Afternoon Tutorial

The Dataminer's Guide to Scalable Mixed-Membership and Nonparametric Bayesian Models

Dr. Amr Ahmed
Google

Dr. Alex Smola
Carnegie Mellon University

Abstract

Large amounts of data arise in a multitude of situations, ranging from bioinformatics to astronomy, manufacturing, and medical applications. For concreteness our tutorial focuses on data obtained in the context of the internet, such as user generated content (microblogs, e-mails, messages), behavioral data (locations, interactions, clicks, queries), and graphs. Due to its magnitude, much of the challenges are to extract structure and interpretable models without the need for additional labels, i.e. to design effective unsupervised techniques. We present design patterns for hierarchical nonparametric Bayesian models, efficient inference algorithms, and modeling tools to describe salient aspects of the data.

Instructors

Dr. Amr Ahmed is a Research Scientist at Google. He received his PhD from Carnegie Mellon University in 2011. His thesis "Modeling Users and Content: Structured Probabilistic Representation and Scalable Online Inference Algorithms" was awarded the prestigious ACM SIGKDD Doctoral Dissertation award in 2012. He spent a year as a Research Scientist at Yahoo! Research before joining Google. He authored over 40 papers on topics that are core to this tutorial (including a best-paper runner-up award at WSDM 2012) and co-presented 3 tutorials at web and machine learning conferences.

Dr. Alex Smola received his PhD from the University of Technology in Berlin in 1998. Subsequently he was research group leader and professor at the Australian National University and Senior Principal Researcher at National ICT Australia. From 2008 until 2012 he was Principal Research Scientist at Yahoo. Since 2012 he is a visiting researcher at Google and since 2013 a full professor at the Machine Learning Department of Carnegie Mellon University. He has written over 180 papers (that won several best paper awards at ICML, WSDM and SIGIR) and authored or edited 5 books. His work covers a broad range of subjects from statistical learning theory, convex optimization, and functional analysis to practical algorithms for scalable data classification, regression, clustering, and topic models. His recent work focuses on distributed, very large scale latent variable models for user profiling and content recommendation.

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
ACM 978-1-4503-2174-7/13/08.