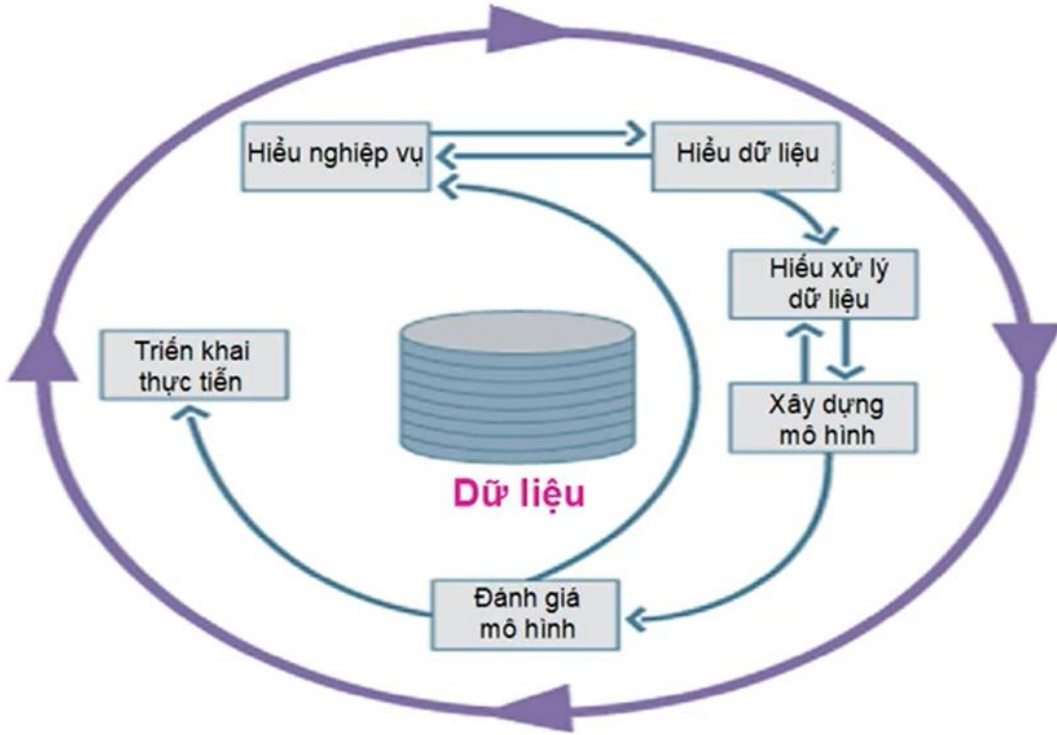


# Tự luận

-----oOo-----

Câu 1: Trình bày quy trình mô hình chuẩn công nghiệp khai phá dữ liệu CRISP-DM. Tại sao bước hiểu kinh doanh và hiểu dữ liệu là quan trọng. Nêu điểm giống và khác nhau (càng nhiều càng tốt) giữa bài toán khai phá dữ liệu và bài toán CSDL.

Câu hỏi	Trả lời	Vị trí
Quy trình mô hình chuẩn công nghiệp khai phá dữ liệu CRISP-DM	<div><ul style="list-style-type: none"><li>• <b>Bước 1: Hiểu nghiệp vụ</b> Ví dụ: Phát hiện bất thường trong cảm biến Internet vạn vật. Phải hiểu như thế nào là bất thường / không thất thường,... Hoặc bài toán phát hiện rời bỏ dịch vụ: liên quan lòng trung thành của khách hàng → đo lòng trung thành</li></ul></div>	

	<ul style="list-style-type: none"> <li>• <b>Bước 2: Hiểu dữ liệu</b> Biết được các giá trị số, mô tả dữ liệu đó. Các dữ liệu này liên quan đến nghiệp vụ ntn → giúp ta hiểu nghiệp vụ hơn và tương tác để tìm hiểu thêm về nghiệp vụ</li> <li>• <b>Bước 3: Tiền xử lý dữ liệu</b> Xây dựng tập dữ liệu đầu vào cho bài toán: thu thập, lựa chọn, chuyển đổi và làm sạch dữ liệu</li> <li>• <b>Bước 4: Xây dựng mô hình</b> Lựa chọn mô hình, xác định tham số tối ưu cho mô hình, kết hợp với việc chuẩn bị dữ liệu nhằm đạt được mô hình có kết quả tối quá trình 3, 4 có thể lặp</li> <li>• <b>Bước 5: Đánh giá mô hình.</b> Mô hình có thể tốt hoặc ko tốt. Nếu ko tốt có thể quay lại nghiệp vụ để đánh giá lại. Nếu Tốt có thể triển khai mô hình thực tiễn hoặc cung cấp thêm tri thức để ta hiểu nghiệp vụ hơn và quay lại bước 1</li> <li>• <b>Bước 6: Triển khai:</b> Triển khai mô hình và đưa vào sử dụng</li> </ul>	
Tại sao bước hiểu kinh doanh và hiểu dữ liệu là quan trọng	<ul style="list-style-type: none"> <li>• <b>Hiểu kinh doanh/ Hiểu nghiệp vụ:</b> Trả lời 5 yếu tố: Ta đã biết được gì, Cần quyết định điều gì, Cái gì cố gắng để đạt được, Cái gì cản trở giải bài toán, Cái gì tìm hiểu thêm được Cung cấp một số nội dung khái quát nhất để hiểu bài toán: Xác định mục tiêu kinh doanh, định danh dữ liệu cần thiết từ dữ liệu sẵn có, tính toán chi phí, lợi nhuận</li> <li>• <b>Hiểu dữ liệu:</b> Giúp hiểu bài toán, kiểu dữ liệu, kích thước dữ liệu, đối tượng dữ liệu, đặc trưng =&gt; Đầu vào thực sự cho bài toán phân tích dữ liệu</li> </ul> <p>Do đó, bước hiểu kinh doanh và hiểu dữ liệu là quan trọng nhất, bởi vì 2 bước này là tiền đề giúp ta nắm bắt được bản chất gốc rễ của dữ liệu, từ đó có thể đưa ra phương hướng tiếp cận và xử lý dữ liệu hiệu quả, chính xác, để tìm lỗi xử lý dữ liệu nếu có phát sinh.</p>	

Nêu điểm giống và khác nhau (càng nhiều càng tốt) giữa bài toán khai phá dữ liệu và bài toán CSDL	So sánh	Khai phá dữ liệu	Cơ sở dữ liệu
	Giống nhau	Khai phá dữ liệu là bước phát triển của công nghệ CSDL, hai lĩnh vực cùng nghiên cứu liên quan đến đối tượng chung là CSDL	
	Dữ liệu	Đa dạng kiểu dữ liệu: <ul style="list-style-type: none"><li>• CSDL quan hệ</li><li>• Kho dữ liệu</li><li>• CSDL giao dịch</li><li>• CSDL mở rộng và kho chứa thông tin (CSDL quan hệ-đối tượng, dữ liệu không gian và thời gian ,...)</li></ul>	Dữ liệu thuộc hệ quản trị CSDL
		Dữ liệu lớn	Dữ liệu với kích thước bất kì
		=> Chỉ làm việc với dữ liệu lớn và rất lớn, dạng bất kì	Kích thước bất kỳ, nhưng dữ liệu đó phải được lưu trữ CSDL nào đó
	Truy vấn	Tính đầy đủ về tri thức	Giả thiết tri thức “đầy đủ” không còn có tính cốt lõi → Cần cải tiến, nâng cấp miền tri thức
		Câu hỏi phức tạp hơn, cần khai thác miền tri thức để có thể đưa ra được câu trả lời, là một quá trình tương tác mới có thể đưa ra kết quả	Câu hỏi đơn giản, mang tính chất truy vấn
		Câu trả lời xấp xỉ, gần đúng	Câu trả lời chính xác, tri thức miền đầy đủ

Câu 2: Khái niệm luật kết hợp và các độ đo liên quan. Khái niệm tập phổ biến và tiếp cận chung hai bước tìm luật kết hợp trong một CSDL giao dịch.

