Boost Multi-class sLDA Model for Text Classification

Maciej Jankowski(✉)

Faculty of Cybernetics, Military University of Technology in Warsaw,
Warsaw, Poland
maciej.jankowski@wat.edu.pl

Abstract. Text classification is an important problem in Natural Language Processing. It differs from many other classification tasks by the large number of features that have to be used during training. One of the solution for reducing dimensionality of feature space, is the usage of Latent Dirichlet Allocation. After this step, the smaller problem can be solved using standard classifiers. In [11], authors propose combination of LDA and Softmax classifier called Multi-class sLDA, that does both tasks simultaneously. However, to use the method, we have to choose a number of topics - hyperparameter of the model. This step requires analysis and human supervision. In this paper, we propose Boost Multi-class sLDA model, based on ensemble of many Multi-class sLDA models, that does not require the choice of topic number. Moreover, our model achieves significantly better classification accuracy, than Multi-class sLDA for any number of topics.

1 Introduction

Topic models are very popular methods of text analysis. The most popular algorithm for topic modeling is Latent Dirichlet Allocation (LDA) [5]. Recently, many new methods were proposed, that enable the usage of this model in large scale processing. One of the extension to the LDA is Supervised Latent Dirichlet Allocation (sLDA) [9], which adds to LDA a response variable Y, connected to each document. This response variable is real valued and drawn from a linear regression. For classification purposes, continuous output is not appropriate, therefore a new model called Multi-class sLDA was proposed in [11]. This model can be successfully used for classification, competing with state of the art classifiers. Its advantage lies in the intermediate step that reduces dimension and can possibly find useful features like synonyms and polysemes.

One of the problem is, that choice of the number of topics K, is not happening automatically and requires some previous analysis. A few methods were proposed so far to automatize this step [6,10,12,15], but none of them works very well in supervised tasks. In this paper, we develop an ensemble algorithm that consists of many Multi-class sLDA models with different numbers of topics. We show, that this ensemble works better than any single model. In addition to

L. Rutkowski et al. (Eds.): ICAISC 2018, LNAI 10841, pp. 633–644, 2018.
https://doi.org/10.1007/978-3-319-91253-0_59

improving accuracy, the usage of this ensemble allows us to avoid the manual step of choosing the number of topics.

The rest of this paper is structured as follows: in Sect. 2, we briefly discuss the problem of dimensionality reduction in text analysis and the role LDA-based methods to achieve this goal. In Sect. 3, we provide definition of Multi-class sLDA model. We also describe in detail one of the approach to carry out inference, estimation and prediction in this model. In Sect. 4 we introduce our approach called Boost Multi-class sLDA. In Sect. 5, we study the performance of our model. Finally in Sect. 6 we summarize our findings.

2 Dimensionality Reduction for Text Classification

Classification of texts is a challenging subject. Treating words as individual features, leads to a very large set. This may pose problems for classifiers. Some features may be correlated (synonyms), others may have multiple meanings. Computational cost for large number of features may be prohibitive for large corpora. Because of those reasons, we are interested in preprocessing raw text in a way that reduces its dimension. One way to achieve that is to use LDA [5] in an unsupervised manner. However, in the context of classification, choosing the best number of topics, may not be possible using only unsupervised methods. If for example, our task is to classify movies as good or bad, we want to find topics that are somehow related to sentiments. However, if the dominant structure in reviews is genre, this is something that may be found. It is possible, that best value of K for LDA, is not the best value of K for Multi-class sLDA.

3 Multi-class sLDA Model

In this section, we describe Multi-class Supervised Latent Dirichlet Allocation (Multi-class sLDA) [11], a supervised method for classification, that builds on previous models for Topic Modeling [5] and [9]. Notation used in this paper is summarized in Table 2. We use $Dir(\alpha)$ for Dirichlet distribution with parameter α, and $Mult(1, \tau)$ for multinomial distribution with single trial and probability of success of each outcome described by vector τ.

Let K be a fixed number of topics. For a given text corpus $\mathcal{T}$, we define V to be a number of words in dictionary, M be a number of documents in corpus and C be a number of classes, to which each document can belong. We further assume, that there is a corpus dependent parameter $\alpha \in \mathbb{R}^K$, parameter $\beta \in \mathbb{R}^{K \times V}$ and parameter $\eta \in \mathbb{R}^{C \times K}$. Each document $d \in \{1, \ldots, M\}$, is assumed to be generated from the following process:

1. Draw topic proportions $\theta_d | \alpha \sim Dir(\alpha)$
2. For each word $w_{d,n}, n \in \{1, 2, \ldots, N_d\}$:
   (a) Draw topic assignment $z_{d,n} | \theta \sim Mult(1, \theta_d)$
   (b) Draw word $w_{d,n} | z_{d,n}, \beta_{z_{d,n}} \sim Mult(1, \beta_{z_{d,n}})$

3. Draw class label $c_d|z_d, \eta \sim softmax(\bar{z}_d, \eta)$, where

$$\bar{z}_d = \frac{1}{N_d} \sum_{n=1}^{N_d} z_{d,n}$$

is the empirical topic frequencies ($\bar{z}_d \in \mathbb{R}^K$), and the softmax distribution is given by

$$p(c_d|\bar{z}_d, \eta) = \frac{\exp(\eta_{c_d}^T \bar{z}_d)}{\sum_{l=1}^{C} \exp(\eta_l^T \bar{z}_d)}, \qquad c_d = 1, \ldots, C$$

In the above definition, $z_{d,n} \in \mathbb{R}^K$, is an indicator vector with all elements equal to zero except one that equals 1, for example $(0, \ldots, 0, 1, 0, \ldots, 0)$. Each of those vectors, denote a single integer between 1 and K. When we write $\beta_{z_{d,n}}$, then subscript means the integer denoted by indicator vector (e.g. $\beta_{(0,1,0)}$ means β_2 and $\beta_{(0,0,1)}$ means β_3). Similarly, we often write $\beta_{i,w_{d,n}}$. Since $w_{d,n}$ is also indicator variable, when we write $\beta_{i,(0,0,1)}$ we mean $\beta_{i,3}$.

The classification problem is as follows: we have input of M vectors, each constitutes a topic proportion θ_d. For each θ_d, we want to estimate the probability of the class label taking on each of the C different possible values.

3.1 Multi-class sLDA Computation

Standard approach to finding parameters of mixture models is to use Expectation Maximization algorithm [1]. In this algorithm, we alternate between two steps: E - where we find posterior distribution of mixture component given parameters, and M - in which we estimate parameters using distribution from E step. In this form, the algorithm cannot be used for Multi-class sLDA, because E step is intractable [5]. The main challenge is therefore, to approximate the distribution $p(z|w_d)$, in an efficient way. One of the solution, called Variational Inference, is based on approximating the distribution with a family of simpler distributions. Standard approach in LDA based models is to use fully factorized distribution. The method is known as mean field approximation. We will not present details of this algorithm, for references see [5,9,11,17]. Instead, we will focus on those details, that are needed to develop Boost Multi-class sLDA algorithm.

3.2 Approximate Inference

In inference part (E-step in variational EM), we approximate values of parameters $\gamma_d \in \mathbb{R}^{N_d \times K}$ (parameters of Dirichlet prior for θ) and $\phi_d \in \mathbb{R}^K$ (Dirichlet prior for z). Inference is carried out for a single document, therefore index d responsible for a document does not vary, but is fixed. Let $\pi = \{\alpha, \beta_d, \eta\}$ denote the set of model parameters and $q(\theta_d, z_d|\gamma, \phi)$ is join variational distribution of θ_d and z_d which factorizes as follows

$$q(\theta_d, z_d|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N_d} q(z_n|\phi_n) \tag{1}$$

The variational objective function $\mathcal{L}$ also known as the evidence lower bound (ELBO), is an expectation with respect to latent variables z that follow an approximating distribution q. For Multi-class sLDA, ELBO is

$$\begin{aligned}\mathcal{L}_d(\gamma_d, \phi_d; \pi) = \mathbb{E}_q[\log p(\theta_d|\alpha)] + \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(Z_{d,n}|\theta_d)] \\ + \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(w_{d,n}|Z_{d,n}, \beta_d)] + \mathbb{E}_q[\log p(c|Z_d, \eta)] + \mathbf{H}(q) \qquad (2)\end{aligned}$$

where $\mathbf{H}(q)$, is the entropy of the variational distribution. Maximizing this lower bound with respect to γ and ϕ leads to the following pair of updates of an iterative fixed-point method

$$\gamma_{d,i} = \alpha_i + \sum_{n=1}^{N_d} \phi_{d,n,i}, \qquad i \in \{1, \dots, K\} \qquad (3)$$

$$\phi_{d,n,i} \propto \beta_{i,w_n} \exp\left(\Psi(\gamma_{d,i}) + \frac{1}{N}\eta_{c,i} - (h^T \phi_n^{(old)})^{-1} h_i\right) \qquad (4)$$

where

$$h^T \phi_n = \sum_{c=1}^{C} \prod_{n=1}^{N_d} \Big(\sum_{k=1}^{K} \phi_{dnk} \exp\big(\frac{1}{N_d}\eta_{ck}\big)\Big)$$

3.3 Estimation

Main goal of estimation (M-step in variational EM), is to find maximum likelihood estimates of topics β_d, $d \in \{1, \dots, M\}$, and class coefficients η_c, $c \in \{1, \dots, C\}$. Corpus log-likelihood is

$$\mathcal{L}(\mathcal{T}) = \sum_{d=1}^{M} \log p(w_d, c_d|\alpha, \eta, \beta) \qquad (5)$$

Optimizing with respect to $\beta_{k,i}$, $k \in \{1, \dots, K\}, i \in \{1, \dots, V\}$, leads to the following update rule for β

$$\beta_{k,i} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} w^i_{d,n} \phi_{d,n,i} \qquad (6)$$

where

$$w^i_{d,n} = \begin{cases} 1, & \text{if } w_{d,n} = (0, \dots, 0, \underset{i}{1}, 0, \dots, 0) \\ 0, & \text{if } w_{d,n} = (0, \dots, 0, \underset{j \neq i}{1}, 0, \dots, 0) \end{cases}$$

