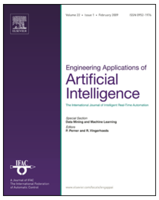
Engineering Applications of Artificial Intelligence 35 (2014) 176–186

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



A variational Bayes model for count data learning and classification

Ali Shojaee Bakhtiari [a], Nizar Bouguila [b,*]

[a] Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada H3G 1T7
[b] The Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada H3G 1T7

ARTICLE INFO

Article history:
Received 21 September 2013
Received in revised form
27 March 2014
Accepted 29 June 2014
Available online 20 July 2014

Keywords:
Latent topic models
Count data
Generalized Dirichlet distribution
Variational Bayes
Text classification
Visual scene categorization

ABSTRACT

Several machine learning and knowledge discovery approaches have been proposed for count data modeling and classification. In particular, latent Dirichlet allocation (LDA) (Blei et al., 2003a) has received a lot of attention and has been shown to be extremely useful in several applications. Although the LDA is generally accepted to be one of the most powerful generative models, it is based on the Dirichlet assumption which has some drawbacks as we shall see in this paper. Thus, our goal is to enhance the LDA by considering the generalized Dirichlet distribution as a prior. The resulting generative model is named latent generalized Dirichlet allocation (LGDA) to maintain consistency with the original model. The LGDA is learned using variational Bayes which provides computationally tractable posterior distributions over the model's hidden variables and its parameters. To evaluate the practicality and merits of our approach, we consider two challenging applications namely text classification and visual scene categorization.

1. Introduction

Count data appear in many domains (e.g. data mining, computer vision, machine learning, pattern recognition, and bioinformatics) and applications. Examples include textual documents and images modeling and classification where each document or image can be represented by a vector of frequencies of words (Nigam et al., 2000) or visual words (Csurka et al., 2004), respectively. The extraction of knowledge hidden in count data is a crucial problem which has been the topic of a significant amount of research in the past. The naive Bayes assumption, through the consideration of the multinomial distribution, was extensively used for count data modeling (Nigam et al., 2000). However, serious deficiencies such as, being prone to training bias, the need for the assumption of independence for features and failure to model text well, were observed with the application of the multinomial distribution as thoroughly discussed in Madsen et al. (2005) and Bouguila and Ziou (2007a). The most widely used solution to overcome these deficiencies is the consideration of the Dirichlet distribution as a conjugate prior to the multinomial which generally offers better flexibility, generalization and modeling capabilities (Madsen et al., 2005; Bouguila and Ziou, 2007a; Mei et al., 2007). Despite many favorable features, it has been pointed out that the Dirichlet distribution has some shortcomings, also. The main disadvantages of the Dirichlet distribution are its very restrictive negative covariance matrix and the fact that the elements with similar mean values must have absolutely the same variance which is not always the case in real-life applications (Bouguila, 2008). To overcome those deficiencies, research has been focused on providing a transition from the Dirichlet assumption to better modeling assumptions (Bouguila, 2011). The context of this paper is majorly about this transition as well, where the ultimate goal is to have more accurate data modeling.

One of the immediate applications of proper data modeling is classification. It covers a vast extent of problems such as placement of textual data into appropriate library entries or classifying objects into their relevant categories. In this context, one of the most challenging tasks is the classification of visual scenes without going deep inside their semantics. The challenge behind the former is that visual scenes are generally composed of a huge number of minute objects. The presence of this ever occurring objects makes it extremely complicated to develop useful classifiers based on the semantics alone. After all one would expect to see roads, trees, sun and the sky recurring in scenes both taken inside the city or in the suburb. The need to consider the presence of recurring data singletons, whether words, visual words or visual objects, led to the so-called topic based models. Latent semantic indexing (LSI) (Deerwester et al., 1990) is the first successful model proposed to extract recurring topics from data. It was proposed for textual documents modeling using mainly singular value decomposition (SVD). A generative successful extension of LSI called probabilistic latent semantic indexing (PLSI) was proposed in Hofmann (2001). And a hierarchical extension of PLSI was proposed in Vinokourov and

[*] Corresponding author. Tel.: +1 5148482424; fax: +1 5148483171.
E-mail addresses: al_sho@encs.concordia.ca (A.S. Bakhtiari), nizar.bouguila@concordia.ca (N. Bouguila).

