UBND TỈNH THÁI BÌNH
TRƯỜNG ĐẠI HỌC THÁI BÌNH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

ĐỀ CƯƠNG CHI TIẾT HỌC PHẦN

KHAI PHÁ DỮ LIỆU

Ngành đào tạo: Công nghệ thông tin
Bậc đào tạo: Đại học

(Ban hành kèm theo Quyết định Số: 640/QĐ-ĐHTB*, ngày 14/12/2019)*

1. Tên học phần: Khai phá dữ liệu - Mã học phần IT553079

2. Số tín chỉ: 3 (3,0)

3. Trình độ: Cho sinh viên năm thứ 3

4. Phân bổ thời gian

- Lên lớp:

Lý thuyết: 45 tiết (3 tiết lên lớp/tuần, 1 tiết = 50 phút)

Thực hành:

- Tự học: (45 x 2) = **90 giờ**

5. Điều kiện tiên quyết: Công nghệ phần mềm, Phân tích thiết kế hướng đối tượng, cơ sở dữ liệu.

6. Mục tiêu của học phần

6.1. Kiến thức

Sau khi hoàn tất môn học, sinh viên phải đạt được các yêu cầu sau:

- Hiểu 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 Mô hình và mẫu dữ liệu
- Nắm được các vấn đề về dữ liệu trong giai đoạn tiền xử lý cho các tác vụ khai phá dữ liệu. Dữ liệu và độ đo
- Tìm hiểu các bài toán khai phá dữ liệu phổ biến như hồi qui, phân loại, gom cụm, và khai phá luật kết hợp
- Sử dụng các giải thuật và công cụ khai phá dữ liệu để phát triển ứng dụng khai phá dữ liệu
- Được chuẩn bị về kiến thức để có thể nghiên cứu trong lĩnh vực khai phá dữ liệu.

6.2. Kỹ năng

- Khả năng hiểu ý nghĩa và vai trò của khai phá dữ liệu trong giải quyết các bài toán thực tế trong tình hình kinh tế- xã hội-khoa học-kỹ thuật ngày nay
- Khả năng nhận dạng và hiểu các vấn đề liên quan đến dữ liệu sẽ được khai phá và quá trình khai phá dữ liệu
- Khả năng ứng dụng của khai phá dữ liệu vào các hoạt động cụ thể của các đơn vị, tổ chức
- Khả năng phân tích và xử lý dữ liệu cho quá trình khai phá dữ liệu
- Khả năng phát triển các kỹ thuật khai phá dữ liệu
- Khả năng phát triển ứng dụng khai phá dữ liệu
- Khả năng vận dụng các tiện ích hỗ trợ khai phá dữ liệu được cung cấp phổ biến ngày nay như Weka, MS SQL Server….
- Khả năng tham gia phân tích và xử lý dữ liệu cho quá trình khai phá dữ liệu
- Khả năng tham gia phát triển các kỹ thuật khai phá dữ liệu
- Khả năng tham gia phát triển ứng dụng khai phá dữ liệu

6.3. Về năng lực tự chủ và tự chịu trách nhiệm:

- Có thái độ nghiêm túc trong học tập;
- Có đạo đức, lương tâm nghề nghiệp, có trách nhiệm với công việc, dám làm, dám chịu trách nhiệm.
- Có ý thức tổ chức kỷ luật, chủ động trong quá trình học tập.

7. Mô tả các nội dung học phần

Giới thiệu các kiến thức cơ bản về khai phá dữ liệu và quá trình khám phá tri thức, các giai đoạn chính của quá trình khai phá dữ liệu và khám phá tri thức. Học phần cũng cung cấp cho người học các bài toán chính (task) trong KPDL như phân lớp, phân cụm, hồi quy, chuỗi thời gian, luật kết hợp…cũng như cách sử dụng các công cụ hỗ trợ xây dựng các ứng dụng KPDL.

8. Nhiệm vụ của sinh viên

- Dự lớp: Sinh viên phải tham gia tối thiểu 80% số tiết học trên lớp.

- Có đầy đủ điểm thường xuyên, điểm đánh giá nhận thức, làm bài tập ở nhà theo yêu cầu của giảng viên.

- Có đủ 3 bài kiểm tra định kỳ.

- Tham gia dự kỳ thi kết thúc học phần.

- Nghiên cứu tài liệu trước khi lên lớp.

9. Tài liệu học tập

-Sách, giáo trình chính:

[1]. Giáo trình khai phá dữ liệu, Nguyễn Hà Nam, Nguyễn Trí Thành, Hà Quang Thụy, NXB ĐHQGHN, 2013. Chương: 1-6, 10.

-Tài liệu thamkhảo:

[2]. Nhập môn phát hiện tri thức và khai phá dữ liệu, Nguyễn Đức Thuần, NXB Thông tin và Truyền thông, 2013

[3] Ho Tu Bao, *Introduction to Knowledge Discovery and Data Mining*, National Center for Natural Science and Technology, 2002

[4]. Morgan Kaufman, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2002

10. Tiêu chuẩn đánh giá sinh viên

10.1. Tiêu chí đánh giá

STT	Điểm thành phần	Quy định	Trọng số	Ghi chú
1	Điểm thường xuyên, đánh giá nhận thức, thái độ thảo luận, chuyên cần, làm bài tập ở nhà.	- Số tiết dự học/Tổng số tiết: 10%. - Số bài tập đã làm/Tổng số bài tập được giao: 10%.	10%	
2	Điểm kiểm tra định kỳ 3 điểm kiểm tra viết 45’	- 3 bài kiểm tra thực hành 1 tiết trên lớp.	30%	
3	Thi kết thúc học phần	- Thi viết (90’)	60%	

10.2. Cách tính điểm

- Sinh viên không tham gia đủ 80% số tiết học trên lớp không được thi lần đầu.

- Điểm thành phần để điểm lẻ đến một chữ số thập phân.

- Điểm kết thúc học phần làm tròn đến phần nguyên.

11. Thang điểm: 10

12. Nội dung chi tiết học phần

Chương	Nội dung	LT	TH	KT
1	**Chương 1: Tổng quan về Khai phá dữ liệu**	9	0	0

Chương	Nội dung	LT	TH	KT
	1.1. Khái niệm Khai phá dữ liệu 1.2. Các ứng dụng của khai phá dữ liệu 1.3. Các bước của quá trình khai phá dữ liệu 1.4. Nhiệm vụ chính trong khai thác dữ liệu *1.4.1. Phân lớp (phân loại - classification)* *1.4.2. Hồi qui (regression)* *1.4.3. Phân nhóm (clustering)* *1.4.4. Tổng hợp (summarization)* *1.4.5. Mô hình hoá sự phụ thuộc* *1.4.6. Phát hiện sự biến đổi và độ* 1.5. Các phương pháp khai phá dữ liệu *1.5.1. Các thành phần của giải thuật KPDL* *1.5.2. Phương pháp suy diễn / quy nạp* *1.5.3. Phương pháp ứng dụng K-láng giềng gần* *1.5.4. Phương pháp sử dụng cây quyết định và luật* *1.5.5. Phương pháp phát hiện luật kết hợp* 1.6.Lợi thế của khai phá dữ liệu so với phương pháp cơ bản *1.6.1. Học máy (Machine Learning)* *1.6.2. Phương pháp hệ chuyên gia* *1.6.3. Phát kiến khoa học* *1.6.4. Phương pháp thống kê* 1.7. Lựa chọn phương pháp 1.8. Những thách thức trong ứng dụng và nghiên cứu trong kỹ thuật khai phá dữ liệu			

Chương	Nội dung	LT	TH	KT
	1.8.1. Các vấn đề về cơ sở dữ liệu *1.8.1. Một số vấn đề khác*			
2	**Chương 2: Tiền xử lý dữ liệu** 2.1. Mục đích 2.2. Làm sạch dữ liệu *2.2.1. Thiếu giá trị* *2.2.2. Dữ liệu nhiễu* *2.2.3. Bài tập áp dụng* 2.3. Tích hợp và biến đổi dữ liệu *2.3.1. Tích hợp dữ liệu* *2.3.2. Biến đổi dữ liệu* *2.3.3. Thu nhỏ dữ liệu* *2.3.4. Bài tập áp dụng*	6	0	0
3	**Chương 3: Khai phá luật kết hợp** 3.1. Khái niệm về luật kết hợp 3.2. Thuật toán Apriori + *Tư tưởng thuật toán* + *Mô tả giải thuật* + *Bài tập áp dụng* 3.3. Thuật toán FP-Growth + *Tư tưởng thuật toán* + *Mô tả giải thuật* + *Bài tập áp dụng* 3.4. So sánh và đánh giá 2 thuật toán + *So sánh* + *Đánh giá* 3.5. Kết luận	8	0	1
4	**Chương 4: Phân lớp và dự đoán** 4.1. Khái niệm cơ bản	11	0	1

Chương	Nội dung	LT	TH	KT
	4.1.1. Phân lớp *4.1.2. Dự đoán* 4.2.Phân lớp sử dụng cây quyết định *4.2.1. Khái niệm Cây quyết định* *4.2.2. Mô hình phân lớp cây quyết định* *4.2.3. Lựa chọn thuộc tính* *4.2.4 Bài tập áp dụng* 4.3. Phân lớp dựa trên xác suất có điều kiện. *4.3.1. Các khái niệm cơ bản về xác suất* *4.3.2. xác suất có điều kiện* *4.3.3. Phân lớp dựa trên Công thức Bayes* *4.3.4. Bài tập áp dụng* 4.4. Các phương pháp phân lớp khác.			
5	**Chương 5: Phân cụm dữ liệu** 5.1. Khái niệm và mục tiêu của phân cụm dữ liệu *5.1.1. Phân cụm dữ liệu là gì?* *5.1.2.. Các mục tiêu của phân cụm dữ liệu* 5.2. Những kỹ thuật phân cụm dữ liệu *5.2.1. Phương pháp phân cụm phân hoạch* *5.2.2. Phương pháp phân cụm phân cấp* *5.2.3. Phương pháp phân cụm dựa trên mật độ* *5.2.4. Phương pháp phân cụm dựa trên lưới* *5.2.5. Phương pháp phân cụm dựa trên mô hình* *5.2.6. Phương pháp phân cụm có dữ liệu ràng buộc* 5.3. Một số thuật toán phân cụm dữ liệu *5.3.1. Thuật toán k-mean* *5.3.2. Thuật toán PAM*	8	0	1

Chương	Nội dung	LT	TH	KT
	5.3.3. Thuật toán AGNES *5.3.4. Thuật toán DIANA*			

13. Hình thức và nội dung từng tuần:

HTTCDH	Nội dung	Thời gian (tiết)	Yêu cầu SV chuẩn bị và địa chỉ tư liệu	Ghi chú
Nội dung 1: *(Tuần 1)*				
Lý thuyết	**Chương 1: Tổng quan về Khai phá dữ liệu** 1.1. Khái niệm Khai phá dữ liệu 1.2. Các ứng dụng của khai phá dữ liệu 1.3. Các bước của quá trình khai phá dữ liệu 1.4. Nhiệm vụ chính trong khai thác dữ liệu *1.4.1. Phân lớp (phân loại - classification)* *1.4.2. Hồi qui (regression)* *1.4.3. Phân nhóm (clustering)* *1.4.4. Tổng hợp (summarization)* *1.4.5. Mô hình hoá sự phụ thuộc* *1.4.6. Phát hiện sự biến đổi và độ*	**3**	*- Chuẩn bị tài liệu giáo trình môn học* *- Nghiên cứu và đọc giáo trình*	
Nội dung 2: *(Tuần 2)*				
Lý thuyết	**Chương 1: Tổng quan về Khai phá dữ liệu (tiếp)** 1.5. Các phương pháp khai phá dữ liệu *1.5.1. Các thành phần của giải thuật KPDL* *1.5.2. Phương pháp suy diễn / quy nạp* *1.5.3. Phương pháp ứng dụng K-láng giềng gần* *1.5.4. Phương pháp sử dụng cây quyết định và luật* *1.5.5. Phương pháp phát hiện luật kết hợp*	**3**	*- Chuẩn bị tài liệu giáo trình môn học* *- Nghiên cứu và đọc giáo trình.*	

