Overcoming Intractability in Unsupervised Learning

Sanjeev Arora

Computer Science Department, Princeton University
arora@princeton.edu

Abstract

Unsupervised learning – i.e., learning with unlabeled data - is increasingly important given today's data deluge. Most natural problems in this domain – e.g. for models such as mixture models, HMMs, graphical models, topic models and sparse coding/dictionary learning, deep learning – are NP-hard. Therefore researchers in practice use either heuristics or convex relaxations with no concrete approximation bounds. Several nonconvex heuristics work well in practice, which is also a mystery.

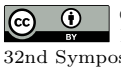
The talk will describe a sequence of recent results whereby rigorous approaches leading to polynomial running time are possible for several problems in unsupervised learning. The proof of polynomial running time usually relies upon nondegeneracy assumptions on the data and the model parameters, and often also on stochastic properties of the data (average-case analysis). We describe results for topic models, sparse coding, and deep learning. Some of these new algorithms are very efficient and practical – e.g. for topic modeling.

1998 ACM Subject Classification F.2 Analysis of Algorithms and Problem Complexity, I.2 Artificial Intelligence

Keywords and phrases machine learning, unsupervised learning, intractability, NP-hardness

Digital Object Identifier 10.4230/LIPIcs.STACS.2015.1

Category Invited Talk



32nd Symposium on Theoretical Aspects of Computer Science (STACS 2015).
Editors: Ernst W. Mayr and Nicolas Ollinger; pp. 1–1

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany