Gretton, Arthur; **Borgwardt, Karsten M.**; **Rasch, Malte J.**; **Schölkopf, Bernhard**; **Smola, Alexander**

A kernel two-sample test. (English) Zbl 1283.62095

J. Mach. Learn. Res. 13, 723-773 (2012).

Summary: We propose a framework for analyzing and comparing distributions, which we use to construct statistical tests to determine if two samples are drawn from different distributions. Our test statistic is the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space (RKHS), and is called the maximum mean discrepancy (MMD). We present two distribution-free tests based on large deviation bounds for the MMD, and a third test based on the asymptotic distribution of this statistic. The MMD can be computed in quadratic time, although efficient linear time approximations are available. Our statistic is an instance of an integral probability metric, and various classical metrics on distributions are obtained when alternative function classes are used in place of an RKHS. We apply our two-sample tests to a variety of problems, including attribute matching for databases using the Hungarian marriage method, where they perform strongly. Excellent performance is also obtained when comparing distributions over graphs, for which these are the first such tests.

MSC:

62G10 Nonparametric hypothesis testing
62G08 Nonparametric regression and quantile regression
60F10 Large deviations

Keywords:

kernel methods; two-sample test; uniform convergence bounds; schema matching; integral probability metric; hypothesis testing

Software:

KDD Cup; UCI-ml

Full Text: Link