Consistent Vector-valued Distribution Regression

Zoltán Szabó (Gatsby Computational Neuroscience Unit, University College London)*

We address the distribution regression problem (DRP): regressing on the (possibly infinite dimensional) domain of probability measures, in the two-stage sampled setup when only samples from the distributions are observable. The DRP formulation offers a unified framework for several important tasks in statistics and machine learning including for example, drug activity prediction, aerosol optical depth estimation, or supervised entropy learning.

Probably one of the most natural approaches for the two-stage sampled DRP task is to handle the sampled distributions as finite sets. This view leads to multi-instance learning (MIL), where one can define kernel learning algorithms based on set kernels (also called multi-instance kernels or ensemble kernels; Haussler, 1999; Gärtner et al., 2002). In this case, the similarity of two sets is measured by the average pairwise point similarities between the sets. From a *theoretical* perspective, very little has been done to establish the consistency of set kernels in learning since their introduction in 1999 (2002): i.e. in what sense (and with what rates) is the learning algorithm consistent, when the number of items per bag, and the number of bags, is allowed to increase?

It is possible, however, to view set kernels in a distribution setting, as they represent valid kernels between (mean) embeddings of empirical probability measures into a reproducing kernel Hilbert space (RKHS). The population limits are well-defined as being dot products between the embeddings of the generating distributions, and for characteristic kernels the distance between embeddings defines a metric on probability measures. When bounded kernels are used, mean embeddings exist for all probability measures.

Based on these properties, here we present a simple (analytically computable) ridge regression approach to DRP: we embed the distributions to a RKHS, and learn the regressor from the embeddings to the outputs. Our contribution is two-fold: firstly, we show that this scheme is consistent in the two-stage sampled setup under mild conditions, for probability measure inputs defined on separable, topological domains endowed with kernels, with vector-valued outputs belonging to an arbitrary separable Hilbert space. Secondly, we establish explicit statistical optimality guarantees for the ridge based, doubly large-scale (distribution input, Hilbert output) estimator in terms of the involved computational effort described by the number of bags and the number of items in the bags. Specially, choosing the kernel on the space of embedded distributions to be linear, we get the consistency of set kernels in regression, which was a 15-year-old open question.

There exist several *heuristics* in the literature to compute the similarity of distributions or bags of samples; these similarities can then form the basis of a learning algorithm. These approaches include (i) kernelized Gaussian divergences by assuming that the training distributions are Gaussians in a RKHS, (ii) positive definite kernels on probability measures such as semigroup kernels, nonextensive information theoretical kernel constructions, and kernels based on Hilbertian metric, (iii) consistent Rényi or Tsallis divergence estimates (which are *not* kernels), or (iv) set metric based techniques. Unfortunately, all these methods are lack of any theoretical guarantee when applied in specific learning tasks, such as regression.

To the best of our knowledge, the only existing technique with consistency guarantees for DRP requires kernel density estimation as an intermediate step (which often scale poorly in practice), and the domain of the distributions to be compact Euclidean. In contrast, our proposed technique is more general by allowing separable topological domains for the distributions and separable Hilbert outputs, and avoids density estimation by working directly on distribution embeddings, which leads to improved scaling properties and performance.

[*] Joint work with Arthur Gretton (Gatsby Computational Neuroscience Unit, University College London), Barnabás Póczos (Machine Learning Department, Carnegie Mellon University), Bharath K. Sriperumbudur (Statistics Department, Pennsylvania State University); the ordering of the second through fourth authors is alphabetical.