Exploring Variational Auto-Encoders for Topic Modelling

Apoorva Jarmale,Kaavya Gowthaman,Oj Sindher,Saarang Pande,Vartika Tewari

[3]

1 Introduction

Topic models are among the most widely used models for learning unsupervised representations of text, with hundreds of different model variants in the literature. However a major challenge is that any change to the topic model requires mathematically deriving a new inference algorithm. This means that applying topic models and developing new models is accompanied by the high computational cost of computing the posterior distribution. Therefore a large body of work has considered approximate inference methods, with the most popular methods being mean field methods, and particularly methods based on collapsed Gibbs sampling. Both mean-field and collapsed Gibbs have the drawback that applying them to new topic models, even if there is only a small change to the modeling assumptions, requires re-deriving the inference methods, which can be mathematically arduous and time consuming, and limits the ability of practitioners to freely explore the space of different modeling assumptions.

2 Proposed Project

In this project we are going to implement one of the state of the art VAE models: ProdLDA[4] and will compare it with basic topic modelling techniques like LDA[1].

2.1 Proposed Techniques

Autoencoding variational Bayes (AEVB) is an effective choice for topic models, because it trains an inference network i.e a neural network that directly maps a document to an approximate posterior distribution,without the need to run further variational updates.This is impressive because in topic models, we expect that a small change in the document will produce only a small change in topics. This is exactly the type of mapping that a universal function approximator like a neural network should be good at representing.

2.1.1 ProdLDA[4]

We would be implementing the ProdLDA[4] paper which has a very effective AEVB inference method. The model described in the paper promises advantages like topic coherence, computational efficiency, and a Black box medthodology i.e does not require rigorous mathematical derivations to handle changes in the model, and can be easily applied to a wide range of topic models.

We also hope to try and implement the NVDM[2] paper for comparison if time permits.

2.2 Data

The 20 Newsgroups[3] data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques.The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). We believe this would be the perfect dataset to work on for Topic Modelling as well.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022.

[2] Yishu Miao, Lei Yu, and Phil Blunsom. "Neural variational inference for text processing". In: *International conference on machine learning*. PMLR. 2016, pp. 1727–1736.

[3] Michael Riffle. *2oNewsgroups Dataset*. `http://qwone.com/~jason/20Newsgroups/`.

[4] Akash Srivastava and Charles Sutton. "Autoencoding variational inference for topic models". In: *arXiv preprint arXiv:1703.01488* (2017).