BOUND DIRICHLET BELIEF NETWORK FOR TOPIC RECONSTRUCTION

Huiqiang Zhong

Supervisor: Dr. Xuhui Fan

School of Mathematics and Statistics
UNSW Sydney

July 2020

Submitted in partial fulfilment of the requirements of the degree of
Master of Statistics

Plagiarism statement

I declare that this thesis is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

I acknowledge that the assessor of this thesis may, for the purpose of assessing it:

- Reproduce it and provide a copy to another member of the University; and/or,
- Communicate a copy of it to a plagiarism checking service (which may then retain a copy of it on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct, and am aware of any potential plagiarism penalties which may apply.

By signing this declaration I am agreeing to the statements and conditions above.

Signed: _____ Date: _____

Abstract

Topic model,especially latent Dirichlet allocation(LDA),has aroused grea research interest in the past two decades which is a useful tools for document classification.The LDA model assume that word of doucment is determined by the topic of document and topic-word matrix.Some multi-layer generative process on word distribution on have been proposed to improve LDA.However it suffer from information decay and complicate samping method. Here we present a Dirichlet-Belief Networks to address problems above. By inserting auxiliary Poisson ran-dom variables into the layerwise connections and appropriate design, we can inference latent parameter in an efficient way.As an Bayesian generatice model,it is more interpretability and gibbs sampling can be used in training model.

Contents

Chapter 1 Introduction 1
1.1 Background and Importance 1
1.2 topic of research paper 2
1.3 Highlight the approach and principal finding 2
1.4 Organization of this paper 2

Chapter 2 Basic language 3
2.1 Bayesian Theorem 3
2.1.1 Prior Probability 3
2.1.2 Likelihood Probability 4
2.1.3 Posterior probability 4
2.1.4 conjugate prior 4
2.2 Basic distribution 4
2.2.1 Beta distribution 4
2.2.2 Dirichlet distribution 5
2.2.3 Poisson and Multinomal distribution 5
2.2.4 conclusion 5
2.3 Markov Chain Monte Carlo and Gibbs Sampling 6
2.3.1 Monte Carlo Method 6
2.3.2 Sampling Method 7
2.3.3 Markov Chain 8
2.3.4 MCMC Sampling and M-H Sampling 9
2.3.5 Gibbs Sampling 10

Chapter 3 Latent Dirichlet Allocation 12
3.1 Model construction 12
3.1.1 Unigram and mixture of unigram model 12
3.1.2 Probabilistic Latent Semantic Analysis 13
3.2 Latent Dirichlet Allocation 13
3.2.1 Generative process of document 14
3.2.2 Gibbs Sampling for LDA 16
3.2.3 Perplexity and Inference 17
3.3 Dirichlet belief networks 17
3.3.1 Research on Topic Distribution 17
3.3.2 Introduction of DIRBN 18

Chapter 4 Conclusion 19

References 20

Chapter 1

Introduction

1.1 Background and Importance

With the development of technology and alomost all information is digitized, we are more and more likely to be overwhelmed by a large amount of information, and it is difficult to find an effective way to find, organize and understand a large amount of information[1].Then, how to effectively obtain better information and how to automatically classify vast amounts of text data, organization and management become more challenging. Therefore, in the face of these problems and needs, the use of computers for language information processing has been extensively studied. As a research hotspot in the field of natural language processing, automatic text classification technology has been rapidly developed and widely used.

Such data is sparse and discrete date which is hard for computer to process.The bag of words models are proposed for samplifying representation in natural language processing.This model can convert a sentence into a vector representation, which is a relatively straightforward method. It does not consider the order of words in the sentence, only considers the number of occurrences of words in the vocabulary in this sentence so that it will miss the relationshaip between each individual vocabulary and induces huge prolem of dimensional disaster. However it still a popular way in representing document.

After representing document by matrix,scientists hope to construct a generative model to Imitate document writing. It reduces a complex precudure into some probabilistic steps and specify a sample distribution for the topic of document[3]. We will introduce a very well-known method in natural language processing - Latent Dirichlet Allocation(LDA) which can extract abstract "themes" from a series of documents through the mechanism of generating models.

This model assumes the implicit mechanism of published documents written by humans: each document is composed of a few topics, and each topic can be composed of a few important words description. In other words, When writing an article, we will first decide the topics related to our document, then each word we write in text is closely related to these topics.Let's follow the generative process of LDA step by step:

- For each document, we extract a theme from the theme distribution.
- Extracting a word from the word distribution corresponding to the abovementioned theme.
- Repeat the above process until each word in the document is traversed.

This generative process is mainly compose of two distributions-the topic distribution of each docuemnt and the word distribution of each topic.In LDA, the topic distribution and word distribution are uncertain. The authors of LDA adopt the Bayesian idea that they should obey a distribution. The topic distribution and word distribution are subject to polynomial distributions,so the topic distribution and word distribution use Dirichlet distribution as their conjugate prior distribution because the polynomial distribution and Dirichlet distribution is a conjugate structure.
As mentioned above,one commonly-used prior for topic distribution and word distribution is diriclet distribution and recently lots of distributions have been imposed on topic distribtion.For example,The correlation topic model(CTM)[4],which exhibits the correlation of topic by introducing logistic noremal distribution.It allows each document to show different topics with different proportions and capture the relationship in groud data.Besides,there are hierarchical document representation based on dirichlet process and Boltzmann machines and neural network.[2].
?

