

School of Library and Information Studies
University of Oklahoma

LIS 4950/5950
Knowledge Discovery in Databases / Data Mining
Draft, Fall 2007

Instructor: William Hanson
Office Phone: 325-3921 **E-Mail:** whanson@ou.edu
Office Hours: To Be Determined (TBD)
Class Times: TBD
Class Location: Bizzel Library, Room TBD
Online Materials: <http://learn.ou.edu>

Course Description:

Knowledge Discovery in Databases / Data Mining

Knowledge Discovery in Databases (KDD), also known as Data Mining, is the ability to discover useful patterns and relationships in large sets of data. KDD is an induction-based set of computer-based methodologies to extract previously unknown patterns in data and can provide a way to both understand and predict behavior. This course will primarily focus on hands-on exercises, supplemented by lecture, group discussion, and assigned readings.

Prerequisite: Junior Standing (TBD).

Course Objectives:

Upon successful completion of this course, the student will be able to:

- Know the basic definition of data mining and be able to recognize when data mining can be used to solve problems
- Understand and be able to apply the knowledge discovery process
- Understand how data mining techniques build models and concepts to solve problems
- Know several data mining strategies and recognize when each strategy is appropriate
- Perform basic statistical and nonstatistical techniques to evaluate the results of knowledge discovery
- Understand the use of spreadsheet based tools to perform data analysis
- Understand and be able to apply software tools to perform data mining, including clustering, rule generation, and neural network tools

Expanded objectives:

Chapter One: The Basics

- Be able to define data mining
- Understand that computers can learn concept definitions
- Know when data mining is an appropriate strategy
- Be able to explain the purpose of data classification models
- Know that data mining has been successful in a number of domains

Chapter Two: Data Mining

- Be able to recognize several data mining techniques
- Understand the difference between supervised and unsupervised techniques
- Understand basic techniques for evaluation, including confusion matrices and lift measurements

Chapter Three: Basic Data Mining Techniques

- Be able to choose an appropriate data mining technique
- Understand how to construct decision trees
- Understand how to generate association rules
- Recognize genetic algorithm techniques

Chapter Four: The iData Analyzer Tool

- Know how to use the iData Analyzer tools
- Be able to use ESX to perform both supervised and unsupervised learning
- Be able to create rules
- Be able to compute and analyze both numerical and categorical significance
- Be able to compute and analyze instance typicality

Chapter Five: Knowledge Discover in Databases

- Know the seven-step KDD process
- Understand how to normalize, convert and smooth data
- Understand how to create and eliminate attributes
- Be able to state the advantages and disadvantages of various methods of dealing with missing data
- Be able to recognize the Cross Industry Standard Process Model for Data Mining (CRISP-DM)

Chapter Six: The Data Warehouse

- Be able to state the differences between transactional databases and decision support databases
- Know the two methods of data warehouse architecture
- Know how on-line analytical processing can be used to analyze multi-dimensional data
- Know how to use Microsoft Excel pivot tables to model multi-dimensional data

Chapter Seven: Formal Evaluation Techniques

- Know how to determine confidence intervals
- Know how to perform hypothesis testing
- Be able to interpret correlation coefficients and scatterplots
- Be able to use statistical evaluation methods to compare different models

Chapter Eight: Neural Networks

- Understand the strengths and weaknesses of neural networks
- Know how feed-forward neural networks operate
- Know how genetic algorithms and back-propagation are used to train neural networks
- Understand how neural networks perform unsupervised clustering

Chapter Nine: Building Neural Networks with iDA

- Be able to perform supervised neural network learning
- Be able to perform unsupervised neural network clustering
- Be able to interpret the meaning of clusters formed in an unsupervised neural network session

Chapter Eleven: Specialized Techniques

- Know how to perform time-series analysis
- Know how data mining can be used for web page development and automation
- Know how data mining can extract useful patterns from unstructured text

