Fall 2016 CSE 4334/5334.001

Data Mining

Syllabus **CSE4334 schedule** **CSE5334 schedule**

Course Information:

- Time: Tue/Thu 2:00-3:20pm (CSE4334); Fri 4-6:50pm (CSE5334.001)
- Classroom: NH 112 (CSE4334); ERB 131 (CSE5334.001)
- Homepage: http://crystal.uta.edu/~cli/cse4334 http://crystal.uta.edu/~cli/cse5334

Instructor: Chengkai Li

- Office hours: Tue/Thu 3:30-4:30pm
- Office: ERB 628
- Phone: (817) 272-0162
- E-mail: cli [AT] uta [DOT] edu
- Homepage: http://ranger.uta.edu/~cli

TA1:

- Office hours:
- Office:
- E-mail:

TA2: Afroza Sultana

- Office hours:
- Office:
- E-mail:

Course Description: This is an introductory course on data mining. Data Mining refers to the process of automatic discovery of patterns and knowledge from large data repositories, including databases, data warehouses, Web, document collections, and data streams. We will study the basic topics of data mining, including data preprocessing, data warehousing and OLAP, data cube, frequent pattern and association rule mining, correlation analysis, classification and prediction, and clustering, as well as advanced topics covering the techniques and applications of data mining in Web, text, big data, social networks, and computational journalism.

Student Learning Outcomes: A solid understanding of the basic concepts, principles, and techniques in data mining; an ability to analyze real-world applications, to model data mining problems, and to assess different solutions; an ability to design, implement, and evaluate data mining software.

Prerequisites:

- For CSE 4334: CSE 3330 Database Systems I and IE 3301 Engineering Probability (or MATH 3313 Introduction to Probability) or consent of instructor.
- For CSE 5334: prerequisites for CSE5334: There is no official prerequisites. You should have sound CSE background from your Bachelor's program (e.g., programming, data structures and algorithms, discrete mathematics, basics of probabilities and statistics). If you don't have database course from anywhere, you are allowed to take the course, but please get the consent of the instructor. You also must get the consent of the instructor if you have CSE deficiency courses to take.

Textbook

- (**Required**) [TSK] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining, Addison-Wesley, 2006. ISBN 0-321-32136-7. (Sample chapters at http://www-users.cs.umn.edu/~kumar/dmbook/index.php)
- (**Required for relevant chapters**) [MRS] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval, Cambridge University Press. 2008. (Free book at http://nlp.stanford.edu/IR-book/)
- (Reference) Jure Leskovec, Anand Rajaraman and Jeff Ullman. Mining of Massive Datasets, 2nd ed., Cambridge University Press. (Free book at http://www.mmds.org/#ver21)
- (Reference) Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, 3rd ed. (2nd

edition is also fine), Morgan Kaufmann Publishers, June 2011. ISBN 9780123814791.
- (Reference) Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R, 1st ed., Springer, 2013. (Free book at http://www-bcf.usc.edu/~gareth/ISL/index.html)
- (Reference) I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005.

Grades

- Pop quizzes 25%
- Final Exam 30% (Time and Location TBD)
- Homework (HW) 10% (Must be done independently)
- Programming Assignments (P) 35% (Must be done independently)

The final letter grades will be based on students' performance. There is no pre-defined cutoffs or distribution of grades. Undergraduate and graduate students are compared in separate groups.

Attendance: At The University of Texas at Arlington, taking attendance is not required but attendance is a critical indicator in student success. Each faculty member is free to develop his or her own methods of evaluating students' academic performance, which includes establishing course-specific policies on attendance. As the instructor of this section, I require all students to attend lectures. However, while UT Arlington does not require instructors to take attendance in their courses, the U.S. Department of Education requires that the University have a mechanism in place to mark when Federal Student Aid recipients "begin attendance in a course." UT Arlington instructors will report when students begin attendance in a course as part of the final grading process. Specifically, when assigning a student a grade of F, faculty report the last date a student attended their class based on evidence such as a test, participation in a class project or presentation, or an engagement online via Blackboard. This date is reported to the Department of Education for federal financial aid recipients.

Announcements: Stay tuned and make sure to check Blackboard frequently. Important announcements will be posted there.

Assignments and Deadlines

- All the assignments must be submitted through Blackboard. We will NOT take hardcopy or email submission, unless the university verifies that Blackboard was malfunctioning or unavailable. If you are not able to submit through Blackboard due to its technical failure, you can email your assignment to us, together with a screenshot showing the technical failure. We will verify with the university.
- Everything is due by 11:59pm on the due date. The deadline is automatically managed by Blackboard. You can still turn in assignment after the deadline. However, you automatically lose 5 points per hour after the due time, till you get 0. (Each individual assignment is 100 points.) We cannot waive the penalty, unless there was a case of illness or other substantial impediment beyond your control, with proof in documents.

