C S 488/508 Introduction to Data Mining

Syllabus

Computer Science, New Mexico State University
Aug. 17, 2022 - Dec. 9, 2022

1 Course time & venue

Class: TueThu 1:00pm - 2:15pm, SH 110.

COVID information

Information on COVID response and what it means for campus classes and activities continues to change. Please continue to refer to university's COVID-related guidelines.

1.1 Exams

Midterm: October 4 (Tuesday), class time
Final: December 6 (Tuesday), 3:50pm-5:50pm US Mountain Time

2 Instructor and teaching assistants

Instructors:
Dr. Huiping Cao
Office: Science Hall 171
Phone: 575-646-4600
Email: hcao@nmsu.edu
Office hours: 10:00am - 12:00pm (Tue) US mountain time or by appointment.
During these hours, you can come to the instructor's office or reach the instructor on zoom meeting: `https://nmsu.zoom.us/j/8549600124`.

Dr. Tuan Le
Office: Science Hall 149
Phone: 575-646-3234
Email: tuanle@nmsu.edu
Office hours: 09:00am-10:00am (Thu) US mountain time or by appointment.
During these hours, you can come to the instructor's office or reach the instructor on zoom meeting: `https://nmsu.zoom.us/j/4048892636`.

Teaching Assistants:
Mr. Fuad Ahmad
Office: Science Hall 128
Email: fuad@nmsu.edu
Office hours: 3:00pm - 4:30pm (Mon Wed) US mountain time or by appointment.
During these hours, you can come to the instructor's office or reach the instructor on zoom meeting: `https://nmsu.zoom.us/j/2760342799`.

Dr. Baokun Li
Office: TBA
Email: bali@nmsu.edu
Office hours: 3:00pm - 4:30pm (Thu) and 9:00am-10:30am (Fri) US mountain time or by appointment.
During these hours, you can come to the instructor's office or reach the instructor on zoom meeting: `https://zoom.us/j/7633205765?pwd=QmFzV2xBQ000eWgwUFhEVmdkdGdXZz09`.

Meeting ID: 7633205765 and Passcode: KMVg0V.

3 Course description

(Catalog description) Techniques for exploring large data sets and discovering patterns in them. Data mining concepts, metrics to measure its effectiveness. Methods in classification, clustering, frequent pattern analysis. Selected topics from current advances in data mining.

4 Course topic list

The course will cover the following topics:

(a) Data, Data Pre-processing, Proximity
(b) Regression: Linear Regression
(c) Classification: Decision Trees, kNN, SVM, Naive Bayes, Bagging, Boosting, Random Forests
(d) Clustering: k-means, Hierarchical Clustering, Gaussian Mixture Models, DBSCAN
(e) Association Analysis: Apriori, FP-Growth, GSP
(f) Anomaly Detection: Distance-based Approaches, Density-based Approaches, Clustering-based Approaches

5 Course learning outcomes

Upon completion of this course, the students are expected to:

- LO1: Explain and recognize different data mining tasks such as data pre-processing, visualization, classification, regression, clustering, association rules, and anomaly detection
- LO2: Apply classical data-mining/machine-learning algorithms for classification, clustering, association rules, and anomaly detection
- LO3: Evaluate and compare the performance of different data-mining/machine-learning algorithms
- LO4: Utilize data mining algorithms to analyze data in real applications using a data mining tool

6 Course practicum requirement

- Use data mining tools and libraries to analyze data in real applications.

7 Prerequisites

- At least a C- in C S 272, or Consent of Instructor.

