Lecture 1: Introduction

CSE 4334/5334 Data Mining

Department of Computer Science and Engineering, University of Texas at Arlington
©Chengkai Li, 2014
Self Introduction

- Naeemul Hassan
- http://idir.uta.edu/~naeemul/

Research interests:
- Database Systems
- Data Mining
- Computational Journalism
My Research

Research Overview

- Skyline Group
- Computational Journalism
- Crowdsourcing
Now it’s your turn

- Name, program
- Prior courses/experiences related to this subject
- What make you decide to take this course?
- What will make you like/hate this course?
- Anything else
Course Page

- [http://idir.uta.edu/~naeemul/cse4334/](http://idir.uta.edu/~naeemul/cse4334/)
  - Syllabus, Schedule (lecture notes), Resources, Accommodation based on disability.

- [Blackboard](http://idir.uta.edu/~naeemul/cse4334/)
  - Announcement (check it on a daily basis)
Basics

- **Lectures:** Tue/Thu, 2-3:20pm, WH 308
- **Instructor:** Naeemul Hassan
  
  Office hours: Tue/Thu 10:00am-12:00pm, ERB 509
  
  Contact: naeemul DOT hassan AT mavs DOT uta DOT edu, (817) 437-4518
  (I do not check voicemails regularly.)

- **TA:** TBD
  
  Office hours: TBD
  
  Email: TBD
Textbook

- **Required Textbook:**
  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:**
The slides highlight the gist of 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.

Studying only the slides is not enough.
- You need to read the book and study the slides carefully.

Many lecture notes are adopted from:
- Jiawei Han (Illinois)
- Vipin Kumar (Minnesota)
Tentative Grading Scheme

- Midterm: 20%
- Final: 30%
- Homework (HW): 20% (Must be done independently)
- Course Project: 30% (Must be done independently)

You are required to attend classes and actively participate in discussions.

- 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.
Homework (HW)

- Problem solving
- Focus on most important topics
- HW1, HW2, HW3, HW4, 5% each
Course Project

- 2 Programming Assignments, 15% each
  - hands-on experience with **big data**, real application
  - Must design, implement (programming), and evaluate
  - open to novel solutions
Blackboard

- Assignment instruction and files
- Submission (we don’t accept email submission or hard-copy)
- Grades
- Questions, Discussion Group
Deadlines

- Everything will be submitted through Blackboard.
- Due time: 11:59pm
- 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).
Regrading

- 7 days after we post scores on Blackboard. 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.
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
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 themes (prominent streak discovery, skyline group, significant fact finding)
Schedule

- http://idir.uta.edu/~naeemul/cse4334/
Your Email

- Make sure your MavMail works. We will only contact you by your MavMail.
- Check it on a daily basis.
Academic Integrity

- **Cheating**
  - Copying another's test or assignment
  - Communication with another during an exam or assignment (i.e. written, oral or otherwise)
  - Giving or seeking aid from another when not permitted by the instructor
  - Possessing or using unauthorized materials during the test
  - Buying, using, stealing, transporting, or soliciting a test, draft of a test, or answer key
Academic Integrity

- **Plagiarism**
  - Using someone else's work in your assignment without appropriate acknowledgement
  - Making slight variations in the language and then failing to give credit to the source
Collusion

Without authorization, collaborating with another when preparing an assignment.
Question? 😊