

CSE5334 DATA MINING

Lecture 1: Introduction

CSE4392/5334 Data Mining, Fall 2009
Department of Computer Science and Engineering, University of Texas at Arlington
©Chengkai Li, 2009

Self Introduction

- Chengkai Li
- <http://ranger.uta.edu/~cli>
- Research interests:
databases, data mining, information retrieval, Web
- Looking for students
Master/PhD project/thesis topics available.

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 2

My Research

[Fall09orientation.ppt](#)

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 3

Now it's your turn

- Name, program/year, where from
- Areas of interest
- Course taken, skills/experiences related to this course
- Why do you want to take this course?
- What do you want to get from the course?
- What would make you like/hate this course?
- Anything else

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 4

Course Page

- <http://crystal.uta.edu/~cli/cse5334>
 - Syllabus, Schedule (lecture notes), Resources, Accommodation based on disability.
- [WebCT](#)
 - Announcement (check it on a daily basis)
 - HW/Project Submissions
 - Grades
 - Discussion Board

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 5

Basics

- **Lectures:** Mon/Wed , 5:30pm-6:50pm, NH 229
- **Instructor:** Chengkai Li
Office hours: Mon/Wed 4:30pm-5:30pm, NH334
Contact: cli [at] uta.edu, (817) 272-0162 (I am not in office everyday, and do not check voice mail. Please reach me in office hours, or by emails.)
- **TA:** Xiaonan (Michael) Li
Office hours: TBA
Contact: xiaonan.li [at] mavs.uta.edu, Phone number TBA

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 6

Textbook

Required Textbook:

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.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, [Introduction to Information Retrieval](http://www-csli.stanford.edu/~hinrich/introduction-retrieval-book.html), Cambridge University Press, 2008. (This book is available online at <http://www-csli.stanford.edu/~hinrich/introduction-retrieval-book.html>)
- I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 2nd ed. 2005.
- T. M. Mitchell, *Machine Learning*, McGraw Hill, 1997.

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington @Chengkai Li, 2009 7

The slides

- The slides highlight the gist of the most important concepts and techniques.
- But
 - It is not meant to be complete. Details may not be included.
 - It may be simplified for ease of explanation.
- You won't do well in the course if you just read the slides
 - You need to read the book and study the slides carefully.
- Some lecture notes are adopted from:
 - Jiawei Han (UIUC)
 - Vipin Kumar (Minnesota)

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington @Chengkai Li, 2009 8

Tentative Grading Scheme

□ Midterm	20%	
□ Final	35%	
□ Homework (HW)	20%	
□ Course Project	25%	(Must be done individually)
□ Bonus Points	5%	

You are required to attend classes and actively participate in discussions (both in-class and WebCT).

Final Letter Grade:

- No pre-defined cutoffs. Will be based on bell curve of your performance.
- Undergraduate and graduate students are compared in separate groups.
- This is the first time this course is offered with a 4392 section.

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington @Chengkai Li, 2009 9

Homework (HW)

- Problem solving
- Focus on most important topics
- Lighter load

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington @Chengkai Li, 2009 10

Course Project (P1-P4)

- Multiple (tentatively 4) Stages
- More hands-on experience
- Programming Assignments
- Mainly implementation
- Some small open-ended problems

- Alternatively, students are encouraged to do a semester-long open-ended research project. Talk to me if you prefer this.

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington @Chengkai Li, 2009 11

Bonus points: class participation

- Class participation (5%)
 - In-class discussion
 - WebCT discussion group
 - The instructor and TA will post discussion topics from time to time.
 - You are highly encouraged to initiate your own thread.

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington @Chengkai Li, 2009 12

WebCT

- Assignment instruction and files
- Student assignment submission (we don't accept email submission or hard-copy)
- Discussion Group
- Grades

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 13

Deadlines

- Everything will be submitted through WebCT.
- Due time: 11:55pm
- Late submission: 5-point deduction per hour, till you get 0. (The raw score of each assignment is 100. So there is no point to submit it after 20 hours).

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 14

How to Submit through WebCT

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.

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 15

Regrading

- 7 days after we post scores on WebCT. TA will handle regrade requests. Won't consider it after 7 days.
- If not satisfied with the results, 7 days to request again. Instructor will handle it, and the decision is final.

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 16

Topics in Textbook

Part 1: Introduction

- Data Preprocessing
- Data Warehouse and OLAP Technology: An Introduction
- Advanced Data Cube Technology and Data Generalization
- Mining Frequent Patterns, Association and Correlations
- Classification and Prediction
- Cluster Analysis

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 17

Topics in Textbook

Part 2: Advanced Applications and Current Research

- Mining data streams, time-series, and sequence data
- Mining graphs, **social networks** and multi-relational data
- Mining object, spatial, multimedia, text and Web data
 - Mining complex data objects
 - Spatial and spatiotemporal data mining
 - Multimedia data mining
 - **Text mining**
 - **Web mining**
- Applications and trends of data mining
 - Mining business & biological data
 - Visual data mining
 - Data mining and society: Privacy-preserving data mining
- **Additional (often current) themes could be added to the course**

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 18

Schedule

- <http://crystal.uta.edu/~cli/cse5334>

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 19

Conferences and Journals on Data Mining

- KDD Conferences
 - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
 - SIAM Data Mining Conf. (SDM)
 - (IEEE) Int. Conf. on Data Mining (ICDM)
 - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
 - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Other related conferences
 - ACM SIGMOD
 - VLDB
 - (IEEE) ICDE
 - WWW, SIGIR
 - ICML, CVPR, NIPS
- Journals
 - Data Mining and Knowledge Discovery (DAMI or DMKD)
 - IEEE Trans. On Knowledge and Data Eng. (TKDE)
 - KDD Explorations
 - ACM Trans. on KDD

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 20

Where to find papers

- [Google](#)
 - [Google Scholar](#)
 - [DBLP Bibliography](#)
 - [CiteSeer](#)
 - Services through UTA Library
- <http://library.uta.edu/JDBC/DBs/dbejournal.jsp>
- [ACM Digital Library](#)
 - [IEEE Xplore](#)
 - [Other Computer Science articles](#)

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 21

Get bored?

- Do you watch Youtube?

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 22

<http://www.youtube.com/watch?v=gCzew6qLa8U>

<http://www.youtube.com/watch?v=463gKcXDvzQ>

Don't do it. It's not worth it.

We are very serious about this.

read & sign the statement

Lecture 1: Introduction CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 23