LAYERWISE INFERENCE FOR DIRICHLET BELIEF NETWORKS

Huiqiang Zhong

Supervisor: Dr. Xuhui Fan

School of Mathematics and Statistics

UNSW Sydney

August 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 random 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

With the development of technology, alomost all information is digitized and we are likely to be overwhelmed by a large amount of information. As a result 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 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 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 documents.

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[2]. We will introduce a well-known method in natural language processing called 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 above-mentioned theme.
- Repeat the above process until each word in the document is traversed.

This generative process is mainly compose of two distributions including 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 distribution of word and topic distribution should obey a prior distribution. The topic distribution and word distribution are subject to multinomial distributions,so the topic distribution and word distribution use Dirichlet distribution as their conjugate prior distribution because the multinomial 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) [3] ,which exhibits the correlation of topic by introducing logistic normal 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.[4]. In hierarchical dirichlet process[5],The author assume that the documents can be divided into set of groups ,and we can find the latent structure to decribe the data in the same group. However the

number of group is unkonwn and need to be inferred. HDP can automatically derive the number of the optimal topics, without specifying the number of groups.

On the other hand,The total counts of document is oftenly observed by bag of models and Poisson Factor Analysis model[6] proposes Beta-Gamma-Gamma-Poisson Model which can assign count of words into differenct topics.This kind of hierarchical structure is useful in sharing information between differnt topics.

Compare to numerous studies on topic distribution,research on word distribution has not received much attention. Zhao [4] introduces an alterative model named Dirichlet belief model to replace dirichlet prior distribution on word distribution. The output of DBN model is parameterised by Dirichlet-dsitributed hidden units which is connected with gamma weights.However,in order to obtain effieient Gibbs sampling,DirBN model back-propagates the observed information matric from output layer to upper layer by CRT[7] model which results in information decay and limination on network layer.

In this paper,we will introduce Bounded-Dirichlet Belief Network(UDBN). One of our key contribution is the introduction of of auxiliary Poisson random variable in the layerwise connection. These auxiliary variables are appropriately designed to faciliate the model inference.Besides,the application of auxiliary circumvent the complicated bottom-up back-propagation and enable full Gibbs Sampling on inference.

The paper is organised as follows. In Chapter 2 we will introduce basic knowledge of Bayesian Theorem such as prior distribution ,posterior distribution,common distribution used in this paper.Besides,MCMC and inference method including Gibbs Sampling also will be presented. In Chapter 3,we will discuss the Latent Dirichlet Allocation,including the genertive process of documents and the inference of LDA model.Beisdes, we will introduce the Dirichlet belief network,which is proposed to alternate the prior distribution of

word matrix.In the last chapter,we propose a new layerwise inference of Dirichlet Belief Network which introdece auxiliary Poisson random variable,further introducing the principle and the Gibbs samping process on this model.
we demonstrate the modelling advantage of out layerwise Dirichlet Belief Network in the applications of topic modelling and relational modelling.

Chapter 2

Preliminaries

In this chapter we outline the background of Latent Dirichlet Allocation and Inference method. We begin by reviseing 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

There was a hit-and-run accident in a certain city and there were only two-color cars in the city, 15 % blue and 85 % green. When the accident happen, a person saw it at the scene, and he identified it as a blue car. However, according to the analysis by experts, the probability that the evident was correct is 80 %. So, what is the probability that the car involved in the accident is a blue car?

Let B be the event where the car is blue in the city, G be the event where the car is green, and E be the event where the car is observed to be blue. Then, from the known information

$$P(B) = 0.15$$

$$P(G) = P(-B) = 0.85$$

If no witness sees the car of the perpetrator, then we can only guess blindly. Therefore, the probability that the car of the perpetrator is blue can only be the probability that the car is blue in the entire city, which is the prior probability $P(B) = 0.15$, because at this time we have no other evidence and can only make a rough estimate.

The witness said he saw the car and said it was blue. Note that there are two situations.

(a) The car is indeed blue, and the witness correctly distinguished the olor of car.

$$P(E, B) = P(E|B) * P(B) = 0.15 * 0.8 = 0.12$$

(b) The car is green , but the witness sees it as blue

$$P(E,\ B) = P(E|\ B) * P(\ B) = 0.85 * 0.2 = 0.17$$

What we are asking for is actually the probability that the car is blue under the condition of witnesses:

$$\frac{0.12}{0.12 + 0.17} = 0.41$$

Now the police found a new witness, and he also thinks the car that caused the accident is blue .So what is answer of above question?We have updated our original knowledge. We believe that the probability that the vehicle in the accident is blue is 0.41 rather than 0.15

$$\begin{aligned} P(B|E) &= \frac{P(B, E)}{P(E)} \\ &= \frac{P(E|B)P(B)}{P(E|B)P(B) + P(E|\ B)P(\ B)} \\ &= \frac{0.41 * 0.8}{0.41 * 0.8 + (1 - 0.41) * 0.2} \\ &= 0.735 \end{aligned}$$

Bayesian statistics considers that unknown models or parameters are uncertain and subkect to a certain probability distribution. In particular, we will first make a guess about this probability distribution based on subjective judgment or past experience, which is called the prior distribution.Then given more and more observations,we can modify the guess of the probability distribution and the final probability distribution is called the posterior distribution. As we obtain more and more data or evidence, our knowledge of reality has been updated and improved. Let us review the formule of Bayesian Thoerem:

$$\text{Posterior distribution} = \text{Likelihood info} + \text{Prior distribution}$$

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

- $P(A|B)$ is the conditional probability of A given event B, 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 prior 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.Take the coin toss example , before you toss it, you will judge that the probability of heads is 0.5. This is the so-called prior probability.In our example,we known that car of the city only have two color,and the pribability that the color is blue is 0.15.It is also regarded as priro distibuton.

A reasonable prior distribution is very useful for the estimation of unknown event. Judgments on many practical problems in life are related to people's knowledge, experience, and insights. In this case, if we combine the finite and observed data with the priors obtained from knowledge and experience, we will get better inferences about the unknown.

2.1.2 Likelihood Probability

We assume that the distribution of x is $f(x|\theta).x_1, x_2, \ldots, x_n$ are the samples from the observations,so that the joint distribution of samples is:

$$L(x_1, x_2, \ldots, x_n; \theta) \quad = \quad f(x_1|\theta)f(x_2|\theta)..f(x_n|\theta)$$

This formula can be regarded as a function of θ and L is a probability density function which is called linklihood function.$P(B)$ is a standardized constant. $\frac{P(B|A)}{P(B)}$ is called likelihood, which is an adjustment factor, that is, the adjustment of the occurrence of new information event B. The effect is to make the prior probability closer to the true probability.

- If the probability function $\frac{P(B|A)}{P(B)} > 1$, 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.

In our example,likelihood function is $P(E|B)$ and the standardized constant is $P(B)$.The likelihood is:

$$\begin{aligned} \frac{P(E|B)}{P(B)} &= \frac{P(E|B)}{P(B)} \\ &= \frac{P(E|B)}{P(E|B)P(B) + P(E|\ B)P(\ B)} \\ &= \frac{0.8}{0.15 * 0.8 + 0.85 * 0.2} = 2.79 \end{aligned}$$

The probabiity function is larger than 1,so that the probabilty that the color of car is blue becomes larger.

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

If the prior distribution and the likelihood function can make the prior distribution and the posterior distribution 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 $p(\theta|x)$ and the prior probability $p(\theta)$ of a random variable θ belong to the same distribution cluster, then $p(\theta|x)$ and $p(\theta)$ are called conjugate distributions, and also called $p(\theta)$ is the conjugate prior of the likelihood function $p(x|\theta)$.

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 a priori probability density distribution $Beta(\alpha,\beta)$, and 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:

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

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.So we can define the probility density function:

$$f(x_1, x_2, , x_n|\alpha_1, , \alpha_n) = \frac{1}{B(\alpha)} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$$

where the parameter $\alpha_1, \ldots, \alpha_n > 0$ and $\sum_{i=1}^{n} = 1$
Besides,the normalizing constant $B(\alpha)$ is the multivariate beta function:

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

For example,carry out K times of test, but This experiment will have n results and the probability of result is subject to a priori probability density distribution $Dirichlet(\alpha_1, , \alpha_N)$, The results of the experiment record the number of occurrences of each situation $[x_1, x_2, , x_n]$, the posterior probability density

distribution of the probability of the test is subject to $Dirichlet(\alpha_1+x_1,,\alpha_n+x_n)$.Prove:

Prior distribution:

$$f(p_1,p_2,,p_n|\alpha_1,\ldots,\alpha_n)=\frac{1}{B(\alpha)}\prod_{i=1}^{n}x_i^{\alpha_i-1}$$

likelihood Function:

$$f(x_1,x_2,,x_n|k,p_1,p_2,\ldots,p_n)=\frac{k!}{x_1!,,x_n}\prod_{i=1}^{n}p_i^{x_i}$$

Posterior distribution:

$$f(p_1,p_2,\ldots,p_n|k,x_1\ldots x_n,\alpha_1\ldots\alpha_n) = \frac{f(x_1,x_n|k,p_1\ldots p_n)*f(p_1p_n|\alpha_1\ldots\alpha_n)}{\int_x f(k,x_1\ldots x_n,\alpha_1,\ldots,\alpha_n)}$$

Given that:

$$\begin{aligned}\int_x f(k,x_1\ldots x_n,\alpha_1,\ldots,\alpha_n) &= \int_p f(x_1,x_n|k,p_1\ldots p_n)*f(p_1p_n|\alpha_1\ldots\alpha_n)\\ &= \frac{k!}{x_1!,,x_n}\frac{1}{B(\alpha)}\int_p\prod_{i=1}^{n}p_1^{\alpha_1+x_1}\\ &= \frac{k!}{x_1!,,x_n!}\frac{B(\alpha+x)}{B(\alpha)}\end{aligned}$$

$$\begin{aligned}f(x_1,x_n|k,p_1\ldots p_n)*f(p_1p_n|\alpha_1\ldots\alpha_n) &= \frac{k!}{x_1!,,x_n}\prod_{i=1}^{n}p_i^{x_i}\frac{1}{B(\alpha)}\prod_{i=1}^{n}p_i^{\alpha_i-1}\\ &= \frac{k!}{x_1!,,x_n!}\frac{1}{B(\alpha)}\prod_{i=1}^{n}p_i^{x_i+\alpha_i-1}\end{aligned}$$