Regrading: Regrading request must be made within 7 days after we post scores on Blackboard. TA will handle regrade requests. If student is not satisfied with the regarding results, you get 7 days to request again. The instructor will regrade, and the decision is final.

Drop Policy: Students may drop or swap (adding and dropping a class concurrently) classes through self-service in MyMav from the beginning of the registration period through the late registration period. After the late registration period, students must see their academic advisor to drop a class or withdraw. Undeclared students must see an advisor in the University Advising Center. Drops can continue through a point two-thirds of the way through the term or session. It is the student's responsibility to officially withdraw if they do not plan to attend after registering. Students will not be automatically dropped for non-attendance. Repayment of certain types of financial aid administered through the University may be required as the result of dropping classes or withdrawing. For more information, contact the Office of Financial Aid and Scholarships (http://wweb.uta.edu/aao/fao/).

Disability Accommodations: UT Arlington is on record as being committed to both the spirit and letter of all federal equal opportunity legislation, including *The Americans with Disabilities Act (ADA)*, *The Americans with Disabilities*

Amendments Act (ADAAA), and *Section 504 of the Rehabilitation Act*. All instructors at UT Arlington are required by law to provide "reasonable accommodations" to students with disabilities, so as not to discriminate on the basis of disability. Students are responsible for providing the instructor with official notification in the form of a letter certified by the Office for Students with Disabilities (OSD). Only those students who have officially documented a need for an accommodation will have their request honored. Students experiencing a range of conditions (Physical, Learning, Chronic Health, Mental Health, and Sensory) that may cause diminished academic performance or other barriers to learning may seek services and/or accommodations by contacting:
The Office for Students with Disabilities, (OSD) www.uta.edu/disability or calling 817-272-3364. Information regarding diagnostic criteria and policies for obtaining disability-based academic accommodations can be found at www.uta.edu/disability.
Counseling and Psychological Services, (CAPS) www.uta.edu/caps/ or calling 817-272-3671 is also available to all students to help increase their understanding of personal issues, address mental and behavioral health problems and make positive changes in their lives.

Non-Discrimination Policy: The University of Texas at Arlington does not discriminate on the basis of race, color, national origin, religion, age, gender, sexual orientation, disabilities, genetic information, and/or veteran status in its educational programs or activities it operates. For more information, visit uta.edu/eos.

Title IX Policy: The University of Texas at Arlington ("University") is committed to maintaining a learning and working environment that is free from discrimination based on sex in accordance with Title IX of the Higher Education Amendments of 1972 (Title IX), which prohibits discrimination on the basis of sex in educational programs or activities; Title VII of the Civil Rights Act of 1964 (Title VII), which prohibits sex discrimination in employment; and the Campus Sexual Violence Elimination Act (SaVE Act). Sexual misconduct is a form of sex discrimination and will not be tolerated. For information regarding Title IX, visit www.uta.edu/titleIX or contact Ms. Jean Hood, Vice President and Title IX Coordinator at (817) 272-7091 or jmhood@uta.edu.

Academic Integrity: Students enrolled all UT Arlington courses are expected to adhere to the UT Arlington Honor Code:

I pledge, on my honor, to uphold UT Arlington's tradition of academic integrity, a tradition that values hard work and honest effort in the pursuit of academic excellence.
I promise that I will submit only work that I personally create or contribute to group collaborations, and I will appropriately reference any work from other sources. I will follow the highest standards of integrity and uphold the spirit of the Honor Code.

UT Arlington faculty members may employ the Honor Code in their courses by having students acknowledge the honor code as part of an examination or requiring students to incorporate the honor code into any work submitted. Per UT System Regents' Rule 50101, §2.2, suspected violations of university's standards for academic integrity (including the Honor Code) will be referred to the Office of Student Conduct. Violators will be disciplined in accordance with University policy, which may result in the student's suspension or expulsion from the University. Additional information is available at https://www.uta.edu/conduct/.

Electronic Communication: UT Arlington has adopted MavMail as its official means to communicate with students about important deadlines and events, as well as to transact university-related business regarding financial aid, tuition, grades, graduation, etc. All students are assigned a MavMail account and are responsible for checking the inbox regularly. There is no additional charge to students for using this account, which remains active even after graduation. Information about activating and using MavMail is available at http://www.uta.edu/oit/cs/email/mavmail.php.