http://dx.doi.org/10.1016/j.engappai.2014.06.023
0952-1976/

Girolami (2002). However, PLSI is only generative at the words layer and does not provide a probabilistic model at the level of documents. Therefore, two major problems arise with PLSI. Firstly, the number of parameters increases with the number of documents. Secondly, it is not clear how one can learn a document outside of the training phase. To overcome these shortcomings, the authors in Blei et al. (2003a) proposed the LDA model which has so far proven to be a reliable and versatile approach for data modeling. LDA has received a particular attention in the literature and several applications (e.g. natural scene classification Fei-Fei and Perona, 2005) and extensions have been proposed. Examples of extensions include the hierarchical version of LDA (Blei et al., 2003b), used for instance in Sivic et al. (2008) for hierarchical object classification, the online version proposed in Hoffman et al. (2010), and the discriminative supervised version described in Lacoste-Julien et al. (2008). Of course, these extension efforts are useful for several real-life applications and scenarios, but have ignored an important aspect of LDA namely the fact that it considers the Dirichlet distribution, including its drawbacks, for generating latent topics. Previously other researchers tried to develop latent topic models based on the conjugate priors other than Dirichlet (Caballero et al., 2012). Their model however is based on Gibbs sampling and Markov chain Monte Carlo (MCMC) method (Robert and Casella, 2004). The advantage of the MCMC method is its relative ease of derivation. However, it has been shown that sampling methods require much more computation time than deterministic methods such as variational Bayes. Therefore, where it is possible to derive an analytic form, deterministic models are more preferable. In this work we shall focus on deriving an extension to the LDA model using the generalized Dirichlet assumption using the variational Bayes method.

Recently, the second author has shown that the generalized Dirichlet is a good alternative to the Dirichlet when using finite mixture models for count data clustering (Bouguila, 2008). Like the Dirichlet, the generalized Dirichlet distribution is a conjugate prior to the multinomial distribution which is a crucial property in the LDA model. Moreover, the generalized Dirichlet has a more versatile covariance matrix and also it lifts the variance limitations facing Dirichlet vectors (Bouguila, 2008). The goal of this work is to propose an extension of LDA based on the generalized Dirichlet distribution. To maintain consistency with the LDA model we call our model, latent generalized Dirichlet allocation (LGDA). We shall develop a variational Bayes estimation approach inspired from the one proposed in Blei et al. (2003a), yet with the generalized Dirichlet assumption. The Dirichlet distribution is a special case of the generalized Dirichlet distribution (Connor and Mosimann, 1969; Bouguila and Ziou, 2007b), therefore it is expectable that the LGDA will provide good modeling capabilities. In the experimental results we shall elaborate the conjunctions between the two models further. We shall compare the two models via two challenging applications namely text and visual scene classification.

The rest of the paper is organized as follows. In Section 2, we introduce the LGDA model and give the detailed derivations to learn its parameters. Section 3 is devoted to the presentation of the results of applying both LDA and LGDA. The applications concern text and visual scene classification and are used to show the strengths and weaknesses of both models. Finally, conclusion and some thoughts about future directions follow in Section 4.

2. Latent generalized Dirichlet allocation

2.1. The model

Like LDA, LGDA is a fully generative probabilistic model over a corpus. A corpus in our case is a collection of M documents (or images) denoted by $\boldsymbol{M} = (\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_M})$. And each document $\boldsymbol{w}_m$ is a sequence of N_m words $\boldsymbol{w}_m = (w_{m1}, \dots, w_{mN_m})$. In what follows, for sheer convenience, we drop the index m wherever we are not referring to a specific document. The word $w_n = (w_n^1, \dots, w_n^V)$ is considered as a binary vector drawn from a vocabulary of V words, so that $w_n^j = 1$ if the j-th word is chosen and zero, otherwise. The model proceeds with generating every single word (or visual word) of the document (or the image) through the following steps:

1. Choose $N \propto Poisson(\zeta)$.
2. Choose $(\theta_1, \dots, \theta_d) \propto GenDir(\overrightarrow{\xi})$.
3. For each of the N words w_n:
   (a) choose a topic $z_n \propto Multinomial(\overrightarrow{\theta})$,
   (b) choose a word w_n from $p(w_n | z_n, \mu_w)$.

