Bayesian Parameter Estimation in LDA

Z.Y. Liu
School of Software, Beijing Institute of Technology
Institute of Information Engineering
Chinese Academy of Sciences
China

W.P. Wang
Institute of Information Engineering
Chinese Academy of Sciences
China

Y. Wang, W.Y. Lu
School of Software
Beijing Institute of Technology
China

Z.Z. Ji
School of Computer Science and Technology
Beijing Institute of Technology
China

***Abstract*--Latent Dirichlet Allocation (LDA) probabilistic topic model is widely used in text mining, natural language processing and so on. But LDA's mathematical theory is particularly complex, thus it is very difficult to understand LDA for a novice. In order to more quickly and easily learn LDA, and further promote its application, this paper will deeply analyze LDA from the perspective of Bayesian parameter estimation. At first we explain the advantage of Bayesian parameter estimation by an instance, and then introduce a simple Bayesian Unigram model. Next based on the simple Bayesian Unigram model and PLSA model, a full Bayesian probabilistic topic model—LDA is presented.**

Keywords-bayesian; parameter estimation; LDA; topic model

I INTRODUCTION

In text mining, natural language processing and so on, probabilistic topic models are receiving extensive attention. The probabilistic topic model can understand text from semantic level, while the traditional Vector Space Model (VSM) [1] only represents text based on the "words" or "terms" appeared in the documents. Under VSM, feature space is established according to the observed terms without any semantic consideration. As a result, this will obviously lead to deviation of the meaning of text. To address the shortcoming, probabilistic topic models have be proposed, which extract semantic topics using co-occurrence of terms in documents, and are used to transform documents locating in term space to the ones in topic space, and then enable us to obtain the understanding of the documents on a meaning level rather than by lexical congruence.

The origin of probabilistic topic model is Probabilistic Latent Semantic Analysis (PLSA) [2], PLSA model based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. The parameters of PLSA model are estimated by Maximum Likelihood Estimation (MLE) method [3].

But Bayesian scholars think that the unknown parameter should be the random variable, and have a prior distribution; therefore a new probabilistic topic model named LDA [4] is proposed. The parameters of LDA model have the prior distribution, and are estimated by Bayesian method. LDA model has attracted many scholars' attention since its start, but its mathematical theory is too complex to understand quickly.

In order to more quickly understand LDA model, and further promote its application, this paper will start from Bayesian parameter estimation, and then analyze the mathematical theory of LDA model from the perspective of Bayesian parameter estimation. This paper is organized as follows. In Section 2 we introduce Bayesian parameter estimation based on an illustrative example. Compared to related PLSA model, LDA model with Bayesian parameter estimation is presented in Section 3. Finally, Section 4 presents summary.

II BAYESIAN PARAMETER ESTIMATION

Bayesian parameter estimation is an important method for parameter estimation. The key difference between the Bayesian parameter estimation method and the classical parameter estimation method is whether prior beliefs are adopted. Bayesian parameter estimation method attach importance to the overall information and sample information, also note that prior information collection and collation, and make it quantitative, form the prior distribution, participate in statistical inference, so that improve the quality of statistical inference[5].

To illustrate the superiority of Bayesian parameter estimation method, an example which is a shooter experiment [6] is as follows:

The shooter experiment: One player shot n times, hit r times, and then how to estimate the player's hit rate according to his result?

Usually to estimate hit rate θ with $\frac{r}{n}$, but this approach is flawed. For example: when n = r = 1, estimate $\hat{\theta}$ =1, but when n = r = 100, also estimate $\hat{\theta}$ = 1; and when n = 100, r = 0, estimate $\hat{\theta} = 0$, but when n = 1, r = 0, also estimate $\hat{\theta} = 0$.

If one player shot 100 times, and hit 100 times, then we intuitively felt his shot level was very high. However the other one shot 1 time, hit 1 one time, and we cannot infer that their shot level is same, but estimating θ with $\frac{r}{n}$, the same estimation result is obtained.

According to probability theory, if one player's hit rate is θ, when he shot n times, hit x times, then the probability of this event is shown as follows:

$$P(X = x|\theta) = C_n^x \theta^x (1-\theta)^{n-x} \quad (1)$$

Where θ is a random variable, the probability in eqn (1) is the conditional probability of X given θ.

To estimate θ by the sample X, according to Bayesian theorem the conditional probability density of θ given X can be written as:

$$\pi(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{m(X)} \quad (2)$$

Where π(θ|X) is a combination of a prior belief that the probability distribution function is π(θ) and observations X with the likelihood p(X|θ). m(X) is the marginal density function of X, which is not relevant to θ. And this conditional probability π(θ|X) is named as posterior probability. The posterior probability is used to estimate the parameter θ given a sample X.

And if we do not know about the shooter, the assignment of his hit rate θ in [0,1] has the same possibility, then the prior distribution π(θ) follows Uniform distribution in [0,1] (Bayesian assumption), that is,

$$\pi(\theta) = \begin{cases} 1, & \theta \in [0,1] \\ 0, & \theta \overline{\in} [0,1] \end{cases} \quad (3)$$

Thus, π(θ|X) can be expressed as:

$$\pi(\theta|X = x) = \frac{\theta^x (1-\theta)^{n-x}}{\int_0^1 \theta^x (1-\theta)^{n-x} d\theta}, \quad 0 \le \theta \le 1 \quad (4)$$

And eqn(4) indicates that the posterior distribution π(θ|X) follows Beta distribution.

Then the parameter θ will be estimated by the expectation of this posterior distribution π (θ|X), that is,

$$\hat{\theta} = E\{\theta|X = x\} = \frac{1}{B(x+1, n-x+1)} \int_0^1 \theta \cdot \theta^x (1-\theta)^{n-x} d\theta$$
$$= \frac{B(x+2, n-x+1)}{B(x+1, n-x+1)} = \frac{x+1}{n+2} \quad (5)$$

Back to the example of the shooter, when n = r = 1, $\hat{\theta} = \frac{1+1}{1+2} = \frac{2}{3}$, and when n = r = 100, $\hat{\theta} = \frac{100+1}{100+2} = \frac{101}{102}$. Obviously this estimate is more reasonable than the above $\frac{r}{n}$.

Summarize the core idea of Bayesian parameter estimation method:

Regard the unknown-parameter θ as a random variable. When θ is known, the joint distribution density p(x1,x2,…,xn; θ) of sample x1,x2,…,xn is regarded as the conditional probability density of x1,x2,…,xn given θ, denoted by p(x1,x2,…,xn|θ) or p(X|θ).

Try to determine a prior distribution π(θ) (based on a prior belief of the parameter θ).

Combine the conditional distribution density p(x1, x2,…,xn|θ) and prior distribution π(θ), posterior distribution density π(θ|x1,x2,…,xn) can be obtained according to Bayesian theorem.

Use the posterior distribution density π(θ|x1, x2,…,xn) to estimate the parameter θ.

For the prior distribution π(θ), if following the conjugate distribution of p (X|θ), Bayesian parameter estimation of the above shooter example is as follows:

In the example, one player shot n times, hit x times, this event follows Binomial distribution, and the conjugation distribution of Binomial distribution is the Beta distribution, denoted by B (a,b). Assign this Beta distribution to the prior distribution π(θ), then π (θ|X) can be expressed as:

$$\pi(\theta|X = x) = \frac{1}{B(x+a, n-x+b)} \theta^{x+a-1} (1-\theta)^{n-x+b-1}, \quad 0 < \theta < 1 \quad (6)$$

Then the parameter θ will be estimated by the expectation of this posterior distribution π (θ|X), that is,

$$\hat{\theta} = E\{\theta|X = x\} = \frac{x+a}{n+a+b} \quad (7)$$

This estimation result shows that using B (a, b) as a prior distribution, i.e. shot (a+b) times before this shooting, hit a times, now shot n times, hit x times, thus, in total shot (n + a + b), hit (x + a) times.

When a = b = 1, B (a, b) is a Uniform distribution in [0, 1], so this method is consistent with the above Bayesian Assumption: prior distribution follows Uniform distribution when we do not know about the shooter.

In this shooter experiment, assuming the results of each shot are not "hit" or "un-hit", but multiple rings, that is, the target surface is divided into multiple rings (k rings), every shot can hit a target ring, and the hit probability of each ring is different, denoted by θ = (θ1,…,θk). The result X = (x1,…,xk) of n-shooting follows Multinomial distribution Mult(θ), and the conjugate distribution of Multinomial distribution is Dirichlet distribution[7], so assign Dirichlet distribution to prior distribution π(θ), and using the conditional distribution p(X|θ) and the prior distribution π(θ), the posterior distribution π(θ|X) can be obtained according to Bayesian theorem. Then the expectation of the posterior distribution π(θ|X) is used to estimate θ.

Process similar to the above shooter experiment can be used in text modeling. While text modeling, let us assume that there are k words in a vocabulary, and the word probabilities are denoted by θ = (θ1,…,θk) . For a document which contains N words, the N observed words is a multivariate discrete random variable that follows Multinomial distribution with parameter θ. And the conjugate prior distribution of θ is Dirichlet distribution.

This process of text modeling can be described with Bayesian Unigram Model, and its graphical model is depicted in Figure 1.

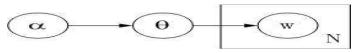

FIGURE I: BAYESIAN UNIGRAM MODEL.

Where θ refers to the probabilities of the words in a vocabulary, w refers to the observed word in a document which contains N words，α is the parameter of the prior Dirichlet distribution which expresses a prior belief.

Under this Bayesian Unigram Model, the words of every document are drawn independently from a Dirichlet distribution.

III BAYESIAN PARAMETER ESTIMATION IN LDA

Bayesian Unigram Model in section 2 is too simple to be used to text modeling. Because documents are mixtures of topics, and a topic is a "bag-of-words". Each document should be generated by first choosing a topic and then generating words based on this topic. This idea of text modeling can be described with Mixture of Unigrams Model [8], and its graphical model is depicted in Figure 2.

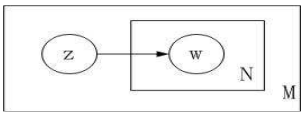

FIGURE II: MIXTURE OF UNIGRAM MODEL.

Where z refers to a topic, w refers to an observed word in the document, M is the size of a corpus, and N is the length of a document.

Probabilistic latent semantic analysis (PLSA) proposed by Hoffman in 1999 is a mathematical topic model. And it is regarded as the actual origin of topic model. The graphical model of PLSA Model is depicted in Figure 3.

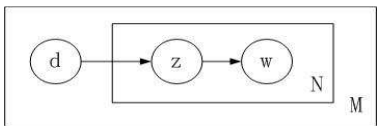

FIGURE III: GRAPHICAL MODEL REPRESENTATION OF PLSA.

Where d refers to a document index, z refers to a latent topic, w refers to an observed word in the document, M is the size of a corpus, and N is the length of a document.

PLSA model assumes the following generative process for each document in a corpus:

Choose a document index d~p(d);

For each each of the N words in the document d

Choose a latent topic z~p(z|d);

Choose a word w~p(w|z)

According the generative process of PLSA model we can find the two key parameters which are p(z|d) and p(w|z). The p(z|d) indicates the probability distribution over topics for each document, and the p(w|z) indicates the probability distribution over words associated with each topic. In PLSA model, the two parameters are estimated by MLE (Maximum Likelihood Estimation) method.

But for the two parameters: p (z|d) and p(w|z) in PLSA, Bayesian scholars have their different opinion, that is, the two parameters are random variables, so they should have their prior distributions. Further the PLSA model is extended based Bayesian opinion. In the extended PLSA model, p(z|d) and p(w|z) have their conjugate prior distributions —Dirichlet distribution. This new extended model is LDA model [9].

