Seminar: 11/14/2022

Lin Wang
Assistant Professor in Department of Statistics at Purdue University

Title: Balanced Subsampling for Regression with Big Data

Abstract: The dramatic growth of big datasets presents a new challenge to data storage and analysis. Data reduction, or subsampling, that extracts useful information from datasets is a crucial step in big data analysis. I will introduce a series of balanced subsampling approaches for big data with a focus on regression models under different settings. The merits of the proposed approaches are three-fold: (i) they are easy to implement and fast; (ii) they are suitable for distributed parallel computing and ensure the subsamples selected in different batches have no common data points; and (iii) they outperform existing methods in minimizing the mean squared errors of the estimated parameters (in parametric regression) and the mean integrated squared errors of the regression function (in nonparametric regression). Theoretical results and extensive numerical results show that the proposed approaches are superior to existing subsampling approaches. The advantages of the balanced subsampling approaches are also illustrated through the analysis of real-life examples.