VE484: Data Mining

Prof. Qiang (Shawn) Cheng

UM-SJTU Joint Institute

Summer Term 2016

Syllabus

Lectures. Dr. Qiang Cheng

- Tu Th F(odd week) 8:00 – 9:40am, F410 (Dong Xia Yuan)
- Attendance is required.
- Office hours: TuTh: 10:00am-11:30pm,
  Location: Michigan Institute 210.
  Other times, appointment needed.
- Email: qcheng888@yahoo.com

TA: Dai Tao

- Office hours: Wed. 20:00-22:00
- Office: Yu Liming Center, Mobile: 13122113552,
  Email: suafeng@sjtu.edu.cn
- Grades home works and/or tests, holds office hours, and answers questions regarding the homework and grading.
- TA is a student who needs to study, do thesis…

Background and Books

Prereq. Data structure, or some Programming courses, or instructor's permission (see me after class).

Textbook. 1. *Mining of Massive Datasets*, (downloadable from the internet), by Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, 2014.

2. Data Mining and Analysis-Fundamental Concepts and Algorithms, (downloadable from the internet at http://www.cs.rpi.edu/~zaki/PaperDir/DMABOOK.pdf), by Mohammed Zaki and Wagner Meira Jr., 2014.

Recommended books for references:

1. Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann,* 2011. (http://web.engr.illinois.edu/~hanj/bk2/toc.pdf)
2. Pang-NingTan, Michael Steinbach ,and Vipin Kumar, *Introduction to Data Mining*, 2nd edition, Addison, 2006. (http://www-users.cs.umn.edu/~kumar/dmbook/index.php)

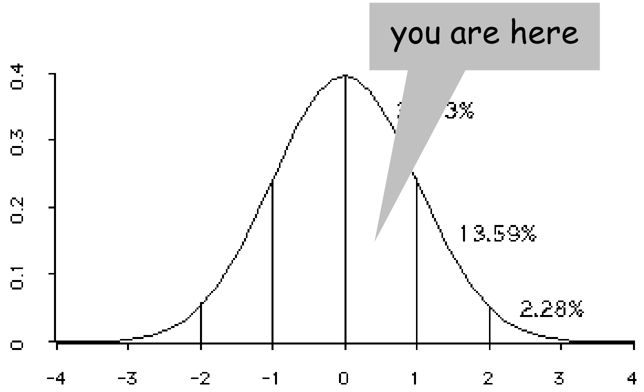
Grades

Grading.

- "Weekly" or biweekly problem sets, due time: 8:00am, the firstThur after one week (or two weeks for lab/machine problems) of the assignment date.
  No late homework will be accepted.
- Class participation, staff discretion for borderline cases.
- Optional in-class presentation: encouraged, ~20-30min, on applications, problems, techniques, or results related to class materials.
  May lead to bonus points (up to 10%).
- Three quizzes, one midterm exam, one final (time: TBD)
- Grade determination: tests 30%+40%, HW: 20%, quizzes: 10%
- Letter grade: 1) above 90%, then definitely A;
  2) otherwise, depending on the standing in the class

subject to change

Course grades.



Collaboration

Collaboration policy. (ask if unsure)

- Course materials are always permitted.
- You are encouraged to attend office hours as needed.
- External resources are encouraged, e.g., Google, Yahoo.

"Collaboration permitted" problem sets.

- Default permission level, unless otherwise stated.
- Can form study group of up to 3 students.
- Study group may work on problems jointly, but you must write up solutions individually.

"No collaboration" problem sets.

- Can always consult course staff.

You need “independently” work out problems in tests and quizzes:

- No text book or class note is permitted.
- No other book is permitted.

Overview of Contents

1. Introduction to data mining
2. Review of linear algebra and introduction to Matlab
3. Know your data
4. Data preprocessing
5. Graph data
6. High-dimensional data
7. Kernel methods
8. Dimensionality reduction
9. Sequence mining
10. Graph mining
11. Clustering: basic methods and more advance methods
12. Classification and regression:basic methods and more advanced methods, including linear discriminant analysis, decision trees, support vector machines, etc.

Note: The focus will be on fundamental techniques and basic skills. Some advanced materials may be omitted (left to more advanced courses or discussion with instructor)