Construction of Dirichlet Mixture Allocation total probability model based on multiple class text analysis

Xiaoming Ling$^1$ and Feng Zhang$^2$
$^1$ National Green Coating Equipment and Technology Research Centre, Lanzhou Jiaotong University, Lanzhou, Gansu, 730000, PR China
$^2$ National Green Coating Equipment and Technology Research Centre, Lanzhou Jiaotong University, Lanzhou, Gansu, 730000, PR China

*Corresponding author’s e-mail: lingxm24@lzjtu.edu.cn

Abstract. The LDA model is a total probability generation model for analyzing a large number of documents. It extends PLSA, another text analysis model. In this model, each document is treated as a topic hybrid model, and the topic's proportional prior distribution is a Dirichlet distribution. The LDA model does not reflect complex dependencies between underlying topics. Based on LDA, this paper introduces a new topic generation model, DMA (Dirichlet Mixture Allocation), which models document collections more accurately than LDA when documents are obtained in multiple classes. In this paper, we build a probabilistic topic model of DMA, use the method of variational inference to approximate each parameter of the model, study the model, and finally solve the estimation of each parameter.

1. Introduction
LDA model introduces Dirichlet distribution to hidden variables[1]. As a total probability document generation model, compared with PLSA, LDA randomly assigns values to test data and assigns probability to a new document. Moreover, compared with naive Bayesian model LDA does not produce overfitting phenomenon[1-3]. The subject model gradually reveals the intrinsic properties of data at the full probability level, which is not only used in text analysis, but also gradually applied in many aspects such as image and scene classification[4-6].

Although the LDA model with the introduction of Dirichlet distribution is obviously better than the naive Bayesian model PLSA[1], the LDA model is relatively single in the topic distribution, and it is not suitable for the generation of multiple types of documents. In real life, a large biological corpus can contain multiple classes, such as biomaterial science, bioinformatics, molecular biology, etc. In this case, we need multimodal modeling, and the LDA model clearly does not support this type of data set[7].

In view of the modeling of multimodal and related knowledge of DP process, a probabilistic generation model——DMA had been proposed. DMA is a further extension on the basis of LDA model. The single-mode modeling of LDA model is promoted to multi-mode modeling by using DP process. Because of the multimodal nature of mixed priors, DMA overcomes the limitations of LDA and has powerful capabilities in fitting multiple classes hidden within a data set, helping to understand the entire data set. Because DMA is an extension of LDA, it has the good features of LDA. However, precise DMA calculations are difficult, and MCMC approximations are slow when dealing with large documents. Therefore, we use the method of variational reasoning to approximate its posteriori
estimation and use the method of exponential distribution family to calculate the parameter estimation of the model[8].

This article is organized as follows: In the next section, we'll look at the DMA model in detail. In section 3, the method based on variational reasoning is used to approximate a posteriori estimation, and the parameter estimation is obtained by using exponential distribution family. Finally, we give our conclusion in the fourth section.

2. Dirichlet mixture allocation
In this section, DMA is described in detail. First, identify the language of the text collection, such as words, topics, and documents. In fact, the DMA model is not necessarily bound to text; it can be applied to other problems involving data sets, such as images. DMA, as a probabilistic generation model, defines documents that are retrieved probabilistically from different classes of topics, and each document is a mixture of potential topics specified by its topic ratio.

Specifically, the form has the following definition:

The topic distribution $q$ of each document is sampled from the Dirichlet mixed distribution of quantities:

$$DM(q|p,a) = \sum_{r=1}^{R} p_r \text{Dirichlet}(q|a_r)$$

Where, $p = \{p_1, \ldots, p_R\}$ is the mixing coefficient of Dirichlet mixing distribution, and $\sum_{r=1}^{R} p_r = 1$.

$a = \{a_1, \ldots, a_R\}$ represents the Dirichlet coefficient represented by each row of each Dirichlet distribution in the Dirichlet mixed distribution. $a_r = \{a_{r,1}, \ldots, a_{r,k}\}^T$ is a one-dimensional vector and is the corresponding Dirichlet parameter of the topic distribution $q$.

Each word $\{1, 2, \ldots, V\}$ in the vocabulary is assigned to different potential topics by polynomial distribution $b_z$. $V$ is the number of words in the word list, $b = (b_1, \ldots, b_k)$ indicating the matrix of potential topics.

In a document, $w = (w_1, w_2, \ldots, w_N)$ is a mixture of potential topics, which represents the position of a word in the document, such as $w_n$ is the n word in the document (note that $w_n$ can be repeated). In a document of N words, words $w_n$ are produced in two steps:

1. Sample a potential topic index $z_n \sim \text{Multinomial}(q)$, $z_n$ indicate which topic the n word $w_n$ in the document belongs to. It plays the role of an indicator, $z_n$ is a vector that is 1 at the topic and 0 at the rest.

