STAT GU 4241, Statistical Machine Learning

Time: Mon-Wed 10:10-11:25 am. **Location:** 140 Uris Hall.

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

Assistant Instructor: Tianjun Ke. *email:* tk3103@columbia.edu,

Office hours:
- In Person, Tuesday, Thursday 4:30-5:30pm, Uris 324
- Zoom (Meeting ID: 850 7538 9505 Passcode: 1HYEGh) Monday 1:30-2:30pm .

Ed Time: Tuesday, Thursday 8-9pm.

Evaluation: Homeworks, allowing collaboration (however turn in separate solutions), and two take home exams. Programming questions will assume Python, however you may choose any language.

All assignments, including exams, will count the same towards the final grade.

Course Overview:

This is an introduction to machine learning from a statistical perspective. While we will cover many of the same introductory elements of machine learning as courses in other departments, the statistical perspective emphasizes the distinction between what may be *learned from samples* and the *underlying population patterns* being estimated. Major families of algorithms will be 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 procedures will be introduced in homeworks, rather than in class, while class material will focus instead on general insights on the huge variety of procedures available in machine learning. I will strive to give you some sense of where machine learning stands at the moment in the grand vision of AI, and 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.