Supervising Latent Topic Model for Maximum-Margin Text Classification and Regression

Wanhong Xu

Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213
wanhong@cmu.edu

Abstract. In this paper, we investigate the text classification and regression problems: given a corpus of text documents as training, each of which has a response label, the task is to train a predictor for predicting its response of any given document. In previous work, many researchers decompose this task into two separate steps: they first use a generative latent topic model to learn low-dimensional semantic representations of documents; and then train a max-margin predictor using them as features. In this work we demonstrate that it is beneficial to combine both steps of learning low-dimensional representations and training a predictor into one step of minimizing a singe learning objective. We present a novel step-wise convex optimization algorithm which solves this objective properly via a tight variational upper bound. We conduct an extensive experimental study on public available movie review and 20 Newsgroups datasets. Experimental results show that compared with state of art results in the literature, our one step approach can train noticeably better predictors and discover much lower-dimensional representations: a 2% relative accuracy improvement and a 95% relative number of dimensions reduction in the classification task on the Newsgroups dataset; and a 5.7% relative predictive R^2 improvement and a 55% relative number of dimensions reduction in the regression task on the movie review dataset.

1 Introduction

With tremendous text information made available online, there is a growing demand to analyze and manage large corpuses of electronic text. Learning low-dimensional semantic representations of text documents is a common and often necessary step for various applications and text analyses. For example, this low-dimensional semantic representation has been used for structurally browsing a text corpus and categorizing and clustering text documents in information retrieval domain.

A recent trend in learning low-dimensional semantic representations focuses on generative latent probabilistic models based on so-called *topics*. The belief behind those latent topic models is that a document consisting of a large number of words might be concisely described as a smaller number of semantic themes. A topic is a probability distribution over words of a vocabulary, and is used to statistically describe a semantic theme. Then, a document is semantically represented as a mixture of topics.

In the literature, most popular latent topic models are Latent Dirichlet Allocation (LDA) [2] and Probabilistic Latent Semantic Analysis (pLSA) [5]. Particularly, LDA

M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI 6118, pp. 403–414, 2010.

is a Bayesian version of pLSA. Both models are unsupervised to simultaneously discover topics and low-dimensional topic representations of documents, and have been successfully used in various applications [7,3].

Unfortunately, semantic representations produced in this unsupervised manner may not necessarily be good features for classification and regression tasks. The reason is that topics are learned without considering document labels to be predicted, for example the categories of news postings. Those unsupervised learned topics describe semantic themes that generally happen in all documents and don't describe semantic themes discriminative across document categories. Therefore, semantic representations of documents based on general topics are not well distinctive against document categories and those two-step approaches of building predictors subsequently on them would result in sub-optimal performance. We believe that topics should contain as much discriminative information as possible from document labels such that semantic representations based on them are suitable for prediction.

In this paper, we propose an approach to integrate the latent topic model for learning low-dimensional semantic representations and support vector machine or regression for training max-margin predictors into one single learning objective. By coupling them together, we are able to supervise the latent topic model with benefits of the maximum margin principle, and guide it to discover topics describing discriminative information and generate semantic representations more suitable for prediction tasks. Due to the optimization hardness of the single learning objective, we propose a tight variational upper bound for the single learning objective and develop a novel step-wise convex algorithm for optimizing the upper bound. For both classification and regression tasks, we present experiments showing that our approach can achieve better predictive power and is able to discover much lower dimensional representations than two-step approaches and also three state of art methods.

2 Preliminaries

We first introduce notations that will be used throughout the paper; then review latent topic models and max-margin classifier and regressor.

A text document d is a sequence of N words $<w_1 w_2 \dots w_N>$, where each word is from a fixed vocabulary with totally V words. Following a common bag-of-word assumption, we represent this document as a bag of words. The document could have a response label y, which is either a categorical class, or a continuous real number. Let D be a corpus of M labeled documents. The problem in this work is to learn a good predictor using D as training data for predicting the response label of a new document.

2.1 Latent Topic Model

To represent a document by semantic topics, we define a K-topic vocabulary T. Each topic t of T is a multinomial distribution of all words in the vocabulary, i.e. $\{p(w|t)\}_{w=1..V}$, simply denoted as $\beta_{t,:}$. We also let β be the set of all topics, i.e., $\{\beta_{t,:}\}_{t=1..K}$. Then, we can represent the document by a topic mixture proportion $\theta = \{p(t)\}_{t=1..K}$.

This topic representation implies a generative process to documents. For each word w_n in a document d, we

1. draw a topic assignment $z_n|\theta \sim Mult(\theta)$, where z_n has the 1-of-K representation;
2. draw the word $w_m|z_n, \beta \sim Mult(\beta_{z_n,:})$.

Both LDA and pLSA are based on the above generative process. The difference is that LDA introduces a Dirichlet prior to θ for alleviating the potential overfitting problem of pLSA. For classification and regression tasks, both models show no difference in prediction performance. Therefore, to keep our approach simple, we recruit pLSA as the topic model in our approach.

