18

Discriminative Training of Mixed Membership Models

Jun Zhu

Department of Computer Science and Technology, State Key Laboratory of Intelligent Technology and Systems; Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China

Eric P. Xing

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

CONTENTS

Mixed membership models have shown great promise in analyzing genetics, text documents, and social network data. Unlike most existing likelihood-based approaches to learning mixed membership models, we present a discriminative training method based on the maximum margin principle to utilize supervising side information such as ratings or labels associated with documents to discover more predictive low-dimensional representations of the data. By using the linear expectation operator, we can derive efficient variational methods for posterior inference and parameter estimation. Empirical studies on the 20 Newsgroup dataset are provided. Our experimental results demonstrate qualitatively and quantitatively that the max-margin-based mixed membership model (topic model in particular for modeling text): 1) discovers sparse and highly discriminative topical representations; 2) achieves state-of-the-art prediction performance; and 3) is more efficient than existing supervised topic models.

18.1 Introduction

Mixed membership models are hierarchical extensions of finite mixture models where each data point exhibits multiple components. They have been successfully applied to analyze genetics (Pritchard et al., 2000), social networks (Airoldi et al., 2008), and text documents. For text analysis, probabilistic latent aspect models such as latent Dirichlet allocation (LDA) (Blei et al., 2003) have recently gained much popularity for stratifying a large collection of documents by projecting every document into a low-dimensional space spanned by a set of bases that capture the semantic aspects, also known as *topics*, of the collection. LDA posits that each document is an admixture of latent topics, of which each topic is represented as a unigram distribution over a given vocabulary. The document-specific admixture proportion vector $\boldsymbol{\theta}$ is modeled as a latent Dirichlet random variable, and can be regarded as a low-dimensional representation of the document in a topical space. This low-dimensional representation can be used for downstream tasks such as classification, clustering, or merely as a tool for structurally visualizing the otherwise unstructured document collection.

LDA is typically built on a discrete bag-of-words representation of input contents, which can be texts (Blei et al., 2003), images (Fei-Fei and Perona, 2005), or multi-type data (Blei and Jordan, 2003). However, in many practical applications, we can easily obtain useful side information besides the document or image contents. For example, when online users post their reviews for products or restaurants, they usually associate each review with a rating score or a thumbs-up/thumbs-down opinion; web sites or pages in the public Yahoo! Directory[1] can have their categorical labels; and images in the LabelMe (Russell et al., 2008) database are organized by a visual ontology and additionally each image is associated with a set of annotation tags. Furthermore, there is an increasing trend towards using online crowdsourcing services (such as Amazon Mechanical Turk[2]) to collect large collections of labeled data with a reasonably low price. Such side information often provides useful high-level or direct summarization of the content, but it is not directly utilized in the original LDA to influence topic inference. One would expect that incorporating such information into latent aspect modeling could guide a topic model towards discovering secondary (or non-dominant) but semantically more salient statistical patterns (Chechik and Tishby, 2002) that may be more interesting or relevant to the user's goal, such as making predictions on unlabeled data.

To explore this potential, developing new topic models that appropriately capture side information mentioned above has recently gained increasing attention. Representative attempts include the supervised topic model (sLDA) (Blei and McAuliffe, 2007), which captures real-valued document ratings as a regression response; multi-class sLDA (Wang et al., 2009), which directly captures discrete labels of documents as a classification response; and discriminative LDA (DiscLDA) (Lacoste-Julien et al., 2008), which also performs classification, but with a mechanism different from that of sLDA. All these models focus on the document-level side information such as document categories or review rating scores to supervise model learning. More variants of supervised topic models can be found in a number of applied domains, such as the aspect rating model (Titov and McDonald, 2008) for predicting ratings for each aspect of a hotel. In computer vision, various supervised topic models have been designed for understanding complex scene images (Sudderth et al., 2005; Fei-Fei and Perona, 2005).

It is worth pointing out that among existing supervised topic models for incorporating side information, there are two classes of approaches, namely, *downstream supervised topic models* (DSTM) and *upstream supervised topic models* (USTM). In a DSTM, the response variable is predicted based on the latent representation of the document, whereas in a USTM the response variable is being conditioned to generate the latent representation of the document. Examples of USTM

[1] See http://dir.yahoo.com/.

[2] See https://www.mturk.com/.

include DiscLDA and the scene understanding models (Sudderth et al., 2005; Fei-Fei and Perona, 2005), whereas sLDA is an example of DSTM. Another distinction between existing supervised topic models is the training criterion, or more precisely, the choice of objective function in the optimization-based learning. The sLDA models are trained by maximizing the *joint* likelihood of the content data (e.g., text or image) and the responses (e.g., labeling or rating), whereas DiscLDA models are trained by maximizing the *conditional* likelihood of the responses given contents.

In this chapter, we present maximum entropy discrimination latent Dirichlet allocation (MedLDA), a supervised topic model leveraging the maximum margin principle for making more effective use of side information during estimation of latent topical representations. Unlike existing supervised topic models mentioned above, MedLDA employs an arguably more discriminative max-margin learning technique within a probabilistic framework; and unlike the commonly adopted two-stage heuristic which first estimates a latent topic vector for each document using a topic model and then feeds them to another downstream prediction model, MedLDA integrates the mechanism behind max-margin prediction models (e.g., SVMs) with the mechanism behind hierarchical Bayesian topic models (e.g., LDA) under a unified constrained optimization framework. It employs a composite objective motivated by a tradeoff between two components—the negative log-likelihood of an underlying topic model which measures the goodness-of-fit for document contents, and a measure of prediction error on training data. It then seeks a regularized posterior distribution of the predictive function in a feasible space defined by a set of *expected* max-margin constraints generalized from the SVM-style margin constraints. Our proposed approach builds on earlier developments in maximum entropy discrimination (MED) (Jaakkola et al., 1999; Jebara, 2001) and partially observed maximum entropy discrimination Markov network (PoMEN) (Zhu et al., 2008). In MedLDA, because of the influence of both the likelihood function over content data and max-margin constraints induced by the side information, the discovery of latent topics is therefore coupled with the max-margin estimation of model parameters. This interplay can yield latent topical representations that are more discriminative and more suitable for supervised prediction tasks, as we demonstrate in the experimental section. We also present an efficient variational approach for inference under MedLDA, with a running time comparable to that of an unsupervised LDA and lower than other likelihood-based supervised LDAs. This advantage stems from the fact that MedLDA can directly optimize a margin-based loss instead of a likelihood-based one, and thereby avoids dealing with the normalization factor resultant from a full probabilistic generative formulation, which generally makes learning harder.

Finally, although we have focused on topic models, we emphasize that the methodology we develop is quite general and can be applied to perform max-margin learning for various mixed membership models, including the relational model (Airoldi et al., 2008). Moreover, the ideas can be extended to nonparametric Bayesian models (Zhu et al., 2011a; Zhu, 2012; Xu et al., 2012).

The rest of this chapter is structured as follows. Section 18.2 introduces the preliminaries that are needed to present MedLDA. Section 18.3 presents the MedLDA model for classification, together with an efficient algorithm. Section 18.4 presents empirical studies of MedLDA. Finally, Section 18.5 concludes this chapter with future research directions discussed.

18.2 Preliminaries

We begin with a brief overview of the fundamentals of mixed membership models, support vector machines, and maximum entropy discrimination (Jaakkola et al., 1999), which constitute the major building blocks of the proposed MedLDA.

18.2.1 Hierarchical Bayesian Mixed Membership Models

A general formulation of mixed membership models was presented in Erosheva et al. (2004), which characterizes these models in terms of assumptions at four levels: *population*, *subject*, *latent variable*, and *sampling scheme*. Population level assumptions describe the general structure of the population that is common to all subjects. Subject level assumptions specify the distribution of observed responses given individual membership scores. Latent variable level assumptions are about whether the membership scores are fixed or random. Finally, the last level of assumptions specify the number of distinct observed characteristics (attributes) and the number of replications for each characteristic.

(1) **Population Level**. Assume that there are K components or basis subpopulations in the populations of interest. For each subpopulation k, we denote by $f(x_{dn}|\beta_{kn})$ the probability distribution of the nth response variable for the dth subject, where $\boldsymbol{\beta}_k$ is an M-dimensional vector of parameters. Within a subpopulation, the observed responses are assumed to be independent across subjects and characteristics.

(2) **Subject Level**. For each subject d, a membership vector $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK})$ represents the degrees of the subject's membership to the various subpopulations. The distribution of the observed response x_{dn} for each subject given the membership scores $\boldsymbol{\theta}_d$ is then $p(x_{dn}|\boldsymbol{\theta}_d) = \sum_k \theta_{dk} f(x_{dn}|\beta_{kn})$. Conditional on the mixed membership scores, the response variables x_{dn} are independent of each other, and also independent across subjects.

(3) **Latent Variable Level**. With respect to the membership scores, one could assume they are either fixed unknown constants or random realizations from some underlying distribution. For Bayesian mixed membership models, which are our focus, the latter strategy is adopted, that is, assume that $\boldsymbol{\theta}_d$ are realizations of latent variables from some distribution $D_{\boldsymbol{\alpha}}$, parameterized by a vector $\boldsymbol{\alpha}$. The probability of observing x_{dn} is then $p(x_{dn}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \big(\sum_k \theta_{dk} f(x_{dn}|\beta_{kn}) \big) D_{\boldsymbol{\alpha}}(d\boldsymbol{\theta})$.

(4) **Sampling Scheme Level**. Suppose R independent replications of M distinct characteristics are observed for the dth subject. The conditional probability of observing $\mathbf{x}_d = \{x^r_{d1}, \ldots, x^r_{dM}\}_{r=1}^R$ given the parameters is then

$$p(\mathbf{x}_d|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \left(\prod_{n=1}^{M} \prod_{r=1}^{R} \sum_{k=1}^{K} \theta_{dk} f(x^r_{dn}|\beta_{kn}) \right) D_{\boldsymbol{\alpha}}(d\boldsymbol{\theta}). \tag{18.1}$$

Hierarchical Bayesian mixed membership models have been widely used in analyzing various forms of data, including discrete text documents (Blei et al., 2003), population genetics (Pritchard et al., 2000), social networks (Airoldi et al., 2008), and disability survey data (Erosheva, 2003). Below, we will study the mixed membership models for discrete text documents (i.e., topic models) as a test bed for mixed membership modeling ideas. But we emphasize that the methodology we will develop is applicable to a broad range of hierarchical Bayesian models.

18.2.2 Hierarchical Bayesian Topic Models

Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a Bayesian mixed membership model for modeling discrete text documents. In LDA, the components or subpopulations are *topics*, of which each topic is a multinomial distribution over the M words in a given vocabulary, i.e., $\boldsymbol{\beta}_k \in \mathcal{P}$, where $\mathcal{P}$ is the space of probability distributions with an appropriate dimension which will be omitted when the context is clear; and the membership scores $\boldsymbol{\theta}_d$ for document d is a mixing proportion vector over the K topics. We denote the vector of words appearing in document d as $\mathbf{w}_d = (w_{d1}, \ldots, w_{dN_d})$.

