psd

MASTER'S THESIS PRESENTATION

Online Variational Bayes Algorithm for Latent Dirichlet Allocation with Sparse Update

WHEN
May 4, 2022
3:00 PM

WHERE
Zoom Meeting
For ZOOM presentations, details will be provided in an email announcement for this seminar.

Zehao Shao, MS candidate

Topic models are hierarchical Bayesian models of document collections. They can uncover the main themes that pervade a corpus and then use those themes to help organize, search, and explore the documents. In topic modeling, a topic is a distribution over a fixed vocabulary and each document exhibits the topics with different proportions. Both the topics and the topic proportions of documents are hidden variables. Inferring the conditional distribution of these variables given an observed set of documents is the central computational problem. Latent Dirichlet Allocation (LDA) is a Bayesian probabilistic topic model of text documents.

We develop our algorithm on the basis of online LDA in Hoffman et al. (2010), a stochastic gradient optimization algorithm for topic modeling. We take advantage of sparsity during the update of parameters by introducing an alternative parameter that represents the non-smoothing portion of the variational Dirichlet parameter in the Latent Dirichlet Allocation. It is simple and is intuitively sensible.We study the performance of our online LDA with sparse update mainly by fitting a 100-topic topic model to articles downloaded from Wikipedia. We evaluate the best parameters of the model and compare the performance of online LDA with and without sparse update by held-out perplexity on test sets.

stat.uchicago.edu

THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS