THE UNIVERSITY OF CHICAGO

Department of Statistics

STATISTICS COLLOQUIUM

EDGAR DOBRIBAN

Statistics Department
Wharton University of Pennsylvania

How to Deal with Big Data? Understanding Large-scale Distributed Regression

MONDAY, October 15, 2018 at 4:30 PM
Eckhart 133, 5734 S. University Avenue
Refreshments before the seminar at 4:00PM in Jones 111

ABSTRACT

Modern massive datasets pose an enormous computational burden to practitioners. Distributed computation has emerged as a universal approach to ease the burden: Datasets are partitioned over machines, which compute locally, and communicate short messages. Distributed data also arises due to privacy reasons, such as in medicine. It is important to study how to do statistical inference and machine learning in a distributed setting. In this talk, we present results about one-step parameter averaging in statistical linear models under data parallelism. We do linear regression on each machine, and take a weighted average of the parameters. How much do we lose compared to doing linear regression on the full data? Here we study the performance loss in estimation error, test error, and confidence interval length in high dimensions, where the number of parameters is comparable to the training data size. We discover several key phenomena. First, averaging is not optimal, and we find the exact performance loss. Second, different problems are affected differently by the distributed framework. Estimation error and confidence interval length increases a lot, while prediction error increases much less. These results match numerical simulations and a data analysis example. We rely on recent results from random matrix theory, where we develop a new calculus of deterministic equivalents as a tool of broader interest.

This is joint work with Yue Sheng, available at https://arxiv.org/abs/1810.00412.

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.