1.2 topic of research paper

Announce the research topic/question being addressed in research paper

1.3 Highlight the approach and principal finding

DEscrition of your paper approach and why you chose it
brief summary of your major findings

1.4 Organization of this paper

Chapter 2

Basic language

In this chapter we outline the basic background of Latent Dirichlet Allocation . We begin by reviswing the Bayesian Theorem associated with functional analysis in LDA. Then some common statistic distributions will be introduced which are used in inferencing posterior distribution and latend parameters.Sampling techniques such as Gibbs Sampling which is frequently used to obtain results of model will be listed in last subsection.

2.1 Bayesian Theorem

We can never be completely sure of this world, because it is a constantly changing being, and change is the essence of reality. However, what we can do is, as expressed by this theorem, as we obtain more and more data or evidence, our knowledge of reality has been updated and improved

$$\begin{aligned} P(A|B) &= \frac{P(A,B)}{P(B)} \\ &= \frac{P(B|A) * P(A)}{P(B)} \end{aligned}$$

- $P(A|B)$ is the conditional probability of A after the occurrence of B is known, and is also excluded from the posterior probability of A due to the value obtained from B, indicating the confidence that event A will occur after event B occurs.
- $P(A)$ is the a priori probability or edge probability of A, and represents the confidence that event A occurs.
- $P(B|A)$ is the conditional probability of B after the occurrence of A. It is also called the posterior probability of B because of the value obtained from A, and is also considered as a likelihood function.
- $P(B)$ is the prior probability or edge probability of B, which is called a standardized constant.
- $P(B|A)P(B)$ is called the standard likelihood ratio (there are many names, and no unified standard name is found), which indicates the degree of support provided by event B for the occurrence of event A.

2.1.1 Prior Probability

A priori probability refers to an event that has not yet occurred, and an estimate of the probability of the event occurring, describing a variable in the

absence of something.Prior information comes from experience and historical data

2.1.2 Likelihood Probability

- If the "probability function" P(B—A)/P(B)¿1, it means that the "prior probability" is enhanced and the probability of the occurrence of event A becomes greater;
- If "probability function" = 1, it means that event B does not help to determine the possibility of event A;
- If the "probability function" ¡1, it means that the "prior probability" is weakened, and the probability of event A becomes smaller.

2.1.3 Posterior probability

The posterior probability refers to the probability that the cause of the event is caused by a factor under the condition that the event has occurred. It is the conditional probability after considering an event.

2.1.4 conjugate prior

The posterior probability distribution function has the same form as the prior probability distribution function If the prior distribution and the likelihood function can make the prior distribution and the posterior distribution (posterior distributions) have the same form, then the prior distribution and the likelihood function are said to be conjugate. Therefore, conjugate refers to the prior probability distribution and likelihood function. If the posterior probability p(—x) and the gas prior probability p() of a random variable belong to the same distribution cluster, then p(—x) and p() are called conjugate distributions, and , Also called p() is the conjugate prior of the likelihood function p(x—).
The conjugate prior and posterior have the same form. This can easily form an iteration in the calculation process. According to the new observation data, the original posterior probability becomes a new prior probability, and then a new posterior probability is updated. The parameters of this posterior probability are more accurate. . This process greatly simplifies Bayesian analysis

2.2 Basic distribution

2.2.1 Beta distribution

The beta distribution can represent the probability distribution of a probability. When you don't know the specific probability of a thing, it can be called the probability of the occurrence of all probabilities
Definition The Beta Function,For each positive α and β.define:

$$P(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma\alpha+\beta}$$

For example,carry out N times of Bernoulli test, the probability of success of the test p is subject to an a priori probability density distribution $Beta(\alpha, \beta)$, the test result appears K times of success of the test, the posterior probability density distribution of the probability p of the test success is $Beta(\alpha+K, \beta+N-K)$.Prove:
Prior distribution:

$$f(p|\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma\alpha+\beta} x^{\alpha-1}(1-x)^{\beta-1}$$

Likelihood Function:

$$f(n_1, n_2, , N|p) = p^K(1-p)^{N-K}$$

Posterior distribution:

$$f(p|n_1, n_2, ..N, \alpha, \beta) = \frac{f(n_1, n_2, , N|p)f(p|\alpha,\beta)}{f(n1, n2, ...N, \alpha, \beta)}$$

Given that:

$$\begin{aligned} f(n1, n2, ...N, \alpha, \beta) &= \int_p f(n_1, n_2, , N|p)f(p|\alpha,\beta) \\ &= \frac{1}{B(\alpha,\beta)} \int_p p^{\alpha+K-1}(1-p)^{\beta+N-K-1} \\ &= \frac{B(\alpha+k, \beta+N-K)}{B(\alpha,\beta)} \end{aligned}$$

$$\begin{aligned} f(n_1, n_2, , N|p)f(p|\alpha,\beta) &= \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} * \frac{f(n_1, n_2, , N|p)f(p|\alpha,\beta)}{f(n1, n2, ...N, \alpha, \beta)} \\ &= \frac{1}{B(\alpha,\beta)} p^{\alpha+K-1}(1-p)^{\beta+N-K-1} \end{aligned}$$

