Kernel Two-Sample and Independence Tests

Arthur Gretton*

Gatsby Unit, CSML, University College London, UK

Abstract:

Many problems in unsupervised learning require the analysis of features of probability distributions. At the most fundamental level, we might wish to test whether two distributions are the same, based on samples from each - this is known as the two-sample or homogeneity problem. We address this problem by mapping probability distributions to elements in a reproducing kernel Hilbert space (RKHS). Given a sufficiently rich RKHS, these representations are unique: thus comparing feature space representations allows us to compare distributions without ambiguity. The distance between feature mappings of two random variables to an RKHS is denoted the maximum mean discrepancy (MMD). A second important testing problem is to discover whether two random variables drawn from a joint distribution are independent. It turns out that any dependence between pairs of random variables can be encoded as a kernelized distance between the RKHS mapping of the joint distribution, and that of the product of the marginals: this statistic is called the Hilbert Schmidt Independence Criterion (HSIC). We may again formulate a hypothesis test to determine whether the dependence is statistically significant. Being expressed in terms of kernels, the tests can be used on any domain where kernels have been defined: tests can be conducted on distributions over strings, graphs, groups, semigroups, and compact manifolds.

Finally, a link is established between the MMD and HSIC, on one hand, and the energy distance and distance covariance, on the other. In the case where the energy distance is computed with a semimetric of negative type, a positive definite kernel, termed distance kernel, may be defined such that the MMD corresponds exactly to the energy distance. Conversely, for any positive definite kernel, we can interpret the MMD as an energy distance with respect to a negative-type semimetric. This result extends to distance covariance, when distance kernels on the product space are used to establish an equivalence of the distance covariance and HSIC.

Key words: nonparametric hypothesis testing, reproducing kernel Hilbert space, energy distance