2. Sample word $w_n \sim \text{Multinomial}(b_{z_n})$, indicating that the n word $w_n$ in the document was sampled from the n word in the $z$ topic.

The sampling above is done at two levels, one is the distribution from which the underlying topic variables in a single document are extracted, which is similar to the LDA model. The other section represents topic document modeling to illustrate the use of Dirichlet hybrid distribution to model multi-class structures hidden in document collections. Figure 1 shows the probability graph models of LDA,
and Figure 2 shows the DMA respectively. The difference between the two is that for the prior of the topic probability distribution, LDA is a single Dirichlet distribution, while DMA is a multiple class mixture of multiple Dirichlet distributions.

According to Figure 1, formula (2) is the subject index $z$, the document subject ratio $q$, and the joint distribution of the documents’ words $w$ when using DMA to generate documents:

$$p(z, q, w | p, a, b) = p(q | p, a) \prod_{n=1}^{N} p(z_n | q) p(w_n | b, z_n)$$  (2)

The joint distribution of formula (2) is explained as follows:

$p(q | p, a)$ represents the probability. Formula (3) is the Dirichlet mixed distribution:

$$p(q | p, a) = \sum_{r=1}^{R} p_D(q | a) = \sum_{r=1}^{R} [p(q | a)] \prod_{k=1}^{K} [G(a_{(r, k)})]^{-1} \prod_{k=1}^{K} [q^{(a_{(r, k)} - 1)}]$$  (3)

Where $G$ is the gamma function.

$p(z_n | q)$ represents the probability of the topic index to a certain topic based on the probability of the existing topic. Formula (4) represents the probability of the index to the first topic:

$$p(z_n = k | q), k=1, 2, ..., K$$  (4)

$p(w_n | b, z_n)$ represents the probability of drawing a word $w_n$ from a topic when sampling to the topic index $z_n$. Formula (5) indicates the probability when the word $w_n$ is $v$:

$$p(w_n = v | b, z_n) = p(w_n = v | b_{z_n}) = b_{z_n, v}$$  (5)

The only value that can be observed in the document is $w_n$, so its likelihood function $p(w | p, a, b)$ can be obtained by the marginal distribution of formula (2):

$$p(w | p, a, b) = \int p(q | p, a) \prod_{n=1}^{N} p(z_n | q) p(w_n | b, z_n) dq$$  (6)

The likelihood function of the document set $D$ is expressed by formula (7):

$$p(D | p, a, b) = \sum_{m=1}^{M} p(w_m | p, a, b) = \sum_{m=1}^{M} [p(q_m | p, a)] \prod_{n=1}^{N} p(z_{m, n} | b, z_{m, n}) dq_m$$  (7)

We obtain the joint distribution of potential variables $z$ and the likelihood function of the whole document set by analyzing the model diagram of DMA.

3. Inference and learning
The required parameters will be obtained by means of vartional inference.

The Dirichlet mixing prior is represented by introducing variables $r$, in DMA, the topic will be sampled from the Dirichlet mixing distribution, which is controlled by $r$ and $p$ parameters. By adding $r$ the joint distribution in the previous section, it can be understood that the current document is sampled from the $r$ Dirichlet distribution. Formula (8) is the auxiliary variable $r$, topic index $z$, document topic ratio $q$, and the joint distribution of the word $w$ of the document:

$$p(r, z, q, w | p, a, b) = p(r | p) p(q | a_r) \prod_{n=1}^{N} p(z_n | q) p(w_n | b, z_n)$$  (8)

Explain formula (8):

$p(r | p)$ represents the probability of drawing the $r$ Dirichlet distribution after disassembling the Dirichlet mixture with DP, $p$ represents dismantling the mixture by folding sticks, and $p = (p_1, p_2, ..., p_N)$ represents the probability weight vector generated by the process of breaking sticks, where $N$ can be infinity.

$p(q | a_r)$ represents the probability distribution of topic $q$ in the case of the $r$ Dirichlet distribution currently drawn.
\( p(z_n | q) \) and \( p(w_n | b, z_n) \) have been explained in Formula (4) and Formula (5), they does not repeat.

Since the observable values are the words \( w_n \) in the document, the posteriori estimation of the potential variable corresponding to formula (9) is:

\[
p(r, z, q | w, p, a, b) = \frac{p(r, z, q, w | p, a, b)}{p(w | p, a, b)}
\]

(9)

The exact calculation of the denominator is difficult and formula (10) can treat the denominator as a constant:

\[
p(r, z, q | w, p, a, b) \propto p(r, z, q, w | p, a, b)
\]

(10)

Formula (10) is introduced into the method of variational inference to approximate and then a posteriori estimate is solved.

