Computer Science 6973: Knowledge Discovery (and Data Mining)

Instructor: Dr. McGovern

Spring 2007

1 General Information

Class time: TR 3-4:15

Class location: TBA

Prerequisites: CS 5033. Note that prior programming experience is assumed. Also, I assume a mathematical background and an understanding of statistics and probability.

Required Materials: Principles of Data Mining by David J. Hand, Heikki Mannila and Padhraic Smyth

Instructor: Dr. McGovern

- *Office:* EL 144A
- *Phone:* 325-5427 (voice mail available)
- *URL for class:* http://learn.ou.edu
- *Personal URL:* http://www.cs.ou.edu/∼amy
- *Email:* amcgovern@ou.edu
- *Office hours:* To be determined. Also by appointment. Also available via AIM at *dramymcgovern.*

2 Learning goals and topics

In today's information-rich world, finding automatic ways to sift through the wealth of information and to identify the salient patterns is critical. This course will cover knowledge discovery/data mining approaches to addressing this issue. The process of knowledge discovery is about making sense of data that is too complex for humans to understand. Knowledge discovery is a interactive process between a human and a computer. Data mining techniques highlight salient patterns in the data. These patterns are examined by the human to provide feedback. Data mining refines the patterns and the cycle continues until the human is satisfied.

By the end of the semester, students will be able to:

- Recognize a knowledge discovery/data mining task
- Choose an appropriate knowledge representation (static vs dynamic, discrete vs continuous, propositional vs relational)
- Identify the elements of a successful knowledge discovery algorithm
- Apply current knowledge discovery/data mining algorithms to new situations
- Create new and successful knowledge discovery/data mining algorithms
- Evaluate and analyze the performance of new approaches

This will be a seminar style class where we will read both the book and discuss current papers. In addition, we will have a class-wide project (on the new netflix data where we will aim to improve the prediction of movies that people will enjoy) that we will work on and discuss. A general schedule follows and the specific schedule is available from my website.

Weeks 1

Topics: Introduction, seminar style classes, project discussions.
Readings: Barn-raising paper, Introduction to Knowledge Discovery (KDD Cup paper)

Weeks 2-5

Topics: Statistical Background and Data Representation
Readings: Chapters 1-4 of the book

Week 6-10

Topics: Building the data mining toolbox and study of specific algorithms: association rules, naive bayes classifiers, clustering, linear regression, dependency networks
Readings: Chapters 5-13

Week 11-15

Topics: Applications, specific systems, advanced techniques, project discussion
Readings: Current papers: including fraud detection, analyzing football coach's social network, severe weather detection, ...

3 Course Policies

Attendance: We will discuss concepts and examples in class that are not in the text book. Another student's notes are an inadequate substitute for class attendance. You are responsible for everything that is announced in class.

Class Web Page: This class will use Desire2Learn software for our web page. The URL for the home page is http://learn.ou.edu. Login with your 4+4 (first four letters of your last name followed by the last four digits of your student number), using your standard OU password. If you have difficulty logging in, call 325-HELP. This software provides a number of useful features, including a list of assignments and announcements, an electronic mailing list, newsgroups, and grade book. All handouts are available from Desire2Learn. I update this web site several times a week. You should check the site daily. You are responsible for things posted on the site with a 24 hour delay.

Class Email Alias: Urgent announcements will be sent through email. It is your responsibility to:

- Have your university supplied email account properly forwarded to the location where you read email.
- Make sure that your email address in Desire2Learn is correct, and forwards email to the place where you read it. I'll send out a test message during the first week of class. If you do not receive this message, it is your responsibility to get the problem resolved immediately.
- Have your email program set up properly so that replying to your email will work correctly the first time. You can send email to yourself and reply to yourself to test this. I will not make any attempt to get bounced email messages delivered.

If you need assistance in accomplishing any of these tasks, contact 325-HELP.

Newsgroups and Email: The newsgroup on Desire2Learn should be the primary method of communication outside of class. This allows everyone in the class to benefit from the answer to your question, and provides students with more timely answers since I check Desire2Learn at least once a day. Matters of personal interest should be directed to email instead of to the newsgroup, e.g. informing me of an extended personal illness. Posting guidelines for the newsgroup are available on Desire2Learn.

Academic Misconduct: All work submitted for an individual grade, including homework and individual projects, should be the work of that single individual, and not her friends, and not her tutor. It is acceptable to ask a fellow student for help as long as that help does not consist of copying any computer code, or solutions to other assignments.

