Data Mining

CS57300, Spring 2019, Tuesday & Thursday 4:30-5:45pm, WANG 2599
Course email: datamining.purdue@gmail.com

Instructor

Ming Yin

Office: LWSN 2142B
Email: mingyin@purdue.edu
Office Hours: Wednesday 4-5pm

Teaching Assistants

Hao Ding, ding209@purdue.edu
Mahak Goindani, mgoindan@purdue.edu
Omkar S. Patil, patilo@purdue.edu
Office Hours: Monday 6-7pm (Mahak), Thursday 1-2pm (Omkar), Friday 3-4pm (Hao)
Office Hour Location: HAAS G50

Course Description

Data mining has emerged at the confluence of artificial intelligence, statistics, and databases as a technique for automatically discovering summary knowledge in large datasets. This course introduces students to the process and main techniques in data mining, including basic data visualization and exploratory analysis, predictive modeling, descriptive modeling, and pattern mining approaches. Data mining systems and applications will also be covered, along with selected topics in current research.

Learning Objectives

Upon completing the course, students should be able to:

- Understand the basic data mining process and identify key elements of data mining algorithms
- Recognize different types of data mining tasks
- Implement and apply basic algorithm and standard models
- Understand how to evaluate performance, as well as formulate and test hypotheses

Prerequisites

STAT516 or an equivalent introductory statistics course, CS 381 or an equivalent course that covers basic programming skills (e.g., STAT 598G).

Textbook

The primary text the class is:

- D. Hand, H. Mannila, P. Smyth (2001). *Principles of Data Mining*. MIT Press. (referred to as "PDM")

Course Schedule

The following schedule is tentative and subject to change.

Date	Topic	Reading
Jan 8	Introduction & Course Overview	
Jan 10	Background & Basics (Probability)	PDM Chapter 4.1-4.3, Appendix
Jan 15	Background & Basics (Linear algebra, sampling)	PDM Chapter 4.7
Jan 17	Background & Basics (Statistical inference, hypothesis testing)	PDM Chapter 4.4-4.6
Jan 22	Elements of Data Mining Algorithm	
Jan 24	Data Exploration and Visualization	PDM Chapter 3
Jan 29	Predictive Modeling Overview	PDM Chapter 6.1-6.3, 10.1-10.2
Jan 31	Predictive Modeling: Naïve Bayes	PDM Chapter 10.8
Feb 5	Predictive Modeling: Nearest neighbor, neural network	PDM Chapter 10.3, 10.6, 11.4
Feb 12	Predictive Modeling: Logistic regression, SVM	
Feb 14	Predictive Modeling: Search & Optimization 1	PDM Chapter 8.1-8.3
Feb 19	Predictive Modeling: Search & Optimization 2	
Feb 21	Predictive Modeling: Evaluation	PDM Chapter 10.10
Feb 26	Predictive Modeling: Wrap-up	
Feb 28	Review	
Mar 5	*In-class Midterm*	
Mar 7	Predictive Modeling: Ensembles (Bagging & Boosting)	
Mar 12	*No class (Spring break)*	
Mar 14	*No class (Spring break)*	
Mar 19	Predictive Modeling: Ensembles (Boosting & Random Forest)	
Mar 21	*Final project pitch*	
Mar 26	Guest lecture: Deep learning (Professor Yexiang Xue)	
Mar 28	Descriptive Modeling Overview	
Apr 2	Descriptive Modeling: K-means & Hierarchical clustering	PDM Chapter 9.3-9.5
Apr 4	Descriptive Modeling: GMM & Expectation maximization	PDM Chapter 9.2, 9.6
Apr 9	Descriptive Modeling: Evaluation	
Apr 11	Pattern Mining	PDM Chapter 11
Apr 16	Data Mining: Recent topics	
Apr 18	*No class (Project Day)*	

Apr 23	Final project presentation (Session 1)	
Apr 25	Final project presentation (Session 2)	

Grading

- Assignment: 55%
- Midterm exam: 20%
- Final project: 25% (proposal + pitch: 5%, final presentation: 10%, final report: 10%)

Final Project

Final project serves as an opportunity for students to get hands-on experience in data mining and practice the techniques and algorithms they learn in this course in real-life data mining scenarios. Projects are open-ended. Students are asked to:

- Identify a real-life scenario where data mining can be useful
- Define the data mining task and collect the necessary data
- Apply and implement algorithms learned from this course on the specified data mining task
- Interpret the knowledge obtained from the data mining process and evaluate the performance of implemented data mining algorithms

Students should complete the project in teams of 2-4 people. Tasks related to the final project include:

- Submit a project proposal which identifies the data mining task that the team aims to work on
- Give a pitch presentation on the project proposal in class
- Give a final presentation on the project in class, reporting the results of the project
- Submit a final project report summarizing the project

More detailed instruction on the final project will be provided through project guidelines.

Late Policy

Assignments need to be submitted by the due date listed. Each student gets **three** extension days which can be applied to any combination of assignments during the semester, **except for Assignment 1**, without penalty. Students must explicitly state in the assignment submission the number of extension days used, and cannot be rearranged after they are applied.

Beyond extension days, a late penalty of 10% per day will be applied to assignments that are submitted after the due date. However, assignments will NOT be accepted if they are more than 5 days late.

No extensions or late days are allowed for any project-related due date.