Distribution Regression: Computational-Statistical Efficiency Tradeoff*

Zoltán Szabó (Gatsby Unit, University College London)

Abstract: In this talk I am going to focus on the distribution regression problem: regressing to vector-valued outputs from probability measures. Many important machine learning and statistical tasks fit into this framework, including multi-instance learning or point estimation problems without analytical solution such as hyperparameter or entropy estimation. Despite the large number of available heuristics in the literature, the inherent two-stage sampled nature of the problem makes the theoretical analysis quite challenging: in practice only samples from sampled distributions are observable, and the estimates have to rely on similarities computed between sets of points. To the best of our knowledge, the only existing technique with consistency guarantees for distribution regression requires density estimation as an intermediate step (which often performs poorly in practice), and the domain of the distributions to be compact Euclidean. I propose a simple, analytically computable, ridge regression based alternative to distribution regression by embedding the distributions to a reproducing kernel Hilbert space, and learning the regressor from the embeddings to the outputs. I am going to present the main ideas why this scheme is consistent in the two-stage sampled setup under mild conditions (on separable topological domains enriched with kernels) and present an exact computational-statistical efficiency tradeoff description showing that the studied estimator is able to match the one-stage sampled minimax optimal rate. Specifically, this result answers a 16-year-old open question by establishing the consistency of the classical set kernel [Haussler, 1999; Gärtner et. al, 2002] in regression, and also covers more recent kernels on distributions, including those due to [Christmann and Steinwart, 2010]. [Joint work with Bharath Sriperumbudur, Barnabás Póczos, Arthur Gretton.]

Preprint: `http://arxiv.org/abs/1411.2066`
Code: `https://bitbucket.org/szzoli/ite/`

Bio: Zoltán Szabó is a Research Associate at the Gatsby Unit, University College London (2013 - present). He holds a double PhD in Computer Science and Applied Mathematics from the Eötvös Loránd University (2009-2012; Budapest, Hungary). His primary research interests are information theory, statistical machine learning, empirical processes and kernel methods with applications in remote sensing (sustainability), distribution regression, structured sparsity, independent subspace analysis and its extensions, collaborative filtering.

*Carnegie Mellon University: ML Lunch Seminar, 30 November 2015; abstract.