Câu hỏi	Trả lời	Vị trí
Khái niệm luật kết hợp	<p>Luật kết hợp:</p> <p>Gọi <math>A \rightarrow B</math> là luật kết hợp nếu <math>A \subseteq I, B \subseteq I</math> và <math>A \cap B = \emptyset</math></p> <p>Khai phá luật kết hợp:</p> <p>Là tìm ra các mẫu có tần suất cao, các mẫu kết hợp, liên quan hoặc các cấu trúc tồn tại giữa các tập hợp đối tượng trong cơ sở dữ liệu hoặc các giao dịch, cơ sở dữ liệu quan hệ hoặc các kho chứa thông tin khác. Nói cách khác chúng ta đi tìm tất cả các tập phổ biến từ trong dữ liệu</p>	Slide chương 4 trang 6
Các độ đo liên quan đến luật kết hợp	<p>Độ hỗ trợ của luật kết hợp <math>A \rightarrow B</math>  <math>= \text{support}(A \rightarrow B) = P(A \cup B)</math></p> $1 \geq \text{support}(A \rightarrow B) = \text{support}(AB) = P(A \cup B) \geq 0$ <p>Độ tin cậy của luật kết hợp <math>A \rightarrow B</math>  <math>= \text{confidence}(A \rightarrow B) = P(B A)</math></p> $1 \geq \text{confidence}(A \rightarrow B) = \text{confidence}(AB) = P(B A) \geq 0$	
Khái niệm tập phổ biến	<p>Với độ hỗ trợ tối thiểu <math>s_0</math>, tập mục A là tập mục phổ biến nếu <math>\text{support}(A) \geq s_0</math>          Hay xác suất A xuất hiện trong CSDL <math>D \geq s_0</math>          Khi đó A được gọi là tập mục phổ biến</p>	Slide chương 4 trang 5
<p>Khái quát:</p> <p>Hai bước tìm luật kết hợp trong một CSDL giao dịch</p>	<p>Hai bước tìm luật kết hợp trong một CSDL giao dịch:</p> <ul style="list-style-type: none"> <li>Tìm mọi tập mục phổ biến theo min-sup:              Sử dụng quy hoạch động theo nguyên lý tia Apriori.              Nguyên lý tia Apriori: Với tập mục không phổ biến thì không cần phải kiểm tra mọi tập bao nó</li> <li>Sinh luật mạnh từ các tập mục đó              Với mỗi tập mục phổ biến, sinh ra các tập mục con khác rỗng của nó              Với mỗi tập mục con, sinh luật và kiểm tra xem nó có thỏa mãn điều kiện luật mạnh hay không</li> </ul>	Slide chương 4 trang 11

Câu 3: Phát biểu bài toán tìm tập phổ biến và luật mạnh. Trình bày các nội dung cơ bản của thuật toán Apriori tìm tập mục phổ biến (input-output-nội dung với các bước chính). Giải thích sơ bộ về tính đúng đắn của thuật toán.

Câu hỏi	Trả lời	Vị trí
Bài toán tìm tập phổ biến	<p>Bài toán khai thác tập phổ biến lớp bài toán rất quan trọng trong lĩnh vực khai phá dữ liệu. Mục tiêu của nó là tìm tất cả các tập mẫu, liên kết, tương quan hoặc cấu trúc nhân quả có độ phổ biến cao trong tập hợp tất cả các hạng mục hoặc đối tượng trong cơ sở dữ liệu giao dịch, cơ sở dữ liệu quan hệ và các kho thông tin dữ liệu khác</p> <p><b>Đầu vào</b> là các 1 danh sách chứa các bản ghi với mỗi bản ghi là tập các mục được tương tác và 1 tham số làm cột mốc đánh giá tần suất xuất hiện của tập mục có lớn hơn để giữ lại hay không.</p> <p><b>Đầu ra</b> sẽ là những tập mục thoả mãn đề bài cũng như tần suất xuất hiện của từng tập mục.</p> <p>Khai phá mẫu phổ biến : FIM  Cho trước một CSDL giao dịch D, độ hỗ trợ tối thiểu <math>s_0 &gt; 0</math>  Tìm tất cả các tập mục phổ biến từ D: <math>\{ A \subseteq I : \text{support}(A) \geq s_0 \}</math></p>	
Bài toán tìm luật mạnh	<p>Khai phá luật kết hợp mạnh  Cho trước một CSDL giao dịch D, độ hỗ trợ tối thiểu <math>s_0 &gt; 0</math>, độ tin cậy tối thiểu <math>c_0 &gt; 0</math>  Tìm tất cả các luật mạnh từ D: <math>\{ A \rightarrow B : \text{support}(AB) \geq s_0, S(AB)/S(A) \geq c_0 \}</math></p>	
Nội dung thuật toán Apriori	<p>Nội dung thuật toán Apriori:</p> <ul style="list-style-type: none"> <li>Input: CSDL giao dịch <math>D = \{t \mid t \text{ giao dịch}\}</math> và độ hỗ trợ tối thiểu <math>\text{minsup} &gt; 0</math></li> <li>Output: Tập hợp tất cả các tập phổ biến</li> </ul> <p>Các bước chính:</p> <ul style="list-style-type: none"> <li>Sinh ra tất cả các tập phổ biến có độ dài là 1</li> <li>Lặp lại các bước sau đến khi không còn các tập mục phổ biến mới <ul style="list-style-type: none"> <li>Từ các tập phổ biến có độ dài k (khởi tạo <math>k=1</math>), sinh ra các tập mục phổ biến có độ dài <math>k + 1</math> cần xét: tự kết nối dựa trên tập mục đã có</li> <li>Loại bỏ các tập phổ biến có độ dài <math>k + 1</math> chứa các tập con là các tập không phổ biến có độ dài là k.</li> <li>Tính độ hỗ trợ của các tập phổ biến có độ dài <math>k + 1</math> bằng cách duyệt qua tất cả các giao dịch.</li> <li>Thu lại các tập mục phổ biến có độ dài <math>k + 1</math> thoả mãn.</li> </ul> </li> </ul>	

	<ul style="list-style-type: none"> <li>Tập tất cả các tập phổ biến là hợp của tất cả các tập mục phổ biến thu được từ mỗi lần lặp trên</li> </ul> <p>----- Bài Huyền -----</p> <p>Có 2 nhận định dễ thấy như sau:</p> <ul style="list-style-type: none"> <li>Nếu X là tập mục con của Y và X không đủ phổ biến thì dẫn tới Y cũng vậy</li> <li>Nếu Y là tập mục cha của X và Y đủ phổ biến thì dẫn tới X cũng vậy.</li> </ul> <p>Thuật toán này sẽ lấy nền tảng từ 2 tính chất này.</p> <p>Các bước chính của thuật toán:</p> <ul style="list-style-type: none"> <li>Khởi tạo tập các tập mục ứng viên tồn tại trong đầu vào (mỗi mục trong đầu vào ở lần đầu khởi tạo sẽ tạm thời sẽ coi như 1 tập mục ứng viên) rồi (0)</li> <li>Tính toán tần suất xuất hiện của các tập mục này bằng việc truy quét đầu vào.(1)</li> <li>Loại bỏ các tập mục có tần suất không qua ngưỡng cho phép.(2)</li> <li>Từ các tập mục còn lại thêm các mục vào tập sao cho đảm bảo không trùng lặp với các mục vốn có trong tập (item_set) và các tập mục con của tập mục sinh ra có tần suất xuất hiện qua ngưỡng cho phép.(3)</li> <li>Quay lại bước (1)</li> <li>Tiếp tục lặp lại đến khi nào không còn sinh ra được tập mục mới nữa.</li> </ul>	
Tính đúng đắn của thuật toán	Tìm đúng tập phổ biến và tìm hết tập phổ biến.	

Câu 4: Phát biểu bài toán phân lớp. Trình bày thuật toán phân lớp Bayes (mô hình phân lớp, cơ sở lý thuyết, giải pháp chính, ưu điểm- nhược điểm).

