Department of Applied and Computational Mathematics and Statistics Colloquium

Min Yang

Department of Mathematics, Statistics
& Computer Science
University of Illinois, Chicago

On Data Reduction of Big Data

Extraordinary amounts of data are being produced in many branches of science. Proven statistical methods are no longer applicable with extraordinary large datasets due to computational limitations. A critical step in big data analysis is data reduction. Existing investigations in the context of linear regression focus on subsampling-based methods. However, not only is this approach prone to sampling errors, it also leads to a covariance matrix of the estimators that is typically bounded from below by a term that is of the order of the inverse of the subdata size. We propose a novel approach, termed information-based optimal subdata selection (IBOSS). Compared to leading existing subdata methods, the IBOSS approach has the following advantages: (i) it is significantly faster; (ii) it is suitable for distributed parallel computing; (iii) the variances of the slope parameter estimators converge to 0 as the full data size increases even if the subdata size is fixed, that is, the convergence rate depends on the full data size; (iv) data analysis for IBOSS subdata is straightforward and the sampling distribution of an IBOSS estimator is easy to assess. Theoretical results and extensive simulations demonstrate that the IBOSS approach is superior to subsampling-based methods, sometimes by orders of magnitude. The advantages of the new approach are also illustrated through analysis of real data.

Wednesday, November 14, 2018
4:15 PM – 5:15 PM
127 Hayes-Healy Center

Colloquium Tea 3:45 PM to 4:15 PM Hurley Hall Globe Area