So that:

$$\begin{aligned} f(p|n_1, n_2, ..N, \alpha, \beta) &= \frac{1}{\alpha+K-1}(1-p)^{\beta+N-K-1} p^{\alpha+K-1}(1-p)^{\beta+N-K-1} \\ &= Beta(\alpha+K, \beta+N-K) \end{aligned}$$

2.2.2 Dirichlet distribution

Dirichlet distribution is a generalization of the beta distribution in high dimensions

2.2.3 Poisson and Multinomal distribution

2.2.4 conclusion

Taking a coin toss as an example, after you get a coin, normally the probability of a uniform coin appearing on both sides should be the same, both 0.5, but

you are not sure about the material and weight distribution, you need to judge its Is it really evenly distributed

2.3 Markov Chain Monte Carlo and Gibbs Sampling

2.3.1 Monte Carlo Method

The Monte Carlo method is a calculation method. The principle is to understand a system through a large number of random samples, and then get the value to be calculated.

It is very powerful and flexible, yet quite easy to understand and easy to implement. For many problems, it is often the simplest calculation method, and sometimes even the only feasible method. As a random sampling method, Markov Chain Monte Carlo (hereinafter referred to as MCMC) has a wide range of applications in the fields of machine learning, deep learning, and natural language processing, and is the basis of many complex algorithms.

The early Monte Carlo methods were designed to solve some summation or integration problems that are not very easy to solve. Such as points:

$$Y = \int_a^b f(x)$$

We can get the answer by Newton-Leibniz formula if the function is simple enough.However, In most cases, it is difficult to find the original function of $f(x)$. Of course, we can use the Monte Carlo method to simulate the conversion approximation

Then we can sample$n$ values in the $[a, b]$, interval: $x_0, x_1, \ldots, _n$, and use their average values to represent all $f(x)$ values in the $[a, b]$ interval. So our approximate solution to the definite integral above is:

$$Y = \frac{b-a}{n} \sum^{n} f(x_i)$$

The above method has an implicit assumption that the distribution follows a uniform distribution from a to b, but the actual situation is will subject to various types of distribution.We can improve our method as follows

$$\begin{aligned} Y &= \int_a^b f(x) \\ &= \int_a^b \frac{f(x)}{p(x)} p(x) \\ &= \frac{1}{n} \sum^{n} \frac{f(x_i)}{p(x_i)} \end{aligned}$$

Our question now turns to how to find the distribution of x

2.3.2 Sampling Method

The key to the Monte Carlo method is to get the probability distribution of x. If the probability distribution of x is found, we can sample n sample sets based on this probability distribution based on the probability distribution and bring it into the Monte Carlo summation formula to solve.

For the common uniform distribution, $uniform(0,1)$,It is very easy to sample samples, generally through the linear congruential generator can easily generate between (0,1) Pseudo-random number samples. For other common probability distributions, whether they are discrete distributions or continuous distributions, their samples can be infered by $uniform(0,1)$ sample conversion.

Assuming that x is a continuous random variable, subject to a random distribution$f(x)$, its cumulative distribution function is $F(X)$.We assume that $Y = F(X)$ is subject to 0-1 uniform distribution and $F^{-1}(Y)$ have same distribution with X.For example: PDF of Expoential distribution:

$$f(x) = \lambda e^{\lambda * x}$$

CDF of Expoential distribution:

$$F(x) = 1 - exp^{\lambda * x}$$

Inverse sampling inference:

$$\begin{aligned} u & \quad Uniform(0,1) \\ F(F^{-1}(Y)) &= 1 - exp^{\lambda * F^{-1}(Y)} = u \\ F^{-1}(Y) &= -\frac{\log(1-u)}{\lambda} \end{aligned}$$

But many times, our probability distribution is not a common distribution, which means that we can't easily get a sample set of these unusual probability distributions. We need to Reject-Sampling method.

The basic idea of Reject-Sampling is to cover the "smaller probability distribution" with a "larger probability distribution". This "larger probability distribution" is more easier to sample (standard distribution), and at the same time accept this sample with a certain probability. The sample can be seen as a sample from the "smaller probability distribution"

We can solve some cases by Reject-Sampling when the probability distribution is not common.However,in the case of high dimensions, Rejection Sampling will have two problems. The first is that the proposal distribution q is difficult to find, and the second is that it is difficult to determine a reasonable value of k. These two problems will lead to a high rejection rate and an increase in useless calculations.

From the probability density function $p(X)$ of a known distribution, we want to get samples Xthat subject to this distribution.

- Sample Y from proposal distribution $p(x)$

- Sample U from uniform(0-1) distribution
- if $U < \frac{f(Y)}{c*g(Y)}$,then we accept Y,else continue.

