Topic Models, Latent Space Models, Sparse Coding, and All That: A systematic understanding of probabilistic semantic extraction in large corpus

Eric Xing
School of Computer Science
Carnegie Mellon University

Abstract

Probabilistic topic models have recently gained much popularity in informational retrieval and related areas. Via such models, one can project high-dimensional objects such as text documents into a low dimensional space where their latent semantics are captured and modeled; can integrate multiple sources of information—to "share statistical strength" among components of a hierarchical probabilistic model; and can structurally display and classify the otherwise unstructured object collections. However, to many practitioners, how topic models work, what to and not to expect from a topic model, how is it different from and related to classical matrix algebraic techniques such as LSI, NMF in NLP, how to empower topic models to deal with complex scenarios such as multimodal data, contractual text in social media, evolving corpus, or presence of supervision such as labeling and rating, how to make topic modeling computationally tractable even on web-scale data, etc., in a principled way, remain unclear. In this tutorial, I will demystify the conceptual, mathematical, and computational issues behind all such problems surrounding the topic models and their applications by presenting a systematic overview of the mathematical foundation of topic modeling, and its connections to a number of related methods popular in other fields such as the LDA, admixture model, mixed membership model, latent space models, and sparse coding. I will offer a simple and unifying view of all these techniques under the framework multi-view latent space embedding, and online the roadmap of model extension and algorithmic design toward different applications in IR and NLP. A main theme of this tutorial that tie together a wide range of issues and problems will build on the "probabilistic graphical model" formalism, a formalism that exploits the conjoined talents of graph theory and probability theory to build complex models out of simpler pieces. I will use this formalism as a main aid to discuss both the mathematical underpinnings for the models and the related computational issues in a unified, simplistic, transparent, and actionable fashion.

Tutorial Abstracts of ACL 2012, page 3,
Jeju, Republic of Korea, 8 July 2012.