Kernel-based Regression on Massive Data

Lei Shi
School of Mathematical Sciences, Fudan University, Shanghai, China
Email: leishi@fudan.edu.cn

Kernel methods are attractive in data analysis as they can model nonlinear similarities between observations and provide means to rich representations, both of which are useful for the regression problems in general domains. Despite their popularity, they suffer from two primary inherent drawbacks. One drawback is the positive definiteness requirement of the kernel functions, which greatly restricts their applications to some real data analysis. The other drawback is their poor scalability in massive data scenarios. In this talk, we aim to address these two problems by considering the distributed and Nyström subsampling approaches for coefficient-based regularized regression. Distributed learning and Nyström subsampling are two effective approaches to analyze big data, which serve as standard tools for reducing computational complexity in machine learning problems where massive data sets are involved. Coefficient-based regularized regression can provide a simple paradigm for designing indefinite kernel methods. We show that combinations of these two schemes with coefficient-based regularized regression are not only computationally efficient but also statistically consistent with a mini-max optimal rates of convergence.