

i&gt;¿

# Fall 2011 CSE4334 / 5334 Data Mining

## Course

### Information:

**Instructor:** [Chengkai Li](#)

**TA:** Saravanan Thirumuruganathan

- |  |  |  |
|--|--|--|
| • Time: Tue/Thu<br>9:30-10:50pm  | • Office hours: Tue/Thu<br>11am - 1pm  |  |
| • Classroom: ERB 130   | • Office: ERB 652  | • Office Hours: Friday 10am-12pm                   |
| • Class number: 84814<br>(CSE4334) / 82252<br>(CSE5334)  | • Phone: (817) 272-0162  | • Office: ERB 504                                  |
| • Homepage:<br><a href="http://crystal.uta.edu/~cli/cse5334">http://crystal.uta.edu/~cli/cse5334</a> | • E-mail: <a href="mailto:cli@uta.edu">cli [AT] uta [DOT] edu</a>                  | • Phone: 817-201-5046                              |
|  | • Homepage:<br><a href="http://ranger.uta.edu/~cli">http://ranger.uta.edu/~cli</a> | • E-mail:<br>saravanan.thirumuruganathan@gmail.com |

**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% | (Thursday, Oct. 13th, 9:30am-10:50am, ERB130) |
| • Final Exam     | 35% | (Thursday, Dec. 15th, 8am-10:30pm, ERB 130)   |
| • Homework (HW)  | 15% | (Must be done independently)                  |
| • Course Project | 30% | (Must be done independently)                  |

**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 school verifies that Blackboard was malfunctioning or unavailable.
- 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 from the school.

**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.

**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	
08/25	1	Course Overview			<a href="#">[PPT]</a>
08/30	2	Introduction (Chapter 1)			<a href="#">[PPT]</a>
09/01	3	Prominent Streak Discovery	HW1		<a href="#">[PPT]</a>
Data Warehousing, OLAP, Data Cube (Chapter 3, 4)					
09/06	4	Data Warehousing and OLAP			
09/08	5	Data Cube			<a href="#">[PDF]</a>
09/13	6	Set Query			
Classification and Prediction (Chapter 6)					
09/15	7	Decision Tree	P1	HW1	<a href="#">[PDF]</a>

09/20	8	Decision Tree (cont'd)			
09/22	9	Evaluating Classification Models			<a href="#">[PDF]</a>
09/27	10	Evaluating Classification Models (cont'd)			<a href="#">[PDF]</a>
09/29	11	Bayesian Classifiers	HW2		<a href="#">[PDF]</a>
10/04	12	Nearest Neighbor Classifiers			<a href="#">[PDF]</a> <a href="#">[PPT]</a>
10/06	13	Support Vector Machine			<a href="#">[PDF]</a> <a href="#">[PPT]</a>
Frequent Pattern and Association Rule Mining (Chapter 5)					
10/11	14	Association Rule Mining		HW2	
10/13		Midterm Exam (Thursday, Oct. 13th, 9:30am-10:50am, ERB130)			
10/18	15	Correlation Analysis			<a href="#">[PDF]</a> <a href="#">[PPT]</a>
Clustering (Chapter 7)					
10/20	16	Overview of Clustering, Similarity/Dissimilarity Measure		P1	
10/25	17	K-means			<a href="#">[PDF]</a> <a href="#">[PPT]</a>
10/27	18	K-means (cont'd)			
11/01	19	Hierarchical clustering			<a href="#">[PDF]</a> <a href="#">[PPT]</a>
11/03	20	Hierarchical clustering (cont'd)	P2, HW3		
Text and Web Mining					
11/08	21	Vector Space Model			<a href="#">[PDF]</a>
11/10	22	Document Classification			<a href="#">[PDF]</a>
11/15	23	Document Clustering			<a href="#">[PDF]</a>
11/17	24	MapReduce	P3	HW3	<a href="#">[PDF]</a> <a href="#">[PPT]</a>
11/22	25	MapReduce		P2	
11/24		Thanksgiving Holidays			
11/29	26	MapReduce			
12/01	27	Link Analysis: PageRank			<a href="#">[PDF]</a>

12/06	28	Link Analysis (cont'd)			
12/08	29	Final Review		P3	<a href="#">[PDF]</a>
12/15		Final Exam (Thursday, Dec. 15th, 8am-10:30pm, ERB 130)			