Sparse Supervised Topic Model: Midterm Report

Yining Wang Chun-Liang Li
Machine Learning Department
Carnegie Mellon University
{yiningwa,chunlial}@cs.cmu.edu

Kevin Lin
Department of Statistics
Carnegie Mellon University
kevinl1@andrew.cmu.edu

Abstract

In this paper we propose the *sparse supervised topic model* (SSTM), a graphical model that learns topic structures of a given document collection and also a *sparse* linear prediction model for response vairables associated with documents. Our model jointly learns the topics and the classifier and encourages a sparse classifier by concentrating all the relevant information for prediction into a small set of topics. Experimental results show that our proposed SSTM model has good interpretability on both classification and regression tasks while still achieves reasonable performance in terms of prediction accuracy.

1 Introduction

Topic models build interpretable topic structures on collections of documents. Commonly used topic models such as Latent Dirichlet Allocation (LDA, [3]) model a collection of documents by representing each document as a mixture of topics where each topic is represented by its own distribution over the words. However, apart from the raw documents, various forms of side information are usually available in practice and can be incorporated into the model. Examples of these include document categories for news articles and rating scores for movie reviews. Such side knowledge could provide useful information for both topic learning and supervised prediction.

There has been extensive study in incorporating side information (response) in supervised topic modeling. Supervised LDA (sLDA, [2]) captures a real-valued linear regression response for each document. The goal is to infer the latent topic representations which is representative of documents as well as predictive for predictions tasks. This idea has been generalized from regression tasks to multi-class classification by multi-class sLDA [16]. The multi-class sLDA replaces the linear response in sLDA with label responses drawn from a softmax regression. The other extended variant of topic models with classification is discriminative LDA (DiscLDA, [7]) which introduces an auxiliary parameters for discrimination built upon the original LDA model. Moreover, DiskLDA optimizes conditional likelihood instead of likelihood or Bayesian posterior, both of which might not be optimal for classification or regression tasks. The idea of using side information was also pushed to the boundary by incorporating the notion of max-margin into the modeling. MedLDA [17] learns both the latent topic representations and supervised classifiers in max-margin sense.

Although the aforementioned supervised topic models give promising performance in practice, the supervised models produced are usually not interpretable because they generally output a large number of latent topics. Ideally, we would like to have a sparse prediction model where only a small number of relevant topics are used to form each prediction.

The notion of sparse topic model has been well studied in [19, 13]. Sparse topical coding (STC, [19]) learns a unsupervised latent topic representations by relaxing normalizing constraint for probabilities to proportions, which allows it to control the sparsity by using l_1 regularization. In [13], it further reduces the training complexity and provides a linear-time convergence algorithm. These approaches are under the unsupervised learning setting and aim to model sparse topic representations for documents. This is different from our goal of the project which aims to find a few topics relevant to the response of each document.

An ad-hoc approach to solve the sparse supervised topic model is as follows. First, use the typical unsupervised topic model[1] to model the documents and then apply any classifier with l_1 regularization [14] to select the topics to achieve a sparse model. We call this the pipeline approach. In this project, we aim to jointly solve the latent topic modeling problem and the sparse regression/classification problem rather than addressing them by the ad-hoc two-step approach. We unify these two steps with a Bayesian graphical model. To the best of our knowledge, there is no thorough study on jointly solving latent topic model and sparse prediction models. A key challenge is that we need to assume a Laplacian prior to impose a sparse structure, but this prior is not conjugate with most likelihood choices and this makes our problem non-trivial. In this project, we would use the data augmentation trick presented in [11] to get around this problem.

2 Sparse Supervised Topic Model

Suppose V is the vocabulary size. Let D denote the number of documents and K denote the number of topics. For each topic k, we use a topic distribution vector $\boldsymbol{\phi}_k \in \Delta^{V-1}$ to represent the word distribution for the kth topic. Here Δ^{V-1} is the probability simplex of a V-dimensional vector. To faciliate a Bayesian treatment, we impose a Dir($\boldsymbol{\beta}$) prior on each topic distribution vector $\boldsymbol{\phi}_k$.

