Data Mining and Analytics (INFO254)

Spring 2017

University of California, Berkeley

School of Information

Instructor: Prof. Zachary Pardos

pardos@berkeley.edu

South Hall 202

Office Hours: Mondays 1pm-2:30pm in Tolman Hall 4641

Thursdays immediately after class in South Hall 211

GSI: TBD

Office Hours: Wednesdays 2 - 3 PM in South Hall 110

Class Time: Tuesdays and Thursdays, 3:30 - 5:00 PM

Location: 210 South Hall

Course Description

The goal of Data Mining and Analytics is to introduce you 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 two GSIs and I to work closely with you or your group to understand the objectives and help you work through software logistics and make connections to the core concepts. Tuesdays are lecture days, introducing the concepts and algorithms which will be used in the upcoming project. The primary objective is for everyone to leave the class with hands-on, contemporary data mining skills they can confidently apply in research or industry. There will be a written midterm and the final will be in the form of a group final 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, neural networks, recurrent neural networks, support vector machines) and clustering techniques (e.g. *k*-means, spectral, skip-gram)
- Condcut manual feature engineering (from domain knowledge) vs. machine induced featurization (representation learning)
- Provide an overview of issues in research and practice that will shape the complexion of data science across a variety of domains.

Grading

Homeworks/Labs: 35%

Midterm: 25%

Final Project: 35%

Q & A assignments: 5%

Late Policy: Late submissions (including 1 minute late), will be penalized 20% up to one week after the original due date. Another 10% penalty will be added for each subsequent week. Extensions will be handled on a case by case basis but will not be granted on the same day as the due day.

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 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: The in-class lab periods are a time to collaborate on the assignment. You may share code, strategies, ideas during this time but all assignments are to be turned in individually and completed individually outside of the class period unless otherwise stated in the assignment. The final project will be a team project.

Date	Topic	Readings
Tuesday, January 17th	Course Introduction	
Thursday, January 19th	No Class. Office Hours.	
Tuesday, January 24th	Data Pre-processing Lecture	Section 2.1
Thursday, January 26th	Data Pre-processing Lab	Sections 3.1, 3.2, and 3.5 [install python, pandas & ipython notebook]
Tuesday, January 31st	Clustering Lecture	Sections 10.1, 10.2, 10.6
Thursday, February 2nd	Clustering Lab	
Tuesday, February 7th	Decision Trees Lecture	Sections 8.1 - 8.2.3
Thursday, February 9th	Decision Trees Lab	

Tuesday, February 14th	Neural Networks Lecture	Sections 9.2 (skim) and 9.3 In Spring '18 offering this will be - moved to the 2nd half of the semester and - replaced with SVM and regressions
Thursday, February 16th	SVM and Neural Networks Lab	[paper] A Practical Guide to Support Vector Classification
Tuesday, February 21st	Ensemble Learning Lecture	Section 8.6, [paper] Ensemble Selection from Libraries of Models
Thursday, February 23rd	Kaggle Competition	
Tuesday, February 28th	Cross-validation and Error metrics Lecture	Section 8.5
Thursday, March 2nd	Continue Kaggle Competition	
Tuesday, March 7th	Midterm review	
Thursday, March 9th	Midterm	
Tuesday, March 14th	Skip-gram Lecture	
Thursday, March 16th	Skip-gram Lab	The amazing power of word vectors [blog]
Tuesday, March 21st	Guest Lectur	
Thursday, March 23rd	Dataset 1 slide presentations for final project	
Tuesday, March 28th	Spring Break, No Class	
Thursday, March 30th	Spring Break, No Class	
Tuesday, April 4th	Dimensionality Reduction & Data Visualization	
Thursday, April 6th	Dimensionality Reduction & Data Visualization Lab	
Sunday, April 9th		One page project proposal due @ 11:59pm
Tuesday, April 11th	Deep Learning Lecture (RNN / Keras / mini CNN)	
Thursday, April 13th	Deep Learning Lab (RNN) - extra credit	
Tuesday, April 18th	Advanced Clustering Lecture + group work	Self-tuning Spectral Clustering (bCourses)
Thursday, April 20th	Final project group work	
Tuesday, April 25th	No class, presentations moved to next week	(In Spring '18 offering this will be mid-project presentations)
Thursday, April 27th	No class, presentations moved to next week	
Tuesday, May 2nd	Final Project Presentations	
Thursday, May 4th	Final Project Presentations	
Tuesday, May 9th	Final Project Papers Due at 11:59 PM	

Final week topics and final project schedule to be determined.