Given a corpus D with M documents, pLSA minimizes the following negative log likelihood to learn topics β and also estimate topic proportion θ for each document:

$$\min_{\Theta,\beta} -L(D;\Theta,\beta) = -\sum_{m=1}^{M}\sum_{n=1}^{N_m}\log\sum_{z_{m,n}=1}^{K} p(z_{m,n}|\theta_m)p(w_{m,n}|\beta_{z_{m,n},:}),$$

where $\Theta = \{\theta_m\}_{m=1..M}$ denotes the collection of all topic proportions. We let $\bar{z}_m = \frac{1}{N_m}\sum_{n=1}^{N_m} z_{m,n}$, which is the empirical topic proportion of document m, and let $\bar{Z} = \{\bar{z}_m\}_{m=1..M}$ denote the set of all empirical topic proportions of D.

2.2 Max-margin Classification and Regression

The empirical topic proportion $\bar{z}_m$ from the latent topic model and the document label y_m are used to build max-margin predictors, for example support vector machine (SVM) [4] for classification and support vector regressor (SVR) [10] for regression.

For example, if $y_m \in \{-1, 1\}$ is a binary categorical label, we can learn a SVM $<\omega, b>$ from the corpus D for classification by minimizing the following loss function:

$$\min_{\omega,b} C(\omega, b, \bar{Z}) = \frac{1}{M}\sum_{m=1}^{M}\max\{0, 1 - y_m(\omega^T\bar{z}_m + b)\} + \lambda||\omega||_2,$$

where the first term measures the classification error of $<\omega, b>$ and the second is a penalty term on ω to avoid over-fitting. Similarly, if y_m is continous, we can learn a SVR $<\omega, b>$ with a pre-defined precision ϵ from Corpus D for regression by minimizing the following loss function:

$$\min_{\omega,b} R(\omega, b, \bar{Z}) = \frac{1}{M}\sum_{m=1}^{M}\{\max\{0, y_m - \epsilon - \omega^T\bar{z}_m - b\} + \max\{0, \omega^T\bar{z}_m + b - y_m - \epsilon\}\} + \lambda||\omega||_2,$$

where the first term measures the prediction error of $<\omega, b>$ on the ϵ precision, and the second is a penalty term on ω too.

Because $\bar{z}_m$ is a hidden variable, we need to minimize the expected two loss functions $E_{\bar{Z}}(C(\omega, b, \bar{Z}))$ and $E_{\bar{Z}}(R(\omega, b, \bar{Z}))$. But, it is hard to compute them because of the non-integrable max function in two loss functions. To circumvent this difficulty, two-step approaches minimize $C(\omega, b, E(\bar{Z}))$ and $R(\omega, b, E(\bar{Z}))$ instead. Because $E(\bar{Z})$ has a closed form Θ, this alternative makes possible using standard SVM and SVR solver.

However, we find that both $R(\omega, b, \bar{Z})$ and $C(\omega, b, \bar{Z})$ are convex on $\bar{Z}$. Therefore, $C(\omega, b, E(\bar{Z}))$ and $R(\omega, b, E(\bar{Z}))$ are lower bounds of $E_{\bar{Z}}(C(\omega, b, \bar{Z}))$ and $E_{\bar{Z}}(R(\omega, b, \bar{Z}))$ respectively according to Jensen's inequality. It is obvious that minimizing lower bounds of $E_{\bar{Z}}(C(\omega, b, \bar{Z}))$ and $E_{\bar{Z}}(R(\omega, b, \bar{Z}))$ in two-step approaches are problematic and doesn't guarantee $E_{\bar{Z}}(C(\omega, b, \bar{Z}))$ and $E_{\bar{Z}}(R(\omega, b, \bar{Z}))$ themselves to be minimized too.

3 Framework of Supervising Latent Topic Model

As discussed in the introduction, the goal is to learn semantic topics β describing discriminative themes among documents in the corpus D such that representations of documents by those discriminative topics could be suitable for prediction purpose.

To achieve this goal, two requirements should be satisfied together. On one hand, topics β should be common to make possible a to-be-predicted document be well described by them too. This suggests that the negative log likelihood of the latent topic model $-L(D;\Theta,\beta)$, which measures how data fits those topics β, should be minimized. On the other hand, the empirical topic proportions $\bar{Z}$ can be used to build good predictors. It implies that the expected loss functions $E_{\bar{Z}}(C(\omega,b,\bar{Z}))$ and $E_{\bar{Z}}(R(\omega,b,\bar{Z}))$ of SVM and SVR, which measure how well they are used for prediction, should be minimized too.

To satisfy both requirements together, we intuitively couple them linearly by a positive tradeoff η to a single learning objective. For classification and regression tasks, we need to minimize the following single learning objectives respectively:

$$\min_{\Theta,\beta,\omega,b} -L(D;\Theta,\beta) + \eta E_{\bar{Z}}(C(\omega,b,\bar{Z})) \tag{1}$$

$$\min_{\Theta,\beta,\omega,b} -L(D;\Theta,\beta) + \eta E_{\bar{Z}}(R(\omega,b,\bar{Z})) \tag{2}$$

Therefore, in the single learning objectives, the expected losses of max-margin predictors are used to penalize or supervise the latent topic model to generate discriminative topics suitable for prediction. We name this framework maximum margin latent topic model, shortly denoted as MMpLSA.

To solve those two learning problems, we have to remove or integrate out hidden variables $z_{m,n}$ so that convex optimization could be applied to them.

In the log likelihood function, it is easy to remove hidden variables $z_{m,n}$ by utiltizing the multinomial distribution of $z_{m,n}$. We have the following alternative form of the log likelihood function:

$$-L(D;\Theta,\beta) = -\sum_{m=1}^{M}\sum_{n=1}^{N_m}\log\sum_{z_{m,n}=1}^{K} p(z_{m,n}|\theta_m)p(w_{m,n}|\beta_{z_{m,n},:}) = -\sum_{m=1}^{M}\sum_{n=1}^{N_m}\log\theta_m^T\beta_{:,w_{m,n}},$$

where $\beta_{:,w_{m,n}}=\{\beta_{k,w_{m,n}}\}_{k=1..K}$ is the vector of probabilities of word $w_{m,n}$ in all topics respectively.

However, no closed forms exist for $E_{\bar{Z}}(C(\omega,b,\bar{Z}))$ and $E_{\bar{Z}}(R(\omega,b,\bar{Z}))$ because of non-integrable max functions involved in them. One way to circumvent this closed form trouble as two-step approaches is using $C(\omega,b,E(\bar{Z}))$ and $R(\omega,b,E(\bar{Z}))$ to replace $E_{\bar{Z}}(C(\omega,b,\bar{Z}))$ and $E_{\bar{Z}}(R(\omega,b,\bar{Z}))$ in the learning objectives. But, this way will have the same problem as two step approaches: minimizing lower bounds of single learning objectives doesn't guarantee they will be small too.

Instead, in this work, we propose tight closed form upper bounds of $E_{\bar{Z}}(C(\omega,b,\bar{Z}))$ and $E_{\bar{Z}}(R(\omega,b,\bar{Z}))$ such that when tight upper bounds are minimized, single learning objectives are approximately minimized too.

3.1 Variational Upper Bounds for Expected Max-margin Loss Functions

We first propose a variational upper bound for the max function $\max\{0, x\}$, and then give upper bounds of $E_{\bar{Z}}(C(\omega, b, \bar{Z})$ and $E_{\bar{Z}}(R(\omega, b, \bar{Z}))$.

The function $\max\{0, x\}$ has an upper bound $g_\gamma(x)$ [12], which is defined as below:

$$g_\gamma(x) = \frac{x}{2} + \frac{1}{\gamma}\log(\exp(-\frac{\gamma x}{2}) + \exp(\frac{\gamma x}{2})). \tag{3}$$

It is easy to show that $\max\{0,x\} < g_\gamma(x)$ when $\gamma > 0$. This upper bound is tight because for any x, $\lim_{\gamma\to\infty}(g_\gamma(x) - \max\{0,x\}) = 0$. Let $f_\gamma(x) = \frac{1}{\gamma}\log((\exp(-\frac{\gamma x}{2}) + \exp(\frac{\gamma x}{2})))$. It is a concave function to the variable x^2 [6], and thus its first order Taylor expansion at the variable x^2 is a global upper bound,

$$f_\gamma(x) \leq f_\gamma(\psi) + \frac{1}{4\gamma\psi}\tanh(\frac{\gamma\psi}{2})(x^2 - \psi^2). \tag{4}$$

Note that this upper bound is exact whenever $\psi^2 = x^2$. Combining Eq.3 and 4 yields the desired variational upper bound of $\max\{0, x\}$,

$$\max\{0, x\} < \frac{x}{2} + f_\gamma(\psi) + \frac{1}{4\gamma\psi}\tanh(\frac{\gamma\psi}{2})(x^2 - \psi^2), \tag{5}$$

where ψ is a variational variable which gives the upper bound one degree of freedom to tightly approximate the max function.

Based on the variational upper bound of max function, we then give the variational upper bound of $E_{\bar{Z}}(C(\omega, b, \bar{Z}))$. For a document m and its empirical topic proportion $\bar{z}_m$, we have $E(\bar{z}_m) = \theta_m$ and $E(\bar{z}_m\bar{z}_m^T) = (N_m - 1)/N_m \cdot \theta_m\theta_m^T + 1/N_m\,\mathrm{diag}(\theta_m)$, denoted as Ω_m. Putting them with Eq. 5 and knowledge of $y_m^2 = 1$ together leads to an expected upper bound of $max\{0,1 - y_m(\omega^T\bar{z}_m + b)\}$:

$$\begin{aligned}
&E(max\{0,1 - y_m(\omega^T\bar{z}_m + b)\}) \\
<&E\{\frac{1 - y_m(\omega^T\bar{z}_m + b)}{2} + f_\gamma(\psi_m) + \frac{\tanh(\gamma\psi_m/2)}{4\gamma\psi_m}[(1 - y_m(\omega^T\bar{z}_m + b))^2 - \psi_m^2]\} \\
=&\frac{1 - y_m(\omega^T\theta_m + b)}{2} + f_\gamma(\psi_m) + \frac{\tanh(\gamma\psi_m/2)}{4\gamma\psi_m}[\omega^T\Omega_m\omega + 2\omega^T\theta_m(b - y_m) + (b - y_m)^2 - \psi_m^2].
\end{aligned}$$

We define this upper bound as $B_\gamma(\theta_m, \omega, b, \psi_m)$, and then the variational upper bound of $E_{\bar{Z}}(C(\omega, b, \bar{Z}))$ is $\frac{1}{M}\sum_{m=1}^{M} B_\gamma(\theta_m, \omega, b, \psi_m) + \lambda||\omega||_2$.

Similarly, for regression, we have an expected upper bound $\max\{0, y_m - \epsilon - \omega^T\bar{z}_m - b\} + \max\{0, \omega^T\bar{z}_m + b - y_m - \epsilon\}$:

$$\begin{aligned}
&E(\max\{0, y_m - \epsilon - \omega^T\bar{z}_m - b\} + \max\{0, \omega^T\bar{z}_m + b - y_m - \epsilon\}) \\
<&-\epsilon + f_\gamma(\psi_m) + \frac{\tanh(\gamma\psi_m/2)}{4\gamma\psi_m}[\omega^T\Omega_m\omega + 2\omega^T\theta_m(b - y_m + \epsilon) + (b - y_m + \epsilon)^2 - \psi_m^2] \\
&+ f_\gamma(\psi_m^*) + \frac{\tanh(\gamma\psi_m^*/2)}{4\gamma\psi_m^*}[\omega^T\Omega_m\omega + 2\omega^T\theta_m(b - y_m - \epsilon) + (b - y_m - \epsilon)^2 - \psi_m^{*2}].
\end{aligned}$$

We define this upper bound as $U_\gamma(\theta_m, \omega, b, \psi_m, \psi_m^*)$, and then the variational upper bound of $E_{\bar{Z}}(R(\omega, b, \bar{Z}))$ is $\frac{1}{M}\sum_{m=1}^{M} U_\gamma(\theta_m, \omega, b, \psi_m, \psi_m^*) + \lambda||\omega||_2$.

4 Optimization Procedure for Classification

Armed with the variational upper bound of the expected classification loss function proposed in the previous section, we can approximately minimize the single learning objective for classification (Eq. 1) by minimizing its following upper bound:

$$\begin{aligned}
\min_{\Theta,\beta,\omega,b,\Psi} \; & -\sum_{m=1}^{M}\sum_{n=1}^{N_m} \log \theta_m^T \beta_{:,w_{m,n}} + \lambda_1 \sum_{m=1}^{M} B_\gamma(\theta_m, \omega, b, \psi_m) + \lambda_2 ||\boldsymbol{\omega}||_2 \\
subject\ to: \; & \theta_{m,k} > 0, && m = 1 \ldots M, k = 1 \ldots K; \\
& \beta_{k,v} > 0, && k = 1 \ldots K, v = 1 \ldots V; \\
& \Sigma_{k=1}^{K} \theta_{m,k} = 1, && m = 1 \ldots M; \\
& \Sigma_{v=1}^{V} \beta_{k,v} = 1, && k = 1 \ldots K;
\end{aligned}$$

where λ_1 and λ_2 are absorbed terms of the objective tradeoff η and constant $1/M$ and parameter λ in the variational upper bound of the expected classification loss. Because Θ and β are parameters of multinomial distributions, self-explained constraints as above must be satisfied in this minimization.

It could be proved that this objective upper bound is variable-wise convex[1]. Therefore, we could iteratively minimize it with respect to one of variables with the rest of variables fixed. Because every iteration reduces its overall value, this iterative minimization procedure will cause the value of objective upper bound to converge to a local minimum. Next, we describe this iterative procedure below starting from the simplest iterating step:

OPTIMIZE Ψ: Because variational variables are uncoupled to each other in the objective upper bound, we can divide the optimization for Ψ into M subproblems, one per variational variable. There is only one term involving the variational variables in the objective upper bound. Therefore, the objective upper bound is simplified to the below for optimizing each variational variable:

$$\min_{\psi_m} B_\gamma(\theta_m, \omega, b, \psi_m),$$

which turns out to have a closed form solution $\psi_m = \sqrt{\omega^T \Omega_m \omega + 2\omega^T \theta_m (b - y_m) + (b - y_m)^2}$.

OPTIMIZE ω AND b: The first term of the objective upper bound doesn't involve ω and b and also all constraints don't. With them dropped, the optimization for $<\omega,b>$ is simplified to the following unconstrained optimization problem:

$$\min_{\omega,b} \lambda_1 \sum_{m=1}^{M} B_\gamma(\theta_m, \omega, b, \psi_m) + \lambda_2 ||\boldsymbol{\omega}||_2,$$

which tries to choose $<\omega,b>$ for good prediction. The Hessian matrix of $<\omega,b>$ is:

$$H(<\omega,b>) = \lambda_1 \sum_{m=1}^{M} \left\{ \frac{\tanh(\gamma\psi_m/2)}{2\gamma\psi_m} \begin{bmatrix} \Omega_m & \theta_m \\ \theta_m^T & 1 \end{bmatrix} \right\} + 2\lambda_2 \begin{bmatrix} I_{K\times K} & 0_{K\times 1} \\ 0_{1\times K} & 0 \end{bmatrix},$$

[1] Due to the space limitation, we skip the proof in this writing. Please refer to [11] for details.

where $I_{K\times K}$ is the identity matrix and $0_{K\times 1}$ and $0_{1\times K}$ are vectors with only 0s. This Hessian matrix involves γ, which is supposed to be large enough for well approximating the max function as shown in Sec 3.1. But when γ is big, the Hessian matrix could be ill-conditioned, which will lead to the instability of our algorithm solving this optimization problem. Our stable solution is that we first solve the minimization problem for a small γ to get optimal ω and b, and based on them we solve this problem again for a bigger γ to update optimal ω and b, and so on. In implementation, we start γ from 10 and increase it by 20 until it reaches 200.

OPTIMIZE Θ: Topic proportion θ_ms are uncoupled to each other. Therefore, we can divide the optimization for Θ into M subproblems, one per topic proportion. By dropping the third term of objective upper bound without involving Θ and constraints on β, the optimization for Θ is simplified to the following constrained optimization problem:

$$\begin{aligned}\min_{\theta_m} \; & -\sum_{n=1}^{N_m} \log \theta_m^T \beta_{:,w_{m,n}} + \lambda_1 B_\gamma(\theta_m, \omega, b, \psi_m) \\ subject\ to: \; & \theta_{m,k} > 0, \qquad k = 1 \dots K; \\ & \Sigma_{k=1}^{K} \theta_{m,k} = 1.\end{aligned}$$

The Hessian of θ_m is:

$$H(\theta_m) = \sum_{n=1}^{N_m} \frac{\beta_{:,w_{m,n}} \beta_{:,w_{m,n}}^T}{(\theta_m^T \beta_{:,w_{m,n}})^2} + \lambda_1 \frac{\tanh(\gamma\psi_m/2)}{2\gamma\psi_m} \cdot \frac{N_m - 1}{N_m} \cdot \omega\omega^T,$$

which involves γ too and has the same ill-conditioned problem when γ is large as the Hessian Matrix of $<\omega,b>$. We use the same solution for stability as optimizing $<\omega,b>$.

OPTIMIZE β: Only the first term of objective upper bound involves β. By keeping it and also constraints on β, we simplify the optimization for β to the following constrained optimization problem:

$$\begin{aligned}\min_{\beta} \; & -\sum_{m=1}^{M} \sum_{n=1}^{N_m} \log \theta_m^T \beta_{:,w_{m,n}} \\ subject\ to: \; & \beta_{k,v} > 0, \qquad k = 1 \dots K, v = 1 \dots V \\ & \Sigma_{v=1}^{V} \beta_{k,v} = 1, \quad k = 1 \dots K.\end{aligned}$$

Based on those optimization steps, the iterative optimization procedure for classification is given in Algorithm 1. We discuss the implementation detail in the next section.

4.1 Implementation

In the beginning of the optimization procedure, for each topic t, we initialize its multinomial word distribution $\beta_{t,:}$ by sampling a dirichlet distribution with parameter $(1, \dots, 1)$; we also initialize each value of ω and b by sampling a standard normal distribution.

Every time when the optimization procedure optimizes β and ω and b, we do cross validation to check whether currently learned β are good topics for prediction and $<\omega, b>$ is a good classifier. In the cross validation of β, topic proportion θ_c of a document c

Alg1: Optimization Procedure for Classification

Input: corpus D, model parameters λ_1 and λ_2, and topic number K.

Output: β, ω and b.

```
Initialize β, ω and b;
repeat
    for γ = 10; γ < 200; γ = γ + 20 do
        repeat
            Optimize Θ; Optimize Ψ;
        until convergence
    end for
    Optimize β;
    for γ = 10; γ < 200; γ = γ + 20 do
        repeat
            Optimize ω and b; Optimize Ψ;
        until convergence
    end for
    Cross Validation on β.
until convergence
```

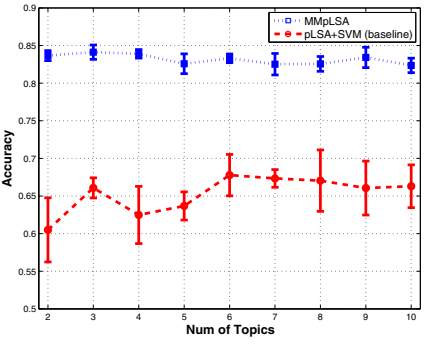

Fig. 1. Experimental results of text classification: Mean and variance of accuracies of the proposed approach MMpLSA and the two-step approach pLSA+SVM. MMpLSA performs significantly better than pLSA+SVM.

preselected for cross validation is estimated by minimizing the negative log likelihood with β fixed:

$$\begin{aligned} \min_{\theta_c} \; & -\sum_{n=1}^{N_c} \log \theta_c^T \beta_{:,w_{c,n}} \\ subject\ to: \; & \theta_{c,k} > 0, \qquad k = 1 \dots K; \\ & \Sigma_{k=1}^{K} \theta_{c,k} = 1. \end{aligned}$$

Let $\bar{z}_c$ be its empirical topic proportion. We predict its label by $sign(E(\omega^T \bar{z}_c + b))$, i.e., $sign(\omega^T \theta_c + b)$.

5 Optimization Procedure for Text Regression

Similar to classification, by utilizing the variational upper bound of the expected regression loss function, we can approximately minimize the single learning objective for regression (Eq. 2) by minimizing its following upper bound:

$$\begin{aligned} \min_{\Theta,\beta,\omega,b,\Psi,\Psi^*} \; & -\sum_{m=1}^{M}\sum_{n=1}^{N_m} \log \theta_m^T \beta_{:,w_{m,n}} + \lambda_1 \sum_{m=1}^{M} U_\gamma(\theta_m, \omega, b, \psi_m, \psi_m^*) + \lambda_2 ||\boldsymbol{\omega}||_2 \\ subject\ to: \; & \theta_{m,k} > 0, && m = 1 \dots M, k = 1 \dots K; \\ & \beta_{k,v} > 0, && k = 1 \dots K, v = 1 \dots V; \\ & \Sigma_{k=1}^{K} \theta_{m,k} = 1, && m = 1 \dots M; \\ & \Sigma_{v=1}^{V} \beta_{k,v} = 1, && k = 1 \dots K. \end{aligned}$$

Due to space limitation, please refer to [11] for the detail of the optimization procedure, which shares many commons with the optimization procedure for classification.

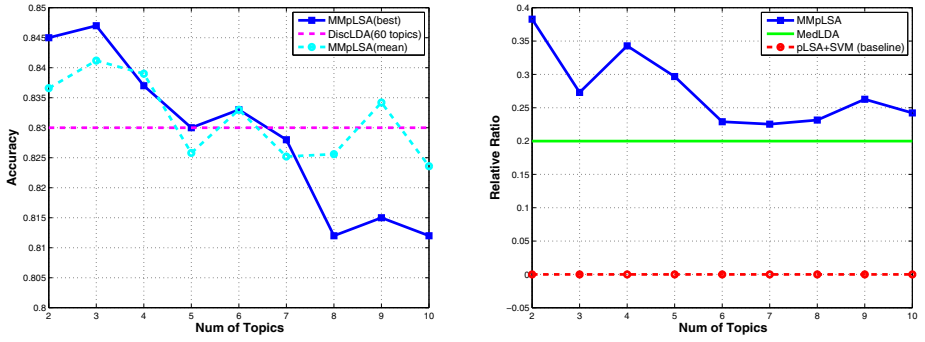

Fig. 2. Experimental results of text classification: (Left) Accuracy mean of five runs of MMpLSA and the accuracy of the run with best cross validation at different numbers of topics v.s. the best result 0.830 of state of art model DiscLDA achieved at 60 topics; MMpLSA achieved the best accuracy 84.7% with 3 topics, a 2% percent relative accuracy improvement and a 95% relative number of topics reduction; (Right) Relative improvement ratio of accuracy mean of MMpLSA against pLSA+SVM v.s. best relative ratio achieved by MedLDA

6 Experiments

In this section, we conducted experiments to evaluate our apporach MMpLSA with state of art methods and also a baseline from two-step approaches; we report extensive performance results on both text classification and regression. Our experiments are able to demonstrate the advantages of applying the max-margin principle to supervise latent topic models. MMpLSA can learn from data a compacter latent representation that contains more plentiful information for prediction.

