Simple Consistent Distribution Regression on Compact Metric Domains*

Zoltán Szabó, Arthur Gretton†
Gatsby Computational Neuroscience Unit
CSML, University College London
Alexandra House, 17 Queen Square
London - WC1N 3AR, UK
zoltan.szabo@gatsby.ucl.ac.uk
arthur.gretton@gmail.com

Barnabás Póczos
Machine Learning Department
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue Pittsburgh PA 15213 USA
bapoczos@cs.cmu.edu

Bharath Sriperumbudur
Statistical Laboratory
Center for Mathematical Sciences
Wilberforce Road, Cambridge CB3 0WB
bharathsv.ucsd@gmail.com

Abstract

In a standard regression model, one assumes that both the inputs and outputs are finite dimensional vectors. We address a variant of the regression problem, the distribution regression task, where the inputs are probability measures. Many important machine learning tasks fit naturally into this framework, including multi-instance learning, point estimation problems of statistics without closed form analytical solutions, or tasks where simulation-based results are computationally expensive. Learning problems formulated on distributions have an inherent two-stage sampled challenge: only samples from sampled distributions are available for observation, and one has to construct estimates based on these sets of samples. We propose an algorithmically simple and parallelizable ridge regression based technique to solve the distribution regression problem: we embed the distributions to a reproducing kernel Hilbert space, and learn the regressor from the embeddings to the outputs. We show that under mild conditions (for probability measures on compact metric domains with characteristic kernels) this solution scheme is consistent in the two-stage sampled setup. Specially, we establish the consistency of set kernels in regression (a 15-year-old open question) and offer an efficient alternative to existing distribution regression methods, which focus on compact domains of Euclidean spaces and apply density estimation (which suffers from slow convergence issues in high dimensions).

Acknowledgments This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350.

*UCL-Duke Workshop on Sensing and Analysis of High-Dimensional Data (SAHD), London, United Kingdom, 4-5 September 2014; abstract.

†The ordering of the second through fourth authors is alphabetical.