So that:

$$\begin{aligned}f(p_1,p_2,\ldots,p_n|k,x_1\ldots x_n,\alpha_1\ldots\alpha_n) &= \frac{1}{B(\alpha+x)}\prod_{i=1}^{n}p_i^{x_i+\alpha_i-1}\\ &= Dirichlet(x+\alpha)\end{aligned}$$

2.2.3 Poisson and Multinomial distribution

The Poisson process is a continuous time random process with discrete values. Poisson process is used to describe the number of occurrences of events in a period of time, so it is also a counting process. Given the time-dependent random variable as N_t, which is used to describe the number of occurrences of an event from time 0 to time t.

In short,according to the average number of occurrences of a random event in a certain period of time or in a certain space in the past, poisson distribution can predict the probability that the random event will occur k times in the same long time or the same large space in the future. Its probability mass function is:

$$P(X = K|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

st.

- λ is the average number of random events in a certain period of time or in a certain space in the past.
- $k > 0$.

Multinomial Distribution is a generalization of Binomial Distribution. Binomial does n Bernoulli experiments, which stipulates that there are only two results for each experiment. However if we do n experiments now, there can be as many as m results for each experiment, and the probability of occurrence of m results exclusive and the sum of probability is 1, then the probability that one of the results happens X times subject to a multinomial distribution.

$$f(x_1, x_2, , x_n|k, p_1, p_2, \ldots, p_n) = \frac{k!}{x_1!, , x_n} \prod_{i=1}^{n} p_i^{x_i}$$

where $\sum_{i=1}^{n} = k$.

There is some transformation betweem poisson distribution and multinomial distribution[6].we assume that:

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

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

We introduce another equivalent augmentation[8] as :

$$x_{pi} \sim Pois(\sum_{k=1}^{K} \alpha_{pk}\theta_{ki})$$

$$\eta_{pik} = \frac{\alpha_{pk}\theta ki}{\sum_{k=1}^{K} \alpha_{pk}\theta_{ki}}$$

$$f(x_{pi1} \dots x_{piK}) \sim Mult(x_{pi}|\eta_{pi1} \dots \eta_{piK})$$

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 event through a large number of random samples, and then get the value to be calculated.

It is very powerful and flexible method,and quite easy to understand and 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 has a wide range of applications in the fields of machine learning, deep learning, and natural language processing. It is also 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 :

$$Y = \int_{a}^{b} f(x)$$

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: $x_1, \ldots, x_n$, 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 and bring it into the Monte Carlo summation formula to solve.

For the common uniform distribution $\mathcal{U}(0, 1)$,It is very easy to get samples, generally through the linear congruential generator that 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 sample conversion of uniform distribution.

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 $\mathcal{U}(0,1)$ 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:

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

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 [9] is to cover the smaller probability distribution with a larger probability distribution. This larger probability distribution $q(x)$ called proposal distribution is usually a standard distribution Such as uniform distribution, Gaussian distribution, which makes it easier to sample. Then we introduce a constant k which make $k * q(x) \leq p(x)$.

In each sample:

- Sampling a value x from proposal distribution $q(x)$
- Get a sample μ_0 from uniform distribution $[0, k * q(\mu_0)]$.
- If $\mu_0 < p(x)$,we retain the value otherwise we discard this value. The resulting data is an approximate sample of th 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

