syllabus

schedule

WEEK 1

Introduction and Applications

January 11

- Topics
  - A Course Overview
  - Data Vocabulary
- Suggested Reading
- Assignment 1 is assigned - *Review & Exploring Data*

WEEK 2

Mining for Association Rules

January 18

- Topics
  - Definitions of Frequent Itemsets
  - Determining Frequent Itemsets
  - Creating Association Rules
- Suggested Reading
- Submissions
  - Assignment 1 is **due**
  - Assignment 2 is assigned - *Association Rules*

WEEK 3

Accessing, Storing, and Computing with "Big" Data

January 25

- Topics
  - Distributed Filesystems and Storage
  - Structured Data Analysis (SQL)
  - Introducing the MapReduce Paradigm
  - Distributed Computation
- Suggested Reading - Chapter 2, Sections 2.1-2.4
- Submissions
  - Assignment 2 is **due**
  - Assignment 3 is assigned - *Map Reduce Problem*

WEEK 4

Large Scale Data (Pre)-Processing

February 1

- Topics
  - Basics of Linear Algebra and Probability Theory
  - The Multiple Places Where Data Lives & Multi-source Joins
  - Covariance, Correlation, and Cosine Similarity
  - Dimensionality Reduction and Feature Selection
- Suggested Reading
  - Linear Algebra Review
  - Dimensionality Reduction
  - Map Reduce, Sections 2.5-2.7

WEEK 5

February 8

- Topics
  - Introducing the Gaussian Distribution
  - Parameter Estimation of a Distribution
  - Anomaly and Outlier Detection
  - Unsupervised Modeling with k-Means and Clustering
- Suggested Reading
  - Maximum Likelihood
  - Unsupervised Clustering, Chapter 7 - 7.2
- Submissions
  - Assignment 3 is **due**
  - Assignment 4 is assigned - *Parameter Estimation & Clustering*

WEEK 6

Mining Small-ish to Medium-sized Data - Statistical Learning

February 15

- Topics
  - The Bayesian Framework
  - Naive Bayes Classification
  - Tree-based Algorithms - Random Forests
- Suggested Reading
  - Bayes Theorem
  - Naïve Bayes
  - Tree Algorithms - Chapters 3.1 - 3.3
- Submissions
  - Assignment 4 is **due**
  - Assignment 5 is assigned - *Bayesian Framework & ML Libraries*

WEEK 6

Course Review, Midterm Preparation

February 22

- Topics
  - Course Review and Midterm Preparation
- Submissions
  - Assignment 5 is **due**

WEEK 7

Midterm Exam

February 29

- Topics
  - Linear Algebra Review
  - MapReduce Problems
  - Principle Component Analysis
  - Parameter Estimation
  - Unsupervised Clustering
  - Bayesian Framework
- Suggested Preparation

WEEK 8

No Instruction This Week - Spring Break

March 7

- Have a nice break!

WEEK 9

Foundations of Machine Learning

March 14

- Topics
  - Algorithmic Evaluation with Confusion Matrices, Thresholds, ROC Curves
  - The Objective Function, Regularization, and Constraints

        - Logistic Regression - Precursor to Modern Data Mining
        - Batch Data Processing - Gradient Descent
        - The Bias and Variance Tradeoff
    - **In-Class Colabs**: Logistic Regression with MNIST
    - Suggested Reading
        - Evaluation Metrics, Chapter 8.5
        - Logistic Regression ([1], [2])
    - Submissions
        - Assignment 5 is **due**
        - Assignment 6 is assigned - *Evaluation Metrics*

WEEK 10

Mining Images with Deep Learning

March 22

- Topics
    - Working with Tensors - Reviewing Multivariate Calculus
    - Deep Learning - A Historical Perspective
    - The Backpropation Algorithm
    - Convolutional Neural Networks
- Suggested Reading
- Submissions
    - Assignment 6 is **due**

WEEK 11

Mining Text with Self Supervision

March 28

- Topics
    - Some Basic Approaches
    - Semi-Supervised Learning
    - The Concept of an Embedding Space
    - The Attention Mechanism
    - Large Language Models - From BERT to ChatGPT
- Suggested Reading
    - Information Retrieval, Chapter 13
- Submissions
    - Project proposals are **due**

WEEK 12

Data Mining Applications

April 4

- Topics
    - Social Network Data Mining
    - Recommendation Sciences
    - Time Series Analysis
- Suggested Reading

WEEK 13

Project Presentations

April 11

- Project Presentations and Outbriefs
    - Mining for Anomalous Behavior
    - Mining in Operational Logistics
    - Mining to Notify and Alert
- Submissions
    - Final projects are **due**, including presentation slides and writeup

---

WEEK 13

Industry Day

April 18

- Mining for Anomalous Behavior
- Mining in Operational Logistics
- Mining to Notify and Alert

- Submissions
  - Final projects are **due**, including presentation slides and writeup

---

WEEK 14

Final Exam

April 25

- Topics
  - Objective Functions
  - Logistic Regression
  - Association Rule Mining
  - Evaluation Metrics
  - Backpropagation
  - Convolutions and Recurrence

grading criterion

Labs & Participation	10%
Data Mining Project	10%
Assignments	20%
Midterm Exam	30%
Final Exam	30%

course meeting times

Lectures

- Tues, 6pm-9:20pm
- Room TBD

Office Hours

- Professor, Thurs, 8:30-9:30pm
- TA, Date/Time TBD

suggested textbooks

Introduction to Data Mining, 2nd Edition Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, 2018

Mining of Massive Data Sets, 3rd Edition Jure Leskovec, Anand Rajaraman, and Jeff Ullman, 2014

Deep Learning Ian Goodfellow, Yoshua Bengio, and Aaron Courville, 2016