chúng như tri thức.

2.2. Quy trình KDD (7 bước)

1. **Làm sạch dữ liệu (Data Cleaning):** Loại bỏ nhiễu, xử lý dữ liệu không nhất quán.
2. **Tích hợp dữ liệu (Data Integration):** Kết hợp dữ liệu từ nhiều nguồn khác nhau.
3. **Lựa chọn dữ liệu (Data Selection):** Chọn dữ liệu liên quan đến nhiệm vụ phân tích.

4. **Biến đổi dữ liệu (Data Transformation):** Chuẩn hóa và chuyển đổi định dạng dữ liệu.
5. **Khai phá dữ liệu (Data Mining):** Áp dụng các thuật toán để tìm kiếm mẫu.
6. **Đánh giá mẫu (Pattern Evaluation):** Đánh giá độ thú vị và ý nghĩa của các mẫu.
7. **Trình bày tri thức (Knowledge Presentation):** Trực quan hóa và trình bày tri thức khai thác được.

Quy trình KDD là **lặp lại và tương tác** – mỗi bước có thể cần quay lại bước trước.

3. KHAI PHÁ DỮ LIỆU – KHÁI NIỆM CHÍNH

3.1. Định nghĩa Khai Phá Dữ Liệu

Khai phá dữ liệu (Data Mining – DM) là quá trình:

- **"Trích xuất hoặc khai phá tri thức từ lượng lớn dữ liệu"** (J. Han et al.)
- **"Trích xuất phi tầm thường các thông tin ẩn, chưa được biết trước và có khả năng hữu ích từ dữ liệu"**

Các tên gọi khác: Knowledge Discovery/Mining in Databases (KDD), Knowledge Extraction, Data/Pattern Analysis, Business Intelligence, Information Harvesting.

3.2. Nguồn dữ liệu trong DM

DM xử lý dữ liệu từ nhiều nguồn và kiểu khác nhau:

- **Kiểu dữ liệu:** có cấu trúc, phi cấu trúc, bán cấu trúc.
- **Nguồn:** Flat files, CSDL quan hệ, NoSQL, CSDL giao dịch, Data Warehouse, CSDL không gian/thời gian, văn bản, đa phương tiện, Web, mạng xã hội.
- **Chế độ:** Batch (xử lý theo lô) hoặc Streaming (xử lý thời gian thực – IoT, BI).

3.3. Tri thức khai phá được

- Mô tả các lớp dữ liệu.
- Mẫu phổ biến (frequent patterns), luật kết hợp (association rules).
- Phân loại và dự đoán (classification and prediction).

- Mô hình phân cụm (clustering model).
- Ngoại lệ (outliers), xu hướng (trends).

Hai loại tri thức chính:

- **Mô tả (Descriptive):** Mô tả đặc điểm chung của các đối tượng trong tập dữ liệu.
- **Dự đoán (Predictive):** Khả năng suy luận/dự đoán thông tin mới dựa trên dữ liệu hiện có.

3.4. DM là sự hội tụ của nhiều lĩnh vực

DM đứng ở giao điểm của: **Thống kê**, **Học máy**, **Công nghệ cơ sở dữ liệu**, **Trực quan hóa**, **Khoa học thông tin** và các lĩnh vực khác.

4. NHIỆM VỤ KHAI PHÁ DỮ LIỆU

4.1. Các nhiệm vụ chính

1. **Mô tả dữ liệu (Data description)**
2. **Phân lớp (Classification)**
3. **Dự đoán (Prediction)**
4. **Phân cụm (Clustering)**
5. **Khai phá luật kết hợp (Association rule mining)**
6. **Phân tích xu hướng (Trend analysis)**
7. **Phát hiện ngoại lệ (Outlier detection)**
8. **Phân tích tương đồng (Similarity analysis)**

4.2. 5 yếu tố chính của một nhiệm vụ DM

1. **Dữ liệu liên quan (Task-relevant data):** Nguồn dữ liệu, kiểu dữ liệu, thuộc tính được chọn.
2. **Tri thức kỳ vọng (Expected knowledge):** Loại tri thức cần khai phá (phân lớp, phân cụm, luật kết hợp,...).

