Data Mining - Course Introduction

Joe Burdis
Fall 2024
CUNY Graduate Center

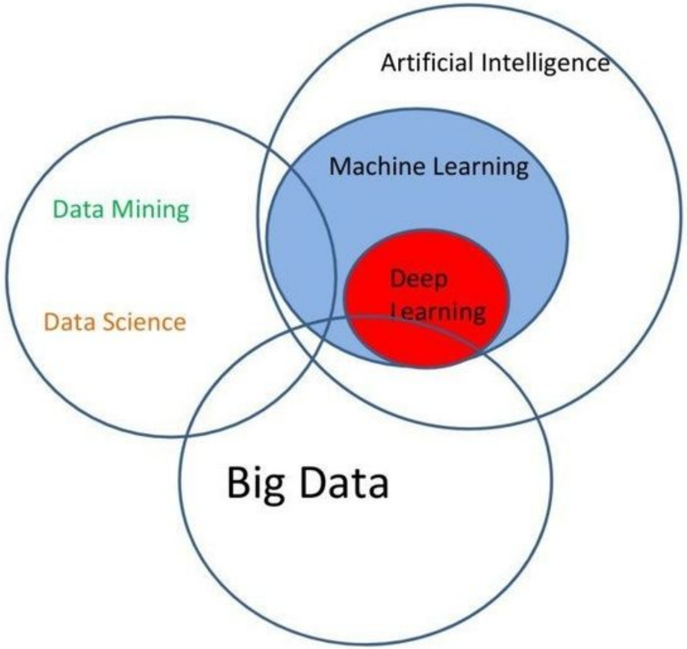
Outline

1. What is Data Mining?
2. Introductions and Survey for the class
3. Introduction to Python Data Science Libraries

What is Data Mining?

Data mining is the process of discovering meaningful patterns and extracting valuable insights from large and complex datasets. In today's world, data mining goes beyond simply finding patterns. Sitting at the intersection of machine learning, statistics, and database systems, it encompasses a wide range of tasks, including:

- **Predictive modeling**: Forecasting future outcomes or behaviors based on historical data.
- **Anomaly detection**: Identifying unusual or unexpected data points that may indicate critical events or errors.
- **Association rule learning**: Discovering interesting relationships between variables in large databases.
- **Clustering**: Grouping similar data points together to understand underlying structures and segments.
- **Classification**: Assigning categories or labels to data points based on their characteristics.
- **Exploratory Data Analysis (EDA):** Employing a broad set of techniques to summarize, visualize, and investigate datasets, often as a precursor to more targeted analysis or modeling. This can involve statistical summaries, data visualization, and interactive exploration to gain a deeper understanding of the data's structure, content, and potential insights.



Source: Technological Singularity

Survey

1. What is your goal from taking this class? E.g.
   a. Gain broad knowledge base
   b. Job in Data Mining
   c. ML Engineering
   d. ML/AI Research
2. What is your related background? E.g. statistics, mathematics, computer science, engineering, professional experience, etc.
3. Related to question #1, what is your preference for the course? E.g. deep vs broad vs somewhere in between. Include time series, spatial or graph data, docker, and kubernetes, etc.
4. Special topic interests?

Grading Policy

- Homework (50%)
- Midterm Exam (20%)
- Final Project (30%)

Course Information

Instructor: Joe Burdis

Email: joe.burdis@gmail.com

Slides and Python code: Blackboard

Main Textbook: Mining of Massive Datasets by Leskovec-Rajaraman-Ullman, Cambridge University Press.

References:

- Wes McKinney: Python for Data Analysis, 2nd Edition, O'Reilly Media, Inc.
- Aurélien Géron: Hands-On Machine Learning with Scikit-Learn and TensorFlow, O'Reilly Media, Inc.

Prerequisites: undergraduate-level programming, calculus, linear algebra, probability, and algorithms

Massive Datasets

- Databases, data warehouses
- Sensor data
- Time series data
- Text data
- Images, videos, audios
- Graph data
- …

Challenges of Data Mining

- Diversity of data
- Efficiency and Scalability
- Mining Methodology
- Security and Social challenges

We won’t restrict ourselves to only working with massive data, but we will highlight corresponding challenges throughout

Scope of This Course

- Widely-used data analysis techniques
- Mining symbolic or numerical data
- Challenges and approaches to mining massive datasets
- Parallel and probabilistic algorithms
- Recent highlights of the data mining community
- Comparison between traditional data mining techniques and predictive modeling and vector search.