Đại Học Quốc Gia TP.HCM
Trường Đại Học Bách Khoa
Khoa Khoa Học & Kỹ Thuật Máy Tính

Vietnam National University – HCMC
Ho Chi Minh City University of Technology
Faculty of Computer Science & Engineering

Đề cương môn học

KHAI PHÁ DỮ LIỆU
(Data Mining)

Số tín chỉ	**3** (3.0.6)			MSMH	CO3029
Số tiết	Tổng: **45**	LT: **45**	TH:	TN:	BTL/TL: **x**
Môn ĐA, TT, LV					
Tỉ lệ đánh giá	BT: **5%**	TN:	KT: **15%**	BTL/TL: **40%**	Thi: **40%**
Hình thức đánh giá	- *Kiểm tra: trắc nghiệm + tự luận, 45-60 phút/bài* - *Thi: trắc nghiệm + tự luận, 120 phút* - *Bài tập lớn: tự luận + báo cáo theo nhóm, 30 phút/bài báo cáo* - *Bài tập: tự luận, 30 phút/bài*				
Môn tiên quyết					
Môn học trước	- Hệ cơ sở dữ liệu				CO2013
Môn song hành					
CTĐT ngành	Khoa Học Máy Tính và Kỹ Thuật Máy Tính				
Trình độ đào tạo	Đại học				
Cấp độ môn học	**3** *[Có thể dạy vào năm 3-4]*				
Ghi chú khác	*3 tiết/buổi, tổ chức trình bày nhóm về đề tài bài tập lớn từ tuần 11 đến tuần 15.*				

1. Mô tả môn học (Course Description)

Môn học này nhằm giới thiệu quá trình khám phá tri thức, các khái niệm, công nghệ, và ứng dụng của khai phá dữ liệu. Ngoài ra, môn học này cũng trình bày các vấn đề tiền xử lý dữ liệu, các tác vụ khai phá dữ liệu, các giải thuật và công cụ khai phá dữ liệu mà có thể được dùng hỗ trợ nhà phân tích dữ liệu và nhà phát triển ứng dụng khai phá dữ liệu. Các chủ đề cụ thể của môn học bao gồm: tổng quan về khai phá dữ liệu, các vấn đề về dữ liệu được khai phá, các vấn đề tiền xử lý dữ liệu, hồi qui dữ liệu, phân loại dữ liệu, gom cụm dữ liệu, khai phá luật kết hợp, phát triển ứng dụng khai phá dữ liệu, và các đề tài nghiên cứu nâng cao trong khai phá dữ liệu.

Course Description:

This course aims to introduce the knowledge discovery process as well as concepts, technologies, and applications of data mining. It is also to discuss data preprocessing issues, data mining tasks, algorithms and tools that can be used to support data analysts and data mining application developers. In particular, its major topics are an overall view about data mining, issues related to data which are going to be mined, data preprocessing issues, data regression, data classification, data clustering, association rules mining, data mining application development, and other research topics of interest in the data mining area.

2. Tài liệu học tập

Sách, Giáo trình chính:

[1] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques", Third Edition, Morgan Kaufmann Publishers, 2012.
[2] David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.

Sách tham khảo:

[3] David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.
[4] Graham J. Williams, Simeon J. Simoff, "Data Mining: Theory, Methodology, Techniques, and Applications", Springer-Verlag, 2006.
[5] ZhaoHui Tang, Jamie MacLennan, "Data Mining with SQL Server 2005", Wiley Publishing, 2005.
[6] Oracle, "Data Mining Concepts", B28129-01, 2008.
[7] Oracle, "Data Mining Application Developer's Guide", B28131-01, 2008.
[8] Ian H.Witten, Frank Eibe, Mark A. Hall, "Data mining: practical machine learning tools and techniques", Third Edition, Elsevier Inc, 2011.
[9] Florent Messeglia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.
[10] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", Second Edition, Springer Science + Business Media, LLC 2005, 2010.

