Special Colloquium

Informative sampling of large database

Prof. Wei Zheng (Indiana University–Purdue University Indianapolis)

Abstract: For many tasks of data analysis, a large database of explanatory variables is readily available, however, the responses are missing and expensive to obtain. A natural remedy is to judiciously select a sample of the data, for which the responses are to be measured. In this paper, we adopt the classical criteria in design of experiments to quantify the information of a given sample. Then, we provide a theoretical justification for approximating the optimal sample problem by a continuous problem, for which fast algorithms can be further developed with the guarantee of global convergence. Our approach exhibits the following features: (i) The statistical efficiency of any candidate sample can be evaluated without knowing the exact optimal sample; (ii) It can be applied to a very wide class of statistical models; (iii) It can be integrated with a broad class of information criteria; (iv) It is scalable for big data.

Friday, January 20 at 3:00 PM in SEO 636