2.3.3 Markov Chain

This is a random process from state to state in the state space. This process requires the "no memory" attribute: the probability distribution of the next state can only be determined by the current state, and the events before it in the time series have nothing to do with it. This special type of "no memory" is called the Markov attribute. Markov chains have many applications as statistical models of actual processes

In order to obtain a theoretical result, let's look at a smaller example (this will facilitate our subsequent calculation demonstration), assuming that in a region, people either live in the city or live in the countryside. The matrix below tells us some laws (or tendencies) of population migration. For example, the first column of row 1 indicates that 90 percent of the population currently living in cities will choose to continue to live in cities next year.

$$H_x = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \tag{2.1}$$

As a simple start, try to calculate the odds of people living in cities today who will live in the countryside after 2 years. Analysis shows that 90 of people currently living in the city will continue to choose to live in the city after 1 year, and 10 will move to the country. Then another year passed, and 10 of those who chose to stay in the city last year moved to the countryside. And 98 of those who moved to the village last year will choose to stay in the village. This analysis process is shown in the figure below, and finally the probability that people who live in the city will live in the countryside after 2 years = $0.90\times0.10+0.10\times0.98$

In fact, you will find that our calculation process is to do the square of the matrix. As shown in the following figure, you will find that the calculation in the second row and first column of the result matrix is performing the operation that is processed above. On this basis, we can also continue to calculate the situation after n years, that is, the result of calculating the self-multiplication of matrix A n time https://zhuanlan.zhihu.com/p/37121528

The algorithm is as follows:

Enter the Markov chain state transition matrix , set the state transition times threshold , the required number of samples ;Sampling from any simple probability distribution to get the initial state value ;sample from the conditional probability distribution ;The sample set is the corresponding sample set that meets our stationary distribution.

If it is assumed that we can obtain the Markov chain state transition matrix corresponding to the stationary distribution of the samples we need to sample,

then we can use the Markov chain sampling to obtain the sample set we need, and then perform Monte Carlo simulation.
But an important question is, given a random distribution at will, how to get the Markov chain state transition matrix P corresponding to it?

2.3.4 MCMC Sampling and M-H Sampling

Mostly, Target stationary distribution $\pi(x)$and a certain Markov chain state transition matrix Q does not satisfy the detailed stationary condition:

$$\pi(i)Q(i,j) \neq \pi(j)Q(j,i)$$

We introduce a $\alpha(i,j)$ so that the above formula can take the equal sign.

$$\pi(i)Q(i,j)\alpha(i,j)\pi(j)Q(j,i)\alpha(j,i)$$

But how can we get the α,by symmetry:

$$\begin{aligned} \alpha(i,j) &= \pi(j)Q(j,i) \\ \alpha(j,i) &= \pi(i)Q(i,j) \end{aligned}$$

α is generally called acceptance rate, and the value is between $[0-1]$, which can be understood as a probability value. This is much like accept-reject sampling, where a common distribution is obtained through a certain acceptance-rejection probability, and here is a common Markov chain state transition matrix Q through a certain acceptance-rejection probability Obtaining the target transition matrix p, the two solutions to the problem are similar.
MCMC sampling algorithm is as follows:

(a) Enter any given Markov chain state transition matrix Q, target stable distribution $\pi(x)$, set the threshold of state transition times $n1$, the number of required samples $n2$;
(b) Get the initial state value x_0 from any simple probability distribution;
(c) For t =0 in $n_1 + n_2 - 1$
- Get the sample value x_* from the conditional probability distribution $Q(x_*|x_0)$.
- Sample U from uniform distribution.

- if $u < \pi(x_*)Q(x_*|x_0)$then accept x_*.

But this sampling algorithm is still more difficult to apply in practice, because in the third step, because accept rate α may be very small, such as 0.1, most of our sampled values are rejected and the sampling efficiency is very low. It is possible that we have sampled millions of Markov chains and have not yet converged, that is, the above n_1 should be very large, which is unacceptable. At this time, it is our Metropolis-Hastings sampling Method.
M-H sampling solves the problem of low MCMC sampling acceptance rate in the previous section.

We can expand both sides of

$$\pi(i)Q(i,j)\alpha(i,j)\pi(j)Q(j,i)\alpha(j,i)$$

, and at this time the detailed stationary condition is also satisfied. We expand the equation by C times to make $c\alpha(ij)$ (accurately, the maximum expansion of both sides is 1), so that we can Improves the acceptance rate of jumps in sampling, so we can take:

$$\alpha = min\frac{Q(j,i)\alpha(j,i)}{Q(i,j)\alpha(i,j)}, 1$$

In the era of big data, M-H sampling faces two major challenges:

(a) Our data features are very many, M-H sampling due to the existence of the acceptance rate calculation formula [formula], the calculation time required in high dimensions is very considerable, and the algorithm efficiency is very low. At the same time, [Formula] is generally less than 1. Sometimes it is hard to calculate but it is rejected. Can it be done without refusing to transfer?

(b) Due to the large feature dimension, it is often difficult to find the joint distribution of each feature dimension of the target, but it is convenient to find the conditional probability distribution between each feature. At this time, can we only have convenient sampling in the case of conditional probability distribution between various dimensions?

2.3.5 Gibbs Sampling

Gibbs Sampling Method is a one special MCMC technique used for sampling variales in large dimensions by sampling each variable from its conditional distribution iterative.