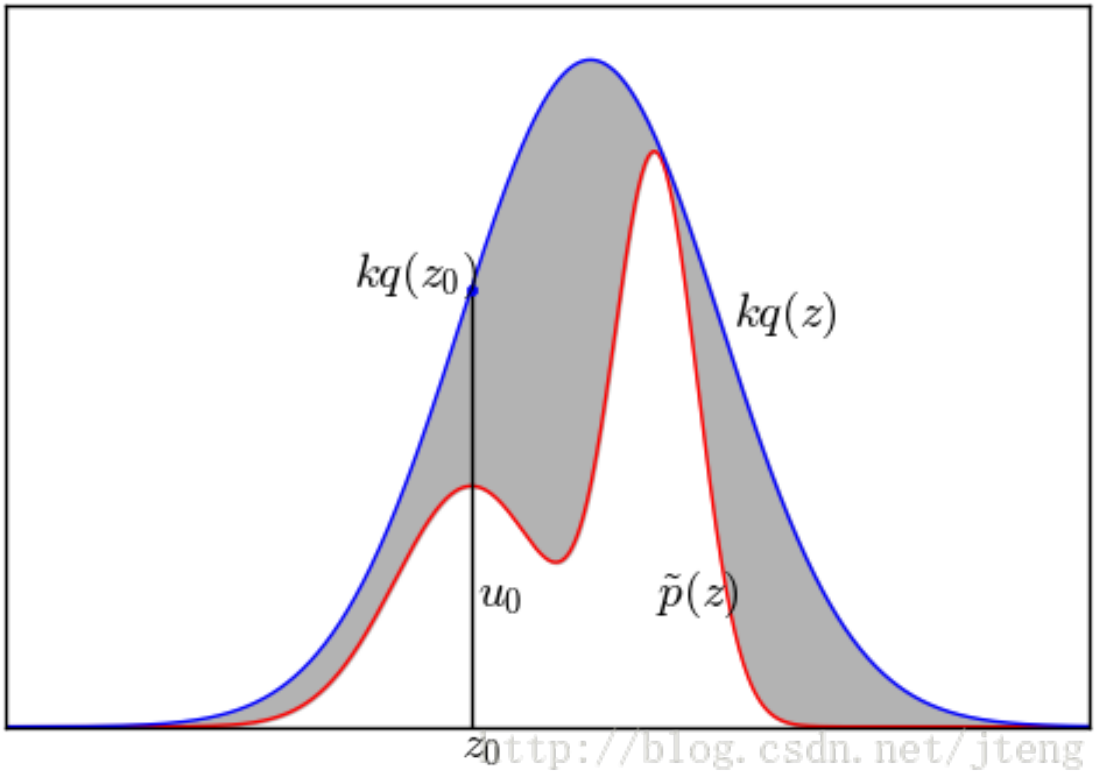

Figure 2.1: Reject-Sampling Method

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.

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 a powerful statistical models.

At each step of the Markov chain, the system can change from one state to another according to the probability distribution, or it can maintain the

current state. The change of state is called transition, and the probability associated with different state changes is called transition probability.

In order to obtain a theoretical result, let's look at a smaller example which will facilitate our calculation demonstration later. Assuming that in a region, people either live in the city or live in the countryside. The matrix below tells us the transition matrix of population migration. For example, the number of the first column and first row indicates that 90 % of the population currently living in cities will choose to live in cities next year and 80 % of people who live in country will continue to stay in country.

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

We assume that half of the people live in the city and the other half are in the countryside. As a simple start, We try to estimate the probability that people who live in the city will still live in the city after two 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% of people 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 80 % of those who moved to the village last year will choose to stay in the village. This analysis process is shown below. One year later:

$$H_x = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \times \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.55 & 0.45 \end{bmatrix}$$

One year later:

$$H_x = \begin{bmatrix} 0.55 & 0.45 \end{bmatrix} \times \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.585 & 0.415 \end{bmatrix}$$

In fact, you will find that our calculation process is to do the square of the matrix. As shown in the formula 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

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 and The sample set is the corresponding sample set that meets our stationary distribution.

If 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 and M-H Sampling

Mostly,target stationary distribution $\pi(x)$and a certain Markov chain state transition matrix Q does not satisfy the detailed balance 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 $[0 \sim 1]$, 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[10].
MCMC 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 n_1, the number of required samples n_2;

(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, the 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. Metropolis-Hastings sampling also called M-H sampling can solve the problem of low sampling acceptance rate in the previous section.
We can expand both sides of

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

At this time the detailed stationary condition is also satisfied. We expand the equation by C times to make $c * \alpha(i,j)$ (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:

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

Compare to basic MCMC method,Metropolic Hasting sampling greatly improves the efficiency of sampling.However, in the era of big data, M-H sampling still faces two major challenges:

(a) Our data features are very many, due to the existence of the acceptance rate, the calculation time of M-H sampling required in high dimensions is very considerable, and the algorithm efficiency is very low. At the same time, $\alpha$is generally less than 1. 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 [11] is another 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) [12]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 Preliminaries,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.

$$p(w_1)p(w_2)..p(w_n) = \prod_{k=1}^{V} p_k^{n_k}$$

The disadvantage of the method of the unigram model is that there are no relationship between the word in the document and it is hard to inference the distriution of word in defferent documents with different topic. The mixture of unigram[13] method improves it by introducing topic for each document. The model samples a topic from distribution $p(z)$ before generating each word.

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

z represents a topic, $p(z)$ 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. pLSA is a topic model, which is a method of modeling the hidden topics in the text. PLSA means that after the document d is given, the topic z corresponding to the document

