
Performance guarantees for kernel-based learning on probability distributions*

Zoltán Szabó (Gatsby Unit, University College London)

Abstract

In this talk I will present a novel result concerning the theory of distribution regression (DR). The DR problem addresses regression to vector-valued outputs from probability measures. Many important machine learning and statistical tasks fit into this framework, including multi-instance learning, and point estimation problems without analytical solution. Despite the large number of available heuristics, the inherent two-stage sampled nature of the problem (in practice only samples from sampled distributions are observable) makes the theoretical analysis quite challenging. To the best of our knowledge, the only existing technique with consistency guarantees for DR requires density estimation (which often performs poorly in practice), and the domain of the distributions to be compact Euclidean. I am going to present a simple, analytically computable, kernel ridge regression-based alternative to DR with an exact computational-statistical efficiency tradeoff analysis. The established result shows that the studied estimator is not only consistent (which specifically answers a 17-year-old open question), but it is also able to match the one-stage sampled minimax optimal rate. Moreover, this distribution-regression algorithm performs as well in practice as the state-of-the-art, task-specific solution in an aerosol prediction problem. [Joint work with Bharath Sriperumbudur, Barnabás Póczos, Arthur Gretton.]

Preprint: <http://arxiv.org/abs/1411.2066>,
Code: <https://bitbucket.org/szoli/ite/>.

*Max Planck Institute, 15 March 2016; abstract.