Our proposed Sparse Supervised Topic Model (SSTM) is a Bayesian graphical model. Under SSTM, each document d is generated based on the following procedure:

1. Draw a topic mixing distribution vector $\boldsymbol{\theta}_d$ according to a Dirichlet prior with parameter $\boldsymbol{\alpha}$: $\boldsymbol{\theta}_d|\boldsymbol{\alpha} \sim \text{Dir}(\boldsymbol{\alpha})$.
2. For the n-th word in the document d:
   (a) Draw a topic assignment z_{dn}[2] according to $\boldsymbol{\theta}_d$: $z_{dn}|\boldsymbol{\theta}_d \sim \text{Mult}(\boldsymbol{\theta}_d)$
   (b) Draw the word w_{dn} according to the corresponding topic distribution $\boldsymbol{\beta}_{z_{dn}}$.
3. Draw a response y_d based on the empirical topic assignments $\bar{\boldsymbol{z}}_d = \frac{1}{m}\sum_{n=1}^{m} z_{dn}$ and the model $\boldsymbol{\eta}$. Furthermore,
   - For the regression task we adopt a linear regression model $y_d \sim \mathcal{N}(\boldsymbol{\eta}^\top \bar{\boldsymbol{z}}_d, \sigma^2)$
   - For the classification task we adopt a probit regression model $p(y_d = 1|\boldsymbol{\eta}, \bar{\boldsymbol{z}}_d, \sigma^2) = \Phi(\sigma^{-1}\boldsymbol{\eta}^\top \bar{\boldsymbol{z}}_d)$, where $\Phi(\cdot)$ is the CDF of standard Gaussian distribution.

   Finally, To enforce sparsity on the prediction model, we impose a Laplace prior on $\boldsymbol{\eta}$, i.e., $\boldsymbol{\eta} \sim \text{Laplace}(0, \nu^2)$.

We summarize the generative process as a plate diagram in Figure 1. Let $\mathbf{Y} = \{y_d\}_d$, $\mathbf{W} = \{w_{dn}\}_{dn}$ denote the response variable and words of all documents. Let $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_k\}_k$ and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_d\}_d$ denote the collection of all topic distribution vectors $\boldsymbol{\phi}_k$ and document topic mixing vectors $\boldsymbol{\theta}_d$. Suppose $\mathbf{Z} = \{\boldsymbol{z}_d\}_d$ is the collection of word topic assignments for all documents. Based on the described generative process, the posterior distribution of the proposed SSTM is,

$$p(\boldsymbol{\Theta}, \boldsymbol{\Phi}, \mathbf{Z}, \boldsymbol{\eta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \nu^2, \mathbf{Y}, \mathbf{W}) \propto p(\boldsymbol{\eta}|\nu^2) \prod_{k=1}^{K} p(\boldsymbol{\phi}_k|\boldsymbol{\beta}) \prod_{d=1}^{D} \left(p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) p(y_d|\boldsymbol{\eta}, \bar{\boldsymbol{z}}_d, \sigma) \prod_{n=1}^{N} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|\boldsymbol{\phi}_{z_{dn}}) \right).$$

Again, we reiterate that our proposed model differs from existing models since 1) it models the conditional distribution of the *response* using a sparse combination of topics Z and 2) the notion of sparsity comes from a Laplacian prior which is generally avoided in most methods due to its lack of conjugacy with other distributions. To our knowledge, no existing method achieves this notion of sparsity when predicting responses by using a unified generative model.

[1] Could use either typical topic model or sparse topic model.

[2] z_{dn} is a K-dimensional indicator vector. (ie., only one element is 1, all others are 0)

