UNIVERSITY OF LOUISVILLE

COLLEGE OF BUSINESS

KNOWLEDGE DISCOVERY IN DATABASES
CIS 445-01-5360
Fall 2016

I. Professor / Instructor	
Instructor	Dr. J. Zurada
Contact information	Office: 306 College of Business Telephone: 502-852-4681 Fax: 502-852-4799 Email: jozef.zurada@louisville.edu
Office hours	Tuesdays and Thursdays 10:00a.m. to 10:55a.m., 3:00pm to 3:55pm; And by appointment

II. Course Information	
Class time / Room	Tuesdays and Thursdays 4:00p.m. to 5:15p.m. 003 College of Business
Required text	• CoursePack: *Knowledge Discovery in Databases* (available at Gray's Bookstore only). It contains chapters 1-5 from book 1 listed under References below; chapters 5, 6, 8, and 10 from book 2; and chapters 7, 11, and 13 from book 3. • Other materials, such as Book 4 and one or two chapters from Book 8, will be posted on Blackboard.
References	1. *Data Mining: Concepts, Models, Methods, and Algorithms*, M. Kantardzic, IEEE Press/Wiley, 2011. 2. *Seven Methods for Transforming Corporate Data into Business Intelligence*, V. Dhar, and R. Stein, Prentice Hall, 1997. 3. *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*, M.J.A. Berry & G.S. Linoff, John Wiley & Sons, 2004. 4. *Fuzzy Logic Toolbox*, User's Guide, MathWorks, 2013. 5. *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, G.J. Myatt & W.P. Johnson, 2nd edition, Wiley, 2014.

	6. *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods and Applications*, G.J. Myatt and W.P. Johnson, Wiley, 2009. 7. *Making Sense of Data III: A Practical Guide to Design Interactive data Visualizations Exploratory Data Analysis and Data Mining*, G.J. Myatt & W.P. Johnson, 2011. 8. *Data Mining: A Tutorial-Based Primer*, R.J. Roiger & M.W. Geatz, Addison-Wesley, 2003. 9. *Data Mining: Concepts and Techniques*, J. Han and M. Kamber, Morgan Kaufmann Publishers, 2001. 10. *Data Mining: Practical Learning Tools and Techniques*, I.H. Witten & E. Frank, Morgan Kaufmann, 2011. 11. *Principles of Data Mining*, D. Hand, H. Mannila, and P. Smith, The MIT Press, Cambridge, MA, 2001. 12. *Applied Data Mining: Statistical Methods for Business and Industry*, P. Giudici, Wiley, 2003. 13. *Discovering Knowledge in Data: An Introduction to Data Mining*, D.T. Larose, Wiley, 2005. 14. *Introduction to Business Data Mining*, D. Olson, Y. Shi, McGraw-Hill, 2007. 15. *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*, G. Shmueli, N.R. Patel, and P.C. Bruce, Wiley, 2010. • There are many books on knowledge discovery in databases (KDD) and data mining (DM) available on the market. However, many of them are either written for (a) business executives or business people and cover the material in a superficial and shallow manner or (b) computer scientists or graduate computer science students who have strong mathematical foundations (6-8 above). Other books (5-7 and 13-15 above) cover relevant data mining topics, but use Excel based software products such as Excel with XLMiner. However, references 5-7 deserve attention. There is no really a single good textbook that covers all the concepts listed in the course description below and tentative course outline at the level appropriate for senior undergraduate CIS students. References 1 through 3, from which the CoursePack was assembled, seem to be the best of all I have found. Reference 10

