McCallum Graduate School of Business
Bentley University

Bentley University

McCallum Graduate School of Business
PhD Quant III/ MA 710 Data Mining

Syllabus

INSTRUCTOR CONTACT INFORMATION
Dominique Haughton: dhaughton@bentley.edu

COURSE MEETING:
PhD Quant III: Mondays, 6:30 - 9:30 (approximately), in person in the ATC conference room or if necessary on Centra/Saba (remotely)
MA 710: Thursdays 7:30-9:50, in person in Smith 218 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. All participants need to be ready to present all readings. To alleviate the work load and to help with the difficulty of the material, we will arrange the class into teams; all teams will need to prepare a deck with their understanding of each reading and we will randomly decide which team presents which reading at the next class meeting. Details will be given at the first class meeting. Note that in the PhD class, each team will consist of one person, yielding 6 teams. I expect around 7 to 8 teams in the MS class.

It is of the upmost importance that participants be capable of presenting material in good English, to respond to questions and to participate actively in class discussions. Information on how to get help in English as a second language as needed is available at the end of this syllabus.

Grading will rely on regular in-class presentations, 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: 60%
Final project and expository paper: 40%

COURSE MATERIALS (TEXTBOOKS, SOFTWARE, ETC.)

Textbooks:

Note that needed chapters from the Kantardzic and Larose books below are arranged into a Wiley e-textbook, available on Vital Source at
https://www.vitalsource.com/products/data-mining-etext-for-bentley-university-daniel-t-larose-v9781119355670

- **K: Data Mining: Concepts, models, methods, and algorithms, second edition,** by Mehmed Kantardzic, Wiley 2011, selected chapters
- **L: Discovering Knowledge in Data, second edition,** by Larose and Larose, Wiley 2015, selected chapters

The following book is available at a 20% discount at
http://support.sas.com/publishing/discounts.html

- **SASEM: Predictive Modeling with SAS® Enterprise Miner™: Practical Solutions for Business Applications**, Second Edition, by Kattamuri S. Sarma, Ph.D.

Software tools:

- SAS and SAS Enterprise Miner
- Salford Systems evaluation version (MARS, Treenet and Random Forests)
- For Genetic Algorithms, Weka, SAS PROC GA, or R package GA
- Pajek and Gephi for social networks
- R (Rattle package)
- Matlab Kohonen toolbox (for self-organizing maps)

Internet resources:

- Salford Systems documentation at https://www.salford-systems.com/
- Genetic algorithms: http://www.obitko.com/tutorials/genetic-algorithms/

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

- **Week 1:** **Introductions and formation of teams**
  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 11; SASEM chapters 2 and 3
- **Week 2:** Decision Trees II
  K chapter 6; SASEM chapters 4
- **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 12; SASEM chapter 5
- **Week 4:** Neural Nets II; Model evaluation; Self-organizing (Kohonen) maps
  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.
  K chapter 7; L chapter 20; Case: Kohonen maps of Vietnamese provinces; SASEM chapter 7; L chapters 15 and 18
- **Week 5:** Ensemble models
  L chapters 25 and 26; K chapter 8
- **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.
L chapter 27; K chapter 13

- **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 23; K chapter 10
- **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 documentation; Case: Determinants of the international digital divide: an analysis using MARS, Journal of Global Information Technology Management, 9(4), 47-71 (Deichmann, Eshghi, Haughton, Masnaghetti, Teebagy, Sayek, Topi) (2006)
- **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 documentation; Case: MARS models and Lunar New Year expenditures in Vietnam; Case: Kickstarter rock music projects.
- **Week 11:** 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?
  SASEM chapter 9; K chapter 11
- **Week 12:** Text Mining II
  Case: An application of text mining to the imputation of missing key player description in a customer database, by M. Skaletsky and D. Haughton. Case: Movie analytics: text mining of movie reviews.
- **Week 13:** Social networks I
  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?

Pajek tutorial; Case: “Reciprocity in social networks - A case study In Tamil Nadu, India”, *Case Studies in Business, Industry and Government Statistics*, **5(2)**, 126-131 (Arumugam, Haughton, Vasanthi, Zhang) (2014)

- **Week 14:** Social Networks II
  Case: 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**

ACADEMIC INTEGRITY STATEMENT

Learning is a privilege that demands responsibility. At Bentley, students and faculty are members of an academic community that supports integrity both inside and outside the classroom. The expectation at Bentley is that students will take advantage of the opportunity for intellectual development and, in doing so, will conduct themselves in a manner consistent with the standards of academic integrity. When these standards are violated or compromised, individuals and the entire Bentley community suffer. Students who engage in acts of academic dishonesty not only face university censure but also may harm their future educational and employment opportunities. In other words, don’t bring unauthorized materials into exams, don’t plagiarize someone else’s work, and make sure that your collaborations are conducted in accordance with university and course policy.

All students have access to Bentley’s academic integrity policy on Blackboard (via the Academic Integrity course page) and the Undergraduate Student Handbook/Graduate Catalog. The best way to avoid a problem is to consult with your instructor before taking any action that might constitute a violation.”

ESOL CENTER

The ESOL Center offers writing and English language support to students who are English Speakers of Other Languages (ESOL). Our faculty tutors specialize in working with multilingual writers and can provide feedback and strategies on writing for all your course and career-related writing. You're welcome to come in for help at any stage of the writing process, from the brainstorming and organizing point through the final drafting stage. In addition, you can receive support related to source documentation, Power Point slide reviews, oral presentation practice, and pronunciation along with conversation fluency and enrichment.

The ESOL Center is located on the lower level of the Bentley Library, room 026. Day and evening appointments can be scheduled through https://bentleyesol.mywconline.net or by dropping by the ESOL Center to see if a faculty tutor is available. Because of the high demand for appointments, however, we encourage scheduling a time in advance whenever possible.

WRITING CENTER

The Writing Center offers one-on-one tutoring to students of all years and skill levels. Located on the lower level of the Bentley library (room 023), the Writing Center provides a welcoming and supportive environment in which students can work on writing from any class or discipline.

Writers are encouraged to visit at all stages of the writing process; they can come with a draft, an outline, or just some initial thoughts and questions.

Staffed by highly skilled student tutors, the Writing Center is open six days a week. Drop-ins are welcome, but appointments are encouraged and can be made online at bentley.mywconline.net or by phone at 781.891.3173. For hours and additional information, visit our website at bentley.edu/writing-center.