3. **Tri thức nền (Background knowledge):** Kiến thức lĩnh vực (domain knowledge) hỗ trợ quá trình huấn luyện và đánh giá.

4. **Độ đo thú vị (Interestingness measures):** Điểm số để đánh giá và so sánh các mô hình; cần đơn giản, chắc chắn, hữu ích và mới lạ.

5. **Đánh giá và trình bày mẫu (Pattern evaluation & knowledge presentation):** Luật, bảng, biểu đồ, đồ thị, cây, khối dữ liệu,...

4.3. 4 thành phần chính của một thuật toán DM

1. **Cấu trúc mô hình/mẫu (Model/pattern structure):** Hàm tổng quát mô tả toàn bộ hoặc một tập con dữ liệu (ví dụ: $Y = aX + b$ là cấu trúc; $Y = 3X + 2$ là mô hình cụ thể).

2. **Hàm điểm số (Score function):** Đánh giá mức độ hiệu quả của mô hình/mẫu; ví dụ: likelihood, SSE, tỷ lệ phân loại sai.

3. **Phương pháp tối ưu hóa và tìm kiếm (Optimization & Search method):** Tìm cấu trúc và tham số phù hợp nhất với hàm điểm số; ví dụ: greedy strategy, heuristics, genetic algorithms.

4. **Chiến lược quản lý dữ liệu (Data management strategy):** Tùy thuộc kích thước, kiểu dữ liệu: tải vào bộ nhớ (nhỏ-vừa) hoặc xử lý phân tán từ đĩa (lớn/big data).

5. QUY TRÌNH KHAI PHÁ DỮ LIỆU

5.1. CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining, khởi xướng 09/1996) là tiêu chuẩn mở phổ biến nhất cho quy trình KDD:

1. **Business Understanding:** Xác định mục tiêu kinh doanh và yêu cầu bài toán.

2. **Data Understanding:** Thu thập, khám phá và kiểm tra dữ liệu ban đầu.

3. **Data Preparation:** Chuẩn bị tập dữ liệu cuối cùng cho việc mô hình hóa.

4. **Modeling:** Áp dụng các kỹ thuật DM, hiệu chỉnh tham số.

5. **Evaluation:** Đánh giá mô hình theo mục tiêu kinh doanh.

6. **Deployment:** Triển khai mô hình vào sản xuất.

SEMMA (SAS Institute): Sample $\rightarrow$ Explore $\rightarrow$ Modify $\rightarrow$ Model $\rightarrow$ Assess.

5.2. Hệ thống Khai Phá Dữ Liệu

Các thành phần của một hệ thống DM:

- **Nguồn dữ liệu:** CSDL, kho dữ liệu, Web, tài liệu.
- **Máy chủ DB/Data Warehouse:** Chuẩn bị dữ liệu tích hợp cho DM.
- **Knowledge base:** Tri thức lĩnh vực/nền.
- **Data mining engine:** Thực thi các nhiệm vụ DM.
- **Pattern evaluation module:** Đánh giá độ thú vị bằng score functions và ngưỡng.
- **User interface:** Tương tác người dùng – chỉ định nhiệm vụ, đánh giá, trực quan hóa.

6. VAI TRÒ VÀ ỨNG DỤNG

DM là công nghệ hiện đại được ứng dụng rộng rãi và "vô hình" trong cuộc sống hàng ngày:

- **Thương mại và tài chính:** Phân tích giỏ hàng, phát hiện gian lận, chấm điểm tín dụng.
- **Y tế:** Chẩn đoán bệnh, dự đoán dịch bệnh.
- **Viễn thông:** Phân tích lưu lượng mạng, phát hiện tấn công.
- **Khoa học:** Phân tích gen, khám phá thuốc mới.
- **Marketing:** Cá nhân hóa quảng cáo, phân khúc khách hàng.