8 Topics and outline

Week 1:	(8/15-8/19)	Syllabus; Introduction
Week 2:	(8/22-8/26)	Introduction (cont.); Review: math, Python
Week 3:	(8/29-9/2)	Review Python (cont.); Data, Data proximity
		9/2: Last Day to Drop a Course without "W" (refund)
Week 4:	(9/5-9/9)	Data exploration; Classification - Basics
Week 5:	(9/12-9/16)	Classification - Decision trees, Issues
Week 6:	(9/19-9/23)	Classification - Logistic regression, KNN, NB
Week 7:	(9/26-9/30)	Classification - SVM, ANN; Midterm review
Week 8:	(10/3-10/7)	**10/4: Midterm**; Classification - Ensemble
Week 9:	(10/10-10/14)	Cluster analysis - concepts; k-Means; hierarchical; Midterm analysis
		10/14: Last Day to Drop Course with "W" (no refund)
Week 10:	(10/17-10/21)	Clustering - Hierarchical (cont.), DBSCAN, Evaluation
Week 11:	(10/24 -10/28)	Clustering - Evaluation, GMM, Spectral
Week 12:	(10/31-11/4)	Association analysis - Concepts, Apriori, FPGrowth
Week 13:	(11/7-11/11)	Association analysis - Sequential patterns; Anomaly detection
Week 14:	(11/14-11/18)	Anomaly detection; Avoid false discoveries
Week 15:	(11/21-11/25)	**Thanksgiving break**
Week 16:	(11/28-12/2)	Final review
Week 17:	(12/5-12/9)	Final's week

We will generally follow the above outline. The pace may be slightly faster or slower.
Refer to the Important Dates for Students at records.nmsu.edu for all deadlines (do not use the NMSU Academic Calendar).

9 Preparation

On average, students should plan to spend at least 10-20 hours per week reviewing the course material and/or working on homework.

10 Textbook and Software

10.1 Textbook

Introduction to Data Mining, by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar (2nd edition). ISBN: 978-0-13-312890-1.

10.2 Software

- You are required to use Python on all your homework and projects. Find information for Pythan at `https://www.python.org/` Please make sure that your code can run on CS servers. If you want to install Python on your own computer, make sure the version is compatible with the version on CS servers.
- You may want to use GitHub to manage your project files. The core of Git is available below (you may have it, or it may be bundled with a graphical user interface): Git 2.9.2 or higher. Online: `https://git-scm.com`
  If you want to have a GUI available; the following are suggested:
  - SourceTree 2.1.3 or higher. Available for OS X and Windows. Online: `https://www.sourcetreeapp.com`
  - SmartGit 7.1.4 or higher. Available for all platforms. Online: `http://www.syntevo.com/smartgit/`

11 Assignments and projects

There will be individual assignments, employing the techniques we learn in this course. Students are encouraged to submit assignments on time, and follow the instructions in the assignment requirements. Late submissions of assignments are allowed. However, late assignments will be assessed **a penalty of 20% per day**.

12 Grading

The total grade for the course is a combination of your exams, assignments, and participation. The weighting will be

Assignments (25%) + Project (40%) + Midterm (15%) + Final Exam (15%) + Participation (5%) .

- Assignments need to be done individually.
- The project will be a semester-long project with several stages. You need to work in a group on this project.
- There will be one midterm and one final exam.
- **Students in CS 508 will be given more work (e.g., extra homework/exam/project questions) than students in CS 488. The extra workload is about 10%.**
- Participation will be calculated using spontaneous participation quizzes.

Final letter grades will be reported. Final letter grades will be based on the **final points**

Letter grade	final points	Letter grade	final points	Letter grade	final points
A+	≥97 points	A	[93,97) points	A-	[90,93) points
B+	[87,90) points	B	[83,87) points	B-	[80,83) points
C+	[77,80) points	C	[73,77) points	C-	[70,73) points
D+	[67, 70) points	D	[63, 67) points	D-	[60, 63) points
Fail	[0, 60) points				

13 Communication

This course will use Canvas announcements for general communication and NMSU email for direct communication. For direct communication, talk with me during office hours, or send email to the address given in the beginning of the syllabus. You can expect a response within 24 hours M-F.

