LogisticLDA: Regularizing Latent Dirichlet Allocation by Logistic Regression*

Jia-Cheng Guo[a], Bao-Liang Lu[a,b], Zhiwei Li[c], and Lei Zhang[c]

[a] Center for Brain-Like Computing and Machine Intelligence
Department of Computer Science and Engineering
[b] MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems
Shanghai Jiao Tong University
800 Dong Chuan Road, Shanghai 200040, China
bllu@sjtu.edu.cn
[c] Microsoft Research Asia
49 Zhichun Road, Haidian District, Beijing 100080, China
{zli, leizhang}@microsoft.com

Abstract. We present in this paper a supervised topic model for multi-class classification problems. To incorporate supervisory information, we jointly model documents and their labels in a graphical model called LogisticLDA, which mathematically integrates a generative model and a discriminative model in a principled way. By maximizing the posterior of document labels using logistic normal distributions, the model effectively incorporates the supervisory information to maximize inter-class distance in the topic space, while documents still enjoy the interchangeability characteristic for ease of inference. Experimental results on three benchmark datasets demonstrate that the model outperforms state-of-the-art supervised topic models. Compared with support vector machine, the model also achieves comparable performance, but meanwhile it discovers a topic space, which is valuable for dimension reduction, topic mining and document retrieval.

Keywords: supervised topic model, text categorization, graph model, statistical learning

1 Introduction

As an unsupervised method, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) seeks to simultaneously find a set of basis (i.e. topics) and embed documents to the latent space spanned by this basis. Due to its inherent capability of producing interpretable and semantically coherent topics, LDA has been widely used in text analysis and shown promising performance in tasks like topic mining, browsing, and accessing document similarity. In contrast, when LDA is applied to text classification tasks, it is often used as only a dimension reduction step to extract features for consecutive discriminative models (e.g. SVM). Because the objective of LDA (as well as other unsupervised topic models) is to infer the best set of latent topics that can explain the document collection rather than separate different classes, the training of LDA is actually independent to supervisory information. This paradigm greatly limits its applications in classification tasks.

To address this issue, some topic models which integrate supervisory information have been proposed in recent years. Mimno and McCallum categorized these models to two types: upstream and downstream models. In "upstream" models, hidden topics are generated by conditioning on supervisory information(i.e. $p(z|y)$, where z means a hidden topic and y means the category of

* J.C. Guo and B.L. Lu are supported by the National Natural Science Foundation of China (Grant No. 60773090 and Grant No. 90820018), the National Basic Research Program of China (Grant No. 2009CB320901), and the National High-Tech Research Program of China (Grant No. 2008AA02Z315).

a document). Examples of upstream models include the Dirichlet Multinomial Regression model (DMR) (Mimno and McCallum, 2008), and the Theme Model (Li and Perona, 2005). Although upstream models explicitly model the class distribution $p(z|y)$ (i.e. reducing intra class variance) in the topic space, they disregard the inter class information (without maximizing inter-class distance), which is very crucial for classification problems. Though Discriminative LDA model (DiscLDA) (Lacoste-Julien et al., 2008) has similar modelling, its unique training scheme make it more like a hybrid of "upstream" and "downstream" models.

In "downstream" models, the supervisory information and words are simultaneously generated from hidden topics (i.e. $p(w, y|z)$, where w means a word), like the supervised LDA (sLDA) proposed by Blei and McAuliffe (2007), and Topic over Time Model (Wang and McCallum, 2006). Although supervisory information in upstream models looks to be more deeply involved in the likelihood function than in downstream models, downstream models are more like discriminative models because the posterior of supervisory information explicitly appears in models. As a regression model, the discriminative information captured in the model can help find a topic space in which class labels can be generated with minimal errors. However, it's unclear that how to adapt sLDA to multi-class classification problem.