In above z_n is a $d+1$ dimensional binary vector of topics defined so that $z_n^i = 1$ if the i-th topic is chosen and zero, otherwise. We define $\overrightarrow{\theta} = (\theta_1, \dots, \theta_{d+1})$, where $\theta_{d+1} = 1 - \sum_{i=1}^{d} \theta_i$. We define matrix μ_w so that a chosen topic is attributed to a multinomial μ_w over the vocabulary of words so that $\mu_{w(ij)} = p(w^j = 1 | z^i = 1)$, from which every word is randomly drawn. $p(w_n | z_n, \mu_w)$ is a single draw multinomial probability conditioned on z_n and $GenDir(\overrightarrow{\xi})$ is a d-variate generalized Dirichlet distribution with parameters $\overrightarrow{\xi} = (\alpha_1, \beta_1, \dots, \alpha_d, \beta_d)$ and probability distribution function given by

$$p(\theta_1, \dots, \theta_d | \overrightarrow{\xi}) = \prod_{i=1}^{d} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \theta_i^{\alpha_i - 1} \left(1 - \sum_{j=1}^{i} \theta_j\right)^{\gamma_i} \tag{1}$$

where $\gamma_i = \beta_i - \alpha_{i+1} - \beta_{i+1}$. It is straightforward to show that when $\beta_i = \alpha_{(i+1)} + \beta_{(i+1)}$, the generalized Dirichlet distribution is reduced to Dirichlet distribution (Bouguila and Ziou, 2007b). We define $\overrightarrow{\theta} = (\theta_1, \dots, \theta_{d+1})$, where $\theta_{d+1} = 1 - \sum_{i=1}^{d} \theta_i$. With the above parameters, the mean and the variance matrix of the generalized Dirichlet elements are as follows (Bouguila and Ziou, 2007b):

$$E(\theta_i) = \frac{\alpha_i}{\alpha_i + \beta_i} \prod_{k=1}^{i-1} \frac{\beta_k}{\alpha_k + \beta_k} \tag{2}$$

$$Var(\theta_i) = E(\theta_i) \left(\frac{\alpha_i + 1}{\alpha_i + \beta_i + 1} \prod_{k=1}^{i-1} \frac{\beta_k + 1}{\alpha_k + \beta_k} + 1 - E(\theta_i) \right) \tag{3}$$

and the covariance between θ_i and θ_j is given by

$$Cov(\theta_i, \theta_j) = E(\theta_j) \left(\frac{\alpha_i}{\alpha_i + \beta_i + 1} \prod_{k=1}^{i-1} \frac{\beta_k + 1}{\alpha_k + \beta_k} + 1 - E(\theta_i) \right) \tag{4}$$

It can be seen from Eq. (4) that the covariance matrix of the generalized Dirichlet distribution is more general than the covariance matrix of the Dirichlet distribution and unlike Dirichlet distribution it is possible for two elements inside the random vector to be positively correlated. Also unlike Dirichlet two elements with the same mean value can have different variances. Generalized Dirichlet distribution, like the Dirichlet distribution, belongs to the exponential family of distributions (see Appendix A). This means that the generalized Dirichlet distribution has a conjugate prior that can be developed in a formal way, which is an important property that we shall used in the following for the learning of our model. It turns out also that generalized Dirichlet like Dirichlet is the conjugate prior of the multinomial distribution. This implies that if $(\theta_1, \dots, \theta_d)$ follows a generalized Dirichlet distribution with parameters $\overrightarrow{\xi}$, and $\overrightarrow{N} = (n_1, \dots, n_{d+1})$ follows a multinomial with parameter $\overrightarrow{\theta}$, then the posterior distribution $p(\overrightarrow{\theta} | \overrightarrow{\xi}, \overrightarrow{N})$ also follows a generalized Dirichlet distribution with parameters ξ' given as follows (Bouguila, 2008):

$$\alpha'_i = \alpha_i + n_i \tag{5}$$

$$\beta'_i = \beta_i + \sum_{l=i+1}^{d+1} n_l \tag{6}$$