ECE 365: Data Science and Engineering

Fall 2021

https://courses.grainger.illinois.edu/ece365/fa2021/index.html

Instructors: Venugopal Veeravalli, Subhonmesh Bose and Zhizhen Zhao.

Course Coordinator: Venugopal Veeravalli

Prerequisites: ECE 313 (or campus equivalent on basic undergrad probability) and some basic linear algebra (vectors, matrices, eigendecomposition, etc.). 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 2021 the application areas will be:

- *Machine learning for power systems:* Grid operation relies on efficient processing of data and identifying patterns in them. In this module, we explore applications of machine learning in grid operations. Specifically, we explore regression and classification tasks such as those that arise in load prediction, consumer electricity usage, recognizing valid power system measurements, and virtual bidding markets.
- *Data science and cryo-electron microscopy single particle analysis:* Cryo-electron microscopy is widely used to resolve 3D structures of macromolecules in their native states. In this module, we will understand how the image data is collected and explore how data science and machine learning are used to infer the protein conformations from noisy observational data.

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

Part 2 (Weeks 6-10): Smart Grid

Lecture 1: Introduction to power systems, basics of neural networks
Lecture 2: Neural networks and load prediction
Lecture 3: Power flow equations
Lecture 4: SVM for detecting corrupt power system measurements
Lecture 5: Detecting network structure
Lecture 6: Basics of electricity markets, virtual bidding
Lecture 7: Trading strategies for virtual bidding
Lecture 8: Wrapping up virtual bidding, understand customer data
Lecture 9: Logistic regression for customer data analysis
Lecture 10: Customer billing and cost savings from solar

Labs
Lab 1: Day-ahead load prediction in ERCOT markets
Lab 2: Detecting bad sensors in power system measurements
Lab 3: Virtual bidding in NYISO's markets
Lab 4: Analyze customer data from Austin, Texas.

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

Part 3 (Weeks 11-15):

Lecture 1: Introduction to cryo-electron microscopy single particle analysis
Lecture 2: Basics of the image formation in cryo-electron microscopy
Lecture 3: Automatic particle picking: semi-supervised and unsupervised classification
Lecture 4: Geometry of the single particle image data and image denoising
Lecture 5: Single particle reconstruction using common-lines
Lecture 6: Single particle reconstruction using maximum likelihood estimation and maximum a posteriori estimation
Lecture 7: Generative adversarial learning for single particle reconstruction
Lecture 8: Variational autoencoder for exploring structural heterogeneity
Lecture 9: From electron density map to protein atomic structures

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

Labs
Lab 1: Exploring cryo-electron microscopy image data
Lab 2: Automatic particle picking
Lab 3: Cryo-electron microscopy single particle reconstruction
Lab 4: Visualizing protein structural heterogeneity