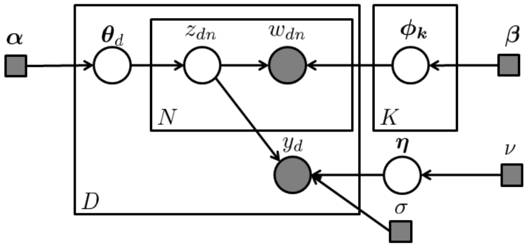

Figure 1: Sparse supervised topic model

3 Parameter inference and prediction of SSTM

In this section, we develop (partially) collapsed Gibbs samplers to infer parameters of both the regression SSTM (SSTM^r) and classification SSTM (SSTM^c) models. For both SSTM^r and SSTM^c, the collapsed posterior distribution of $\boldsymbol{\eta}$ and $\mathbf{Z}$ are well-known [12, 6]:

$$p(\boldsymbol{\eta}, \mathbf{Z}|\mathbf{W}, \mathbf{Y}) \propto p(\boldsymbol{\eta}|\nu^2) \prod_{k=1}^{K} \frac{B(\boldsymbol{c}_k + \boldsymbol{\beta})}{B(\boldsymbol{\beta})} \prod_{d=1}^{D} p(y_d|\boldsymbol{\eta}, \bar{\boldsymbol{z}}_d) \frac{B(\boldsymbol{h}_d + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}, \tag{1}$$

where $B(\boldsymbol{x}) = \frac{\prod_i \Gamma(x_i)}{\Gamma(\sum_i x_i)}$ is the multivariate Beta function; c_k^j is the number of times the jth word is associated with the kth topic and $\boldsymbol{c}_k = \{c_k^j\}_{j=1}^V$; h_d^k is the number of times words are associated with the kth topic in the dth topic and $\boldsymbol{h}_d = \{h_d^k\}_{k=1}^K$.

3.1 Representation with data augmentation

Because the Laplace prior on $\boldsymbol{\eta}$ is not conjugate with the likelihood model, we employ the idea of *data augmentation* to rewrite the Laplacian prior as a mixture model, with the help of an auxiliary variable λ.

Theorem 1 (Data augmentation for Laplacian prior, [10])**.** *Suppose $\boldsymbol{\eta} \sim$ Laplace$(0, \nu^2)$. Then the prior distribution for each η_j can be written as*

$$p(\eta_j|\nu^2) = \frac{e^{-|\eta_j|/\nu}}{2\nu} = \int_0^{+\infty} \frac{e^{-\eta_j^2/2\lambda}}{\sqrt{2\pi\lambda}} \cdot \frac{e^{-\lambda/2\nu^2}}{2\nu^2} \mathrm{d}\lambda =: \int_0^{+\infty} p(\eta_j, \lambda_j|\nu^2) \mathrm{d}\lambda_j. \tag{2}$$

For Probit regression, we employ the following data augmentation trick from [1] to enforce a Gaussian posterior distribution on classification models $\boldsymbol{\eta}$:

Theorem 2 (Data augmentation for Probit regression, [1])**.** *Suppose $p(y = 1|\boldsymbol{\eta}, \boldsymbol{x}, \sigma^2) = \Phi(\sigma^{-1}\boldsymbol{\eta}^\top \boldsymbol{x})$. Consider a data augmentation variable $\gamma \in \mathbb{R}$ and define likelihood $\phi(\boldsymbol{\eta}, \boldsymbol{\gamma}|\boldsymbol{x}, y)$ as*

$$\phi(\boldsymbol{\eta}, \gamma; \boldsymbol{x}, y) := I(y\gamma > 0) \cdot \mathcal{N}(\gamma; \boldsymbol{\eta}^\top \boldsymbol{x}, \sigma^2). \tag{3}$$

The likelihood $p(y|\boldsymbol{\eta}, \boldsymbol{x}, \sigma^2)$ can then be expressed as

$$p(y|\boldsymbol{\eta}, \boldsymbol{x}, \sigma^2) = \int_{-\infty}^{+\infty} \phi(\boldsymbol{\eta}, \gamma; \boldsymbol{x}, y) \mathrm{d}\gamma. \tag{4}$$

3.2 Collapsed Gibbs sampling for SSTM^r

Incorporating data augmentation variables $\{\lambda_k\}_{k=1}^K$, the posterior can be written as

$$p(\boldsymbol{\eta}, \boldsymbol{\lambda}, \mathbf{Z}|\mathbf{W}, \mathbf{Y}) = \frac{1}{Z(\mathbf{W}, \mathbf{Y})} \prod_{k=1}^{K} p(\eta_k, \lambda_k|\nu^2) \frac{B(\boldsymbol{c}_k + \boldsymbol{\beta})}{B(\boldsymbol{\beta})} \prod_{d=1}^{D} p(y|\boldsymbol{\eta}^\top \bar{\boldsymbol{z}}_d, \sigma^2) \frac{B(\boldsymbol{h}_d + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}. \tag{5}$$