6.1 State of Art Approaches

There have been moderate efforts on supervising latent topic models for classification and regression in the literature. The most earliest work is sLDA [1], which supervises LDA for regression by assuming a normal distribution of the response y of a document and also assuming y linearly dependent on its empirical topic proportion $\bar{z}$, i.e., $y \sim N(\mu^T \bar{z}, \sigma^2)$. But this normality assumption doesn't hold for many real datasets. DiscLDA [8] is a discriminative variant for classification. It assumes that topic proportions of each class after a linear transformation should be nearby to each other. Parameters of linear transformations are learned by maximizing the conditional likelihood of the response classes. But DiscLDA can't guarantee that topic proportions of different classes after linear transformations are well separated, which is critical for classification.

Applying the max-margin principle to supervise latent topic models could avoid drawbacks of sLDA and DiscLDA. For example, SVR doesn't require document labels to be normally distributed and SVM could help forcing topic proportions of different classes to be well separated by a good margin. The most recent MedLDA [13] is an approach utilizing the max-margin principle. However, MedLDA uses the lower

bounds of expected SVM and SVR loss functions to supervise latent topic models. It extremely simplifies inference algorithms, but it is problematic as we discuss in Section 3. Our approach MMpLSA also recruits the max-margin principle to supervise latent topic models. But different to MedLDA, we propose tight variational upper bounds of expected loss functions. Based on upper bounds, we develop a stepwise convex algorithm for optimization, totally different to EM algorithms used by those existing approaches.

6.2 Text Classification

To be able to compare MMpLSA with DiscLDA and MedLDA, we also evaluated MMpLSA on the *20 Newsgroups* dataset containing postings to Usenet newsgroups. As DiscLDA and MedLDA, we formulated the same classification problem for distinguishing postings from two newsgroups: *alt.atheism* and *talk.relgion.misc*, a hard task due to the content similarity between them. We also used the training/testing split provided in the *20 Newsgroups* dataset to make possible a fair comparison among them.

To obtain a baseline from two-step approaches, we first fit all the data to pLSA model, and then used empirical topic proportions as features to train a linear SVM for prediction. This baseline is denoted as pLSA+SVM for the rest of section.

For both pLSA+SVM and MMpLSA, 30% of training postings were randomly chosen for cross validation. For the number of topics from 2 to 10, we ran the experiment five times and report accuracies in the Fig. 1. We can observe that MMpLSA performs much better than unsupervised pLSA+SVM. In other words, supervising the latent topic model can discover discriminative topics for better classification.

We further compared MMpLSA with MedLDA and DiscLDA. Lacoste-Julien et al.[8] reported that DiscLDA achieves best accuracy 83.0% at 60 topics. Zhu et al. [13] didn't report the accuracy of MedLDA, but reported the relative improvement ratio of MedLDA against a two-step approach. The best relative improvement ratio is around 0.2, achieved at 20 topics. We show results of comparison between MMpLSA and them in the Fig. 2.

Fig. 2 (Left) reports both the accuracy mean of five runs of MMpLSA and the accuracy of the run with best cross validation against the best accuracy of DiscLDA. We can see that when the number of topics is small, MMpLSA is noticeably better than DiscLDA. MMpLSA achieved the best accuracy 84.7% with 3 topics, a 2% relative accuracy improvement and a 95% relative number of topics reduction. Therefore, compared with DiscLDA, the max-margin principle used by MMpLSA helps in discovering much fewer topics but with more discriminative information. However, when the number of topics increased, the performance of MMpLSA downgraded. The possible reason is that discriminative information is limited and using more than necessary topics to describe it could cause over-fitting.

Fig. 2 (Right) illustrates relative improvement ratio of accuracy mean of MMpLSA against pLSA+SVM v.s. best relative improvement ratio of MedLDA achieved at 20 topics. MMpLSA is better than MedLDA in all cases. It suggests that MMpLSA has advantages of learning more discriminative topics by using the upper bound of expected classification loss in optimization not the lower bound as MedLDA.

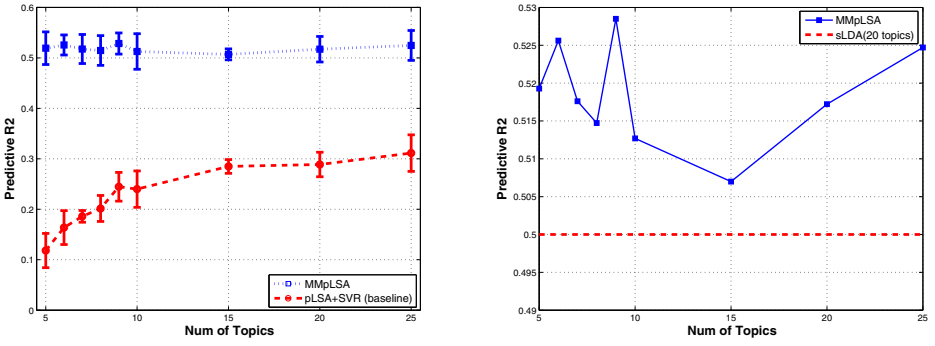

Fig. 3. Experimental results of text regression: (Left) Mean and variance of Predictive R^2s from an 5-fold experiment of the proposed approach MMpLSA and the two-step approach pLSA+SVR; (Right) Predictive R^2 mean of MMpLSA at different numbers of topics v.s. the best result 0.50 of state of art approach sLDA achieved at 20 topics. MMpLSA achieved the best pR^2 0.5285 with 9 topics, a 5.7% relative pR^2 improvement and a 55% relative number of topics reduction.