For the same word that appears for multiple times, there are multiple place holders in $\mathbf{w}_d$. Thus, $\mathbf{w}_d$ can be seen as a replication of appearing words. Let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1; \ldots; \boldsymbol{\beta}_K]$ denote the $K \times M$ matrix of topic parameters. Under LDA, the likelihood of a document corresponds to the following generative process:

1. For document d, draw a topic mixing proportion vector $\boldsymbol{\theta}_d$: $\boldsymbol{\theta}_d|\boldsymbol{\alpha} \sim \text{Dir}(\boldsymbol{\alpha})$;
2. For the nth word in document d, where $1 \leq n \leq N_d$,
   (a) Draw a topic assignment z_{dn} according to $\boldsymbol{\theta}_d$: $z_{dn}|\boldsymbol{\theta}_d \sim \text{Mult}(\boldsymbol{\theta}_d)$;
   (b) Draw the word w_{dn} according to z_{dn}: $w_{dn}|z_{dn}, \boldsymbol{\beta} \sim \text{Mult}(\boldsymbol{\beta}_{z_{dn}})$,

where z_{dn} is a K-dimensional indicator vector (i.e., only one element is 1; all others are 0), an instance of the topic assignment random variable Z_{dn}, and $\text{Dir}(\boldsymbol{\alpha})$ is a K-dimensional Dirichlet distribution, parameterized by $\boldsymbol{\alpha}$. With a little abuse of notations, we have used $\boldsymbol{\beta}_{z_{dn}}$ to denote the topic that is selected by the non-zero element of z_{dn}.

Let $\mathbf{z}_d = \{z_{dn}\}_{n=1}^{N_d}$ denote the set of topic assignments for all the words in document d. For a corpus $\mathcal{D}$ that contains D documents, we let $\boldsymbol{\Theta} = \{\theta_d\}_{d=1}^{D}$, $\mathbf{Z} = \{\mathbf{z}_d\}_{d=1}^{D}$, and $\mathbf{W} = \{\mathbf{w}_d\}_{d=1}^{D}$. According to the above generative process, an *unsupervised* LDA defines the joint distribution

$$p(\boldsymbol{\Theta}, \mathbf{Z}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \left(\prod_{n=1}^{N} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|z_{dn}, \boldsymbol{\beta}) \right). \tag{18.2}$$

For LDA, the learning task is to estimate the unknown parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Maximum likelihood estimation (MLE) is usually applied, which solves the problem

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \ \log p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}), \ \text{ s.t} : \boldsymbol{\beta}_k \in \mathcal{P}. \tag{18.3}$$

Once an LDA model is given (i.e., after learning), we can apply it to perform exploratory analysis for discovering underlying patterns. This task is done by deriving the posterior distribution using Bayes' rule, that is,

$$p(\boldsymbol{\Theta}, \mathbf{Z}|\mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\Theta}, \mathbf{Z}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})}. \tag{18.4}$$

Computationally, however, the likelihood $p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ is intractable to compute exactly. Therefore, approximate inference algorithms based on variational (Blei et al., 2003) or Markov chain Monte Carlo (MCMC) methods (Griffiths and Steyvers, 2004) have been widely used for parameter estimation and posterior inference under LDA.

Note that we have restricted ourselves to treat $\boldsymbol{\beta}$ as an unknown parameter, as done in Blei and McAuliffe (2007); Wang et al. (2009). Extension to a Bayesian treatment of $\boldsymbol{\beta}$ (i.e., by putting a prior over $\boldsymbol{\beta}$ and inferring its posterior) can be easily done in LDA as shown in the literature (Blei et al., 2003), where posterior inference is to find $p(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\beta}|\mathbf{W}, \boldsymbol{\alpha})$ by using Bayes' rule. As we shall see, MedLDA can also be easily extended to the full Bayesian setting under a general framework of regularized Bayesian inference.

The LDA described above does not utilize side information for learning topics and inferring topic vectors $\boldsymbol{\theta}$, which could limit their power for predictive tasks. To address this limitation, supervised topic models (sLDA) (Blei and McAuliffe, 2007) introduce a response variable Y to LDA for each document, as shown in Figure 18.1. For regression, where $y \in \mathbb{R}$, the generative process of sLDA is similar to LDA, but with an additional step—*Draw a response variable:* $y|\mathbf{z}_d, \boldsymbol{\eta}, \delta^2 \sim \mathcal{N}(\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d, \delta^2)$ *for each document* d—where $\bar{\mathbf{z}}_d = \frac{1}{N} \sum_n z_{dn}$ is the average topic assignment over all the words in document d; $\boldsymbol{\eta}$ is the regression weight vector; and δ^2 is a noise variance parameter. Then, the joint distribution of sLDA is

$$p(\boldsymbol{\Theta}, \mathbf{Z}, \mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2) = p(\boldsymbol{\Theta}, \mathbf{Z}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}) p(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta}, \delta^2), \tag{18.5}$$

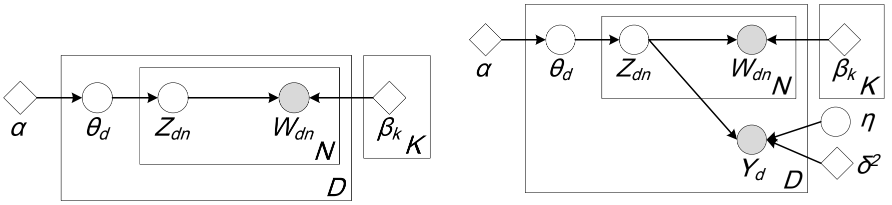


FIGURE 18.1
Graphical illustration of LDA (left) (Blei et al., 2003); and supervised LDA (right) (Blei and McAuliffe, 2007).

where $\mathbf{y} = \{y_d\}_{d=1}^D$ is the set of labels and $p(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta}, \delta^2) = \prod_d p(y_d|\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d, \delta^2)$ due to the model's conditional independence assumption. In this case, the likelihood is $p(\mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2)$ and that task of posterior inference is to find the posterior distribution $p(\boldsymbol{\Theta}, \mathbf{Z}|\mathbf{W}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2)$ by using Bayes' rule. Again, due to the intractability of the likelihood, variational methods were used to do approximate inference and MLE.

By changing the likelihood model of Y, sLDA can deal with various types of responses, such as discrete ones for classification (Wang et al., 2009) using the multi-class logistic regression

$$p(y|\mathbf{z}_d, \boldsymbol{\eta}) = \frac{\exp(\boldsymbol{\eta}_y^\top \bar{\mathbf{z}}_d)}{\sum_{y'} \exp(\boldsymbol{\eta}_{y'}^\top \bar{\mathbf{z}}_d)}, \tag{18.6}$$

where $\boldsymbol{\eta}_y$ is the vector of parameters associated with class y. However, posterior inference in an sLDA classification model can be more challenging than that in the sLDA regression model. This is because the non-Gaussian probability distribution in Equation (18.6) is highly nonlinear in $\boldsymbol{\eta}$ and $\mathbf{z}$, and its normalization factor can make the topic assignments of different words in the same document strongly coupled. If we perform fully Bayesian inference, the likelihood is non-conjugate with the commonly used priors, e.g., a Gaussian prior over $\boldsymbol{\eta}$, and this imposes further challenges on posterior inference. Variational methods were successfully used to approximate the normalization factor (Wang et al., 2009) in an EM algorithm, but they can be computationally expensive as we shall demonstrate in the experimental section.[3]

DiscLDA (Lacoste-Julien et al., 2008) is another supervised topic model for classification. DiscLDA is an upstream model, and the unknown parameter is the transformation matrix used to generate the document latent representations conditioned on class labels. This transformation matrix is learned by maximizing the conditional marginal likelihood of the text given class labels.

This progress notwithstanding, most current developments of supervised topic models have been built on a likelihood-driven probabilistic inference paradigm. In contrast, the max-margin-based techniques widely used in learning discriminative models (Vapnik, 1998; Taskar et al., 2003) have been rarely exploited to learn supervised topic models. Our work in Zhu et al. (2012) presents the first formulation of max-margin supervised topic models,[4] followed by various work on image

[3] For fully Bayesian sLDA, a Gibbs sampling algorithm was developed in Zhu et al. (2013) by exploring data augmentation techniques.

[4] A preliminary version was first published in 2009 (Zhu et al., 2009).

annotation (Yang et al., 2010), classification (Wang and Mori, 2011), and entity relationship extraction (Li et al., 2011). In this chapter, we present a novel formulation of MedLDA under the general framework of regularized Bayesian inference. Below, we briefly review the max-margin principle using the example of support vector machines.

18.2.3 Support Vector Machines

Depending on the nature of the response variable, the max-margin principle can be exploited in both classification and regression. Below we use document classification as an example to recapitulate the ideas behind SVMs, which we will shortly leverage to build our max-margin topic models.

Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_D, y_D)\}$ be a training set, where $\mathbf{x} \in \mathcal{X}$ are inputs such as document-feature vectors, and y are categorical response values taking values from a finite set $\mathcal{Y} = \{1, \cdots, L\}$. We consider the general multi-class classification where L is greater than 2. The goal of SVMs is to find a discriminant function $h(y, \mathbf{x}; \boldsymbol{\eta}) \in \mathcal{F}$ that could make accurate predictions with the argmax rule $\hat{y} = \arg\max_y h(y, \mathbf{x}; \boldsymbol{\eta})$. One common choice of the function family $\mathcal{F}$ is linear functions, that is, $h(y, \mathbf{x}; \boldsymbol{\eta}) = \boldsymbol{\eta}_y^\top \mathbf{f}(\mathbf{x})$, where $\mathbf{f} = (f_1, \cdots, f_I)^\top$ is a vector of feature functions $f_i : \mathcal{X} \to \mathbb{R}$, and $\boldsymbol{\eta}_y$ is the corresponding weight vector associated with class y. Formally, the linear SVM finds an optimal linear function by solving the following constrained optimization problem (Crammer and Singer, 2001):[5]

$$\begin{aligned} \min_{\boldsymbol{\eta}, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{\eta}\|_2^2 + C\sum_{d=1}^{D} \xi_d \\ \text{s.t.}: \quad & h(y_d, \mathbf{x}_d; \boldsymbol{\eta}) - h(y, \mathbf{x}_d; \boldsymbol{\eta}) \geq \ell_d(y) - \xi_d, \forall d, \forall y, \end{aligned} \tag{18.7}$$

where $\boldsymbol{\eta} = [\boldsymbol{\eta}_1^\top, \cdots, \boldsymbol{\eta}_L^\top]^\top$ is the concatenation of all subvectors; $\boldsymbol{\xi}$ are non-negative slack variables that tolerate some errors in the training data; C is a positive regularization constant; and $\ell_d(y)$ is a non-negative function that measures the cost of predicting y if the ground truth is y_d. It is typically assumed that $\ell_d(y_d) = 0$, i.e., no cost for correct predictions. The quadratic programming (QP) problem can be solved in a Lagrangian dual formulation. Samples with non-zero Lagrange multipliers are called support vectors.

18.2.4 Maximum Entropy Discrimination

The standard SVM formulation does not consider uncertainties of unknown variables, and it is thus far difficult to see how to incorporate the max-margin principle into Bayesian mixed membership models or topic models in particular. One significantly further step towards uniting the principles behind Bayesian generative modeling and max-margin learning is the maximum entropy discrimination (MED) formalism (Jebara, 2001), which learns a distribution of all possible classification models that belong to a particular parametric family, subject to a set of margin-based constraints. For instance, the MED classification model learns a distribution $q(\boldsymbol{\eta})$ through solving the following optimization problem:

$$\begin{aligned} \min_{q(\boldsymbol{\eta}) \in \mathcal{P}, \boldsymbol{\xi}} \quad & \mathrm{KL}(q(\boldsymbol{\eta}) \| p_0(\boldsymbol{\eta})) + C\sum_{d=1}^{D} \xi_d \\ \text{s.t.}: \quad & \mathbb{E}_q[h(y_d, \mathbf{x}_d; \boldsymbol{\eta})] - \mathbb{E}_q[h(y, \mathbf{x}_d; \boldsymbol{\eta})] \geq \ell_d(y) - \xi_d, \forall d, \forall y, \end{aligned} \tag{18.8}$$

where $p_0(\boldsymbol{\eta})$ is a prior distribution over the parameters, and $\mathrm{KL}(p\|q) \triangleq \mathbb{E}_p[\log(p/q)]$ is the Kullback-Leibler (KL) divergence.

[5] The formulation implies that $\xi_d \geq 0$, since all possible predictions including y_d are included in the constraints.

As studied in Jebara (2001), this MED problem leads to an entropic-regularized posterior distribution of the SVM coefficients, $q(\boldsymbol{\eta})$; and the resultant predictor $\hat{y} = \arg\max_y \mathbb{E}_{q(\boldsymbol{\eta})}[h(y, \mathbf{x}; \boldsymbol{\eta})]$ enjoys several nice properties and subsumes the standard SVM as special cases when the prior $p_0(\eta)$ is standard normal. Moreover, as shown in Zhu and Xing (2009) and Zhu et al. (2011b), with different choices of the prior over $\boldsymbol{\eta}$, such as a sparsity-inducing Laplace or a nonparametric Dirichlet process, the resultant $q(\boldsymbol{\eta})$ can exhibit a wide variety of characteristics and are suitable for diverse utilities such as feature selection or learning complex non-linear discriminating functions. Finally, the recent developments of the maximum entropy discrimination Markov network (MaxEnDNet) (Zhu and Xing, 2009) and partially observed MaxEnDNet (PoMEN) (Zhu et al., 2008) have extended the basic MED to the much broader scenarios of learning structured prediction functions with or without latent variables.

In applying the MED idea to learn a supervised topic model, a major difficulty is the presence of heterogeneous latent variables in the topic models, such as the topic vector $\boldsymbol{\theta}$ and topic indicator Z. In the sequel, we present a novel formalism called maximum entropy discrimination LDA (MedLDA) that extends the basic MED to make this possible, and at the same time discovers latent discriminating topics present in the study corpus based on available discriminant side information.

18.3 MedLDA: Max-Margin Supervised Topic Models

Now we present a new class of supervised topic models that explicitly employ labeling information in the context of document classification.[6] To make our methodology general, we formalize MedLDA under the framework of regularized Bayesian inference (Zhu et al., 2011a), which can in principle be applied to any Bayesian mixed membership models with a slight change of adding some posterior constraints to consider the supervising side information.

18.3.1 Bayesian Inference as a Learning Model

As shown in Equation (18.4), Bayesian inference can be seen as an information processing rule that projects the prior p_0 and empirical data to a posterior distribution via the Bayes' rule. Under this classic interpretation, a natural way to consider supervising information is to extend the likelihood model to incorporate it, as adopted in sLDA models.

A fresh interpretation of Bayesian inference was given by Zellner (1988), which provides a novel and more natural interpretation of MedLDA, as we shall see. Specifically, the posterior distribution by Bayes' rule is in fact the solution of an optimization problem. For instance, the posterior $p(\Theta, \mathbf{Z}|\mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ of LDA is equivalent to the optimum solution of

$$\min_{q(\Theta,\mathbf{Z})\in\mathcal{P}} \mathrm{KL}(q(\Theta, \mathbf{Z})\|p_0(\Theta, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta})) - \mathbb{E}_q[\log p(\mathbf{W}|\Theta, \mathbf{Z}, \boldsymbol{\beta})]. \tag{18.9}$$

We will use $\mathcal{L}_0(q(\Theta, \mathbf{Z}), \boldsymbol{\alpha}, \boldsymbol{\beta})$ to denote the objective function. In fact, we can show that the optimum objective value is the negative log-likelihood $-\log p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})$. Therefore, the MLE problem can be equivalently written in the variational form

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \big(\min_{q(\Theta,\mathbf{Z})\in\mathcal{P}} \mathcal{L}_0(q(\Theta, \mathbf{Z}), \boldsymbol{\alpha}, \boldsymbol{\beta})\big) = \min_{\boldsymbol{\alpha},\boldsymbol{\beta},q(\boldsymbol{\Theta},\mathbf{Z})\in\mathcal{P}} \mathcal{L}_0(q(\Theta, \mathbf{Z}), \boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{18.10}$$

which is the same as the objective of the EM algorithm (Blei et al., 2003) if no mean field assumptions are made. For the case where $\boldsymbol{\beta}$ is random, we have the same equality as above but with $\boldsymbol{\beta}$

[6] For regression, MedLDA can be developed as in Zhu et al. (2009).

moved from the set of unknown parameters into the distributions. For the fully Bayesian models (either treating $\boldsymbol{\alpha}$ as random too or leaving it pre-specified), we can solve an optimization problem similar as above to infer the posterior distribution.

18.3.2 Regularized Bayesian Inference

For the standard Bayesian inference, the posterior distribution is determined by a prior distribution and a likelihood model through the Bayes' rule. Either the prior or the likelihood model *indirectly* influences the behavior of the posterior distribution. However, under the above optimization formulation of Bayes' rule, we can have an additional channel of bringing in additional side information to *directly* regularize the properties of the desired posterior distributions. Let $\mathcal{M}$ be a model containing all the variables (e.g., $\boldsymbol{\Theta}$ and $\mathbf{Z}$ for LDA) whose posterior distributions we are trying to infer. Let $\mathcal{D}$ be the data (e.g., $\mathbf{W}$) whose likelihood model is defined, and let $\boldsymbol{\tau}$ be hyperparameters. One formal implementation of this idea is the *regularized Bayesian inference* as introduced in Zhu et al. (2011a), which solves the constrained optimization problem

$$\begin{aligned}\min_{q(\mathcal{M}),\boldsymbol{\xi}} \; & \mathrm{KL}(q(\mathcal{M})\|p_0(\mathcal{M}|\boldsymbol{\tau})) - \mathbb{E}_q[\log p(\mathcal{D}|\mathcal{M},\boldsymbol{\tau})] + U(\boldsymbol{\xi}) \\ \mathrm{s.t.}: \; & q(\mathcal{M}) \in \mathcal{P}_{\text{post}}(\boldsymbol{\xi}),\end{aligned} \tag{18.11}$$

where $\mathcal{P}_{\text{post}}(\boldsymbol{\xi})$ is a subspace of distributions that satisfy a set of constraints. We assume $\mathcal{P}_{\text{post}}(\boldsymbol{\xi})$ is non-empty for all $\boldsymbol{\xi}$. The auxiliary parameters $\boldsymbol{\xi}$ are usually nonnegative and interpreted as slack variables. $U(\boldsymbol{\xi})$ is a convex function, which usually corresponds to a surrogate loss (e.g., hinge loss) of a prediction rule, as we shall see. Under the above formulation, Zhu et al. (2011a) presented the infinite latent SVM models for classification and multi-task learning. Below, we present MedLDA as another instantiation of regularized Bayesian models.

18.3.3 MedLDA: A Regularized Bayesian Model

Let $\mathcal{D} = \{(\mathbf{w}_d, y_d)\}_{d=1}^{D}$ be a given fully-labeled training set, where the response variable Y takes values from the finite set $\mathcal{Y}$. MedLDA consists of two parts. The first part is an LDA likelihood model for describing input documents. We choose to use an unsupervised LDA, which defines a likelihood model for $\mathbf{W}$. The second part is a mechanism to consider supervising signal. Since our goal is to discover latent representations $\mathbf{Z}$ that are good for classification, one natural solution is to connect $\mathbf{Z}$ directly to our ultimate goal. MedLDA obtains such a goal by building a classification model on $\mathbf{Z}$. One good candidate of the classification model is the max-margin method which avoids defining a normalized likelihood model.

Formally, let $\boldsymbol{\eta}$ denote the parameters of the classification model. As in MED, we treat $\boldsymbol{\eta}$ as random variables and want to infer the joint posterior distribution $q(\boldsymbol{\eta}, \Theta, \mathbf{Z}|\mathcal{D}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, or $q(\boldsymbol{\eta}, \Theta, \mathbf{Z})$ for short. The classification model is defined as follows. If the latent topic representation $\mathbf{z}$ is given, MedLDA defines the linear discriminant function as

$$F(y, \boldsymbol{\eta}, \mathbf{z}; \mathbf{w}) = \boldsymbol{\eta}^\top \mathbf{f}(y, \bar{\mathbf{z}}), \tag{18.12}$$

where $\mathbf{f}(y, \bar{\mathbf{z}})$ is an LK-dimensional vector whose elements from $(y-1)K$ to yK are $\bar{\mathbf{z}}$ and all others are zero; and $\boldsymbol{\eta}$ is an LK-dimensional vector concatenating L class-specific sub-vectors. In order to predict on input data, MedLDA defines the *effective discriminant function* using the expectation operator

$$F(y; \mathbf{w}) = \mathbb{E}_{q(\boldsymbol{\eta},\mathbf{z})}[F(y, \boldsymbol{\eta}, \mathbf{z}; \mathbf{w})], \tag{18.13}$$

which is a linear functional of q.

With the above definitions, a natural prediction rule for a given posterior distribution q is

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} F(y; \mathbf{w}). \tag{18.14}$$

Then, we would like to "regularize" the properties of the latent topic representations to make them suitable for a classification task. Here, we adopt the framework of regularized Bayesian inference and impose the following max-margin constraints on the posterior distributions:

$$F(y_d; \mathbf{w}_d) - F(y; \mathbf{w}_d) \geq \ell_d(y),\ \forall y \in \mathcal{Y},\ \forall d. \tag{18.15}$$

That is, we want to find a "posterior distribution" that can predict correctly on all the training data using the prediction rule (18.14). However, in many cases, these hard constraints would be too strict. In order to learn a robust classifier for the datasets which are not separable, a natural generalization is to impose the soft max-margin constraints

$$F(y_d; \mathbf{w}_d) - F(y; \mathbf{w}_d) \geq \ell_d(y) - \xi_d,\ \forall y \in \mathcal{Y},\ \forall d, \tag{18.16}$$

where $\boldsymbol{\xi} = \{\xi_d\}$ are non-negative slack variables. Let

$$\mathcal{L}_1(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}), \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathrm{KL}(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}) || p_0(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta})) - \mathbb{E}_q[\log p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\beta})].$$

We define the soft-margin MedLDA model as solving

$$\begin{aligned} \min_{q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}) \in \mathcal{P}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}} \ & \mathcal{L}_1(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}), \boldsymbol{\alpha}, \boldsymbol{\beta}) + \frac{C}{D} \sum_{d=1}^{D} \xi_d \\ \text{s.t. :} \quad & \mathbb{E}_q[\boldsymbol{\eta}^\top \Delta \mathbf{f}(y, \bar{\mathbf{z}}_d)] \geq \ell_d(y) - \xi_d,\ \xi_d \geq 0, \forall d, \forall y, \end{aligned} \tag{18.17}$$

where the prior is $p_0(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p_0(\boldsymbol{\eta}) p_0(\boldsymbol{\Theta}, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta})$, and $\Delta \mathbf{f}(y, \bar{\mathbf{z}}_d) = \mathbf{f}(y_d, \bar{\mathbf{z}}_d) - \mathbf{f}(y, \bar{\mathbf{z}}_d)$. By removing slack variables, problem (18.17) can be equivalently written as

$$\min_{q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}) \in \mathcal{P}, \boldsymbol{\alpha}, \boldsymbol{\beta}} \mathcal{L}_1(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}), \boldsymbol{\alpha}, \boldsymbol{\beta}) + C\mathcal{R}(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z})), \tag{18.18}$$

where

$$\mathcal{R} = \frac{1}{D} \sum_d \arg\max_y \left(\ell_d(y) - \mathbb{E}_q[\boldsymbol{\eta}^\top \Delta \mathbf{f}(y, \bar{\mathbf{z}}_d)] \right)$$

is the hinge loss, an upper bound of the prediction error on training data.

Based on the equality in Equation (18.10), we can see the rationale underlying MedLDA, which is that we want to find latent topical representations $q(\boldsymbol{\Theta}, \mathbf{Z})$ and a model parameter distribution $q(\boldsymbol{\eta})$ which on one hand tends to predict as accurate as possible on training data, while on the other hand tends to explain the data well. The two parts are closely coupled by the expected margin constraints.

Although in theory we can use either sLDA (Wang et al., 2009) or LDA as a building block of MedLDA to discover latent topical representations, as we have discussed in Section 18.2.2, inference under sLDA could be harder and slower because the probability model of discrete Y in Equation (18.6) is nonlinear over $\boldsymbol{\eta}$ and Z, both of which are latent variables in our case, and its normalization factor strongly couples the topic assignments of different words in the same document. Therefore, we choose to use LDA that only models the likelihood of document contents $\mathbf{W}$ but not document label Y as the underlying topic model to discover latent representations Z. Even with this likelihood model, document labels can still influence topic learning and inference because they induce margin constraints pertinent to the topical distributions. As we shall see, the resultant MedLDA classification model can be efficiently learned by utilizing existing high-performance SVM solvers. Moreover, since the goal of max-margin learning is to directly minimize a hinge loss (i.e., an upper bound of the empirical loss), we do not need a normalized distribution model for response variables Y.

Note that we have taken a full expectation to define $F(y; \mathbf{w})$ instead of taking the mode as done in latent SVMs (Felzenszwalb et al., 2010; Yu and Joachims, 2009), because expectation is a nice linear functional of the distributions under which it is taken, whereas taking the mode involves the highly nonlinear $argmax$ function for discrete Z, which could lead to a harder inference task. Furthermore, due to the same reason to avoid dealing with a highly nonlinear discriminant function, we did not adopt the method in Jebara (2001) either, which uses log-likelihood ratios to define the discriminant function when considering latent variables in MED. Specifically, in our case, the max-margin constraints would be

$$\forall d,\ \forall y,\ \log \frac{p(y_d|\mathbf{w}_d, \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(y|\mathbf{w}_d, \boldsymbol{\alpha}, \boldsymbol{\beta})} \geq \ell_d(y) - \xi_d, \tag{18.19}$$

which are highly nonlinear due to the complex form of the marginal likelihood $p(y|\mathbf{w}_d, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Our linear expectation operator is an effective tool to deal with latent variables in the context of maximum margin learning. In fact, besides the present work, we have successfully applied this operator to other challenging settings of learning latent variable structured prediction models with nontrivial dependence structures among output variables (Zhu et al., 2008) and learning nonparametric Bayesian models (Zhu et al., 2011b;a).

18.3.4 Optimization Algorithm for MedLDA

Although we have used the simple linear expectation operator to define max-margin constraints, the problem of MedLDA is still intractable to directly solve due to the intractability of $\mathcal{L}_1$. Below, we present a coordinate descent algorithm with a further constraint on the feasible distribution $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z})$. Specifically, we impose the fully factorized mean field constraint that

$$q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}) = q(\boldsymbol{\eta}) \prod_{d=1}^{D} q(\boldsymbol{\theta}_d|\gamma_d) \prod_{n=1}^{N} q(z_{dn}|\boldsymbol{\phi}_{dn}), \tag{18.20}$$

where $\boldsymbol{\gamma}_d$ is a K-dimensional vector of Dirichlet parameters and each $\boldsymbol{\phi}_{dn}$ parameterizes a multinomial distribution over K topics. With this constraint, we have

$$F(y; \mathbf{w}_d) = \mathbb{E}_q[\boldsymbol{\eta}]^\top \mathbf{f}(y, \bar{\boldsymbol{\phi}}_d),$$

where $\bar{\boldsymbol{\phi}}_d = \mathbb{E}_q[\bar{\mathbf{z}}_d] = 1/N \sum_n \boldsymbol{\phi}_{dn}$; and the objective can be effectively evaluated since

$$\mathcal{L}_1(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}), \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathrm{KL}(q(\boldsymbol{\eta}) \| p_0(\boldsymbol{\eta})) + \mathcal{L}_0(q(\boldsymbol{\Theta}, \mathbf{Z}), \boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{18.21}$$

where $\mathcal{L}_0$ can be computed as in Blei et al. (2003). By considering the unconstrained formulation (18.18), our algorithm alternates between the following steps:

1. **Solve for** $q(\boldsymbol{\eta})$: When $q(\boldsymbol{\Theta}, \mathbf{Z})$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are fixed, the subproblem (in an equivalent constrained form) is to solve

$$\begin{aligned} \min_{q(\boldsymbol{\eta}) \in \mathcal{P}, \boldsymbol{\xi}} \ & \mathrm{KL}(q(\boldsymbol{\eta}) \| p_0(\boldsymbol{\eta})) + \frac{C}{D} \sum_{d=1}^{D} \xi_d \\ \text{s.t.} : \ & \mathbb{E}_q[\boldsymbol{\eta}]^\top \Delta \mathbf{f}(y, \bar{\boldsymbol{\phi}}_d) \geq \ell_d(y) - \xi_d, \forall d, \forall y. \end{aligned} \tag{18.22}$$

By using Lagrangian methods, we have the optimum solution

$$q(\boldsymbol{\eta}) = \frac{1}{\Psi} p_0(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^\top (\sum_d \sum_y \mu_d^y \Delta \mathbf{f}(y, \bar{\boldsymbol{\phi}}_d)) \right), \tag{18.23}$$

where the Lagrange multipliers $\boldsymbol{\mu}$ are the solution of the dual problem:

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & -\log \Psi + \sum_d \sum_y \mu_d^y \Delta \ell_d(y) \\ \text{s.t.}: \quad & \sum_y \mu_d^y \in \left[0, \frac{C}{D}\right], \forall d. \end{aligned} \tag{18.24}$$

We can choose different priors in MedLDA for various regularization effects. Here, we consider the normal prior. For the standard normal prior $p_0(\boldsymbol{\eta}) = \mathcal{N}(0, I)$, we can get: $q(\boldsymbol{\eta})$ is a normal with a shifted mean, i.e., $q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\lambda}, I)$, where $\boldsymbol{\lambda} = \sum_d \sum_y \mu_d^y \Delta \mathbf{f}(y, \bar{\boldsymbol{\phi}}_d)$, and the dual problem is

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & -\frac{1}{2}\| \sum_d \sum_y \mu_d^y \Delta \mathbf{f}(y, \bar{\boldsymbol{\phi}}_d) \|_2^2 + \sum_d \sum_y \mu_d^y \Delta \ell_d(y) \\ \text{s.t.}: \quad & \sum_y \mu_d^y \in \left[0, \frac{C}{D}\right], \forall d. \end{aligned} \tag{18.25}$$

The primal form of problem (18.25) is a multi-class SVM (Crammer and Singer, 2001):

$$\begin{aligned} \min_{\boldsymbol{\lambda}, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2 + \frac{C}{D} \sum_{d=1}^{D} \xi_d \\ \text{s.t.}: \quad & \boldsymbol{\lambda}^\top \mathbb{E}[\Delta \mathbf{f}_d(y)] \geq \Delta \ell_d(y) - \xi_d, \ \forall d, \ \forall y. \end{aligned} \tag{18.26}$$

We denote the optimum solution by $q^*(\boldsymbol{\eta})$ and its mean by $\boldsymbol{\lambda}^*$.

2. **Solve for $\boldsymbol{\phi}$ and γ**: By keeping $q(\boldsymbol{\eta})$ at its previous optimum solution and fixing $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, we have the subproblem as solving

$$\min_{\boldsymbol{\phi}, \boldsymbol{\gamma}} \ \mathcal{L}_0(q(\boldsymbol{\Theta}, \mathbf{Z}), \boldsymbol{\alpha}, \boldsymbol{\beta}) + \frac{C}{D} \sum_{d=1}^{D} \max_{y \in \mathcal{Y}} \left(\ell_d(y) - (\boldsymbol{\lambda}^*)^\top \Delta \mathbf{f}(y, \bar{\boldsymbol{\phi}}_d) \right). \tag{18.27}$$

Since q is fully factorized, we can perform the optimization on each document separately. We observe that the constraints in MedLDA are not dependent on $\boldsymbol{\gamma}$ and $q(\boldsymbol{\eta})$ is also not directly connected with $\boldsymbol{\gamma}$. Thus, optimizing $\mathcal{L}$ with respect to $\boldsymbol{\gamma}_d$ leads to the same update rule as in LDA:

$$\boldsymbol{\gamma}_d \leftarrow \boldsymbol{\alpha} + \sum_{n=1}^{N} \boldsymbol{\phi}_{dn}. \tag{18.28}$$

For $\boldsymbol{\phi}$, the constraints do affect its solution. Although in theory we can solve this subproblem using Lagrangian dual methods, it would be hard to derive the dual objective function (if possible at all). Here, we choose to update $\boldsymbol{\phi}$ using sub-gradient methods. Specifically, let $g(\boldsymbol{\phi}, \gamma)$ be the objective function of problem (18.27). The sub-gradient is

$$\frac{\partial g(\boldsymbol{\phi}, \boldsymbol{\gamma})}{\partial \boldsymbol{\phi}_{dn}} = \frac{\partial \mathcal{L}_0}{\partial \boldsymbol{\phi}_{dn}} + \frac{C}{ND}(\boldsymbol{\lambda}_{\bar{y}_d}^* - \boldsymbol{\lambda}_{y_d}^*), \tag{18.29}$$

where $\bar{y}_d = \arg\max_y (\ell_d(y) + (\boldsymbol{\lambda}^*)^\top \mathbf{f}(y, \bar{\boldsymbol{\phi}}_d))$ is the loss-augmented prediction. By setting the sub-gradient equal to zero, we can get

$$\boldsymbol{\phi}_{dn} \propto \exp\Big(\mathbb{E}[\log \boldsymbol{\theta}_d | \boldsymbol{\gamma}_d] + \log p(w_{dn}|\boldsymbol{\beta}) + \frac{C}{ND}(\boldsymbol{\lambda}_{y_d}^* - \boldsymbol{\lambda}_{\bar{y}_d}^*) \Big). \tag{18.30}$$

We can see that the first two terms in Equation (18.30) are the same as in unsupervised LDA (Blei et al., 2003), and the last term is due to the max-margin formulation of MedLDA and reflects our intuition that the discovered latent topical representation is influenced by the margin constraints. Specifically, for those examples that are misclassified (i.e., $\bar{y}_d \neq y_d$), the last term will not be zero, and it acts as a regularization term that biases the model towards discovering latent representations that tend to make more accurate prediction on these difficult examples. Moreover, this term is fixed for words in the document and thus will directly affect the latent representation of the document (i.e., $\boldsymbol{\gamma}_d$) and therefore leads to a discriminative latent representation. As we shall see in Section 18.4, such an estimate is more suitable for the classification task: for instance, MedLDA needs many fewer support vectors than the max-margin classifiers that are built on raw text or the topical representations discovered by LDA.

3. **Solve for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$**: The last substep is to solve for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with $q(\boldsymbol{\eta})$ and $q(\boldsymbol{\Theta}, \mathbf{Z})$ fixed. This subproblem is the same as the problem of estimating $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in LDA, since the constraints do not directly act on $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Therefore, we have the same update rules:

$$\beta_{kw} \propto \sum_d \sum_n \mathbb{I}(w_{dn} = w)\phi_{dn}^k, \tag{18.31}$$

where $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if the condition holds, 0 otherwise. For $\boldsymbol{\alpha}$, the same gradient descent algorithm as in Blei et al. (2003) can be applied to find a numerical solution.

The above formulation of MedLDA has a slack variable associated with each document. This is known as the *n-slack* formulation (Joachims et al., 2009). Another equivalent formulation, which can be more efficiently solved, is the so called *1-slack* formulation. The 1-slack MedLDA can be written as follows:

$$\begin{aligned} \min_{q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z}),\boldsymbol{\alpha},\boldsymbol{\beta},\xi} \quad & \mathcal{L}_1(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}), \boldsymbol{\alpha}, \boldsymbol{\beta}) + C\xi \\ \text{s.t.}: \quad & \frac{1}{D}\sum_d \mathbb{E}_q[\boldsymbol{\eta}^\top \Delta\mathbf{f}_d(\bar{y}_d)] \geq \frac{1}{D}\sum_d \Delta\ell_d(\bar{y}_d) - \xi, \forall(\bar{y}_1, \cdots, \bar{y}_D). \end{aligned} \tag{18.32}$$

By using the above alternating minimization algorithm and the cutting plane algorithm for solving the 1-slack as well as n-slack multi-class SVMs (Joachims et al., 2009), which is implemented in the SVM^{struct} package,[7] we can solve the 1-slack or n-slack MedLDA model efficiently, as we shall see in Section 18.4.3. SVM^{struct} provides the solutions of the primal parameters $\boldsymbol{\lambda}$ as well as the dual parameters $\boldsymbol{\mu}$, which are needed to do inference.

18.4 Experiments

In this section, we provide qualitative as well as quantitative evaluation of MedLDA on topic estimation and document classification. For MedLDA and other topic models (except DiscLDA, whose implementation details are explained in Footnote 12), we optimize the K-dimensional Dirichlet parameters $\boldsymbol{\alpha}$ using the Newton-Raphson method (Blei et al., 2003). For initialization, we set $\boldsymbol{\phi}$ to be uniform and each topic $\boldsymbol{\beta}_k$ to be a uniform distribution plus a very small random noise; we set

[7] See http://svmlight.joachims.org/svm_multiclass.html.

the posterior mean of $\boldsymbol{\eta}$ to be zero. We have released our implementation for public use.[8] In all the experimental results, we also report the standard deviation for a topic model with five randomly initialized runs.

18.4.1 Topic Estimation

We begin with an empirical assessment of topic estimation by MedLDA on the 20 Newsgroups dataset with a standard list of stopwords[9] removed. The dataset contains about 20,000 postings in 20 related categories. We compare this with unsupervised LDA.[10] We fit the dataset to a 110-topic MedLDA model, which exploits the supervised category information, and a 110-topic unsupervised LDA, which ignores category information.

Figure 18.2 shows the 2D embedding of the inferred topic proportions $\boldsymbol{\theta}$ by MedLDA and LDA using the t-SNE stochastic neighborhood embedding method (van der Maaten and Hinton, 2008), where each dot represents a document and each color-shape pair represents a category. Visually, the max-margin based MedLDA produces a good separation of the documents in different categories, while LDA does not produce a well-separated embedding, and documents in different categories tend to mix together. This is consistent with our expectation that MedLDA could produce a strong connection between latent topics and categories by doing supervised learning, while LDA ignores supervision and thus builds a weaker connection. Intuitively, a well-separated representation is more discriminative for document categorization. This is further empirically supported in Section 18.4.2. Note that a similar embedding was presented in Lacoste-Julien et al. (2008), where the transformation matrix in their model is pre-designed. The results of MedLDA in Figure 18.2 are *automatically* learned.

It is also interesting to examine the discovered topics and their relevance to class labels. In Figure 18.3a we show the top topics in four example categories as discovered by both MedLDA and LDA. Here, the semantic meaning of each topic is represented by the first ten high probability words.

To visually illustrate the discriminative power of the latent representations, i.e., the topic proportion vector $\boldsymbol{\theta}$ of documents, we illustrate and compare the per-class distribution over topics for each model at the right side of Figure 18.3a. This distribution is computed by averaging the expected topic vector of the documents in each class. We can see that MedLDA yields sharper, sparser, and fast decaying per-class distributions over topics. For the documents in different categories, we can see that their per-class average distributions over topics are very different, which suggests that the topical representations by MedLDA have a good discrimination power. Also, the sharper and sparser representations by MedLDA can result in a simpler max-margin classifier (e.g., with fewer support vectors), as we shall see in Section 18.4.2. All these observations suggest that the topical representations discovered by MedLDA have a better discriminative power and are more suitable for prediction tasks (see Section 18.4.2 for prediction performance). This behavior of MedLDA is in fact due to the regularization effect enforced over $\boldsymbol{\phi}$ as shown in Equation (18.30). On the other hand, the fully unsupervised LDA seems to discover topics that model the fine details of documents with no regard for their discrimination power (i.e., it discovers different variations of the same topic which results in a flat per-class distribution over topics). For instance, in the class *comp.graphics*, MedLDA mainly models documents using two salient, discriminative topics (T69 and T11), whereas LDA results in a much flatter distribution. Moreover, in the cases where LDA and MedLDA discover comparably the same set of topics in a given class (like *politics.mideast* and *misc.forsale*), MedLDA results in a sharper low-dimensional representation.

[8] See http://www.ml-thu.net/$\sim$jun/software.shtml.

[9] See http://mallet.cs.umass.edu/.

[10] We implemented LDA based on the public variational inference code by David Blei, using the same data structures as MedLDA for fair comparison.

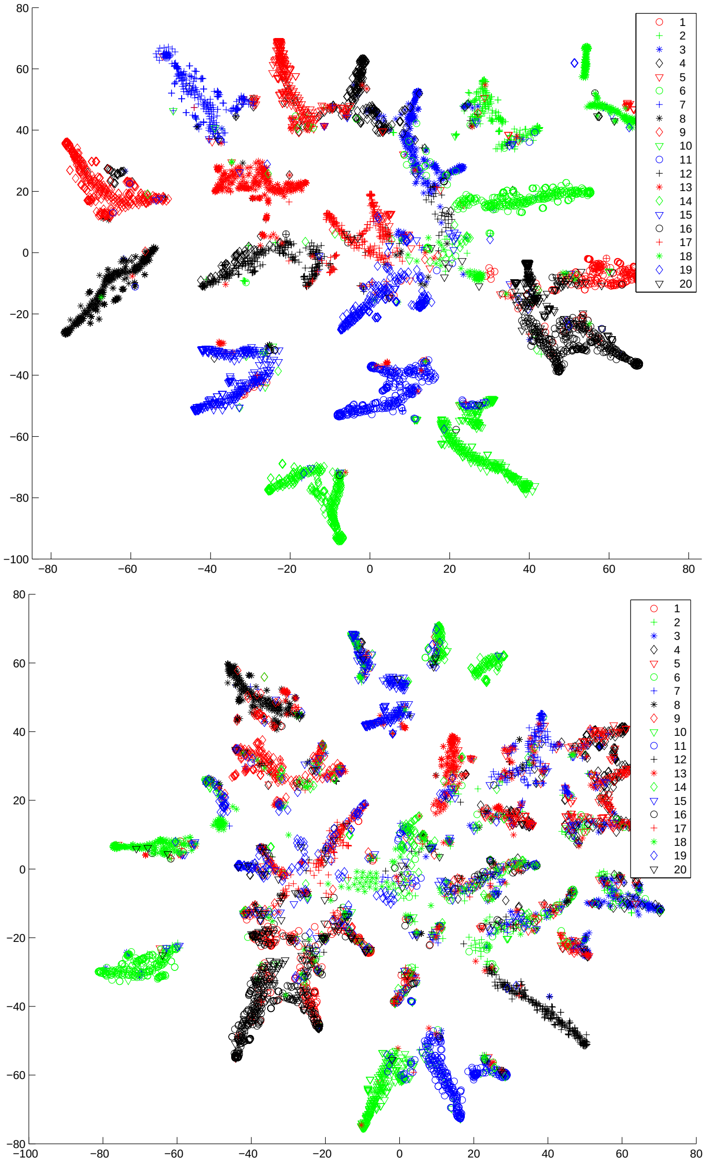
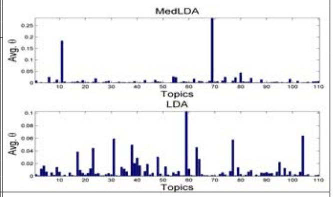
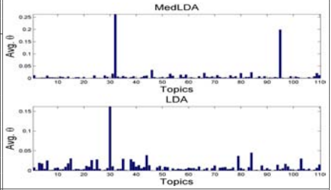
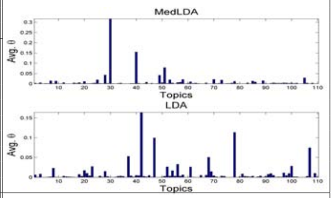
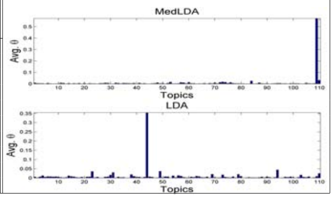
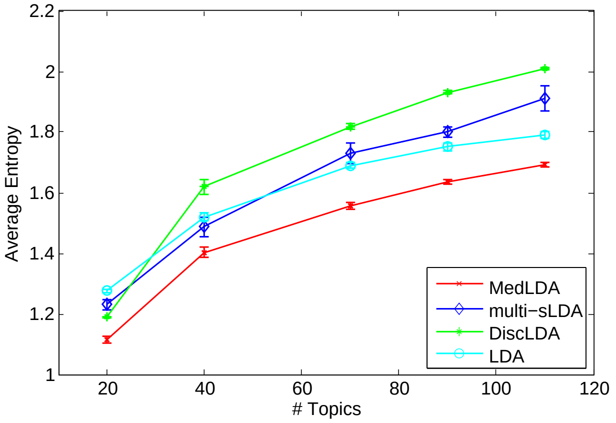


FIGURE 18.2
t-SNE 2D embedding of the topical representation by MedLDA (above) and unsupervised LDA (below). The mapping between each index and category name can be found in: http://people.csail.mit.edu/jrennie/20Newsgroups/.

