COMP7118: Data Mining

Xiaofei Zhang

Fall, 2019

E-mail: xiaofei.zhang@memphis.edu
Office Hours: 11:30-12:30 Tu
Office: DH318

Web: eCourseware
Class Hours: 17:30–18:55 TTh
Class Room: Psychology Bldg 362

Course Description

Data mining has emerged as a major frontier field of study in recent years. Aimed at extracting useful and interesting patterns and knowledge from large data repositories such as databases and the Web, the field of data mining integrates techniques from database, statistics and artificial intelligence. This course will provide a broad overview of the field and focus on a series of advanced topics. The following topics will be covered:

- Knowledge discovery in databases (association rule, clustering, classification, data warehouses)
- Text mining (topic modeling, word embedding, computing journalism)
- Graph mining (PageRank, frequent graph patterns, summarization, linkage prediction)

Recommended Materials

- Data Mining: Concepts and Techniques. Jiawei Han, Micheline Kamber and Jian Pei : Morgan Kaufmann Publishers (3rd edition)
- Data Mining: The Textbook. Charu C. Aggarwal. Springer.

Prerequisites

The official pre-requisite of the course is COMP 3160. However, as Data Mining is a diverse field, it draws on different aspects of the knowledge in fields such as Databases, Artificial Intelligence, Statistics and Management. The following is a checklist of material that will be used in the course. If you do not know all of them, don't worry, but do try to read up on your own.

- Basic computer algorithms (COMP 4030)
- Undergraduate level statistics/probability (ISDS 2710/MATH 4611)
- Database systems (COMP 7115/ISDS 7605)

Programming skills

- Familiar with any of {C,C++,Java,Python,Matlab}
- UNIX skills will be helpful (but not necessary)

Assessments & Grading

- 3 assignments: weight 30%
- Project: weight 40 %
  - Proposal: 5 %
  - Implementation: 20 %
  - Report: 10 %
  - Presentation: 10 %
- Paper presentation: weight 10%
- Final: weight 15%

Note: A list of suggested projects will be provided for inspiration. Students are encouraged to propose their own projects. A proposal is required to justify its motivation, feasibility, and deliverable outcome.

Grading Scale

We will calculate final letter grades in two different ways; then each student will receive the higher of the two letter grades. One way is a fixed grading scale, with the following cutoffs:

$A \geq 90\%$ $A- \geq 82\%$ $B+ \geq 74\%$ $B \geq 66\%$ $B- \geq 58\%$ $C+ \geq 50\%$ $C \geq 42\%$

The other way is a curve, with the following percentages of students receiving each grade:

$A : 18\%$ $A- : 18\%$ $B+ : 18\%$ $B : 18\%$ $B- : 18\%$ $C+ : 5\%$ $C : 5\%$

However, we will feel free to give an F to any student who clearly did not put effort into the course (or an A+ to any student with truly exceptional performance).

Course Policies

Testing Policy

No early or late exams are allowed unless under extreme situations.

Plagiarism/Cheating Policy

Plagiarism or cheating behavior in any form is unethical and detrimental to proper education and will not be tolerated. All work submitted by a student (projects, programming assignments, lab assignments, quizzes, tests, etc.) is expected to be a student's own work. The plagiarism is incurred when any part of anybody else's work is passed as your own (no proper credit is listed to the sources in your own work) so the reader is led to believe it is therefore your own effort. Students are allowed and encouraged to discuss with each other and look up resources in the literature, but appropriate references must be included for the materials consulted, and appropriate citations made when the material is taken verbatim.

If plagiarism or cheating occurs, the student will receive a failing grade on the assignment and (at the instructor's discretion) a failing grade in the course. The course instructor may also decide to forward the incident to the Office of Student Conduct for further disciplinary action. For further information on U of M code of student conduct and academic discipline procedures, please refer to: http://www.memphis.edu/studentconduct/misconduct.htm

Lecture Schedule (tentative)

Week	Topic	Highlights
1	No class	Math review and course overview will be provided
2	Association	Apriori FP-growth Pattern evaluation
3, 4	Clustering	Representative approach Hierarchical approach Probabilistic model-based approach High-dimensional clustering Outlier, statistical/distance/density model
5	Classification	Support Vector Machine Neural network Recurrent neural network
6, 7, 8, 9	Text Mining	Document Representation Topic modeling Applications
10, 11, 12	Graph Mining	PageRank Frequent graph pattern Graph summarization Linkage predication
13, 14	Project presentation	