CS 521: DATA MINING TECHNIQUES

Instructor: Abdullah Mueen
Meetings: Tuesdays, Thursdays, 9:30AM-10:45AM
Zoom Meeting ID: https://unm.zoom.us/j/4596552426
Meeting ID: 459 655 2426
Passcode: mEeTmUeEn
Office: Farris 3020
Email: mueen in the CS or UNM domains;
Office Hours: Thursdays and Wednesdays, 12:30PM-2:00PM at the same Zoom link

Description: This course covers data mining topics from basic to advanced level. Topics include data cleaning, clustering, classification, outlier detection, association-rule discovery, tools and technologies for data mining and algorithms for mining complex data such as graphs, text and sequences. Students will work on a data mining project to gather hands-on experience.

The course learning objectives include

- Learning basic data mining algorithms and their applications
- Learning about the tools and technologies available for analyzing various types of data
- Gaining hands-on experience in cleaning, managing and processing complex data.

Zoom Meeting Link:

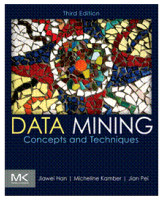
https://unm.zoom.us/j/4596552426?pwd=aVpDQjhrbG9ReHI3UHkvbWNoKzA2UT09

Meeting ID: 459 655 2426

Passcode: mEeTmUeEn

Office Hours:

Book: Data Mining: Concepts and Techniques, 3rd ed. By Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791. We will be occasionally referring to this book by Charu Aggarwal. The book is freely available to download in campus network.



Grading: There will be two exams. One midterm on topics from weeks 1-7 and the final on the remainder of the topics. The exams are worth 25% each. Students will pick group-projects and apply mining algorithms. Project is worth 25%. There will be in-class evaluation conducted by Zoom Polls on covered topics at any class, worth 10% of the course. There will be three to

five programming assignments, together they are 15% of the course. The assignments will be on applying different techniques on real-data selected by the instructor.

Policy:

1. Internet issue at student's end is the student's responsibility. Instructor expects that students have fully functional Internet connection. If there is an internet issue at the instructor's end, students are expected to wait in the meeting room until instructor rejoins or until class time is over, whichever comes first.
2. Instructor assumes permission to record your voice and video during the lecture when discussions are on. Talk to the instructor if otherwise.
3. We will follow an Online Course Etiquette Appreciative Agreement.

Lecture Schedule: A tentative weekly distribution of topics is given below. There will be re-arrangement for holidays and exams.

Week 1:	Ch. 1, 2: What is Data Mining? Types of Data.
Week 2:	Ch. 3: Data Preprocessing. Cleaning, Integration, Reduction and Transformation
Week 3:	Ch. 6, 7: Mining Frequent Patterns (FP), Associations and Correlations. Apriori, FP Tree
Week 4:	Ch. 8: Basic Classification: Decision Tree, Bayes Classifier, Rule Based, Goodness measures
Week 5:	Ch. 8, 9: Advanced Classification: Boosting, Bagging, Random Forest, Lazy Learners, FP based classification
Week 6:	Ch. 10: Basic Clustering: Hierarchical, Partitioning, Density-based, Grid-based
Week 7:	Ch. 11: Advanced Clustering: Subspace clustering, Co-clustering, Fuzzy clustering, Expectation-Maximization clustering
Week 8:	Ch. 12: Outlier Detection: Statistical and Proximity based methods
Week 9:	Ch. 13: Mining Complex Data Types: Sequences (real and discrete)
Week 10	Mining Complex Data Types: Graphs and Trees
Week 11:	Mining Complex Data Types: Text, Logs, Reviews
Week 12:	Ch. 4,5: Data Mining Systems: Data warehousing, Data cubing, Business Intelligence systems
Week 13:	Data Mining Tools: Weka, Vowpal-wabbit, Pivot-tables, Matlab Statistics Toolbox
Week 14:	Web Mining: Web search, Computational advertising, User behavior modeling, Fraud detection

Project: Each group will do one project. A group can have at most three students (individual projects are not allowed). A project consists of two phases of *equal weights*.

> **Proposal:** Each team will select a project involving mining moderate to large sized data. In the proposal, each team will describe their project in the form of a research article containing sections: Abstract, Introduction, Dataset, Tasks, Expected Results. Each team will describe the dataset in detail, the objective of the project divided in two to four tasks and expected results in connection to the data domain. Teams are encouraged to select datasets from active research projects. In addition, teams can select dataset from online sources including active or past Kaggle competitions, open data archives, etc. Evaluation criteria will focus on feasibility, relevance, and complexity of the tasks. **Due: October 8, 2020**.
>
> **Submission:** Each team will submit a report and code on completed projects. Each team will describe their solution to the proposed tasks, describe results obtained and discuss the relevance of the results to the data domain. Teams will discuss how the obtained results match or mismatch with the Expected Results in Proposal. **Due: December 3, 2020.**

Exam:

There will be two exams **on Oct 8, 2020** and **Dec 3, 2020**. Exams will be on UNM Learn.