In this paper, we develop a new downstream topic model for multi-class classification problems. Following the work of Blei and McAuliffe, we also employ LDA to project documents into a topic space (topic simplex). In the topic space, for each class we place a corresponding class prototype, and for each document, we calculate the memberships that this document embedding belongs to each class. Here the membership function is defined as a normal distribution centered at each class prototype. The membership scores of this document are then combined together as the parameter of a multinomial distribution to generate the corresponding label. Because labels are observed, to best explain the observed data, the multinomial parameters (i.e. membership scores) for each multinomial distribution, have to be as sparse as possible. This sparse attribute in turn imposes an implicit constraint to the model that the class prototypes should be apart from each other. In this way, the model effectively incorporates the supervisory information to maximize inter-class distance in the topic space, while documents still enjoy the interchangeability characteristic for ease of inference. The proposed LogisticLDA model can be regarded as an extension of sLDA to multi-class classification. As we will show in Section 2, the objective function (log-likelihood) of this model logically consists of two terms. One is the log-likelihood of documents in LDA model, and the other one is the posterior of document labels given the embeddings of the documents. The posterior in the second term is just the probability used in logistic regression. By choosing a logistic normal distribution as the distribution to generate categories of documents, a logistic regression loss can be plugged in LDA model in a principled way. This is the reason why the model is named as LogisticLDA.

2 LogisticLDA Model

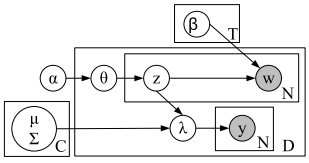


Figure 1: A graphical model representation of LogisticLDA

LogisticLDA is a supervised topic model for multi-class classification problems by extending LDA model. Its graphical model representation is shown in Figure 1. In LDA model, each word of a document is assumed to arise from a set of latent topics($\beta_{1:T}$), which are *Multinomial* distributions over vocabulary. Documents in a corpus share the same set of *T* topics, but each document has its own mixture proportion (θ). The mixture proportion is an embedding of a document in the

learnt latent topic space. Usually, it is used as low-dimensional features in discriminative models to classify documents (Blei et al., 2003) and LDA is only used as a dimension reduction step.

The basic idea of LogisticLDA is to regularize the embeddings of documents (θ) by incorporating supervisory information, and consequently propagating its influence to topic distributions ($\beta_{1:T}$). To support a good classification, embeddings of documents should have some good properties. For example, the embedding of each document should be close to its category center but far from centers of other categories. Adding this requirement to the objective function of a topic model will constrain the embeddings of documents, and consequently the change of document embeddings will further influence the distributions of topics.

To incorporate supervisory information in a more principled way rather than using a heuristic regularization approach, we add to the model a set of label variables y associated with each document, and jointly model the documents and their labels. We assume each category follows a Gaussian distribution $N(\mu_i, \Sigma^{-1})$, in which μ_i is a category prototype in the topic simplex. For each document, we calculate the likelihoods that the document embedding θ is generated by these categories in the topic simplex $p(y = k|\theta, \mu_i, \Sigma^{-1})$. By normalizing these probabilities, we get a C-dimensional *Multinomial* distribution $p(y = k|\lambda)$, whose parameter λ reflects the probabilities that the document belongs to each category. That is, $\lambda_i \propto N(\theta, \mu_i, \Sigma^{-1})$. Such a distribution $p(y = k|\theta, \mu_i, \Sigma^{-1})$ is actually the logistic normal distribution taken in Correlated Topic Model (Blei and Lafferty, 2005). This multinomial distribution $p(y = k|\lambda)$ is used to generate the document label y.

As document labels are observed variables, to best generate the data, the multinomial parameter λ needs to be as sparse as possible. That is, the parameter component corresponding to the document's category should be large while other components should be small. As λ is calculated from Gaussian distributions, $\lambda_i \propto N(\theta|\mu_i, \Sigma^{-1})$, the sparse property of the multinomial parameter in turn imposes an implicit constraint to the model that the category prototypes should be apart from each other and inter-class distance is thus maximized. This explains why maximizing the posterior of document labels will adjust the category distributions and reduce the classification error.

Finally, due to the fact that the scale of a single label's multinomial distribution probability is far smaller than the the scale of document's word generative probability, we proportionally augment document's labels to the number of words in the document. So that, during the training process, the regularization from label won't be simply omitted for the gain of document's word generative probability.

