- **Section Number:** 01
- **Course Name:** Machine Learning & Advanced Analytics for Biomedicine
- **Course Number:** CCTS 40500
- **Units:** 100 units
- **Instructor/s:** Ishanu Chattopadhyay (ishanu@uchicago.edu)
- **Prerequisites/Remarks:** Basic familiarity with coding in python, Linux commandline, RCC cluster environment familarity will be helpful, but not necessary
- **Enrollment limit:** None
- **Is the course undergrad, graduate, or mixed level:** Mixed Level
- **Cross list: CCTS 20500, BIOS 29208**

Course Description

Easy accessibility to data is rapidly transforming scientific research, and advanced analytics powered by sophisticated learning algorithms is uncovering new insights, and solutions to hard problems. The goal of this course is to provide an introductory overview of the key concepts in machine learning, outlining the potential applications in biomedicine and sociology. Beginning from basic statistical concepts, we will discuss implementations of standard and state of the art classification and prediction algorithms, and go on to discuss more advanced topics in unsupervised learning, deep learning architectures, and stochastic time series analysis. We will also cover emerging ideas in data-driven causal inference, and demonstrate applications in uncovering etiological insights from large scale datasets including clinical databases of electronic health records, publicly available sequence and omics datasets in biology, and large scale geospatial datasets in sociology.

Learning Goals

The acquisition of hands-on skills will be emphasized over machine learning theory. On successfully completing the course, students will have acquired enough knowledge of the underlying machinery to intuit and implement solutions to non-trivial data science problems. Rudimentary knowledge of probability theory, and basic exposure to scripting languages such as python is required.

Planned Assessments

Students will be required to turn in solutions to take-home modeling problems, with scripted software in Python. There will be one mid-term assignment, and one final assignment/project. Homework problems will be assigned regularly but not weekly. Final assessment will depend on the performanace on the assignments and the homework problems, and the level of engagement and interest perceived by the instructor.

Tentative Syllabus

1) Introduction to Automated Inference, Machine Learning, Probability Theory, & Statistical Modeling of Data
2) Review of Linear Algebra & Basic Probability theory
3) Bayesian Inference
4) Linear and Logistic Regression, with discussion of LASSO and Ridge Regression
5) Concepts of Overfitting, & Regularization
6) The SkLearn Python Library
7) Support Vector Machines & Support Vector Regression
8) Decision Trees, Random Forests, Extremely Randomized Trees, & Boosting
9) Convolutional Neural Nets (CNN) in Image Classification
10) Introduction to the Tensorflow Library
11) Recurrent Nets (RNN) and Long Short-term Memory (LSTM) Architectures
12) Introduction to Stochastic Processes, and Time-series Modeling (ARIMA, GARCH)
13) Predictive State Representations (PSR) & Observable Operator Models (OOM)
14) Probabilistic Finite Automata in Modeling Stochastic Time Series
15) Zero-knowledge Anomaly Detection
16) ML-enabled Tools In Personalized Medicine, and social policy optimization
17) Exploring Diagnosis & Screening With Electronic Health Records

Texts

[1] C. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer, 2006.
[2] I. Goodfellow, Y. Bengio, A. Courville, and F. Bach, *Deep Learning*, MIT Press, 2016.
[3] K. Murphy and F. Bach, *Machine Learning: A Probabilistic Perspective*, Adaptive Computation and Machi, MIT Press, 2012.
[4] G. Varoquaux, *Scikit-learn tutorial: statistical-learning for sientific data processing*. http://gael-varoquaux.info/scikit-learn-tutorial/, 2010.