CSC529: Advanced Data Mining

Session: Spring
5:45-9:00, 3/30 – 6/08
Class: CDM 220
Office: CDM 837

Instructor: Jonathan F. Gemmell
Email: jonathan.gemmell@gmail.com
Cell Phone: (312) 810-3167
Office Hours: See website
Please email ahead of time

Course Description: The course is for students with prior background in data mining or machine learning techniques. The course will cover advanced modeling techniques, such as ensemble learning, extended linear models, probabilistic graphical models, mixture and latent variable models, and matrix factorization. First, the theoretical foundations of these techniques will be presented and augmented with in-class examples and homework problems. Second, the state-of-the-art research related to these techniques will be presented and augmented with paper reviews that highlight the practical applications of these advanced data mining techniques. Applications of the models will be presented in popular domains, including social computing and health informatics.

Course Learning Goals: At the end of the course, students should be able to:

- understand the basics behind each data mining method as well as the respective cons and pros
- understand how information in real world applications can be formulated and represented as different genres of data, such as matrices, sequences, data streams, graphs/networks
- select, combine, and apply specific data mining techniques for certain data types and challenges, and understand, explain, and interpret the obtained results, and
- identify recent trends and open directions in the field of data mining.

Prerequisites: CSC 424 and (IS 467 –formerly IS567, ECT 584, CSC 578, or CSC478)

Course Management System: DePaul University's Desire2Learn system (d2l.depaul.edu).

Recommended books:

A. *Data Mining: Practical Machine Learning Tools and Techniques* by Witten, Frank, and Hall, 3rd Edition, ISBN 978-0-12-374756-0
    a. This book has a focus of practical applications and the use of the WEKA toolkit.
B. *Probabilistic Graphical Models, Principles and Techniques* by Daphne Koller and Nir Friedman, ISBN 978-0-262-01319-2
    a. This book has a focus on theoretical foundations of probabilistic graphical models
        i. http://mitpress.mit.edu/books/probabilistic-graphical-models

b. eBook available through the DePaul library at
    i. http://depaul.worldcat.org/title/data-mining-practical-machine-learning-tools-and-techniques/oclc/706802868&referer=brief_results

C. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Hastie, Tibshirani, Friedman
    a. This book has a focus on theoretical foundations of data mining; PDF available at http://www-stat.stanford.edu/~tibs/ElemStatLearn

News Widget: The primary form of communication for this class will be the news widget on the D2L. Please make sure you subscribe to the widget and that DePaul has you correct email.

Forums: The class forum is the preferred place to ask questions about the class. If you have questions about a quiz question, the exams or lecture notes, please post them there. I read these frequently. All students should subscribe to the forums so that you receive email updates.

Software: The statistical packages used in this course are R and SAS. They are available in all DePaul labs. You can also access SAS remotely by using our CDM terminals (suitable for fast connections). More information about the software will be posted on the course website.

Grades:			**Final Grades:**	
	Readings:	10%	A:	90% - 100%
	Exam Preps:	10%	B:	80% - 90%
	Case Studies:	30%	C:	70% - 80%
	Project:	20%	D:	60% - 70%
	Final Exam:	30%	F:	less than 60

*Plusses and minuses are given to the upper and lower 3 %

Class Contract: All students must complete the class contract. It is meant to ensure that you have read the syllabus and understand the requirements for the course.

Readings: There are 10 assigned readings throughout the quarter. Readings are due every week before class. Readings are drawn from academic journals to the popular press, from theoretical concepts to practical applications and from interesting recent findings to important historical results.

Exam Preps: There are 10 exam preps. Preps may be taken as many as 10 times and only the highest grade is recorded. Preps are taken online through the D2L. They are due at 5:45 on the day of the class in which they will be reviewed. There will never be an extension for the exam preps.

The purpose of the online preps is not to test your knowledge. You may (and are encouraged) to work through the problems with your fellow classmates, seek out the help of tutors, ask questions on the forums or even offer potential solutions on the forums.

