THE UNIVERSITY OF CHICAGO

Department of Statistics

STATISTICS COLLOQUIUM

LUCAS JANSON

Department of Statistics
Harvard University

Model-X Knockoffs For High-Dimensional Controlled Variable Selection

MONDAY, November 13, 2017 at 4:30 PM
Eckhart 133, 5734 S. University Avenue
Refreshments before the seminar at 4:00PM in Jones 111

ABSTRACT

Many contemporary large-scale applications involve building interpretable models linking a large set of potential covariates to a response in a nonlinear fashion, such as when the response is binary. Although this modeling problem has been extensively studied, it remains unclear how to effectively select important covariates while controlling the fraction of false discoveries, even in high-dimensional logistic regression, not to mention general high-dimensional nonlinear models. To address such a practical problem, we propose a new framework of model-X knockoffs, which reads from a different perspective the knockoff procedure (Barber and Candès, 2015) originally designed for controlling the false discovery rate in linear models. The key innovation of our method is to construct knockoff variables probabilistically instead of geometrically. This enables model-X knockoffs to deal with arbitrary (and unknown) conditional models and any dimensions, including when the dimensionality p exceeds the sample size n, while the original knockoffs procedure is constrained to homoscedastic linear models with n greater than or equal to p. Our approach requires the design matrix be random (independent and identically distributed rows) with a covariate distribution that is known, although we show preliminary evidence that our procedure is robust to unknown/estimated distributions. As we require no knowledge/assumptions about the conditional distribution of the response, we effectively shift the burden of knowledge from the response to the covariates, in contrast to the canonical model-based approach which assumes a parametric model for the response but very little about the covariates. To our knowledge, no other procedure solves the controlled variable selection problem in such generality, but in the restricted settings where competitors exist, we demonstrate the superior power of knockoffs through simulations. Finally, we apply our procedure to data from a case-control study of Crohn's disease in the United Kingdom, making twice as many discoveries as the original analysis of the same data.

For further information and inquiries about building access for persons with disabilities, please contact Jonathan Rodriguez at 773.702.8333 or send him an email at jgrodriguez@galton.uchicago.edu. If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/subscribe/statseminars.