IFT6269 - Mid-project Report - Latent Dirichlet Allocation

Patrice Béchard [*1] **Théophile Gervet** [*2]

Abstract

The project we have chosen is to review the article from Blei, Ng and Jordan on Latent Dirichlet Allocation (Blei et al., 2003), implement the model and experiment with real data from the *20 newsgroups* text dataset.

1. What has been done

We already have started to write the final report for the project. We first provide a description of the Latent Dirichlet Allocation generative probabilistic model, explaining the setting of the problem and providing both the graphical model and the joint distribution associated with the model.

In the following section, we provide the mathematical developments on how to learn the parameters of the model to make accurate predictions using the EM algorithm and we highlight an important problem where the posterior distribution to our model is intractable, forcing us to rely on approximate inference to estimate the parameters. The first method used to solve this problem consist of using variational inference. We provide an in-depth explanation on how we define a new function which replaces the real posterior distribution, which we use to conduct the EM algorithm. We then reformulate both the E-step and the M-step applied to the situation and provide a pseudo-code for the algorithm.

An implementation of the Latent Dirichlet Allocation using variational inference has already been done. We tested the algorithm on sets of documents from two topics (*atheism* and *space*) taken from the *20 newsgroups* text dataset and compared the obtained results with a pre-made implementation of the Latent Dirichlet Allocation provided within the Scikit-Learn library. Table 1 presents some words associated with each topic for both our implementation of the LDA using variational inference and the Scikit-Learn implementation.

We see that the results for both implementation are pretty similar, which is a good sign that our implementation is working as it should be. We also have implemented a function which presents the most likely topic assignment for each word in a document.

Table 1. Words associated to topics for our implementation of LDA using variational inference and Scikit-learn's implementation of LDA.

Our implementation using variational inference		Scikit-learn implementation	
SPACE	ATHEISM	SPACE	ATHEISM
space	one	space	people
launch	would	nasa	god
nasa	people	launch	don
satellite	god	earth	just
system	think	data	think
mission	like	orbit	like
year	thing	shuttle	does
orbit	say	satellite	know
program	know	lunar	say
data	atheist	moon	atheism
also	could	program	time
earth	make	edu	believe

2. What still has to be done

A lot of things are still left for us to do. First, we will present an explanation of the EM algorithm for the model using Gibbs sampling as we have already done with variational inference and will provide a pseudocode of the algorithm as well. There is a lot of documentation about this precise subject in the literature (Griffiths, 2002)(Darling, 2011)(Heinrich, 2005). We will implement the Latent Dirichlet Allocation using Gibbs sampling to do approximate inference on the parameters of the model and experiment with similar data to look at the difference between both of our implementations as well as the one provided by Scikit-Learn. We will also expand our experimentation on more than two subjects and provide various cool visualisations.

[*]Equal contribution [1]Université de Montréal, Montréal, Canada [2]McGill University, Montréal, Canada. Correspondence to: Patrice Béchard <patrice.bechard@umontreal.ca>, Théophile Gervet <theophile.gervet@umontreal.ca>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

References

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Darling, William M. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011.

Griffiths, Tom. Gibbs sampling in the generative model of latent dirichlet allocation. Technical report, Department of Psychology, Stanford University, Stanford, CA, 2002.

Heinrich, Gregor. Parameter estimation for text analysis. Technical report, vsonix GmbH + University of Leipzig, Leipzig, Germany, 2005.