McCallum Graduate School of Business
Bentley University

Bentley University

McCallum Graduate School of Business
PhD Quant III Data Mining

Syllabus

INSTRUCTOR CONTACT INFORMATION
Dominique Haughton: dhaughton@bentley.edu

COURSE MEETING:
Tuesdays, 5:30 - 8:30 (approximately), in person in the ATC conference room or on Centra/Saba (remotely)

COURSE DESCRIPTION
This course will introduce participants to some of the most recent data mining techniques, with an emphasis on: 1. getting a general understanding of how the method works, 2. understanding how to perform the analysis using suitable available software, 3. understanding how to interpret the results in a business research context, and 4. developing the capacity to critically read published research articles which make use of the technique. Contents may vary according to the interest of participants.

LEARNING OBJECTIVES

- **Knowledge**: a working knowledge of recent data mining techniques, how to interpret them and when it is appropriate to use them
- **Skills**: the ability to use various data mining software tools to build models, and to write reports, presentations and expository papers based on the results
- **Perspectives**: an understanding of the role of data mining in business and society

HOW THE COURSE WILL BE TAUGHT

The course will be run as a seminar, with some hands-on work using the software packages mentioned below, and a large importance allocated to participants presenting their understanding of the readings in class.

Grading will rely on regular in-class presentations, weekly summaries and reports on software usage, and an expository paper or project report written by participants organized into teams and

presented at a public mini data mining conference to be held at the end of the course. For example, such an expository paper could cover recent literature on genetic algorithms applied to predictive modelling in database marketing, or review recent trends in web mining (anchoring the discussion on class readings but expanding it to other published work as well). In some other cases, the paper could describe the results of a real-life model building project.

GRADING/PERFORMANCE EVALUATION
Regular class presentations and class participation: 30%
Weekly summaries and reports on software usage: 30%
Final project and expository paper: 40%

ACADEMIC INTEGRITY STATEMENT
The Bentley Honor Code applies. See
http://www.bentley.edu/ugcatalogue/honesty/student_responsibilities_under_the_honor_code.cfm

STATEMENT REGARDING LEARNING DISABILITIES

LIST OF TOPICS (CAN VARY ACCORDING TO PARTICIPANTS' NEEDS)

- **Week 1:** Decision Trees I and SAS Enterprise Miner
  Typical problem: identify predictors of future good credit status for potential customers; decision trees seek to use predictors to subdivide the database into segments with mostly good credit or mostly bad credit customers.
  L chapter 6
  G chapter 10, Case: Customer Relationship Management
- **Week 2:** Decision Trees II
  K chapter 7
  G chapter 11, Case: Credit Scoring
- **Week 3:** Neural Nets I
  Typical problem: find an equation to describe salaries of baseball players in terms of a number of predictors on the quality of their playing; neural nets seek to emulate the process that goes on in the human brain in order to improve on traditional linear regression models.
  L chapter 7, K chapter 9
- **Week 4:** Neural Nets II; Self-organizing (Kohonen) maps I
  Typical problem: given 20 measures of living standards of a province in a country, represent the provinces on a two-dimensional grid so that provinces with higher living standards can be identified more easily; Kohonen maps make it possible to do this with the help of informative graphs.
  G chapter 12, Case: Forecasting TV audiences
  L chapter 9
- **Week 5:** Self-organizing (Kohonen) maps II
  Case: Kohonen maps of Vietnamese provinces

- **Week 6:** Genetic Algorithms I
  Typical problem: among a class of possible solutions, find the solution that will provide the best predictive model so as to optimize response to an offer in the most responsive decile of the database identified by the predictive model; genetic algorithms seek to mimic human evolutionary processes to obtain better solutions to difficult optimizing problems.
  K chapter 10
- **Week 7:** Genetic Algorithms II
  Genetic algorithms and database marketing: Genalytics white paper; *Evolutionary computation for database marketing*, by Bhattacharyya
- **Week 8:** Association (Market Basket) Analysis
  Typical problem: if customers buy cheese at a supermarket, are they likely to also buy crackers? If shoppers visit a particular page when e-shopping, are they more likely to then turn to the ordering page? Association analysis provides algorithms to help identify such associations.
  L chapter 10
  G chapter 7, Case: Market Basket Analysis
- **Week 9:** Beyond regression analysis: MARS (Multivariate Adaptive Regression Splines) models I
  Typical problem: when building a model for the monetary value of a customer, it can happen for example that as the age of the customer increases, the monetary value increases, but only up to a certain age. Beyond that age, the value decreases, or perhaps remains constant. MARS models help tease out these effects automatically, in the presence of a large number of predictors.
  Salford Systems walkabout; *Application of multiple adaptive regression splines (MARS) in direct response modelling*, by Joel Deichmann, Abdolreza Eshghi, Dominique Haughton, Selin Sayek, Nicholas Teebagy, 2002.
- **Week 10:** MARS models II; TreeNet and RandomForests models
  TreeNet and RandomForests are extensions of Decision Trees that seek to achieve better predictive power.
  Salford Systems walkabouts; MARS models and Lunar New Year expenditures in Vietnam, Haughton and Nguyen, 2009
- **Week 11:** Web Mining I
  Typical problem: can we predict which web pages are likely to be most relevant for a query submitted by a site visitor for whom we have past query and navigation data? Can we better understand the factors that drive how customers make purchases on the web?
  G chapter 8, Case: Web clickstream analysis
  G chapter 9, Case: Profiling website visitors
- **Week 12:** Web Mining II; Text Mining I

