Course Project: Latent Dirichlet Allocation (LDA) for topic modeling

The **latent Dirichlet allocation** (**LDA**) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.

Model description

The variable names are defined as follows:

M denotes the number of documents

N is number of words in a given document (document *i* has N_i words)

α is the parameter of the Dirichlet prior on the per-document topic distributions

β is the parameter of the Dirichlet prior on the per-topic word distribution

θ_i is the topic distribution for document *i*

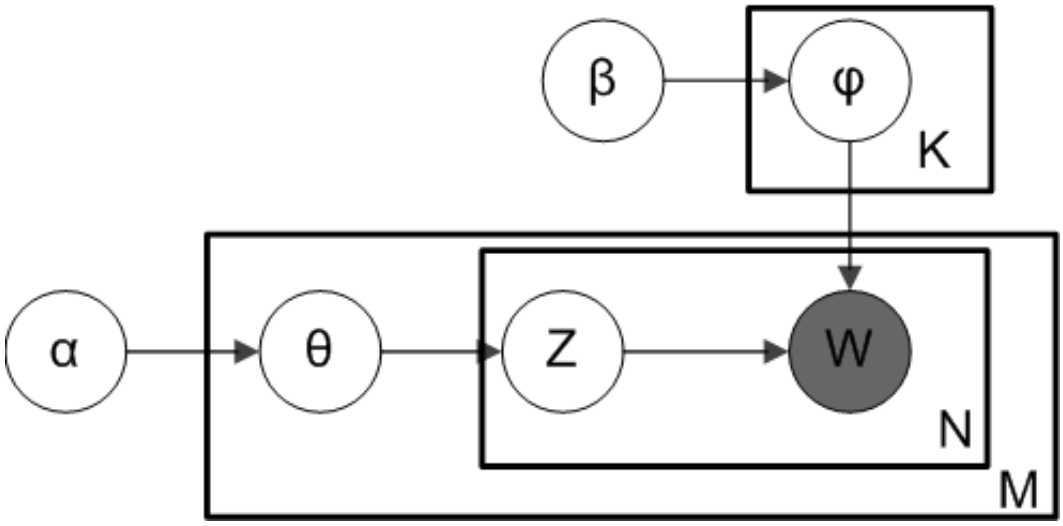
ϕ_k is the word distribution for topic *k*

z_{ij} is the topic for the *j*-th word in document *i*

w_{ij} is the specific word.

To actually infer the topics in a corpus, we imagine a generative process whereby the documents are created. We imagine the generative process as follows. Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over all the words. LDA assumes the following generative process for a corpus **D** consisting of M documents each of length N_i:

1. Choose $\theta_i \sim \mathrm{Dir}(\alpha)$, where i∈{1,...,M} and $\mathrm{Dir}(\alpha)$ is a Dirichlet distribution with a symmetric parameter α which typically is sparse ($\alpha < 1$)
2. Choose $\phi_k \sim \mathrm{Dir}(\beta)$, where $k \in 1, \ldots, K$ and β typically is sparse
3. For each of the word positions i, j, where $i \in 1, \ldots, M$, and $j \in 1, \ldots, N_i$
    1. Choose a topic $z_{ij} \sim \mathrm{Multinomial}(\theta_i)$
    2. Choose a word $w_{ij} \sim \mathrm{Multinomial}(\phi_{z_i,j})$



Through a variational inference (VI), we can efficiently fit the model and infer the topics.

Task

Implement LDA with a variational EM algorithm, infer the topics and evaluate the performance using the demo dataset 'text_data.txt' with K=3 topics. There are 300 documents comprising 100 word in the vocabulary, where each row represents a document, each colum represents a word and the i,j-th entry value is the times of the j-th word showing up in the i-th document. Compare the results with the true document distribution stored in file 'topic_doc_true.txt'. If possible, apply your implementation to the real dataset 'cora_data.txt', which contains. Discuss the results.

Requirements

Submit a notebook or pdf that containes the outcomes generated from your implementation. Clearly describe the observations and explain the results.

The code should be uploaded to Github and the Github link should be provided in the report.

Reference

A general introduction to topic models: http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf

Slides: http://www.cs.columbia.edu/~blei/talks/Blei_ICML_2012.pdf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859-877.

Contact TA: Cai Mingxuan mcaiad@ust.hk