Clustered sender-receiver pairs in relational topic models

Kayla Schaefer

Spring 2015

1 Background

In machine learning, we tend to focus on models that break a network into patterns of connections between latent and observed factors. Such models are then used to explore social relationships, neural nets, or text documents. I intend to focus on topic modeling in text for my report.

The basic model for text is Latent Dirichlet Allocation, proposed by Blei, Ng, and Jordan in 2003. LDA is a hierarchical Bayesian mixture model based on a Dirichlet Process, which is depicted visually in Figure 1. The outer box represents M documents, while the inner box represents N words and their topics within a document.[1].

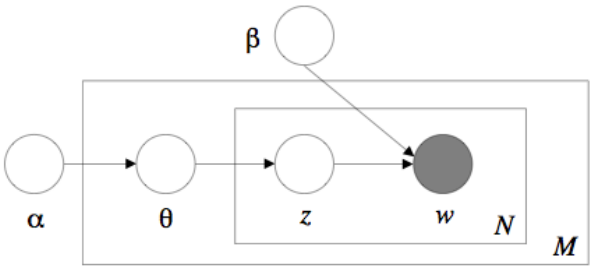

Figure 1: Graphical model representation of LDA.

The K topics are latent variables that are composed from an infinite mixture of probabilities α/K. Each document in a corpus of work is then modeled as a finite mixture over topics. This can be expanded on in the Relational Topic Model, which not only models topics in text, but also "explicitly ties the content of the documents with the connections between them" [2]. Thus, RTM can predict links between new documents based only on the text in the document.

2 Purpose of report

This report would like to continue building on LDA and RTM to be able to cluster authors of text as well. Specifically, we look to look to model email correspondence where each document in the corpus links two possible authors. For a given document, we have a document sender and a receiver. Our goal is to be able to cluster the sender-receiver pairs based on the topic of the document sent. This will be achieved with nested Dirichlet Processes to ensure overlap in the support of senders and receivers for each topic. After establishing the full details of this model, we will implement it on an example email set and compare the results of a holdout test set of documents to the results from existing models by using a Chib estimate of perplexity.

References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993-1022.

[2] Chang, J., & Blei, D. M. (2009). Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics*, 81-88.