Under LDA model, p(z|d) and p(w|z) are estimated by Bayesian estimation method. That is, using the conditional distribution and the prior distribution of parameters, the posterior distribution can be obtained according to Bayesian theorem, then the expectation of the posterior distribution is used to estimate parameters.

The graphical model of LDA Model is depicted in Figure 4.

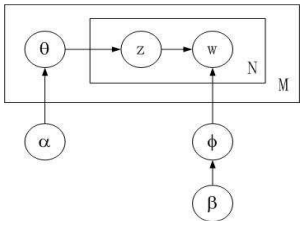

FIGURE IV: GRAPHICAL MODEL REPRESENTATION OF LDA.

Where θ refers to document-topic distribution, ϕ refers to topic-word distribution, α and β are the parameters for the two Dirichlet distributions which express the prior beliefs, M is the size of a corpus, N is the length of a document, z refers to a latent topic, w refers to the observed word in a document which contains N words.

The LDA graphical model can be divided into two key physics process:

1) α→θ→z generates the corresponding topics of all words in a document. α→θ corresponds to Dirichlet distribution, and θ→z Multinomial distribution, so this process is a Dirichlet-Multinomial conjugate structure.

2) β→ϕ→w generates every word of a document based on a special topic. β→ϕ corresponds to Dirichlet distribution, and ϕ→w Multinomial distribution, so this process is also a Dirichlet-Multinomial conjugate structure.

The two Dirichlet-Multinomial conjugate structures cooperate with each other to to generate the document. The generative process for each document is as follows:

1. Choose ϕ_k~ Dirichlet (β), k∈[1, K]
2. For all documents m∈[1, M]
   a) Choose θ_m~ Dirichlet (α)
   b) Choose N_m~ Poiss (ξ)
   c) For the nth word in document m, n∈[1, N_m]
      i. Choose a topic $z_{m,n}$~ Multinomial (θ_m)
      ii. Choose a word $w_{m,n}$~ Multinomial ($\phi z_{m,n}$)

That is, to make a new document, at first LDA chooses a distribution over topics, then for each word in the document, chooses a topic at random according to this distribution, and draws a word from that topic.

IV SUMMARY

LDA probabilistic topic model is becoming more and more useful in many applications, such as text mining, natural

language processing and so on. So it is very necessary to deeply learn LDA. This paper starts from Bayesian parameter estimation, introduces the related mathematical knowledge of LDA which are prior distribution, posterior distribution, Bayesian estimation, Multinomial distribution, Dirichlet distribution, conjugate distribution, etc., and analyzes the limitations of the simple Bayesian Unigram model and PLSA model for text modeling. Finally, a full Bayesian probabilistic topic model — LDA is presented based on these analyses. Through this paper you can know that LDA is an excellent document model based on Bayesian theorem.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (61272361), also supported by Key Project of National Defense Basic Research Program of China (B11201320), National HeGaoJi Key Project (2013ZX01039 - 002-001-001), National High-Tech Research and Development Program of China (2012AA011002).

REFERENCES

[1] G.Salton, A.Wang, C.S. Yang. A Vector Space Model for Automatic Indexing. Communication of the ACM, 18(11), pp. 613-620, 1975.

[2] T.Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning Research, 42(1), pp. 177-196, 2001.

[3] G.Heinrich. Parmeter Estimation for Text Analysis. http://www.arbylon.net/ publications/text -est.pdf.

[4] D.Blei, A.Ng, M.Jordan. Latent Dirichlet Allocation. Machine Learning Research, 3(3), pp. 993-1022, 2003.

[5] S.Mao, X.Lv. Mathematical statistics, China Renmin University Press, 2011.

[6] Y.Zhang, H.Chen. Bayesian Statistical Inference. Science Press, 1994.

[7] L.Wei, W.Zhang. Bayesian Analysis, China Science & Technology University Press, 2013.

[8] K. Nigam, A. McCallum, S. Thrun, T. Mitchell. Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2/3), pp.103–134, 2000.

[9] M.Steyvers, T.Griffths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2006.