NEW YORK UNIVERSITY STERN
Leonard N. Stern School of Business

DATA MINING FOR BUSINESS INTELLIGENCE

C20.0057.01: Spring 2009

SYLLABUS

Professor	**Foster Provost, Information Systems Group**
Office, Hours	TBD
Email	fprovost@stern.nyu.edu **Begin subject: [DM class] …** **← note!**
Telephone	Office 212-99-80806, Fax: 212-99-54228
Course Webpage	Accessible from Blackboard
Classroom	TBD
Meeting times	TBD
First/Last Class	TBD
Final Exam	TBD
Course Assistants & Office Hours	Rong Zheng, KMEC 8-181, **rzheng@stern.nyu.edu**

1. Course Overview

This course will change the way you think about data and its role in business.

Businesses, governments, and individuals create massive collections of data as a by-product of their activity. Increasingly, decision-makers rely on intelligent technology to analyze data systematically to improve decision-making. In many cases automating decision-making processes is necessary because of the volume of data and the speed with which new data are generated. The course is suitable for those interested in working with and getting the most out of data, as well as those interested in understanding data mining from a strategic business perspective.

We will examine how data analysis technologies can be used to improve decision-making. We will study real-world examples and cases, from finance, marketing, operations, and electronic commerce, to place data-mining techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science. In addition, we will work "hands-on" with data mining software.

Specifically, the goal of this course is three-fold; after taking this course you should:

1. *Approach business problems data-analytically.* Think carefully & systematically about whether & how data can improve business performance.

2. *Be able to interact competently on the topic of data mining for business intelligence.* Know the basics of data mining processes, algorithms, & systems well enough to interact with CTOs, expert data miners, and business analysts. Be able to envision data-mining opportunities.

3. *Have had hands-on experience mining data.* Be prepared to follow up on ideas or opportunities that present themselves, e.g., by performing pilot studies.

2. Instruction Method

This is primarily a lecture-based course, but student participation is an essential part of the learning process in the form of active technical and case discussion. The course will explain with real-world examples the uses and some technical details of various data mining techniques. The emphasis primarily is on understanding the application of data mining techniques, and secondarily on the variety of techniques and the mechanics of how they work. Each class session has materials you must read prior to class. You should be prepared to be called on to discuss the readings. You are expected to attend every class session, and to arrive prior to the starting time.

Homework Assignments

There will be a total of seven assignments, each comprising questions to be answered and some including hands-on tasks. The last assignment will be done in your teams (discussed below). Completed assignments must be handed in *prior* to the start of the class on the due date. If submitted by email, they must arrive at least one hour prior to the start of class. They will be graded and returned promptly.

The hands-on tasks will be based on data that we will provide. You will mine the data to get hands-on experience in formulating problems and using the various techniques discussed in class. You will use these data to build and evaluate predictive models. The final assignment will include a "competition": one part of the data will be held back to evaluate the models you mine.

For the hands-on assignments you will use the (award-winning) toolkit Weka, part of the Pentaho open source business intelligence suite:

http://www.cs.waikato.ac.nz/ml/weka/
http://www.pentaho.com

Tutorials/demonstrations of Weka will be given in class. In order to use Weka you must have access to a computer on which you can install software (they don't let us install software on the machines in the computer labs). If you do not have such a computer, please see me immediately so we can make alternative arrangements. The first hands-on assignment will be very easy, ensuring that you can install the software and get it running, before moving on to more challenging assignments.

Term Project

A term project report will be prepared by student teams. Student teams should be of 2 or 3 people. *You should decide on your teams by the end of the third class, and submit them to me.*

Teams are encouraged to interact with the instructor and TA electronically or face-to-face in developing their project reports. You will submit a pre-proposal for your project around mid-term. Each team will present its project in the last class meeting. We will discuss the project requirements and presentations in class.

Final Exam

The final exam date is noted in this document's header and in the schedule of classes. The subject matter covered on the final will be discussed in class.

4. Requirements and Grading

You should attend all class sessions—the sessions build on previous discussions.

Answers to homework questions should be well thought out and communicated precisely. Points will be deducted for sloppy language and irrelevant discussion.

The points to be addressed in the term-project analysis will be discussed in class. The material needed for the term project will be handed out during the term. The analysis should be between 10 and 20 double-spaced pages.

