Distributed Learning for Kernel Mode-Based Regression

Tao Wang*

April 23, 2023

Abstract

We in this paper propose a parametric kernel mode-based regression built on mode value, which can achieve robust and efficient estimators when the data have outliers or heavy-tailed distributions. We show that the resultant estimators can arrive at the highest asymptotic breakdown point of 0.5. We then utilize such a regression for massive datasets by combining it with the distributed statistical learning technique, which can greatly reduce the required amount of primary memory while simultaneously incorporating heterogeneity into the estimation procedure. By approximating the local kernel objective function using a least squares format, we are able to preserve compact statistics for each worker machine and employ them to rebuild the estimate of the entire dataset with asymptotically minimal approximation error. With the help of a Gaussian kernel, an iteration algorithm built on expectation-maximization procedure is introduced, which could substantially lessen the computational burden. The asymptotic properties of the developed mode-based estimators are established, where we prove that the suggested estimator for massive datasets is statistically as efficient as the global mode-based estimator using the full dataset. We further conduct a shrinkage estimation based on the local quadratic approximation and demonstrate that the resulting estimator has the oracle property by employing an adaptive LASSO approach. The finite sample performance of the developed method is illustrated using simulations as well as real data analysis.

Keywords: Distributed learning, Massive data, Mode-based, Variable selection.

JEL Classifications: C01, C14, C55.

*Department of Economics, University of Victoria, Victoria, BC V8W 2Y2, Canada. E-mail: taow@uvic.ca.