3. Mục tiêu môn học (Course Goals)

Sau khi học đạt môn học này, sinh viên có thể:

- Minh họa được các bước trong quá trình khám phá tri thức
- Mô tả các khái niệm cơ bản, công nghệ và ứng dụng của khai phá dữ liệu
- Giải thích các tác vụ khai phá dữ liệu phổ biến như hồi qui, phân loại, gom cụm, và khai phá luật kết hợp
- Nhận dạng được các vấn đề về dữ liệu trong giai đoạn tiền xử lý cho các tác vụ khai phá dữ liệu
- Sử dụng các giải thuật và công cụ khai phá dữ liệu để phát triển ứng dụng khai phá dữ liệu

Course Goals:

Upon successful completion, students will be able to:

- Demonstrate the steps in the overall knowledge discovery process
- Describe basic concepts, technologies, and applications of data mining
- Explain popular data mining tasks including regression, classification, clustering, and frequent itemset and association rules mining
- Identify data related issues in the data preprocessing phase for data mining tasks
- Use data mining algorithms and tools for data mining application development

4. Chuẩn đầu ra môn học (Course Outcomes)

STT	Chuẩn đầu ra môn học	CDIO
L.O.1	Minh họa được các bước trong quá trình khám phá tri thức	
	L.O.1.1 – So sánh quá trình khám phá tri thức và quá trình khai phá dữ liệu L.O.1.2 - Liệt kê được các bước trong quá trình khám phá tri thức L.O.1.3 – Nêu ví dụ thực tế về quá trình khám phá tri thức	

L.O.2	Mô tả các khái niệm cơ bản, công nghệ và ứng dụng của khai phá dữ liệu	
	L.O.2.1 – Liệt kê các tác vụ khai phá dữ liệu L.O.2.2 – Mô tả được các thành phần của tác vụ khai phá dữ liệu tổng quát L.O.2.3 – Mô tả được các thành phần của giải thuật khai phá dữ liệu tổng quát L.O.2.4 – Mô tả được quy trình khai phá dữ liệu chuẩn L.O.2.5 – Liệt kê được các ứng dụng của khai phá dữ liệu trong ít nhất 1 lĩnh vực thực tế L.O.2.6 – Phân biệt được hệ thống khai phá dữ liệu với các dạng hệ thống khác như hệ cơ sở dữ liệu diễn dịch, hệ thống truy hồi thông tin, hệ thống học máy, …	
L.O.3	Giải thích các tác vụ khai phá dữ liệu phổ biến như hồi qui, phân loại, gom cụm, và khai phá tập mẫu thường xuyên và luật kết hợp	
	L.O.3.1 – Giải thích tác vụ hồi qui dữ liệu L.O.3.2 – Giải thích tác vụ phân loại dữ liệu L.O.3.3 – Giải thích tác vụ gom cụm dữ liệu L.O.3.4 – Giải thích tác vụ khai phá tập mẫu thường xuyên và luật kết hợp	
L.O.4	Nhận dạng được các vấn đề về dữ liệu trong giai đoạn tiền xử lý cho các tác vụ khai phá dữ liệu	
	L.O.4.1 - Xác định được các mô tả thống kê của tập dữ liệu cho trước L.O.4.2 – Mô tả được vấn đề và giải pháp nhận diện nhiễu và phần tử ngoại biên trong tập dữ liệu cho trước L.O.4.3 – Thực hiện được các biến đổi dữ liệu trên tập dữ liệu cho trước L.O.4.4 – Thực hiện được các thu giảm dữ liệu trên tập dữ liệu cho trước	
L.O.5	Sử dụng các giải thuật và công cụ khai phá dữ liệu để phát triển ứng dụng khai phá dữ liệu	
	L.O.5.1 – Khai phá được mô hình hồi qui dữ liệu/mô hình phân loại dữ liệu/mô hình gom cụm dữ liệu/tập mẫu thường xuyên và luật kết hợp tương ứng trong ứng dụng khai phá dữ liệu L.O.5.3 – Sử dụng được thư viện khai phá dữ liệu trong ứng dụng khai phá dữ liệu L.O.5.3 – Minh họa được việc sử dụng kết quả khai phá dữ liệu trong một chương trình ứng dụng cụ thể	

Course outcomes:

No.	Course outcomes	CDIO
L.O.1	Demonstrate the steps in the overall knowledge discovery process	
	L.O.1.1 – Compare a knowledge discovery process with a data mining process L.O.1.2 – List the steps of a knowledge discovery process L.O.1.3 – Give a practical example of a knowledge discovery process	
L.O.2	Describe basic concepts, technologies, and applications of data mining	
	L.O.2.1 – List data mining tasks L.O.2.2 – Describe each component of a data mining task in general L.O.2.3 – Describe each component of a data mining algorithm in general	

	L.O.2.4 – Describe a standardized data mining process L.O.2.5 – List data mining applications in at least one application domain L.O.2.6 – Determine the differences between a data mining system and other systems such as deductive database systems, information retrieval systems, machine learning systems, and so on	
L.O.3	Explain popular data mining tasks including regression, classification, clustering, and frequent itemset and association rules mining	
	L.O.3.1 – Explain data regression L.O.3.2 – Explain data classification L.O.3.3 – Explain data clustering L.O.3.4 – Explain frequent itemset and association rules mining	
L.O.4	Identify data related issues in the data preprocessing phase for data mining tasks	
	L.O.4.1 – Determine statistical descriptives of a given data set L.O.4.2 – Describe noise and outlier detection problems and solutions of a given data set L.O.4.3 – Conduct data transformation on a given data set L.O.4.4 – Conduct data reduction on a given data set	
L.O.5	Use data mining algorithms and tools for data mining application development	
	L.O.5.1 – Build data regression models/data classification models/data clustering models/frequent itemsets and association rules in data mining applications L.O.5.2 – Utilize data mining libraries for data mining application development L.O.5.3 – Demonstrate using mining models/patterns in a particular data mining application	

5. Hướng dẫn cách học - chi tiết cách đánh giá môn học

Để đáp ứng mục tiêu của môn học, sinh viên cần thực hiện tốt các đòi hỏi sau đây:

a. *Có mặt tại lớp phải hơn 75% từ tuần 1 đến tuần 8 và 100% từ tuần 9 đến tuần 15.*
b. *Sau mỗi chương, sinh viên làm các bài tập của chương.*

Về đánh giá, có tất cả 4 cột điểm:

a. *Bài tập: : 5%*
b. *Bài tập lớn : 40%*
c. *Kiểm tra : 15%*
d. *Thi cuối kỳ : 40%*

Hình thức làm bài như sau:

a. *Bài tập: tự luận, có thể được thực hiện trên lớp và/hoặc về nhà sau mỗi chương*
b. *Bài tập lớn: được thực hiện theo nhóm ngoài lớp từ tuần 2 đến tuần 10, báo cáo tự luận và trình bày nhóm trên lớp từ tuần 11 đến tuần 15.*
c. *Kiểm tra: trắc nghiệm + tự luận, được thực hiện trên lớp vào tuần thứ 9, thời gian làm bài 45 phút-60 phút.*
d. *Thi cuối kỳ: trắc nghiệm + tự luận, được thực hiện theo lịch thi cuối kỳ, thời gian làm bài 120 phút.*

6. Dự kiến danh sách Cán bộ tham gia giảng dạy

- TS. Võ Thị Ngọc Châu - Khoa: Khoa học & Kỹ Thuật Máy Tính
- TS. Trần Minh Quang - Khoa: Khoa học & Kỹ Thuật Máy Tính
- Th.S. Dương Ngọc Hiếu - Khoa: Khoa học & Kỹ Thuật Máy Tính
- Th.S. Huỳnh Văn Quốc Phương - Khoa: Khoa học & Kỹ Thuật Máy Tính
- Th.S. Trương Quang Hải - Khoa: Khoa học & Kỹ Thuật Máy Tính