7. TÓM TẮT

- DM/KDD: Trích xuất mẫu thú vị từ tập dữ liệu lớn; tri thức khai phá phải hợp lệ, hữu ích, mới lạ và có thể hiểu được.
- Quy trình KDD gồm 7 bước lặp lại; DM là thành phần cốt lõi.
- Các nhiệm vụ DM: mô tả, dự đoán, phân lớp, phân cụm, luật kết hợp, ngoại lệ, xu hướng.
- 5 yếu tố: dữ liệu liên quan, tri thức kỳ vọng, tri thức nền, độ đo thú vị, trình bày tri thức.

- 4 thành phần thuật toán: cấu trúc mô hình, hàm điểm số, tối ưu hóa/tìm kiếm, quản lý dữ liệu.
- DM là sự hội tụ của thống kê, học máy, CSDL, trực quan hóa.

8. CÂU HỎI TỰ LUẬN

Câu 1. Khai phá dữ liệu (Data Mining) là gì? Tại sao lại có câu nói "We are data rich, but information poor"? Trình bày mối quan hệ giữa DM và KDD.

Câu 2. Trình bày đầy đủ 7 bước trong quy trình KDD. Tại sao quy trình này mang tính **lặp lại và tương tác**? Cho ví dụ một bước có thể cần quay lại bước trước.

Câu 3. Phân biệt hai loại tri thức trong DM: **mô tả (descriptive)** và **dự đoán (predictive)**. Cho ví dụ cụ thể cho mỗi loại.

Câu 4. DM được mô tả là "sự hội tụ của nhiều lĩnh vực". Hãy trình bày vai trò của: thống kê, học máy, công nghệ CSDL và trực quan hóa trong DM.

Câu 5. Liệt kê và giải thích 8 nhiệm vụ chính của khai phá dữ liệu. Phân loại chúng thành nhiệm vụ mô tả và nhiệm vụ dự đoán.

Câu 6. Trình bày **5 yếu tố chính** mô tả một nhiệm vụ DM. Với mỗi yếu tố, nêu một ví dụ cụ thể trong bài toán phân loại email spam.

Câu 7. Trình bày **4 thành phần chính** của một thuật toán DM. Phân tích từng thành phần trong thuật toán K-means.

Câu 8. Giải thích khái niệm **cấu trúc mô hình (model structure)** và **cấu trúc mẫu (pattern structure)**. Sự khác biệt giữa mô hình và mẫu là gì?

Câu 9. **Hàm điểm số (score function)** trong thuật toán DM cần đáp ứng những yêu cầu nào? Cho ví dụ 3 hàm điểm số khác nhau và chỉ ra ưu/nhược điểm của mỗi loại.

Câu 10. Trình bày quy trình **CRISP-DM**. Tại sao đây là tiêu chuẩn phổ biến nhất trong công nghiệp? So sánh CRISP-DM với SEMMA.

Câu 11. Mô tả kiến trúc của một **hệ thống khai phá dữ liệu** điển hình. Vai trò của từng thành phần là gì?

Câu 12. Giải thích **chiến lược quản lý dữ liệu** trong thuật toán DM. Cách tiếp cận khác nhau như thế nào giữa tập dữ liệu nhỏ/vừa và lớn?

Câu 13. Kể tên và mô tả ngắn gọn ít nhất 5 ứng dụng thực tiễn của DM trong cuộc sống. Với mỗi ứng dụng, nêu rõ nhiệm vụ DM tương ứng.

Câu 14. Phân tích **vai trò lịch sử** của DM trong sự phát triển của công nghệ cơ sở dữ liệu từ những năm 1960 đến hiện tại.

Câu 15. Tại sao DM được mô tả là "ubiquitous and invisible" (phổ biến và vô hình) trong cuộc sống hiện đại? Trình bày ít nhất 3 ứng dụng mà người dùng thường xuyên tiếp xúc mà không nhận ra.

Câu 16. So sánh **học có giám sát (supervised learning)**, **học không giám sát (unsupervised learning)** và **học tăng cường (reinforcement learning)** trong bối cảnh DM. Mỗi phương pháp phù hợp với nhiệm vụ DM nào?

