MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification

Jun Zhu[†*] JUN-ZHU@MAILS.TSINGHUA.EDU.CN
Amr Ahmed[†] AMAHMED@CS.CMU.EDU
Eric P. Xing[†] EPXING@CS.CMU.EDU
[*]Dept. of Comp. Sci & Tech, TNList Lab, Tsinghua University, Beijing 100084 China
[†]School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213 USA

Abstract

Supervised topic models utilize document's side information for discovering predictive low dimensional representations of documents; and existing models apply likelihood-based estimation. In this paper, we present a max-margin supervised topic model for both continuous and categorical response variables. Our approach, the maximum entropy discrimination latent Dirichlet allocation (MedLDA), utilizes the max-margin principle to train supervised topic models and estimate predictive topic representations that are arguably more suitable for prediction. We develop efficient variational methods for posterior inference and demonstrate qualitatively and quantitatively the advantages of MedLDA over likelihood-based topic models on movie review and 20 Newsgroups data sets.

1. Introduction

Statistical topic models have recently gained much popularity in managing a large collection of documents by discovering a low dimensional representation that captures the latent semantic of the collection. This low dimensional representation can then be used for tasks like classification and clustering or merely as a tool to structurally browse the otherwise unstructured collection. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is an example of such models for textual documents. LDA posits that each document is an admixture of latent topics where the topics are unigram distribution over a given vocabulary. The admixture proportion is document-specific and is distributed as a latent Dirichlet random variable.

When LDA is used for classification tasks, the document-specific mixing proportions are fed, usually, to a downstream classifier like an SVM. This two-step procedure is rather suboptimal as the side information of the documents, such as the category of a document or a numerical rating of a movie review, is not used in discovering the low-dimensional representation of the documents and thus can result in a sub-optimal representation for prediction. Developing a low dimensional representation that retains as much information as possible about the response variable has been studied in text modeling (McCallum et al., 2006) and image analysis (Blei & Jordan, 2003). Recently, supervised variants of LDA have been proposed, including the supervised LDA (sLDA) (Blei & McAuliffe, 2007) and the discriminative LDA (DiscLDA) for classification (Lacoste-Jullien et al., 2008). While sLDA and DiscLDA share the same goal (uncovering the latent structure in a document collection while retaining predictive power for supervised tasks), they differ in their training procedures. sLDA is trained by maximizing the joint likelihood of data and response variables while DiscLDA is trained to maximize the conditional likelihood of response variables.

In this paper, we propose a *max-margin discriminative* variant of supervised topic models for both regression and classification. In contrast to the above two-stage procedure of using topic models for prediction tasks, the proposed *maximum entropy discrimination latent Dirichlet allocation* (MedLDA) is an integration of max-margin learning and hierarchical Bayesian topic models by optimizing a single objective function with a set of *expected* margin constraints. MedLDA is a special instance of PoMEN (i.e., partially observed maximum entropy discrimination Markov network) (Zhu et al., 2008b), which was proposed to combine max-margin learning and structured hidden variables in Markov networks, for discovering latent topic presentations of documents. In MedLDA, the parameters for the regression or classification model are learned in

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

a max-margin sense; and the discovery of latent topics is coupled with the max-margin estimation of the model parameters. This interplay yields latent topic representations that are more suitable for supervised prediction tasks. We develop an efficient and easy-to-implement variational method for MedLDA, and in fact its running time is comparable to that of an unsupervised LDA for classification. This property stems from the fact that the MedLDA classification model directly optimizes the margin and does not suffer from a normalization factor which generally makes learning hard as in fully generative models such as sLDA.

The paper is structured as follows. Sec. 2 presents the MedLDA for both regression and classification, with efficient variational EM algorithms. Sec. 3 generalizes MedLDA to other latent variable topic models. Sec. 4 presents empirical comparison between MedLDA and likelihood-based topic models. Finally, Sec. 5 concludes this paper with future research directions.

2. Max-Entropy Discrimination LDA

In this section, we present the MedLDA model for both regression and classification. We first review the supervised topic models.

2.1. (Un)Supervised Topic Models

The unsupervised LDA (latent Dirichlet allocation) (Blei et al., 2003) is a hierarchical Bayesian model, where topic proportions for a document are drawn from a Dirichlet distribution and words in the document are repeatedly sampled from a topic which itself is drawn from those topic proportions. Supervised topic models (sLDA) (Blei & McAuliffe, 2007) introduce a response variable to LDA for each document, as illustrated in Figure 1.

Let K be the number of topics and M be the number of terms in a vocabulary. β denotes a $K \times M$ matrix and each β_k is a distribution over the M terms. For the regression problem, where the response variable $y \in \mathbb{R}$, the generative process of sLDA is as follows:

1. Draw topic proportions $\theta|\alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   (a) Draw a topic assignment $z_n|\theta \sim \mathrm{Mult}(\theta)$.
   (b) Draw a word $w_n|z_n, \beta \sim \mathrm{Multi}(\beta_{z_n})$.
3. Draw a response variable: $y|z_{1:N}, \eta, \delta^2 \sim N(\eta^\top \bar{z}, \delta^2)$, where $\bar{z} = 1/N \sum_{n=1}^N z_n$.