Campus Carry: Effective August 1, 2016, the Campus Carry law (Senate Bill 11) allows those licensed individuals to carry a concealed handgun in buildings on public university campuses, except in locations the University establishes as prohibited. Under the new law, openly carrying handguns is not allowed on college campuses. For more information, visit http://www.uta.edu/news/info/campus-carry/.

Student Feedback Survey: At the end of each term, students enrolled in face-to-face and online classes categorized as "lecture," "seminar," or "laboratory" are directed to complete an online Student Feedback Survey (SFS). Instructions on how to access the SFS for this course will be sent directly to each student through MavMail approximately 10 days before the end of the term. Each student's feedback via the SFS database is aggregated with that of other students enrolled in the course. Students' anonymity will be protected to the extent that the law allows. UT Arlington's effort to solicit, gather, tabulate, and publish student feedback is required by state law and aggregate

results are posted online. Data from SFS is also used for faculty and program evaluations. For more information, visit http://www.uta.edu/sfs.

Final Review Week: For semester-long courses, a period of five class days prior to the first day of final examinations in the long sessions shall be designated as Final Review Week. The purpose of this week is to allow students sufficient time to prepare for final examinations. During this week, there shall be no scheduled activities such as required field trips or performances; and no instructor shall assign any themes, research problems or exercises of similar scope that have a completion date during or following this week unless specified in the class syllabus. During Final Review Week, an instructor shall not give any examinations constituting 10% or more of the final grade, except makeup tests and laboratory examinations. In addition, no instructor shall give any portion of the final examination during Final Review Week. During this week, classes are held as scheduled. In addition, instructors are not required to limit content to topics that have been previously covered; they may introduce new concepts as appropriate.

Emergency Exit Procedures: Should we experience an emergency event that requires us to vacate the building, students should exit the room and move toward the nearest exit. When exiting the building during an emergency, one should never take an elevator but should use the stairwells. Faculty members and instructional staff will assist students in selecting the safest route for evacuation and will make arrangements to assist individuals with disabilities.

Student Support Services: UT Arlington provides a variety of resources and programs designed to help students develop academic skills, deal with personal situations, and better understand concepts and information related to their courses. Resources include tutoring, major-based learning centers, developmental education, advising and mentoring, personal counseling, and federally funded programs. For individualized referrals, students may visit the reception desk at University College (Ransom Hall), call the Maverick Resource Hotline at 817-272-6107, send a message to resources@uta.edu, or view the information at http://www.uta.edu/universitycollege/resources/index.php.

Schedule

As the instructor for this course, I reserve the right to adjust this schedule in any way that serves the educational needs of the students enrolled in this course.

University calendar: Fall 2016

Date	#	Lecture	Assignment		Lecture Notes	Required Reading
			Out	Due		
08/26	1	Course Overview			[PDF]	
Overview, Data, and Text						
09/02	2	Modeling Text Data: vector space model, search engine			[PDF]	
09/02	3	Modeling Text Data: vector space model, search engine				
09/09	4	The Life-Cycle of Data: data types, data extraction, curation, integration, wrangling, retrieval, mining	P1			
09/09	5	Data Mining, Big Data, Data Science, Applications, Tools, Datasets	HW1			
09/16	6	Similarity Measures				
Research and Application						
09/16	7	Computational Journalism (guest lecture)				
Classification and Prediction						
09/23	8	Decision Tree				
09/23	9	Decision Tree				
09/30	10	Bayesian Classifiers	HW2	HW1		
09/30	11	Bayesian Classifiers				
10/07	12	Support Vector Machine, Nearest Neighbor Classifiers				
10/07	13	Text Mining: classification				
10/14	14	Evaluating Classification	P2	P1		

		Models				
10/14	15	Evaluating Classification Models				
10/21	16	Evaluating Classification Models				
Web and Graph Mining, Large-Scale Data Processing						
10/21	17	Web Mining: link analysis (PageRank)				
10/28	18	Web Mining: link analysis (PageRank)				
10/28	19	Large-Scale Data Processing (MapReduce)				
11/04	20	Large-Scale Data Processing (MapReduce)	HW3	HW2		
11/02	Last day to drop class					
Clustering						
11/04	21	Overview of Clustering				
11/11	22	K-means	P3	P2		
11/11	23	Hierarchical clustering				
11/18	24	Text Mining: clustering				
Frequent Pattern and Association Rule Mining						
11/18	25	Association Rule Mining				
11/24	Thanksgiving					
12/02	26	Correlation Analysis				
12/02	27	Computational Journalism		HW3		
Research and Application: Computational Journalism						
12/06				P3		
12/?	Final Exam (Time and Location TBD)					