Câu hỏi	Trả lời	Vị trí
Phát biểu bài toán phân lớp	<p>Bài toán phân lớp là quá trình phân lớp 1 đối tượng dữ liệu vào 1 hay nhiều lớp đã cho trước nhờ 1 mô hình phân lớp</p> <p>Mô hình này được xây dựng dựa trên 1 tập dữ liệu được xây dựng trước đó có gán nhãn (hay còn gọi là tập huấn luyện).</p> <p>Quá trình phân lớp là quá trình gán nhãn cho đối tượng dữ liệu.</p>	Chương 5 trang 9

	<p><b>Đầu vào</b></p> <ul style="list-style-type: none"> <li>• Tập dữ liệu <math>D = \{d_i\}</math></li> <li>• Tập các lớp <math>C_1, C_2, \dots, C_k</math> mỗi dữ liệu <math>d</math> thuộc một lớp <math>C_i</math></li> <li>• Tập ví dụ <math>D_{\text{exam}} = D_1 + D_2 + \dots + D_k</math> với <math>D_i = \{d \in D_{\text{exam}} : d \text{ thuộc } C_i\}</math></li> <li>• Tập ví dụ <math>D_{\text{exam}}</math> đại diện cho tập <math>D</math></li> <li>• <math>D</math> gồm <math>m</math> dữ liệu <math>d_i</math> thuộc không gian <math>n</math> chiều</li> </ul> <p><b>Đầu ra</b></p> <ul style="list-style-type: none"> <li>• Mô hình phân lớp: ánh xạ từ <math>D</math> sang <math>C</math></li> </ul>	
Thuật toán Bayes	<p><b>Mô hình phân lớp:</b>  Bộ phân lớp Bayes là một giải thuật thuộc lớp giải thuật phân lớp thống kê, nó có thể dự đoán xác suất của một phần tử dữ liệu thuộc vào một lớp là bao nhiêu.</p>	Slide chương 5 trang 49-55
	<p><b>Cơ sở lý thuyết:</b>  <math display="block">P(A   B) = P(B   A) \cdot \frac{P(A)}{P(B)}</math> <ul style="list-style-type: none"> <li>- <math>P(A)</math>: Xác suất của sự kiện <math>A</math> xảy ra</li> <li>- <math>P(B)</math>: Xác suất của sự kiện <math>B</math> xảy ra</li> <li>- <math>P(B   A), P(A   B)</math>: Xác suất (có điều kiện) của sự kiện <math>B</math> xảy ra, nếu biết rằng sự kiện <math>A</math> đã xảy ra</li> <li>- <math>P(A   B)</math>: Xác suất (có điều kiện) của sự kiện <math>A</math> xảy ra, nếu biết rằng sự kiện <math>B</math> đã xảy ra</li> </ul> <p>Phân lớp Bayes được dựa trên định lý Bayes:  Gọi <math>X</math> là một chứng cứ (evidence) (trong bài toán phân lớp thì <math>X</math> là một phần tử dữ liệu), <math>H</math> là một giả thiết để cho <math>X</math> thuộc về một lớp <math>C</math> nào đó. Trong bài toán phân lớp chúng ta muốn xác định giá trị <math>P(H   X)</math>, là xác suất để giả thiết <math>H</math> là đúng với chứng cứ <math>X</math> thuộc vào lớp <math>C</math></p> <math display="block">P(H   X) = \frac{P(X   H) \cdot P(H)}{P(X)}</math> </p>	

	<p><b>Giải pháp chính:</b></p> <p>Xác suất tiên nghiệm: là xác suất thu được từ tập dữ liệu học</p> <p>Xác suất hậu nghiệm: là tính toán thông qua xác suất tiên nghiệm</p> <p>Giả thuyết Naive Bayes: Xác suất xuất hiện của thuộc tính độc lập với ngữ cảnh và vị trí của nó trong đối tượng</p> <p>Mô tả thuật toán Naive Bayes:</p> <p>Tính xác suất tiên nghiệm cho từng nhãn (số lượng đối tượng có nhãn C / tổng số các đối tượng)</p> <p>Tính xác suất tiên nghiệm cho giá trị đặc trưng <math>f_i</math> thuộc lớp C</p> <p>Với mỗi giá trị X đầu vào, tính giá trị xác suất hậu nghiệm: xác suất đối tượng đó nhận nhãn C <math>p(C X)</math>, nếu vượt qua ngưỡng cho trước thì X thuộc lớp C</p>					
	<table><tr><th>Ưu điểm</th><th>Nhược điểm</th></tr><tr><td><ul style="list-style-type: none"><li>Giả định độc lập (nhược điểm cũng là ưu điểm): hoạt động tốt cho nhiều bài toán   miền dữ liệu và ứng dụng</li><li>Đơn giản nhưng đủ tốt để giải quyết nhiều bài toán</li><li>Cho phép kết hợp tri thức tiên nghiệm (prior knowledge) và dữ liệu quan sát được (observed data)</li><li>Huấn luyện mô hình (ước lượng tham số) dễ và nhanh</li><li>Tính gia tăng (incrementality): Không phải ước lượng lại toàn bộ tham số mô hình khi có dữ liệu huấn luyện</li><li>Đầu ra không chỉ là một phân lớp cụ thể mà là một phân bố xác suất trên tập các lớp: có thể áp dụng cho bài toán phân lớp với một đối tượng có thể thuộc nhiều lớp.</li></ul></td><td><ul style="list-style-type: none"><li>Giả định độc lập (naïve assumption): Giả định này sai trong hầu hết các trường hợp thực tế trong đó các thuộc tính trong các đối tượng thường phụ thuộc lẫn nhau. Ngoài ra không kết hợp được các thuộc tính phức hợp (hợp của 2, 3 thuộc tính đơn lẻ chẳng hạn)</li><li>Mô hình không được huấn luyện bằng một phương pháp tối ưu mạnh và chặt chẽ: Tham số của mô hình là các ước lượng xác suất điều kiện đơn lẻ.</li></ul></td></tr></table>	Ưu điểm	Nhược điểm	<ul style="list-style-type: none"><li>Giả định độc lập (nhược điểm cũng là ưu điểm): hoạt động tốt cho nhiều bài toán   miền dữ liệu và ứng dụng</li><li>Đơn giản nhưng đủ tốt để giải quyết nhiều bài toán</li><li>Cho phép kết hợp tri thức tiên nghiệm (prior knowledge) và dữ liệu quan sát được (observed data)</li><li>Huấn luyện mô hình (ước lượng tham số) dễ và nhanh</li><li>Tính gia tăng (incrementality): Không phải ước lượng lại toàn bộ tham số mô hình khi có dữ liệu huấn luyện</li><li>Đầu ra không chỉ là một phân lớp cụ thể mà là một phân bố xác suất trên tập các lớp: có thể áp dụng cho bài toán phân lớp với một đối tượng có thể thuộc nhiều lớp.</li></ul>	<ul style="list-style-type: none"><li>Giả định độc lập (naïve assumption): Giả định này sai trong hầu hết các trường hợp thực tế trong đó các thuộc tính trong các đối tượng thường phụ thuộc lẫn nhau. Ngoài ra không kết hợp được các thuộc tính phức hợp (hợp của 2, 3 thuộc tính đơn lẻ chẳng hạn)</li><li>Mô hình không được huấn luyện bằng một phương pháp tối ưu mạnh và chặt chẽ: Tham số của mô hình là các ước lượng xác suất điều kiện đơn lẻ.</li></ul>	
Ưu điểm	Nhược điểm					
<ul style="list-style-type: none"><li>Giả định độc lập (nhược điểm cũng là ưu điểm): hoạt động tốt cho nhiều bài toán   miền dữ liệu và ứng dụng</li><li>Đơn giản nhưng đủ tốt để giải quyết nhiều bài toán</li><li>Cho phép kết hợp tri thức tiên nghiệm (prior knowledge) và dữ liệu quan sát được (observed data)</li><li>Huấn luyện mô hình (ước lượng tham số) dễ và nhanh</li><li>Tính gia tăng (incrementality): Không phải ước lượng lại toàn bộ tham số mô hình khi có dữ liệu huấn luyện</li><li>Đầu ra không chỉ là một phân lớp cụ thể mà là một phân bố xác suất trên tập các lớp: có thể áp dụng cho bài toán phân lớp với một đối tượng có thể thuộc nhiều lớp.</li></ul>	<ul style="list-style-type: none"><li>Giả định độc lập (naïve assumption): Giả định này sai trong hầu hết các trường hợp thực tế trong đó các thuộc tính trong các đối tượng thường phụ thuộc lẫn nhau. Ngoài ra không kết hợp được các thuộc tính phức hợp (hợp của 2, 3 thuộc tính đơn lẻ chẳng hạn)</li><li>Mô hình không được huấn luyện bằng một phương pháp tối ưu mạnh và chặt chẽ: Tham số của mô hình là các ước lượng xác suất điều kiện đơn lẻ.</li></ul>					



Câu 5: Trình bày thuật toán phân lớp cây quyết định (mô hình phân lớp, cơ sở lý thuyết, giải pháp chính, ưu điểm- nhược điểm, ví dụ đơn giản).

Câu hỏi	Trả lời	Vị trí
Thuật toán cây quyết định	<p><b>Mô hình phân lớp:</b>            Cây quyết định: Cây gán nhãn ở cả nút lẫn cung</p> <ul style="list-style-type: none"> <li>• Lá: giá trị biểu thị lớp, chính xác có một cung vào và không có cung ra.</li> <li>• Nút trong: tên thuộc tính; có chính xác một cung vào và một số cung ra (gắn với điều kiện kiểm tra giá trị thuộc tính của nút)</li> <li>• Gốc không có cung vào</li> </ul> <p>Cây quyết định có thể dùng để phân lớp các đối tượng dựa vào dãy các luật.</p>	Slide chương 5 trang 32-34
	<p><b>Cơ sở lý thuyết:</b>            Xây dựng cây quyết định</p> <ul style="list-style-type: none"> <li>• Phương châm: <i>chia để trị và đệ quy</i>. Mỗi nút tương ứng với một tập các ví dụ học và gốc sẽ là toàn bộ tập dữ liệu.</li> <li>• Sử dụng các độ đo thuộc tính để sắp xếp các nút đặc trưng.</li> </ul>	
	<p><b>Giải pháp chính:</b>            Tại mỗi bước, một thuộc tính tốt nhất sẽ được chọn ra dựa trên một tiêu chuẩn nào đó (chúng ta sẽ bàn sớm). Với mỗi thuộc tính được chọn, ta chia dữ liệu vào các child node tương ứng với các giá trị của thuộc tính đó rồi tiếp tục áp dụng phương pháp này cho mỗi child node. Việc chọn ra thuộc tính tốt nhất ở mỗi bước như thế này được gọi là cách chọn greedy (tham lam). Cách chọn này có thể không phải là tối ưu, nhưng trực giác cho chúng ta thấy rằng cách làm này sẽ gần với cách làm tối ưu. Ngoài ra, cách làm này khiến cho bài toán cần giải quyết trở nên đơn giản hơn.            Sau mỗi câu hỏi, dữ liệu được phân chia vào từng child node tương ứng với các câu trả lời cho câu hỏi đó. Câu hỏi ở đây chính là một thuộc tính, câu trả lời chính là giá trị của thuộc tính đó. Để đánh giá chất lượng của một cách phân chia, chúng ta cần đi tìm một phép đo.            Một số phép đo: độ lợi thông tin, chỉ số gini</p> <p><b>Thuật toán ID3:</b></p> <ul style="list-style-type: none"> <li>• Tạo một nút gốc Root cho cây quyết định</li> </ul>	

	<ul style="list-style-type: none"><li>• Nếu toàn bộ tập ví dụ học đều thuộc cùng 1 lớp thì trả lại nút gốc 1 nút đơn với nhãn của tập ví dụ học</li><li>• Nếu danh sách thuộc tính là rỗng thì trả về cho nút gốc 1 nút đơn với nhãn bằng giá trị nhãn phổ biến nhất trong tập ví dụ học</li><li>• Với các giá trị còn lại:<ul style="list-style-type: none"><li>+ Gán thuộc tính từ danh sách thuộc tính mà phân lớp tập ví dụ học tốt nhất cho A</li><li>+ Gán A cho thuộc tính quyết định nút gốc</li><li>+ Thực hiện lặp lại với mỗi giá trị vi có thể có của A:<ul style="list-style-type: none"><li>○ Cộng thêm một nhánh con của gốc, mang giá trị phù hợp với <math>A = v_i</math></li><li>○ Tìm tập ví dụ học con <math>E_i</math> chứa tập các ví dụ học có giá trị A bằng với vi</li><li>○ Nếu <math>E_i</math> rỗng, dưới mỗi nhánh dưới thêm 1 nút lá với giá trị nhãn bằng giá trị nhãn phổ biến nhất trong tập ví dụ học</li><li>○ Nếu <math>E_i</math> không rỗng, tiếp tục thực hiện thuật toán ID3 với tập ví dụ học là <math>E_i</math>, tập nhãn đầu vào và tập thuộc tính là tập thuộc tính đầu vào không có thuộc tính A</li></ul></li></ul></li><li>• Trả về cây vừa tạo được</li></ul>					
	<table><tr><th>Ưu điểm</th><th>Nhược điểm</th></tr><tr><td><ul style="list-style-type: none"><li>• Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.</li><li>• Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả</li><li>• Có thể làm việc với cả dữ liệu số và dữ liệu phân loại</li><li>• Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê</li><li>• Có khả năng là việc với dữ liệu lớn</li></ul></td><td><ul style="list-style-type: none"><li>• Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thậm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.</li><li>• Cây quyết định hay gặp vấn đề overfitting</li></ul></td></tr></table>	Ưu điểm	Nhược điểm	<ul style="list-style-type: none"><li>• Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.</li><li>• Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả</li><li>• Có thể làm việc với cả dữ liệu số và dữ liệu phân loại</li><li>• Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê</li><li>• Có khả năng là việc với dữ liệu lớn</li></ul>	<ul style="list-style-type: none"><li>• Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thậm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.</li><li>• Cây quyết định hay gặp vấn đề overfitting</li></ul>	
Ưu điểm	Nhược điểm					
<ul style="list-style-type: none"><li>• Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.</li><li>• Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả</li><li>• Có thể làm việc với cả dữ liệu số và dữ liệu phân loại</li><li>• Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê</li><li>• Có khả năng là việc với dữ liệu lớn</li></ul>	<ul style="list-style-type: none"><li>• Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thậm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.</li><li>• Cây quyết định hay gặp vấn đề overfitting</li></ul>					

Câu 6: Trình bày thuật toán phân lớp k- láng giềng gần nhất (căn cứ, giải pháp chính, ưu điểm- nhược điểm, ví dụ đơn giản).

Câu hỏi	Trả lời		Vị trí
Thuật toán k láng giềng gần nhất	<b>Giới thiệu:</b> Khi đưa vào một phần tử dữ liệu mới, giải thuật sẽ tìm k phần tử dữ liệu gần nhất rồi dựa trên nhãn lớp của từng láng giềng này để quyết định nhãn của phần tử dữ liệu mới		Slide chương 5 trang 54-55
	<b>Giải pháp: thuật toán phân lớp kNN</b> Cho trước <ul style="list-style-type: none"><li>Một tập D các đối tượng dữ liệu biểu diễn bản ghi các đặc trưng</li><li>Một đo đo khoảng cách hoặc tương tự nào đó</li><li>Một số <math>k &gt; 0</math> (số láng giềng gần nhất)</li></ul> Phân lớp đối tượng mới $X_c$ được biểu diễn <ul style="list-style-type: none"><li>Tính khoảng cách (độ tương tự) từ X tới tất cả dữ liệu thuộc D</li><li>Tìm k dữ liệu thuộc D gần X nhất</li><li>Dùng nhãn lớp của k-láng giềng gần nhất để xác định nhãn lớp của X: nhãn nhiều nhất trong k-láng giềng gần nhất</li></ul>		
	Ví dụ:		
	<b>Ưu điểm</b>	<b>Nhược điểm</b>	
	<ul style="list-style-type: none"><li>Không tốn thời gian để học: Độ phức tạp tính toán quá trình training = 0</li><li>Đơn giản</li><li>Không cần giả sử về phân phối của các class</li></ul>	<ul style="list-style-type: none"><li>Nhạy cảm với nhiễu nếu k nhỏ</li><li>kNN là thuật toán mà mọi tính toán đều nằm ở khâu test. Trong đó, việc tính khoảng cách tới từng điểm dữ liệu trong training set sẽ tốn nhiều thời gian, đặc biệt với các CSDL có số chiều lớn và nhiều điểm dữ liệu</li><li>k lớn <math>\rightarrow</math> độ phức tạp tăng</li></ul>	

Câu 7: Trình bày thuật toán phân lớp SVM (mô hình phân lớp, cơ sở lý thuyết, giải pháp chính, ưu điểm- nhược điểm, nhận xét).

