Course syllabus 2301694 Special topics in applied mathematics

Data Mining research essentials

1.	**Course Number**	2301694
2.	**Course Credit**	3
3.	**Course Title**	Special topics in applied mathematics: Data Mining research essentials
4.	**Faculty of Science**	Department of Mathematics
5.	**Semester**	First
6.	**Academic Year**	2007
7.	**Instructors**	Assistant Professor Krung Sinapiromsaran, Ph. D., Tel:02-218-5225, krung.s@chula.ac.th
8.	**Condition**	
8.1.	**Prerequisite**	-
8.2.	Corequisite	-
8.3.	Concurrent	-
9.	**Status**	Elective
10.	**Curriculum**	Computational Science
11.	**Degree**	Master of Science

12. Hours/week

Lecture	2 hours/week
Lab	2 hours/week
Self-study	6 hours/week

13. Course Description

Database concept; SQL language; data preparation; statistics for data mining; knowledge representation: tables, trees, rules, instance-based, clusters; credibility and comparing data mining methods; the minimum description length principle.

14. Course Outline

14.1. Learning Objectives/Behavioral objectives: Student can

- create and design a database for given data analysis tasks
- write the SQL statements to request data from the DBMS
- apply the data preparation techniques such as feature (attributes) selections, discretization
- apply statistics to mine data
- describe and explain the use of given knowledge representation
- evaluate the data mining models via train-validation-test, cross-validation and other techniques
- compare the data mining models using different criteria

14.2. Learning Contents

Chapter 1: Database concepts and Data Manipulation Language	6 hours

- Database design Entity/Relationship Model
- Relational Databases
- SQL language
- Queries and Reports

Chapter 2: Data Preparation	9 hours

- Data Cleaning: Missing value, Noisy data
- Data Integration and Transformation
- Discretization and Concept Hierarchy generation
- AOI: Attribute-Oriented Induction

Chapter 3: Statistics	9 hours

- Descriptive Statistics
- Bivariate Statistics
- Multiple regression and correlation
- Principle Component Analysis

Chapter 4: Knowledge representation	9 hours

- Decision tables
- Decision trees
- Classification rules
- Association rules
- Instance-based representation
- Clusters

Chapter 5:Credibility and comparing data mining methods	12 hours

- Training, validation and testing
- Predicting performance
- Cross-validation
- Other estimates: Leave-one-out, Bootstrap
- Predicting probabilities: Quadratic loss function, Informational loss function

- Cost matrix: Lift charts, ROC curves, Recall-precision curves
- Cost curves
- The minimum description length principle

14.3. Method

Week	*Date*	*Detail*
1	4, 6 June 2007	Database design Entity/Relationship Model Relational Databases
2	11, 13 June 2007	SQL language Queries and Reports
3	18, 20 June 2007	Data Cleaning: Missing value, Noisy data Data Integration and Transformation
4	25, 27 June 2007	Discretization and Concept Hierarchy generation
5	2, 4 July 2007	AOI: Attribute-Oriented Induction
6	9, 11 July 2007	Descriptive Statistics and Bivariate Statistics
7	16, 18 July 2007	Multiple regression and correlation
8	23 July 2007	Midterm exam on Monday 23 July 2007, 8:30 – 11:30 am.
9	1 August 2007	Principle Component Analysis
10	6, 8 August 2007	Decision tables and Decision trees
11	15 August 2007	Classification rules and Association rules
12	20, 22 August 2007	Instance-based representation and Clusters
13	27, 29 August 2007	Training, validation and testing Cross-validation Other estimates: Leave-one-out, Bootstrap
14	3, 5 September 2007	Predicting probabilities: Quadratic loss function, Informational loss function
15	10, 12 September 2007	Cost matrix: Lift charts, ROC curves, Recall-precision curves
16	17, 19 September 2007	Cost curves The minimum description length principle
17	24 September 2007	Final exam on Monday 24 September 2007, 8:30 – 11:30 am.

14.4. Media

Board, LCD projector, computer with Internet connection

14.5. Assignment through Network System

-

14.6. Evaluation

14.6.1. Assessment of academic knowledge

Midterm 50 points on 23 July 2007, 8:30 – 11:30 am.
Final 50 points on 24 September 2007, 8:30 – 11:30 am.

14.6.2. Assessment of work or classroom activities

-

14.6.3. Assessment of the assigned tasks

Project 10 points.

15. Reading List

15.1. Required Text

1. Ian H. Witten and Eibe Frank, DATA MINING: Practical Machine Learning Tools and Techniques, second edition, Morgan Kaufmann publishers, 2005.

15.2. Supplementary Texts

1. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann publishers, 2001.
2. Christopher J. Date, An Introduction to Database Systems, fifth edition, Addison-Wesley Publishing Company, 1990.

15.3. Research Articles/Academic Articles

Any related research articles or papers

15.4. Electronic Media or Websites

1. http://en.wikipedia.org/wiki/
2. http://www.kdnuggets.com/
3. http://www.autonlab.org/tutorials/

16. Teaching Evaluation

16.1. Teaching type

Lecture 4