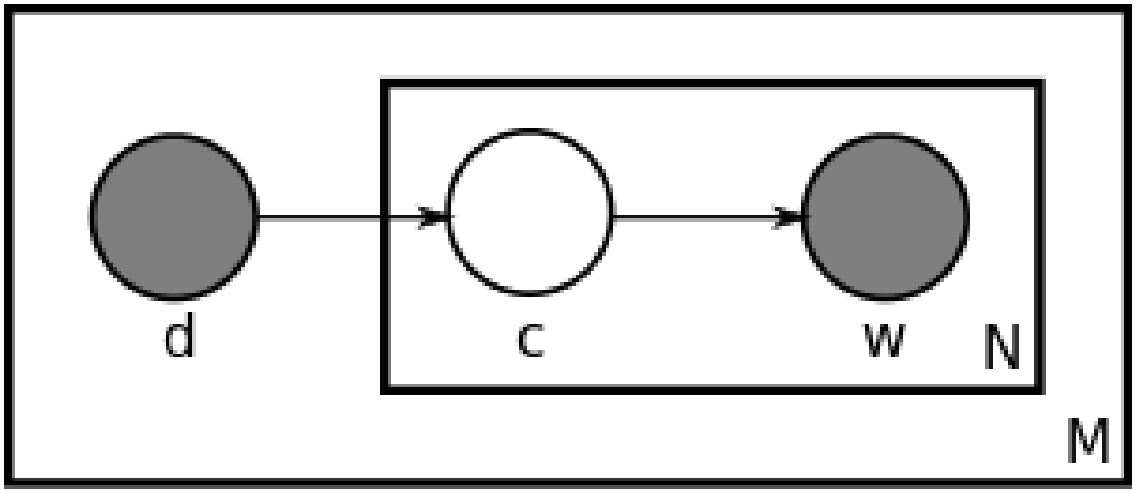

Figure 3.1: pLSA model

needs to be selected with a certain probability, and then the word w is selected from the topic z with a certain probability[14].

PLSA models the joint distribution of d and w by the following formula:

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

(a) Choose a document with probability p;

(b) Get the topic z by with probability $p(z|d)$;

(c) Generate a word with probability $p(w_n|z_k)$;

Imagine that we wants to write N documents, and we needs to determine the word in each position in each document. Suppose we has K optional topics and V optional words. Therefore, we made K dice with V sides, each corresponding to a topic, and each side of the dice corresponds to The selected term. Then, each time a document is written, a K-sided document-topic dice will be made; each time a word is written, the dice will be thrown to select the topic; after the result of the theme is obtained,we can get words by tossing the corresponding topic-word dice. Repeating this method N times, then finish writing all documents.

It is easy to find that for a new document, we cannot know what its corresponding $P(d)$. 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 topic with a fixed probability and then find the corresponding word distribution according to the extracted topic and then extract a vocabulary according to Word distribution. 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 adopts the Bayesian idea that they should obey a prior distribution. The topic distribution and word distribution are both multinomial distributions, because the multinomial 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:

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 multinomial distribution of topics θ_i.

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

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

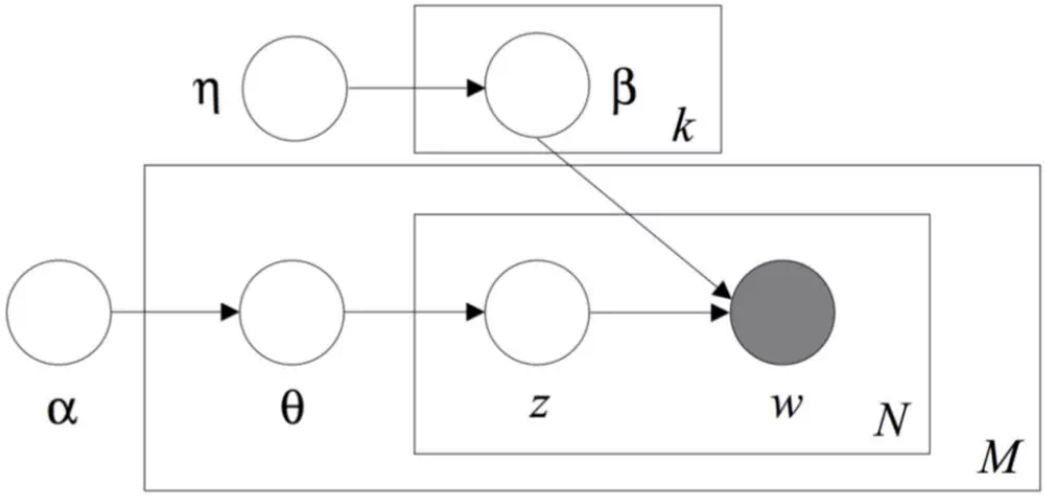

Figure 3.2: The caption of this figure.

This probability map can be decomposed into two parts:

(a)

$$\alpha \rightarrow \theta \rightarrow z$$

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)

$$\eta \rightarrow \beta \rightarrow w|k$$

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[18],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[19] 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 prior distribution of topics and word matrices should obey Dirichlet distribution, which is an effective assumption, but it also brings many limitations. Recently,people introduce different generative models based on Latent Dirichle Model

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, in 2006 Blei proposes a related topic model called Correlated Topic Model, CTM[3], the model extracts the topic from the Logistic Normal distribution, succeddfully overcoming the disadvantage that LDA model cannot extract relatation of information between documents. 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[15]. 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 model, and the experiment confirms that the optimal number of topics selected by HDP model 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 [6] introduces poisson distribution to generate the words of document. The total count of word will be assign different latent topic.

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

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

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)[16] 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 distributions 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.

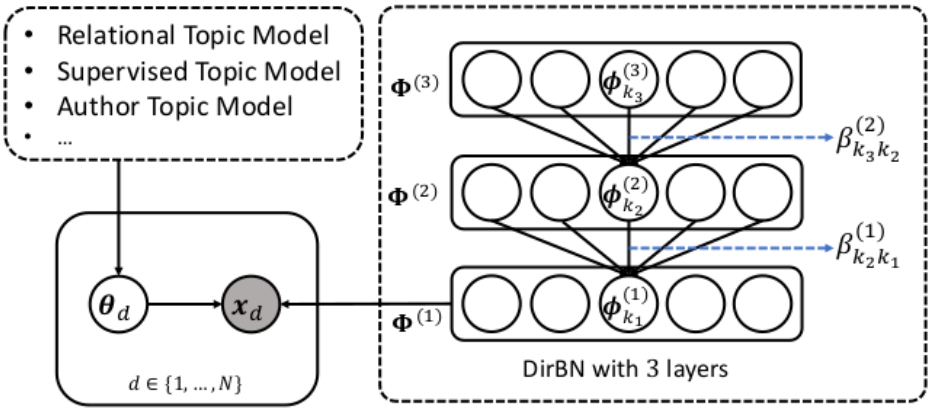

Figure 3.3: DirBN Model

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 [17]. Therefore, little informa- tion might be available after a few layer back-propagations

3.3.3 Inference of DirBN

The training of DirBN model introduce several data augmentation techniques in the inference of latent variables. Given topic-doc matrix θ and topic-word

matrix ϕwe can sample the topic of each word ini each document.Then we get the topic-word count matrix $x_{k_1}^{(1)} = [x_{1k_1}^{(1)}, \ldots, x_{Vk_1}^{(1)}]$ which is also the input count vector of DirBN inference process. The process of Inference involve two key parts:

(a) Propagating the input count vector from Bottom to top.Due to the incon-jugate property between these latent distri-butions, direct efficient Gibbs sampling over these random variables is difficult. Insteadchoose to first back propagates the observed counting information into each layer and then proceed forward variable Gibbs sampling.

(b) Updating latent parameters from top to bottom.After back propagata-tion of latent count,we can update the variables by conjugate posterior distribution.

Chapter 4

Bound Dirichlet Belief Model

4.1 Introduction

In this work, we propse Bound Dirichlet Belief Network to address the issue of DirBN model and enlarge the modelling capability. By inserting auxiliary Poisson random variables into the layerwise connections and appropraite design.we assume that there are N objects to be modelled in this word,where $\pi_{i'}$ is used to denote the i'-th latent distribution in the l-th layer.There are L hidden layers in the deep architure.

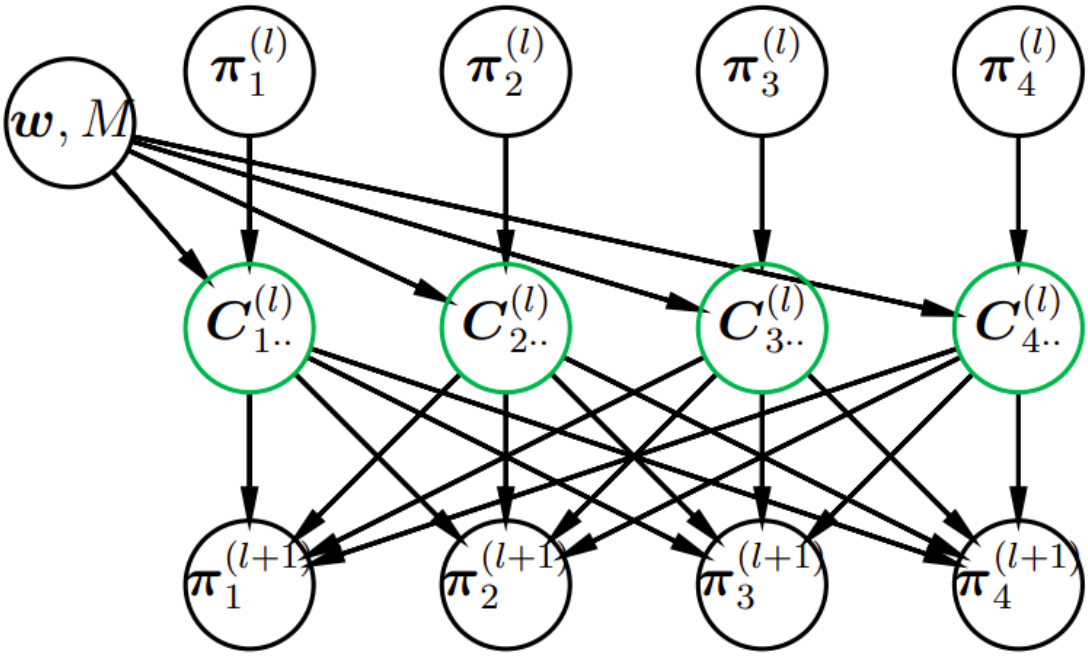

Figure 4.1: UBDN Model

$$\pi^1_{i'.} \sim Dirichlet(\beta)$$

$$C^1_{i'..} \sim Multi(M^l; \pi^1_{i'.} \otimes w^l_{i'.})$$

$$\pi^2_i \sim Dirihlet(\alpha_d + \sum_{i'} C_{i'.i})$$

$$\dots$$

$$C^l_{i'..} \sim Multi(M^l; \pi^l_{i'.} \otimes w^l_{i'.})$$

$$\pi^{l+1}_i \sim Dirihlet(\alpha_d + \sum_{i'} C^l_{i'.i})$$

- Information pass from Topic i' to Topic i;
- $\pi^l_{i'.}$ is the Topic-Word matrix in layer l;
- All layer has same Topic number;
- $w^l_{i'i}$ is the probability that information pass from topic i' to i;
- This is the reconstruction process;;
- $\sum_{i'} C_{i'.i}$ is a (V*K) by 1 Vector;

