MLDS 421-0_SEC1: Data Mining

Northwestern University

Instructor: Ashish Pujari | apujari@northwestern.edu

Schedule: Winter 2024, Fridays 2 PM – 5 PM

Course Overview

Welcome to this course on data mining, where we delve into the art of uncovering valuable insights from extensive datasets. This intricate process involves the identification of anomalies, trends, patterns, and correlations within large data sets, enabling the anticipation of future outcomes. Positioned at the crossroads of machine learning, statistics, and database systems, data mining methods play a pivotal role in extracting meaningful information. In the business realm, these techniques aid in deciphering patterns, ranging from customer preferences, and purchasing behavior to fraud detection and spam filtering.

Throughout this course, we will explore essential topics, including Exploratory Data Analysis (EDA), Dimensionality Reduction, Association Rule Mining, Recommendation Engines, and Clustering methods. Emphasizing practical application, the primary mode of instruction will involve Python notebooks, allowing for a hands-on approach to mastering the intricacies of data mining.

Learning Objectives

1. Gain proficiency in Exploratory Data Analysis (EDA) techniques to effectively preprocess and analyze data for meaningful insights.
2. Explore Dimensionality Reduction methods to efficiently handle high-dimensional datasets and enhance model performance.
3. Master the principles of Association Rule Mining to uncover relationships and patterns within datasets, facilitating informed decision-making.
4. Develop expertise in designing and implementing Recommendation Engines for personalized content suggestions based on user behavior.
5. Acquire skills in Clustering methods to categorize and group similar data points, enabling the identification of inherent structures within datasets.
6. Utilize Python notebooks as the primary tool for implementing and experimenting with data mining algorithms, fostering hands-on proficiency.

Prerequisites

- Know your computer (Setting environment variables, Using the Mac/PC terminal, traversing applications/folders, updating security preferences)
- Programming for Analytics - Python

Course Materials

Class topics are sourced from multiple sources including the following recommended books. While reading assignments will be supplemented, the books jointly cover the class topics, with both construct and methods. Note: The hands-on exercises in class as well as assignments have been custom designed for this course and are based on public data sources.

Recommended Books:

- Data Mining: Concepts and Techniques – Jiawei Han, et al.
- Elements of Statistical Learning
- Python Machine Learning - Sebastian Raschka
- Hands on Machine Learning - Aurélien Géron

Software

This course will require working in

- Python Anaconda, Jupyter Notebooks
- Python libraries – Pandas, Sklearn, Matplotlib, Seaborn, etc.

Note: These software applications work best on PC's/Macs. Ensure the computer you are using provides you with the authority to perform these installs. Some work-related computers may not permit such installations without admin rights.

INFRASTRUCTURE

DeepDish @ Northwestern University

Google Colab

Course Work

Assignments (60%)

The course will include 4 assignments involving data mining problems and coding which account for 60% of your grades.

- All the code and solutions in the individual assignments must be your own.
- If the same code is identified across assignments, students will be penalized.
- You may reuse code-snippets from large language models, code repositories, lecture examples, and other reference material.
- Collaboration is only permitted for team projects.

Late Policy

- 20% for the first day it is late.
- 10% for every additional day it is late.

Team Project (30%)

The goal of the team project is to apply data mining techniques and best practices to real world problems. Students will team up as groups of 3 or 4 to collaborate on a public dataset or a Kaggle competition. Students will present their approach, algorithms, and findings as a team during the final presentation in Week 10. More details about the project will be provided in the course material.

The final project accounts for **30%** of your overall grade, and the project report will include the following sections:

- Abstract
- Paper Review and Model Approach
- Data Analysis and Model Development
- Findings and Conclusion

Class Participation (10%)

Class participation includes but is not limited to the following activities:

- Class attendance and quizzes
- Engaging in class and online discussions

Course Schedule

Week 1: Introduction to Data Mining

- Course Orientation
- Data Mining Overview
- Data Mining Applications
- Software Installation

Week 2: Dimensionality Reduction 1

- Curse of Dimensionality
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Project Teams

Week 3: Dimensionality Reduction 2

- Kernel PCA
- T-SNE
- U-MAP
- Project Review Checkpoint 1
- Assignment 1 due

Week 4: Cluster Analysis I

- Cluster Analysis Overview
- Measures of Similarity and Dissimilarity
- Expectation–Maximization
- Partitioning-based Clustering
    - K-Means, K-Median, PAM(K-Medoid), K-Modes, K-Prototypes

Week 5: Cluster Analysis II

- Hierarchical-based Clustering
    - Agglomerative and Divisive clustering
- Density-based Clustering
    - DB-SCAN
- Model-based Clustering
    - Gaussian Mixture Models
- Assignment 2 due

Week 6: Association Rules Mining

- Frequent Itemsets
- Association Rules
- APriori, FPGrowth

Week 7: Recommender Systems

- Content-based Recommenders
- Collaborative Filtering
    - Memory-based, Model-based
- Hybrid Recommenders
- Assignment 3 due

Week 8: Bayesian Networks

- Probabilistic Graphical Models
- Bayes Nets Representation
- Bayes Nets Inference

Week 9: Graph Mining

- Graph Applications
- Graph Data Model and Algorithms
- Graph Analytics
- Assignment 4 due

Week 10 (Final Project)

- Team presentations (20 mins per team)
- Q & A discussion
- Final Project due