To estimate the unknown constants $(\alpha, \beta, \eta, \delta^2)$, sLDA maximizes the joint likelihood $p(\mathbf{y}, \mathbf{W}|\alpha, \beta, \eta, \delta^2)$, where $\mathbf{y}$ is the vector of response variables in a corpus $\mathcal{D}$ and $\mathbf{W}$ are all the words. Given a new document, the expected response value is the prediction:

$$E[Y|w_{1:N}, \alpha, \beta, \eta, \delta^2] = \eta^\top E[\bar{Z}|w_{1:N}, \alpha, \beta, \delta^2], \qquad (1)$$

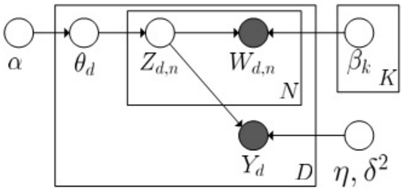


Figure 1. Supervised topic model (Blei & McAuliffe, 2007).

where $E[X]$ is an expectation w.r.t the posterior distribution of the r.v. X or its variational approximation.

DiscLDA (Lacoste-Jullien et al., 2008) is a discriminative variant of supervised topic models for classification, where the unknown parameters (i.e., a linear transformation matrix) are learned by maximizing the conditional likelihood of the response variables.

Below, we present a max-margin variant of the supervised topic models, which can discover predictive topic representations that are more suitable for supervised prediction tasks, e.g., regression and classification.

2.2. Learning MedLDA for Regression

Instead of learning a point estimate of η as in sLDA, we take a Bayesian-style approach and learn a distribution $q(\eta)$ in a max-margin manner. For prediction, we take the average over all the possible models:

$$E[Y|w_{1:N}, \alpha, \beta, \delta^2] = E[\eta^\top \bar{Z}|w_{1:N}, \alpha, \beta, \delta^2]. \qquad (2)$$

Now, the question underlying the averaging prediction rule (2) is how we can devise an appropriate loss function and constraints to integrate the max-margin concepts into latent topic discovery. In the sequel, we present the *maximum entropy discrimination latent Dirichlet allocation* (MedLDA) based on the PoMEN (i.e., partially observed maximum entropy discrimination Markov networks) (Zhu et al., 2008b) framework. PoMEN is an elegant combination of max-margin learning with structured hidden variables in Markov networks. The MedLDA is a special case of PoMEN to learn latent topic models to discover latent semantic structures of document collections.

For regression, the MedLDA is defined as an integration of a Bayesian sLDA, where the parameter η is sampled from a prior $p_0(\eta)$, and the ϵ-insensitive support vector regression (SVR) (Smola & Schölkopf, 2003). Thus, MedLDA defines a joint distribution: $p(\theta, \mathbf{z}, \eta, \mathbf{y}, \mathbf{W}|\alpha, \beta, \delta^2) = p_0(\eta) p(\theta, \mathbf{z}, \mathbf{y}, \mathbf{W}|\alpha, \beta, \eta, \delta^2)$, where the second term is the same as in the sLDA, that is, $p(\theta, \mathbf{z}, \mathbf{y}, \mathbf{W}|\alpha, \beta, \eta, \delta^2) = \prod_{d=1}^D p(\theta_d|\alpha) (\prod_{n=1}^N p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta)) p(y_d|\eta^\top \bar{z}_d, \delta^2)$. The marginal likelihood on $\mathcal{D}$ is $p(\mathbf{y}, \mathbf{W}|\alpha, \beta, \delta^2)$. Since directly optimizing the log marginal likelihood is intractable, as in sLDA, we optimize an upper bound $\mathcal{L}(q)$, where $q(\theta, \mathbf{z}, \eta|\gamma, \phi)$ is a variational distribution to approximate the posterior $p(\theta, \mathbf{z}, \eta|\alpha, \beta, \delta^2, \mathbf{y}, \mathbf{W})$.

Thus, the integrated learning problem is defined as:

$$\text{P1(MedLDA}^r\text{)}: \quad \min_{q,\alpha,\beta,\delta^2,\xi,\xi^\star} \; \mathcal{L}(q) + C\sum_{d=1}^{D}(\xi_d + \xi_d^\star)$$

$$\text{s.t. } \forall d: \quad \begin{cases} y_d - E[\eta^\top \bar{Z}_d] \leq \epsilon + \xi_d, \; \mu_d \\ -y_d + E[\eta^\top \bar{Z}_d] \leq \epsilon + \xi_d^\star, \; \mu_d^\star \\ \xi_d \geq 0, \; v_d \\ \xi_d^\star \geq 0, \; v_d^\star \end{cases}$$

where $\mu, \mu^\star, v, v^\star$ are lagrange multipliers; $\mathcal{L}(q) = -E[\log p(\theta, \mathbf{z}, \eta, \mathbf{y}, \mathbf{W}|\alpha, \beta, \delta^2)] - \mathcal{H}(q(\mathbf{z}, \theta, \eta))$; $\mathcal{H}(q)$ is the entropy of q; $\xi, \xi^\star$ are slack variables absorbing errors in training data; and ϵ is the precision.

The rationale underlying the MedLDA^r is that: let the current model be $p(\theta, \mathbf{z}, \eta, \mathbf{y}, \mathbf{W}|\alpha, \beta, \delta^2)$, then we want to find a latent topic representation and a model distribution (as represented by the distribution q) which on one hand tend to predict correctly on the data with a sufficient large margin, and on the other hand tend to explain the data well (i.e., minimizing an variational upper bound of the negative log-likelihood). This interplay will yield a topic representation that is more suitable for max-margin learning, as explained below.

2.2.1. Variational EM-Algorithm

The constrained problem P1 is generally intractable. Thus, we make additional independence assumptions about q. As in standard topic models, we assume that $q(\theta, \mathbf{z}, \eta|\gamma, \phi) = q(\eta)\prod_{d=1}^{D} q(\theta_d|\gamma_d)\prod_{n=1}^{N} q(z_{dn}|\phi_{dn})$, where γ_d is a K-dimensional vector of Dirichlet parameters and each ϕ_{dn} is a categorical distribution over K topics. Then, $E[Z_{dn}] = \phi_{dn}$, $E[\eta^\top \bar{Z}_d] = E[\eta]^\top(1/N)\sum_{n=1}^{N}\phi_{dn}$. We can develop an EM algorithm, which iteratively solves the following two steps:

E-step: infer the posterior distribution of the hidden variables (i.e., θ, $\mathbf{z}$, and η).

M-step: estimate the unknown parameters (i.e., α, β, and δ^2).

The essential difference between MedLDA and sLDA lies in the E-step to infer the posterior distribution of $\mathbf{z}$ and η because of the margin constraints in P1. As we shall see in Eq. (4), these constraints will bias the expected topic proportions towards the ones that are more suitable for max-margin learning. Since the constraints in P1 are not on the unknown parameters (α, β, and δ^2), the M-step is similar to that of the sLDA. We outline the algorithm in Alg. 1 and explain it in details below. Specifically, we formulate a Lagrangian[1] L for P1 and iteratively solve the following steps:

Algorithm 1 Variational MedLDA^r

```
Input: corpus $\mathcal{D} = \{(\mathbf{y}, \mathbf{W})\}$, constants $C$ and $\epsilon$, and topic number $K$.
Output: Dirichlet parameters $\gamma$, posterior distribution $q(\eta)$, parameters $\alpha$, $\beta$ and $\delta^2$.
repeat
  /**** E-Step ****/
  for $d = 1$ to $D$ do
    Update $\gamma_d$ as in Eq. (3).
    for $i = 1$ to $N$ do
      Update $\phi_{di}$ as in Eq. (4).
    end for
  end for
  Solve the dual problem D1 to get $q(\eta)$, $\mu$ and $\mu^\star$.
  /**** M-Step ****/
  Update $\beta$ using Eq. (5), and update $\delta^2$ using Eq. (6).
  $\alpha$ is fixed as $1/K$ times the ones vector.
until convergence
```

Optimize L over γ: Since the constraints in P1 are not on γ, we can get the same update formula as in sLDA for each document d separately:

$$\gamma_d \leftarrow \alpha + \sum_{n=1}^{N}\phi_{dn} \qquad (3)$$

Optimize L over ϕ: For each document d and each word i, by setting $\partial L/\partial \phi_{di} = 0$, we have:

$$\phi_{di} \propto \exp\Big(E[\log\theta|\gamma] + E[\log p(w_{di}|\beta)] + \frac{y_d}{N\delta^2}E[\eta] - \frac{2E[\eta^\top\phi_{d,-i}\eta] + E[\eta\circ\eta]}{2N^2\delta^2} + \frac{E[\eta]}{N}(\mu_d - \mu_d^\star)\Big), \qquad (4)$$

where $\phi_{d,-i} = \sum_{n\neq i}\phi_{dn}$ and the result of exponentiating a vector is a vector of the exponentials of its corresponding components. The first two terms in the exponential are the same as those in unsupervised LDA.

The essential differences of MedLDA^r from the sLDA lie in the last three terms in the exponential of ϕ_{di}. Firstly, the third and fourth terms are similar to those of sLDA, but in an expected version since we are learning the distribution $q(\eta)$. The second-order expectations $E[\eta^\top\phi_{d,-i}\eta]$ and $E[\eta\circ\eta]$ mean that the co-variances of η affect the distribution over topics. This makes our approach significantly different from a point estimation method, like sLDA, where no expectations or co-variances are involved in updating ϕ_{di}. Secondly, the last term is from the max-margin regression formulation. For a document d, which lies around the decision boundary, i.e., a support vector, either μ_d or $\mu_d^\star$ is non-zero, and the last term biases ϕ_{di} towards a distribution that favors a more accurate prediction on the document. Moreover, the last term is fixed for words in the document and thus will directly affect the latent representation of the document, i.e., γ_d. Therefore, the latent representation by MedLDA^r is more suitable for max-margin learning.

Optimize L over $q(\eta)$: Let A be the $D\times K$ matrix whose rows are the vectors $\bar{Z}_d^\top$. Set the partial

[1] $L = \mathcal{L}(q) + C\sum_{d=1}^{D}(\xi_d + \xi_d^\star) - \sum_{d=1}^{D}\mu_d(\epsilon + \xi_d - y_d + E[\eta^\top\bar{Z}_d]) - \sum_{d=1}^{D}(\mu_d^\star(\epsilon + \xi_d^\star + y_d - E[\eta^\top\bar{Z}_d]) + v_d\xi_d + v_d^\star\xi_d^\star) - \sum_{d=1}^{D}\sum_{i=1}^{N}c_{di}(\sum_{j=1}^{K}\phi_{dij} - 1)$, where the last term is due to the normalization condition $\sum_{j=1}^{K}\phi_{dij} = 1, \; \forall i, d$

derivative $\partial L/\partial q(\eta) = 0$, then we get:

$$q(\eta) = \frac{p_0(\eta)}{Z} \exp\big(\eta^\top \sum_{d=1}^{D} (\mu_d - \mu_d^\star + \frac{y_d}{\delta^2}) E[\bar{Z}_d] - \eta^\top \frac{E[A^\top A]}{2\delta^2} \eta\big)$$

