Scalable Learning on Distributions and Functions

Junier B. Oliva

Carnegie Mellon University
*Contact author: joliva@cs.cmu.edu.com

Keywords: Kernel Methods, Nonparametric Statistics, Scalability

A great deal of attention has been applied to studying new and better ways to perform learning tasks involving static finite vectors. Indeed, over the past century the fields of statistics and machine learning have amassed a vast understanding of various learning tasks like density estimation, clustering, classification, and regression using simple real valued vectors. However, we do not live in a world of simple objects. From the contact lists we keep, the sound waves we hear, and the distribution of cells we have, complex objects such as sets, functions, and distributions are all around us. Furthermore, with ever-increasing data collection capacities at our disposal, not only are we collecting more data, but richer and more bountiful complex data are becoming the norm.

In this presentation we analyze functional regression problems where input covariates, and possibly output responses, are functions from a nonparametric function class. Such problems cover a large range of interesting applications including time-series prediction problems, and also more general tasks like studying a mapping between two separate types of distributions.

However, previous nonparametric estimators for functional regression problems scale badly computationally with the number of input/output pairs in a data-set. Yet, given the complexity of functional data it may be necessary to consider large data-sets in order to achieve a low estimation risk.

To address this issue, we present two novel scalable nonparametric estimators: the Double-Basis Estimator (2BE) for function-to-real regression problems; and the Triple-Basis Estimator (3BE) for function-to-function regression problems. Both the 2BE and 3BE can scale to massive data-sets. We show an improvement of several orders of magnitude in terms of prediction speed and a reduction in error over previous estimators in various synthetic and real-world data-sets.