Câu hỏi	Trả lời		Vị trí
Thuật toán SVM	<b>Mô hình phân lớp:</b> Thuật toán SVM là thuật toán thuộc lớp giải thuật phân lớp thống kê. Nó có khả năng xử lý cả dữ liệu tuyến tính và phi tuyến. Bản chất của giải thuật này là nó xây dựng một siêu phẳng để phân chia dữ liệu thành 2 nửa. Trong trường hợp dữ liệu không tuyến tính thì nó sẽ sử dụng một hàm nhân (kernel function) để chuyển đổi tập dữ liệu ban đầu sang một không gian mới có số chiều lớn hơn để xử lý		Slide chương 5 trang 58-60
	<b>Cơ sở lý thuyết:</b> SVM thực chất là một bài toán tối ưu , mục tiêu của thuật toán này là tìm được một không gian F và siêu phẳng quyết định f trên F sao cho sai số phân loại là thấp nhất.		
	<b>Giải pháp chính:</b> Tìm một siêu phẳng: $\alpha_{SVM} \cdot d + b$ phân chia dữ liệu thành hai miền. Siêu phẳng tốt nhất là siêu phẳng có margin tới các miền dữ liệu là lớn nhất Phân lớp một tài liệu mới: xác định dấu của <ul style="list-style-type: none"><li><math>f(d) = \alpha_{SVM} \cdot d + b</math></li><li>Thuộc lớp dương nếu <math>f(d) &gt; 0</math></li><li>Thuộc lớp âm nếu <math>f(d) &lt; 0</math></li></ul> Chú ý: Nếu là phân lớp đa lớp, coi 1 lớp mang nhãn 1, các lớp còn lại mang nhãn -1 rồi tiếp tục phân lớp dữ liệu		
	<b>Ưu điểm</b>	<b>Nhược điểm</b>	
	Là một kĩ thuật phân lớp khá phổ biến, SVM thể hiện được nhiều ưu điểm trong số đó có việc tính toán hiệu quả trên các tập dữ liệu lớn. Có thể kể thêm một số ưu điểm của phương pháp này như:	<ul style="list-style-type: none"><li>Bài toán số chiều cao: Trong trường hợp số lượng thuộc tính (p) của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu (n) thì SVM cho kết quả khá tồi.</li></ul>	

	<ul style="list-style-type: none"> <li>Xử lý tốt trên không gian số chiều cao</li> <li>Tiết kiệm bộ nhớ: Do chỉ có một tập hợp con của các điểm được sử dụng trong quá trình huấn luyện và ra quyết định thực tế cho các điểm dữ liệu mới nên chỉ có những điểm cần thiết mới được lưu trữ trong bộ nhớ khi ra quyết định.</li> <li>Linh hoạt: Phân lớp thường là phi tuyến tính. Khả năng áp dụng Kernel mới cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính từ đó khiến cho hiệu suất phân loại lớn hơn.</li> </ul>	<ul style="list-style-type: none"> <li>Chưa thể hiện rõ tính xác suất: Việc phân lớp của SVM chỉ là việc cố gắng tách các đối tượng vào hai lớp được phân tách bởi siêu phẳng SVM. Điều này chưa giải thích được xác suất xuất hiện của một thành viên trong một nhóm là như thế nào. Tuy nhiên hiệu quả của việc phân lớp có thể được xác định dựa vào khái niệm margin từ điểm dữ liệu mới đến siêu phẳng phân lớp mà chúng ta đã bàn luận ở trên.</li> </ul>	
--	--	---	--

Câu 8: Trình bày hai bộ độ đo phổ biến nhất (hồi tưởng-chính xác, chính xác-hệ số lỗi) được sử dụng trong đánh giá các bộ phân lớp dữ liệu. So sánh hai bộ độ đo này.

Câu hỏi	Trả lời	Vị trí													
Ma trận nhầm lẫn	Dùng trong đánh giá bộ phân lớp nhị phân	Slide chương 6 trang 15													
	<table><tr><th colspan="2" rowspan="2">Số lượng các mẫu qua phân lớp và trong thực tế</th><th colspan="2">Giá trị thực</th></tr><tr><th>Thuộc lớp dương</th><th>Thuộc lớp âm</th></tr><tr><th rowspan="2">Giá trị qua bộ phân lớp nhị phân</th><th>Thuộc lớp dương</th><td>TP</td><td>FP</td></tr><tr><th>Thuộc lớp âm</th><td>FN</td><td>TN</td></tr></table>		Số lượng các mẫu qua phân lớp và trong thực tế		Giá trị thực		Thuộc lớp dương	Thuộc lớp âm	Giá trị qua bộ phân lớp nhị phân	Thuộc lớp dương	TP	FP	Thuộc lớp âm	FN	TN
	Số lượng các mẫu qua phân lớp và trong thực tế				Giá trị thực										
			Thuộc lớp dương	Thuộc lớp âm											
	Giá trị qua bộ phân lớp nhị phân		Thuộc lớp dương	TP	FP										
Thuộc lớp âm		FN	TN												
<ul style="list-style-type: none"><li>TP: số ví dụ dương P mà thuật toán phân đúng (T) cho dương P</li><li>TN: số ví dụ âm N mà thuật toán phân đúng (T) cho âm N</li><li>FN: số ví dụ dương P mà thuật toán phân sai (F) cho âm N</li><li>FP: số ví dụ âm N mà thuật toán phân sai (F) cho dương</li></ul>															

Độ hồi tưởng - chính xác	<p>Độ hồi tưởng R (Recall, còn ký hiệu là p) hoặc TPR (True Positive Rate) :</p> $R = TPR = \frac{TP}{TP+FN}$ <p>Độ chính xác (precision):</p> $P = \frac{TP}{TP+FP}$ <p>Tỷ lệ dương giả FPR (False Positive Rate):</p> $FPR = \frac{FP}{(FP+TN)}$ <p>Tỷ lệ âm giả FNR (False Negative Rate):</p> $FNR = \frac{FN}{TP+FN}$ <p>Kết hợp R và P độ đo F :</p> $F_{\beta} = \frac{((\beta^2+1) \times P \times R)}{(\beta^2 \times P + R)}$		Slide chương 6 trang 15
Độ chính xác – nhầm lẫn	<p>Dựa trên ma trận nhầm lẫn:</p> $A = Accuracy = \frac{\text{Số dự báo chính xác}}{\text{Tổng số dự báo}} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$ $E = Error = \frac{\text{Số dự báo sai}}{\text{Tổng số dự báo}} = \frac{(FP + FN)}{(TP + TN + FP + FN)}$		Slide chương 6 trang 20
So sánh	<b>Accuracy - Error rate</b>	<b>Precision - Recall</b>	Slide chương 6 trang 21
	Dễ thực hiện	Phức tạp hơn Accuracy	
	Đánh giá độ chính xác của mô hình	Đánh giá độ tin cậy và tính ổn định của mô hình	

	Sử dụng 1 yếu tố cho độ đo của mình (tỉ lệ chính xác)	Sử dụng nhiều yếu tố cho phép đo (T/F, P/N)	
	Không xác định được độ lệch dữ liệu	Xác định được độ lệch dữ liệu, nhạy cảm với dữ liệu (ví dụ)	
Ví dụ	<p>Tập test có 9990 ví dụ lớp 0 và 10 ví dụ lớp 1.  Kiểm thử: mô hình dự đoán cả 9999 ví dụ là lớp 0 và 1 ví dụ lớp 1. TP=1 FN = 9 TN=9990 FP= 0  Theo phương án (recall, precision) có  <math>= 1/10=0.1; =1/1=1; f1 = 2*0.1/(0.1+1.0)= 0.18</math>  Theo phương án (accuracy, error rate) có  accuracy=0.9991; error rate = 9/10000 = 0.0009</p>		

Câu 9: Phát biểu bài toán phân cụm. Trình bày thuật toán phân cụm K-mean gán cứng (nội dung cơ bản, ưu điểm- nhược điểm).

