Latent Dirichlet Allocation: An overview of the model and use cases

Technical Review

Milad Parvaneh
Email address: sparva2@illinois.edu

November 20, 2022

Abstract

Latent Dirichlet Allocation has been well-known to be a suitable method for many text mining/analysis applications. Examples of such applications are text classification, topic extraction, and document modeling. In this brief technical review, we revisit some of this method's principal aspects and elaborate on some of its representative formulations.

Introduction

Text mining and analysis have been growing for years, with many practical applications. Topic extraction, text classification, and document clustering are some examples of these applications. Early approaches to tackle these tasks were largely developed by information retrieval researchers, which included exhibiting a document by a vector of real numbers. These numbers were chosen to be ratios of some count. These ratios are known to be useful, but not necessarily adequate, features representing a document. One classic example is the TF-IDF scheme [Salton and McGill, 1983]. Using such a scheme would enable us to express documents of any given length by fixed-length vectors.

Although the TF-IDF scheme had advantages, it suffered from capturing the inter- and intra-document structures. Such drawbacks were later addressed by more complex expression methods, among them Latent Semantic Indexing (LSI). An improved version of LSI was also developed, known as probabilistic LSI (pLSI), in which each word is modeled as a sample drawn from a mixture model. The reader is referred to [Deerwester et al., 1990] and [Hofmann, 1999], among others, for further reviews.

Upon further improvements, the Probabilistic Latent Semantic Analysis (PLSA) [Hofmann, 2013] and Latent Dirichlet Allocation (LDA) [Blei et al., 2003] were developed, which have been known to be the leading methods for topic modeling within the text mining/analysis literature. LDA, generally speaking, is viewed as a generative probabilistic model in the sense that it imposes a Dirichlet distribution as a prior. Dirichlet distributions are probability distributions of the continuous multivariate type that are typically utilized as prior distributions in Bayesian statistics.

Model overview

The parameters used to describe the model are listed in Table 1. In order to simplify notation, we rewrite

$$\phi_w^{(j)} = P(w \mid z = j) \tag{1}$$

$$\theta_j^{(d)} = P(z = j \mid d) \tag{2}$$

One characteristic feature of LDA, compared with PLSA, is that the topic mixture θ is sampled from a Dirichlet prior, as the method name suggests. The Dirichlet distribution has $\alpha_1, \ldots, \alpha_K$ parameters where each can be considered a counter for how many times a related topic is sampled in a given document. This probability distribution is formulated as

$$\begin{aligned} P(\theta^{(d)} \mid \alpha) &= \mathrm{Dir}(\theta^{(d)} \mid \alpha) && (3) \\ &= \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{j=1}^{K} (\theta_j^{(d)})^{\alpha-1} && (4) \end{aligned}$$

Here, $\Gamma(x)$ is the Gamma function. The reader is referred to [Blei et al., 2003] and [Lu et al., 2011] for detailed formulations.

In a similar manner, the word distributions ϕ can be drawn from a Dirichlet prior for which we use parameters $\beta_1, \ldots, \beta_M$. M is the size of the vocabulary. By integrating over ϕ and θ, we can express the probability of the collection C as

$$P(C \mid \alpha, \beta) = \int_{\phi^{(1)}} \cdots \int_{\phi^{(K)}} \prod_{i=1}^{K} P(\phi^{(i)} \mid \beta) \prod_{d \in D} \int P(\theta^{(d)} \mid \alpha) \prod_{w \in V} \left(\sum_{j=1}^{K} \theta_j^{(d)} \phi_w^{(j)} \right)^{c(w,d)} \mathrm{d}\theta^{(d)} \mathrm{d}\phi^{(1)} \ldots \mathrm{d}\phi^{(K)} \tag{5}$$

As for estimating the parameters of LDA, we need to employ approximate inference methods. Examples are expectation propagation, variational techniques, and Gibbs sampling. Exact inference cannot be utilized as LDA consists of a multi-level

Table 1: LDA model parameters.

Terms	Represented by
Vocabulary	V
Collection of documents	C
Any given topic	z (finite set of K topics)
Distribution of words for a given topic z	$P(w \mid z)$
Distribution of topics for a given document d	$P(z \mid d)$

hierarchical Bayesian model with Dirichlet distributions as priors.

Given all these extra complexities introduced by LDA, it then becomes capable of overcoming overfitting issues that came with PLSA. LDA is also better at evaluating previously unseen documents. In addition, LDA has been shown to exhibit better performance in classification tasks as long as its hyper-parameters are properly optimized.

Conclusion

In this brief overview, we revisited some primary characteristic features of LDA. Also, we scanned over some representative formulations of the model. LDA is a robust method for many text mining and analysis tasks. Moreover, one can find many improved versions that are developed on top of the original model.

Here, we mentioned some major use cases of LDA relating to text mining/analysis applications. However, the reader is referred to [Lu et al., 2011], among others, for further details on task performance and hyper-parameter optimization.

References

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

T. Hofmann. Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*, 2013.

Y. Lu, Q. Mei, and C. Zhai. Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. *Information Retrieval*, 14(2):178–203, 2011.

G. Salton and M. J. McGill. *Introduction to modern information retrieval.* mcgraw-hill, 1983.