Regression on Probability Measures: A Simple and Consistent Algorithm*

Zoltán Szabó (Gatsby Computational Neuroscience Unit, University College London)†

Abstract

We address the distribution regression problem: we regress from probability measures to Hilbert-space valued outputs, where only samples are available from the input distributions. Many important statistical and machine learning problems can be phrased within this framework including point estimation tasks without analytical solution, or multi-instance learning. However, due to the two-stage sampled nature of the problem, the theoretical analysis becomes quite challenging: to the best of our knowledge the only existing method with performance guarantees requires density estimation (which often performs poorly in practise) and the distributions to be defined on a compact Euclidean domain. We present a simple, analytically tractable alternative to solve the distribution regression problem: we embed the distributions to a reproducing kernel Hilbert space and perform ridge regression from the embedded distributions to the outputs. We prove that this scheme is consistent under mild conditions (for distributions on separable topological domains endowed with kernels), and construct explicit finite sample bounds on the excess risk as a function of the sample numbers and the problem difficulty, which hold with high probability. Specifically, we establish the consistency of set kernels in regression, which was a 15-year-old-open question, and also present new kernels on embedded distributions. The practical efficiency of the studied technique is illustrated in supervised entropy learning and aerosol prediction using multispectral satellite images.

Preprint: `http://arxiv.org/abs/1411.2066`

Code: `https://bitbucket.org/szzoli/ite/`

*Centre for Research in Statistical Methodology (CRiSM) Seminars, Department of Statistics, University of Warwick, UK; May 29, 2015; abstract.

†Joint work with Bharath K. Sriperumbudur (Department of Statistics, Pennsylvania State University), Barnabás Póczos (Machine Learning Department, Carnegie Mellon University), Arthur Gretton (Gatsby Computational Neuroscience Unit, University College London).