Câu hỏi	Trả lời	Vị trí
Phát biểu bài toán phân cụm	<p><b>Bài toán:</b>  Bài toán phân cụm dữ liệu là bài toán gom các dữ liệu về các nhóm sao cho các điểm dữ liệu trong cùng 1 nhóm sẽ có sự tương đồng với nhau theo một tiêu chí nào đó.  Cho tập dữ liệu <math>D = \{d_i\}</math>  Cần:  Phân các dữ liệu thuộc D thành các cụm:</p> <ul style="list-style-type: none"> <li>Các dữ liệu trong một cụm: “tương tự” nhau (gần nhau)</li> <li>Dữ liệu hai cụm: “không tương tự” nhau (xa nhau)</li> </ul>	Slide chương 6 trang 11

	<p><b>Độ đo thể hiện quan hệ giữa các điểm dữ liệu:</b></p> <ul style="list-style-type: none"> <li>• Độ đo tương đồng <ul style="list-style-type: none"> <li>+ Biểu diễn: vector n chiều</li> <li>+ Giá trị nhị phân: Ma trận kề, độ đo Jaccard</li> <li>+ Giá trị rời rạc <math>[0, m]</math>: Chuyển m giá trị thành nhị phân, độ đo Jaccard</li> <li>+ Giá trị thực : độ đo cosin hai vector</li> </ul> </li> <li>• Độ đo khác biệt <ul style="list-style-type: none"> <li>+ Đối ngẫu độ đo tương đồng</li> <li>+ Thuộc tính nhị phân: đối xứng, không đối xứng</li> <li>+ Giá trị rời rạc: hoặc tương tự trên hoặc dạng đơn giản (q thuộc tính giống nhau)</li> <li>+ Giá trị thực: Khoảng cách Manhattan, Euclide, Mincowski</li> <li>+ Tính xác định dương, tính đối xứng, tính bất đẳng thức tam giác.</li> </ul> </li> </ul>	
	<p><b>Một số phương pháp phân cụm:</b></p> <ul style="list-style-type: none"> <li>• Phân cụm mô hình và phân cụm phân vùng <ul style="list-style-type: none"> <li>+ Mô hình: Kết quả là mô hình biểu diễn các cụm dữ liệu</li> <li>+ Vùng: Danh sách cụm và vùng dữ liệu thuộc cụm</li> </ul> </li> <li>• Phân cụm đơn định và phân cụm xác suất <ul style="list-style-type: none"> <li>+ Đơn định: Mỗi dữ liệu thuộc duy nhất một cụm</li> <li>+ Xác suất: Danh sách cụm và xác suất một dữ liệu thuộc vào các cụm</li> </ul> </li> <li>• Phân cụm phẳng và phân cụm phân cấp <ul style="list-style-type: none"> <li>+ Phẳng: Các cụm dữ liệu không giao nhau</li> <li>+ Phân cấp: Các cụm dữ liệu có quan hệ phân cấp cha- con</li> </ul> </li> <li>• Phân cụm theo lô và phân cụm tăng <ul style="list-style-type: none"> <li>+ Lô: Tại thời điểm phân cụm, toàn bộ dữ liệu đã có</li> <li>+ Tăng: Dữ liệu tiếp tục được bổ sung trong quá trình phân cụm</li> </ul> </li> </ul>	



Thuật toán phân cụm k-mean	<p><b>Giới thiệu:</b>          Giải thuật k-means thuộc lớp phân cụm phẳng.          Đầu vào: Tập dữ liệu D, số lượng các cụm k          Đầu ra: Tập dữ liệu đã được phân thành k cụm  <b>Mô tả:</b>          Chọn một độ đo khoảng cách hoặc độ đo tương tự nào đó</p> <ul style="list-style-type: none"> <li>• Chọn ngẫu nhiên k phần tử dữ liệu trong tập dữ liệu S làm trọng tâm đại diện cho k cụm</li> <li>• Lặp các bước sau cho đến khi gặp điều kiện dừng</li> </ul> <p>Với mọi dữ liệu d trong S:</p> <ul style="list-style-type: none"> <li>+ Tính khoảng cách similarity của điểm đó với k trọng tâm của k cụm</li> <li>+ Phân các phần tử trong D vào các cụm dựa vào độ tương đồng của nó với trọng tâm các cụm (phân vào cụm có độ tương đồng lớn nhất)</li> <li>• Tính lại trọng tâm của các cụm</li> </ul> <p>Điều kiện dừng: 1 trong 2 điều kiện sau</p> <ul style="list-style-type: none"> <li>• Sau mỗi vòng lặp không có sự thay đổi trọng tâm của cụm</li> <li>• Điều kiện dừng cưỡng bức: khống chế số lần lặp hoặc khi giá trị mục tiêu đủ nhỏ</li> </ul>		Slide chương 6 trang 17
	<b>Ưu điểm</b>	<b>Nhược điểm</b>	
	<ul style="list-style-type: none"> <li>• Đơn giản, dễ hiểu, dễ cài đặt</li> <li>• Hiệu quả về thời gian: tuyến tính <math>O(tkn)</math>, t số lần lặp, k số cụm, n là số phần tử</li> <li>• Một thuật toán phân cụm phổ biến nhất, cho ra kết quả tốt với nhiều bài toán, đặc biệt là với bài toán các cụm có dạng cầu hoặc elipse</li> <li>• Có khả năng mở rộng và dễ dàng sửa đổi với những dữ liệu mới</li> <li>• Đảm bảo hội tụ sau một số lần lặp nhất định, đảm bảo các cụm đều có dữ liệu và số cụm ổn định (k cho trước)</li> </ul>	<ul style="list-style-type: none"> <li>• Phải “tính trung bình được”: dữ liệu phân lớp thì dựa theo tần số</li> <li>• Cần cho trước k : số cụm</li> <li>• Nhạy cảm với ngoại lệ (cách xa so với đại đa số dữ liệu còn lại): ngoại lệ thực tế, ngoại lệ do quan sát sai (làm sạch dữ liệu)</li> <li>• Nhạy cảm với mẫu ban đầu: cần phương pháp chọn mẫu thô tốt</li> <li>• Không thích hợp với các tập dữ liệu không siêu-ellip hoặc siêu cầu (các thành phần con không ellip/cầu hóa)</li> </ul>	

	<ul style="list-style-type: none"> <li>Các cụm được tách biệt rõ ràng, không có hiện tượng 1 đối tượng nằm trong 2 cụm</li> <li>Cho tối ưu cục bộ tốt</li> </ul>		
--	--	--	--

Câu 10: Trình bày thuật toán phân cụm phân cấp HAC (độ đo tương tự cụm, nội dung cơ bản, ưu điểm- nhược điểm).

