ECE 398BD: Making Sense of Big Data

Spring 2017

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

Instructors: Venu Veeravalli, Olgica Milenkovic and Minh Do.

Course Coordinator: Venu Veeravalli

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 Spring 2017 the application areas will be:

- *Biological Data Analytics:* It may be argued that biology and medical sciences are the two disciplines with the fastest growing datasets and data repositories. It is nowadays common to refer to data being of genomic, rather than astronomic size. What is known as –Omics data gives invaluable information about the structure and composition of our genomes, our unique genetic markers, the communication activity between genes and other molecules, the structure of our building block proteins and many other health related issues. In this part of the course we will cover diverse topics of relevance in bioinformatics, ranging from de Bruijn graphs (used to stitch DNA sequence fragments produced by experiments into a complete DNA sequence) to suffix trees (used for efficient data representation) and community detection (used to identify cancer gene communities). You will also get acquainted with modern biological data acquisition technologies, data libraries and publicly available data processing software.

- *Audio and Video data analytics*: 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.

Course Plan

<u>Part 1 (Weeks 1-5): Foundations of Machine Learning</u>

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.

<u>Part 2 (Weeks 6-10): Audio and Video Analytics</u>

Lecture 1: Signal acquisition and sampling. Examples of audio, image and video sensors.
Lecture 2: Audio spectral analysis: DFT, short-time Fourier transform
Lecture 3: Audio content identification. Example: Shazam system
Lecture 4: Visual feature extraction: color histograms, SIFT features
Lecture 5: Image search: query by example using color histogram and feature matching
Lecture 6: Video analytics: scene change detection,
Lecture 7: Video analytics continued: activity detection
Lecture 8: Novel video analytics methods and algorithms based on compressed sensing
Lecture 9: Concluding lecture.

Labs
Lab 1: Audio and video acquisition, spectral analysis, and color histogram
Lab 2: Audio content identification (develop Shazam on cloud)
Lab 3: Visual feature extraction, and foreground/background segmentation in video
Lab 4: Putting the whole system together.

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

Part 3 (Weeks 11-15): Biological Data Analytics

Lecture 1: Introduction to bioinformatics. Biological data.
Lecture 2: Sequence alignment. Global vs local alignment. Dynamic programming.
Lecture 3: The Smith-Waterman and Needlman-Wunsch algorithms. BLAST.
Lecture 4. Suffix trees and the Burrows-Wheeler transform. Bowtie2.
Lecture 5: Dynamic programming for sequence folding prediction. Vienna and Mfold. Stochastic grammars for folding models.
Lecture 6: Sanger sequencing. Overview of Next Generation and Third Generation Sequencing technologies.
Lecture 7: Basics of graph theory. Genome assembly via de Bruijn Graphs. EULER and IDBA_UD.
Lecture 8: Statistical read error-correction for Illumina, PacBio and Oxford Nanopore sequencers. Quake.
Lecture 9: Biological data repositories and databases.
Lecture 10: Biological data compression. Reference-based compression. CRAM. Context-tree weighting.

Labs
Lab 1: Sequence alignment and applications of BLAST.
Lab 2: Bowtie and DNA forensics.
Lab 3: Genome assembly. Influence of sequencing errors on assembler accuracy.
Lab 4: -Omics data compression.
Lab 5: Genomic sequence amplification and primer selection.

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