where $E[A^\top A] = \sum_{d=1}^{D} E[\bar{Z}_d \bar{Z}_d^\top]$, and $E[\bar{Z}_d \bar{Z}_d^\top] = \frac{1}{N^2}(\sum_{n=1}^{N} \sum_{m \neq n} \phi_{dn}\phi_{dm}^\top + \sum_{n=1}^{N} \mathrm{diag}\{\phi_{dn}\})$. Plugging $q(\eta)$ into L, we get the dual problem of P1:

$$\begin{aligned} \mathrm{D1}: \ & \max_{\mu,\mu^\star} \ -\log Z - \epsilon \sum_{d=1}^{D} (\mu_d + \mu_d^\star) + \sum_{d=1}^{D} y_d(\mu_d - \mu_d^\star) \\ & \text{s.t. } \ \forall d: \ \mu_d, \mu_d^\star \in [0, C]. \end{aligned}$$

In MedLDA^r, we can choose different priors to introduce some regularization effects. For the standard normal prior: $p_0(\eta) = \mathcal{N}(0, I)$, the posterior is also a normal: $q(\eta) = \mathcal{N}(\mu_\eta, \Sigma)$, where $\mu_\eta = \Sigma\big(\sum_{d=1}^{D}(\mu_d - \mu_d^\star + \frac{y_d}{\delta^2})E[\bar{Z}_d]\big)$ is the mean and $\Sigma = (I + 1/\delta^2 E[A^\top A])^{-1}$ is a $K \times K$ co-variance matrix. Computation of Σ can be achieved robustly through Cholesky decomposition of $\delta^2 I + E[A^\top A]$, an $O(K^3)$ procedure. Another example is the Laplace prior, which can lead to a shrinkage effect (Zhu et al., 2008a) that is useful in sparse problems. In this paper, we focus on the normal prior.

For the standard normal prior, the dual problem D1 is a quadratic programming problem:

$$\begin{aligned} & \max_{\mu,\mu^\star} \ -\frac{1}{2} a^\top \Sigma a - \epsilon \sum_{d=1}^{D} (\mu_d + \mu_d^\star) + \sum_{d=1}^{D} y_d(\mu_d - \mu_d^\star) \\ & \text{s.t. } \ \forall d: \ \mu_d, \mu_d^\star \in [0, C], \end{aligned}$$

where $a = \sum_{d=1}^{D}(\mu_d - \mu_d^\star + \frac{y_d}{\delta^2})E[\bar{Z}_d]$. This problem can be solved with any standard QP solvers, although they may not be so efficient. To leverage recent developments in support vector regression, we note that the following primal form of D1 can be reformulated as a standard SVR problem and solved by using existing algorithms like SVM-light (Joachims, 1999) to get μ_η and the dual parameters μ and $\mu^\star$:

$$\min_{\mu_\eta, \xi, \xi^\star} \ \frac{1}{2}\mu_\eta^\top \Sigma^{-1} \mu_\eta - \mu_\eta^\top (\sum_{d=1}^{D} \frac{y_d}{\delta^2} E[\bar{Z}_d]) + C \sum_{d=1}^{D} (\xi_d + \xi_d^\star)$$

$$\text{s.t. } \forall d: \left\{ \begin{array}{r} y_d - \mu_\eta^\top E[\bar{Z}_d] \leq \epsilon + \xi_d, \mu_d \\ -y_d + \mu_\eta^\top E[\bar{Z}_d] \leq \epsilon + \xi_d^\star, \mu_d^\star \\ \xi_d, \geq 0, v_d \\ \xi_d^\star \geq 0, v_d^\star \end{array} \right.$$

Now, we estimate the unknown parameters α, β, and δ^2. Here, we assume α is fixed.

Optimize L over β. The update equations are the same as for sLDA:

$$\beta_{k,w} \propto \sum_{d=1}^{D} \sum_{n=1}^{N} 1(w_{dn} = w)\phi_{dnk}, \qquad (5)$$

Optimize L over δ^2. This step is similar to that of sLDA but in an expected version. The update rule is:

$$\delta^2 \leftarrow \frac{1}{D}\big(\mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top E[A]E[\eta] + E[\eta^\top E[A^\top A]\eta]\big), \quad (6)$$

where $E[\eta^\top E[A^\top A]\eta] = \mathrm{tr}(E[A^\top A]E[\eta\eta^\top])$.

2.3. Learning MedLDA for Classification

For classification, the response variables y are discrete. For brevity, we only consider the multi-class classification, where $y \in \{1, \cdots, M\}$. The binary case can be easily defined based on a binary SVM and the optimization problem can be solved similarly.

As we have stated, *fully* generative topic models, such as the sLDA, have a normalization factor, which can make the learning generally intractable, except for some special cases like the normal distribution as in the regression case. In (Blei & McAuliffe, 2007), variational methods or high-order Taylor expansion is applied to approximate the normalization factor of a GLM. In our max-margin formulation, since our target is to directly minimize a hinge loss, we do not need a fully generative model. Instead, we define a *partially* generative model on $(\theta, \mathbf{z}, \mathbf{W})$ only as in the unsupervised LDA, and for the classification model (i.e., from Z_d to Y_d), we apply the max-margin principle, which does not require a normalized distribution. Thus, in this case, the likelihood of the corpus $\mathcal{D}$ is $p(\mathbf{W}|\alpha, \beta)$.

