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.

Health and Safety Statement

The University is committed to protecting the health and safety of the entire community – students, faculty, and staff. For this reason, it is mandatory that everyone wears a mask in University buildings and observe appropriate social distancing, including in classrooms. Masks have been provided to students, faculty, and staff and classrooms have been specifically assigned to allow for social distancing to support these requirements. **You must wear a mask appropriately (e.g. over the nose and mouth) if you are attending class in person, and you must do this for every class session and for the entire duration of each class session**. If you fail to do this, you will be politely reminded of the requirement and then asked to leave if you do not comply. Students who refuse to adhere to the requirement for mask wearing or social distancing in the course will be in violation of the COVID-19 Community Commitment and will be referred to the Student Conduct system through a COVID-19 Concern Report. Such referrals will lead to student conduct hearings and may result in disciplinary action.

Students who feel unable to wear a mask may contact the Office of Disability Resources to explore options for accommodations. Students requiring accommodations may be asked to participate in the course through synchronous or asynchronous learning as part of this accommodation.

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: Health Analytics	(Guest Lecture: tentatively Prof. Tim Dye 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 27 (15:25-17:05 for 100 minutes) *only pre-approved students can take the exam at an alternative time or location**
*** Course Project presentation: 10 min. pp, December 1, 6, and 8**

Instructor and grading

Instructor: Prof. Jiebo Luo, Wegmans Hall Rm 3101, x65784
Lectures: TR 15:25-16:40, Gavett Hall Room 202 (in-person only). The lecture notes will be available on Blackboard shortly after each class.
Office hours: before classes (14:00-15:00) or by appointment (use cs email)
TA: Hanjia Lyu, Junyu Chen, office hours (tentatively): M/W, 1-3pm, Computer Science VISTA Lab (3504 Wegmans Hall)

The midterm exam will be given in class only. The answers to the exam problems will need to be handwritten (unless for disability accommodation).

Grading (total 100%)

- homework assignments 35% (5% for each of the 5 assignments, plus a small programming project 10%, **programming can NOT be done in R**)
- midterm 30%
- final project & presentation 30% (presentation counts 10%)
- class participation/effort 5%

The 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, while other programming language is allowed)**

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