ECE 398BD: Making Sense of Big Data

Spring 2016

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

Instructors: Minh Do, Venu Veeravalli, and Pramod Viswanath.

Course Coordinator: Pramod Viswanath

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

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 based around three topics, drawn from diverse areas of engineering. The topics for the Spring 2016 offering are the following.

- Machine learning form the basis of all big data algorithms. In this topic, we study the basic tools of machine learning, including classification, regression, clustering and dimensionality reduction. Each topic will be illustrated with simple real data sets, and should help students apply the correct statistical tools to other data sets such as those encountered in the remainder of the course.
- Social networks are a central part of the fabric of Web 2.0. In this topic we study models for messages ("rumors") spread across social networks. The goal is to understand the simple analytics to localize the source of the rumor. The topic has applications in identifying the source of computer viruses or real viruses (eg: identifying the so-called ``patient zero" in epidemics such as the recent Ebola scare). Using this application area, algorithms for analytics on graphs are studied in detail – examples include graph centrality and clique/community detection.
- Audio and Video data are widely available online. For example, camera phones that generate millions of pixels in milliseconds are carried around all the time by billions of people worldwide. Surveillance cameras in a typical company site generate about terabytes of video every day. These ubiquitous visual recording devices generate big unstructured data that provide gold mine for analytics. In this part of the course, students will learn how these data types are acquired, sampled and stored. Concrete analytics problem involving audio recognition (similar to the commercially available Shazam software) and object detection and monitoring system will be studied.

Each of these topics 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.

The idea is that each of the three topics/stories will be taken through the four cycles highlighted above. They are distinct enough to bring about different aspects of the four issues but similar in the sense that the overall plot of what is being covered is similar. An important component of this course is that it is project based. The three topics will be covered sequentially, each ending with a project and a concrete (software) laboratory outcome. The technical material needed is developed on a need-basis, drawing from previous courses (prerequisites), and online material (corequisites). The three topics are synchronized in the sense that the technical material is streamlined and once covered, will be amortized in later parts of the course.

Course Plan

Part 1 (Weeks 1-5): Fundamentals of Machine Learning by Venugopal Veeravalli

Lecture 1: Introduction to the course; Review of linear algebra and probability
Lecture 2: Elements of Machine Learning
Lecture 3: k-Nearest Neighbor Classifier and Bayes Classifier
Lecture 4: Linear Classifiers and Linear Discriminant Analysis
Lecture 5: Kernel Tricks and Support Vector Machines
Lecture 6: Linear Regression
Lecture 7: K-means Clustering
Lecture 8: SVD and Eigen-Decomposition
Lecture 9: PCA and Applications

Labs
Lab 1: Introduction to Python
Lab 2: Linear Classification
Lab 3: Kernel Tricks and SVMs
Lab 4: Linear Regression and Clustering
Lab 5: PCA

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

Part 2 (Weeks 6-10): Social Network Analytics by Pramod Viswanath

Lecture 1 Social network modeling as graphs. Static models and dynamic (spreading) models.
Lecture 2: Key quantities: missing links, source identification, community detection
Lecture 3: Identifying the source of rumors. Natural algorithms based on distance centrality.
Lecture 4: Source identification with time stamps and sampling.
Lecture 5: Spreading algorithms to hide the source of the rumor.

Lecture 6: Missing links in the graph – recommending friendships.
Lecture 7: Different algorithmic approaches to recommending friendships.
Lecture 8: Multiple communities in a social network. Cliques and Dense subgraphs.
Lecture 9: Community detection algorithms. Number of communities and users within a community.
Lecture 10: Introduction to linear and convex relaxations of the community detection problems and corresponding algorithms.

Labs
Lab 1: Implement distance centrality algorithms to identify source of rumors
Lab 2: Implement algorithms that can use timing meta-data to infer the source of the message.
Lab 3: Implement recommendation/prediction algorithms on social network graphs
Lab 4: Implement community detection algorithm based on near-clique identification.

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

Part 3 (Weeks 11-15): Audio and Visual Analytics by Minh N. Do

Lecture 1: Introduction to audio and visual analytics. Example applications.
Lecture 2: Signal acquisition and sampling. Audio and visual sensors.
Lecture 3: Audio spectral analysis: DFT, short-time Fourier transform
Lecture 4: Audio content identification. Example: Shazam system
Lecture 5: Visual global feature extraction: color histograms
Lecture 6: Visual local feature extraction: keypoint detection and description, SIFT
Lecture 7: Image search: query by example using color histogram and feature matching
Lecture 8: Video analytics: background modeling and subtraction
Lecture 9: Video analytics: motion detection and tracking
Lecture 10: Introduction to deep learning for audio and visual recognition

Labs
Lab 1: Audio and visual acquisition, inspection, and visualization in Python and OpenCV.
Lab 2: Spectral analysis; k-means clustering and classification of spectral feature vectors
Lab 3: Audio content identification; implement the Shazam algorithm
Lab 4: Image feature extraction in OpenCV/Python. Feature matching and visual search
Lab 5: Background subtraction; motion detection and tracking in video

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