

IDS 472: Statistics for Information Systems and Data Mining

Spring 2004, Call 58438, Monday and Wednesday 3:30-4:45 PM, Room DH118
Yair M. Babad, UH 2403, Phone 312-996-8094, Cell 847-809-0487, Fax 312-413-0385
e-mail: ybabad@uic.edu, URL: <http://www.uic.edu/~ybabad>
Office Hours Monday and Wednesday, 5:00-6:00 PM

Updated: 11/15/2003 21:36:02

COURSE OBJECTIVE & PHILOSOPHY

One of the most profound results of the information technology revolution is the explosion in data and information availability. Effective use of this information is for many organizations a critical need; this includes the operational use of the information, as well as its use for prediction, planning and control. This course is an intermediate-level course devoted to the latter task: discovering meaningful patterns in data. In essence, Data Mining (DM) is a user-centric, interactive process that leverages analytical and statistical technologies and computing power. It is widely used in business, to a large extent for Customer Relationship Management (CRM) and credit scoring, as well as for industrial quality assurance, market research and process control.

TEXTBOOK AND READING MATERIAL

The text used is Data Mining: Concepts, Models, Methods and Algorithms by Mehmed Kantardzic, Wiley, 2003, ISBN 0-471-22852-4 (K). Recommended text is Data Mining: A Tutorial-Based Primer by Richard J. Roiger & Michael W. Geatz, Addison Wesley, 2003, ISBN 0-201-74128-8 (RG), which comes with an Excel data mining software and several datasets that will be used in the class.

Other resources of interest are the following: Those of you interested in more advanced text with data mining pseudo-code algorithms, will find Data Mining: Introductory and Advanced Topics by Margaret H. Dunham, Prentice Hall, 2003, ISBN 0-13-088892-3 to be of interest. Very good business cases may be found in Applied Data Mining: Statistical Methods for Business and Industry by Paolo Giudici, Wiley, 2003, ISBN 0-470-84679-8, and in Mastering Data Mining: The Art and Science of Customer Relationship Management by Michael J. A. Berry and Gordon S. Linoff, Wiley, 2000, ISBN 0-471-33123-6.

We will not "read the text in class". Rather, certain issues will be emphasized, a discussion will be held, and your questions will be answered and discussed. You must read on your own and be familiar IN ADVANCE OF EACH CLASS with the assigned material as given in the schedule, and with the class notes available in my web page, and be ready to participate in the class discussions.

In my web page you will find PowerPoint presentations and other material that I will use or introduce in class. You are advised to print these presentations (probably with 3 or 6 slides per page, framed, in black and white printing format) prior to class, so that you can use them in class in lieu of notes. You are responsible for knowing the contents of these materials as well as the textbook material (and of course whatever is discussed in class).

COMMUNICATIONS & PREREQUISITES

A common theme in the IDS sequence of courses is the development of your communications skills and the use of available computer technology and common software tools. You are expected to be familiar with word-processing and spreadsheet tools, and submit your work using such tools. All communications will use electronic mail. The assignments and other course materials can be printed out from the World Wide Web, at my URL given above.

I maintain a web page for this class. To this end, get to my URL listed above, select this class (IDS 472 for Spring 2004), and you will find yourself in an "announcement file" for this course. This file includes references to related documents, such as this syllabus, homework, and PowerPoint presentation of class material, in addition to the latest announcements related to the class.

It is expected that you have completed all the prerequisites for admission to this class (IDS 371 – Business Statistics II, or two semesters of statistics). Alternatively, you have to get my approval to join the class. These prerequisites may be checked, and students lacking them may be dropped from this class. The course assumes that different students have different levels of understanding and background of the course's topics, yet we will present the topics at advanced level. Students with little familiarity of the material are expected to prepare themselves to fully understand the material and contribute to course work and discussions.

HOMEWORK, PROJECT, QUIZZES AND EXAMS

The course work will include homework, a team project, quizzes and exams. The assignment will be announced in class, as well as in the announcements file noted above. There will be 3 homework assignments, to be done individually; these will be submitted only electronically to my e-mail. Note that this prevents any submission of hand-written assignment solutions. Late submission of homework will not be accepted and will not be graded, unless cleared IN ADVANCE with me.