Email Communications: Your NMSU email account is the official means of communicating with the university. Information critical to your success at NMSU is delivered to you via this account, and you are expected to follow rules and policies provided to you via this communication method. Please be advised that due to privacy and security concerns, we are unable to respond to emails from or about students that do not originate from an official NMSU email address.

14 Classroom policies

Attendance is highly correlated to student performance, so if you hope to do well you should be sure to log into the course and complete assignments.

15 Netiquette

Netiquette Expectations

- Always be polite and respectful in online correspondences.
- Do not attack. Listen gently to other views.

- Pay attention to your word choice. Be sensitive to others.
- Use spell check.
- Avoid using all caps. Use of all caps is considered shouting.

16 Required Technical Skills

Taking an online course requires technical skills as well as other soft skills. However, at a minimum you will need to meet certain technology responsibilities to complete work for this course. If you have questions about technical requirements for the course, please contact me immediately.

To begin in this course, you must:

- Read this syllabus carefully and contact me immediately if you have any questions. You are responsible for the content and assignments in this syllabus.
- Be able to obtain access to an internet connection, preferably broadband, and a working computer for the duration of this course.
- Be proficient with Microsoft© Office applications.
- Be able to send and receive emails and email attachments in and out of class.
- Be able to change your Canvas Notification settings.
- Be able to maintain backups of all work you create for this course.

17 Technology Requirements

To fully participant in the course you will need the following:

- Access to a Windows or Macintosh desktop computer or laptop with internet access, sound, and speakers.
- Canvas Learning Management System
- NMSU Canvas Login
- Canvas student FAQ
- Microsoft Office 2007 or higher
- Adobe Reader (for reading PDF files)
- Headset with microphone.
- Web cam.
- Zoom
- Respondus Lockdown Browser (LDB) and Monitor

18 Web Browsers

Use only the latest version of Google Chrome or Mozilla Firefox for Canvas. Safari, Internet Explorer, and Microsoft Edge have known issues that can interfere with performing basic tasks within Canvas. The links to download the recommended browsers as well as instructions on how to ensure you have the latest version are listed below:

- Download Google Chrome
- How to update Google Chrome
- Download Mozilla Firefox
- How to update Mozilla Firefox

Canvas does not fully support mobile devices; while there is a free Canvas mobile app available through iTunes store, a lot of functionality is unavailable when using a mobile phone. When you take this course, it is assumed you have access to a computer or laptop for full access to functionality in this course.

19 Accessibility of learning Tools

This course uses several software programs and technologies. Please read the following for more information about their accessibility. Note: A Voluntary Product Accessibility Template (VPAT) is a standardized form developed by the Information Technology Industry Council to show how a software product meets key regulations of Section 508 of the Rehabilitation Act.

- Canvas Accessibility Standards and help.
- Canvas-compatible Screen Readers: VoiceOver (Mac), JAWS (PC)
- Adobe Products Accessibility for Adobe Connect 9, Adobe Acrobat, and more.
- Apple Products VPATs and Accessibility features for Safari Web Browser, Mac OS X, and more.
- Microsoft Products: Section 508 and Microsoft Accessibility for Office, Skype, and more.
- Google VPATs and Accessibility Products and Features for Google Earth, Chrome Web Browser, Google Docs, and more.
- Firefox Web Browser: Section 508 (version 3.5 and up)

20 Syllabus Student Resources & Policy

The provost office has detailed resources and policy information at `https://provost.nmsu.edu/faculty-and-staff-resources/syllabus/policies.html`. It covers topics related to

- COVID
- Discrimination & Disability Accommodation
- Other NMSU Information (important phone numbers and information)
- Academic Misconduct
- Important Dates
- Grading Policy
- Recording Lectures
- Student Support Services
- Privacy Policies

21 Disclaimer

The information in this syllabus is provided on a good-faith basis. If in my judgment things need to be changed, they will be changed. You will be given adequate notice through an announcement on Canvas if and when such changes occur.