The purpose of the exam preps is to prepare you for the exam (which will test your knowledge). The questions are very (VERY!) indicative to what will be on the exams. Thus, when taking a prep, you should endeavor to learn the underlying material and not simply get the right answers in order to rack up points.

Case Studies: Throughout the quarter, you will perform three case studies: Case Study 1 (week 4) – Ensemble Classifiers or Ensemble Clustering. Case Study 2 (week 6) – Kernel PCA or Kernel SVM. Case Study 3 (week 8) – Bayesian Networks or Markov Models. For each case study, you will identify a dataset that would be appropriate for one of the techniques presented in class. See the D2L for the case study handout.

Project: For the final project, groups will be assigned an advanced data-mining topic. Collaboratively the group will create a discussion of the algorithm, a tutorial on how to use the algorithm, and several examples of how to employ the algorithm. The final report should be composed as follows: See the D2L for the final project handout.

Exams: The final exam is given during week 11; please see the calendar. The exam may not be rescheduled accept under extreme extenuating circumstances. Exams are closed-book and closed-notes.

Proctored Exams: DL students may take the exams 1) in class, 2) on campus in the DL office or 3) with a proctor. DL students must take the exam in the allotted window. Exams may not be rescheduled accept under extreme extenuating circumstances. Registration begins on the D2L by using the "Proctored Exam" tab. Please see: http://www.cdm.depaul.edu/onlinelearning/Pages/FAQ.aspx#exams

Attendance: Attendance is not required. However, it is expected that you will attend every class; it is the single most important action you can take in mastering the course objectives. You are responsible for all material covered, assignments delivered or received, and announcements made in class sessions that you miss. For distance learning students, this means viewing the classes in a timely manner, participate in the discussion forum, and being sure to email or call in any questions that you have.

Online Teaching Evaluation: Evaluations are a way for students to provide valuable feedback regarding their instructor and the course. Detailed feedback will enable the instructor to continuously tailor teaching methods and course content to meet the learning goals of the course and the academic needs of the students. They are a requirement of the course and are key to continue to provide you with the highest quality of teaching. The evaluations are anonymous; the instructor and administration do not track who entered what responses. A program is used to check if the student completed the evaluations, but the evaluation is completely separate from the student's identity. Since 100% participation is our goal, students are sent periodic reminders over three weeks. Students do not receive reminders once they complete the evaluation. Students complete the evaluation online in CampusConnect.

Academic Integrity Policy: This course will be subject to the academic integrity policy passed by faculty. More information can be found at http://academicintegrity.depaul.edu/

Plagiarism: The university and school policy on plagiarism can be summarized as follows: Students in this course should be aware of the strong sanctions that can be imposed against someone guilty of plagiarism. If proven, a charge of plagiarism could result in an automatic F in the course and possible expulsion. The strongest of sanctions will be imposed on anyone who submits as his/her own work any assignment which has been prepared by someone else. If you have any questions or doubts about what plagiarism entails or how to properly acknowledge source materials be sure to consult the instructor.

Incomplete: An incomplete grade is given only for an exceptional reason such as a death in the family, a serious illness, etc. Any such reason must be documented. Any incomplete request must be made at least two weeks before the final, and approved by the Dean of the College of Computing and Digital Media. Any consequences resulting from a poor grade for the course will not be considered as valid reasons for such a request.

Resources for Students with Disabilities: Students who feel they may need an accommodation based on the impact of a disability should contact the instructor privately to discuss their specific needs. All discussions will remain confidential.

To ensure that you receive the most appropriate accommodation based on your needs, contact the instructor as early as possible in the quarter (preferably within the first week of class), and make sure that you have contacted the Center for Students with Disabilities (CSD) at:

Student Center, LPC, Suite #370
Phone number: (773)325.1677
Fax: (773)325.3720
TTY: (773)325.7296