7. Nội dung chi tiết

Tuần / Chương	Nội dung	Chuẩn đầu ra chi tiết	Hoạt động dạy và học	Hoạt động đánh giá
1	**Chương 1: Tổng quan về khai phá dữ liệu** 1. Quá trình khám phá tri thức 2. Các khái niệm 3. Ý nghĩa và vai trò của khai phá dữ liệu 4. Ứng dụng của khai phá dữ liệu 5. Tóm tắt Yêu cầu tự học đ/v sinh viên (6giờ)	L.O.1.1 – So sánh quá trình khám phá tri thức và quá trình khai phá dữ liệu	- Giảng lý thuyết - Câu hỏi trên lớp theo cá nhân	Kiểm tra giữa kỳ và thi cuối kỳ
		L.O.1.2 - Liệt kê được các bước trong quá trình khám phá tri thức	- Giảng lý thuyết - Câu hỏi trên lớp theo cá nhân	Bài tập lớn
		L.O.1.3 – Nêu ví dụ thực tế về quá trình khám phá tri thức	- Giảng lý thuyết - Câu hỏi trên lớp theo cá nhân	Kiểm tra giữa kỳ
		L.O.2.1 – Liệt kê các tác vụ khai phá dữ liệu	- Giảng lý thuyết - Câu hỏi trên lớp theo cá nhân	Bài tập
		L.O.2.2 – Mô tả được các thành phần của tác vụ khai phá dữ liệu tổng quát	- Giảng lý thuyết - Câu hỏi trên lớp theo cá nhân	Thi cuối kỳ
		L.O.2.3 – Mô tả được các thành phần của giải thuật khai phá dữ liệu tổng quát	- Giảng lý thuyết - Câu hỏi trên lớp theo cá nhân	Thi cuối kỳ
2	**Chương 2: Các vấn đề tiền xử lý dữ liệu** 3.1. Tổng quan về giai đoạn tiền xử lý dữ liệu 3.2. Tóm tắt mô tả về dữ liệu 3.3. Làm sạch dữ liệu 3.4. Tích hợp dữ liệu 3.5. Biến đổi dữ liệu 3.6. Thu giảm dữ liệu 3.7. Rời rạc hóa dữ liệu	L.O.4.1 – Determine statistical descriptives of a given data set	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Kiểm tra giữa kỳ
		L.O.4.2 – Describe noise and outlier detection problems and solutions of a given data set	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Kiểm tra giữa kỳ
		L.O.4.3 – Conduct data transformation on a given data set	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Bài tập lớn và thi cuối kỳ
		L.O.4.4 – Conduct data reduction on a given data set	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Bài tập lớn

	3.8. Tạo cây phân cấp ý niệm 3.9. Biểu diễn dữ liệu 3.10. Tóm tắt Yêu cầu tự học đ/v sinh viên (6 giờ)			
3	**Chương 3: Hồi qui dữ liệu** 3.1. Tổng quan về hồi qui 3.2. Hồi qui tuyến tính 3.3. Hồi qui phi tuyến 3.4. Ứng dụng 3.5. Các vấn đề với hồi qui 3.6. Tóm tắt Yêu cầu tự học đ/v sinh viên (6 giờ)	L.O.3.1 – Explain data regression	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Kiểm tra giữa kỳ và thi cuối kỳ
		L.O.5.1 – Build data regression models/data classification models/data clustering models/frequent itemsets and association rules in data mining applications	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Bài tập và bài tập lớn
4, 5	**Chương 4: Phân loại dữ liệu** 4.1. Tổng quan về phân loại dữ liệu 4.2. Phân loại dữ liệu với cây quyết định 4.3. Phân loại dữ liệu với mạng Bayesian 4.4. Phân loại dữ liệu với mạng Neural 4.5. Các phương pháp phân loại dữ liệu khác 4.6. Tóm tắt Yêu cầu tự học đ/v sinh viên (12 giờ)	L.O.3.2 – Explain data classification	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Kiểm tra giữa kỳ và thi cuối kỳ
		L.O.5.1 – Build data regression models/data classification models/data clustering models/frequent itemsets and association rules in data mining applications	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Bài tập và bài tập lớn
6, 7	**Chương 5: Gom cụm dữ liệu** 5.1. Tổng quan về gom cụm dữ liệu 5.2. Gom cụm dữ liệu bằng phân hoạch 5.3. Gom cụm dữ liệu bằng phân cấp 5.4. Gom cụm dữ liệu dựa trên mật độ 5.5. Gom cụm dữ liệu dựa trên mô hình 5.6. Các phương	L.O.3.3 – Explain data clustering	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Kiểm tra giữa kỳ và thi cuối kỳ
		L.O.5.1 – Build data regression models/data classification models/data clustering models/frequent itemsets and association rules in data mining applications	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Bài tập và bài tập lớn