Câu 17. Trình bày sự khác biệt giữa **Statistics (thống kê)** và **Data Mining**. Thống kê mô tả (descriptive statistics) và thống kê suy diễn (inductive statistics) đóng vai trò gì trong DM?

Câu 18. Các nguồn dữ liệu nào có thể được sử dụng trong DM? Thảo luận về thách thức khi khai phá từ dữ liệu **streaming (thời gian thực)** so với dữ liệu **batch**.

Câu 19. Giải thích khái niệm **độ đo thú vị (interestingness measure)**. Một độ đo thú vị cần thỏa mãn những tiêu chí nào? Tại sao nó quan trọng trong DM?

Câu 20. Phân tích mối quan hệ giữa **DM và công nghệ cơ sở dữ liệu (DB technologies)**. Công nghệ CSDL hỗ trợ DM như thế nào? Cho ví dụ về các hệ thống DM tích hợp vào DBMS.

9. CÂU HỎI TRẮC NGHIỆM

Câu 1. Định nghĩa nào sau đây mô tả chính xác nhất về Khai Phá Dữ Liệu (DM)?

A. Quá trình truy vấn dữ liệu từ cơ sở dữ liệu quan hệ.

B. Quá trình trích xuất tri thức ẩn, chưa biết và có khả năng hữu ích từ lượng lớn dữ liệu.

C. Quá trình làm sạch và chuẩn hóa dữ liệu trước khi phân tích.

D. Quá trình thiết kế và quản lý cơ sở dữ liệu.

Câu 2. Bước nào sau đây **không** thuộc quy trình KDD?

A. Data Cleaning.

B. Data Mining.

C. Data Normalization.

D. Pattern Evaluation.

Câu 3. Trong quy trình KDD, bước **Data Mining** đứng ở vị trí thứ mấy?

A. Thứ 3.

B. Thứ 4.

C. Thứ 5.

D. Thứ 7.

Câu 4. Tri thức **mô tả (descriptive)** trong DM có đặc điểm:

A. Dự báo giá trị tương lai dựa trên dữ liệu hiện tại.

B. Mô tả các đặc điểm chung của đối tượng trong tập dữ liệu.

C. Phân loại đối tượng mới vào các lớp đã biết.

D. Xác định các luật nhân quả trong dữ liệu.

Câu 5. Phân cụm (Clustering) thuộc loại tri thức nào?

A. Dự đoán (Predictive).

B. Mô tả (Descriptive).

C. Cả hai.

D. Không thuộc loại nào.

Câu 6. DM là sự hội tụ của nhiều lĩnh vực. Lĩnh vực nào **không** được đề cập trong slide?

A. Thống kê (Statistics).

B. Học máy (Machine Learning).

C. Kinh tế học (Economics).

D. Công nghệ cơ sở dữ liệu (DB Technology).

Câu 7. Số yếu tố chính mô tả một nhiệm vụ DM là:

A. 3.

B. 4.

C. 5.

D. 7.

Câu 8. **Hàm điểm số (score function)** trong thuật toán DM dùng để:

A. Quản lý dữ liệu từ đĩa vào bộ nhớ.

B. Đánh giá mức độ hiệu quả/phù hợp của mô hình/mẫu với tập dữ liệu.

C. Tìm kiếm các mẫu phổ biến trong tập dữ liệu.

D. Trực quan hóa kết quả khai phá.

Câu 9. Thành phần nào của thuật toán DM giúp cải thiện **khả năng mở rộng (scalability)** cho dữ liệu lớn?

A. Model/pattern structure.

B. Score function.

C. Optimization and search method.

D. Data management strategy.

Câu 10. CRISP-DM được khởi xướng vào năm nào?

A. 1991.

B. 1993.

C. 1996.

D. 2000.

Câu 11. Bước nào trong CRISP-DM đánh giá mô hình theo **mục tiêu kinh doanh**?

A. Modeling.

B. Data Preparation.

C. Evaluation.

D. Deployment.

Câu 12. **Pattern evaluation module** trong hệ thống DM có nhiệm vụ:

A. Lưu trữ tri thức nền (background knowledge).

B. Áp dụng độ đo thú vị và ngưỡng để lọc mẫu có ý nghĩa.

C. Trực quan hóa kết quả khai phá cho người dùng.

D. Kết nối với các nguồn dữ liệu.

Câu 13. Phân loại (Classification) và dự đoán (Prediction) thuộc loại nhiệm vụ DM nào?