An independent and identically distributed variational distribution is constructed on the auxiliary variable \( r \) and the potential variables \( q \) and \( z \) respectively, where the distribution is parameterized by free variational parameters. In particular, the variational distribution of Formula (11) is chosen:

\[
q(r, z, q | l, g, f) = q(r | l) q(q | g) \prod_{n=1}^{N} q(z_n | f_n)
\]

(11)

\( q(r | l) \) is the polynomial distribution with a parameter of \( l \), \( q(q | g) \) is the Dirichlet distribution with a parameter of \( g \), and \( q(z_n | f_n) \) is the polynomial distribution with a parameter of \( f_n \).

The next step is divided into two steps, the first is the inference process using exponential distribution family method to solve the variational parameter \( l, g, f \).

Formula (12) and (13) is used to solve the variational parameter \( g \), and \( p(q | z) \) is changed into the form of exponential distribution family:

\[
p(q | z) \propto p(q | p, a) p(z | q)
\]

\[
\sum_{r=1}^{R} \prod_{k=1}^{K} (q_{d,k}^{a_{d,k} - 1} \prod_{n=1}^{N} q_{d,k}^{d(z_{d,n,k})})
\]

\[
= \sum_{r=1}^{R} \exp(\log p_r) \exp(\log \prod_{k=1}^{K} (q_{d,k}^{a_{d,k} - 1} \prod_{n=1}^{N} q_{d,k}^{d(z_{d,n,k})}))
\]

\[
= \sum_{r=1}^{R} \exp((\log q_{d,1}, \ldots, \log q_{d,K}) \cdot h(z, a, p) + \log p_r)
\]

(12)

\[
h(z, a, p) = \left\{ \begin{array}{l}
    a_{r,1} - 1 + \sum_{n=1}^{N} d(z_{d,n,1}), \\
    \ldots, \\
    a_{r,K} - 1 + \sum_{n=1}^{N} d(z_{d,n,K})
\end{array} \right.
\]

(13)

Equation (14) is used to solve the variational parameters \( g \):
\[ h(g_d) = \sum_{r=1}^{R} E_{q(z|y)q(y|g)} h(z,a,p) \]
\[ = \left\{ \sum_{r=1}^{R} \left( E_{q(z|y)} \left( \sum_{n=1}^{N} d(z_{d,n,L}) \right) + E_{q(y|g)} (a_{r,L}-1) \right) \right\} \]
\[ \vdots \]
\[ = \left[ \sum_{r=1}^{R} (a_{r,L} - 1 + \sum_{n=1}^{N} d(z_{d,n,L}) y_{d,n,L}) \right] \]
\[ = \left[ \sum_{r=1}^{R} (a_{r,L} + \sum_{n=1}^{N} d(z_{d,n,L}) y_{d,n,L}) \right] \]
\[ \vdots \]
\[ g_d = \left\{ \sum_{r=1}^{R} (a_{r,L} + \sum_{n=1}^{N} d(z_{d,n,L}) y_{d,n,L}) \right\} \]
\[ \vdots \]
\[ = \sum_{r=1}^{R} (a_{r,L} + \sum_{n=1}^{N} f_{d,n}) \]
\[ = \sum_{r=1}^{R} a_{r,L} + \sum_{n=1}^{N} f_{d,n} \]

The solution of the variational parameter \( l, f \) can be obtained.

The final optimal variational parameters can be obtained from equation (16):
\[ l \propto p \exp \left( \log G(\sum_{k=1}^{K} a_{i,k}) - \sum_{k=1}^{K} \log G(a_{i,k}) \right) \]
\[ f_{k} \propto b_{k,w} \exp (Y(g_{k}) - Y(\sum_{k=1}^{K} g_{k})) \]

The optimal solution of the variational parameters is obtained, and \( l^* \) can be used as an estimate of the topic ratio \( q \) of document \( w \), and which topics in document \( w \) are explained. At the same time, we can use \( f^* \) to represent the estimate of the subject index \( z \) in document \( w \).

We study the model and solve the estimation of model parameter \( a, b, p \). By means of variational reasoning, the maximization of ELOB is obtained, and the variational EM method is used to update the lower bound and model parameters until convergence.

ELOB is represented by Jasons inequality:
\[
\log p(w | p, a, b) \geq L(l^*, g^* f^* | p, a, b)
\]
\[
= E_q[\log p(r | p)] + E_q[\log q | a]
+ E_q[\log p(z | q)] + E_q[\log p(w | z, b)] - E_q[\log q^*]
\]
(17)

By summing all documents on formula (17), the lower bound of logarithmic likelihood of this set can be obtained by formula (18):
\[
L(D | p, a, b) = \sum_{m=1}^{M} L(l^*, g^* f^* | p, a, b)
\]
(18)

In the above article, the E step of EM algorithm is replaced by the exponential distribution family method when solving the variational parameters of. As follows, the model parameter \( p, a, b \) is solved by M step. By calculating the lower bound of logarithmic likelihood, we can get the optimal estimation of \( p_r^* \) and \( b_r^* \):
\[
p_r^* = \sum_{m=1}^{M} l_{mr}
\]
(19)
\[
b_{k_r}^* = \sum_{m=1}^{M} \sum_{n=1}^{N} d(w_{mn})
\]
(20)

Parameter \( a \) is estimated through the maximization of ELOB, and \( a_r \) represents the parameter of the \( r \) Dirichlet distribution. Formula (21) maximizes \( a \) through formula (18):
\[
L_{[a]} = \sum_{m=1}^{M} \left\{ \log G(\sum_{k=1}^{K} a_{rk}) - \sum_{k=1}^{K} \log G(a_{rk}) + \sum_{k=1}^{K} (a_{rk} - 1)(Y(g_{mk}) - Y(\sum_{k=1}^{K} g_{mk})) \right\}
\]
(21)

Newton iteration method is used for Equation (21), and the gradient and hessian value obtained are expressed by formula (22) and formula (23):
\[
\frac{\partial L_{[a]}}{\partial a_{rk}} = \sum_{m=1}^{M} \left[ l_{mr} Y(\sum_{k=1}^{K} a_{rk}) - Y(a_{rk}) + (Y(g_{mk}) - Y(\sum_{k=1}^{K} g_{mk})) \right]
\]
(22)
\[
\frac{\partial^2 L_{[a]}}{\partial a_{rk} \partial a_{rk'}} = \sum_{m=1}^{M} \left[ Y'(\sum_{k=1}^{K} a_{rk}) - d(k_r, k_r') Y(a_{rk}) \right]
\]
(23)

Thus, the optimal model parameter \( p^*, a^*, b^* \), \( g^* f^* \) of each document can be obtained by the coordinate ascending method in formula (17), and the lower bound of ELOB can be updated by calculating formula (18). The above two steps are computed iteratively until the lower bound of the logarithm converges.

4. Conclusions

We introduced DMA model, a new topic model designed to model collections of documents containing multiple classes. DMA uses Dirichlet blending as a priority for document subject proportions, so it is well suited to the multi-class structure of such document collections. By considering the unimodal Dirichlet prior distribution in LDA, DMA can be regarded as a generalization of LDA. As a result, DMA overcomes the problem of underfitting when LDA models multiclass document collections. In addition, LDA is a PLSI Bayesian generalization using specific priors, so the DMA does not experience fitting problems. DMA is based on the LDA model and provides a useful inference mechanism in areas involving multiple level structures. The main advantages of being a generative model include its modularity and extensibility. As a probabilistic module, linear discriminant analysis can be easily embedded into a more complex model. Instead of a single Dirichlet distribution in the LDA model, DMA uses a mixture of Dirichlet distributions. This allows for richer structure in the underlying topic
space, and in particular allows for forms of document clustering that are different from clustering implemented through shared topics. With the construction of more thematic models, the future development will tend to the comparison of model parameter approximation and the application and promotion of models in various aspects such as image and scene classification.

References
[1] Blei, D, Ng, A, Jordan, M. (2003) Latent Dirichlet allocation. Journal Machine Learning Research, vol.3: 993–1022.
[2] Griffiths, T.L, Steyvers, M. (2004) Finding scientific topics. Proc Natl A cad, Sci USA vol, 101 Suppl 1: 5228–5235.
[3] Blei, D.M. (2012) Probabilistic topic models. Communications of the ACM, 55 (4): 77-84.
[4] Sivic, J, Russell, B,E, fros, A, Zisserman, A, Freeman, W.T. (2005) Discovering objects and their locations in images. Proceedings of Tenth IEEE International Conference onComputer Vision.
[5] Russell, B, Efros, A, Sivic, J, Freeman, W, Zisserman, A. (2006) Using multiple segmentations to discover objects and their extent in image collection. Proceedings of Eleventh IEEE International Conference on Computer Vision.
[6] Fei-Fei, L, Perona, P. (2005) A bayesian hierarchical model for learning natural scene categories. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.
[7] Yu, K, Yu, S, Tresp, V. Dirichlet enhanced latent semantic analysis. In: Conference in Artificial Intelligence and Statistics. pp. 437–444.
[8] Hoffman, M.M, Blei, D, Paisley, J. (2013) Stochastic Variational Inference. Journal of Machine Learning Research.