HTTCDH	Nội dung	Thời gian (tiết)	Yêu cầu SV chuẩn bị và địa chỉ tư liệu	Ghi chú
Nội dung 3: *(Tuần 3)*				
Lý thuyết	**Chương 1: Tổng quan về Khai phá dữ liệu (tiếp)** 1.6. Lợi thế của khai phá dữ liệu so với phương pháp cơ bản *1.6.1. Học máy (Machine Learning)* *1.6.2. Phương pháp hệ chuyên gia* *1.6.3. Phát kiến khoa học* *1.6.4. Phương pháp thống kê* 1.7. Lựa chọn phương pháp 1.8. Những thách thức trong ứng dụng và nghiên cứu trong kỹ thuật khai phá dữ liệu *1.8.1. Các vấn đề về cơ sở dữ liệu* *1.8.2. Một số vấn đề khác*	**3**	*- Chuẩn bị tài liệu giáo trình môn học* *- Nghiên cứu và đọc giáo trình.*	
Nội dung 4: *(Tuần 4)*				
Lý thuyết	**Chương 2: Tiền xử lý dữ liệu** 2.1. Mục đích 2.2. Làm sạch dữ liệu *2.2.1. Thiếu giá trị* *2.2.2. Dữ liệu nhiễu* *2.2.3. Bài tập áp dụng*	**3**	*- Chuẩn bị tài liệu giáo trình môn học*	
Nội dung 5: *(Tuần 5)*				
Lý thuyết	**Chương 2: Tiền xử lý dữ liệu (tiếp)** 2.3. Tích hợp và biến đổi dữ liệu *2.3.1. Tích hợp dữ liệu* *2.3.2. Biến đổi dữ liệu* *2.3.3. Thu nhỏ dữ liệu* *2.3.4. Bài tập áp dụng*	**3**	*- Chuẩn bị tài liệu giáo trình môn học* *- Nghiên cứu và đọc giáo trình.*	

HTTCDH	Nội dung	Thời gian (tiết)	Yêu cầu SV chuẩn bị và địa chỉ tư liệu	Ghi chú
Nội dung 6: *(Tuần 6)*				
Lý thuyết	**Chương 3: Khai phá luật kết hợp** 3.1. Khái niệm về luật kết hợp 3.2. Thuật toán Apriori *+ Tư tưởng thuật toán* *+ Mô tả giải thuật* *+ Bài tập áp dụng*	**3**	*Chuẩn bị bài tập thực hành chương 4.*	
Nội dung 7: *(Tuần 7)*				
Lý thuyết	**Chương 3: Khai phá luật kết hợp (tiếp)** 3.3. Thuật toán FP-Growth *+ Tư tưởng thuật toán* *+ Mô tả giải thuật* *+ Bài tập áp dụng*	**3**	*- Chuẩn bị tài liệu giáo trình môn học* *- Nghiên cứu và đọc giáo trình.*	
Nội dung 8: *(Tuần 8)*				
Lý thuyết	**Chương 3: Khai phá luật kết hợp (tiếp)** 3.4. So sánh và đánh giá 2 thuật toán *+ So sánh* *+ Đánh giá* 3.5. Kết luận	**2**	*Chuẩn bị bài tập thực hành chương 3 trên Weka.*	
Kiểm tra – Đánh giá	KT lý thuyết 1 tiết	**1**		
Nội dung 9: *(Tuần 9)*				
Lý thuyết	**Chương 4: Phân lớp và dự đoán** 4.1. Khái niệm cơ bản *4.1.1. Phân lớp* *4.1.2. Dự đoán* 4.2.Phân lớp sử dụng cây quyết định	**3**	*- Chuẩn bị tài liệu giáo trình môn học*	