	not only covers data mining, but also describes software Weka.
Course description	• This course introduces the process, concepts, models, and methods of knowledge discovery in large volumes of data. It discusses how traditional statistical methods, SQL queries, and OLAP techniques can be complemented by data mining techniques such as neural networks, decision trees, genetic algorithms, fuzzy logic, and clustering. The course also covers aspects of data preprocessing, transformation, and cleansing; data and feature reduction; developing and testing models; assessing the discovered knowledge; and interpreting and using the results. Through tutorials and projects, students are exposed to data mining software packages such as SAS Enterprise Miner, Weka, and MatLab Fuzzy Logic Toolbox which businesses and scientific community use to build successful data mining applications for making decisions and achieving competitive advantage. • This course is a natural extension of the database design course (CIS-310), in which students learned the basic principles of the database and data warehouse design, data organization and management, SQL, PL/SQL, triggers, cursors, and some OLAP techniques to name a few. This course introduces more sophisticated techniques that businesses use to process data in order to make good and sound business decisions and be more competitive. These techniques go beyond SQL and PL/SQL queries as well as OLAP. • During less than 40 hours that are allocated for this course we will be able to discuss only selected aspects of the knowledge discovery in databases (KDD) and data mining (DM) field. The course material will be covered through a series of lectures, demos, and discussing several successful applications of data mining.
Prerequisites	CIS-310 and MGMT-201 (statistics); MATH-107, and MATH-111 or MATH-205 (calculus) would also be helpful. To really understand the presented material and fully benefit from the course, the students <u>should</u> be familiar with the basic descriptive and inferential statistics concepts, analytical geometry, and matrix and vector algebra. The elementary knowledge of precalculus and calculus will also be helpful. I will try to review the

	necessary concepts briefly during the course and introduce the new ones as needed.
Learning objectives	• Understand the data mining process, and requirements in its every phase to build a successful application. • Learn the fundamental concepts concerning neural networks, fuzzy logic, decision trees, and other data DM methods and to be able to use them in practice for simple business decision making. • Gain some experience with commercial DM packages such as SAS Enterprise Miner, Weka, and the MatLab Fuzzy Logic Toolbox.
Introduction to the field	• Current database technology and the computer hardware allow us to gather, store, access, and manipulate massive volumes of raw data in an efficient and inexpensive manner. In addition, the amount of data collected and warehoused in all industries is growing every year at a phenomenal rate. Nevertheless, our ability to discover critical, non-obvious nuggets of useful information in data, that could influence or help in the business decision making process, is still limited. • Knowledge discovery in databases (KDD) is a multidisciplinary field that focuses on the overall process of information discovery from large volumes of warehoused data. The field combines the database concepts and theory, machine learning, pattern recognition, statistics, artificial intelligence, uncertainty management, and high-performance computing to name a few. Furthermore, the problem of information discovery involves many steps, ranging from data manipulation and retrieval to mathematical and statistical inference, search, and uncertain reasoning. In particular, KDD is a process that typically includes the following stages: learning the application domain; data acquisition, preparation, selection, and cleansing; model and hypothesis development; data mining (DM); assessing the discovered knowledge (testing and verification); interpretation and using discovered knowledge; and visualization of results. • KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns (rules, relationships, correlation, trends, descriptions of rare events) in data by applying computer-based methodology.

	• KDD & DM have become so important that many prominent universities offer an MS degree just in this field. Several top business schools including Stern School of Business, New York University, regularly teach this course in a traditional MBA degree program as well. Other selected schools offer a Ph.D. degree in KDD & DM or business analytics. The UofL College of Business is introducing a 4-course sequence in Business Analytics starting Fall 2017.
Teaching / Learning pedagogy	• Topics noted in the "Tentative Course Outline" section of the syllabus will be covered mainly through lectures, class activities, and demos. Additional assignments will include homeworks, in-class activities, tutorials, labs, and projects. I will post as much course related material as I can on the Blackboard at https://blackboard.louisville.edu/webapps/login. No hard copies of these materials will be given to the students. Students need to come to class prepared with the materials printed off Blackboard and read before the class starts. Blackboard is also equipped with the discussion forum through which students can post messages regarding the assignments or concepts covered in class.
Final drop date	See: http://louisville.edu/calendars/academic/undergrad-grad
Expectations of outside time required for class	To be successful you should allow about 5.0 hours for reading, research and study time each week.

