



SLING Efficient algorithms for sustainable machine learning

Mon, March 21th, 2022, 3:00 p.m., DIBRIS Room 706, Via Dodecaneso 35, Genova.

Statistical Learning & Optimization

Learning-augmented count-min sketches via Bayesian nonparametrics

Abstract.

The count-min sketch (CMS) is a time and memory efficient randomized data structure that provides estimates of tokens' frequencies in a data stream, i.e. point queries, based on randomly hashed data. Learning-augmented CMSs aim at improving the CMS by means of learning models that allow to better exploit data properties. We focus on the learning-augmented CMS of Cai, Mitzenmacher and Adams (NeurIPS, 2018), which relies on Bayesian nonparametric (BNP) modeling of a data stream via Dirichlet process (DP) priors; this is referred to as the CMS-DP. We present a novel and rigorous approach of the CMS-DP, and we show that it allows to consider more general classes of nonparametric priors than the DP prior. We apply our approach to develop a novel learning-augmented CMS under power-law data streams, which relies on BNP modeling of the stream via Pitman-Yor process (PYP) priors; this is referred to as the CMS-PYP. Applications to synthetic data and real data show that the CMS-PYP outperforms both the CMS and CMS-DP in the estimation of low-frequency tokens; this is known to be a critical feature in natural language processing, where it is indeed common to encounter power-law data.

Speaker

Stefano Favaro

Università degli Studi di Torino



SLING Efficient algorithms for sustainable machine learning

Mon, March 21th, 2022, 3:00 p.m., DIBRIS Room 706, Via Dodecaneso 35, Genova.



Stefano Favaro is Professor of Statistics at University of Torino, and Carlo Alberto Chair of Statistics and Machine Learning at Collegio Carlo Alberto. He is also Associate Member at the Department of Statistics of University of Oxford, and Research Fellow at the Institute for Applied Mathematics and Information Technologies “Enrico Magenes” of CNR. In 2019 Stefano Favaro was awarded by a ERC Consolidator Grant on nonparametric Bayes and empirical Bayes methods for discrete functional estimation. Beyond nonparametric Bayesian methods, his research

interests include statistical machine learning, data confidentiality and fairness, learning-augmented recovery algorithms, and mathematics of deep learning. Since 2020 Stefano Favaro serves as an Associate Editor for Bernoulli and Statistical Science.