In summary, each document and its supervisory side information arise from the following generative process under the LogisticLDA model:

1. Draw topic proportion $\theta \sim Dir(\alpha)$
2. For each of the N words w_n:
   (a) Draw topic assignment $z_n \sim Mult(\theta)$
   (b) Draw word $w_n|z_n, \beta_{1:K} \sim Mult(\beta_{z_n})$
3. Compute $\lambda \sim softmax(\bar{z}; \mu_{1:C}, \Sigma^{-1})$, which is $\lambda_i = \frac{\exp(-(\bar{z}-\mu_i)^T\Sigma^{-1}(\bar{z}-\mu_i)/2)}{\sum_{j=1}^{C}\exp(-(\bar{z}-\mu_j)^T\Sigma^{-1}(\bar{z}-\mu_j)/2)}$
4. For each of the N labels y_n, draw topic assignment $y_n \sim Mult(\lambda)$

Here we define $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$. Note that the expectation of $\bar{z}$ is θ.

2.1 Discussions

Why $\bar{z}$ not θ? As our discussion, a more straightforward formulation is to directly associate λ to the topic mixture θ rather than the empirical topic frequencies $\bar{z}$ in the proposed model. However, this association makes the derivation of training algorithm much more complex. As the expectation of

$\bar{z}$ is θ, the two formulations actually have almost equivalent functionalities. Blei et al. has another argument to this problem(Blei and McAuliffe, 2007).

Modeling of categories. In previous research on supervised embedding of documents, there are some common criteria, e.g. distances among documents in the same category (intra-class variance) should be small while distances among documents in different categories (inter-class distance) should be large, and large-margin rules. However, these criteria are difficult to be introduced to topic models because they will also introduce extra dependencies among documents while we wish to keep the document exchangeability property to make the model inferable. In LogisticLDA, by introducing the multinomial distribution $p(y = k|\lambda)$ to generate observed labels, we convert dependencies between documents to dependencies between documents and categories. In this way, we effectively incorporate the supervisory information to maximize inter-class distance in the topic space, while documents still enjoy the interchangeability characteristic for easy inference.

A concern about the single Gaussian per category assumption is that single Gaussian cannot deal with complex decision boundary as the embeddings of documents may not be linearly separable. However, if the number of topics is not too small (e.g. 100 topics), the possible forms of topic space are huge, as well as embeddings of documents in it. Thus, by regulating the topic space with label information, we should be able to find a topic space under which documents are linearly classifiable. Our experiments on various datasets indicated the reasonableness of this assumption.

Reason of the name. Logically, the log-likelihood of this model consists of two terms[1]:

$$L(\boldsymbol{w}, \boldsymbol{y}; \alpha, \mu, \Sigma^{-1}) = L_{LDA} + \sum_{n=1}^{N} E[\log(p(y_n|\bar{z}, \mu, \Sigma^{-1}))]$$

where $\boldsymbol{w}$ denotes a document. The first term is the log-likelihood of the document in LDA model, and the second term is the posterior of the document label given the embedding of this document. Thus, LogisticLDA could be treated as a combination of a generative model and a discriminative model. The generative model seeks to simultaneously find a set of basis (i.e. topics) and embed documents to the latent space spanned by this basis, while the discriminative model regulates the embedding of documents by forcing category prototypes apart from each other. If we assume the variance of each category distribution is I, by removing the quadratic term $exp(-\frac{1}{2}\bar{z}^T\bar{z})$, $p(y = k|\bar{z}, \mu_{1:C}, I)$ can be rewritten as

$$p(y = k|\bar{z}, \mu_{1:C}, I) = \frac{\exp(\mu_k^T\bar{z} - \frac{1}{2}\mu_k^T\mu_k)}{\sum_{j=1}^{C}\exp(\mu_j^T\bar{z} - \frac{1}{2}\mu_j^T\mu_j)}$$

It is the probability of logistic regression. Thus, this model is named as logisticLDA.

3 Approximate Inference

Given parameters α, β, μ and Σ^{-1}, the joint probability of a document and its label is given by:

$$p(\boldsymbol{w}, \boldsymbol{y}|\alpha, \beta_{1:T}, \mu_{1:C}, \Sigma^{-1}) = \int d\theta p(\theta|\alpha) \sum_{Z_{1:n}} \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n) \prod_{n=1}^{N} p(y_n|\bar{z}, \mu_{1:C}, \Sigma^{-1}) \quad (1)$$

However, due to the coupling between hidden variables, θ and z, in the integration, the computation is intractable. So we adopt a variational method to approximate it (Bishop, 2006).

We use the same approximate distribution as in LDA,

$$q(\theta, z_{1:n}|\gamma, \phi_{1:N}) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \quad (2)$$

[1] Mathematically, only by introducing a variational distribution, the log-likelihood can be decomposed to these two terms.

where $q(\theta|\gamma)$ is a K-dimension Dirichlet distribution, each $q(z_n|\phi_n)$ is a K-dimension multinomial distribution, and γ,ϕ_n are corresponding variational parameters. Then a lower bound of log-likelihood of the model is given by:

$$\log(p(\boldsymbol{w},\boldsymbol{y}|\alpha,\beta_{1:T},\mu_{1:C},\Sigma^{-1})) \geq E_q[\log p(\theta|\alpha)] + \sum_{n=1}^{N} E_q[\log p(z_n|\theta)] + \sum_{n=1}^{N} E_q[\log p(w_n|z_n)] + \sum_{n=1}^{N} E_q[\log p(y_n|\mu_{1:C},\Sigma^{-1})] + H(q) \quad (3)$$

Except for the fourth term of the right hand side in Eq. 3, all other terms are identical to the terms in the variational estimation of LDA. For the clarity of discussion, in the following text, we constrain the problem to a single label classification, as the extension to multi-label problem is straight forward. The fourth term is the expectation of log-likelihood of label y under approximate distribution $q(\theta, z_{1:N})$, which can be written as,

$$E_q[\log p(y|\mu_{1:C},\Sigma^{-1})] = E_q[-\log(1+\sum_{i=1,i\neq y}^{C}\exp(\bar{z}^T\Sigma^{-1}(\mu_i-\mu_y)+(\mu_y^T\Sigma^{-1}\mu_y-\mu_i^T\Sigma^{-1}\mu_i)/2))] \quad (4)$$

However, this term is still difficult to calculate under the approximated distribution $q(\theta, z_{1:N})$. So in order to calculate it, we give up some flexibilities of the model. First, we enforce the quadratic term $\mu_i^T\Sigma^{-1}\mu_i = 1$ to avoid a non-convex optimization problem. Actually, this constraint is equal to put the class centers onto unit sphere instead of on the simplex. Second, we set $\Sigma^{-1} = \mathrm{I}$, where I is the identity matrix. Finally, we get a much simpler equation:

$$E_q[\log p(y|\mu_{1:C},\Sigma^{-1})] = E_q[-\log(1+\sum_{i=1,i\neq y}^{C}\exp(\bar{z}^T(\mu_i-\mu_y)))] \quad (5)$$

To compute this term, we apply Taylor expansion twice to find a approximation of it. We first expand the logsum at ξ, and then expand each $\exp(\bar{z}^T(\mu_i-\mu_y))$ term at ρ_i. The approximation is

$$E_q[\log p(y|\mu_{1:C},\Sigma^{-1})] \approx -\log\xi - \frac{1}{\xi}(1+\sum_{i=1,i\neq y}^{c} e^{\rho_i} + e^{\rho_i}(E_q[\bar{z}^T(\mu_i-\mu_y)]-\rho_i) \quad (6)$$

where $E_q[\bar{z}^T(\mu_i-\mu_y)] = \frac{1}{N}\sum_{n=1}^{N}\phi_n^T(\mu_i-\mu_y)$. Two new variational parameters, ξ and ρ, are introduced into the variational inference.

3.1 E-Step

In E step of variational EM, we now need to evaluate four set of variational parameters, γ, ϕ, ξ and ρ. We use coordinate ascent, repeatedly optimizing the target function with respect to each parameter while holding the others fixed. Due to the simplification to the model in derivation, each coordinate can be optimized analytically.