Starting from a two-dimensional data distribution, assuming that $\pi(x_1, x_2)$ is a two-dimensional joint data distribution, observe the first two points with the same feature size$A(x_1^{(1)}, x_2^{(1)}$ and $B(x_1^{(1)}, x_2^{(2))}$.For example:

$$\pi(x_1^{(1)}, x_2^{(1)}\pi(x_2^{(2)}|x_1^{(1)}) = \pi(x_1^{(1)})\pi(x_2^{(2)}|x_1^{(1)})\pi(x_2^{(1)}|x_1^{(1)}$$

$$\pi(x_1^{(1)}, x_2^{(1)}\pi(x_2^{(1)}|x_1^{(1)}) = \pi(x_1^{(1)})\pi(x_2^{(2)}|x_1^{(1)})\pi(x_2^{(1)}|x_1^{(1)}$$

Since the right sides of the two formulas are equal, we have:

$$\pi(x_1^{(1)}, x_2^{(1)})\pi(x_2^{(2)}|x_1^{(1)}) = \pi(x_1^{(1)}, x_2^{(1)})\pi(x_2^{(1)}|x_1^{(1)})$$

Observing the above detail balance formula , we find that on the straight line of $x_1 = x_1^{(1)}$, if the conditional probability distribution $\pi(x_2|x_1^{(1)})$ is used as the state transition probability of the Markov chain, the transition between any two points meets Detail balanceconditions! In the same way, on the straight line of $x_2 = x_2^{(1)}$, if the conditional probability distribution $\pi(x_1|x_2^{(1)})$ is used as the state transition probability of the Markov chain, the transition between any two points also meets the detail balance condition.

With the transition matrix, we can infer a two-dimensional Gibbs sample, which requires a conditional probability between two dimensions. The algorithm is as follows:

(a) Given stationary distribution $\pi(x_1, x_2)$,Set the threshold value of the number of state transitions n_1, the number of required samples n_2.

(b) Randomly initialize values $x_1^{(1)}$ and $x_2^{(1)}$.

(c) for t in [0,n1+n2-1]:

- Get sample $x_2^{(t+1)}$ from conditional distribution $p(x_2|x_1^t)$
- Get sample $x_1^{(t+1)}$ from conditional distribution $p(x_1|x_2^t)$

Chapter 3

Latent Dirichlet Allocation

The topic model is an important tool for text mining. In recent years, it has received a lot of attention in industry and academia. In the field of text mining, a large amount of data is unstructured, and it is difficult to obtain relevant and expected information directly from the information. A method of text mining: the topic model can identify the topics in the document and mine Information is hidden in the corpus, and has a wide range of uses in scenarios such as topic aggregation, extracting information from unstructured text, and feature selection Latent Dirichlet Allocation (LDA) [5]is the most representative model among them.This popular method can be applied in various domians such as Discovery of theme patterns hidden in the corpus,Classification of the document according to the theme.

3.1 Model construction

3.1.1 Unigram and mixture of unigram model

Before intrducing Topic-word Model,Let us review some basic model for short test. N represents the number of words in the document to be generated, w_n represents the nth word w generated, and $p(w)$ represents the distribution of the word w, which can be obtained through statistical learning of the corpus, such as giving a book to count each word in The probability of occurrence in the book.The whole test can be represented by Vector $W = (w_1, w_2, \ldots, w_n)$. We assume that words are independent of each other, so that the probability that this document will be generated is:

$$\begin{aligned} p(W) &= p(w_1, w_2, \ldots, w_n) \\ &= p(w_1)p(w_2)..p(w_n) \end{aligned}$$

As mentioned in section of Introduction,each text can be converted into a vector $N = (n_1, n_2, \ldots, n_V)$ by Bag of Word model and V is number of Vocabulary. We further assume that the text matrix is subject to a multinomial distribution.

$$p(w_1)p(w_2)..p(w_n) = \prod_{k=11}^{V} p_k^{n_k}$$

The disadvantage of the method of the unigram model is that the generated text has no theme and is too simple. The mixture of unigram[**?**] method

improves it. The model samples a topic from distribution $p(z)$ before generate each word .

$$p(W|Z) = \sum_{z} p(z) \prod_{n=2}^{N} p(w_n|z)$$

z represents a theme, $p(z)$ represents the probability distribution of the theme, z is generated by $p(z)$ according to probability; $p(w|z)$ represents the distribution of w given z, which can be regarded as a $k * V$ matrix, k is the number of topics, V is the number of words, each line represents the probability distribution of words corresponding to this topic, that is, the probability of each word contained in topic z, generated by this probability distribution with a certain probability.

3.1.2 Probabilistic Latent Semantic Analysis

Another widely used topic model is PLSA model. In the PLSA model, the topic is actually a probability distribution on words. Each topic represents a probability distribution on a different word, and each document can be regarded as a probability distribution on the topic. Each document is generated by such a two-layer probability distribution, which is also the core idea of the generation model proposed by PLSA[**?**].
PLSA models the joint distribution of d and w by the following formula:

$$p(d, w_n) = p(d) \sum_{k=1}^{K} p(w_n|z)p(z|d)$$

It is easy to find that for a new document, we cannot know what its corresponding $P(d)$ is, so although the PLSA model is a generative model on a given document, it cannot generate a new unknown document.Another problem with this model is that as the number of documents increases, the parameters of $P(z|d)$ also increase linearly, which leads to the problem of overfitting the model no matter how much training data there is. These two points have become two major flaws that limit the more widespread use of the PLSA model.

