CMPT 741: Data Mining

Fall 2024

Instructor: Ke Wang

Office: TASC1 9235

wangk@cs.sfu.ca

Teaching Materials

1. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, “Introduction to Data Mining”, Addison Wesley, 2006.
2. Anand Rajaraman, Jure Leskovec, and Jeffrey D. Ullman, “Mining of Massive Data Sets”.
3. Reference: Jiawei Han, Micheline Kamber, Jian Pei, ``Data Mining: Concepts and Techniques”
4. PPT Slides: SFU canvas Files

Course Information

- In-person class. Recording from Fall 2023 is available at Canvas/Pages
- sfu.canvas/Files: slides, assignments, quizzes, midterm, exam
- Grading
  - Weekly in-class quiz - 15%
  - Two assignments - 20%
  - Midterm – 25%
  - Final exam - 40%

Topics

- Data Mining
  - Introduction (1,4)
  - Classification (supervised learning) (1,4) (Assignment 1)
  - Clustering (unsupervised learning) (1,2,,4)
  - Association Rule Mining (1,3, 4) (Midterm)
- Big Data
  - Recommendation Systems (2,4)
  - Find Similar Sets (2,4) (Assignment 2)
  - Link Analysis (2,4)
  - Data Privacy and Security (4)
  - Applications (4) (Final)
- Numbers indicate the sources on the previous slide
- Helpful: knowledge on machine learning, database, statistics, algorithms, data structures.

Office Hours (online only)

- Instructor: Friday 10:00-11:00am, https://sfu.zoom.us/j/8360430493
- TA:
  - Ehsan Hoseinzade, ehoseinz@sfu.ca, Monday 7-8pm, https://sfu.zoom.us/j/6319855111
  - Anandharaju Durai Raju, aduraira@sfu.ca
- Start from week of Sept 16, 2024.

Marking, Submission, Grading

- Partial marking for minor mistakes, e.g., typos, etc. No mark for major mistakes, e.g., wrong concepts, not using concepts/methods from class, not readable answers, etc.
- Stick to given question description. No mark is given for adding new assumptions to simplify the question.
- Only the submissions received in the sfu.coursys before deadline will be marked.
- Grading is relative to the class performance.

Warnings

- Academic honesty policy
  - https://www.sfu.ca/students/academicintegrity/review.html. Zero tolerance for cheating and plagiarism
- Class attendances essential
  - Statistics show those attending class regularly do better. Q&A and additional examples in class.
- Seek help *early* through office hours, chat with TA and instructors frequently to confirm understanding, etc.

Useful links

- Data sets:

https://archive.ics.uci.edu/ml/datasets.php

- Software open sources: https://sourceforge.net/projects/weka/