A. Mô tả (Descriptive).

B. Dự đoán (Predictive).

C. Không có phân loại.

D. Phân cụm.

Câu 14. Câu nói "We are data rich, but information poor" phản ánh điều gì?

A. Thiếu thiết bị lưu trữ dữ liệu.

B. Lượng dữ liệu lớn nhưng khả năng rút ra tri thức hữu ích còn hạn chế.

C. Dữ liệu chất lượng thấp do thiếu chuẩn hóa.

D. Cơ sở dữ liệu không đủ lớn để khai phá.

Câu 15. Trong hệ thống DM, **Knowledge base** lưu trữ:

A. Kết quả của quá trình khai phá dữ liệu.

B. Tri thức lĩnh vực/nền hỗ trợ quá trình DM.

C. Tập dữ liệu gốc chưa xử lý.

D. Các mô hình đã huấn luyện.

Câu 16. Theo SEMMA, 5 bước lần lượt là:

A. Sample, Explore, Mine, Model, Assess.

B. Select, Extract, Modify, Mine, Assess.

C. Sample, Explore, Modify, Model, Assess.

D. Summarize, Explore, Modify, Mine, Apply.

Câu 17. Phương pháp tối ưu hóa và tìm kiếm trong thuật toán DM có mục tiêu:

A. Lưu trữ dữ liệu hiệu quả trên đĩa.

B. Tìm cấu trúc và tham số mô hình tối ưu nhất theo hàm điểm số.

C. Trực quan hóa không gian tham số.

D. Đánh giá mức độ thú vị của mẫu.

Câu 18. Nhiệm vụ **Outlier detection** trong DM nhằm mục đích:

A. Nhóm các đối tượng tương đồng thành cụm.

B. Tìm các đối tượng bất thường, không tuân theo đặc tính chung của dữ liệu.

C. Dự đoán giá trị tương lai của thuộc tính số.

D. Khai phá các mẫu phổ biến giữa các item.

Câu 19. Nguồn dữ liệu nào sau đây **không** được đề cập trong slide C1?

A. Flat files.

B. Social networks.

C. Blockchain databases.

D. Time series databases.

Câu 20. Theo tháp quản lý dữ liệu (Data Management Pyramid), từ dưới lên, thứ tự đúng là:

A. Data Sources $\rightarrow$ DM $\rightarrow$ OLAP $\rightarrow$ Data Warehouses $\rightarrow$ Making Decisions.

B. Data Sources $\rightarrow$ Data Warehouses $\rightarrow$ OLAP/Statistical Analysis $\rightarrow$ DM $\rightarrow$ Visualization $\rightarrow$ Making Decisions.

C. DM $\rightarrow$ Data Sources $\rightarrow$ Visualization $\rightarrow$ OLAP $\rightarrow$ Making Decisions.

D. Data Sources $\rightarrow$ Statistical Analysis $\rightarrow$ DM $\rightarrow$ Data Warehouses $\rightarrow$ Making Decisions.

Câu 21. Số bước trong quy trình KDD theo Fayyad et al. (1996) là:

A. 5.

B. 6.

C. 7.

D. 9.

Câu 22. Phương pháp tìm kiếm **greedy strategy** trong DM có đặc điểm:

A. Luôn tìm được nghiệm tối ưu toàn cục.

B. Chọn lựa chọn tốt nhất tại mỗi bước mà không quay lại, có thể bỏ lỡ nghiệm tối ưu toàn cục.

C. Duyệt toàn bộ không gian tìm kiếm.

D. Sử dụng các toán tử tiến hóa (crossover, mutation) để tìm nghiệm.

Câu 23. Trong DM, **Trend analysis** được sử dụng để:

A. Nhóm các đối tượng tương đồng.

B. Phát hiện các quy tắc kết hợp giữa các mục.

C. Phát hiện xu hướng thay đổi của dữ liệu theo thời gian.

D. Đánh giá mức độ thú vị của mẫu.

Câu 24. Loại dữ liệu nào sau đây là dữ liệu **phi cấu trúc (unstructured)**?

