NEW YORK UNIVERSITY STERN
Leonard N. Stern School of Business

# DATA MINING FOR BUSINESS INTELLIGENCE

# B20.3336.31: Spring 2011

## SYLLABUS <draft Oct. 2010>

| | |
|---|---|
| Professor | **Claudia Perlich,** Information, Operations & Management Sciences Department |
| Office; Hours | TBD |
| Email | cperlich@stern.nyu.edu<br>**Begin subject: [DM GRAD] …** **← note!** |
| Telephone | Office 212-99-80806, Fax: 212-99-54228 |
| Course Webpage | Accessible from Blackboard |
| Classroom | TBD |
| Meeting time | Tuesdays, 6pm-9pm |
| First/Last Class | Tues Feb 8th / Tues May 3rd |
| Final Quiz | Take home after last class |
| Course Assistant<br>CA Office Hours | TBD |

## 1. Course Overview

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

Businesses, governments, and individual nhs 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 analytical and decision-making processes is necessary because of the volume of data and the speed with which new data are generated.

We will examine how data mining technologies can be used to improve decision-making. We will study the fundamental principles and techniques of data mining, and we will examine real-world examples and cases 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.

After taking this course you should:

1. *Approach business problems data-analytically.* Think carefully & systematically about whether & how data can improve business performance, to make better-informed decisions for management, marketing, investment, etc.

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, consultants, etc. Envision 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. Focus and interaction

The course will explain through lectures and real-world examples the fundamental principles, uses, and some technical details of data mining techniques. The emphasis primarily is on understanding the business application of data mining techniques, and secondarily on the variety of techniques. We will discuss the mechanics of how the methods work as is necessary to understand the fundamental concepts and business application.

This is primarily a lecture-based course, but student participation is an essential part of the learning process in the form of active discussion. I will expect you to be prepared for class discussions by having satisfied yourself that you understand what we have done in the prior classes. You are expected to attend every class session, to arrive prior to the starting time, to remain for the entire class, and to follow basic classroom etiquette, including having all electronic devices turned off and put away for the duration of the class (this is Stern policy, see below) and refraining from chatting or doing other work or reading during class.

The Blackboard site for this course will contain lecture notes, reading materials, assignments, extra-class discussions, and late-breaking news. You should check the Blackboard site daily, and I will assume that you have read all announcements and class discussion. You should use the Blackboard discussion board to ask any questions you have about class material, and you should try to answer your classmates' questions. Your class participation grade will include your contributions to the discussion board. You will not be penalized for being wrong in trying to participate on the discussion board (or in class).

I will use Blackboard as a primary means of communication. It is your responsibility to check Blackboard (and your email) at least once a day during the week (M-F), and you will be expected to be aware of any announcements within 24 hours of the time the message was sent.

**I will check my email at least once a day during the week (M-F). I get a tremendous amount of email, and can not process it all daily. *Your email will get priority if you include the special tag in the subject header as indicated in the information block on the first page of this syllabus***. I sort/filter based on this tag in order to make sure to process class mail first. I do not guarantee to be able to process your email in a timely fashion if you do not include the special tag.

In general, we will follow Stern default policies unless I state otherwise. I will assume that you have read them and agree to abide by them:

**http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511**

## 3. Lecture Notes and Readings

- Lecture notes: For most classes I will hand out lecture notes, which will outline the primary material for the class. You will be expected to flesh these out with your own note taking, and to ask questions about any material in the notes that is unclear after our class discussion. Depending on the direction our class discussion takes, we may not cover all material in the notes. If the notes themselves are not adequate to explain a topic we skip, you should ask about it on the discussion board.

Other readings are intended to supplement the material we learn in class. They give alternative perspectives on and additional details about the topics we cover:

- 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.*

- Supplemental book (optional):
  Data Mining Techniques, Second Edition
  by Michael Berry and Gordon Linoff , Wiley, 2004
  ISBN: 0-471-47064-3
  - *available as ebook for free: http://site.ebrary.com/lib/nyulibrary*
  - *available from Amazon*

Many students find this book to be an excellent supplemental resource. In the class schedule below I suggest the most important sections to read to supplement each class module.

## 4. Requirements and Grading

You should attend all class sessions—the class sessions are the main source of content for this class and each class builds on previous discussions. Please arrive on time; late arrivals distract the entire class. Absences and tardiness will be reflected heavily in your class participation grade.

Answers to homework questions should be well thought out and communicated precisely, avoiding sloppy language, poor diagrams, and irrelevant discussion.

At NYU Stern we seek to teach challenging courses that allow students to demonstrate differential mastery of the subject matter. Assigning grades that reward excellence and reflect differences in performance is important to ensuring the integrity of our curriculum. In my experience, students generally become engaged with this course and do excellent or very good work, receiving As and Bs, and only one or two perform only adequately or below and receive C's or lower. Note that the actual distribution for this course and your own grade will depend upon how well each of you actually performs this particular semester.

**Information Sheet**
The last page of this syllabus is an information sheet. Please detach it, fill it out completely, and return it by the end of the first class.

**Homework Assignments**
There will be a total of seven assignments, each comprising questions to be answered and hands-on tasks. Except as explicitly noted otherwise, you are expected to complete your assignments on your own—without interacting with others.

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. Emailed assignments should be sent to the TA, cc'ing the instructor. Assignments 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, and to find patterns in the data.

