A Comparison of Posterior Approximation Algorithms for Latent Dirichlet Allocation

Cindy Cook
STAT 540 Project, Fall 2015

With an increase in the amount of digitized text, topic modeling has become an important area of research. One of the most popular methods for topic modeling is Latent Dirichlet Allocation (LDA) first described in [3]. LDA is a probabilistic hierarchical generative model, which uses the Bayesian framework to discover hidden structure, or *topics*, within a text Corpus. Given a document the distribution of latent topics is found through estimation of the posterior distribution. However, this is very challenging and (computationally demanding) since the posterior distribution of the latent variables given a document is intractable. The development of several fast and accurate approximation methods applied to LDA has provided a user with several options including collapsed Gibbs sampling [7], collapsed variational Bayesian inference [11], maximum likelihood estimation [9], and maximum a posteriori estimation [6]. In [3], variational methods are applied to solve the problem. These methods were coded in C by Blei (refer to [4]) and have since been utilized in the *topicmodels* package in R [8]. Another package, *lda*, in R has been developed by Chang, which uses his own C code with collapsed Gibbs sampling for posterior approximation [5]. On the website Cross Validated, there is a discussion of which package is better for users, with no definitive answer [1].

For this project I propose to conduct a simulation study, which will focus on a comparison of the collapsed Gibbs sampling algorithm and the variational Bayesian algorithm as applied to LDA. I will start by first gaining an understanding of what these algorithms are theoretically doing. Then I will explore their performances empirically. To this aim, I will simulate data following the generative process defined in [3]. I will vary the number of topics, i.e. groups: $5, 10, 25, 50$, the hyperparameters , i.e. values: $0.001, 0.01, 0.1$, and the length of words in the vocabulary, i.e. 0 to $5,000$ by $1,000$ to follow the simulation set up in [10]. Mukherjee and Blei focused their study on comparing the collapsed and non-collapsed variational Bayes algorithms. I will focus on comparing an algorithm I develop to compute LDA with collapsed Gibbs sampling and compare it to the existing results from the LDA function in the *topicmodels* package in R. I will use both perplexity and precision/recall measures to compare the two algorithms as shown in [2]. I will also take into account the computational costs. With the research I have done so far, it is known that the collapsed Gibbs algorithm will take longer to gain convergence, but should be more accurate. In what situations this very general statement remains true is the ultimate goal of this research project.

References

[1] Anonymous. (2012), "Two R Packages for Topic Modeling, lda and topicmodels?" *Cross Validated*: http://stats.stackexchange.com/questions/24441/two-r-packages-for-topic-modeling-lda-and-topicmodels.

[2] Asuncion, A., Welling, M., Smyth, P., and Teh, Y. (2009), "On Smoothing and Inference for Topic Models." *UAI*: 27-34.

[3] Blei, D., Ng, A., and Jordan, M. (2003), "Latent Dirichlet Allocation." *Journal of Machine Learning Research*: 3 993-1022.

[4] Blei, D. (2004),"LDA-C."

[5] Chang, J. (2015), "Package 'lda'." *CRAN*.

[6] Chien, J., and Wu, M. (2008), "Adaptive Bayesian Semantic Analysis." *Audio, Speech, and Language Processing*, IEEE Transactions on: 16(1), 198207.

[7] Griffiths, L. and Steyvers, M. (2004), "Finding Scientific Topics." *PNAS*: 1(Suppl 1), 5228-5235.

[8] Grun, B. and Hornik, K. (2015), "Package 'topicmodels'." *CRAN*.

[9] Hofmann, T. (2001), "Unsupervised Learning by probabilistic Latent Semantic Analysis." *Machine Learning*: 42(1), 177196.

[10] Mukherjee, I. and Blei, D. (2009), "Relative Performance Grantees for Approximate Inference in Latent Dirichlet Allocation." *NIPS*: 21, 11291136.

[11] Teh, Y. W., Newman, D., and Welling, M. (2007), "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation." *NIPS*: 3, 1353-1360.