Class	MedLDA			LDA			Average θ per class
comp.graphics	T 69	T 11	T 80	T 59	T 104	T 31	
	image	graphics	db	image	ftp	card	
	jpeg	image	key	jpeg	pub	monitor	
	gif	data	chip	color	graphics	dos	
	file	ftp	encryption	file	mail	video	
	color	software	clipper	gif	version	apple	
	files	pub	system	images	tar	windows	
	bit	mail	government	format	file	drivers	
	images	package	keys	bit	information	vga	
	format	fax	law	files	send	cards	
	program	images	escrow	display	server	graphics	
sci.electronics	T 32	T 95	T 46	T 30	T 84	T 44	
	ground	audio	source	power	water	sale	
	wire	output	rs	ground	energy	price	
	power	input	time	wire	air	offer	
	wiring	signal	john	circuit	nuclear	shipping	
	don	chip	cycle	supply	loop	sell	
	current	high	low	voltage	hot	interested	
	circuit	data	dixie	current	cold	mail	
	neutral	mhz	dog	wiring	cooling	condition	
	writes	time	weeks	signal	heat	email	
	work	good	face	cable	temperature	cd	
politics.mideast	T 30	T 40	T 51	T 42	T 78	T 47	
	israel	turkish	israel	israel	jews	armenian	
	israeli	armenian	lebanese	israeli	jewish	turkish	
	jews	armenians	israeli	peace	israel	armenians	
	arab	armenia	lebanon	writes	israeli	armenia	
	writes	people	people	article	arab	turks	
	people	turks	attacks	arab	people	genocide	
	article	greek	soldiers	war	arabs	russian	
	jewish	turkey	villages	lebanese	center	soviet	
	state	government	peace	lebanon	jew	people	
	rights	soviet	writes	people	nazi	muslim	
misc.forsale	T 109	T 110	T 84	T 44	T 94	T 49	
	sale	drive	mac	sale	don	drive	
	price	scsi	apple	price	mail	scsi	
	shipping	mb	monitor	offer	call	disk	
	offer	drives	bit	shipping	package	hard	
	mail	controller	mhz	sell	writes	mb	
	condition	disk	card	interested	send	drives	
	interested	ide	video	mail	number	ide	
	sell	hard	speed	condition	ve	controller	
	email	bus	memory	email	hotel	floppy	
	dos	system	system	cd	credit	system	

(a)



(b)

FIGURE 18.3
Top topics under each class as discovered by the MedLDA and LDA models (a). The average entropy of $\boldsymbol{\theta}$ over documents on 20 Newsgroups data (b).

A quantitative measure for the sparsity or sharpness of the distributions over topics is the entropy. We compute the entropy of the inferred topic proportion for each document and take the average over the corpus. Here, we compare MedLDA with LDA, sLDA for multi-class classification (multi-sLDA) (Wang et al., 2009),[11] and DiscLDA (Lacoste-Julien et al., 2008).[12] For DiscLDA, as in Lacoste-Julien et al. (2008), we fix the transformation matrix and set it to be diagonally sparse. We use the standard training/testing split[13] to fit the models on training data and infer the topic distributions on testing documents. Figure 18.3b shows the average entropy of different models on testing documents when different topic numbers are chosen. For DiscLDA, we set the class-specific topic number $K_0 = 1, 2, 3, 4, 5$ and correspondingly $K = 22, 44, 66, 88, 110$. We can see that MedLDA yields the smallest entropy, which indicates that the probability mass is concentrated on quite a few topics, consistent with the observations in Figure 18.3a. In contrast, for LDA the probability mass is more uniformly distributed on many topics (again consistent with Figure 18.3a), which results in a higher entropy. For DiscLDA, although the transformation matrix is designed to be diagonally sparse, the distributions over the class-specific topics and shared topics are flat. Therefore, the entropy is also high. Using automatically learned transition matrices might improve the sparsity of DiscLDA.

18.4.2 Prediction Accuracy

We perform binary and multi-class classification on the 20 Newsgroup dataset. To obtain a baseline, we first fit all the data to an LDA model, and then use the latent representation of the training[14] documents as features to build a binary or multi-class SVM classifier. We denote this baseline as *LDA+SVM*.

Binary Classification

As in Lacoste-Julien et al. (2008), the binary classification is to distinguish postings of the newsgroup *alt.atheism* and the postings of the group *talk.religion.misc*. The training set contains 856 documents with a split of 480/376 over the two categories, and the test set contains 569 documents with a split of 318/251 over the two categories. Therefore, the *naive baseline* that predicts the most frequent category for all test documents has accuracy 0.672.

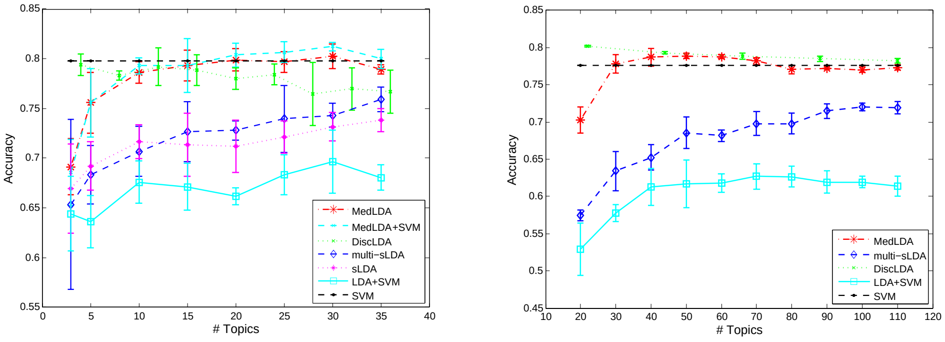
We compare the binary MedLDA with sLDA, DiscLDA, LDA+SVM, and the standard binary SVM built on raw text features. For supervised LDA, we use both the regression model (sLDA) (Blei and McAuliffe, 2007) and classification model (multi-sLDA) (Wang et al., 2009). For the sLDA regression model, we fit it using the binary representation (0/1) of the classes, and use a threshold 0.5 to make prediction. For MedLDA, to see whether a second-stage max-margin classifier can improve the performance, we also build a method of *MedLDA+SVM* similar to LDA+SVM. For DiscLDA, we fix the transition matrix. Automatically learning the transition matrix can yield slightly better results, as reported in Lacoste-Julien (2009). For all the above methods that utilize the class label information, they are fit *ONLY* on the training data.

[11] We thank the authors for providing their implementation, on which we made necessary slight modifications, e.g., improving the time efficiency and optimizing $\boldsymbol{\alpha}$.

[12] DiscLDA is a conditional model that uses class-specific topics and shared topics. Since the code is not publicly available, we implemented an in-house version by following the same strategy as in Lacoste-Julien et al. (2008) and share K_1 topics across classes and allocate K_0 topics to each class, where $K_1 = 2K_0$, and we varied $K_0 = \{1, 2, \cdots\}$. We should note here that Lacoste-Julien et al. (2008) and Lacoste-Julien (2009) gave an optimization algorithm for learning the topic structure (i.e., a transformation matrix), however, since the code is not available, we resorted to one of the fixed splitting strategies mentioned in the paper. Moreover, for the multi-class case, the authors only reported results using the same fixed splitting strategy we mentioned above. For the number of iterations for training and inference, we followed Lacoste-Julien (2009). Moreover, following Lacoste-Julien (2009) and personal communication with the first author, we used symmetric Dirichlet priors on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and set the Dirichlet parameters to 0.01 and $0.1/(K_0 + K_1)$, respectively.

[13] See http://people.csail.mit.edu/jrennie/20Newsgroups/.

[14] We use the training/testing split in: http://people.csail.mit.edu/jrennie/20Newsgroups/.



(a)

(b)

FIGURE 18.4
Classification accuracy of different models for: (a) binary and (b) multi-class classification on the 20 Newsgroup data.

We use the SVM-light (Joachims, 1999), which provides both primal and dual parameters, to build SVM classifiers and to estimate the posterior mean of $\boldsymbol{\eta}$ in MedLDA. The parameter C is chosen via 5-fold cross-validation during training from $\{k^2 : k = 1, \cdots, 8\}$. For each model, we run the experiments five times and take the average as the final results. The prediction accuracy of different models with respect to the number of topics is shown in Figure 18.4(a). For DiscLDA, we follow Lacoste-Julien et al. (2008) and set $K = 2K_0 + K_1$, where K_0 is the number of class-specific topics, K_1 is the number of shared topics, and $K_1 = 2K_0$. Here, we set $K_0 = 1, \cdots, 8, 10$.

We can see that the max-margin MedLDA outperforms the likelihood-based downstream models, including multi-sLDA, sLDA, and LDA+SVM. The best performances of the two discriminative models, MedLDA and DiscLDA, are comparable. However, MedLDA is easier to learn and faster in testing, as we shall see in Section 18.4.3. Moreover, the different approximate inference algorithms used in MedLDA (i.e., variational approximation) and DiscLDA (i.e., Monte Carlo sampling methods) can also make the performance different. We tried the collapsed variational inference (Teh et al., 2006) for MedLDA and it can give slightly better results. However, the collapsed variational method is computationally more expensive. Finally, since MedLDA already integrates the max-margin principle into its training, our conjecture is that the combination of MedLDA and SVM does not further improve the performance much on this task. We believe that the slight differences between MedLDA and MedLDA+SVM are due to the tuning of regularization parameters. For efficiency, we do not change the regularization constant C during training MedLDA. The performance of MedLDA would be improved if we selected a good C in different iterations because the data representation is changing.

Multi-class Classification

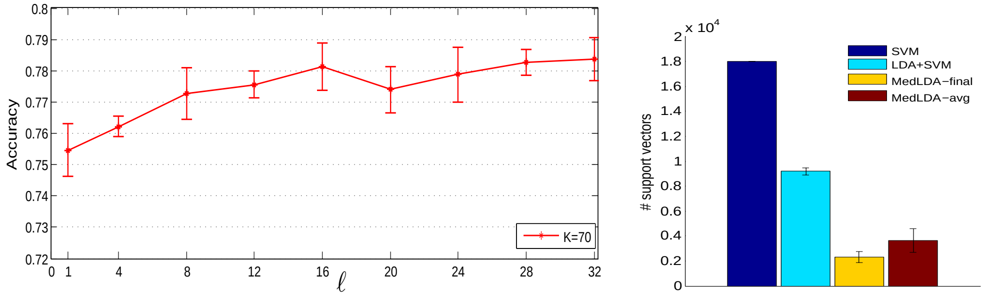
We perform multi-class classification on 20 Newsgroups with all the 20 categories. The dataset has a balanced distribution over the categories. For the test set, which contains 7,505 documents in total, the smallest category has 251 documents and the largest category has 399 documents. For the training set, which contains 11,269 documents, the smallest and the largest categories contain 376 and 599 documents, respectively. Therefore, the naive baseline that predicts the most frequent category for all the test documents has the classification accuracy 0.0532.

We compare MedLDA with LDA+SVM, multi-sLDA, DiscLDA, and the standard multi-class SVM built on raw text. We use the SVM^{struct} package with a cost function as $\Delta\ell_d(y) \triangleq \ell\mathbb{I}(y \neq y_d)$ to solve the sub-step of learning $q(\boldsymbol{\eta})$ and build the SVM classifiers for LDA+SVM. The parameter

ℓ is selected with 5-fold cross-validation. The average results, as well as standard deviations over 5 randomly initialized runs, are shown in Figure 18.4(b). For DiscLDA, we use the same equation as in Lacoste-Julien et al. (2008) to set the number of topics and set $K_0 = 1, \cdots, 5$. We can see that supervised topic models discover more predictive representations for classification, and the discriminative max-margin MedLDA and DiscLDA perform comparably, slightly better than the standard multi-class SVM (about 1.3 ± 0.3 percent improvement in accuracy). However, as we have stated and will show in Section 18.4.3, MedLDA is simpler to implement and faster in testing than DiscLDA. As we shall see shortly, MedLDA needs much fewer support vectors than standard SVM.

Figure 18.5(a) shows the classification accuracy on the 20 Newsgroups dataset for MedLDA with 70 topics. We show the results with ℓ manually set to $1, 4, 8, 12, \cdots, 32$. We can see that although the common $0/1$-cost works well for MedLDA, we can get better accuracy by using a larger cost to penalize wrong predictions. The performance is quite stable when ℓ is set to be larger than 8. The reason why ℓ affects the performance is that ℓ as well as C control: 1) the scale of the posterior mean of $\boldsymbol{\eta}$ and the Lagrangian multipliers $\boldsymbol{\mu}$, whose dot-product regularizes the topic mixing proportions in Equation (18.30); and 2) the goodness-of-fit of the MED large-margin classifier on the data. For practical reasons, we only try a small subset of candidate C values in parameter search, which can also influence the difference on performance in Figure 18.5(a). Performing very careful parameter search on C could possibly shrink the difference. Finally, for a small ℓ (e.g., 1 for the $0/1$-cost), we usually need a large C in order to obtain good performance. But, our empirical experience with SVM^{struct} shows that the multi-class SVM with a larger C (and smaller ℓ) is typically more expensive to train than the SVM with a larger ℓ (and smaller C). That is one reason why we choose to use a large ℓ.

Figure 18.5(b) shows the number of support vectors for MedLDA, LDA+SVM, and the multi-class SVM built on raw text features, which are high-dimensional ($\sim$60,000 dimensions for the 20 Newsgroup data) and sparse. Here we consider the traditional n-slack formulation of multi-class SVM and n-slack MedLDA using the SVM^{struct} package, where a support vector corresponds to a document-label pair. For MedLDA and LDA+SVM, we set $K = 70$. For MedLDA, we report both the number of support vectors at the final iteration and the average number of support vectors over all iterations. We can see that both MedLDA and LDA+SVM generally need many fewer support vectors than the standard SVM on raw text. The major reason is that both MedLDA and LDA+SVM use a much lower-dimensional and more compact representation for each document. Moreover, MedLDA needs (about 4 times) fewer support vectors than LDA+SVM. This could be because



(a) (b)

FIGURE 18.5
Sensitivity to the cost parameter ℓ for the MedLDA (a); and the number of support vectors for n-slack multi-class SVM, LDA+SVM, and n-slack MedLDA (b). For MedLDA, we show both the number of support vectors at the final iteration and the average number during training.

MedLDA makes use of both text contents and the supervising class labels in the training data, and its estimated topics tend to be more discriminative when being used to infer the latent topical representations of documents, i.e., using these latent representations by MedLDA, the documents in different categories are more likely to be well-separated, and therefore the max-margin classifier is simpler (i.e., needs fewer support vectors). This observation is consistent with what we have observed on the per-class distributions over topics in Figure 18.3a. Finally, we observe that about 32% of the support vectors in MedLDA are also the support vectors in multi-class SVM on the raw features.

18.4.3 Time Efficiency

Now, we report empirical results on time efficiency in training and testing. All the following results are achieved on a standard desktop with a 2.66GHz Intel processor. We implement all the models in C++ language.

Training Time

Figure 18.6 shows the average training time together with standard deviations on both binary and multi-class classification tasks with 5 randomly initialized runs. Here, we do not compare with DiscLDA because learning the transition matrix is not fully implemented in Lacoste-Julien (2009), but we will compare the testing time with it. From the results, we can see that for binary classification, MedLDA is more efficient than multi-class sLDA and is comparable with LDA+SVM. The slowness of multi-class sLDA is because the normalization factor in the distribution model of y strongly couples the topic assignments of different words in the same document. Therefore, the posterior inference is slower than that of LDA and MedLDA, which uses LDA as the underlying topic model. For the sLDA regression model, it takes even more training time due to the mismatch between its normal assumption and the non-Gaussian binary response variables, which prolongs the E-step.

For multi-class classification, the training time of MedLDA is mainly dependent on solving a multi-class SVM problem. Here, we implemented both 1-slack and n-slack versions of multi-class SVM (Joachims et al., 2009) for solving the sub-problem of estimating $q(\boldsymbol{\eta})$ and Lagrangian mul-

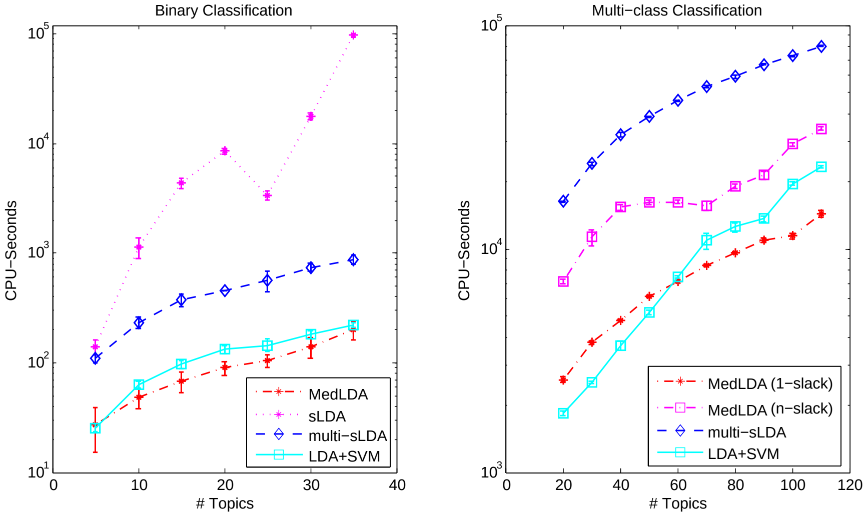
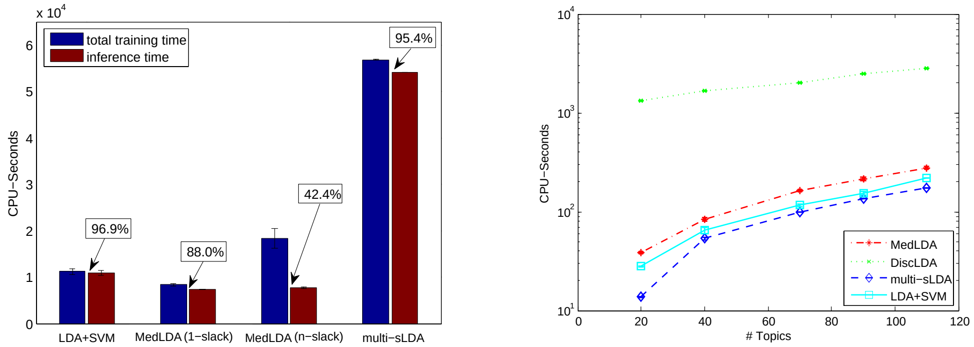


FIGURE 18.6
Training time (CPU seconds in log-scale) of different models for both binary (left) and multi-class classification (right).

tipliers in MedLDA. As we can see from Figure 18.6, the MedLDA with 1-slack SVM as the subsolver can be very efficient, comparable to unsupervised LDA+SVM. The MedLDA with n-slack SVM solvers is about three times slower. Similar to the binary case, for the multi-class supervised sLDA (Wang et al., 2009), because of the normalization factor in the category probability model (i.e., a softmax function), the posterior inference on different topic assignment variables (in the same document) is strongly correlated. Therefore, the inference is about ten times slower than that on LDA and MedLDA, which takes LDA as the underlying topic model.

We also show the time spent on inference and the ratio it takes over the total training time for different models in Figure 18.7(a). We can clearly see that the difference between 1-slack MedLDA and n-slack MedLDA is on the learning of SVMs. Both methods have similar inference time. We can also see that for LDA+SVM and multi-sLDA, more than 95% of the training time is spent on inference, which is very expensive for multi-sLDA. Note that LDA+SVM takes a longer inference time than MedLDA because we use more data (both training and testing) to learn unsupervised LDA.



(a)

(b)

FIGURE 18.7
The inference time and total training time for learning different models, as well as the ratio of inference time over total training time (a). For MedLDA, we consider both the 1-slack and n-slack formulations; for LDA+SVM, the SVM classifier is the fast 1-slack formulation; and (b) Testing time of different models with respect to the number of topics for multi-class classification.

Testing Time

Figure 18.7(b) shows the average testing time with standard deviation on the 20 Newsgroup testing data with five randomly initialized runs. We can see that MedLDA, multi-class sLDA, and unsupervised LDA are comparable in testing time, faster than that of DiscLDA. This is because all three models of MedLDA, multi-class sLDA, and LDA are *downstream* models (see the Introduction for definition). In testing, they do exactly the same tasks, i.e., inferring the overall latent topical representation and doing prediction with a linear model. Therefore, they have comparable testing time. However, DiscLDA is an *upstream* model, for which the inference to find the category-dependent latent topic representations is done multiple times. Therefore, in principle, the testing time of an upstream topic model is about $|\mathcal{C}|$ times slower than that of its downstream counterpart model, where $\mathcal{C}$ is the finite set of categories. The results in Figure 18.7(b) show that DiscLDA is roughly twenty times slower than other downstream models. Of course, the different inference algorithms can also make the testing time different.