Câu hỏi	Trả lời	Vị trí
Thuật toán phân cụm phân cấp HAC	<b>Độ đo tương tự cụm</b> <ul style="list-style-type: none"> <li>Độ tương tự giữa hai đại diện cụm</li> <li>Độ tương tự cực đại giữa hai dữ liệu thuộc hai cụm: single-link</li> <li>Độ tương tự cực tiểu giữa hai dữ liệu thuộc hai cụm: complete-link</li> <li>Độ tương tự trung bình giữa hai cụm dữ liệu</li> </ul>	Slide chương 6 trang 23-27
	Thuộc thuật toán phân cụm phân cấp Nội dung cơ bản Input: $D = \{d\}$ tập dữ liệu, độ đo tương tự sim và có thể số cụm $k$ và ngưỡng phân cụm $q > 0$ Output: $G$ : Tập các cụm phân cấp của $D$ Thuật toán: <ul style="list-style-type: none"> <li>Khởi tạo <math>G</math> là tập các cụm chỉ chứa một dữ liệu</li> <li>Lặp lại cho đến khi đủ lượng cụm tối thiểu: <math> G  \leq k</math> <ul style="list-style-type: none"> <li>Tìm hai cụm mà có độ tương tự với nhau cao nhất</li> <li>So sánh độ tương tự của hai cụm đó với ngưỡng đo <math>q</math>: <ul style="list-style-type: none"> <li>Nếu nhỏ hơn <math>q \rightarrow</math> dừng vì độ tương tự giữa các cụm quá bé</li> <li>Nếu <math>\geq q \rightarrow</math> gộp hai cụm lại thành 1 cụm</li> </ul> </li> </ul> </li> </ul> Chú ý: <ul style="list-style-type: none"> <li><math>G</math> là tập các cụm trong phân cụm</li> <li>Điều kiện <math> G  &lt; k</math> có thể thay thế bằng <math> G  = 1</math></li> </ul>	

	Ưu điểm	Nhược điểm	
	<ul style="list-style-type: none"> <li>Không cần giả định trước một số lượng cụm cụ thể</li> <li>Có thể tìm ra được số cụm tối ưu</li> </ul>	<ul style="list-style-type: none"> <li>Một khi đã hợp hai cụm lại làm một, ta không thể tách lại chúng được nữa</li> <li>Các hàm mục tiêu không thể rút gọn được.</li> <li>Kết quả thu được phụ thuộc vào độ đo tương tự và tập dữ liệu đầu vào</li> </ul>	

Câu 11: Trình bày các phương pháp đánh giá thuật toán phân cụm

Câu hỏi	Trả lời	Vị trí
Phương pháp đánh giá thuật toán phân cụm	<p>Một số phương pháp đánh giá phân cụm điển hình</p> <ul style="list-style-type: none"> <li>Người dùng kiểm tra <ul style="list-style-type: none"> <li>Nghiên cứu trọng tâm và miền phủ</li> <li>Luật từ cây quyết định</li> <li>Đọc các dữ liệu trong cụm</li> </ul> </li> <li>Đánh giá theo các độ đo tương tự/khoảng cách <ul style="list-style-type: none"> <li>Độ phân biệt giữa các cụm</li> <li>Phân ly theo trọng tâm</li> </ul> </li> <li>Dùng thuật toán phân lớp <ul style="list-style-type: none"> <li>Coi mỗi cụm là một lớp</li> <li>Học bộ phân lớp đa lớp (cụm)</li> <li>Xây dựng ma trận nhầm lẫn khi phân lớp</li> <li>Tính các độ đo: entropy, tinh khiết, chính xác, hồi tưởng, độ đo F và đánh giá theo các độ đo này</li> </ul> </li> </ul>	Slide chương 6 trang 36

Câu 12: Định nghĩa hệ thống tư vấn, các tính chất và so sánh bài toán lọc cộng tác trong hệ tư vấn với bài toán phân lớp

Câu hỏi	Trả lời		Vị trí
Định nghĩa hệ thống tư vấn	Hệ thống tư vấn là các công cụ phần mềm và kỹ thuật cung cấp các tư vấn về các mục có khả năng cao là hữu ích nhất đối với một người dùng đích. Trong đó: Mục (item) là các sản phẩm/bài viết/trang web/bản nhạc/bộ phim/ con người / tổ chức / v.v. Phân loại hệ thống tư vấn → chủ động; bị động (hệ thống hỏi đáp)		Slide 6: Phân cụm dữ liệu và hệ tư vấn trang 43  Video: ngày 4/4, tại 1:01:01
Tính chất hệ thống tư vấn	<p><b>Tính có liên quan</b></p> <ul style="list-style-type: none"><li>+ Các mục tư vấn cần liên quan tới người dùng: biện minh</li></ul> <p><b>Tính mới lạ</b></p> <ul style="list-style-type: none"><li>+ Tư vấn các mục người dùng chưa hoặc khó quan sát</li><li>+ Tránh tư vấn lặp các mục có tính phổ biến</li></ul> <p><b>Tính “may mắn bất ngờ”</b></p> <ul style="list-style-type: none"><li>+ Tạo ngạc nhiên cho người dùng</li><li>+ Không chỉ là chưa quan sát được</li></ul> <p><b>Tính đa dạng gia tăng</b></p> <ul style="list-style-type: none"><li>+ Các mục tư vấn cần đa dạng, tránh cùng thuộc một thể loại</li><li>+ Lựa chọn tư vấn mục cùng thể loại theo các tư vấn khác nhau</li></ul> <p><b>Tính giải trình</b></p> <ul style="list-style-type: none"><li>+ Nên có giải trình mục được tư vấn</li><li>+ “tư vấn phim”: về đạo diễn, về diễn viên, về thể loại ưa chuộng của người dùng</li></ul>		
So sánh bài toán lọc cộng tác trong hệ tư vấn với bài toán phân lớp	Phân lớp	Lọc cộng tác	Slide chương 6 trang 45, video 4/4 ở 1:13
	<ul style="list-style-type: none"><li>- Biến độc lập và biến phụ thuộc tách rời nhau</li><li>- Các dòng dữ liệu học và các dòng dữ liệu đánh giá tách rời nhau</li><li>- Xây dựng mô hình và đánh giá mô hình độc lập nhau</li></ul>	<ul style="list-style-type: none"><li>- Không chia ranh giới biên phụ thuộc và biến độc lập</li><li>- Không chia ranh giới giữa dòng dữ liệu học và dòng dữ liệu đánh giá</li></ul>	

Câu 13: Giới thiệu chung về hệ thống các kỹ thuật lọc trong hệ tư vấn. Tư vấn xã hội (định nghĩa hẹp và rộng), tư vấn nhóm

Câu hỏi	Trả lời	Vị trí
<p>Hệ thống các kỹ thuật lọc trong hệ tư vấn</p>	<div data-bbox="541 358 1577 773"> <pre> graph TD     Root[Kỹ thuật hệ thống tư vấn] --&gt; Content[Kỹ thuật dựa trên nội dung]     Root --&gt; Collaborative[Kỹ thuật dựa trên cộng tác]     Root --&gt; Knowledge[Kỹ thuật dựa trên tri thức]     Root --&gt; Demographic[Kỹ thuật dựa trên nhân khẩu học]     Root --&gt; Hybrid[Kỹ thuật kết hợp]     Collaborative --&gt; Model[Kỹ thuật dựa trên mô hình]     Collaborative --&gt; Memory[Kỹ thuật dựa trên ghi nhớ]     Model --&gt; Clustering[Phân cụm]     Model --&gt; Association[Luật kết hợp]     Model --&gt; Bayesian[Mạng Bayes]     Model --&gt; Neural[Mạng nơ-ron]     Model --&gt; VV[v.v.]     Memory --&gt; User[Hướng người dùng]     Memory --&gt; Item[Hướng mục] </pre> </div> <p>Kỹ thuật lọc trong hệ tư vấn được phân loại thành 4 nhóm chính</p> <ul style="list-style-type: none"> <li>- Dựa trên nội dung</li> <li>- <b>Dựa trên cộng tác (gồm kỹ thuật dựa trên mô hình, kỹ thuật dựa trên ghi nhớ)</b>  <b>Lọc cộng tác dựa trên ghi nhớ bao gồm cộng tác hướng mục và hướng người dùng</b></li> <li>- Dựa trên tri thức</li> <li>- Dựa trên nhân khẩu học</li> <li>- Kết hợp (của 1 số trong các phương pháp trên)</li> </ul> <p>Kỹ thuật lọc này đòi hỏi tài nguyên ít nhất: chỉ sử dụng ma trận hữu ích → độc lập miền <math>S_u</math> là tập các mục được người dùng <math>u</math> đánh giá, <math>S_i</math> là tập người dùng đã đánh giá tập <math>i</math></p>	<p>Slide chương 6 trang 47</p> <p>Video 4/4 tại 1:23:00</p>
<p>Tư vấn xã hội</p>	<p>Định nghĩa hẹp: Sử dụng mối quan hệ xã hội</p> <p>Định nghĩa rộng: Sử dụng mọi dữ liệu từ phương tiện xã hội</p>	<p>Slide chương 6 trang 57</p> <p>Học trên lớp</p>

