

Fall 2009 CSE4392 / 5334 Data Mining

Resources: [Google](#) [Google Scholar](#) [CiteSeer](#) [DBLP Bibliography](#) [ACM Digital Library](#) [IEEE Xplore](#) [Other Computer Science articles](#)

TA:

Course Information:

- Time: Mon./Wed. 5:30-6:50pm
- Classroom: [Nedderman Hall](#) 229
- Class number: 85677 (CSE4392) / 84603 (CSE5334)
- Homepage: <http://crystal.uta.edu/~cli/cse5334>

Instructor: [Chengkai Li](#)

- Office hours: Mon./Wed. 4:30pm-5:30pm
- Office: [Nedderman Hall](#) 334
- Phone: (817) 272-0162
- E-mail: [cli \[AT\] uta \[DOT\] edu](mailto:cli@uta.edu)
- Homepage: <http://ranger.uta.edu/~cli>

- Office hours:
- Office:
- Phone:
- E-mail: [\[AT\] mavs \[DOT\] uta \[DOT\] edu](mailto:mavs@uta.edu)
- Homepage:

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 and text.

Prerequisites: CSE 3330/5330 Database Systems I or CSE 4331/5331 Database Systems II or similar courses or consent of instructor

Textbook

- **(Required)** Jiawei Han and Micheline Kamber. [Data Mining: Concepts and Techniques](#), 2nd ed., Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6.
- (Reference) Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Addison-Wesley, 2006. ISBN 0-321-32136-7.
- (Reference) Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, [Introduction to Information Retrieval](#), Cambridge University Press. 2008. (This book is available online at <http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html>)
- (Reference) I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005.
- (Reference) T. M. Mitchell, Machine Learning, McGraw Hill, 1997.

Grades

- | | | |
|------------------|-----|---|
| • Midterm Exam | 20% | (Monday, Oct. 12th, 5:30pm-6:50pm, NH229) |
| • Final Exam | 35% | (Monday, Dec. 7th, 5:30pm-8pm, NH229) |
| • Homework (HW) | 20% | |
| • Course Project | 25% | |

- **Bonus Points** 5% Based on in-class and WebCT participation. Students are expected to attend classes and actively participate in discussions.
-

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

Assignments and Deadlines

- All the assignments must be submitted through WebCT. We will NOT take hardcopy or email submission, unless the school verifies that WebCT was malfunctioning or unavailable.
- Everything is due by 11:55pm on the due date (listed in the [schedule page](#)). The deadline is automatically managed by WebCT. 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 from the school.
- Before the deadline, WebCT allows you to submit as many times as you want. However, you are only allowed to submit once after the deadline. If you have already submitted before the deadline, you will not be allowed to submit again after the deadline.

Regrading: Regrading request must be made within 7 days after we post scores on WebCT. 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.

WebCT: Log in to the WebCT page <http://www.uta.edu/webct> with your NetID and password. We use WebCT for: (1) Announcements; (2) Assignment Submission; (3) Discussion Group; (4) Releasing materials, assignments, scores and grades. Follow these steps exactly during electronic assignment submission.

1. Click button "Upload file" to upload your file.
 2. Fill in your email address (UTA email address only) in the "Notification" box.
 3. Then you must click button "submit assignment". Otherwise, your file will not be submitted.
 4. Verify that your file is indeed submitted into WebCT. (You should see the file name after "Student files". Click the link to download the file and verify it.)
 5. Check your email. You must keep the notification email from WebCT.
 6. If you don't find your submission or don't receive notification within 10 minutes, try step 1-5 again.
 7. If step 6 still fails after you give it another try, email your file to the TA and yourself immediately.
 8. When we grade the assignment, if we cannot find your file in WebCT and we don't receive it in email, we will ask you for the copy of notification email. Without the notification email, we would assume that you didn't submit the assignment.
-

Ethics Policies and Academic Integrity: The College cannot and will not tolerate any form of academic dishonesty by its students. This includes, but is not limited to cheating on examinations, plagiarism, or collusion (explained in the document below). Students are required to read the following document carefully, sign it, return the signed copy to the instructor, and keep a copy for their own records. Hardcopies of this document will be provided to the students in the first class, and also can be picked up in the instructor's office. If you print by yourself, please make it double-sided.

[Statement on Ethics, Professionalism, and Conduct for Engineering Students](#)

Miscellaneous: If you require accommodation based on disability, I would like to meet with you in the privacy of my office during the first week of the semester to ensure that you are appropriately accommodated. Please read the page of [the office for students with disabilities](#).

Schedule:

Date	#	Lecture	Assignment		Lecture Notes
			Out	Due	
08/24	1	Course Overview			
08/26	2	Introduction (Chapter 1)			
08/31	3	Course Project Topics			
Data Warehousing, OLAP, Data Cube (Chapter 3, 4)					
09/02	4	Data Warehousing and OLAP	HW1		
09/07		No Class. Labor Day Holiday.			
09/09	5	Data Cube			
Data Preprocessing (Chapter 2)					
09/14	6	Data and Data Quality			
09/16	7	Data Preprocessing		HW1	
Classification and Prediction (Chapter 6)					
09/21	8	Decision Tree	HW2		
09/23	9	Evaluating Classification Models			
09/28	10	Bayesian Classifiers			
09/30	11	Rule-based			
10/05	12	Nearest Neighbor Classifiers			
10/07	13	Support Vector Machine		HW2	
10/12		Midterm Exam (Monday, Oct. 12th, 5:30pm-6:50pm, NH229)			
Frequent Pattern and Association Rule Mining (Chapter 5)					
10/14	14	Association Rule Mining			
10/19	15	Correlation Analysis			
Clustering (Chapter 7)					
10/21	16	Overview of Clustering, Similarity/Dissimilarity Measure	HW3		

10/26	17	K-means			
10/28	18	Hierarchical			
11/02	19	Hierarchical (cont'd)			
Text and Web Mining					
11/04	20	Document Classification and Clustering			
11/09	21	Latent Semantic Indexing		HW3	
11/11	22	Information Extraction: Wrapper Induction	HW4		
11/16	23	MapReduce			
11/18	24	Link Analysis: PageRank			
11/23	25	Link Analysis (cont'd)			
11/25	26	Social Network Analysis		HW4	
11/30	27	overflow lecture			
12/02		Project Demo (Time and Location: TBD)			
12/02	28	Final Review			
12/07		Final Exam (Monday, Dec. 7th, 5:30pm-8pm, NH22)			

University calendar: [Fall 2009](#)