where:

$$M^{(l)} \sim Pisson(M)$$

$$w^l_{i'.} \sim Dirichlet(\eta)$$

Since $\sum_k w^l_{i'k} = 1$,we can further get :

$$\sum_{i'} C_{i'.i} \sim Multinomial(M^l; \pi^l_{i'})$$

That is to say, the normalized proportion vector can be ragard as an approximator of $\pi^l_{i'}$. We reorganiza the counting variables and calculate the recevied counts information for rach node i. For any node i, it will receive latent counts from all the nodes in this layer. These received counts can then be used to consistute the concentration parameter vector of the receiving node $i's$ Dirichlet distribution.

The number M^l of events records the total intensity of relating the latent representation to the related counts. Larger values of M^l will remind a closer

approximation to π^l. Further, the introduction of Poisson counts help to decompose the additive effect of the weights, which is quite difficult to have closed Gibbs sampling format.

This key contribution needs to be emphasized again. The introduction of the counts variable C enables us to use a layerwise sampling method for all the variables and thus avoid the complicated strategy of backward counts propagation and forward variable sampling. $C^l_{i'}$plays the role of "likelihood" variable for each latent distribution and we can thus easily obtain Gibbs sampling for $\pi^{(l)}_{i'}$. Based on Poisson-Multinomial Equivalence,we can get:

$$C^l_{i'ki} \sim Poisson(M^l \pi^l_{i'k} w^l_{i'i})$$

Combining its likelihood, we can easily obtain all its potential posterior proportional values.

4.2 Inference Process

Unlike DirBN model which inference latent vairiables by propagating count matrix with Chinese Restaurant Table (CRT) distribution.In DBN model,we can directly use Gibbs sampling stratrgy for model inference.

(a) Input matrix is topic-word matrix z which is a K by V matrix and then sample $\pi^L_{i'}$:

$$\pi^L_{i'} \sim Dirichlet(z + C^L_{i'.i})$$

(b) Sampling $C^l_{i'ki}$.

$$P(C^l_{i'ki}|.) \sim \frac{(M^l \pi^l_{i'k} w^l_{i'i})^{C_{i'ki}}}{C^l_{i'ki}!}$$

$$P(\pi^{l+1}_i | C^l_{i'ki}, ..) = \frac{1}{B(\alpha_d + \sum_{i'} C_{i'.i})} \prod^V \pi^{\alpha_d + C^l_{i'ki} - 1}_{ik}$$

The posterior distribution of $C^l_{i'ki}$ is:

$$P(C^l_{i'ki}|.) \propto \frac{(M^l \pi^l_{i'k} w^l_{i'i})^{C_{i'ki}}}{C_{i'ki}!} . \frac{1}{B(\alpha_d + \sum_{i'} C^l_{i'.i})} \prod^{V} \pi_{ik}^{\alpha_d + C^l_{i'ki} - 1}$$

and

$$\sum_{k,i} C^l_{i'ki} = M^l$$

then

$$C^l_{i'ki} \sim Mult(M^l | \frac{C^l_{i'..}}{\sum_{k,i} C^l_{i'ki}})$$

(c) Sampling $\pi^l_{i'}$,its posterior distribution is Dirichlet distribution.Given that :

$$\pi^l_{i'..} \sim Dirihlet(\alpha_d + \sum_{i'} C^{(l-1)}_{i'.i})$$

$$\pi^1_{i'.} \sim Dirichlet(\beta)$$

$$\sum_{i} w_{i'i} = 1$$

then :

$$\sum_{i} C_{i'ki} \sim Multi(M^l; d\pi^l_{i'.})$$

$$\pi^l_{i'.} \sim Dirichlet(\eta + \sum_{i'} C^{l-1}_{i'.i} + \sum_{i} C^l_{i'ki})$$

$$\pi^1_{i'.} \sim Dirichlet(\beta + \sum_{i} C_{i'ki})$$

(d) Sampling $w_{i'}$.Given that:

$$w^l_{i'} \sim Dirichlet(\eta)$$

* Likelihood

$$\sum_{k} C_{i'ki} \sim Multi(M^l; w^l_{i'.})$$

* Prior distribution

$$W_{i'i} \sim Dirichlet(\alpha_w)$$

* Posterior distribution

$$p(w_{i'}|..) \sim Dirichlet(\alpha_w + w_{i'.})$$

4.3 Experiment

The experiment was conducted on a real-word datsets called Tag My News (TMN) [20] which contains 32597 RSS news labelled with 7 categories and there are 13370 unique words in the documents.

For UBD model,we set $M = 200, \eta = \beta = 1$. We use 50 % of data as our training data and make another other half of data as test data. We use perplexity as the evaluation criterion of the model. Obviously as the number of training increases, the value of Perplexity continues to decrease.

