Understanding of Latent Dirichlet Allocation (LDA)

Sampath Lakkaraju
(011818781)
sampath.lakkarjau@sjsu.edu
San Jose Sate University, San Jose.

Abstract—**This paper is aimed for the demonstration of my basic understanding of Topic modelling algorithm Latent Dirichlet Allocation (LDA). The paper will introduce the basics of the Topic modeling, natural language processing and in detailed explain about the assumptions and working of the algorithm. The paper will also briefly touch the simplification made to it and Pseudo code obtained from them**.

Keywords— Latent Dirichlet Allocation, Topic modelling, Latent variables, Dirichlet distribution, Multinomial distribution, corpus, Document, Topic.

1. Introduction

In this digital world huge amount of data is being generated every day. A big portion of this data is for the information storage and exchange. It may be News, Social media content or any other content which are understandable by humans. This data needs to be processed for either better storage or to **achieve** the information it has for any kind of gains. NLP is the field of computer science allows us to obtain information from this type of data. Topic modelling is one of the techniques in NLP which provides the relation of the unknown data to known ones. Numerous algorithms have been developed in the aspect of extracting topic information from data. One of these algorithms and most widely used one is Latent Dirichlet allocation (LDA). It is a algorithm based upon the concept of bag of words and is one of the few which considers numerous topics can be present in a document. This paper explains the basic concepts of the algorithms and pseudo code, understood from the paper written by David M. Blei, Andrew Y. Ng and Michael I. Jordan in 2003 [1].

2. Scope

This paper talks about the Latent Dirichlet Allocation and will not touch any of the topics LDA is based upon such as pLSI and bag of words and the domains under which it falls. Also, the paper explains the working of the algorithm centered around text documents but not the broader aspects for which the algorithm can be applied.

3. Definitions

Natural Language Processing: Natural Language Processing (NLP) is the study of how to program computers to process and analyze large amount of natural language (human language) data [1]

Topic modelling: Topic modelling is a technique of in NLP where statistical models are developed to for gaining information about the collection of documents [2].

Bag of Words: Is a simplified representation of text documents wherein the document is represented as set of words it contains.

4. Latent Dirichlet allocation(LDA)

LDA is a generative statistic model which allows a set of observations to be explained by latent variables (variables that are not directly observed) that can explain the similarities in the data [1]. It is an example of Topic model

The definition put forth by David M. Blei [2] If there are K underlying topics using which documents are generated, and each topic is generated by multinomial distribution over the |V| words in the vocabulary. A document is generated by sampling of

these mixtures and then generating words from these mixtures.

In simple words, LDA assumes that a document is a collection of topics and words in the document are related to the topics. The probability of the word distribution over topics and topic distribution of words id calculated in the LDA.

For precisely defining LDA we need to understand few terminologies that is used.

- Word is the basic unit of the data and is represented by a vector w where the single component is equal to 1 and all other as 0.
- A document is the sequence of **N** words represented by **w =** (w1, w2 , ….., wn)
- A corpus is a collection of **M** documents represented by D = {**w1, w2 , ……, wm)**

Now LDA assumes the following process for generation of each document **w** in a corpus D:

- Select **N** of document using a Poisson distribution (ξ)
- Choose θ from the Dirichlet distribution(α) over K topics.
- Now for N select each word w_n by:
    - A topic z_n is chosen from Mult(θ) (multinomial distribution).
    - Select a word w_n from the $p(w_n|z_n ; \beta)$ (multinomial probability condition on topic for words)

Where α is the Dirichlet prior parameter for per-document topic distribution and β is Dirichlet prior parameter for per topic word document.

Representing the probability of words in a document in an equation:

$$\boldsymbol{p(w)} = \int_{\theta} \left(\prod_{n=1}^{N} \sum_{Z_n=1}^{K} \boldsymbol{p(w_n|z_n;\beta)p(z_n|\theta)} \right) \boldsymbol{p(\theta;\alpha)d\theta}.. \boldsymbol{(eq1)}$$

Where $p(z_n|\theta)$ is the multinomial distribution for topics z_n, p(θ, α) Dirichlet distribution.

The Figure1 represents a graphical model of LDA. The boxes also known as “plates” represents replicates. The outer plate represents documents while the inner plate represents the choice of words and topics which are repeated accordingly. The plate M denotes the total no of documents in the corpus

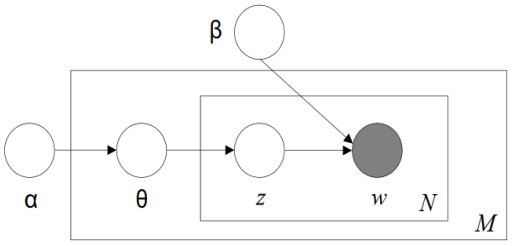

Figure 1:Graphical representation of LDA

while N plate denotes the total no of word in a document.

