ECE 398BD: Making Sense of Big Data

Fall 2017

http://courses.engr.illinois.edu/ece398BD

Instructors: Yoram Bresler, Lav Varshney and Pramod Viswanath.

Course Coordinator: Yoram Bresler

Prerequisites: ECE 313 (or campus equivalent on basic undergrad probability) and some basic linear algebra. General mathematical maturity expected of engineering undergraduates.

Textbook: None. Relevant course notes will be handed out to the students.

Target Audience: Juniors or Seniors

Outline: Big Data is all around us. Petabytes of data is collected by Google and Facebook. 24 hours of video is uploaded on Youtube every minute. Making sense of all this data in the relevant context is a critical question. This course takes a holistic view towards understanding how this data is collected, represented and stored, retrieved and computed/analyzed upon to finally arrive at appropriate outcomes for the underlying context. The course is divided into three parts, with the first part focusing on foundations of machine learning, and the remaining two on specific application areas. Each application topic is covered at four discrete levels.

- We start with the context of where the data comes from, how it is acquired, what are the biases and noise levels in the data leading to statistical and physical models of the data acquired.
  Appropriate data representation mechanisms and distributed storage and computing architectures are discussed next. Based on the type of the data, different compression/ coding methods are appropriate. Images, videos, genomic data, medical imaging data, smart grid data, each bring their own unique characteristics which can be harnessed towards efficient representation.
- Once data is stored and represented efficiently, we look for the right statistical and algorithmic tools to analyze the data. Spectral methods (including Fourier methods and PCA), Clustering algorithms, SVM, Mining algorithms are studied in the specific context of the data.
- Finally, the analyzed data leads to appropriate inferences or visualizations as appropriate to the physical problem we started out with. This closes the loop bringing utility to the original setting and context in which the data was acquired.

For Fall 2017 the application areas will be:

- Computational Social Science
- Network Analytics

Course Plan

Part 1 (Weeks 1-5): Foundations of Machine Learning

Lecture 1: Introduction to the course; Review of Linear Algebra and Probability
Lecture 2: k-Nearest Neighbor Classifiers and Bayes Classifiers
Lecture 3: Linear Classifiers and Linear Discriminant Analysis
Lecture 4: Naïve Bayes, Logistic Regression, SVM
Lecture 5: Kernel Tricks and Model Selection
Lecture 6: K-Means Clustering
Lecture 7: Linear Regression
Lecture 8: SVD and Eigen-Decomposition
Lecture 9: Principal Component Analysis
Lecture 10: Optimization Techniques for Machine Learning. Q&A

Labs (Weeks 1-5)
Lab 1: Introduction to Python and the Canopy environment
Lab 2: Linear Classification: k-NN and LDA
Lab 3: Linear Classification: SVM
Lab 4: Clustering and Linear Regression
Lab 5: Eigen-Decompositions, SVD and PCA

Grading: 30% pre-lab quizzes (in class), 70% labs and lab reports.

Part 2 (Weeks 6-10): TBD

Lecture 1:
Lecture 2:
Lecture 3:
Lecture 4:
Lecture 5:
Lecture 6:
Lecture 7:
Lecture 8:
Lecture 9: Concluding lecture.

Labs
Lab 1:
Lab 2:
Lab 3:
Lab 4:

Grading: 30% pre-lab quizzes (in class), 70% labs and lab reports

Part 3 (Weeks 11-15): TBD