HTTCDH	Nội dung	Thời gian (tiết)	Yêu cầu SV chuẩn bị và địa chỉ tư liệu	Ghi chú
	4.2.1. Khái niệm Cây quyết định *4.2.2. Mô hình phân lớp cây quyết định* *4.2.3. Lựa chọn thuộc tính* *4.2.4 Bài tập áp dụng*			
Nội dung 10: *(Tuần 10)*				
Lý thuyết	**Chương 4: Phân lớp và dự đoán (tiếp)** 4.3. Phân lớp dựa trên xác suất có điều kiện. *4.3.1. Các khái niệm cơ bản về xác suất* *4.3.2. xác suất có điều kiện* *4.3.3. Phân lớp dựa trên Công thức Bayes* *4.3.4. Bài tập áp dụng*	**3**	*- Chuẩn bị tài liệu giáo trình môn học*	
Nội dung 11: *(Tuần 11)*				
Lý thuyết	**Chương 4: Phân lớp dữ liệu (tiếp)** 4.4. Các phương pháp phân lớp khác: Naïve Bayes, láng giềng gần nhất + Ý tưởng + Thuật toán + Bài tập áp dụng	**2**	*- Chuẩn bị tài liệu giáo trình môn học* *- Nghiên cứu và đọc giáo trình.*	
Kiểm tra – Đánh giá	KT lý thuyết 1 tiết	**1**		
Nội dung 12: *(Tuần 12)*				
Lý thuyết	**Chương 5: Phân cụm dữ liệu** 5.1. Khái niệm và mục tiêu của phân cụm dữ liệu *5.1.1. Phân cụm dữ liệu là gì?*	**3**	*- Chuẩn bị tài liệu giáo trình môn học*	

HTTCDH	Nội dung	Thời gian (tiết)	Yêu cầu SV chuẩn bị và địa chỉ tư liệu	Ghi chú
	5.1.2. Các mục tiêu của phân cụm dữ liệu.		*- Nghiên cứu và đọc giáo trình.*	
Nội dung 13: *(Tuần 13)*				
Lý thuyết	**Chương 5: Phân cụm dữ liệu (tiếp)** 5.2. Những kỹ thuật phân cụm dữ liệu *5.2.1. Phương pháp phân cụm phân hoạch* *5.2.2. Phương pháp phân cụm phân cấp* *5.2.3. Phương pháp phân cụm dựa trên mật độ* *5.2.4. Phương pháp phân cụm dựa trên lưới*	**3**	*- Chuẩn bị tài liệu giáo trình môn học*	
Nội dung 14: *(Tuần 14)*				
Lý thuyết	**Chương 5: Phân cụm dữ liệu (tiếp)** *5.2.5. Phương pháp phân cụm dựa trên mô hình* *5.2.6. Phương pháp phân cụm có dữ liệu ràng buộc* 5.3. Một số thuật toán phân cụm dữ liệu *5.3.1. Thuật toán k-mean* *5.3.2. Thuật toán PAM*	**3**	*- Chuẩn bị tài liệu giáo trình môn học* *- Nghiên cứu và đọc giáo trình.*	
Nội dung 15: *(Tuần 15)*				

HTTCDH	Nội dung	Thời gian (tiết)	Yêu cầu SV chuẩn bị và địa chỉ tư liệu	Ghi chú
Lý thuyết	**Chương 5: Phân cụm dữ liệu (tiếp)** *5.3.3. Thuật toán AGNES* *5.3.4. Thuật toán DIANA*	**1**	*- Chuẩn bị tài liệu giáo trình môn học*	
	Ôn tập – Hệ thống	**1**		
Kiểm tra – Đánh giá	- Kiểm tra đánh giá môn học	**1**		

TRƯỞNG KHOA **TRƯỞNG BỘ MÔN**

(Đã ký) *(Đã ký)*