Vector-valued Distribution Regression – Keep It Simple and Consistent[*]

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

Abstract

We tackle the distribution regression problem (DRP): regressing from probability measures to vector-valued outputs in the two-stage sampled case, where the input distributions are only available through samples. Numerous important and challenging machine learning and statistical tasks fit into the studied problem family such as multi-instance learning or point estimation tasks. Although there is a vast number of heuristics in the literature to address the DRP problem, to the best of our knowledge the only existing technique with performance guarantees requires density estimation (which often scales poorly in practice) and the distributions to have densities on compact Euclidean domains. In my talk, I am going to present a simple alternative to solve the DRP problem by embedding the input distributions to a reproducing kernel Hilbert space, followed by ridge regression from the embeddings to the outputs. We prove that the proposed approach is consistent: we derive finite sample excess risk bounds which hold with high probability and establish explicit convergence rates as a function of the problem difficulty and sample numbers. Specifically, we justify the applicability of set kernels in regression, which was a 15-year-old open question, and construct alternative kernels on the embedded distributions. The studied scheme is viable under mild conditions, on separable topological domains endowed with kernels. We demonstrate the efficiency of the method in two applications, supervised entropy learning and aerosol optical depth prediction based on multispectral satellite images.

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