EMTM 554: Data Mining Syllabus

Administrivia

- **Homework 1** is due at the first class. HW2 is due the second class, etc.
- All readings (except those in the textbook), supplemental readings, homeworks, and lecture slides are on Webcafe.
- Readings are also available on Study.net
- **Textbook**: Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*, 2nd edition, Morgan Kaufmann, 2006.
- **Software**: JMP
  - Please make sure you have a copy of JMP from the Statistics course. If you do not, please contact the EMTM office.
- There will be a Quiz on the last day of class.
- **Prerequisite**: the EMTM Statistics course.
- This page: http://www.cis.upenn.edu/~ungar/DBM/syllabus.html

Lecture 1: Overview of Data Mining

- **Topics**
  - What is Data Mining, and what is it used for?
  - Strategic marketing
    - CapitalOne Case study
  - Data Warehousing
    - OLTP, OLAP, ETL, etc.
  - [WebCafe](introDBM.ppt, strategy.ppt, DataWarehousing.ppt)
- **Required readings**
  - Text: Chapt 1 Introduction
  - *Data mining in context* from Mastering data mining (Berry and Linoff)
  - *Why master the art?* from Mastering data mining (Berry and Linoff)
  - *The Long Tail* (Chris Anderson)
  - Capital One (Clemens and Thatcher)
  - Text: Chapt 3 Data Warehouse and OLAP Technology
- **Supplemental readings**
  - Discovering Knowledge in Data - Chapt 1 (Larose)
- **Homework 1**

Lecture 2: Methods

- **Topics**
  - Visualization: PTDD
  - Personalization: collaborative filtering
  - Market segmentation: clustering
  - Prediction: Decision trees and regression methods
• JMP for Data Mining: The good, the bad, and the ugly
  • WebCafe: methods.ppt
• Required readings
  • Text: Chapter 6, sections 1,2,3,6,7,9,11 - Classification and Regression
  • Text: Chapter 7.4.1 k-means - clustering
• Supplemental readings
  • Information Visualization in Data Mining and Knowledge Discovery Chapt 2 (Color Plates are separate)
  • Discovering Knowledge in Data (Larose) chap 6 & 7 Decision trees, Neural networks
  • Homework 2

Lecture 3: Evaluation

• Topics
  • Evaluation: prediction and pitfalls
  • Correlation and causality
    • Gazelle.com case study
  • WebCafe: evaluation.ppt, gazelle.ppt
• Required readings
  • Text: Chapter 6, sections 12,13 - Accuracy and Error Measures
  • Homework 3

Lecture 4: The DBM process & software; CRM and other applications

• Topics
  • The DBM process, CRISP-DM
  • DBM Software and Industries, vertical and horizontal
  • WebCafe: process.ppt, tools.ppt, costing.ppt
• Required readings
  • Data Mining Methodology: The Virtuous Cycle Revisited from Mastering data Mining (Berry and Linoff)
• Supplemental readings
  • http://www.cis.upenn.edu/~ungar/DBM/software.html has pointers to software
  • Knowing What to Sell, When, and to Whom (Kumar, Venkatesan and Reinartz)
  • The Dark Side of Customer Analytics (Davenport and Harris)
  • Churn (Berry and Linhoff)
  • Homework 4

Lecture 5: Text Mining

• Topics
  • Intro to web search
  • Text mining: IR and IE, easy and hard
  • WebCafe: search.ppt, textmining.ppt
• Readings
Text: 10.4 Text Mining
- *Text Mining: Predictive Methods for Analyzing Unstructured Information* - Chapt 1 (Weiss et al.)

- Supplemental readings
  - *Text Mining Case Studies* - Chapt 7 from Text Mining (Weiss et al.)
  - Text: 10.5 web mining

- **Homework 5**

**Lecture 6: Social Networks, Course Summary**

- **Topics**
  - Course summary
  - Social network analysis
    - WebCafe: summary.ppt, socialNets.ppt

- **Readings**
  - *Social Networks* from The Economist special report Jan 28, 2010

- **Final Project**

[return home](http://www.cis.upenn.edu/~ungar/DBM/syllabus.html)

*ungar@cis.upenn.edu*