III. Evaluation			
Grading scale	A+ = 97 to 100% B+ = 87 to 89.99% C+ = 77 to 79.99% D+ = 67 to 69.99% Below 60% = F	A = 92 to 96.99% B = 82 to 86.99% C = 72 to 76.99% D = 62 to 66.99%	A- = 90 to 91.99% B- = 80 to 81.99% C- = 70 to 71.99% D- = 60 to 61.99%
Grading scheme	Grading component	Weighted grading percentage	
	4 Tests	65%*****	
	Assignments: (Labs, Tutorials, Homeworks, and In-class Activities)	35%	
	Total	100%	

***You must get at least 60% total average (a D- grade) on the 4 tests to pass the course.**

IV. Schedule			
The CoursePack and Lecture Notes are denoted by (CP) and (LN), respectively. LN are posted on Blackboard. Additional materials may be posted on Blackboard and/or handed in class.			
Week	**Date**	**Topic**	**Readings**
1	Tue: 8/23	Course Syllabus & Schedule Administration of the Course Data Mining Concepts, Data Mining Applications	 pp. 2-13 (CP), LN pp. 87-106 (CP), LN
	Thu: 8/25	Preparing the Data	pp. 15-26 (CP), LN
2	Tue: 8/30	Preparing the Data	pp. 15-26 (CP), LN
	Thu: 9/1	Data Reduction	pp. 28-43 (CP), LN
3	Tue: 9/6	Data Reduction	pp. 28-43 (CP), LN
	Thu: 9/8	Data Reduction	pp. 28-43 (CP), LN
4	Tue: 9/13	Machine Learning (Learning from Data) and Data Mining Tasks and Tools	pp. 45-70 (CP), LN You may rely on LN only. Read the pages from CP to the extent you feel comfortable and to the extent covered in class.
	Thu: 9/15	Machine Learning (Learning from Data) and Data Mining Tasks and Tools	pp. 45-70 (CP), LN, See comments above.
5	Tue: 9/20	Machine Learning (Learning from Data) and Data Mining Tasks and Tools	pp. 45-70 (CP), LN, See comments above.
	Thu: 9/22	**Test 1**	Will cover the material discussed between Aug 23 and Sep 8.
6	Tue: 9/27	Statistical Methods	pp. 72-84 (CP), LN
	Thu, 9/29	Software: SAS 9.4 and SAS Enterprise Miner 14.1	See the materials posted on Blackboard, LN, Demo
7	Tue: 10/4	**Fall Break (No classes)**	
	Thu: 10/6	Decision Trees and Decision Rules	pp. 146-157 (CP), LN
8	Tue: 10/11	Decision Trees and Decision Rules	pp. 146-157 (CP), LN
	Thu: 10/13	Decision Trees and Decision Rules	pp. 146-157 (CP), LN
9	Tue: 10/18	Artificial Neural Networks	pp. 120-133 (CP), 158-180 (CP), LN
	Thu: 10/20	**Test 2**	Will cover the material discussed between Sep 13 and Oct 13.
10	Tue: 10/25	Artificial Neural Networks	pp. 120-133 (CP), 158-180 (CP), LN

		Weka 3.6 – Software for Data Mining from the University of Waikato, New Zealand.	http://www.cs.waikato.ac.nz/ml/weka/
	Thu: 10/27	Artificial Neural Networks	pp. 120-133 (CP), 158-180 (CP), LN
11	Tue: 11/1	Artificial Neural Networks	pp. 120-133 (CP), 158-180 (CP), LN
	Thu: 11/3	Cluster Analysis	pp. 181-197 (CP), LN
12	Tue: 11/8	**Election Day – No Classes**	
	Thu: 11/10	Association Rules	LN
13	Tue: 11/15	Fuzzy Sets and Fuzzy Logic	pp. 134-144 (CP), LN
	Thu: 11/17	**Test 3**	Will cover the material discussed between Oct 20 and Nov 10.
14	Tue: 11/22	MatLab Fuzzy Logic Toolbox Fuzzy Sets and Fuzzy Logic	See the materials posted on Blackboard. pp. 134-144 (CP), LN
	Thu: 11/24	**Thanksgiving (No Classes)**	
15	Tue: 11/29	MatLab Fuzzy Logic Toolbox	See the materials posted on Blackboard. pp. 134-144 (CP), LN
	Thu: 12/1	Genetic Algorithms	pp. 108-119 (CP), pp. 198-211 (CP), LN
16	Tue: 12/6	Genetic Algorithms Memory-based reasoning, Text Mining, Web Mining, Visualization Methods (**only if time allows**)	pp. 108-119 (CP), pp. 198-211 (CP), LN Additional handout will be provided, if needed.
17	Wed, 12/14, 11:30a.m.-12:45p.m.	**Test 4** Also see: http://louisville.edu/registrar/registration-information/final-exam-schedules	Will cover the material discussed between Nov 17 and Dec 6. Test 4 may be an on-line exam or take-home.