Optimization with respect to γ. Since the forth term in Eq. 3 does not directly involve γ, the updating equation is the same as in LDA,

$$\gamma^{new} \leftarrow \alpha + \sum_{n=1}^{N}\phi_n \quad (7)$$

Optimization with respect to ρ_i. By letting the derivative of the target function to zero, we get ρ_i at,

$$\rho_i^{new} = \frac{1}{N}\sum_{n=1}^{N}\phi_n^T(\mu_i-\mu_y) \quad (8)$$

which is equal to $E_q[\bar{z}^T(\mu_i - \mu_y)]$, the expectation of $\parallel \mu_i - \bar{z} \parallel_2 - \parallel \mu_y - \bar{z} \parallel_2$. This result means that we make the Talyor expansion of e^x at the expectation point of x in Eq. 6. Notice that ρ_i^{new} is actually a minimum point of target function, but as the first order Talyor expansion of $-exp(x)$ is actually an upper bound instead of a desirable lower bound, a minimum point is what we need for better approximation, and please refer to Appendix for discussion of the approximation's tightness.

Optimization with respect to ξ. By letting derivative of target function to zero, we get a maximum of ξ

$$\xi^{new} = 1 + \sum_{i=1, i \neq y}^{C} e^{\rho_i} + e^{\rho_i}((\mu_i - \mu_y)^T \frac{1}{N} \sum_{n=1}^{N} \phi_n - \rho_i) \tag{9}$$

By replacing the solution of ρ_i into this equation, we find the solution of ξ is the approximated value of $E_q[1 + \sum_{i=1, i \neq y}^{C} \exp(\bar{z}^T(\mu_i - \mu_y))]$. Again, it means we apply Taylor expansion of $\log x$ at the expectation point of x in Eq. 6.

Optimization with respect to ϕ_{nj}. ϕ_{nj} is involved in two parts of log-likelihood: a generative term (generating word w_n) and a discriminative term (generating label y). By letting the derivative of log-likelihood with respect to ϕ_{nj} to zero, we get,

$$\phi_{nj}^{new} \propto \beta_{jw_n} \exp(\Psi(\gamma_j) - \Psi(\sum_{j'=1}^{T} \gamma_{j'}) - \frac{1}{\xi}(\sum_{i=1, i \neq y}^{C} e^{\rho_i}(\mu_i - \mu_y))) \tag{10}$$

This update is a critical difference between LogisticLDA and the original LDA. Compared with the updating equation of ϕ_n in LDA, this equation has an additional term, $-\frac{1}{\xi}(\sum_{i=1, i \neq y}^{C} e^{\rho_i}(\mu_i - \mu_y))$, by which supervisory information can regularize the embedding of a document. By applying ξ and ρ_i's solution to this equation, we can find out that the $\frac{e^{\rho_i}}{\xi}$ term is the approximated value of $E_q[\frac{p(\bar{z}|\mu_i, \Sigma^{-1})}{\sum_{j=1}^{C} p(\bar{z}|\mu_j, \Sigma^{-1})}]$, which is the probability that the embedding of current document is generated from category i. $\mu_i - \mu_y$ is the direction from class i to class y. Thus, the physical meanings of the additional term is easy to understand: ϕ_n is pushed away from other class's center according to the probability that it may be misclassified to that class. The larger probability it may be misclassified, the farther its embedding will be pushed away from the prototype of that class. Moreover, since the probability exponentially decreases with the increasing of L2 distance, a fact is that only samples near to the decision boundary will contribute to the estimation of ϕ_n. This behavior is similar as the way that SVM determines its decision boundary, in which decision boundary is determined only by support vectors (Burges, 1998).

3.2 M-step

In M-step, we optimize the three model parameters α,$\beta_{1:T}$, and $\mu_{1:C}$ by maximizing the approximation of log-likelihood.

Estimating topic distribution $\beta_{1:T}$. As β is not directly involved in the forth term of Eq.3, its update is the same as in LDA,

$$\beta_{iw}^{new} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dn}^{i} I(w_{dn} = w) \tag{11}$$

which is proportional to the probability of word w being assigned to topic i. Again, we need to normalize β_i to sum to one. As we have discussed that each ϕ_{dn} has been pushed to a "safe" position, the estimation of $\beta_{1:T}$ will not only take the word's generating probability into account, but also are aligned in a way that it is suitable for classification.