3.2 Latent Dirichlet Allocation

In PLSA, we will extract a subject word with a fixed probability and then find the corresponding word distribution according to the extracted subject word.Then according to Word distribution, extract a vocabulary. It can be seen that in PLSA, the topic distribution and word distribution are uniquely determined. However, in LDA, the topic distribution and word distribution are uncertain. The authors of LDA adopt the Bayesian idea that they should obey a distribution. The topic distribution and word distribution are both polynomial distributions, because the polynomial distribution And Dirichlet distribution is a conjugate structure. In LDA, the topic distribution and word distribution use Dirichlet distribution as their conjugate prior distribution.Thus,On the basis of PLSA, LDA adds two Dirichlet priors for topic distribution and word distribution.

3.2.1 Generative process of document

The LDA model can be represented by the following probability graph model:

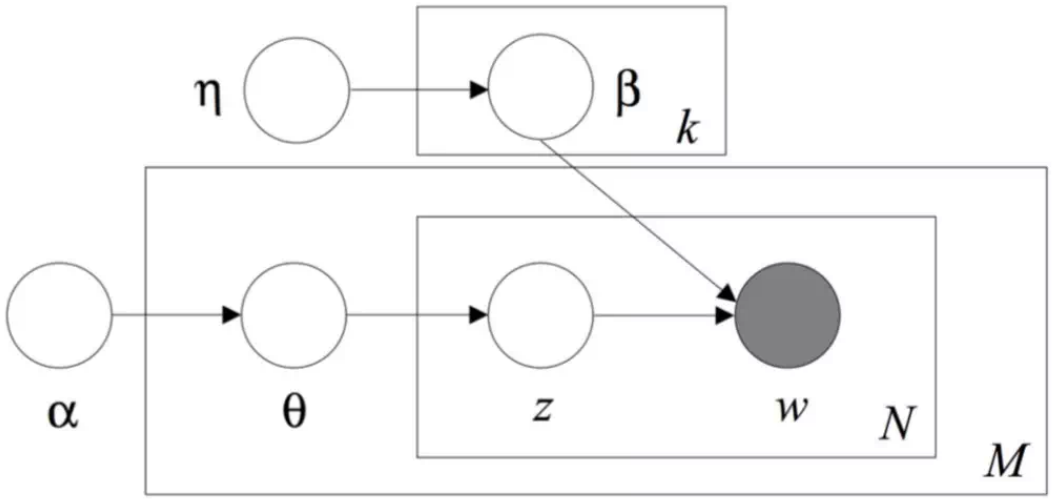

Figure 3.1: The caption of this figure.

In the LDA model, a document is generated as follows:

(a) Sampling from Dirichlet distribution α to generate the topic θ distribution of document i.

(b) Sample the jth word of the topic $z_{i,j}$ of the document i from the polynomial distribution of topics θ_i.

(c) Sampling from Dirichlet distribution eta to generate word distribution$\beta_{z_{i,j}}$ corresponding to topics$z_{i,j}$.

(d) Sampling from the polynomial distribution$\beta_{z_{i,j}}$ of words to finally generate words $w_{i,j}$.

This probability map can be decomposed into two parts:

(a)

$$\alpha \to \theta \to z$$

This process means that when generating the m-th document, a candidate for a topical toll is generated, and then the topic of each word in the document is randomly generated according to the toll

(b)

$$\eta \to \beta \to w|k$$

This process means that the word of the document is generated when the topic number is known, and the vector with the topic k is selected in the topic-word matrix, and then the word is generated according to this vector

In the process of document construction,M documents correspond to M independent Multinomial-Dirichlet conjugate structures, and K topics also correspond to K independent Multinomial-Dirichle conjugate structures. Among

them, $M + K$ conjugate structures are all independent. Let's discuss the document generative process further.
From the first decomposition,We know that $\alpha \to \theta$ represents the theme corresponding to all documents, which is subject to the Dirichlet distribution. And $\theta \to z$ is to generate the theme corresponding to each word, subject to the multinomial distribution. and the Dirichlet distribution is the conjugate distribution of the polynomial distribution. All the whole is a conjugate structure

$$\begin{aligned} f(z_k|\alpha) &= \int f(z_k|p)f(p|\alpha)d_p \\ &= \int \prod_{k=1}^{K} p_k^{n_k} Dir(p|)d_p \\ &= \int \prod_{k=1}^{K} p_k^{n_k} \frac{1}{B(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} d_p \\ &= \frac{1}{B(\alpha)} \int \prod_{k=1} V p_k^{n_k + \alpha_k - 1} d_p \\ &= \frac{B(n_k + \alpha_k)}{B(\alpha)} \end{aligned}$$

Vector $n = (n_{1,2}, ..n_K)$ represents the number of words of topic V in each document. B is the gamma function:

$$B(\alpha) = \frac{\prod_{i=1}^{V} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{V} \alpha_i)}$$

For the second decomposition,We know that $\eta \to \beta$ represents the word-distribution corresponding to all topics, which is subject to the Dirichlet distribution. And $\beta \to w|k$ is to generate the word corresponding to each topic, subject to the multinomial distribution. and the Dirichlet distribution is the conjugate distribution of the polynomial distribution. All the whole is a conjugate structure

$$\begin{aligned} f(w|\beta) &= \int f(w|\beta)f(\beta|\eta)d_\beta \\ &= \int \prod_{k=1}^{V} p_k^{n_k} Dir(p|\alpha)d_p \\ &= \int \prod_{k=1}^{V} p_k^{n_k} \frac{1}{B(\alpha)} \prod_{k=1}^{V} p_k^{\alpha_k - 1} d_p \\ &= \frac{1}{B(\alpha)} \int_{k=1} V p_k^{n_k + \alpha_k - 1} d_p \\ &= \frac{B(n_k + \alpha_k)}{B(\alpha)} \end{aligned}$$