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/ download the "book version" – see HW#1
http://www.pentaho.com

**IMPORTANT**: ***In order to use Weka you must have access to a computer on which you can install software. If you do not have such a computer, please see me immediately so we can make alternative arrangements.*** The first hands-on assignment is very easy, ensuring that you can install the software and get it running, before moving on to more challenging assignments.

A brief demonstration of Weka will be given in class. If you would like more help with Weka, please visit the course assistant during his office hours, or make an appointment with him. He may hold "lab sessions" if enough students are interested.

Generally the course assistant should be the first point of contact for questions about and issues with the homeworks.

**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.

**Term Project**
A term project report will be prepared by student teams. Student teams should comprise 4 students. *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 about half way through the course. Each team will present its project at the end of the semester. We will discuss the project requirements and presentations in class.

**Midterm & Final Quizzes**
The final quiz will be a take-home to be completed during the week following the last class. The midterm quiz format will be discussed in class. The subject matter covered and the exact dates will be discussed in class.

**Regrading**
If you feel that a calculation, factual, or judgment error has been made in the grading of an assignment or exam, please write a formal memo to me describing the error, within one week after the class date on which that assignment was returned. Include documentation (e.g., a photocopy of class notes). I will make a decision and get back to you as soon as I can. Please remember that grading any assignment requires the grader to make many judgments as to how well you have answered the question. Inevitably, some of these go "in your favor" and possibly some go against. In fairness to all students, the entire assignment or exam will be regraded.

*FOR STUDENTS WITH DISABILITIES*: If you have a qualified disability and will require academic accommodation during this course, please contact the Moses Center for Students with Disabilities (CSD, 998-4980) and provide me with a letter from them verifying your registration and outlining the accommodations they recommend. If you will need to take an exam at the CSD, you must submit a completed Exam Accommodations Form to them at least one week prior to the scheduled exam time to be guaranteed accommodation.

***Please read the policies for Stern courses***

***http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511***

***Please keep in mind the Stern Honor Code***

http://www.stern.nyu.edu/mba/studact/mjc/hc.html

# *Class Schedule* ***FROM SPRING 2010 – should be similar in spring 2011***

| Class Number | Date | Module | Topics | Book Sections (Optional) | Deliverables |
|---|---|---|---|---|---|
| 1 | Monday February 8 | **Introduction** | What is DM? Why DM now? DM process, relation to other BI techniques, different data mining tasks | Ch. 1 & 2 | **Info Sheet** |
| 2 | Monday February 22 | **Data Mining Fundamentals**: Predictive Modeling | How do I produce a focused segmentation? What is a model? basic terminology, predictive modeling, classification, regression, tree induction, class-probability estimation, brief toolkit demo | Ch. 4 pp. 116-120 Ch. 6 pp. 165-194, 209 | **HW#1 due**<br>**Try HW#2 hands-on before this class** |
| 3 | Monday March 1 | | How do I know my model is any good? evaluation, in-sample versus out-of-sample, overfitting, cross-validation, ROC analysis, expected value framework, domain knowledge validation, geometric interpretation, linear model versus tree induction, logistic regression | Ch. 3 | **HW#2 due**<br>**Group Lists due** |
| 4 | Monday March 8 | | Bayesian & memory-based reasoning, nearest neighbors, variable normalization, text classification, "Naïve" Bayes, spam filtering | Ch. 8 pp. 257-271 | **HW#3 due** |
| | March 15 | SPRING BREAK<br>**Note: for 2011:** Sat, Feb 19 - Mon, Feb 21 *(Presidents' Day weekend)*<br>Mon, Mar 14 - Sun, Mar 20 *(Spring Break)* | | | |
| 5 | Monday March 22 | **Data Mining Fundamentals**: Descriptive/ Unsupervised Data Mining | unsupervised/descriptive data mining, unsupervised algorithms, associations, clustering | Ch. 9 Ch. 11 | **HW#4 due by 11:59pm Wed March 24** |

| Class Number | Date | Module | Topics | Book Sections (Optional) | Deliverables |
|---|---|---|---|---|---|
| 6 | Monday March 29 | **Data Mining in Action: cases, applications, and practical insight** | Fraud, Customer Retention, Image Classification<br><br>variable selection, feature engineering, neural networks, social networks | Ch. 7 pp. 211-243 | **Project pre-proposal due** |
| | | | | | |
| 7 | Monday April 5 | | DM and On-line Advertising<br><br>ethics of data mining, privacy, what can/do firms know?, what <u>should</u> they do? | Ch. 4 pp. 87-110 (skip pp.90-93) | |
| 8 | Monday April 12 | | knowledge-engineering bottleneck, rule-based systems, knowledge in action<br><br>data mining for credit management<br>data mining process in action, expected value in action, clustering revisited | | **HW#5 due** |
| 9 | Monday April 19 | | **GUEST SPEAKER** | | **HW#6 due** |
| 10 | Monday April 26 | | DM and ecommerce (Amazon), recommender systems, collaborative filtering, DM and competitive advantage | Ch 8 pp 282-285 | |
| 11 | Monday May 3 | | wrapup & review | revisit pp.60-64 & 233 | **Project reports due** |
| 12 | Monday May 10 | | project presentations | | |
| **<u>Final Quiz</u>**: "Take Home" (on Blackboard) | | | | | |