A. Bảng dữ liệu trong CSDL quan hệ.

B. File CSV.

C. Email và văn bản tự do.

D. File JSON.

Câu 25. Quy trình KDD là **lặp lại (iterative)** vì lý do nào?

A. Mỗi bước phải thực hiện đúng một lần.

B. Kết quả của bước sau có thể yêu cầu điều chỉnh các bước trước đó.

C. DM luôn cần nhiều vòng lặp để đạt độ chính xác 100%.

D. Phần cứng máy tính không đủ mạnh để xử lý một lần.

Câu 26. **Association rule mining** trong DM thuộc loại tri thức nào?

A. Dự đoán (Predictive).

B. Mô tả (Descriptive).

C. Phân lớp (Classification).

D. Phân cụm (Clustering).

Câu 27. Trong 4 thành phần của thuật toán DM, thành phần nào ảnh hưởng lớn nhất đến **hiệu suất tính toán** khi làm việc với Big Data?

A. Model/pattern structure.

B. Score function.

C. Optimization and search method.

D. Data management strategy.

Câu 28. Thống kê suy diễn (Inductive Statistics) trong DM chủ yếu liên quan đến nhiệm vụ:

A. Mô tả và tóm tắt dữ liệu.

B. Dự đoán và suy luận từ mẫu sang tổng thể.

C. Phát hiện ngoại lệ.

D. Phân cụm dữ liệu.

Câu 29. Hệ thống Oracle Data Mining thuộc thành phần nào của hệ thống DM?

A. User interface.

B. Pattern evaluation module.

C. Data mining engine tích hợp vào DBMS.

D. Knowledge base.

Câu 30. Bước **Data Transformation** trong KDD bao gồm hoạt động nào?

A. Xóa dữ liệu trùng lặp và sửa lỗi.

B. Kết hợp dữ liệu từ nhiều nguồn vào Data Warehouse.

C. Chuẩn hóa (normalization) và xây dựng thuộc tính mới (feature construction).

D. Trực quan hóa mẫu khai phá được.

Câu 31. Trong một hệ thống DM, **User interface** có chức năng:

A. Thực thi các nhiệm vụ DM trên tập dữ liệu.

B. Cho phép người dùng chỉ định nhiệm vụ, xác minh dữ liệu, đánh giá và trực quan hóa kết quả.

C. Lưu trữ tri thức lĩnh vực.

D. Kết nối với Data Warehouse.

Câu 32. Thuật ngữ nào **không** đồng nghĩa với "Data Mining"?

A. Knowledge Discovery in Databases (KDD).

B. Business Intelligence.

C. Data Warehousing.

D. Information Harvesting.

Câu 33. **Phân lớp (Classification)** trong DM thuộc loại học:

A. Học không giám sát.

B. Học có giám sát.

C. Học tăng cường.

D. Học bán giám sát.

Câu 34. **Phân cụm (Clustering)** trong DM thuộc loại học:

A. Học có giám sát.

B. Học tăng cường.

C. Học không giám sát.

D. Học bán giám sát.

Câu 35. Yếu tố nào trong 5 yếu tố của nhiệm vụ DM cung cấp **ngưỡng** để đánh giá kết quả?

A. Task-relevant data.

B. Expected knowledge.

C. Background knowledge.

D. Interestingness measures.

Câu 36. Bước **Knowledge Presentation** trong KDD thường sử dụng:

A. Chỉ bảng số liệu.

B. Biểu đồ, đồ thị, cây, luật, khối dữ liệu,...

C. Chỉ mã nguồn chương trình.

D. Văn bản mô tả không có hình ảnh.

Câu 37. Theo slide C1, **Data Warehouse** phục vụ DM bằng cách:

A. Thực thi trực tiếp các thuật toán phân lớp.

B. Cung cấp dữ liệu tích hợp, được tổ chức theo nhiều chiều (dimension), sẵn sàng cho khai phá.

C. Lưu trữ kết quả khai phá dữ liệu.

D. Trực quan hóa tri thức khai phá được.

Câu 38. Đặc điểm nào phân biệt DM với truy vấn CSDL thông thường (SQL query)?

