STAT GU 4541, Honors Statistical Machine Learning

Time: Mon-Wed 10:10-11:25 am. **Location:** Hamilton 503.

Instructor: Samory Kpotufe. *email:* skk2175@columbia.edu,
Office hours: 30 mn right after each class.

Assistant Instructor: Victor Daniel. *email:* vdd2111@columbia.edu,

Office hours:
- In Person (to be announced)
- Zoom (to be announced)

Ed Time: Tuesday, Thursday 8-9pm.

Evaluation: 4 Homeworks, allowing collaboration (however turn in separate solutions), 2 take home exams, and class participation (assessed via 5-10 mn pop up quizzes, 2-3 of which will be dropped). Programming questions will assume Python, however you may choose any language.

Homeworks and exams each count for 15 percent of the total grade, while quizzes count for 10 percent. The actual number of quizzes will be determined over time.

Course Overview:

This is an introduction to machine learning from a statistical perspective, aimed at students with good mathematical preparation: you will be able to do well if you've done well in the required courses below (see pre-requisites). Although the topics will be the same as in any introductory course on the subject, the class aims to get into more of the actual mathematical insights behind popular ML procedures. This is not a proof-based course, but you will be expected to be at ease with basic algebraic manipulations commensurate with the pre-requisites. Note that I taught the same course under the number 4241 in Spring 2024 and 2025; the new "honors" designation is to reflect the depth of understanding we're aiming at: while the material requires some work to keep up with, you should find it intellectually rewarding as we proceed to demystify much of ML.

In particular, we emphasize a *statistical perspective*, i.e., a distinction between what may be *learned from samples* and the *underlying population patterns* being estimated: in applications of ML, we are always interested in performance of at the population level (rather than on a test sample), and such population-level performance ultimately drives algorithmic design, by trading-off with computational considerations (e.g., space and time efficiency).

Major families of algorithms are covered, from *unsupervised* procedures for clustering, to *supervised* procedures for classification and regression.

Some keywords: k-means, EM, Gradient Descent, SVMs, Linear and Polynomial Regression, Decision trees, Boosting, Perceptron, Neural Networks (basic introduction).

Note: Many of these algorithms are introduced in homeworks, while we focus in class on major insights and design paradigms. I will strive to give you some sense of where machine learning stands at the moment (w.r.t. the grand vision of AGI), including which questions remain unanswered.

Pre-requisites:

- **Probability:** random variables and random vectors, common distributions such as multivariate Gaussians, moments, conditional probabilities and expectations.

- **Statistics:** point estimation (MoM, MLE), bias, variance, confidence intervals.

- **Linear Algebra:** matrices, eigenvalues and eigenvectors.
- **Multivariate calculus:** gradients, hessians.
- **Programming:** Familiarity with scientific programming, e.g., Matlab, R, Python.

Books: Below are some useful companion books, listed (essentially) in order of difficulty.

- Duda, Richard O., and Peter E. Hart. Pattern classification. John Wiley & Sons, 2006.
- Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. New York: springer, 2009.
- Mohri, Mehryar. "Foundations of machine learning." MIT press, 2018.
- Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning. Cambridge press, 2014.