We now give a collapsed Gibbs sampling algorithm to infer parameters in SSTM^r. The algorithm samples each parameter from its posterior distribution conditioned on the other sampled parameter values. The Gibbs sampler is collapsed in the sense that the topic dictionary $\mathbf{\Phi}$ and mixing vectors $\mathbf{\Theta}$ are integrated out when computing the conditional distribution of $\mathbf{Z}$.

Update of $\mathbf{Z}$: Fix document d and word n. The conditional distribution of $p(z_{dn} = k)$ can be expressed as

$$p(z_{dn} = k|\mathbf{Z}_{-}, \boldsymbol{\eta}, w_{dn} = t) \propto \frac{(c^{t}_{k,-n} + \beta_t)(h^{k}_{d,-n} + \alpha_k)}{\sum_{j=1}^{V} c^{j}_{k,-n} + \beta_j} \exp\left(-\frac{\eta_k^2}{2\sigma^2 n_d^2} + \frac{\eta_k(y_d - \boldsymbol{\eta}^\top \bar{\boldsymbol{z}}_{d,-n})}{\sigma^2 n_d}\right). \tag{6}$$

Here $c^{\cdot}_{\cdot,-n}$, $h^{\cdot}_{\cdot,-n}$ and $\bar{\boldsymbol{z}}_{\cdot,-n}$ denote topic counts without the nth term in the dth document. n_d indicates the number of words in the dth document.

Update of $\boldsymbol{\eta}$: The conditional distribution of $\boldsymbol{\eta}$ is

$$p(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\lambda}, \mathbf{Y}) \propto \exp\left(-\sum_{d=1}^{D} \frac{(y_d - \boldsymbol{\eta}^\top \bar{\boldsymbol{z}}_d)^2}{2\sigma^2} - \sum_{k=1}^{K} \frac{\eta_k^2}{2\lambda_k}\right). \tag{7}$$

Consequently, $\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\lambda}, \mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ where

$$\boldsymbol{\mu} = \mathbf{\Sigma}\left(\sum_{d=1}^{D} \frac{y_d \bar{\boldsymbol{z}}_d}{\sigma^2}\right), \quad \mathbf{\Sigma} = \left(\mathrm{diag}(\lambda_1^{-1}, \cdots, \lambda_K^{-1}) + \sum_{d=1}^{D} \frac{\bar{\boldsymbol{z}}_d \bar{\boldsymbol{z}}_d^\top}{\sigma^2}\right)^{-1}.$$

Update of $\boldsymbol{\lambda}$: Given the regression model $\boldsymbol{\eta}$, the conditional distribution of each augmented variable can be expressed as

$$p(\lambda_k|\eta_k) = \mathcal{GIG}(\lambda_k; p, a, b) = \mathcal{GIG}(\lambda_k; \frac{1}{2}, \frac{1}{\nu^2}, \eta_j^2), \tag{8}$$

where $\mathcal{GIG}(x; p, a, b) = C(p, a, b)x^{p-1}\exp(-\frac{1}{2}(\frac{b}{x} + ax))$ is a generalized inverse Gaussian distribution [4] and $C(p, a, b)$ is a normalizing constant. Subsequently, λ_k^{-1} follows an inverse Gaussian distribution:

$$p(\lambda_k^{-1}|\eta_k) = \mathcal{IG}(\lambda_k; \frac{1}{\nu|\eta_k|}, \frac{1}{\nu^2}), \tag{9}$$

where $\mathcal{IG}(x; a, b) = \sqrt{\frac{b}{2\pi x^3}}\exp(-\frac{b(x-a)^2}{2a^2 x})$ for $a, b > 0$. Note that λ_k^{-1} can be sampled from an inverse Gaussian distribution in $O(1)$ time [8].