18.5 Conclusions and Discussions

We have presented maximum entropy discrimination LDA (MedLDA), a supervised topic model that uses the discriminative max-margin principle to estimate model parameters such as topic distributions underlying a corpus, and infer latent topical vectors of documents. MedLDA integrates the max-margin principle into the process of topic learning and inference via optimizing one single objective function with a set of *expected* margin constraints. The objective function is a trade-off between the goodness-of-fit of an underlying topic model and the prediction accuracy of the resultant topic vectors in a max-margin classifier. We provide empirical evidence which appears to demonstrate that this integration could yield predictive topical representations that are suitable for prediction tasks, such as classification. Our results demonstrate that MedLDA is an attractive supervised topic model, which can achieve state-of-the-art performance for topic discovery and prediction accuracy while needing fewer support vectors than competing max-margin methods that are built on raw text or the topical representations discovered by unsupervised LDA.

The results of prediction accuracy on the 20 Newsgroups dataset show that MedLDA works slightly better than the SVM classifiers built on raw input features. These slight improvements tend to raise the question, “When and why should we choose MedLDA?” We have two possible answers:

1. MedLDA is a topic model. Besides predicting on unseen data, MedLDA can discover semantic patterns underlying complex data. In contrast, SVM models are more like black box machines which take raw input features and find good decision boundaries or regression curves, but that are incapable of discovering or considering hidden structures of complex data.[15] As an extension of SVM, MedLDA performs both exploratory analysis (i.e., topic discovery) and predictive tasks (e.g., classification) simultaneously. So, the first selection rule is that if we want to disclose some underlying patterns besides doing prediction, MedLDA should be preferred to SVM.

2. Even if our goal is prediction performance, MedLDA should also be considered as a competitive alternative. As shown in the synthetic experiments (Zhu et al., 2012) as well as the follow-up work (Yang et al., 2010; Wang and Mori, 2011; Li et al., 2011), depending on the data and problems, max-margin supervised topic models can outperform SVM models, or at least they are comparable if no gains are obtained. One reason that leads to our current results on 20 Newsgroups is that the fully factorized mean field assumption could be too restricted and lead to inaccurate estimates. In fact, we have tried more sophisticated inference methods such as collapsed variational inference (Teh et al., 2006) and collapsed Gibbs sampling,[16] both of which could lead to superior prediction performance.

Finally, MedLDA presents one of the first successful attempts, in the context of Bayesian mixed membership models (or topic models in particular), towards pushing forward the interface between max-margin learning and Bayesian generative modeling. As further demonstrated in others’ work (Yang et al., 2010; Wang and Mori, 2011; Li et al., 2011) as well as our recent work on regularized Bayesian inference (Chen et al., 2012; Zhu et al., 2011a;b; Zhu, 2012; Xu et al., 2012), the max-margin principle could be a fruitful addition to “regularize” the desired posterior distributions of Bayesian models for performing better prediction in a broad range of scenarios, such as image annotation/classification, multi-task learning, social link prediction, low-rank matrix factorization, etc. Of course, the flexibility on performing max-margin learning brings in new challenges. For example, the learning and inference problems of such models need to deal with some non-smooth

[15] Some strategies like sparse feature selection can be incorporated to make an SVM more interpretable in the original feature space, but this is beyond the scope of this discussion.

[16] Sampling methods for MedLDA can be developed by using Lagrangian dual methods. Details are reported in Jiang et al. (2012).

loss functions (e.g., the hinge loss in MedLDA), for which developing efficient algorithms for large-scale applications is a challenging research problem. Moreover, although we have good theoretical understandings of the generalization ability of max-margin methods without latent variables (e.g., SVMs), it is a challenging problem to provide theoretical guarantees for the generalization performance of max-margin models with latent variables.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* : 1981–2014.

Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’03)*. New York, NY, USA: ACM, 127–134.

Blei, D. M. and McAuliffe, J. (2007). Supervised topic models. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. (eds), *Advances in Neural Information Processing Systems 20*. Cambridge, MA: The MIT Press, 121–128.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* : 993–1022.

Chechik, G. and Tishby, N. (2002). Extracting relevant structures with side information. In Becker, S., Thrun, S., and Obermayer, K. (eds), *Advances in Neural Information Processing Systems 15*. Cambridge, MA: The MIT Press, 857–864.

Chen, N., Zhu, J., Sun, F., and Xing, E. P. (2012). Large-margin predictive latent subspace learning for multiview data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 34: 2365–2378.

Crammer, K. and Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* : 265–292.

Erosheva, E. A. (2003). Bayesian estimation of the Grade of Membership model. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds), *Bayesian Statistics 7*. New York, NY: Oxford University Press, 501–510.

Erosheva, E. A., Fienberg, S. E., and Lafferty, J. D. (2004). Mixed-membership models of scientific publications. *Proceedings of National Academy of Sciences* 101 : 5220–5227.

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 10th IEEE Computer Vision and Pattern Recognition (CVPR 2005)*. San Diego, CA, USA: IEEE Computer Society, 524–531.

Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32: 1627 – 1645.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* 101 Suppl 1: 5228–5235.

Jaakkola, T., Meilă, M., and Jebara, T. (1999). Maximum entropy discrimination. In Solla, S. A., Leen, T. K., and Müller, K. -R. (eds), *Advances in Neural Information Processing Systems 12*. Cambridge, MA: The MIT Press, 470–476.

Jebara, T. (2001). Discriminative, Generative and Imitative Learning. Ph.D. thesis, Media Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Jiang, Q., Zhu, J., Sun, M., and Xing, E. P. (2012). Monte Carlo methods for maximum margin supervised topic models. In Bartlett, P., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds), *Advances in Neural Information Processing Systems 25*. Red Hook, NY: Curran Associates, Inc., 1601–1609.

Joachims, T. (1999). Making large-scale SVM learning practical. In Schölkopf, B., Burges, C. J. C., and Smola, A. J. (eds), *Advances in Kernel Methods–Support Vector Learning*. Cambridge, MA: The MIT Press.

Joachims, T., Finley, T., and Yu, C. -N. (2009). Cutting-plane training of structural SVMs. *Machine Learning Journal* 77 : 27–59.

Lacoste-Julien, S. (2009). Discriminative Machine Learning with Structure. Ph.D. thesis, EECS Department, University of California, Berkeley, California, USA.

Lacoste-Julien, S., Sha, F., and Jordan, M. I. (2008). DiscLDA: Discriminative learning for dimensionality reduction and classification. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds), *Advances in Neural Information Processing Systems 21*. Red Hook, NY: Curran Assoiates, Inc., 897–904.

Li, D., Somasundaran, S., and Chakraborty, A. (2011). A combination of topic models with max-margin learning for relation detection. In *Proceedings of TextGraphs-6: Workshop on Graph-based Methods for Natural Language Processing (ACL-HLT 2011)*. The Association for Computational Linguistics, 1–9.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *American Journal of Human Genetics* 67: 170–181.

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision* 77: 157–173.

Sudderth, E. B., Torralba, A., Freeman, W., and Willsky, A. S. (2005). Learning hierarchical models of scenes, objects, and parts. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV 2005), Vol. 2*. Los Alamitos, CA, USA: IEEE Computer Society, 1331–1338.

Taskar, B., Guestrin, C., and Koller, D. (2003). Max-margin Markov networks. In Thrun, S., Saul, L. K., and Schölkopf, B. (eds), *Advances in Neural Information Processing Systems 16*. Cambridge, MA: The MIT Press, 25–32.

Teh, Y. W., Newman, D., and Welling, M. (2006). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In Schölkopf, B., Platt, J., and Hofmann, T. (eds), *Advances in Neural Information Processing Systems 19*. Red Hook, NY: Curran Associates, Inc., 1343–1350.

Titov, I. and McDonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-08)*. Columbus, OH, USA: Association for Computational Linguistics, 308–316.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*9 : 2579–2605.

Vapnik, V. (1998). *Statistical Learning Theory*. New York, NY: John Wiley & Sons.

Wang, C., Blei, D. M., and Fei-Fei, L. (2009). Simultaneous image classificationn and annotation. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*. Los Alamitos, CA, USA: IEEE Computer Society, 1903–1910.

Wang, Y. and Mori, G. (2011). Max-margin latent Dirichlet allocation for image classification and annotation. In *Proceedings of the 22nd British Machine Vision Conference (BMVC 2011)*. BMVA Press.

Xu, M., Zhu, J., and Zhang, B. (2012). Bayesian nonparametric maximum margin matrix factorization for collaborative prediction. In Bartlett, P., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds), *Advances in Neural Information Processing Systems 25*. Red Hook, NY: Curran Associates, Inc., 64–72.

Yang, S., Bian, J., and Zha, H. (2010). Hybrid generative/discriminative learning for automatic image annotation. In Grünwald, P. and Spirtes, P. (eds), *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*. Corvallis, OR, USA: AUAI Press, 683–690.

Yu, C. -N. and Joachims, T. (2009). Learning structural SVMs with latent variables. In Bottou, L. and Littman, L. (eds), *Proceedings of the 26th International Conference on Machine Learning (ICML '09)*. Omnipress, 1169–1176.

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *American Statistician* 42: 278–280.

Zhu, J. (2012). Max-margin nonparametric latent feature models for link prediction. In *Proceedings of the 29th International Conference on Machine Learning (ICML '12)*. Omnipress, 719–726.

Zhu, J., Ahmed, A., and Xing, E. P. (2009). MedLDA: Maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th International Conference on Machine Learning (ICML '09)*. Omnipress, 1257–1264.

Zhu, J., Ahmed, A., and Xing, E. P. (2012). MedLDA: Maximum margin supervised topic models. *Journal of Machine Learning Research (JMLR)* 13 : 2237–2278.

Zhu, J., Chen, N., and Xing, E. P. (2011a). Infinite latent SVM for classification and multi-task learning. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P., Pereira, F., and Weinberger, K. Q. (eds), *Advances in Neural Information Processing Systems 24*. Red Hook, NY: Curran Associates, Inc., 1620–1628.

Zhu, J., Chen, N., and Xing, E. P. (2011b). Infinite SVM: A Dirichlet process mixture of large-margin kernel machines. In *Proceedings of the 28th International Conference on Machine Learning (ICML '11)*. Omnipress, 617–624.

Zhu, J. and Xing, E. P. (2009). Maximum entropy discrimination Markov networks. *Journal of Machine Learning Research* 10 : 2531–2569.

Zhu, J., Xing, E. P., and Zhang, B. (2008). Partially observed maximum entropy discrimination Markov networks. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds), *Advances in Neural Information Processing Systems 21*. Red Hook, NY: Curran Associates, Inc., 1924–1931.

Zhu, J., Zheng, X., and Zhang, B. (2013). Improved Bayesian logistic supervised topic models with data augmentation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.