The grade breakdown is as follows:

1. Homeworks (7): 25%
2. Term Project (1): 20%
3. Participation and Class Contribution: 10%
4. Midterm Quiz: 20%
5. Final Exam: 25%

Late Assignments

Assignments are due prior to the start of the lecture on the due date. Turn in your assignment early if there is any uncertainty about your ability to turn it in on the due date. Assignments up to 24 hours late will have their grade reduced by 25%; assignments up to one week late will have their grade reduced by 50%. After one week, late assignments will receive no credit.

5. Communication, Text, etc.

The Blackboard site for this course will contain lecture notes, reading materials, assignments, and late breaking news. It is accessible via: http://sternclasses.nyu.edu/

Post questions regarding course content to the Blackboard site (unless you are uncomfortable doing so) so that others can benefit from the answers. You are encouraged to contribute by answering/following up on posted questions.

Readings:

1. Textbook:

Data Mining Techniques, Second Edition
by Michael Berry and Gordon Linoff , Wiley, 2004
ISBN: 0-471-47064-3

- ***available from bookstore***
- ***available from Amazon (usually; may be less expensive)***
- ***available as ebook for free: http://site.ebrary.com/lib/nyulibrary***

2. Supplemental readings posted to blackboard or distributed in class. *Note that some of these readings are accessible for free only from an NYU computer. If you can't access a link from home, please try it from school.*

Please keep in mind the Stern Code of Conduct

http://w4.stern.nyu.edu/uc/currentstudents/codeofconduct.cfm?doc_id=5599

Class Schedule from Spring 2008
[Topics, structure & deliverables may change for 2009]

Class Number	Date	Module	Topics	Textbook Readings [Coursepack readings assigned in class]	Homework due dates
1	Tuesday January 22	Introduction	what is DM, why DM now, DM process, relation to other BI techniques, data mining tasks, linear regression revisited	• Ch. 1 & 2, • Ch. 4 pp. 116-120	
2	Thursday January 24				
3	Tuesday January 29				
4	Thursday January 31	Predictive Modeling	what is a model?, basic terminology, classification, tree induction, class-probability estimation	Ch. 6 pp. 165-194, 209	**HW#1 due**
5	Tuesday February 5				
6	Thursday February 7		evaluation, in-sample versus out-of-sample, overfitting, cross-validation, sanity checking toolkit demo	Ch. 3 pp. 43-54	
7	Tuesday February 12				**HW#2 due**
8	Thursday February 14		geometric interpretation, linear model versus tree induction, logistic regression		
9	Tuesday February 19		Bayesian & memory-based reasoning, nearest neighbors, variable normalization, text classification, "naïve" Bayes	Ch. 8 pp. 257-271	**HW#3 due**
10	Thursday February 21				
11	Tuesday February 26		knowledge-engineering bottleneck, rule-based systems, knowledge in action, evaluation	Ch. 3 pp. 54-86	
12	Thursday February 28		evaluation: error costs		
13	Tuesday March 4		neural networks	Ch. 7 pp. 211-243	**HW#4 due**
14	Thursday March 6				
15	Tuesday March 11		TBD		
16	Thursday March 13	MIDTERM QUIZ			
	March 18	SPRING BREAK			
	March 20				

Class Number	Date	Module	Topics	Readings	Homework due dates
17	Tuesday March 25	Descriptive/ Unsupervised Data Mining	descriptive data mining, unsupervised algorithms, associations, clustering	Ch. 9, Ch. 11	
18	Thursday March 27				**Project preproposal due**
19	Tuesday April 1	Combining Data Mining Techniques	data mining process in action, expected value revisited, clustering revisited	Ch. 4 pp. 87-110 (skip pp.90-93)	**HW#5 due**
20	Thursday April 3				
21	Tuesday April 8	Data Mining and Electronic Commerce	DM and competitive advantage, recommender systems, collaborative filtering, and other e-com applications	Ch 8 pp 282-285	
22	Thursday April 10				
23	Tuesday April 15	Targeted marketing case study	variable selection	revisit pp.60-64 & 233	**HW#6 due**
24	Thursday April 17	**Professor Vasant Dhar, formerly of Morgan Stanley**	genetic algorithms, data mining in finance	Ch. 13	
25	Tuesday April 22				
26	Thursday April 24	Ethics & Data Mining	what can/do firms know? what should they do?		
27	Tuesday April 29	wrapup and review	wrapup and review		**2 things due: project report, HW#7 (group)**
28	Thursday May 1	case presentations and competition results	case presentations and competition results		
Final Exam	Thursday May 8 8:00am-9:50	FINAL EXAM			