<p>Tư vấn nhóm</p>	<p>Có các kiểu nhóm: chính thức, không thường xuyên, ngẫu nhiên, tự động</p> <p>Phương pháp tích hợp dự đoán:  Từ các hồ sơ người dùng → Tư vấn cho từng cá nhân → Các mục hoặc đánh giá mục từ các tư vấn cá nhân → Tích hợp thành tư vấn nhóm → Danh sách mục tư vấn cho nhóm</p> <p>Phương pháp tích hợp mô hình:  Từ các hồ sơ người dùng → Tích hợp hồ sơ các cá nhân trong nhóm người dùng → Tư vấn nhóm → Danh sách mục tư vấn cho nhóm</p>	<p>Side chương 6 trang 59</p>
--------------------	--	-------------------------------

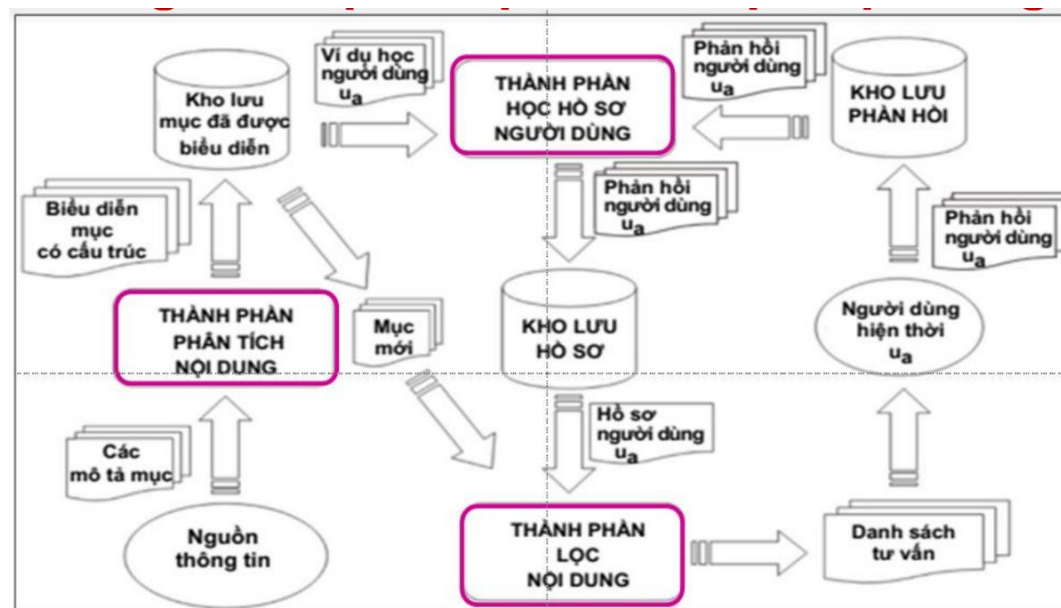
Câu 14: Trình bày sơ đồ khối thủ tục lọc cộng tác hướng người dùng và giải thích sơ bộ.

Câu 15: Trình bày sơ đồ khối thủ tục lọc cộng tác hướng mục và giải thích sơ bộ.

Câu 16: Các thành phần của bộ lọc dựa trên nội dung trong hệ tư vấn.

Câu hỏi	Trả lời	Vị trí
<p>Định nghĩa bộ lọc dựa trên nội dung</p>	<p>Dữ liệu ở dạng nội dung Giả thiết: mối quan tâm mục của người này ít liên quan tới người khác. Mối quan tâm của người theo tính chất của mục.</p> <p>Tổng quát quy trình lọc dựa trên nội dung</p> <pre> graph TD     A[Bắt đầu] --&gt; B[Tiền xử lý dữ liệu và trích xuất đặc trưng mục]     B --&gt; C[Học mô hình sở thích người dùng theo đặc trưng mục dựa trên phân lớp hoặc hồi quy]     C --&gt; D[Xử dụng mô hình để lọc và tư vấn mục]           </pre>	<p>Slide chương 6 trang 51</p>

Các thành phần của bộ lọc dựa trên nội dung



3 thành phần chính:

- Bộ phân tích nội dung:
  - + Thu thập dữ liệu về các mục
  - + Tìm biểu diễn mục dưới dạng có cấu trúc
  - + Sử dụng kỹ thuật trích xuất đặc trưng
  - + Cung cấp đầu vào cho Bộ học hồ sơ và Thành phần lọc
- Bộ lọc hồ sơ người dùng
  - + Thu thập dữ liệu phản hồi của người dùng: bao gồm đánh giá
  - + Tổng quát hóa thành mô hình sở thích của người dùng
  - + Sử dụng kỹ thuật học máy
- Thành phần lọc
  - + Đối sánh biểu diễn mục tiềm năng với mô hình sở thích người dùng
  - + Độ liên quan và chọn các mục có liên quan nhất

Slide chương 6 trang 53



Câu 17: .Các phương thức đánh giá hiệu năng hệ thống tư vấn và các độ đo liên quan.

Câu hỏi	Trả lời	Vị trí
<p>Phương thức đánh giá hiệu năng hệ tư vấn</p>	<ul style="list-style-type: none"> <li>• Người dùng nghiên cứu: <ul style="list-style-type: none"> <li>+ Huy động tập người dùng: Dữ liệu tương tác người dùng-hệ thống</li> <li>+ Lợi thế: hệ thống chạy thực tế. Hạn chế: tuyển dụng người dùng</li> </ul> </li> <li>• Trực tuyến: <ul style="list-style-type: none"> <li>+ Chọn người dùng thực làm việc với hệ thống</li> <li>+ Độ đo tỷ lệ chuyển đổi (conversion rate): tần suất người dùng chọn mục do hệ thống đề xuất</li> <li>+ Chọn 1 từ 2 thuật toán: kiểm thử A/B (A/B test) chọn ngẫu nhiên hai nhóm người dùng A, B, A một thuật toán, B một thuật toán, như nhau về điều kiện và về cùng khoảng thời gian.</li> <li>+ Lợi thế: chọn ngẫu nhiên người dùng → không có thiên vị.</li> <li>+ Hạn chế: không đủ người dùng (khi hệ thống mới làm việc)</li> </ul> </li> <li>• Ngoại tuyến: <ul style="list-style-type: none"> <li>+ Sử dụng bộ dữ liệu lịch sử cho đánh giá: Netflix Prize</li> <li>+ Lợi thế: có sẵn khung và độ đo đánh giá chuẩn</li> <li>+ Hạn chế: dữ liệu quá khứ+hiện tại không phản ánh xu thế sau này</li> <li>+ Chấp nhận rộng rãi và phương pháp phổ biến nhất</li> </ul> </li> </ul>	<p>Slide chương 6 trang 55</p>

Độ đo

Hướng phân lớp:

$$precision@k(u) = \frac{|predicted_k(u) \cap relevant(u)|}{k}$$

$$recall@k(u) = \frac{|predicted_k(u) \cap relevant(u)|}{|relevant(u)|}$$

$$precision@(u) = \frac{|predicted(u) \cap relevant(u)|}{predicted(u)}$$

$$recall@(u) = \frac{|predicted(u) \cap relevant(u)|}{|relevant(u)|}$$

Hướng hồi quy

$$MSE = \frac{\sum_{(u,j) \in E} e_{uj}^2}{|E|}$$

$$RMSE = \sqrt{\frac{\sum_{(u,j) \in E} e_{uj}^2}{|E|}}$$

$$NRMSE = \frac{RMSE}{p_{max} - p_{min}}$$

$$MAE = \frac{\sum_{(u,j) \in E} |e_{uj}|}{|E|}$$

$$NMAE = \frac{MAE}{p_{max} - p_{min}}$$

Trong đó:

sai số toàn phương trung bình (mean squared error: MSE),  
sai số quân phương trung bình (root mean squared error: RMSE, là căn bậc hai của MSE),  
sai số quân phương trung bình chuẩn hóa (normalized RMSE: NRMSE),  
sai số tuyệt đối trung bình (mean-absolute-error: MAE),  
sai số tuyệt đối trung bình chuẩn hóa (normalized MAE: NMAE)