Specifically, for classification, we assume the discriminant function F is linear, that is, $F(y, z_{1:N}, \eta) = \eta_y^\top \bar{z}$, where $\bar{z} = 1/N \sum_n z_n$ as in the regression model, η_y is a class-specific K-dimensional parameter vector associated with the class y and η is a MK-dimensional vector by stacking the elements of η_y. Equivalently, F can be written as $F(y, z_{1:N}, \eta) = \eta^\top \mathbf{f}(y, \bar{z})$, where $\mathbf{f}(y, \bar{z})$ is a feature vector whose components from $(y-1)K + 1$ to yK are those of the vector $\bar{z}$ and all the others are 0. From each single F, a prediction rule can be derived as in SVM. Here, we consider the general case to learn a distribution of $q(\eta)$ and for prediction, we take the average over all the possible models and the latent topics:

$$y^\star = \arg\max_y E[\eta^\top \mathbf{f}(y, \bar{Z})|\alpha, \beta]. \qquad (7)$$

Similar to the regression model, we define the integrated latent topic discovery and multi-class classification model as follows:

$$\mathrm{P2(MedLDA^c)}: \min_{q, q(\eta), \alpha, \beta, \xi} \mathcal{L}(q) + KL(q(\eta)||p_0(\eta)) + C\sum_{d=1}^{D} \xi_d$$

$$\text{s.t. } \forall d, \ y \neq y_d: \ E[\eta^\top \Delta\mathbf{f}_d(y)] \geq 1 - \xi_d; \ \xi_d \geq 0,$$

where $q(\theta, \mathbf{z}|\gamma, \phi)$ is a variational distribution; $\mathcal{L}(q) = -E[\log p(\theta, \mathbf{z}, \mathbf{W}|\alpha, \beta)] - \mathcal{H}(q(\theta, \mathbf{z}))$ is a variational upper bound of $-\log p(\mathbf{W}|\alpha, \beta)$; $\Delta\mathbf{f}_d(y) = \mathbf{f}(y_d, \bar{Z}_d) - \mathbf{f}(y, \bar{Z}_d)$, and ξ are slack variables. $E[\eta^\top \Delta\mathbf{f}_d(y)]$ is the "*expected* margin" by which the true label y_d is favored over a prediction y.

The rationale underlying the MedLDA^c is similar to that of the MedLDA^r, that is, we want to find a latent topic representation $q(\theta, \mathbf{z}|\gamma, \phi)$ and a parameter distribution $q(\eta)$ which on one hand tend to predict as accurate as possible on training data, while on the other hand tend to explain the data well. The KL-term in P2 is a regularizer of the distribution $q(\eta)$.

2.3.1. Variational EM-Algorithm

As in MedLDA^r, we can develop a similar variational EM algorithm. Specifically, we assume that q is fully factorized, as in the standard unsupervised LDA. Then, $E[\eta^\top \mathbf{f}(y, \bar{Z}_d)] = E[\eta]^\top \mathbf{f}(y, 1/N \sum_{n=1}^N \phi_{dn})$. We formulate the Lagrangian[2] L of P2 and iteratively optimize L w.r.t γ, ϕ, $q(\eta)$ and β. Since the constraints in P2 are not on γ or β, their update rules are the same as in MedLDA^r and we omit the details here. We explain the optimization of P2 over ϕ and $q(\eta)$ and show the insights of the max-margin topic model:

Optimize L over ϕ. Again, since q is fully factorized, we can perform the optimization on each document separately. Set $\partial L / \partial \phi_{di} = 0$, then we have:

$$\phi_{di} \propto \exp\big(\ E[\log\theta|\gamma] + E[\log p(w_{di}|\beta)] + \frac{1}{N} \sum_{y \neq y_d} \mu_d(y) E[\eta_{y_d} - \eta_y]\big). \tag{8}$$

The first two terms in Eq. (8) are the same as in the unsupervised LDA and the last term is due to the max-margin formulation of P2 and reflects our intuition that the discovered latent topic representation is influenced by the max-margin estimation. For those examples that are around the decision boundary, i.e., support vectors, some of the lagrange multipliers are non-zero and thus the last term acts as a regularizer that biases the model towards discovering a latent representation that tends 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., γ_d) and will yield a discriminative latent representation, as we shall see in Section 4, which is more suitable for the classification task.

Optimize L over $q(\eta)$: As in the regression model, we get the dual problem of P2:

$$\begin{aligned} \mathrm{D2}: \quad & \max_{\mu} \ -\log Z + \sum_{d=1}^{D} \sum_{y \neq y_d} \mu_d(y) \\ & \text{s.t. } \ \forall d: \ \sum_{y \neq y_d} \mu_d(y) \in [0, C], \end{aligned}$$

and the posterior $q(\eta) = \frac{1}{Z} p_0(\eta) \exp(\eta^\top \mu_\eta)$, where $\mu_\eta = \sum_{d=1}^D \sum_{y \neq y_d} \mu_d(y) E[\Delta \mathbf{f}_d(y)]$.

Again, we can choose different priors in MedLDA^c for different regularization effects. We consider the normal prior in this paper. For the standard normal prior $p_0(\eta) = \mathcal{N}(0, I)$, we can get: $q(\eta)$ is a normal with a shifted mean, i.e., $q(\eta) = \mathcal{N}(\mu_\eta, I)$, and the dual problem D2 is the same as the dual problem of a standard multi-class SVM that can be solved using existing SVM methods (Crammer & Singer, 2001) :

$$\begin{aligned} & \max_{\mu} \ -\frac{1}{2} \| \sum_{d=1}^{D} \sum_{y \neq y_d} \mu_d(y) E[\Delta \mathbf{f}_d(y)] \|_2^2 + \sum_{d=1}^{D} \sum_{y \neq y_d} \mu_d(y) \\ & \text{s.t. } \ \forall d: \ \sum_{y \neq y_d} \mu_d(y) \in [0, C]. \end{aligned}$$

3. MedTM: a general framework

We have presented MedLDA, which integrates the max-margin principle with an *underlying* LDA model, which can be supervised or unsupervised, for discovering predictive latent topic representations of documents. The same principle can be applied to other generative topic models, such as the correlated topic models (CTM) (Blei & Lafferty, 2005), as well as undirected random fields, such as the exponential family harmoniums (EFH) (Welling et al., 2004).