Vector $n = (n_{1,2}, ..n_V)$ represents the number of words generated by each topic V.
We assume two vector:

$$\vec{w} = (\vec{w_1}, \vec{w_2}, ..\vec{w_k})$$
$$\vec{z} = (\vec{z_1}, \vec{z_2}, ..\vec{z_k})$$

$$\begin{aligned} p(\vec{w}\vec{z}|\alpha, \beta) &= p(\vec{w}|\vec{z}, \beta)p(z|\alpha) \\ &= \prod_i^K \frac{B(\beta + n_k)}{\beta} \prod_i^M \frac{n_m + \alpha}{\alpha} \end{aligned}$$

w_k indicates that these words were generated by the kth topic.

3.2.2 Gibbs Sampling for LDA

BY the joint probability distribution $p(\vec{w}\vec{z}|\alpha, \beta)$ in the previous subsection, we can use Gibbs Sampling to sample it.The ith word in the corpus is denoted as z_i, where $i = (m, n)$ is a two-dimensional subscript, which corresponds to the nth word in the m document. According to the Gibbs Sampling algorithm in the second subsection, we need to require the conditional distribution corresponding to any coordinate axis. Assuming the observed word $w_i = t$, then by Bayes' rule, we can easily get:

$$\begin{aligned} p(z_i = k|\vec{z_{\neg_i}}, \vec{w}) &\propto p(z_i = k, w_i = t|\vec{z_{\neg_i}}, \vec{w_{\neg_i}}) \\ &= \int p(z_i = k, w_i = t, \theta, \alpha|\vec{z_{\neg_i}}, \vec{w_{\neg_i}}) d_\theta d_\alpha \\ &= \int p(z_i = k, \theta|\vec{z_{\neg_i}}, \vec{w_{\neg_i}}) p(w_i = t, \alpha|\vec{z_{\neg_i}}, \vec{w_{\neg_i}}) d_\theta d_\alpha \\ &= \end{aligned}$$

Finally, we get the Gibbs Sampling formula of the LDA model as:

$$p(z_i = k|\vec{z_{\neg_i}}, \vec{w}) \propto \frac{n^{(k)}_{m,\neg i} + \alpha_k}{\sum^K (n_{m,\neg i)} + \alpha_k)} * \frac{n^{(t)}_{k,\neg i} + \beta_t}{\sum^V (n_{m,\neg i)} + \beta_t)}$$

The equation on the right can be seen as

$$p(topic|doc) * p(word|topic)$$

We summarize the LDA Gibbs sampling algorithm process. The first is the training process:

(a) Choose the right number of topics , choose the right hyperparameter vector $\vec{\alpha}$ and $\vec{\beta}$.
(b) Corresponding to each word of each document in the corpus, randomly assign a topic number z.
(c) Rescan the corpus, for each word, use Gibbs sampling formula to update its topic

(d) number, and update the number of the word in the corpus. Repeat the Gibbs sampling based on coordinate axis rotation in step 3 until the Gibbs sampling converges.
(e) Calculate the theme of each word in each document in the corpus, get the document theme distribution θ_d, calculate the distribution of each theme in the corpus, get the LDA theme and word distribution β_k.

3.2.3 Perplexity and Inference

In information theory, perplexity is a measure of judging the probability model or probability distribution prediction, and can be used to evaluate the quality of the model.

$$perplexity(D) = exp(-\frac{\sum_{k=1}^{M} p(\vec{w_k})}{\sum_{k=1}^{M} N_k})$$

The denominator is the sum of all the words in the test set, that is, the total length of the test set. Where $p(\vec{w_k})$ refers to the probability of each word of document k in the test set.

$$\begin{aligned} p(\vec{w_k}) &= \prod_{i=1}^{V} p(w_k^{(i)}) \\ &= \prod_{i=1}^{V} \int p(w_k^{(i)}|z_i)p(z_i)d_{z_i} \end{aligned}$$

With the LDA model, for the new document doc, we only need to think that the topic-word matrix is stable and is provided by the model obtained from the training corpus, so we only need to estimate ttopic distribution of he document. The specific algorithm is as follows:

(a) For each word w_i in the current document, randomly initialize a topic number z.
(b) Use Gibbs Sampling formula to resample each word w_i.
(c) Repeat the above process until Gibbs Sampling convergence.
(d) Statistics on the topic distribution in the document, the distribution is θ_i.