V. Additional Work Details	
Blackboard	Posting and correction of grades. Grades will be posted on the course website (Blackboard) and students need to check them periodically. If a posted grade(s) is incorrect, please dispute it within 1 week after it is posted by coming to my office, calling me or sending e-mail. No grades will be changed after one week after the posting.

Participation and class contribution	• I strongly encourage you to participate in classes. The material covered will mainly be based on the CoursePack and lecture notes posted on Blackboard. The student will be fully responsible for all the material covered in class and the work assigned, whether present in class or not. Missing the class discussion may hurt the understanding of the material, and in-class activities (projects) cannot be made-up. (To earn credit for in-class activities you need to be present in class on the day when they are held and actually work the assigned problems with your group.) Due dates for homeworks, labs, projects, etc. will be announced. • It is your responsibility to attend every class and turn assignments on time. If you choose to miss class, come late or leave earlier, ask your colleagues what was covered in class. If you have questions, which require longer answers, about any assigned problems (homeworks, labs, in-class activities, tests, etc.) or material covered in class, please come to my office. Face to face interaction is much more educational, more effective, and less time consuming than my replying to your e-mails.
Changes in the syllabus	Any changes to the syllabus and course schedule will be announced in class and posted on Blackboard.

V1. Student Responsibilities / College and University Issues	
University of Louisville student conduct and responsibilities	This course will abide by University of Louisville student conduct and responsibilities with regards to ethics and related issues: http://louisville.edu/dos/students/studentrightsandrespsonsibilities
College of Business student conduct and responsibilities	This course will abide by College of Business student conduct and responsibilities with regards to ethics and related issues: http://business.louisville.edu/students/college-of-business-academic-dishonesty-policy
Religious holiday conflict policy	http://louisville.edu/diversity/documents/work-restricted-holy-day-policies-and-calendar
University policy on equal access	http://louisville.edu/disability
Title IX/Clery Act Notification concerning sexual misconduct	http://louisville.edu/delphi/resources/syllabus/samples
Classroom policy	• Student conduct. Students are not forced to come to class. If students come to class, they chose to stay for the entire class

	period. Accidental leaving and returning to class, whenever you desire, is not allowed and will not be tolerated. Such behavior is considered Bad Conduct and shows bad manners. Furthermore, such behavior disturbs your classmates and instructor. Cell phones and pagers must be turned off. The use of your laptops is restricted to course related issues only. No food or drink in COB classrooms and all computer labs. The only exception is a bottle or cup of water. • Everything submitted for grading must reflect your own work. Labs, projects, and homeworks should be handed in before class starts. All late assignments will be penalized 10% per calendar day. No assignments will be accepted after solutions are handed in, posted on the web or discussed in class. Prior permission is needed for make-up examinations. "No shows" on test day will automatically receive a "zero" for that test.

VII. Other Notes
Please keep in mind that I have a **hearing disability** which severely impairs the hearing in my right ear and reduces the hearing in my left ear. This is a condition that cannot be easily remedied with the hearing aids I am wearing. To allow me to answer your questions please remain quiet during lectures and ask your questions loudly and clearly. If necessary, I will ask you to repeat the question and approach you.