Data Mining and Analytics (Spring 2019)

Jump to Today

Data Mining & Analytics (INFO 254/154)

Spring 2019

University of California, Berkeley

School of Information

Instructor: Prof. Zachary Pardos (homepage (https://www.ischool.berkeley.edu/people/zachary-pardos))

pardos@berkeley.edu (mailto:pardos@berkeley.edu)

TuTh 2:00 - 3:30 PM — 150 GSPP

Office Hours: Monday 3-4:30pm (zoom (https://berkeley.zoom.us/my/pardos) or by appointment in 2121 Berkeley Way West, Suite 4232)

GSI: **Neha Mittal**

neha01mittal@berkeley.edu

Office Hours: 11.15am on Mondays. Room 302, South Hall. 15-minute slots by appointment via email.

Reader: **Rajvardhan Oak**

rvoak@ischool.berkeley.edu

Class Time: Tuesdays and Thursdays, 2:00 - 3:30 PM

Location:150 GSPP

Course Description

Data Mining and Analytics introduces students to the practical fundamentals and emerging paradigms of data mining and machine learning with enough theory to aid intuition building. The course is project-oriented, with a project beginning in class every Thursday and to be completed outside of class by the following week, or two for longer assignments. The in-class portion of the project is meant to be collaborative and a time for the instructor and GSIs to work closely with project groups to understand the objectives, help work through software logistics, and connect project work to lecture. Tuesday lectures introducing theories, concepts, contexts, and algorithms. Students should expect to leave the class with hands-on, contemporary data mining skills they can confidently apply in research and industry. There will be a written midterm test and a final group project report and presentation. Experience with Python is required.

Course Objectives

- Foster critical thinking about real-world actionability from analytics.
- Develop intuition in various machine learning classification algorithms (e.g. decision trees, feed-forward neural networks, recurrent neural networks, support vector machines) and clustering techniques (e.g. k-means, spectral, skip-gram)
- Conduct manual feature engineering (from domain knowledge) vs. machine induced featurization (representation learning)
- Provide an overview of issues in research and practice that will affect the practice of data science in a variety of domains.

Grading

Homeworks/Labs: 30%

Midterm: 25%

Final Project: 35%

Quizzes: 10%

Late Policy: Late submissions, will be penalized 20% up to one week after the original due date. Another 10% penalty will be added for each subsequent week. Each student's most severe late penalty will be forgiven at the end of the course.

Texts

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*, Third Edition *(3rd ed.)*. Morgan Kaufmann.

Special Needs/Accommodations

I am committed to creating a learning environment welcoming of all students. If you have any special needs, please notify me as soon as possible so that appropriate accommodations can be made.

Agenda and Assignments

Typically, Tuesdays will be lecture days and Thursdays will be lab and quiz days. Labs are due one week after they are assigned (at 1pm) unless otherwise specified. Readings (recommended, not required) are associated with the day they are listed. Preparation materials for the labs will be posted on bCourses. Readings refer to the textbook unless otherwise specified.

Note on collaboration/group work: You may share strategies, ideas and help debug each others assignments but all assignments are to be turned in individually and completed individually otherwise stated in the assignment. Please do not copy and paste code without referring the source material in your comments or write up. The final project will be a team project.

[slides and lab assignment documents]: We have provided a link to the lecture slide deck in advance to help you prepare for class better. Lab assignments and their respective tutorial helper pages will be available at the start of class on the Thursday of the start of the lab.

Date	Topic
Tuesday, 22 January	No Class
Thursday, 24 January	**Course Introduction** (https://docs.google.com/presentation/d/1QnJS9AvAXyMa1I
Tuesday, 29 January	**Data Pre-processing Lecture** (https://docs.google.com/presentation/d/1YoNnA0LaQDqbN
Thursday, 31 January	**Data Pre-processing Lab**
Tuesday, 5 February	**Clustering Lecture** (https://docs.google.com/presentati usp=sharing)
Thursday, 7 February	Clustering Lab
Tuesday, 12 February	**Decision Trees Lecture** (https://docs.google.com/presentation/d/1fY8CZUeCkw2Ue6
Thursday, 14 February	Decision Trees Lab
Tuesday, 19 February (https://docs.google.com/presentation/d/1izlMw4BEZtXG_NKcdQtpyqur27tUQUFUFyXg9gvRE_M/edit? usp=sharing)	**Neural Networks Lecture** (https://docs.google.com/presentation/d/1izlMw4BEZtXG_NI

Thursday, 21 February	SVM and Neural Networks Lab
Tuesday, 26 February	**Ensemble Learning Lecture** (https://docs.google.com/presentation/d/1xPYCFSbNEfyTfT(
Thursday, 28 February	Kaggle Competition
Tuesday, 5 March	**Cross-validation and Error metrics Lecture** (https://docs.google.com/presentation/d/17WqCEnzETtZtRg
Thursday, 7 March	Continue Kaggle Competition
Tuesday, 12 March	**Midterm review** (https://docs.google.com/presentation/d usp=sharing)
Thursday, 14 March	Midterm
Tuesday, 19 March	**Skip-gram Lecture** (https://docs.google.com/presentatic usp=sharing)
Thursday, 21 March	Competition mini report-outs, Skip-gram Lab
Tuesday, 26 March	Spring Break, No Class
Thursday, 28 March	Spring Break, No Class
Tuesday, 2 April	**RNNs, Recommendation & Education** (https://docs.gc XIwZgbi0BMLqWtIQJNWuVMyCNgg2BCwqFQfwcg/edit?usp
Thursday, 4 April	Dataset 1 slide presentations for final project

Tuesday, 9 April	Dimensionality Reduction & Data Visualization (https vWrlDAgcCt_RVTvNLUmiMGialNY4/edit?usp=sharing)
Thursday, 11 April	Quiz (word2vec), finish Dimensionality Reduction & Data V
Tuesday, 16 April	TBD + final project work
Thursday, 18 April	Quiz (RNN/dimensionality reduction), Deep Learning Lab
Tuesday, 23 April	Advanced Clustering Lecture (https://docs.google.com Y/edit?usp=sharing) + final project work
Thursday, 25 April	Final project group work
Tuesday, 30 April	No class, presentations moved to next week
Thursday, 2 May	No class, presentations moved to next week
Tuesday, 7 May	Final Project Presentations
Thursday, 9 May	Final Project Presentations
Tuesday, 14 May	

Course Summary:

Date	Details	
Thu Feb 7, 2019	Lab 1: Data Preprocessing (https://bcourses.berkeley.edu/courses/1477998/assignments/7946382)	due by 11:59am