Vector-valued Distribution Regression: A Simple and Consistent Approach*

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

Abstract

We address the distribution regression problem (DRP): regressing on the domain of probability measures, in the two-stage sampled setup when only samples from the distributions are given. The DRP formulation offers a unified framework for several important tasks in statistics and machine learning including multi-instance learning (MIL), or point estimation problems without analytical solution. Despite the large number of MIL heuristics, essentially there is no theoretically grounded approach to tackle the DRP problem in the two-stage sampled case. To the best of our knowledge, the only existing technique with consistency guarantees requires kernel density estimation as an intermediate step (which often scale poorly in practice), and the domain of the distributions to be compact Euclidean. We analyse a simple (analytically computable) ridge regression alternative to DRP: we embed the distributions to a reproducing kernel Hilbert space, and learn the regressor from the embeddings to the outputs. We show that this scheme is consistent in the two-stage sampled setup under mild conditions, for probability measure inputs defined on separable, topological domains endowed with kernels, with vector-valued outputs belonging to an arbitrary separable Hilbert space. Specially, choosing the kernel on the space of embedded distributions to be linear and the output space to the real line, we get the consistency of set kernels in regression, which was a 15-year-old open question. In our talk we are going to present (i) the main ideas and results of consistency, (ii) concrete kernel constructions on mean embedded distributions, and (iii) two applications (supervised entropy learning, aerosol prediction based on multispectral satellite images) demonstrating the efficiency of our approach.

Paper: `http://arxiv.org/pdf/1402.1754`
Code: `https://bitbucket.org/szzoli/ite/`

Acknowledgments. This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350. The work was carried out while Bharath K. Sriperumbudur was a research fellow in the Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge, UK.

*Statistical Science Seminars, UCL, London, United Kingdom, October 9, 2014; abstract.

†Joint work with Arthur Gretton (Gatsby Computational Neuroscience Unit, University College London), Barnabás Póczos (Machine Learning Department, Carnegie Mellon University), Bharath K. Sriperumbudur (Statistics Department, Pennsylvania State University); the ordering of the second through fourth authors is alphabetical.