ECE 365: Data Science and Engineering

Fall 2020

https://courses.grainger.illinois.edu/ece365/fa2020/index.html

Instructors: Venugopal Veeravalli, Subhonmesh Bose and Ilan Shomorony.

Course Coordinator: Venugopal 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 Fall 2019 the application areas will be:

- *Data science and genomics:* DNA sequencing technologies generate large amounts of data and can provide important insights into the biology of all living organisms. We will explore how data science is used to understand the genetic composition of an organism, how genetic variants determine phenotypes, and how genes regulate cell function.
- *Introduction to natural language processing:* Automatic processing of natural language texts to make sense of the meaning conveyed is of central importance to many human-centered applications of today. In this part of the course we will see how modeling different levels of natural language leads to making sense of the patterns of meaning conveyed by words. We will work with state-of-the-art approaches to natural language processing using publicly available datasets.

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, Kernel Tricks
Lecture 5: Logistic Regression, SVM and Model Selection
Lecture 6: K-Means Clustering and Applications
Lecture 7: Linear Regression and Applications
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% quizzes, 70% labs.

Part 2 (Weeks 6-10): Genomics

Lecture 1: Introduction to DNA sequencing technologies
Lecture 2: Sequence alignment I. Dynamic programming, Smith-Waterman algorithm
Lecture 3: Sequence alignment II. Min-hashes, sketching, and Jaccard similarity
Lecture 4: Genome assembly. De Bruijn graphs and string graphs
Lecture 5: Genome-wide association studies via logistic regression
Lecture 6: Introduction to RNA-seq and the RNA quantification problem
Lecture 7: RNA-seq quantification via the EM algorithm
Lecture 8: Single-cell RNA-seq I. Dimensionality reduction via PCA and t-SNE
Lecture 9: Single-cell RNA-seq II. k-means clustering, Gaussian mixture models

Grading: 30% quizzes, 70% labs.

Labs
Lab 1: Exploring DNA sequencing data
Lab 2: Genome-wide association studies and Manhattan plots
Lab 3: Quantifying RNA via the EM algorithm
Lab 4: Visualizing and clustering single-cell RNA-seq data

Part 3 (Weeks 11-15): Natural Language Processing

Lecture 1: Introduction to NLP. Words as units of text.
Lecture 2: Words in isolation: Bag-of-words models for text processing
Lecture 3: Text as word sequences: Language modeling
Lecture 4. Sequence labeling
Lecture 5: Understanding meaning: Lexical Semantics

Lecture 6: Distributional and distributed semantics
Lecture 7: Discourse
Lecture 8: Application: Machine translation
Lecture 9: Application: Machine translation
Lecture 10: Non-English NLP and Recap

Labs
Lab 1: Word frequency distributions and vocabulary curves
Lab 2: Text classification
Lab 3: Language Modeling
Lab 4: Word-embeddings
Lab 5: Machine translation

Grading: 30% quizzes, 70% labs.