3.3 Dirichlet belief networks

As can be seen from the above, the bright distribution of topics and word matrices obey Dirichlet distribution, which is an effective assumption, but it also brings many limitations. Recently,people introduce different generative models

3.3.1 Research on Topic Distribution

The LDA model assumes that the topics are independent of each other, however, this assumption It is very inconsistent with the actual data set. To overcome this defect, Blei In 2006, a related topic model (Correlated Topic Model, CTM)[**?**], the model extracts the topic from the Logistic Normal distribution,

overcoming the disadvantage that LDA model cannot extract relatation of information between texts.This model mentioned above have been successfully applied in extracting scientific subjects and image extraction.
Another well-known research on topic-distribution is Hierarchical Dirichlet processes[6]. Based on the deformation of Dirichlet Process, HDP is A nonparametric Bayesian model that can automatically train the most suitable from the document set Appropriate number of topics K. Nonparametric characteristics of HDP through Dirichlet process Solve the problem of selecting the number of topics in LDA, and the experiment confirms that The optimal number of topics selected by HDP is equal to the optimal number of topics selected based on perplexity . LDA model assumes that topic of each word is subject to multinomial distribution and the document is converted into count matrix by Bag of Word model.Poisson Factor [**?**] Analysis introduces poisson distribution.

$$x_p i = \sum_{k=1}^{K} x_{pik}$$

$$x_{pik} \sim Pois(\gamma_{pk}\theta_{ki})$$

and topic-distribution is subject to gamma distribution:

$$\theta_{ki} \sim Gamma(r_k, \frac{p_k}{1-p_k})$$

3.3.2 Introduction of DIRBN

Compared to considerable researchs topic model,The word model has not been fully studied and The Dirichlet Belief Network (DBN)[7] introduces a deep generative model where each layer is weighted by sets of topics.
Different from single layer dirichlet,DirBN model proposed a multi-layer Dirchlet layer generative process on word-distribution.The latent distri- butions in each layer of DBN are generated as Dirichlet random variables and can thus be interpreted as categorical distributions.Then the hidden units are connected with gamma-distributed weights.
In comparison to existing deep generative models,DirBN have better Interpretability on topics annd higher modelling accuracy.However,the current formulation of DBN odel suffers from decay during information passing.
DBN's deep architecture is currently limited to only a few layers. In order to obtain efficient Gibbs sam- pling, DBN back-propagates the observed information from the output layer to each hidden layer. The cost of the information back-propagation is that the information would decay in a O(log) rate on passing through one layer to its upper layer [8]. Therefore, little informa- tion might be available after a few layer back-propagations

Chapter 4

Conclusion

In mathematics, certain kinds of mistaken proof are often exhibited, and sometimes collected, as illustrations of a concept of mathematical fallacy. There is a distinction between a simple mistake and a mathematical fallacy in a proof: a mistake in a proof leads to an invalid proof just in the same way, but in the best-known examples of mathematical fallacies, there is some concealment in the presentation of the proof. For example, the reason validity fails may be a division by zero that is hidden by algebraic notation. There is a striking quality of the mathematical fallacy: as typically presented, it leads not only to an absurd result, but does so in a crafty or clever way. Therefore these fallacies, for pedagogic reasons, usually take the form of spurious proofs of obvious contradictions. Although the proofs are flawed, the errors, usually by design, are comparatively subtle, or designed to show that certain steps are conditional, and should not be applied in the cases that are the exceptions to the rules.

The traditional way of presenting a mathematical fallacy is to give an invalid step of deduction mixed in with valid steps, so that the meaning of fallacy is here slightly different from the logical fallacy. The latter applies normally to a form of argument that is not a genuine rule of logic, where the problematic mathematical step is typically a correct rule applied with a tacit wrong assumption. Beyond pedagogy, the resolution of a fallacy can lead to deeper insights into a subject (such as the introduction of Pasch's axiom of Euclidean geometry and the five color theorem of graph theory). Pseudaria, an ancient lost book of false proofs, is attributed to Euclid.

Mathematical fallacies exist in many branches of mathematics. In elementary algebra, typical examples may involve a step where division by zero is performed, where a root is incorrectly extracted or, more generally, where different values of a multiple valued function are equated. Well-known fallacies also exist in elementary Euclidean geometry and calculus.

References

[1] Blei, David M. "Probabilistic topic models." Communications of the ACM 55.4 (2012): 77-84.

[2] Zhao H, Du L, Buntine W, et al. Dirichlet belief networks for topic structure learning[C]//Advances in neural information processing systems. 2018: 7955-7966.

[3] Griffiths T L, Steyvers M. Finding scientific topics[J]. Proceedings of the National academy of Sciences, 2004, 101(suppl 1): 5228-5235.

[4] Blei D M, Lafferty J D. A correlated topic model of science[J]. The Annals of Applied Statistics, 2007, 1(1): 17-35.

[5] David M. Blei, AndrewY. Ng, Michael I. Jordan, LatentDirichlet Allocation, Journal of Machine Learning Research 3, p993-1022,2003

[6] Y.W.Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581, 2006.

[7] Zhao, H., Du, L., Buntine, W., and Zhou, M. Dirichlet belief networks for topic structure learning. In NeurIPS, pp. 7966–7977, 2018

[8] Zhou, M., Cong, Y., and Chen, B. Augmentable gamma belief networks. Journal of Machine Learning Research, 17(163):1–44, 2016.

[9] Barron, K., Kung, E. and Proserpio, D., 2018, June. The Sharing Economy and Housing Affordability: Evidence from Airbnb. In EC (p. 5).

[10] Benzinger, H., Berkson, E. and Gillespie, T.A., Spectral families of projections, semigroups, and differential operators, *Tran. Amer. Math. Soc.* **275** (1983), 431–475.