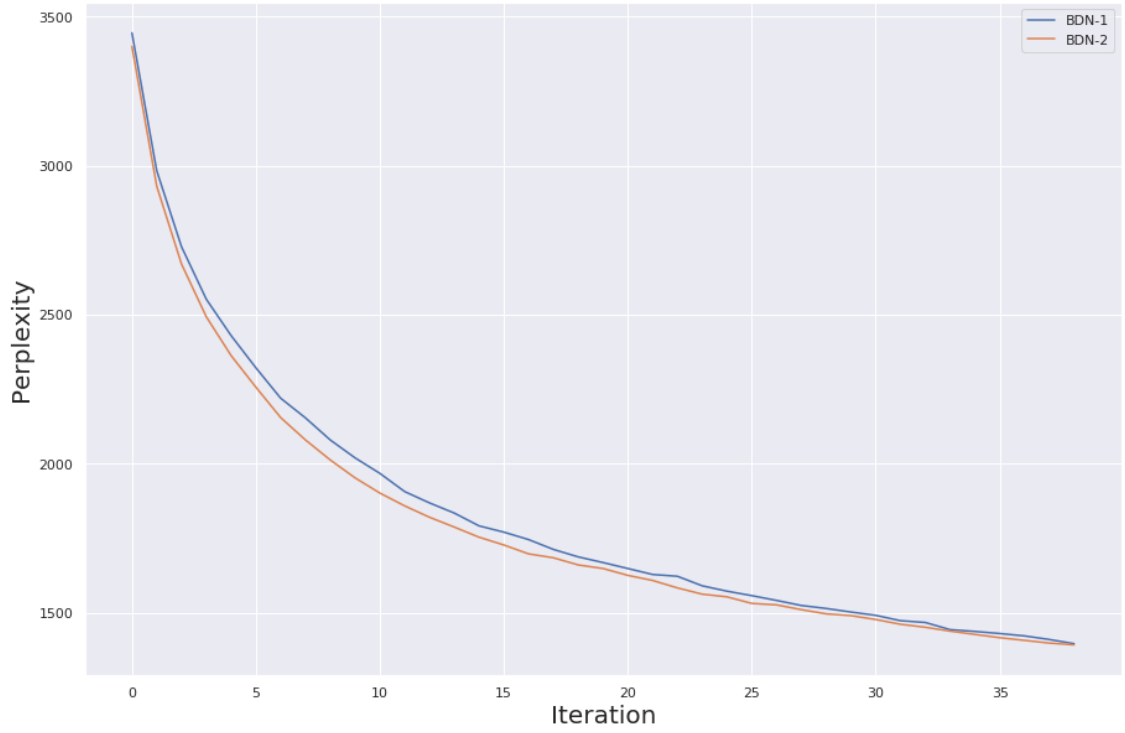

Figure 4.2: Result of experiment

Chapter 5

Conclusion

In this paper, we first introduce the background of Latent Dirichlet Allocation such as bayesian theorem,basic distriubution and the method used in inference process including Gibbs sampling and Metropoli-Hasting sampling.Then we review the generative process of document,inferencing the latent variables in LDA models by Gibbs sampling.

For improving the performance of LDA model,Scientists have done a lot of research on word distribution and topic distribution. However,compared with lots of research on topic structure, word distribution attracts less interest. Zhao[16] proposed Dirichlet belief model to replace dirichlet prior distribution on word distribution,but it is hard for inference latent parameters with complex structure.

We have present a new layerwise inference of Dirichlet Belief Model. With each topic in different layers, DBN model is able to deepen the understanding on topic hierarchies. By inserting auxiliary poisson random variable into the layerwise connections and appropriate design,direct efficient Gibbs sampling over random variables is available.

However,the limitation of this model is that the lengths of latent distrubutions in all the layers are restricted to be the same and need to be fixed.Further directions include introduction of Chinese Restaurant Process on the prior of the length of latent distribution.

References

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

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

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

[4] Zhao, He, et al. "Dirichlet belief networks for topic structure learning." Advances in neural information processing systems. 2018.

[5] Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In Advances in neural information processing systems (pp. 1385-1392).

[6] Zhou, Mingyuan, et al. "Beta-negative binomial process and Poisson factor analysis." Artificial Intelligence and Statistics. 2012.

[7] Griffiths, Thomas L., et al. "Hierarchical topic models and the nested chinese restaurant process." Advances in neural information processing systems. 2004.

[8] W. Buntine and A. Jakulin. Discrete component anal- ysis. In Subspace, Latent Structure and Feature Se- lection Techniques. Springer-Verlag, 2006.

[9] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

[10] Chib, Siddhartha, and Edward Greenberg. "Understanding the metropolis-hastings algorithm." The american statistician 49.4 (1995): 327-335.

[11] Gelman, Andrew, et al. Bayesian data analysis. CRC press, 2013.

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

[13] Nigam, Kamal, et al. "Text classification from labeled and unlabeled documents using EM." Machine learning 39.2-3 (2000): 103-134.

[14] Hofmann, Thomas. "Probabilistic latent semantic analysis." arXiv preprint arXiv:1301.6705 (2013).

[15] 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.

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

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

[18] Liu, Jun S. "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem." Journal of the American Statistical Association 89.427 (1994): 958-966.

[19] Wallach, Hanna M., et al. "Evaluation methods for topic models." Proceedings of the 26th annual international conference on machine learning. 2009.

[20] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics, 3:299–313, 2015.