MSiA 421: Data Mining

Northwestern University

Ashish Pujari | apujari@northwestern.edu | https://linkedin.com/in/apujari

Winter 2023, Wednesdays 6 PM – 8:30 PM

Course Overview

Data mining is the process of extracting useful information from a vast amount of data. It involves finding anomalies, trends, patterns, and correlations within large datasets to predict outcomes. Data mining methods are at the intersection of machine learning, statistics, and database systems. Data Mining assists firms in analyzing and comprehending patterns, from learning about what customers are interested in or want to buy to fraud detection and spam filtering.

This course will cover topics such as Exploratory Data Analysis, Dimensionality Reduction, Association Rule Mining, Recommendation Engines and Clustering methods, and will be primarily taught through Python notebooks.

Learning Objectives

After completing this course, students should be able to:

1. Perform Exploratory Data Analysis (EDA) on raw data sets.
2. Apply advanced techniques, methods and best practices in data analysis and visualization
3. Implement data pre-processing, association mining, recommender systems and clustering algorithms.
4. Work efficiently in groups and evaluate the algorithms on real-world problems.
5. Produce living documents with code, graphs, and text using Jupyter Notebooks.

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.
- Docker

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

Sagemaker Studio Lab

Course Work

Assignments

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 use code from the interactive repository, lecture examples, and other reference material as inspiration but you may not copy and paste any functions into your code
- 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

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 11. 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

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

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

Evaluation

You will pass this course provided you complete the following:

- Assignments (4) - 60 %
- Final Project - 30%
- Class Participation – 10%

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 Recommendation Systems
- 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