THE UNIVERSITY OF CHICAGO

Department of Statistics

STATISTICS COLLOQUIUM

MICHAEL MAHONEY

Department of Statistics
University of California, Berkeley

Terabyte-sized Computational Statistics

MONDAY, November 21, 2016, at 4:00 PM
Eckhart 133, 5734 S. University Avenue
Refreshments following the seminar in Jones 111

ABSTRACT

When dealing with data of terabyte-size scale and beyond, computing even basic descriptive statistics can be a challenge, and computing finer statistical properties such as correlations can be very non-trivial. The talk will provide an overview of recent work at the interface of methods (statistical and algorithmic theory, etc.), implementations (on single machine versus distributed data center versus supercomputer), and applications (in science, e.g., genetics, astronomy, and climate science, as opposed to internet and social media) that aim to provide tools to perform bread-and-butter computational statistics on data up to and beyond terabyte-size scales. A key issue here is that in statistics one is often primarily interested in correlational properties of the data, and thus one must go beyond database-like query/counting operations on flat tables to deal with more complex couplings that are implicit when one is modeling data with matrices. As an example, one of the most straightforward formulations of the machine learning problem of feature selection boils down to the linear algebraic problem of selecting good columns from a data matrix. This formulation has the advantage of yielding features that are interpretable to scientists in the domain from which the data are drawn, an important consideration when machine learning methods are applied to realistic scientific data. While simple, this problem is central to many other seemingly nonlinear learning methods. Moreover, while unsupervised, this problem also has strong connections with related supervised learning methods such as Linear Discriminant Analysis and Canonical Correlation Analysis. We will describe recent work implementing Randomized Linear Algebra algorithms for this feature selection problem (as well as related NMF and PCA problems) in parallel and distributed environments on inputs of size ranging from ones to tens of terabytes, as well as the application of these implementations to specific scientific problems in areas such as mass spectrometry imaging and climate modeling.

For further information and about building access for persons with disabilities, please contact Courtney Tillman at 773.702.8333 or send email (cmtillman@galton.uchicago.edu). If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/arc/statseminars.