THE UNIVERSITY OF CHICAGO

Department of Statistics

STATISTICS COLLOQUIUM

GARVESH RASKUTTI

Department of Statistics
University of Wisconsin-Madison

Variable Selection Using Presence-Only Data with Applications to Biochemistry

MONDAY, February 26, 2018 at 4:30 PM
Eckhart 133, 5734 S. University Avenue
Refreshments before the seminar at 4:00PM in Jones 111

ABSTRACT

In a number of domains, we are presented with a class-action problem involving positive and unlabelled data, referred to as presence-only responses. The application I present today involves studying the relationship between protein sequence and function and presence-only data arises since for many experiments it is impossible to obtain a large set of negative (non-functional) sequences. Furthermore, if the number of variables is large and the goal is variable selection (as in this case), a number of statistical and computational challenges arise due to the non-convexity of the objective. In this talk, I present an algorithm (PUlasso) with provable guarantees for doing variable selection and classification with presence-only data. Our algorithm involves using the majorization-minimization (MM) framework which is a generalization of the well-known expectation-maximization (EM) algorithm. In particular to make our algorithm scalable, our algorithm has two computational speed-ups to the standard EM algorithm. I provide a theoretical guarantee where we first show that our algorithm is guaranteed to converge to a stationary point, and then prove that any stationary point achieves the minimax optimal mean-squared error of slogp/n, where s is the sparsity of the true parameter. I also demonstrate through simulations that our algorithm out-performs state-of-the-art algorithms in the moderate p settings in terms of classification performance. Finally, I demonstrate that our PUlasso algorithm performs well on a biochemistry example.

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.