Speaker: Emmanuel J. Candès, Stanford University

Title: What do we really know about logistic regression?
A modern maximum-likelihood theory

Abstract:

Every student in statistics or data science learns early on that when the sample size largely exceeds the number of variables, fitting a logistic model produces estimates that are approximately unbiased. Every student also learns that there are formulas to predict the variability of these estimates which are used for the purpose of statistical inference; for instance, to produce p-values for testing the significance of regression coefficients. Although these formulas come from large sample asymptotics, we are often told that we are on reasonably safe grounds when $n$ is large in such a way that $n \geq 5p$ or $n \geq 10p$. We show that this is far from the case, and consequently, inferences routinely produced by common software packages are often unreliable.

Consider a logistic model with independent features in which $n$ and $p$ become increasingly large in a fixed ratio. Then we show that (1) the MLE is biased, (2) the variability of the MLE is far greater than classically predicted, and (3) the commonly used likelihood-ratio test (LRT) is not distributed as a chi-square. The bias of the MLE is extremely problematic as it yields wrong predictions for the probability of a case based on observed values of the covariates. We develop a new theory, which asymptotically predicts (1) the bias of the MLE, (2) the variability of the MLE, and (3) the distribution of the LRT. We empirically also demonstrate that these predictions are extremely accurate in finite samples and produce valid inference. Further, an appealing feature is that these novel predictions depend on the unknown sequence of regression coefficients only through a single scalar, the overall strength of the signal. This suggests very concrete procedures to adjust inference; we describe one such procedure learning a single parameter from data and producing accurate inference.

This is joint work with Pragya Sur.