Syllabus

Course Description

Course description

Fundamental concepts and techniques of data mining, including data attributes, data visualization, data pre-processing, mining frequent patterns, association and correlation, classification methods, and cluster analysis. Advanced topics (time permitting) include outlier detection, stream mining, and social media data mining.

If an undergraduate student is registered for CSC 440 (as an advanced version of CSC 240, the undergraduate-level course), she or he needs to meet the same requirements as CSC 440.

Prerequisites: MTH161, MTH 165, CSC171, CSC 172. Some knowledge of artificial intelligence (CSC 242) or probability theory (CSC 262) will be helpful.

Course schedule (tentative, chapters refer to the textbook)

Data Mining: Concepts and Techniques, 3/E
Jiawei Han, Micheline, Kamber, and **Jian Pei**

- **Publisher:** Morgan Kaufmann, 2011
- **ISBN-10:** 0123814790
- **ISBN-13:** 978-0123814791

- Overview and Introduction	(lecture notes, Chap. 1)
- Getting to Know Your Data	(Chap. 2)
- Data Preprocessing	(Chap. 3)
- Review of Linear Algebra, Statistics	(lecture notes)
- Pattern Recognition Concepts	(lecture notes)
- Mining Frequent Patterns	(Chap. 6)
- Association and Correlation	(Chap. 6)
- Advanced Pattern Mining	(Chap. 7)
- Classification	(Chap. 8/9*)
- Cluster Analysis	(Chap. 10/11*)
- Outlier Detection	(Chap. 12)
- Advanced Topics: Social Multimedia Mining	(lecture notes)
- Advanced Topics: Biomedical Informatics	(Guest Lecture: tentatively Prof. Martin Zand of URMC)
- Advanced Topics: Network Mining	(Guest Lecture: tentatively Prof. Gourab Ghoshal or Prof. Gonzalo Mateos)
- Advanced Topics: Influence Mining	(Guest Lecture: TBD*)
- Trends and Research Frontiers	(Chap. 13, lecture notes)

*** Time Permitting**

*** Midterm Exam: October 31**

*** Course Project presentation: 10 min. pp, December 5, 10, 12**

Instructor and grading

Instructor: Prof. Jiebo Luo, Wegmans Hall Rm 3101, x65784
Lectures: TR 15:25-16:40, Gavett Hall 202
Office hours: before classes (13:00-15:00) or by appointment (use cs email)
TA: Haitian Zheng, Wei Zhu, office hours: M/W, 1-3pm, Computer Science VISTA Lab (3504 Wegmans Hall)

Grading (total 100%)

- homework assignments 35% (5% for each of the 5 assignments, plus a small project 10%)
- midterm 30%
- final project & presentation 30% (presentation counts 10%)
- class participation/effort 5%

Expectation for the final project - something "new"

- an existing algorithm applied to new data or new problems
- a new algorithm (or a modified version of an existing algorithm) applied to the same data
- new findings from a comparative study of using different algorithms for the same problem

* note: both the small project and the final project require **programming (Python recommended)**

Textbooks and other resources

- Required textbook

Data Mining: Concepts and Techniques, 3/E
Jiawei Han, **Micheline, Kamber**, and **Jian Pei**

- **Publisher:** Morgan Kaufmann, 2011
- **ISBN-10:** 0123814790
- **ISBN-13:** 978-0123814791

- Recommended reference book

Social Media Modeling and Computing
Steven C.H. Hoi, Jiebo Luo, Susanne Boll, Dong Xu, Rong Jin, Irwin King

- **Publisher:** Springer, 2011
- **ISBN-10:** 0857294350
- **ISBN-13:** 978-0857294357

Mining of Massive Datasets, 2/E
Jure Leskovec, Anand Rajaraman, Jeffery David Ullman

- **Publisher:** Cambridge University Press, 2011
- **ISBN-13:** 978-1107077232