Linear-time Divergence Measures with Applications in Hypothesis Testing*

Zoltán Szabó† (CMAP, École Polytechnique)

Abstract

Maximum mean discrepancy and Hilbert-Schmidt independence criterion are among the most popular and successful techniques in machine learning to measure the difference and the independence of random variables, respectively. Their computational complexity is however rather restrictive, quadratic in the number of samples. In order to mitigate this serious computational bottleneck, I am going to present 3 linear-time kernel-based alternatives with illustrations in hypothesis testing. The power of the new linear-time methods is demonstrated in natural language processing (distinguishing articles from two categories), computer vision (differentiating positive and negative emotions), dependency testing of media annotations (song - year of release, video - caption) and criminal data analysis.

Code:

- Information Theoretical Estimators toolbox:
  - `https://bitbucket.org/szzoli/ite-in-python/`,
- Linear-time Two-sample Testing (NIPS-2016, Oral):
  - `https://github.com/wittawatj/interpretable-test`,
- Linear-time Independence Testing (ICML-2017):
  - `https://github.com/wittawatj/fsic-test`,
- Linear-time Goodness-of-fit Testing (NIPS-2017, Best Paper Award):
  - `https://github.com/wittawatj/kernel-gof`.

*Tao Seminar, INRIA Saclay, France, 13 February 2018; abstract.

†Joint work with Wittawat Jitkrittum[1], Wenkai Xu[1], Kacper Chwialkowski[1], Arthur Gretton[1] and Kenji Fukumizu[2]; [1]Gatsby Unit, University College London; [2]Institute for Statistical Mathematics, Tokyo.