6.3 Text Regression

To compare MMpLSA with sLDA and MedLDA on regression, we evaluated MMpLSA on the public available movie review dataset [9], in which each review is paired with a rating within $[0, 1]$. The regression task is to predict the rating of a movie review.

To obtain a baseline from two-step approaches, we first fit training reviews to pLSA model, and then used empirical topic proportions as features to train a linear SVR . We denote this baseline as pLSA+SVR.

Following sLDA and MedLDA, we also ran an 5-fold experiment on the same dataset to evaluate pLSA+SVR and MMpLSA, and assessed the quality of predictions by Predictive R^2 (pR^2) as sLDA and MedLDA. In this 5-fold experiment, when one fold was for test, the rest were for training with 25% of reviews randomly chosen for tuning parameters.

Fig. 3 (Left) shows the results. We can see that the supervised MMpLSA can get much better results than the unsupervised two-step approach pLSA+SVR. Moreover, the performance of MMpLSA is consistent for numbers of topics ranging from 5 to 25. It suggests that MMpLSA can discover most discriminative information with few topics and simply increasing number of topics won't improve performance.

We further compared MMpLSA with sLDA and MedLDA. Zhu et al. [13] showed that sLDA and MedLDA have similar performance and MedLDA is only better than sLDA when the number of topics is small. For the 5-fold experiment, the best pR^2 mean was 0.50, achieved by sLDA with 20 topics. Fig. 3 (Right) compares the pR^2 mean of MMpLSA to this best result in the literature. MMpLSA is noticeably better than this best result for all numbers of topics. MMpLSA achieved the best pR^2 mean 0.5285 with 9 topics, a 5.7% relative pR^2 improvement and a 55% relative number of topics reduction.

The experimental result shows again that applying the max-margin principle to supervise latent topic models helps in discovering a much compacter semantic representation with more discriminative information for prediction than state of art approaches.

7 Conclusions and Future Work

We have proposed MMpLSA that applies the max-margin principle to supervise latent topic models for both classification and regression. MMpLSA integrates learning latent topic representations and training a max-margin predictor into one single learning objective. This integration generates topics describing discriminative themes in the corpus so that topic representations of documents are more suitable for prediction. Due to the optimization hardness of single learning objectives, we proposed tight variational upper bounds for them and developed step-wise convex procedures for optimizing those upper bounds. We studied the predictive power of MMpLSA on movie review and 20 Newsgroups data sets, and found that MMpLSA performed noticeably better in prediction with significantly fewer topics than state of art models. These results illustrate the benefits of the max-margin supervised latent topic model when dimension reduction and prediction are the ultimate goals. However, discriminative information in documents is always limited. MMpLSA could possibly over-fit documents if it is asked to discover more discriminative topics than real discriminative topics existing in documents. Therefore, one of future work could be introducing priors to topics in the MMpLSA for alleviating possible over-fitting.

References

1. Blei, D.M., McAuliffe, J.D.: Supervised topic models. In: NIPS, pp. 121–128 (2007)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via plsa. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
4. Burges, C.J.C.: A tutorial on support vector machine for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998)
5. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of SIGIR, pp. 491–501 (1999)
6. Jaakkola, T., Jordan, M.: A variational approach to bayesian logistic regression models and their extensions. In: Proceedings of the 1997 Conference on Artificial Intelligence and Statistics (1997)
7. Klie, S.: An application of latent topic document analysis to large-scale proteomics databases. In: German Bioinformatics Conference (2007)
8. Lacoste-Julien, S., Sha, F., Jordan, M.I.: Disclda: Discriminative learning for dimensionality reduction and classification. In: NIPS, pp. 897–904 (2008)
9. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: ACL (2005)
10. Smola, A., Scholkopf, B.: A tutorial on support vector regression. Statistics and Computing, 199–222 (2003)
11. Xu, W.: Supervising latent topic model for maximum-margin text classification and regression. CMU Technical Report (2009)
12. Zhang, T., Oles, F.: Text categorization based on regularized linear classification methods. Information Retrieval, 5–31 (2001)
13. Zhu, J., Ahmed, A., Xing, E.P.: Medlda: Maximum margin supervised topic models for regression and classification. In: ICML, pp. 1257–1264 (2009)