Typical problem: given free unstructured responses to a survey question, can we identify the main themes in the responses, and cluster the responses into some main categories?
BFS chapters 7 and 8; SAS Text Miner documentation
A Review of Two Text-Mining Packages: SAS TextMining andWordStat, by Angelique Davi, Dominique Haughton, Nada Nasr, Gaurav Shah, Maria Skaletsky, and Ruth Spack

- **Week 13:** Text Mining II
  An application of text mining to the imputation of missing key player description in a customer database, by M. Skaletsky and D. Haughton.
- **Week 14:** Social Networks
  Typical problem: if collaboration is defined between two researchers as having published at least one paper together, which factors tend to encourage stronger collaboration networks? Can we identify if a network is changing over time beyond just what one might expect by chance alone?
  Proactive Encouragement of Interdisciplinary Research Teams in a Business School Environment: Strategy and Results, by Adams, Carter, Hadlock, Haughton and Sirbu.
- **Week 15:** **Final Presentations**

READINGS AND OTHER LEARNING MATERIALS (SOFTWARE, ETC.)

Textbooks:

- **K: Data Mining: Concepts, models, methods, and algorithms**, by Mehmed Kantardzic, Wiley 2003, selected chapters (for decision trees and genetic algorithms)
- **L: Discovering Knowledge in Data**, by Daniel Larose, Wiley 2005, selected chapters (for decision trees and association analysis)
- **BFS: Modeling the Internet and the Web,** by P. Baldi, P. Frasconi and P. Smyth, Wiley 2003, selected chapters, for web and text mining
- **G: Applied Data Mining,** P. Giudici, Wiley 2003, selected cases.

Software tools:

- SAS and SAS Enterprise Miner and/or IBM SPSS Modeler
- R
- MARS (one-month free version)
- TreeNet (one-month free version)
- Random Forests (one-month free version)
- For Genetic Algorithms, SAS PROC GA, or R package GA
- Pajek and Gephi for social networks

Internet resources:

- Multiple Adaptive Regression Splines interactive walkabout: http://www.salford-systems.com/walkaboutmars1.php
- Treenet: http://www.salfordsystems.com/faq4TreeNet.php
- Genetic algorithms: http://cs.felk.cvut.cz/~xobitko/ga/

Selected additional research articles:

For Decision Trees

- Direct marketing modeling with CART and CHAID, By Dominique Haughton and Samer Oulabi, *Journal of Direct Marketing*, **7(3)**, 16-26, 1993
- A personalized recommender system based on web usage mining and decision tree induction, by Yoon Ho Cho, Jae Kyeong Kim and Soung Hie Kim, *Expert Systems with Applications*, **23(3)**, 329-342, 2002

For Neural Nets

- *Neural networks as statistical tools for business researchers*, by DeTienne et al., Organizational research methods, 2003

For MARS models

- *Forecasting recession, can we do better on MARS?*, by Sephton, Federal Reserve Bank of Saint Louis, 2001
- *Application of multiple adaptive regression splines (MARS) in direct response modelling*, by Joel Deichmann, Abdolreza Eshghi, Dominique Haughton, Selin Sayek, Nicholas Teebagy, 2002
- *A comparison of two non-parametric schemes, MARS and neural networks*, by De Veaux, Psichogios and Ungar, Computers in chemical engineering, 1993

For Genetic Algorithms

- *Evolutionary computation for database marketing*, by Bhattacharyya, Journal of database management, 2003
- *Using genetic algorithms to find technical trading rules*, by Allen and Karjalainen, Journal of financial economics, 2003
- *Targeting customers with statistical and data-mining techniques*, by Drew, Manni, Betz and Datta, Journal of service research, 2001