3.3 Partially collapsed Gibbs sampling for SSTM^c

Incorporating data augmentation variables $\{\lambda_k\}_{k=1}^{K}$ and $\{\gamma_d\}_{d=1}^{D}$, the collapsed posterior distribution of $\boldsymbol{\eta}$ and $\mathbf{Z}$ can be expressed as

$$p(\boldsymbol{\eta}, \mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\gamma}|\mathbf{Y}, \mathbf{W}) = \frac{1}{Z(\mathbf{Y}, \mathbf{W})} \prod_{k=1}^{K} p(\eta_k, \lambda_k|\nu^2) \frac{B(\boldsymbol{c}_k + \boldsymbol{\beta})}{B(\boldsymbol{\beta})} \prod_{d=1}^{D} \phi(\boldsymbol{\eta}, \gamma_d; \bar{\boldsymbol{z}}_d, y_d) \frac{B(\boldsymbol{h}_d + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}. \tag{10}$$

We now give a Partially Collapsed Gibbs (PCG) sampling algorithm to infer parameters in SSTM^c. The sampling algorithm is partially collapsed in the sense that when computing the conditional distribution of $p(z_{dn})$ the augmented variable γ_d is integrated out. [15] shows that the partially collapsed sampling approach improves convergence rate of the Gibbs sampling algorithm. Note that the inference rule of augmented variable $\boldsymbol{\lambda}$ is exactly the same with the one in SSTM^r.

Update of $\mathbf{Z}$: The conditional distribution of $z_{dn} = k$ can be expressed as

$$p(z_{dn} = k|\mathbf{Z}_{-}, \boldsymbol{\eta}, w_{dn} = t) \propto \frac{(c^{t}_{k,-n} + \beta_t)(h^{k}_{d,-n} + \alpha_k)}{\sum_{j=1}^{V} c^{j}_{k,-n} + \beta_j} \Phi\left(\boldsymbol{\eta}^\top \tilde{\boldsymbol{z}}_d^{n,k}\right)^{\tilde{y}_d} \left(1 - \Phi\left(\boldsymbol{\eta}^\top \tilde{\boldsymbol{z}}_d^{n,k}\right)\right)^{1-\tilde{y}_d}, \tag{11}$$

where $\tilde{y}_d = (1 + y_d)/2 \in \{0, 1\}$ and $\tilde{\boldsymbol{z}}_d^{n,k} = \bar{\boldsymbol{z}}_{d,-n} + \boldsymbol{e}_k/n_d$. Note that in Eq. (11) we integrate out the data augmentation variable γ_d in order to obtain a partially collapsed Gibbs sampler.

Update of $\boldsymbol{\eta}$: The conditional distribution of $\boldsymbol{\eta}$ is

$$p(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \propto \exp\left(-\sum_{d=1}^{D} \frac{(\gamma_d - \boldsymbol{\eta}^\top \bar{\boldsymbol{z}}_d)^2}{2\sigma^2} - \sum_{k=1}^{K} \frac{\eta_k^2}{2\lambda_k}\right). \tag{12}$$

Consequently, $\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\sum_{d=1}^{D} \frac{\gamma_d \bar{\boldsymbol{z}}_d}{\sigma^2}\right), \quad \boldsymbol{\Sigma} = \left(\operatorname{diag}(\lambda_1^{-1}, \cdots, \lambda_K^{-1}) + \sum_{d=1}^{D} \frac{\bar{\boldsymbol{z}}_d \bar{\boldsymbol{z}}_d^\top}{\sigma^2}\right)^{-1}. \tag{13}$$

Update of $\boldsymbol{\gamma}$: The update rule of $\boldsymbol{\gamma}$ is very simple. Suppose ϑ_d follows a normal distribution with zero mean and σ^2 variance. Then

$$\gamma_d \sim \begin{cases} \mathcal{N}_{(0,\infty)}(\eta^\top \bar{\boldsymbol{z}}_d, \sigma), & \text{if } y_d = 1; \\ \mathcal{N}_{(-\infty,0)}(\eta^\top \bar{\boldsymbol{z}}_d, \sigma), & \text{if } y_d = 0, \end{cases} \tag{14}$$

where $\mathcal{N}_{(a,b)}$ is the truncated normal distribution between a and b.

3.4 Prediction

To apply the learned classification/regression model $\boldsymbol{\eta}$ one needs to sample topic assignments $\boldsymbol{z}_d$ for a new document d. To do this, we use a point estimate of the topic dictionary $\widehat{\boldsymbol{\Phi}}$ and then sample $\boldsymbol{z}_d$ from its posterior distribution, integrating out the topic mixing vector $\boldsymbol{\theta}_d$. Similar approaches were also taken in [17, 18] to perform prediction. More specifically, the MAP estimator of $\widehat{\boldsymbol{\Phi}}$ has the form $\widehat{\phi}_{k,t} \propto c_k^t + \beta_t$ for $k \in [K]$ and $t \in [V]$. Afterwards, each column in $\widehat{\boldsymbol{\Phi}}$ is normalized so that the probabilities sum to one. Given the estimate $\widehat{\boldsymbol{\Phi}}$, the latent topic assignment for each word z_{dn} can be sampled from a categorical distribution as $p(z_{dn} = k|\boldsymbol{z}_{d,-n}, \widehat{\boldsymbol{\Phi}}, w_{dn} = t) \propto \widehat{\phi}_{k,t}(h_{d,-n}^k + \alpha_k)$.

4 Experiments

In this section we report experimental results of our proposed SSTM model and its competitors on real world document datasets. We first briefly introduce the datasets we used and implementation details of the algorithms. Quantitative results for classification and regression then follow. Finally, representative words in the learned topics are presented to provide an intuition of the objective of our proposed algorithms.

4.1 Datasets and implementation details

For the classification task we use the BBC news dataset [5]. The dataset consists of 2225 news articles and 9636 terms built from the BBC news website corresponding to news in five topical areas from 2004 to 2005. The five topics include business, entertainment, politics, sports and technology. For the regression task we use the movie rating dataset [9] which is originally created for sentiment analysis. The dataset contains 30286 terms over 5006 movie review documents, each rated according to preference ranging from 0 to 10.

For each dataset we divide the documents into two groups of roughly equal size and use one for training/validating and the other one for held-out testing. We run 5-fold cross-validation to screen parameters on the training/validating dataset for all algorithms and use the parameter setting that gives the best classification/regression performance for cross-validation. For a total of 50 parameter settings the parameter screening step is completed within two days on a Opterion 6380 server with four 16-core CPUs and 256G RAM.

We compare the performance of our proposed SSTM model with a pipeline algorithm and also sLDA, a supervised topic model with dense prediction weights [2]. For the pipeline algorithm, we first run vanilla LDA using collapsed Gibbs sampling and then run ℓ_1 regularized logistic regression for binary classification and LASSO for linear regression, built on learnt latent topic representations. The sLDA model is also trained using collapsed Gibbs sampling. All training and testing routines are implemented in C++.

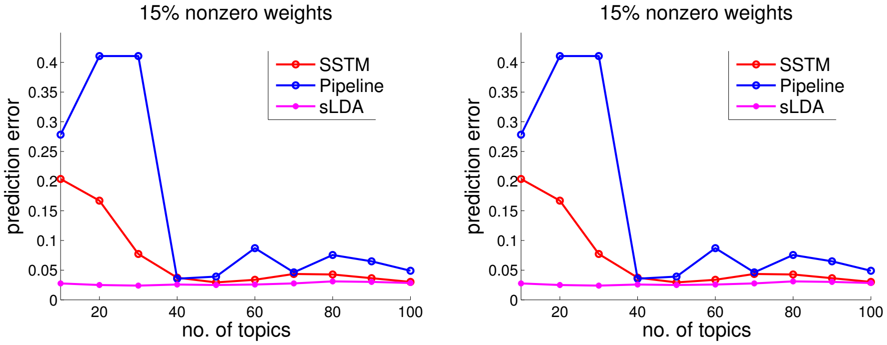

Figure 2: Classification error vs. number of topics (K) under 10% (left) and 15% (right) constrained sparsity level.

