Université Cheikh Anta Diop
Master in Development Policy
Data Mining

Syllabus

INSTRUCTOR CONTACT INFORMATION

Dominique Haughton: dhaughton@bentley.edu

COURSE MEETING:

TBD, in person or on Centra (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: 40%
Weekly summaries and reports on software usage: 20%
Final project and expository paper: 40%

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:** Sampling
  Typical problem: how do we adjust statistical analyses when the data at hand comes from a survey and is not a simple random sample of the population? When should we and should we not use sampling weights?
  SL selected chapters
- **Week 4:** 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:** Bayesian Analysis I
  Typical problem: is it possible to codify prior information one may have at hand on the parameters of a model and incorporate this information into statistical models?
  Handouts
- **Week 7:** Bayesian Analysis II
  Handouts

Case: Bayesian analysis of poverty rates in Vietnam

- **Week 8:** Multilevel models and small area estimation
  Typical problem: given data on household living standards from a survey with a multilevel sampling design (communes within districts within provinces, for example), is it possible to use statistical models that take this multilevel structure into account to obtain better small area estimates of quantities of interest?
  MLWin Tutorial
  Case: Multilevel models and small area estimation in Vietnam (Dominique Haughton, Phong Nguyen, Irene Hudson and John Boland)
- **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 Davu, 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.
- **SL: Sampling: Design and Analysis, second edition** by Sharon Lohr, Brooks/Cole, Cengage Learning 2010.

Software tools:

- SAS and SAS Enterprise Miner; Weka and/or RapidMiner
- MARS (one-month free version)
- TreeNet (one-month free version)
- Random Forests (one-month free version)
- Pajek and SIENA for social networks
- Winbugs for Bayesian analysis
- MLWin for Multilevel models

Internet resources:

- Multiple Adaptive Regression Splines interactive walkabout: http://www.salford-systems.com/walkaboutmars1.php
- Treenet: http://www.salfordsystems.com/faq4TreeNet.php

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

For Bayesian Analysis

- Allenby, G., P. Bakken and P. Rossi, 2004. "How Bayesian methods have changed the face of marketing research", *Marketing Research*, Summer 2004.
- Adams, F., 2006. "Expert elicitation and Bayesian analysis of construction contract risks: an investigation", *Construction Management and Economics*, **24(1)**, 81-96.
- Haughton, D. and Nguyen, P., 2003. "Bayesian analysis of poverty rates: the case of Vietnamese provinces", *Journal of Modern Applied Statistical Methods*, **2(1)**, 189-194.