Chapter Twelve: Rules-Based Systems

- Know that the field of artificial focuses on problems that cannot be solved using traditional computing techniques
- Know that rules-based systems separate problem-solving knowledge from the reasoning mechanism used to apply the knowledge
- Be able to state the steps to develop a rules-based expert system
- Be able to recognize types of problems that can be solved with rules-based systems

Required Textbook

Data Mining: A Tutorial-Based Primer, Roiger and Geatz. Addison-Wesley, 2003. ISBN 0-201-74128-8 (note: accompanying software and datasets are required for the course)

Class Department (Behavior):

1. Drinking is allowed...Eating is not allowed.
2. Pagers and Cell Phones are to be turned off during class.
3. Be on time – to class and with projects. Attendance will be taken at the start of each class.
4. Come to class to **actively** participate. It is essential to your development and understanding of the course – part of your grade depends on it. I will appreciate your effort and you in turn will find the time spent in class more enjoyable.
5. Read the assignment and be prepared for each class session.
6. Do not hesitate to ask questions during class. If it is not clear to you, it probably is not clear to someone else.

Grading:

Item	Points	Weight	Final Grades – Points (%)
Midterm Exam	150	15%	A: 900 + (90% and above)
Final Exam	200	20%	B: 800 – 899 (89% – 80%)
Chapter exercises (11 @ 50)	550	55%	C: 700 – 799 (79% – 70%)
Classroom Preparation (50 pts attendance; 50 pts all other, i.e., participation in class discussions, etc.)	100	10%	D: 600 – 699 (69% – 60%) F: less than 600 (below 60%)
Total	1000	100%	

A grade of A implies that all requirements have been met and substantially exceeded. “A, the highest grade, is given of work of exceptional quality. D is the lowest grade for which credit is given in any under graduate college and means that, although in the judgment of the instructor credit should be allowed for the course, a degree will not be conferred upon a student whose work is all of that level.” (*The University of Oklahoma General Catalog, 2003-2006*, p. 34)

A grade of F implies failure to meet minimal requirements. School of Library and Information Studies policy requires that any student receiving a grade of F be recommended for dismissal from the Bachelor of Arts and Information Studies Program.

Graded Items:

1. This class will focus on “learning by doing”. Accordingly, the majority of your grade will be a series of short exercises to give you plenty of practice at various aspects of data mining.
2. There will be a mid-term, worth 150 points, and a comprehensive final worth 200 points.
3. Tests will be a combination of multiple choice, short answer and essay. Every test will be on material covered in lectures or in the text. **Pop quizzes will be given at random times.**

Due Dates and Late Assignments

All assignments are due at the beginning of class on the due date in the syllabus. Late work will not be accepted except under extraordinary circumstances, and then only by prior arrangement with the instructor. Unless prior arrangements have been made, there will be no make-up exams or quizzes. Any make-up exams will be an alternate (harder) test.

Attendance :

I will take roll at the beginning of each class and **attendance is part of your grade**. Unless you make prior arrangements with me, students missing class on days of quizzes or class exercises will receive zero (0) credit for that day's assignment.

The number one reason for attending class is that your instructor is the person who determines your grade – it's impossible to get a good grade in this class without regular attendance.

(The following is extracted from University Policy 4.19.1 Class Attendance: Students):

“Students are responsible for the content of courses in which they are enrolled. Specific policy concerning attendance requirements and announced and unannounced examinations is the responsibility of the individual instructor. Students have a responsibility to inform faculty prior to absences whenever possible. Faculty should make every effort to find a reasonable accommodation for students who miss class as a result of participation in Provost-approved University-sponsored activities or legally required activities such as emergency military service. Students missing class on account of jury duty must receive such an accommodation. When absences seriously affect a student's class work, the instructor will report this fact to the Admissions and Records Office, where the information will be directed to the dean concerned.”