1. Do not show another student a copy of your projects or homework before the submission deadline. The penalties for permitting your work to be copied are the same as the penalties for copying someone elses work.

2. Make sure that your computer account is properly protected. Use a good password, and do not give your friends access to your account or your computer system. Do not leave printouts, floppy disks or thumb drives around a laboratory where others might access them.

Programming projects will be checked by software designed to detect collaboration. This software is extremely effective and has withstood repeated reviews by the campus judicial processes.

Upon the first documented occurrence of collaborative work, I will report the academic misconduct to the Campus Judicial Coordinator. The procedure to be followed is documented in the University of Oklahoma Academic Misconduct Code[1]. In the unlikely event that I elect to admonish the student, the appeals process is described in http://www.ou.edu/provost/integrity-rights/.

Incompletes: The grade of I is intended for the rare circumstance when a student who has been successful in a class has an unexpected event occur shortly before the end of the class. I will not consider giving a student a grade of I unless the following three conditions have been met. 1. It is within two weeks of the end of the semester. 2. The student has a grade of C or better in the class. 3. The reason that the student cannot complete the class is properly documented and compelling.

[1] http://www.ou.edu/studentcode

Accommodation of Disabilities: The University of Oklahoma is committed to providing reasonable accommodation for all students with disabilities. Students with disabilities who require accommodations in this course are requested to speak with the professor as early in the semester as possible. Students with disabilities must be registered with the Office of Disability Services prior to receiving accommodations in this course. The Office of Disability Services is located in Goddard Health Center, Suite 166, phone 405/325-3852 or TDD only 405/325-4173.

Classroom Conduct: Disruptions of class will not be permitted. Examples of disruptive behavior include:

- Allowing a cell phone or pager to repeatedly beep audibly.
- Answering your cell phone in class (if you do this, I get to talk to the person on the other end!)
- Playing music or computer games during class in such a way that they are visible or audible to other class members.
- Exhibiting erratic or irrational behavior.
- Behavior that distracts the class from the subject matter or discussion.
- Making physical or verbal threats to a faculty member, teaching assistant, or class member.
- Refusal to comply with faculty direction.

In the case of disruptive behavior, I may ask that you leave the classroom and may charge you with a violation of the Student Code of Responsibilities and Conduct.

4 Homework and Projects

Due dates: Unless the assignment specifies otherwise, all homeworks and projects will be due at the beginning of class, 12 noon.

Slack days: You have one slack day for reading summaries and one slack day for a homework. For a summary, the slack day will entitle you to skip the summary for that reading. For the homework, the slack day will give you a 24 hour extension. Please note that you should not skip the reading entirely as you will end up behind! At the end of the semester, unused slack days will count towards class participation points.

Projects: There will be a class-wide project. We will focus on the newly released netflix data where the goal (for $\$1,000,000$) is to improve the netflix predictions by 10%. This project will be collaborative. To participate in the project, you must sign the legal agreements in the first week of class.

Project code: Your project code must be written exclusively by you. Students working on joint projects may certainly help one another and are expected to share code within the project group. However, they may not share beyond the group. Your project code and writeups must be written exclusively by you or your group. Use of any downloaded code or code taken from a book (whether documented or undocumented) is considered academic misconduct and will be treated as such. Exceptions from this policy (such as a project that builds on an existing open-source project) may be granted but you **MUST** speak with me first.

5 Grading and Evaluation

Grade calculation: Your final grade will be determined as follows:

- Final project: 50%
- Homework and summaries: 20%
- Class participation: 30%

Note that class participation is important. This means that attending class, asking questions, and answering questions are all important for your grade. Since this is a small seminar, class participation is expected for all members.

Grade questions: Please note that I will examine the entire project or homework in question and your final grade may end up lower. All disagreements about the grading of projects or homework must be brought to my attention within one week of when the item was returned.

Desire2Learn Grade Summary: Desire2Learn has a grade book that is used to store the raw data that is used to calculate your course grade. It is the responsibility of each student in this class to check their grades on Desire2Learn after each project or homework is returned. If an error is found, bring the grading document to me or the TA, and we will correct Desire2Learn.

Borderline grades: Borderline final grades will be decided by two factors: class participation and your final exam grade. If you are close to a border and you did well on the final, that can push you over a grade boundary. Likewise, being an active participant in class can push you over a grade boundary.