Formally, the max-entropy discrimination topic models (MedTM) can be generally defined as:

$$\begin{aligned} \mathrm{P(MedTM)}: \ & \min_{q(H), q(\Upsilon), \Psi, \xi} \mathcal{L}(q(H)) + KL(q(\Upsilon) \| p_0(\Upsilon)) + U(\xi) \\ & \text{s.t. } \textit{expected} \text{ margin constraints,} \end{aligned}$$

where H are hidden variables (e.g., $(\theta, \mathbf{z})$ in LDA); Υ are the parameters of the model pertaining to the prediction task (e.g., η in sLDA); Ψ are the parameters of the underlying topic model (e.g., the Dirichlet parameter α); and $\mathcal{L}$ is a variational upper bound of the negative log likelihood associated with the underlying topic model. U is a convex function over slack variables. For the general MedTM model, we can develop a similar variational EM-algorithm as for the MedLDA. Note that Υ can be a part of H. For example, the underlying topic model of MedLDA^r is a Bayesian sLDA. In this case, $H = (\theta, \mathbf{z}, \eta)$, $\Upsilon = \emptyset$ and the term $KL(q(\eta) \| p_0(\eta))$ is contained in its $\mathcal{L}$.

4. Experiments

In this section, we provide qualitative as well as quantitative evaluation of MedLDA on text modeling, classification and regression.

4.1. Text Modeling

We study text modeling of the MedLDA on the 20 Newsgroups data set with a standard list of stop words[3] removed. The data set contains postings in 20 related categories. We compare with the standard unsupervised LDA. We fit the dataset to a 110-topic MedLDA^c model, which explores the supervised category information, and a 110-topic unsupervised LDA.

[2] $L = \mathcal{L}(q) + KL(q(\eta) \| p_0(\eta)) + C \sum_{d=1}^D \xi_d - \sum_{d=1}^D v_d \xi_d - \sum_{d=1}^D \sum_{y \neq y_d} \mu_d(y) (E[\eta^\top \Delta \mathbf{f}_d(y)] + \xi_d - 1) - \sum_{d=1}^D \sum_{i=1}^N c_{di} (\sum_{j=1}^K \phi_{dij} - 1)$, where the last term is from the normalization condition $\sum_{j=1}^K \phi_{dij} = 1, \ \forall i, d$.

[3] http://mallet.cs.umass.edu/

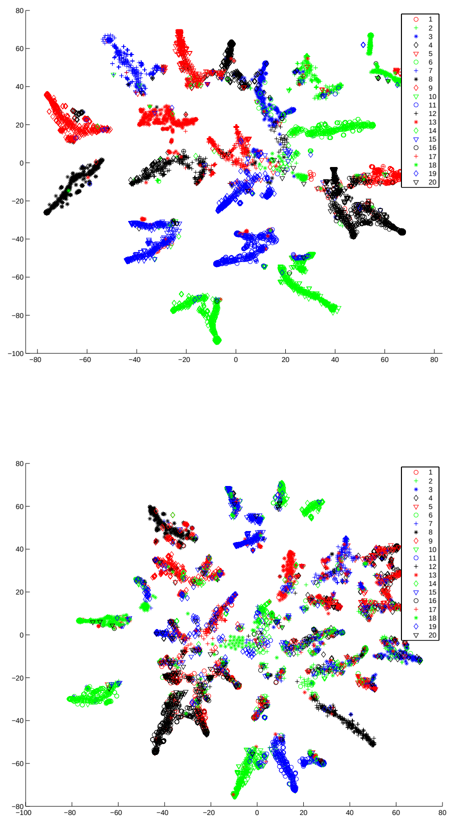


Figure 2. t-SNE 2D embedding of the topic representation by: MedLDA^c (above) and the unsupervised LDA (below).

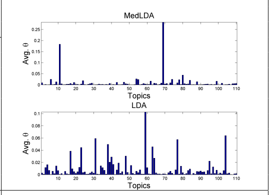
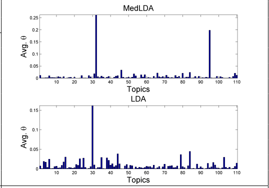
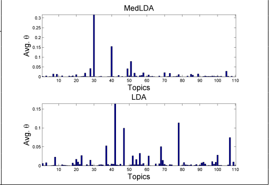
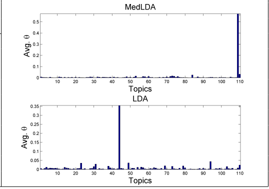
Class	MedLDA			LDA			Average θ per class
comp.graphics	T 69 image jpeg gif file color files bit images format program	T 11 graphics image data ftp software pub mail package fax images	T 80 db key chip encryption clipper system government keys law escrow	T 59 image jpeg color file gif images format bit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

Figure 3. Top topics under each class as discovered by the MedLDA and LDA models

Figure 2 shows the 2D embedding of the expected topic proportions of MedLDA^c and LDA by using the t-SNE stochastic neighborhood embedding (van der Maaten & Hinton, 2008), where each dot represents a document and color-shape pairs represent class labels. Obviously, the max-margin based MedLDA^c produces a better grouping and separation of the documents in different categories. In contrast, the unsupervised LDA does not produce a well separated embedding, and documents in different categories tend to mix together. A similar embedding was presented in (Lacoste-Jullien et al., 2008), where the transformation matrix in their model is pre-designed. The results of MedLDA^c in Figure 2 are *automatically* learned.