	pháp gom cụm dữ liệu khác 5.7. Tóm tắt Yêu cầu tự học đ/v sinh viên (12 giờ)			
8, 9	**Chương 6: Khai phá luật kết hợp** 6.1. Tổng quan về khai phá luật kết hợp 6.2. Biểu diễn luật kết hợp 6.3. Khám phá các mẫu thường xuyên 6.4. Khám phá các luật kết hợp từ các mẫu thường xuyên 6.5. Khám phá các luật kết hợp dựa trên ràng buộc 6.6. Phân tích tương quan 6.7. Tóm tắt Yêu cầu tự học đ/v sinh viên (12 giờ)	L.O.3.4 – Explain frequent itemset and association rules mining	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Kiểm tra giữa kỳ và thi cuối kỳ
		L.O.5.1 – Build data regression models/data classification models/data clustering models/frequent itemsets and association rules in data mining applications	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Bài tập và bài tập lớn
10	**Chương 7: Phát triển ứng dụng khai phá dữ liệu** 7.1. Tổng quan về vấn đề phát triển ứng dụng khai phá dữ liệu 7.2. Qui trình phát triển ứng dụng khai phá dữ liệu 7.3. Các chuẩn dành cho khai phá dữ liệu 7.4. Các công cụ hỗ trợ phát triển ứng dụng khai phá dữ liệu 7.5. Tóm tắt Yêu cầu tự học đ/v sinh viên (6 giờ)	L.O.5.1 – Build data regression models/data classification models/data clustering models/frequent itemsets and association rules in data mining applications	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Bài tập lớn
		L.O.5.2 – Utilize data mining libraries for data mining application development	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Bài tập lớn
		L.O.5.3 – Demonstrate using mining models/patterns in a particular data mining application	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Bài tập lớn
11	**Chương 8: Các đề tài nghiên cứu trong khai phá dữ liệu** 8.1. Hướng dữ	L.O.2.1 – List data mining tasks	- Giảng lý thuyết - Câu hỏi trên lớp theo cá nhân/nhóm	Kiểm tra giữa kỳ và thi cuối kỳ
		L.O.2.5 – List data mining applications in	- Giảng lý thuyết - Câu hỏi trên lớp theo cá	Bài tập

	liệu 8.2. Hướng kỹ thuật 8.3. Hướng ứng dụng 8.4. Tóm tắt Yêu cầu tự học đ/v sinh viên (6 giờ)	at least one application domain	nhân/nhóm	
		L.O.2.6 – Determine the differences between a data mining system and other systems such as deductive database systems, information retrieval systems, machine learning systems, and so on	- Giảng lý thuyết - Câu hỏi/bài tập trên lớp theo cá nhân/nhóm	Kiểm tra giữa kỳ và thi cuối kỳ
12-15	**Báo cáo nhóm của sinh viên về các đề tài bài tập lớn của môn học** Yêu cầu tự học đ/v sinh viên (24 giờ)	L.O.5.1 – Build data regression models/data classification models/data clustering models/frequent itemsets and association rules in data mining applications	- Bài tập lớn theo nhóm - Báo cáo nhóm trên lớp	Bài tập lớn
		L.O.5.2 – Utilize data mining libraries for data mining application development	- Bài tập lớn theo nhóm - Báo cáo nhóm trên lớp	Bài tập lớn
		L.O.5.3 – Demonstrate using mining models/patterns in a particular data mining application	- Bài tập lớn theo nhóm - Báo cáo nhóm trên lớp	Bài tập lớn

8. Thông tin liên hệ

Bộ môn/Khoa phụ trách	Hệ Thống Thông Tin/ Khoa học và Kỹ thuật Máy tính
Văn phòng	Nhà A3
Điện thoại	38647256 - Ext 5847
Giảng viên phụ trách	TS. Võ Thị Ngọc Châu – TS. Trần Minh Quang
Email	chauvtn@cse.hcmut.edu.vn, quangtran@cse.hcmut.edu.vn

Tp. Hồ Chí Minh, ngày 07 tháng 05 năm 2015

TRƯỞNG KHOA CHỦ NHIỆM BỘ MÔN CB PHỤ TRÁCH LẬP ĐỀ CƯƠNG