Minimax-Optimal Distribution Regression

Z. Szabó[1,*], B. Sriperumbudur[2], B. Póczos[3] and A. Gretton[1]

[1] *Gatsby Unit, University College London; zoltan.szabo@gatsby.ucl.ac.uk, arthur.gretton@gmail.com*
[2] *Department of Statistics, Pennsylvania State University; bks18@psu.edu*
[3] *Machine Learning Department, Carnegie Mellon University; bapoczos@cs.cmu.edu*
[*] *Corresponding author*

Abstract. *We focus on the distribution regression problem (DRP): we regress from probability measures to Hilbert-space valued outputs, where the input distributions are only available through samples (this is the 'two-stage sampled' setting). Several important statistical and machine learning problems can be phrased within this framework including point estimation tasks without analytical solution (such as hyperparameter or entropy estimation) and 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 to solve the DRP task 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 DRP task: we embed the distributions to a reproducing kernel Hilbert space and perform ridge regression from the embedded distributions to the outputs. Our main contribution is to prove that this scheme is consistent in the two-stage sampled setup under mild conditions (on separable topological domains enriched with kernels): we present an exact computational-statistical efficiency tradeoff analysis showing that the studied estimator is able to match the* one-stage *sampled minimax-optimal rate. This result answers a 17-year-old open question, by establishing the consistency of the classical set kernel (Haussler, 1999; Gärtner et al., 2002) in regression. We also cover consistency for more recent kernels on distributions, including those due to Christmann and Steinwart (2010). The practical efficiency of the studied technique is illustrated in supervised entropy learning and aerosol prediction using multispectral satellite images.*

Keywords. *Two-Stage Sampled Distribution Regression; Kernel Ridge Regression; Mean Embedding; Multi-Instance Learning; Minimax Optimality*

References

Christmann, A. and Steinwart, I. (2010). Universal kernels on non-standard input spaces. In *Advances in Neural Information Processing Systems (NIPS)*, pages 406–414.

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. (2002). Multi-instance kernels. In *International Conference on Machine Learning (ICML)*, pages 179–186.

Haussler, D. (1999). Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California at Santa Cruz.