Estimating corpus's topic prior α. Optimization of α can take the same algorithm as in LDA. However, we adopt an easier to compute algorithm proposed by T.P. Minka (Minka, 2000) in this

paper. The derivative is computed by:

$$\frac{\partial L}{\partial \alpha_i} = D(\Psi(\sum_{j=1}^{T}\alpha_j) - \Psi(\alpha_j)) + \sum_{d=1}^{D}(\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{T}\gamma_{dj})) \tag{12}$$

Estimating class centers $\mu_{1:C}$. By maximizing the log-likelihood, we can get,

$$\mu_i^{new} \propto \sum_{d=1}^{D} I(y_d = i)\frac{1}{\xi_d}\sum_{j=1,j\neq i}^{C} e^{\rho_{dj}}E_q[\bar{z}_d] - I(y_d \neq i)\frac{1}{\xi_d}e^{\rho_{di}}E_q[\bar{z}_d] \tag{13}$$

Then we normalized μ_i to $\parallel \mu_i \parallel_2 = 1$, as we have chosen to embed class centers on the unit sphere (i.e. $\mu_i^T\mu_i = 1$). By applying the ρ_i and ξ's solution, we can get,

$$\mu_i^{new} \propto \sum_{d=1}^{D} I(y_d = i)E_q[p(y \neq i|\bar{z}_d, \mu_{1:C})]E_q[\bar{z}_d] - I(y_d \neq i)E_q[p(y = i|\bar{z}_d, \mu_{1:C})]E_q[\bar{z}_d] \tag{14}$$

The physical meanings of μ_i is obvious. The μ_i will move towards those false negative documents of class i, and move away from the false positive documents which are easily to misclassified to it. Those documents that lie in the "safe" zone (that is, not near to the decision boundary) are ignored. By focusing only on samples near decision boundaries, we can expect the generated class centers are suitable for classification. Indeed, they are always moved to try to improve the inter-class distances between classes which are likely to be misclassified.

3.3 Prediction

To predict the label of an unseen document, we wish to compute the expected probability of the given document belonging to a certain label i, and choose the one with maximum expected probability. The expected probability of a given document belongs to a certain label is given by,

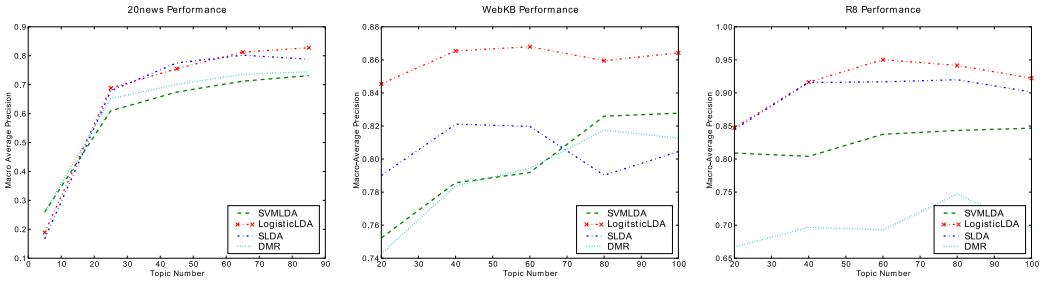
$$E_q[log(p(y = i|\bar{z}, \mu_{1:C}, \Sigma^{-1}))] = -\log(1 + \sum_{j=1,j\neq i}^{C}\exp(E_q[\bar{z}]^T(\mu_j - \mu_i))) \tag{15}$$

where $E_q[\bar{z}] = \frac{1}{N}\sum_{i=1}^{N}\phi_i$, and ϕ_i is estimated same as previous section, with exception that terms depends on y are removed. The resulting updating function is same as standard LDA (Blei et al., 2003).

4 Experiments

4.1 Datasets

We evaluated LogisticLDA on three benchmark datasets: 20 newsgroups, WebKB, and Reuters 21578 (Lewis et al., 2004). Because the numbers of document in different categories in Reuters 21578 are very skewed, we took two subsets of it proposed in (Cardoso-Cachopo and Oliveira, 2007). The first dataset consists of 8 categories, and the second dataset consists of 52 categories. Regarding data split, the 20news-group data-set comes with presplitted data. For other dataset, we use the split provided by (Cardoso-Cachopo and Oliveira, 2007). We applied simple pre-processing to documents: removing stopwords, Porter's stemming and removing words whose document frequencies(DF) are less than 3. Table 1 lists the details of them.



