A Hilbert Space Embedding for Distributions

Alex Smola[1], Arthur Gretton[2], Le Song[1], and Bernhard Schölkopf[2]

[1] NICTA and ANU, Northbourne Avenue 218, Canberra 0200 ACT, Australia,
{alex.smola, le.song}@nicta.com.au
[2] MPI for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany,
{arthur,bernhard.schoelkopf}@tuebingen.mpg.de

While kernel methods are the basis of many popular techniques in supervised learning, they are less commonly used in testing, estimation, and analysis of probability distributions, where information theoretic approaches rule the roost. However it becomes difficult to estimate mutual information or entropy if the data are high dimensional.

We present a method which allows us to compute distances between distributions *without* the need for intermediate density estimation. Our approach allows algorithm designers to specify which properties of a distribution are most relevant to their problems. Our method works by studying the convergence properties of the expectation operator when restricted to a chosen class of functions. In a nutshell our method works as follows: denote by $\mathcal{X}$ a compact domain and let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space on $\mathcal{X}$ with kernel k. Note that in an RKHS we have $f(x) = \langle f, k(x, \cdot)\rangle$ for all functions $f \in \mathcal{H}$. This allows us to denote the expectation operator of a distribution p via

$$\mu[p] := \mathbf{E}_{x\sim p}[k(x, \cdot)] \qquad \text{and hence } \mathbf{E}_{x\sim p}[f(x)] = \langle \mu[p], f\rangle \text{ for } f \in \mathcal{H}.$$

Moreover, for a sample $X = \{x_1, \ldots, x_m\}$ drawn from some distribution p we may denote the empirical counterparts via

$$\mu[X] := \tfrac{1}{m}\sum_{i=1}^{m} k(x_i, \cdot) \qquad \text{and hence } \tfrac{1}{m}\sum_{i=1}^{m} f(x_i) = \langle \mu[X], f\rangle \text{ for } f \in \mathcal{H}.$$

This allows us to compute distances between distributions p, q via $D(p, q) := \|\mu[p] - \mu[q]\|$ and empirical samples X, X' via $D(X, X') := \|\mu[X] - \mu[X']\|$ alike. One can show that under rather benign regularity conditions $\mu[X] \to \mu[p]$ at rate $O(m^{-\frac{1}{2}})$. Such a distance is useful in a number of estimation problems:

- Two-sample tests whether X and X' are drawn from the same distribution.
- Density estimation, where we try to find p so as to minimize the distance between $\mu[p]$ and $\mu[X]$, either by mixture models or by exponential families.
- Independence measures where we compute the distance between the joint distribution and the product of the marginals via $D(p(x, y), p(x) \cdot p(y))$.
- Feature selection algorithms which try to find a subset of covariates x maximally dependent on the target random variables y.

Our framework allows us to unify a large number of existing feature extraction and estimation methods, and provides new algorithms for high dimensional nonparametric statistical tests of distribution properties.