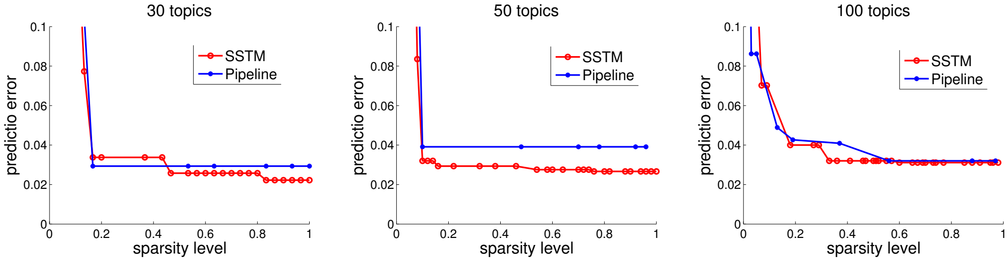

Figure 3: Classification error vs. resulting sparsity level for $K = 30, 50, 100$.

4.2 Prediction error

We first compare the classification error of SSTM, sLDA and the pipeline algorithm on the BBC news dataset in Figure 2 to perform binary classification. We divided the documents associated with the five different subjects into a positive group (all entertainment and sports) and a negative group (all politics, technology and business). The number of topics (K) ranges from 10 to 100 and we constrain the sparsity level of SSTM and the pipeline method. By saying an output classifier has 10% sparsity we mean that at most 10% of the weights are nonzero.

Figure 2 shows that our proposed SSTM model consistently outperforms the pipeline one in terms of classification error. On the other hand, the dense supervised LDA result outperforms both SSTM and pipeline algorithm by a large margin. This is not surprising, because sLDA used far more topics than SSTM and the pipeline method and as a result its model interpretability is sacrificed.

We also compare the classification error of SSTM and pipeline when K is fixed and the sparsity level changes. In Figure 3 we plot the classification error of both algorithms for $K = 30, 50$ and 100 with sparsity level of the output weight vector ranging from 0 (completely sparse) to 1 (no sparsity at all). Figure 3 shows that under most settings the joint SSTM model outperforms the pipeline one. Furthermore, the third figure demonstrates a fast error decay of SSTM than the pipeline model. We conjecture that information relevant to the prediction task is more concentrated in the topics learned by SSTM and hence to achieve the same level of prediction error the pipeline solution uses far more topics than the SSTM solution.

For the regression task, we report the mean square error (MSE) on the movie rating dataset for SSTM, sLDA and the pipeline method in Figure 4. Similar to Figure 2, in Figure 4 the sparsity level of output regression models are constrained while the number of topics changes. One difference is that we use more topics for regression than classification because estimating movie ratings is related to sentiment analysis and is much harder than news categorization. We can see that still in most cases the SSTM model outperforms the pipeline method and in general supervised LDA with dense regression models works much better.

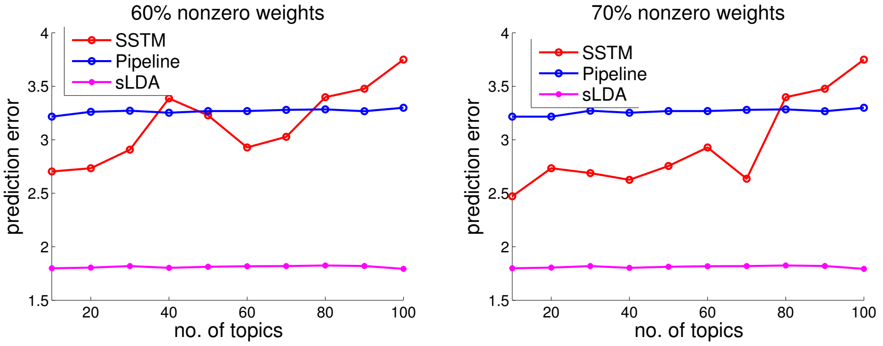

Figure 4: MSE (regression error) vs. number of topics (K) under 60% (left) and 60% (right) constrained sparsity level.