A. DM truy vấn dữ liệu nhanh hơn SQL.

B. DM tìm các mẫu ẩn, chưa biết trước; SQL chỉ truy xuất dữ liệu đã biết cần tìm.

C. DM chỉ hoạt động với CSDL phi quan hệ.

D. DM và SQL query hoàn toàn giống nhau về mục đích.

Câu 39. Trong 4 thành phần của thuật toán DM, thành phần nào xác định **hình dạng tổng quát** của giải pháp?

A. Score function.

B. Model/pattern structure.

C. Data management strategy.

D. Optimization and search method.

Câu 40. Phát biểu nào sau đây về KDD là **sai**?

A. KDD là quy trình lặp lại và tương tác.

B. DM là thành phần cốt lõi và duy nhất trong KDD.

C. KDD bao gồm 7 bước chính.

D. Mục tiêu KDD là tìm tri thức hợp lệ, mới lạ và hữu ích.

ĐÁP ÁN

Câu hỏi tự luận – Hướng dẫn trả lời

Câu	Nội dung cần trình bày
1	DM: trích xuất tri thức ẩn từ lượng lớn dữ liệu. Nghịch lý: thu thập nhiều dữ liệu nhưng ít rút ra được giá trị hữu ích. KDD là quy trình tổng thể; DM là bước cốt lõi trong KDD.
2	7 bước KDD: làm sạch $\rightarrow$ tích hợp $\rightarrow$ lựa chọn $\rightarrow$ biến đổi $\rightarrow$ khai phá $\rightarrow$ đánh giá mẫu $\rightarrow$ trình bày tri thức. Lặp lại vì kết quả bước sau có thể yêu cầu điều chỉnh bước trước (ví dụ: kết quả DM tệ $\rightarrow$ quay lại Data Transformation).
3	Mô tả: mô tả đặc điểm chung – ví dụ: phân tích hành vi mua hàng của nhóm khách hàng. Dự đoán: suy luận thông tin mới – ví dụ: dự đoán giá cổ phiếu, phát hiện gian lận thẻ tín dụng.
4	Thống kê: mô tả và suy diễn từ mẫu; Học máy: xây dựng mô hình từ dữ liệu; Công nghệ CSDL: quản lý, truy cập, truy vấn dữ liệu lớn hiệu quả; Trực quan hóa: giúp người dùng hiểu tri thức khai phá được.
5	8 nhiệm vụ: mô tả, phân lớp, dự đoán, phân cụm, luật kết hợp, xu hướng, ngoại lệ, tương đồng. Mô tả: mô tả, phân cụm, luật kết hợp, tương đồng. Dự đoán: phân lớp, dự đoán, xu hướng, ngoại lệ.
6	5 yếu tố: (1) dữ liệu liên quan – ví dụ: email có/không nhãn spam; (2) tri thức kỳ vọng – phân lớp; (3) tri thức nền – đặc điểm ngôn ngữ spam; (4) độ đo thú vị – tỷ lệ chính xác; (5) trình bày – bộ phân loại email.
7	4 thành phần trong K-means: (1) cấu trúc mô hình: k clusters với tâm r_i; (2) hàm điểm số: SSE; (3) tối ưu hóa: lặp gán + cập nhật tâm (greedy); (4) quản lý dữ liệu: tải tất cả vào RAM (nếu nhỏ/vừa).
8	Model structure: tóm tắt toàn bộ dataset (global view) – ví dụ $Y = aX + b$. Pattern structure: tóm tắt một tập con dữ liệu thỏa điều kiện cụ thể (local view). Model = cấu trúc đã có giá trị tham số; Pattern = mẫu mô tả một nhóm nhỏ đối tượng.
9	Score function cần: độc lập với dataset, dễ tính, phản ánh đúng chất lượng mô hình. Ví dụ: (1) SSE – nhỏ hơn tốt hơn; (2) Likelihood – lớn hơn tốt hơn; (3) Misclassification rate – nhỏ hơn tốt hơn.
10	CRISP-DM: 6 bước (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment). Phổ biến vì mở, tập trung cả yêu cầu kinh doanh lẫn kỹ thuật. SEMMA (SAS): 5 bước, chủ yếu tập trung kỹ thuật.