(a) 20 Newsgroups (b) Web KB (c) R8

Figure 2: Macro average precision(MAP) changes with the number of topics on three datasets

4.2 Experimental Setup

To evaluate the performance of LogisticLDA models, we chose five baseline approaches: Naive Bayes(NB) (Yang, 1999), SVM with linear kernel(SVM) (Joachims, 1998), performing dimension reduction by LDA and then classifying documents by SVM with Gaussian kernels(SVMLDA) (Blei et al., 2003), supervised LDA model(sLDA) (Blei and McAuliffe, 2007), Dirichlet-Multinomial Regression(DMR) (Mimno and McCallum, 2008). Although it was proposed a decade ago, SVM with linear kernel is still one of the state-of-the-art algorithms for text categorization. And we use TF-idf normalized feature for it. For other models we used bag of word representation. Because the implementation of sLDA which can directly support multi-class categorization is unclear, we decomposed the classification problem as C binary classification problems (i.e. one vs. rest). We use the 0-1 regression implementation proposed in (Blei and McAuliffe, 2007). It took more than two weeks to train sLDA classifiers. In contrast, it usually takes a day or two to train our model. For single label multi-class classification, *micro-average precision*(mip) and *macro-average precision*(map) are the two most popular performance measures (Yang, 1999), and we use them as benchmark.

Table 1: Details of datasets

	#Docs	#Words	#Train	#Categories
20NG	18821	34989	11293	20
WebKB	4199	7287	2803	4
R8	7674	7394	5485	8
R52	9100	8396	6532	52

4.3 Experimental Results

We conducted three experiments to evaluate the performance of LogisticLDA.

4.3.1 Influences of Number of Topics For topic models, the number of topics (T) is crucial. We performed experiments to evaluate appropriate numbers of topics. Figure 2 shows the macro-average precisions of topic model related approaches on the three datasets. From these curves we can see, with the increasing of number of topics, the performances of these approaches increase quickly, but drop after it reaches a point. If the number of topics is not big enough, the flexibility of a model is very limit, that is, we cannot get a good topic space to support classification. Once the number of topics grants enough flexibility to the model, but we continue increasing it, the model quickly becomes overfitting on training set. According to this figure, 40 to 80 is an approximate best range of number of topics for almost all approaches on the three datasets. In practice, we can determine the best number of topics by drawing such a figure. In following experiment, we will only report the performance of all topic model based algorithms under their best numbers of topics.

Table 2: Overall performance comparison

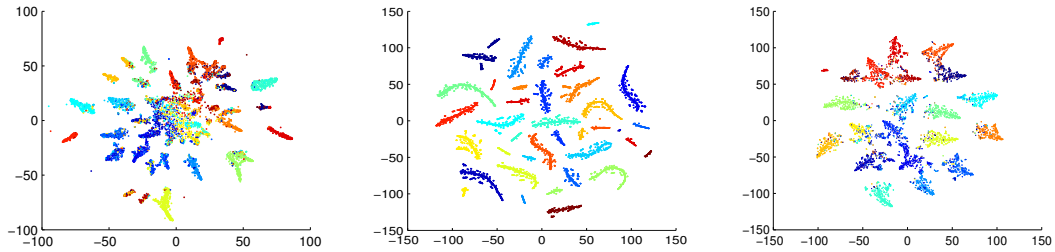
	20NG		WebKB		R8		R52	
	mip	map	mip	map	mip	map	mip	map
NB	0.8021	0.7764	0.8459	0.8286	0.9357	0.9149	0.8744	0.5366
SVM	0.8302	0.8278	0.8793	0.8625	0.9612	0.9329	0.8895	0.6033
SVMLDA	0.7404	0.7333	0.8286	0.827	0.9465	0.8462	0.8801	0.4318
sLDA	0.8048	0.8011	0.8358	0.8211	0.9222	0.919	0.8752	0.5907
DMR	0.7454	0.7444	0.8336	0.8176	0.9177	0.7469	0.7733	0.4043
LogisticLDA	0.8262	0.8245	0.8671	0.8724	0.9534	0.9513	0.9127	0.6164