	Relevant Topics						Irrelevant Topics		
η_j	-1.25	-0.75	-1.16	-1.16	1.24	1.35	0		0
	labour	people	year	game	film	game	people	···	on
	elect	user	economics	people	best	plays	year	···	year
	party	security	growth	mobile	award	win	on	···	time
	tori	site	on	technology	music	player	uk	···	plays
	govern	software	economy	phone	star	england	film	···	people
	people	year	rate	digit	include	against	time	···	go
	tax	microsoft	market	year	actor	first	work	···	company

Table 1: Table showing the top seven words of the 6 relevant topics (topics associated with non-zero coordinates of η) as well as 2 arbitrarily selected irrelevant topics. The 6 relevant topics are (from left to right) about politics, technology, business, technology, entertainment and sports.

4.3 Learned topics

We display the learned words and topics for the BBC dataset. The results shown in Table 1 are the $K = 50, \nu = 0.05, \sigma = 0.5, \alpha = .\vec{1}, \beta = .\vec{1}$. These topics obtained a misclassification rate on testing data of 0.03. The resulting η had only 6 non-zero coordinates, and we display their associated topics. Of the five subjects in the BBC dataset, we have at least one topic dedicated to each subject. As expected, the η_j's for politics, technology and business topics are negative while the η_j's for entertainment and sports are positive.

We also included the top 7 words for two arbitrarily chosen topics associated with $\eta_j = 0$. Based on the top 7 words, we do not strongly feel that these topics contain information useful for prediction. Hence, our algorithm partitioned relevant and irrelevant topics. This is a encouraging result showing that our algorithm automatically concentrated the most relevant words for prediction within a small set of topics. We do not expect the naive pipeline approach to achieve such a sparse set of topics to achieve the same prediction accuracy.

5 Discussion

In general prediction (giving accurate classification/prediction results) and feature selection (picking variables/topics that are most relevant to a prediction task) are quite different and sometimes are competing objectives. There might be concerns over our approach in that the performance on both tasks (prediction and feature selection) could suffer by solving them jointly. However, if topics are first obtained via unsupervised or supervised topic models without sparsity regularization, it could be very hard to select a handful of highly relevant topics because there is no incentive for the first phase of this type of naive approach to concentrate relevant information into a few topics. As a result, many topics could end up relevant to the prediction task, as shown in the third plot in Figure 3.

Another motivation for obtaining sparse weight vectors for supervised topic modeling concerns *model interpretability*. Consider the case when there are thousands of topics used to train a linear classifier. Though the model might perform well, there are too many topics used for prediction to reasonably determine which topics are more important than others for the prediction task.

Finally, we remark that our model can be easily generalized to the multi-task setting where each document has M labels for M different tasks and one wants to build M prediction models for each task based on shared latent topic representations. Under such settings our proposed SSTM model will output a specific small set of relevant topics for each task. On the other hand, this could be very difficult for unsupervised or ℓ_2 regularized supervised topic models because there is no incentive for these methods to concentrate information for different prediction tasks during the topic learning phase.

References

[1] J. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[2] D. Blei and J. McAuliffe. Supervised topic models. In *NIPS*, 2007.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, (3):993–1022, 2003.

[4] L. Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, 1986.

[5] D. Greene and P. Cunningham. Producing accurate interpretable clusters from high-dimensional data. In *European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2005.

[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[7] S. Lacoste-Julien, F. Sha, and M. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008.

[8] J. R. Michael, W. R. Schucany, and R. W. Haas. Generating random variates using transformations with multiple roots. *The American Statistician*, 30(2):88–90, 1976.

[9] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

[10] T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[11] N. Polson and S. Scott. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

[12] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *SIGKDD*, 2008.

[13] K. Than and T. B. Ho. Fully sparse topic models. In *ECML/PKDD*, 2012.

[14] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

[15] D. A. Van Dyk and T. Park. Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796, 2008.

[16] C. Wang, D. Blei, and F.-F. Li. Simultaneous image classification and annotation. In *CVPR*, 2009.

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

[18] J. Zhu, N. Chen, H. Perkins, and B. Zhang. Gibbs max-margin topic models with data augmentation. *Journal of Machine Learning Research*, (15):1073–1110, 2014.

[19] J. Zhu and E. Xing. Sparse topic coding. In *UAI*, 2011.