Course Outline (All readings are from *Roiger and Geatz* unless otherwise stated -- Graded items in **Underline):**

Week	Dates	Subject	Reading
1		Introduction and Course Overview Data mining definitions and computer learning	None 1.1 – 1.3
2		Is data mining the right tool? Data mining process model and application Chapter exercise due	1.4-1.7
3		Supervised and unsupervised techniques	2.1 – 2.4
4		Evaluation techniques Chapter exercise due	2.5
5		Decision Trees and association rules, choosing the right techniques Chapter exercise due	3.1 – 3.3, 3.5
6		Data mining tools – using the iData Analyzer	4.1 – 4.3
7		Specific approaches and techniques Chapter exercise due	4.4 – 4.8
8		Mid-Term exam	
9		KDD process model	5.1 – 5.7
10		Applications of the KDD process Chapter exercise due	5.8 – 5.10
11		Data Warehouses Chapter exercise due	6.1 - 6.4
12		Formal evaluation techniques Chapter exercise due	7.1 – 7.6
13		Neural Networks Chapter exercise due	8.1 – 8.4
14		Building Neural Networks Chapter exercise due	9.1 – 9.3
15		Time series, web mining, text mining Chapter exercise due	11.1 – 11.3
16		Rules-based systems Chapter exercise due	12.1 – 12.4
17		Final Exam – TBD	

Academic Integrity:

(the following statement is extracted from <http://www.ou.edu/provost/integrity> -- all students are encouraged to review the OU policy and to discuss any issues or questions with me)

What does "academic integrity" mean?

Academic integrity means honesty and responsibility in scholarship. Professors have to obey rules of honest scholarship, and so do students. Here are the basic assumptions about academic work at the University of Oklahoma:

- (1) Students attend OU in order to learn and grow.
- (2) Academic assignments exist for the sake of this goal.
- (3) Grades exist to show how fully the goal is attained.
- (4) Thus, all work and all grades should result from the student's own effort to learn and grow. Academic work completed any other way is pointless, and grades obtained any other way are fraudulent.

Academic integrity means understanding and respecting these basic truths, without which no university can exist. Academic misconduct -- "cheating" -- is not just "against the rules." It violates the assumptions at the heart of all learning. It destroys the mutual trust and respect that should exist between student and professor. Finally, it is unfair to students who earn their grades honestly.

Sexual Harassment Policy:

The University of Oklahoma explicitly condemns sexual harassment of students, staff, and faculty. Sexual harassment is unlawful and may subject those who engage in it to university sanctions as well as civil and criminal penalties. Sexual harassment is defined as unwelcome sexual advances, requests for sexual favors, and other verbal or physical conduct of a sexual nature in the following context:

- 1) When submission to such conduct is made either explicitly or implicitly a term or condition of an individual's employment or academic standing, or
- 2) When submission to or rejection of such conduct by an individual is used as the basis for employment or academic decisions affecting such individual, or
- 3) When such conduct has the purpose or effect of unreasonably interfering with an individual's work or academic performance or creating an intimidating, hostile, or offensive working or academic environment.

University of Oklahoma Reasonable Accommodation Policy on Disabilities

Any student in this course who has a disability that may prevent him or her from fully demonstrating his or her abilities should contact me personally as soon as possible so we can discuss accommodations necessary to ensure full participation and facilitate your educational opportunities.

University of Oklahoma Policy on Religious Holidays

It is the policy of the University to excuse absences of students that result from religious observances and to provide without penalty for the rescheduling of examinations and additional required class work that may fall on religious holidays.

Internet Information

Note takers for the course will be posted on the OU Desire to Learn (D2L) site, <http://learn.ou.edu>. Note takers, revisions to this syllabus, announcements, and revisions to the course outline will be posted in the appropriate folder. You are expected to use this Internet site to keep abreast of course changes. You are also encouraged to download, print out, and bring the note taker to class. The D2L site will also include the class discussion board and chat room.

WILLIAM E. HANSON
Instructor, Library and Information Studies