4.3.2 Performance Comparison Table 2 lists performances of all approaches on the four datasets. We mark the best performances under each measure in pink color. Overall, the performances of SVM and LogisticLDA outperform other methods on all the datasets under both the two measures. SVM and LogisticLDA both get 4 best scores. The two measures, *mip* and *map*, can show different aspects of algorithms. *mip* is good at showing the overall performance of an algorithm, while *map* is a more balanced score with equal weight for every class. From Table 2, we can see SVM often get best results under *mip* measure, while LogisticLDA often get best results on *map* measure. This is a characteristic of LogisticLDA as a topic model. As for topic models, the classification task is carried out in the topic space, which is jointly generated by all the documents in the corpus, and small classes may leverage samples from large classes by filtering out noise through topic space. So the result of small classes may be improved.

4.3.3 Topic Space Regulation As the main goal of LogisticLDA is to regulate the topic space by label information of documents, it is important for us to examine the document's distribution in the topic space. For this purpose, we adopted the tSNE tool developed by van der Maaten and Hinton (2008). Fig.3 shows the scatter plot of 2D-embedding of topic distributions provided by tSNE on the 20 newsgroup dataset under topic number $T = 85$, in which each colored dot denotes a document and different colors denote different classes. Compared with standard LDA in Fig. 3(a), there are the noticeable margins between data points of different classes for LogisticLDA on training set in Fig.3(b). This is because when guided with class label, large margin between classes are the natural result of minimizing the exponential loss of logistic regression as in Eq.10. It is noted that the shape of each class is more like a strip instead of a cluster which is demanded by Gaussian distribution in Fig.3(b). This is because by incorporating label information in a discriminative way, the resulting model only requires that the data lie in the inverse direction of $\mu_i - \mu_y$ to increase inter-class distances, as the explanation to Eq.10, rather than close to the corresponding class center in all dimensions. In this way, the task of simultaneously seeking semantic topic space and discriminative embedding can be achieved. In Fig.3(c), we show the embedding results of LogisticLDA on test set, and we can see that most classes are still separated by large margins. This result demonstrates that via the regulating of ϕ_n, the label information has been propagated into topic distributions $\beta_{1:T}$ and make them being aligned in a way that are suitable for classification, and the generalization ability of LogisticLDA is satisfied.

5 Conclusion and Discussion

In this paper, we developed a supervised topic model, LogisticLDA, in which we mathematically integrated a generative model and a discriminative model in a principle way. By maximizing the posterior of document labels using logistic normal distributions, the model effectively incorporates the supervisory information to maximize inter-class distances in the topic space, while documents still enjoy the interchangeability characteristic for easy inference. Compared with discriminative methods, LogisticLDA can achieve comparable classification accuracy, but it can discover a latent



(a) LDA

(b) LogisticLDA's training set result

(c) LogisticLDA's result on test set

Figure 3: tSNE's 2D embeding of estimated topic distribution on 20NG dataset. Each color represents a class in the dataset with each dot corresponding to one document in the dataset

topic space, which is valuable for dimension reduction, topic mining and information retrieval. Good mathematical properties of logistic regression will definitely aspire extensions to LogisticLDA, e.g. multi-Gaussian formulation.

References

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Blei, David and Jon McAuliffe. 2007. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*.

Blei, David M. and John D. Lafferty. 2005. Correlated topic models. In *NIPS*.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3.

Burges, Christopher J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167.

Cardoso-Cachopo, Ana and Arlindo L. Oliveira. 2007. Semi-supervised single-label text categorization using centroid-based classifiers. In *SAC*.

Joachims, Thorsten. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *ECML*.

Lacoste-Julien, Simon, Fei Sha and Michael Jordan. 2008. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*.

Lewis, David D., Yiming Yang, Tony G. Rose and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5.

Li, Fei-Fei and Pietro Perona. 2005. A bayesian hierarchical model for learning natural scene categories. In *CVPR (2)*.

Mimno, David M. and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*.

Minka, Thomas P. 2000. Estimating a dirichlet distribution. Technical report, Microsoft.

Van der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9.

Wang, Xuerui and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven and Dimitrios Gunopulos, editors, *KDD*, pages 424–433. ACM. ISBN 1-59593-339-5.

Yang, Yiming. 1999. An evaluation of statistical approaches to text categorization. *Inf. Retr.*.