Distribution Regression - the Set Kernel Heuristic is Consistent*

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

Abstract

Bag of feature (BoF) representations are omnipresent in machine learning; for example, an image can be described by a bag of visual features, a document might be considered as a bag of words, or a molecule can be handled as a bag of its different configurations. Set kernels (also called multi-instance or ensemble kernels; Gärtner 2002) defining the similarity of two bags as the average pairwise point similarities between the sets, are among the most widely applied tools to handle problems based on such BoF representations. Despite the wide applicability of set kernels, even the most fundamental theoretical questions such as their consistency in specific learning tasks is unknown.

In my talk, I am going to focus on the distribution regression problem: regressing from a probability distribution to a real-valued response. By considering the mean embeddings of the distributions, this is a natural generalization of set kernels to the infinite sample limit: the bags can be seen as i.i.d. (independent identically distributed) samples from a distribution. We will propose an algorithmically simple ridge regression based solution for distribution regression and prove its consistency under fairly mild conditions (for probability distributions defined on locally compact Polish spaces). As a special case, we give positive answer to a 12-year-old open question, the consistency of set kernels in regression. We demonstrate the efficiency of the studied ridge regression technique on (i) supervised entropy learning, and (ii) aerosol prediction based on satellite images.

Preprint: `http://arxiv.org/pdf/1402.1754`
Code: `https://bitbucket.org/szzoli/ite/`

*CSML Lunch Talk Series, London, UK; May 2, 2014; abstract.

†Joint work with Arthur Gretton (Gatsby Computational Neuroscience Unit, University College London), Barnabás Póczos (Machine Learning Department, Carnegie Mellon University), Bharath K. Sriperumbudur (Department of Pure Mathematics and Mathematical Statistics, University of Cambridge); the ordering of the second through fourth authors is alphabetical.