SERM—Data Mining—2011

# Purpose

In this study unit, students will be introduced to basic techniques of data mining, including attribute-value, probabilistic, and relational mining techniques. Students will apply these techniques to large datasets in the various case studies.

# Goals

By the end of this unit students should be able to:

1. Describe what data mining is
2. Identify scenarios in which data mining is likely to be a useful problem solving tool
3. Explain the steps involved in data mining
4. Describe the technological components necessary to do data mining, i.e. hardware and software
5. Demonstrate the ability to use of at least one tool or set of tools to complete a data mining analysis

# Readings

The following text will be used.

Kamber, Micheline and Jiawei Han (2005) *Data Mining: Concepts and Techniques, 2nd Ed.* Morgan Kaufmann. ISBN: 1558609016

# Evaluation

The primary means of evaluating students’ understanding of data mining concepts and practice will be through the data mining analysis report and presentation that they will write and turn in at the end of this unit. In order to produce this report, students will have to:

1. Identify a data set
2. Clean the data and put it in a format suitable for use with a data mining software package
3. Select the variables for analysis and run appropriate mining algorithms
4. Identify interesting patterns and/or outliers that result from the mining operations, if any
5. Analyze these patterns and make a case interpreting what they mean

Data sets may come from the case studies being worked on simultaneously in other study units or from online sources. Successful completion of the data mining report will demonstrate that students have a firm grasp on the basic concepts and skills associated with mining knowledge from existing data.

As a secondary means of evaluation, students receive one short test given midway through the unit.

# Schedule

## Day 1: Data Rich, Information Poor—What is Data Mining and Why is it Useful?

“Data mining” has many different meanings and is a multi-disciplinary practice that has been influenced by many fields. This introductory day will bring out the underlying and unifying themes of KDD (Knowledge Discovery from Data) via a series of illustrative cases. The cases will be chosen in fields relevant to the course of study of the SERM participants, but will also be broad in focus to show the power and flexibility of the analytical methods that data mining represents. During the second half of this session, the steps in the data mining process will be introduced. If time permits, a simple demonstration using the software tools will be done together.

### Relevant Readings

Kamber & Han Chapters 1 & 11

### Assignment

In addition to reading, participants may be asked to write a short piece describing how data mining could be useful in their own chosen area or in their work.

## Day 2: Stage 1—Data Cleaning and Integration

The next several days of the unit will be spent working together through a case study using data from one of the other units of study in the SERM program. The beginning of this day’s study will be spent introducing the case, possibly with the participation of the instructor for the related unit. Students will be asked to help develop a “plan of attack” for the data analysis. Students will get familiar with the analysis tools and the rest of this day will be spent preparing the data for analysis.

### Relevant Readings

Kamber & Han Chapter 2

### Assignment

Students will need to spend time finishing data preparation, in the event that we don’t finish in class.

## Day 3: Stage 2—Data Selection and Transformation

This day will be spent in continued pre-processing of data in preparation for the mining step. If variable selection and data transformation is completed in sufficient time, then the data mining techniques to be used in the following day will be introduced.

### Relevant Readings

Kamber & Han Chapter 3

## Day 4: Stage 3—Data Mining

Although there are many different approaches to mining the data, the methods to be used for the case study will be chosen ahead of time and this day’s lecture will quickly focus on explaining these. Once introduced, the methods will be used right away to complete the data mining step on the data cleaned and pre-processed during the first two days. Hopefully the data set chosen will require that students return to the cleaning and transformation stages in response to the initial results of the mining step. This will highlight the iterative and often exploratory nature of the knowledge discovery process.

Also, the unit project will be introduced and assigned so that the students will have the weekend to work on it. The unit project will be an analysis similar to the group case study, to be completed either individually or perhaps in small groups.

### Relevant Readings

Kamber & Han Chapters 5 & 6

### Assignment

Catch up on readings, and begin working on the unit project. Test will be distributed this day, to be done independently and returned on Day 5.

## Day 5: Stage 4—Pattern Evaluation and Knowledge Presentation

This will be the last day of the group case study. We will spend time discussing the patterns that emerged from the data mining step, and potentially return to earlier stages, re-running the data mining to discover new knowledge. We will spend time as a group discussing what these patterns may mean and outlining our presentation of the results.

### Relevant Reading

Kamber & Han Chapter 7

### Assignment

Continue work on the unit project. Complete the reading.

## Day 6: Further Discussion of Analysis Methods

Now that we will have completed a complete cycle, we will spend a little more time looking at alternative methods of mining, as well as perhaps looking at other sources of data, particularly sources such as real-time data streams from data being gathered in the field, e.g. energy production from wind turbines. There will also be lab time for students to work on their unit projects and ask questions about them.

### Relevant Readings

Selections from Kamber & Han Chapters 8, 9, &10

### Assignment

Continue work on the unit project.

## Day 7: Lab and Discussion

This day will be essentially the same as Day 6.

## Day 8: Class Presentations

Students will present their projects and findings to the class. There will be time for Q&A after each presentation. Students will turn in their completed reports in class on this day.