Câu	Nội dung cần trình bày
11	Kiến trúc gồm: Data sources, DB/DW server (chuẩn bị dữ liệu), Knowledge base (tri thức nền), DM engine (thực thi nhiệm vụ), Pattern evaluation module (đánh giá mẫu), User interface (tương tác).
12	Nhỏ/vừa: tải toàn bộ vào RAM, xử lý trực tiếp. Lớn/Big Data: lưu trên đĩa/hệ thống phân tán; từng phần được tải vào RAM để xử lý đồng thời; cần hỗ trợ indexing, phân tán, bảo mật.
13	Ví dụ: (1) phát hiện spam – phân lớp; (2) gợi ý sản phẩm – luật kết hợp; (3) phân nhóm khách hàng – phân cụm; (4) phát hiện gian lận thẻ – ngoại lệ; (5) dự đoán giá cổ phiếu – dự đoán.
14	Từ 1960s (thu thập, tạo DB) $\rightarrow$ 1970s-80s (DBMS) $\rightarrow$ 1980s-nay (Advanced DB: OO, spatial, temporal) $\rightarrow$ 1980s-nay (Data Warehousing + DM) $\rightarrow$ 1990s-nay (Web-based DB) $\rightarrow$ Tương lai (Integrated systems).
15	Ví dụ: (1) Netflix/Spotify gợi ý nội dung; (2) Thẻ tín dụng tự động phát hiện gian lận; (3) Google tự hoàn thành tìm kiếm. Người dùng không thấy thuật toán chạy ở hậu trường.
16	Supervised: có nhãn lớp – phân lớp, dự đoán. Unsupervised: không nhãn – phân cụm, luật kết hợp. Reinforcement: học từ phần thưởng – không phổ biến trong DM truyền thống.
17	Thống kê truyền thống: giả thuyết trước, kiểm định; DM: không giả thuyết trước, tìm kiếm mẫu. Descriptive statistics: mô tả đặc trưng dữ liệu (hỗ trợ bước Data Summarization trong KDD). Inductive: suy diễn từ mẫu (hỗ trợ xây dựng mô hình dự đoán).
18	Batch: xử lý dữ liệu tĩnh theo lô, kết quả không cần tức thời, dễ xử lý hơn. Streaming: dữ liệu đến liên tục thời gian thực (IoT), cần xử lý nhanh, không lưu toàn bộ. Thách thức streaming: bộ nhớ hạn chế, cần thuật toán online/incremental.
19	Độ đo thú vị đánh giá giá trị của mẫu/luật. Tiêu chí: đơn giản (simple), chắc chắn (certain), hữu ích (useful), mới lạ (novel). Quan trọng vì không phải mọi mẫu đều có ý nghĩa thực tiễn; giúp lọc ra tri thức thực sự có giá trị.
20	DB technologies giúp: quản lý Big Data (paging, phân tán), tổ chức đa chiều (Data Warehouse), hỗ trợ nhiều kiểu dữ liệu, tối ưu truy vấn, bảo mật. Ví dụ: Oracle DM, SQL Server analyzers, IBM Intelligent Miner.

Câu hỏi trắc nghiệm – Đáp án

Câu	ĐA	Câu	ĐA	Câu	ĐA	Câu	ĐA
1	B	11	C	21	B	31	C
2	C	12	B	22	B	32	B
3	C	13	B	23	C	33	D
4	B	14	B	24	B	34	B
5	B	15	B	25	B	35	B
6	C	16	C	26	B	36	C
7	C	17	B	27	D	37	B
8	B	18	B	28	B	38	B
9	D	19	C	29	C	39	B
10	C	20	B	30	C	40	B