Latent Dirichlet Allocation (LDA) is a Bayesian probabilistic topic model. It is widely used as a Natural Language Processing (NLP) technique to classify documents. Given a set of documents in a corpus, each of which contains a set of words, LDA discovers K “topics” for each document, represented as latent factors each taking the form of a distribution of words. The LDA model assumes a generative process for each document d that its topics follow Multinomial(θd) drawn from Dirichlet θd~Dir(α), and each word wdn is chosen from selected topic by wdn ∼ Multinomial (βzn ) drawn from Dirichlet βk ∼ Dir(η).

Two major algorithms for inferring the model parameters are variational inference [2] and gibbs sampling [4]. The former is a simplified approximation method that results in a biased estimate, but is computationally efficient with well-defined numerical convergence criteria, while the latter is a probabilistic technique that converges to an unbiased solution, but its convergence is hard to diagnose. Here we discuss briefly the variational inference EM algorithm, which is used in our customer segmentation implementation.

In the variational inference EM algorithm, the intractable distribution of the hidden variables given the data P(θ,z|w,α,β) is approximated by a simplified, conditionally independent variational distribution as shown below:

Where and {, …, } are free variational parameters. In the E-step, given fixed parameter inputs and , optimal values of and are found using the algorithm as follows. The algorithm indicates a requirement of operations for a single document. Since empirical experiment suggests the number of iterations required is in the order of to achieve convergence, the total number of operations is roughly for a corpus of size M.

|  |
| --- |
| E-step with fixed and inputs: |
| 1. initialize :=1/K for all word n and topic i  2. initialize := for all topic i  3. repeat  4. for n=1 to  5. for i=1 to K  6.  7. normalize to sum to 1  8.  9. until converge |

The M-step is completed by maximizing the resulting lower bound on the log likelihood with respect to and . The two steps repeat until the lower bound of likelihood converges[[1]](#footnote-2). The complexity of the M-step is , where V is the total vocabulary in the corpus.

When applied to customer wire transfer data, we regard data of each customer as a document and each transaction as a word, LDA then learns the “topics” of the customers, formed by the type of transactions they make. Our customer segmentation experiments are implemented using genism LDA package based on onlineldavb.py script, by Hoffman, Blei, Bach: Online Learning for Latent Dirichlet Allocation, NIPS 2010 has been used. The package allows for single processer, multicore parallelization as well as online learning modules.

Reference

[2]  D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[4]  T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

1. Again variational inference provides a tractable lower bound on the log likelihood, which can be maximized with respect to and . [↑](#footnote-ref-2)