CMSC 25025 / STAT 37601:
MACHINE LEARNING AND LARGE SCALE DATA ANALYSIS

Syllabus, Spring 2017

Machine Learning and Large Scale Data Analysis is an advanced undergraduate level course in statistical machine learning for "big data." The course introduces machine learning and the analysis of large data sets using distributed computation and storage infrastructure. Basic machine learning methodology and relevant statistical theory are presented in lectures, and assignments and projects give students hands-on experience with the methods on different types of data. Topics include linear and nonparametric regression and classification, clustering algorithms, graphical models and hierarchical Bayesian modeling, topic models, word embeddings, and recurrent neural networks. Data types include Google image search data, Twitter feeds, political speeches, archives of scientific articles, public records of the city of Chicago, and telescope imagery. Programming is based on Python using Spark on the RCC compute nodes.

Schedule

LECTURES	Tuesday and Thursday 1:30-2:50 pm	Ryerson 251

Contact Information

Instructor:		
John Lafferty		lafferty@uchicago.edu
Office hours:	Tuesday, 5:00-6:00 pm, Jones 226	
Course Assistants:		
Qinqing Zheng		qinqing@cs.uchicago.edu
Lab hours:	Wednesday, 1:30-2:50 pm (CSIL 4)	
Tom Hen		tomhen@gmail.com
Lab hours:	Wednesday, 4:30-5:50 pm (CSIL 4)	
Pramod Mudrakarta		pramodkm@uchicago.edu
Lab hours:	Wednesday, 3:00-4:20 pm (CSIL 4)	
Yi Ding		dingy@gmail.com
Lab hours:	Wednesday, 1:30-2:50 pm (CSIL 4)	
TA group office hours:	Monday, 5:00-6:00 pm, Jones 226	

Prerequisites

Prerequisites: (CMSC 15400 or CMSC 12200), and (STAT 22000 or STAT 23400), or consent of the instructor.

Course Overview

This course is an introduction to machine learning and statistics for analyzing large scale data. The course presents motivation, methods, and some supporting theory for several types of data analysis, including classification and regression, clustering, hierarchical Bayesian modeling, unsupervised feature learning, and graphical modeling. The main objective of the course is for students to gain an understanding of and experience with some essential statistical machine learning methodology. An additional goal is to gain exposure to cloud computing for data analysis, using virtual clusters of compute nodes and distributed storage.

For each topic discussed in the class, an assignment will be given to explore the topic conceptually, and simulated data and other relatively small scale data will be used to gain some experience with different methods. Students will then explore several of the methods on larger scale data, using Spark on the University of Chicago's RCC system.

The course will not follow a textbook closely. However, the following book will be used for portions of the course, and is recommended: "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," by T. Hastie, R. Tibshirani, and J. Friedman (Springer, 2nd edition). The book is available at `www-stat.stanford.edu/~tibs/ElemStatLearn/`.

Course Structure and Grading

The course will have a standard lecture format.

Assignments will be handed out roughly every 10 days, and due in class. Assignments will typically include a mix of problem solving and data analysis (coding). Python will be the course programming language, unless otherwise specified. Assignments will also involve larger scale computation. These parts will be marked as LSDA (large scale data analysis), and will involve more open-ended use of a dataset. For these portions of the assignments, students may work individually or in pairs. LSDA problems will involve working with moderately sized data sets.

An in-class midquarter exam will be given, as well as a final exam. A couple of short 10 minute in-class quizzes may also be given. Each of these components will be weighted as follows to determine a final grade:

- Quizzes: 0%
- Assignments: 65%
- Midquarter exam: 15%
- Final exam: 20%

Policy on Assignments and Projects

Assignments must be submitted before the beginning of class on the day they are due. Both written and programming portions of assignments will be submitted electronically. Students may submit up to two assignments three days late, with no penalty.

Assignments will typically have at least one problem that involves "small scale" computation. Problems marked LSDA require use the programming environment set up for the course, based on Python Spark on the RCC.

Collaboration on homework assignments with fellow students is encouraged. However, such collaboration should be clearly acknowledged, by listing the names of the students with whom you have had any discussions concerning the problem. You may *not* share written work or code—after discussing a problem with others, the solution must be written by yourself. An exception is for the LSDA portions of assignments, where students may work in pairs. The contributions of each person should be clearly stated.

Course Materials, Calendar and Discussion Board

Course materials will be posted on the chalk website, at `chalk.uchicago.edu`, and will be updated throughout the semester. We will also use *Piazza* as a forum for discussion and questions: `piazza.com/uchicago/spring2017/cmsc25025/home`

A preliminary schedule of topics, exams, and assignments follows:

Week	Date	Topic	Out/Due
1	March 28 March 30	course overview background concepts, software platform	 assn 1 out
2	April 4 April 6	clustering, pca, and linear classification	 assn 1 in; assn 2 out
3	April 11 April 13	stochastic gradient descent, sparse coding	
4	April 18 April 20	language models and word embeddings	assn 2 in; assn 3 out
5	April 25 April 27	recursive neural networks	 assn 3 in; assn 4 out
6	May 2 May 4	— mixtures and Bayesian inference	midquarter exam
7	May 9 May 11	topic models	assn 4 in; assn 5 out
8	May 16 May 18	nonparametric regression	 assn 5 in; assn 6 out
9	May 23 May 25	greedy algorithms and random forests	
10	May 30 June 1	review reading period	assn 6 in
11	TBA	final exam	

The methods and data sets to be used on the assignments include the following (subject to change):

- Scanned images of digits (k-means clustering)
- State of the Union speeches (document clustering)
- Twitter feeds (stochastic gradient descent)
- Wikipedia (topic modeling)
- Archives of the Proceedings of the National Academy of Science (topic modeling)
- Lightcurves from the Kepler telescope (nonparametric regression, hypothesis testing)
- Free Music Archive (spectrograms, hashing and search)
- Beer and wine reviews (sentiment analysis, classification)
- City of Chicago crime data (spatial point processes)

Expectations

Statistical machine learning is an exciting, rapidly changing field. The landscape of cloud computing is also changing rapidly–each year this course has been taught, the computing infrastructure has had to be adapted. This is the case again this year, when we will use Spark on the RCC. Students are requested to be active participants to help "break in" this new system. The staff and course assistants will be working hard to help things go as smoothly as possible. The assignments are intended to be engaging and fun—but the data and computing environment will hold some surprises. An adventurous mindset is requested.