There will be a team-oriented data mining course project. The project will include data cleaning and scrubbing, data exploration, model building, and data analysis. The project, and its various segments, will be discussed also during classes. Each team will submit electronically to my e-mail a final project report. Detail of the project, its grading and requirements, and the size of the teams (probably 3-5 students per team) will be announced later in class.

There will be two mid-term exams, but (due to scheduling commitments) no final exam. Rather, each class session (except the first one) may include a brief open book quiz, which stress understanding of the required material. This system allows timely grade progress feedback, and motivates to prepare for each session (and thus increase the probability of quality participation and getting the most from the class sessions).

CLASS ATTENDANCE AND HONOR CODE

You are expected to attend all classes, and are responsible for all announcements made in class or in the announcement file. Makeup of quizzes or reports will be given only by approval **PRIOR** to the quiz or report, except for extreme circumstances. Punctuality is highly regarded; no student, if arriving late, will be given any extra time to complete a quiz, nor will makeup quizzes be offered.

The university's honor code will be adhered to. Cheating will result in an automatic failing grade for the problem, quiz, exam or project for all those participating in the cheating or copying, and may lead to a failing grade in the course for all those students who are deemed to have consciously contributed to the cheating.

GRADING

Grades will be based on the exams (20% of the final grade for each exam), a team project (30%), and on the quizzes and homework (30% - equally weighted, and dropping the worst one). Final grades will be assigned on a curve, and I will exercise my judgment as to the cut points, as well as to the grading of students who miss or come late to many of the classes.

Don't nitpick about the grading. Persons who complain will not be rewarded for it; those who have the decency not to complain would deserve the same break. A request to look at one problem leads to re-grading of the whole paper, which often leads to a lower grade.

No "extra credit" opportunities will be offered or assigned to specific individuals under any circumstances; all students' grades will be based on the same components - this is an equal opportunity course.

TENTATIVE & APPROXIMATE COURSE SCHEDULE

Note: Each of the tasks should reach my e-mail at most by midnight prior to the date specified for the "Task Due". To illustrate: HW A is due (according to the "Task Due" column) on February 11; thus, it should reach my e-mail by midnight of February 10.

Class	Date	Chapter Topic (in Text)	Task Due	Resource
1	Jan 12	Introduction to DM		K 1
2	Jan 14	Data Mining Example Using iDA – Credit Card Promotion		RG 4
	Jan 19	*** No class – King Birthday		
3	Jan 21	Overview of DM Techniques		RG 2-3
4	Jan 26	Overview of DM Techniques – cont.		
5	Jan 28	Preparing the Data		K 2
6	Feb 2	Data Reduction		K 3
7	Feb 4	Data Reduction – cont.		
8	Feb 9	Statistical Methods		K 5
9	Feb 11	Statistical Methods – cont.	HW A	
10	Feb 16	Statistical Methods – cont.		
11	Feb 18	Exam 1		
12	Feb 23	Cluster Analysis		K 6
13	Feb 25	Cluster Analysis – cont.		
14	Mar 1	Cluster Analysis – cont.		
15	Mar 3	Clementine – a DM Tool and Its Use for Data Exploration		(A)
16	Mar 8	The Use of Clementine and iDA for Clustering		
17	Mar 10	Decision Trees and Decision Rules	HW B	K 7
18	Mar 15	Decision Trees and Decision Rules – cont.		
19	Mar 17	Clementine Use for Decision Rules		
	Mar 22, 24	*** No class – Spring Break		
20	Mar 29	Exam 2		

Class	Date	Chapter Topic (in Text)	Task Due	Resource
21	Mar 31	Association Rules – Market Basket Analysis		K 8
	Apr 5	*** No class – Passover		
22	Apr 7	Association Rules – cont.		
23	Apr 12	Association Rules – cont.		
24	Apr 14	Neural Networks		K 9
25	Apr 19	Neural Networks – cont,	HW C, Project	
26	Apr 21	The CRISP-DM Data Mining Methodology		(B)
	Apr 26, 28	*** No class – conference commitments		

(A) Clementine User Manual will be available through the announcements file.

(B) The CRISP-DM Manual will be available through the announcements file.

Homework subject areas:

HW A: Data preparation and reduction

HW B: Clustering and statistical data mining

HW C: Association rules data mining