It is also interesting to examine the discovered topics and their association with class labels. In Figure 3 we show the top topics in four classes as discovered by both MedLDA and LDA. Moreover, we depict the per-class distribution over topics for each model. This distribution is computed by averaging the expected latent representation of the documents in each class. We can see that MedLDA yields sharper, sparser and fast decaying per-class distributions over topics which have a better discrimination power. This behavior is in fact due to the regularization effect enforced over ϕ as shown in Eq. (8). On the other hand, LDA seems to discover topics that model the fine details of documents with no regard to 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 in this class 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.

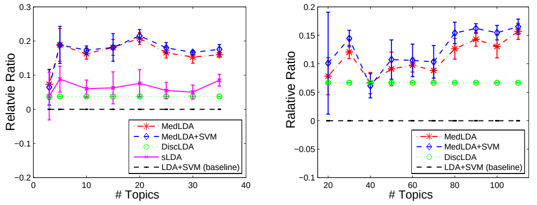
4.2. Prediction Accuracy

In this subsection, we provide a quantitative evaluation of the MedLDA on prediction performance.

4.2.1. Classification

We perform binary and multi-class classification on the 20 Newsgroup data set. To obtain a baseline, we first fit all the data to an LDA model, and then use the latent representation of the training[4] documents as features to build a binary/multi-class SVM classifier. We

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



(a) (b)

Figure 4. Relative improvement ratio against LDA+SVM for: (a) binary and (b) multi-class classification.

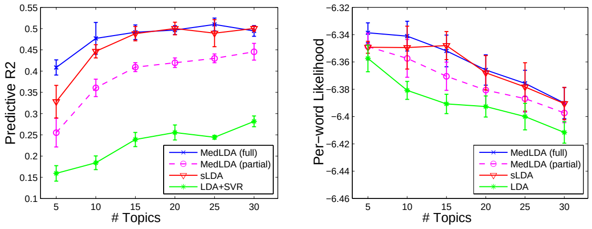


Figure 5. Predictive R^2 (left) and per-word likelihood (right) of different models on the movie review dataset.

denote this baseline by LDA+SVM. For a model $\mathcal{M}$, we evaluate its performance using the relative improvement ratio, i.e., $\frac{precision(\mathcal{M}) \,-\, precision(LDA+SVM)}{precision(LDA+SVM)}$.

Binary Classification: As in (Lacoste-Jullien et al., 2008), the binary classification is to distinguish postings of the newsgroup *alt.atheism* and the postings of the group *talk.religion.misc*. We compare MedLDA^c with sLDA, DiscLDA and LDA+SVM. For sLDA, to the best of our knowledge, the classification model has not been evaluated. Therefore, we fit an sLDA regression model using the binary representation (0/1) of the class, and use a threshold 0.5 to make prediction. For MedLDA^c, to see whether a second-stage max-margin classifier can improve the performance, we also build a method *MedLDA+SVM*, similar to LDA+SVM. For all the above methods that utilize the class label information, they are fit *ONLY* on the training data.

We use the SVM-light (Joachims, 1999) to build SVM classifiers and to estimate $q(\eta)$ in MedLDA^c. The parameter C is chosen via 5 fold cross-validation during the training from $\{k^2 : k = 1, \cdots, 8\}$. For each model, we run the experiments for 5 times and take the average as the final results. The relative improvement ratios of different models w.r.t topic numbers are shown in Figure 4(a). For the recently proposed DiscLDA (Lacoste-Jullien et al., 2008), since the implementation is not available, the results are taken from the original paper for both DiscLDA and LDA+SVM.

We can see that the max-margin based MedLDA^c works better than sLDA, DiscLDA and the two-step method of LDA+SVM. Since MedLDA^c integrates the max-margin principle in its training, the combination of MedLDA and SVM does not yield additional benefits on this task. We believe that the slight differences between MedLDA and MedLDA+SVM are due to tuning of the regularization parameters. For efficiency, we do not change the regularization constant C during training MedLDA^c. The performance would be improved if we select 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 categories. We compare MedLDA^c with MedLDA+SVM, LDA+SVM, and DiscLDA. We use the SVM^{struct} package[5] with a 0/1 loss to solve the sub-step of learning $q(\eta)$ and build the SVM classifiers for LDA+SVM and MedLDA+SVM. The results are shown in Figure 4(b), where the results of DiscLDA are again taken from (Lacoste-Jullien et al., 2008). We can see that all the supervised topic models discover more predictive topics for classification, and the max-margin based MedLDA^c can achieve significant improvements with an appropriate number (e.g., ≥ 80) of topics. Again, we believe that the slight difference between MedLDA^c and MedLDA+SVM is due to parameter tuning.

4.2.2. Regression

We evaluate the MedLDA^r model on the movie review data set. As in (Blei & McAuliffe, 2007), we take logs of the response values to make them approximately normal. We compare MedLDA^r with the unsupervised LDA and sLDA. As we have stated, the underlying topic model in MedLDA^r can be a LDA or a sLDA. We have implemented both, as denoted by *MedLDA (partial)* and *MedLDA (full)*, respectively. For LDA, we use its low dimensional representation of documents as input features to a linear SVR and denote this method by *LDA+SVR*. The evaluation criterion is predictive R^2 (pR^2) as defined in (Blei & McAuliffe, 2007).

