מכון ויצמן למדע

THE WEIZMANN INSTITUTE OF SCIENCE

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Machine Learning and Statistics Seminar

Room 1 ,Ziskind Building
on Wednesday, May 15, 2019
at 11:15

Gal Elidan
HUJI

Learning Rules-first Classifiers

Abstract:

Complex classifiers may exhibit "embarrassing" failures even in "easy" cases where humans can provide a simple justified explanation. Avoiding such failures is obviously of key importance. In this work, we focus on one such setting, where a label is perfectly predictable if the input contains certain features, or rules, and otherwise it is predictable by a linear classifier. We define a hypothesis class that captures this notion and determine its sample complexity. We also give evidence that efficient algorithms cannot achieve this sample complexity. We then derive a simple and efficient algorithm and show that its sample complexity is close to optimal, among efficient algorithms. Experiments on synthetic and sentiment analysis data demonstrate the efficacy of the method, both in terms of accuracy and interpretability. At the end of the talk, I will also give a teaser to demonstrate the non-intuitive nature of the more general problem of embarrassment.