Learning to (approximately) count with Bayesian nonparametrics

A seminar by Mario Beraha

Università degli Studi di Milano Bicocca

Thursday 10 Apr 2025 | 2.30 p.m.
Room Benvenuti
Department of Statistical Sciences

We study how to recover the frequency of a symbol in a large discrete data set, using only a compressed representation, or sketch, of those data obtained via random hashing. This is a classical problem in computer science, with various algorithms available, such as the count-min sketch. However, these algorithms often assume that the data are fixed, leading to overly conservative and potentially inaccurate estimates when dealing with randomly sampled data. In this paper, we consider the sketched data as a random sample from an unknown distribution, and then we introduce novel estimators that improve upon existing approaches. Our method combines Bayesian nonparamet ric and classical (frequentist) perspectives, addressing their unique limitations to provide a principled and practical solution.

Università degli Studi di Padova

Dipartimento di Scienze Statistiche