Figure 5 shows the results together with the per-word likelihood. We can see that the supervised MedLDA and sLDA can get much better results than the unsupervised LDA, which ignores supervised responses. By using max-margin learning, MedLDA (full) can get slightly better results than the likelihood-based sLDA, especially when the number of topics is small (e.g., ≤ 15). Indeed, when the number of topics is small, the latent representation of sLDA alone does not result in a highly separable problem, thus the integration of max-margin training helps in discovering a more discriminative latent representation using the same number of topics. In fact, the number of support vectors (i.e., documents that have at least one non-zero lagrange multiplier) decreases dramatically at $T = 15$ and stays nearly the same for $T > 15$, which with reference to Eq. (4) explains why the relative improvement

[5] http://svmlight.joachims.org/svm_multiclass.html

over sLDA decreased as T increases. This behavior suggests that MedLDA can discover more predictive latent structures for *difficult*, non-separable problems.

For the two variants of MedLDA^r, we can see an obvious improvement of MedLDA (full). This is because for MedLDA (partial), the update rule of ϕ does not have the third and fourth terms of Eq. (4). Those terms make the max-margin estimation and latent topic discovery attached more tightly. Finally, a linear SVR on the empirical word frequency gets a pR^2 of 0.458, worse than those of sLDA and MedLDA.

4.2.3. Time Efficiency

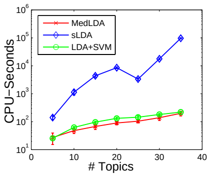


Figure 6. Training time.

For binary classification, MedLDA^c is much more efficient than sLDA, and is comparable with the LDA+SVM, as shown in Figure 6. The slowness of sLDA may be due to the mismatching 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^c is mainly dependent on solving a multi-class SVM problem, and thus is comparable to that of LDA. For regression, the training time of MedLDA (full) is comparable to that of sLDA, while MedLDA (partial) is more efficient.

5. Conclusions and Discussions

We have presented the maximum entropy discrimination LDA (MedLDA) that uses the max-margin principle to train supervised topic models. MedLDA integrates the max-margin principle into the latent topic discovery process via optimizing one single objective function with a set of *expected* margin constraints. This integration yields a predictive topic representation that is more suitable for regression or classification. We develop efficient variational methods for MedLDA. The empirical results on movie review and 20 Newsgroups data sets show the promise of MedLDA on text modeling and prediction accuracy.

MedLDA represents the first step towards integrating the max-margin principle into supervised topic models, and under the general MedTM framework presented in Section 3, several improvements and extensions are in the horizon. Specifically, due to the nature of MedTM's joint optimization formulation, advances in either max-margin training or better variational bounds for inference can be easily incorporated. For instance, the mean field variational upper bound in MedLDA can be improved by using the tighter collapsed variational bound (Teh et al., 2006) that achieves results comparable to collapsed Gibbs sampling (T. Griffiths, 2004). Moreover, as the experimental results suggest, incorporation of a more expressive underlying topic model enhances the overall performance. Therefore, we plan to integrate and utilize other underlying topic models like the fully generative sLDA model in the classification case.

Acknowledgements

This work was done while J.Z. was visiting CMU under a support from NSF DBI-0546594 and DBI-0640543 awarded to E.X.; J.Z. is also supported by Chinese NSF Grant 60621062 and 60605003; National Key Foundation R&D Projects 2003CB317007, 2004CB318108 and 2007CB311003; and Basic Research Foundation of Tsinghua National TNList Lab.

References

Blei, D., & Jordan, M. (2003). Modeling annotated data. *SIGIR*, 127–134.

Blei, D., & Lafferty, J. (2005). Correlated topic models. *NIPS*, 147–154.

Blei, D., & McAuliffe, J. D. (2007). Supervised topic models. *NIPS*, 121–128.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *J. of Mach. Learn. Res.*, 993–1022.

Crammer, K., & Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *J. of Mach. Learn. Res.*, 265–292.

Joachims, T. (1999). Making large-scale SVM learning practical. *Advances in kernel methods–support vector learning, MIT-Press*, 169–184.

Lacoste-Jullien, S., Sha, F., & Jordan, M. I. (2008). DiscLDA: Discriminative learning for dimensionality reduction and classification. *NIPS*, 897–904.

McCallum, A., Pal, C., Druck, G., & Wang, X. (2006). Multi-conditional learning: generative/discriminative training for clustering and classification. *AAAI*, 433–439.

Smola, A., & Schölkopf, B. (2003). A tutorial on support vector regression. *Statistics and Computing*, 199-222.

T. Griffiths, M. S. (2004). Finding scientific topics. *Proc. of National Academy of Sci.*, 5228–5235.

Teh, Y. W., Newman, D., & Welling, M. (2006). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *NIPS*, 1353–1360.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *JMLR*, 2579–2605.

Welling, M., Rosen-Zvi, M., & Hinton, G. (2004). Exponential family harmoniums with an application to information retrieval. *NIPS*, 1481-1488.

Zhu, J., Xing, E., & Zhang, B. (2008a). Laplace maximum margin Markov networks. *ICML*, 1256–1263.

Zhu, J., Xing, E., & Zhang, B. (2008b). Partially observed maximum entropy discrimination Markov networks. *NIPS*, 1977–1984.