THE WEIZMANN INSTITUTE OF SCIENCE
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
Faculty Seminar

Schmidt Auditorium ,Shcmidt Building
on Thursday, Jan 16, 2020at 11:00

NOTE UNUSUAL DAY AND PLACE

Alon Kipnis Stanford University

Higher Criticism for discriminating frequency-tables and testing authorship

Abstract:

The Higher Criticism (HC) test is a useful tool for detecting the presence of a signal spread across a vast number of features, especially in the sparse setting when only few features are useful while the rest are pure noise. We adapt the HC test to the two-sample setting of detecting changes between two frequency tables. We apply this adaptation to authorship attribution challenges, where the goal is to identify the author of a document using other documents whose authorship is known. The method is simple yet performs well without handcrafting and tuning. Furthermore, as an inherent side effect, the HC calculation identifies a subset of discriminating words, which allow additional interpretation of the results. Our examples include authorship in the Federalist Papers, machine-generated texts, and the identity of the creator of the Bitcoin.

We take two approaches to analyze the success of our method. First, we show that, in practice, the discriminating words identified by the test have low variance across documents belonging to a corpus of homogeneous authorship. We conclude that in testing a new document against the corpus of an author, HC is mostly affected by words characteristic of that author and is relatively unaffected by topic structure. Finally, we analyze the power of the test in discriminating two multinomial distributions under a sparse and weak perturbation model. We show that our test has maximal power in a wide range of the model parameters, even though these parameters are unknown to the user.