The α and β are corpus level parameters and are assumed to be generated once in the process. θ is sampled once per document and z and w are sampled once for every word. These three level assumptions will help in considering multiple topics for one document and leads to higher accuracy of topic modelling.

We have discussed how the LDA assumes a document is prepared and its components. For LDA to learn topic representation of the corpus it considers the assumption in reverse order. Working backwards LDA randomly assigns each word in the all the documents to one of the K topics specified. Now the all the word distributions of other documents except one under consideration is assumed as correct. Then the LDA calculates $p(w_n|z_n ; \beta)$ (multinomial probability condition on topic for words) and $p(z_n|\theta)$ (multinomial distribution for topics in the document). A new topic based on the multiplication of the two proportion is assigned to the word. This is repeated till a steady state is achieved where the distribution makes sense.

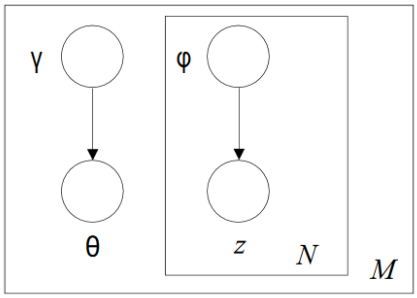

Figure 2: Simplified graphical representation of LDA negelcting external distributions.

The equation 1 has various parameters that need not be considered for every scenario. For simplification the equation 1 and the graphical representations are transformed to the below given figure and equations with help of variational inferences. These transformations are discussed in detailed in the paper [3] and will not be discussed for this paper.

The new graphical representation presented in the figure2 reduces the edges between θ, z, w. the resultant is Dirichlet parameter $\boldsymbol{\gamma}$ and multinomial parameter $\boldsymbol{\varphi}$ [3]**.** As mentioned above the following are the equations which represents a simplified LDA.

$$\boldsymbol{\varphi_{ni} \propto \beta_{iw_n} exp\{E_q[log(\theta_i)|\gamma]\} \dots (eq2)}$$

$$\boldsymbol{\gamma_i = \alpha_i + \sum_{n=1}^{N} \varphi_{ni} \dots (eq3)}$$

As shown in the paper [3] the multinomial update has been computed and can be represented as follows.

$$\boldsymbol{E_q[log(\theta_i)|\gamma] = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_i \dots (eq4)}$$

From these three equations we can generate a pseudo code as follows:

1. *Initialize* $\boldsymbol{\varphi_{ni}^0 := \frac{1}{k}}$ *for all n and i.*
2. *Initialize* $\boldsymbol{\gamma_i := \alpha_i + \frac{N}{k}}$ *for all i.*
3. *Repeat*
    a. *For n=1 to N*
        i. *For i= 1 to K*
            1. $\boldsymbol{\varphi_{ni}^{t+1} := \beta_{iw_n} exp\{\Psi(\gamma_i^t)\}}$
            2. *normalize* $\boldsymbol{\varphi_{ni}^{t+1}}$ *to sum to 1*
    b. $\boldsymbol{\gamma_i = \alpha_i + \sum_{n=1}^{N} \varphi_{ni}}$
4. *Until Convergence*

The following pseudo code is implemented creates the LDA algorithm used for text topic modelling.

5. Implementation

A simple code in the form of jupyter notebook will be submitted which shows the how LDA works. The algorithm will not be developed from scratch, instead will the library **Gensim** implementation will be used**.**

6. Limitations

For using the LDA algorithms one must specify the number of topics that has to be considered, this may either restrict the classification of the document to fewer topics or may lead to unnecessary border classification. Also, the topic distribution cannot specify the correlation among topics [6].

7. Conclusion

LDA algorithm takes in a corpus, assumes each word is related to topics. Tries to establish a formula though which the document has been created. This formula will help us in classifying new documents to similar ones and identify similar documents in the corpus.

8. References

[1] A. Y. N. M. I. J. David M. Blei, "Latent Dirichelt Allocation," Berkely, 2003.

[2] anon, "Natural language processing," wikipedia, dec 2018. [Online]. Available: https://en.wikipedia.org/wiki/Natural_language_processing.

[3] "Topic_model," wikipedia, dec 2018. [Online]. Available: https://en.wikipedia.org/wiki/Topic_model.

[4] "Latent_Dirichlet_allocation," wikipedia, 29 10 2018. [Online]. Available: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation. [Accessed 12 10 2018].

[5] A. Y. N. David M. Blei, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* 2003.

[6] S. Sullivan, *https://www.youtube.com/watch?v=DWJYZq_fQ2A,* youtube.