Terms in ELBO containing η are

$$\mathcal{L}_{[\eta]}(\mathcal{T}) = \sum_{d=1}^{D} \left(\eta_{c_d}^T \bar{\phi}_d - \log \Bigg(\underbrace{\sum_{e=1}^{C} \prod_{n=1}^{N_d} \Bigg(\sum_{k=1}^{K} \phi_{dnk} \exp \big(\frac{1}{N_d} \eta_{ek} \big) \Bigg)}_{u} \Bigg) \right)$$

Optimization for η is done using conjugate gradient. This method requires first derivative

$$\frac{\partial \mathcal{L}_{[\eta]}(\mathcal{T})}{\partial \eta_{ci}} = \sum_{d=1}^{M} I[c_d = c] \bar{\phi}_{di} - \sum_{d=1}^{M} \frac{1}{u} \frac{\partial u}{\partial \eta_{ci}} \tag{7}$$

3.4 Prediction

To classify new unseen document, we calculate approximate probability of each class c given document w_d, and choose a class with the highest probability

$$c^* = \underset{c \in \{1,\ldots,C\}}{\operatorname{argmax}} \ \mathbb{E}_q[\eta_c^T \bar{z}] = \underset{c \in \{1,\ldots,C\}}{\operatorname{argmax}} \ \eta_c^T \bar{\phi}$$

The denominator is constant, therefore we do not need to calculate it. However, it is possible to approximate probability of each class given document w_d

$$p(c|w_d) = \exp \left(\eta_c^T \bar{z} - \log \left(\sum_{l=1}^{C} \exp(\eta_l^T \bar{z}) \right) \right).$$

Because we do not know real value of $\bar{z}$, we approximate the probability as described in [11], using expectation with respect to variational distribution $\mathbb{E}_q$

$$\begin{aligned} p(c|w_d) &\approx \mathbb{E}_q \left[\exp \left(\eta_c^T \bar{z} - \log \left(\sum_{l=1}^{C} \exp(\eta_l^T \bar{z}) \right) \right) \right] \\ &\geq \exp \left(\mathbb{E}_q[\eta_c^T \bar{z}] - \mathbb{E}_q \left[\log \left(\sum_{l=1}^{C} \exp(\eta_l^T \bar{z}) \right) \right] \right) \\ &\geq \exp \left(\eta_c^T \bar{\phi} - \log \left(\sum_{l=1}^{C} \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \phi_{n,k} \exp \big(\frac{1}{N} \eta_{l,k} \big) \right) \right) \right) \end{aligned} \tag{8}$$

4 Boost Multi-class sLDA

One of the problem with topic models is, that we need to know the number of topics, before we start to train the model. This is a standard example of model selection. The simplest approach is to run estimation for different values of K, and test each model on a validation set. Finally choosing the number that gives the best performance. In the context of LDA model, many approaches to this problem have been proposed in the literature [6,10,12,15]. Instead of trying to find the best number of topics, we can train many models and validate them, finally rejecting those that do not pass the acceptance criteria. A very popular approach for criticizing Bayesian models is called Posterior Predictive Checking (PPC) [3]. This idea has been applied to LDA model in [13]. In [8], authors propose a model known as Hierarchical Dirichlet Process (HDP), which chooses the best number of topic automatically and separately for each document. All those methods (with the exception of HDP), were designed for LDA, which is unsupervised model.

4.1 Ensemble

In this section, we develop Boost Multi-class sLDA, an ensemble algorithm for Multi-class sLDA, based on AdaBoost algorithm. We will show, that ensemble of many different models, each with different number of topics, outperforms the best single model. These results, allow us to avoid the manual step of choosing single value of K.

The following description addresses binary classification problem, but it can be easily extended to multiclass case. We therefore assume, that response variable Y takes values in set $\{0,1\}$. We denote each Multi-class sLDA classifier as $\mathcal{M}_l : \{0,1\}^{N_d \times V} \to \{-1,1\}$. This function is defined as

$$\mathcal{M}_l(w_d) = \operatorname*{argmax}_{c \in \{-1,1\}} p(Y = c | w_d),$$

and can be calculated as described in Sect. 3.4. Ensemble of L Multi-class sLDA classifiers is a function $\mathcal{M} : \{0,1\}^{N_d \times V} \to \{-1,1\}$, that depends on the choice of the ensemble algorithm

$$\mathcal{M}(w_d) = f(\mathcal{M}_1(w_d), \ldots, \mathcal{M}_L(w_d))$$

4.2 AdaBoost for Multi-class sLDA

AdaBoost is very popular learning algorithm introduced in [2]. The purpose of the algorithm is to apply classification algorithm to repeatedly modified versions of the data. This way, we obtain many instances of the classifier $\mathcal{M}_1, \ldots, \mathcal{M}_L$, each with different characteristics. To predict outcome for a new unseen document, we apply weighted majority voting

$$\mathcal{M}(w_d) = \operatorname{sign}\left(\sum_{l=1}^{L} \alpha_l \mathcal{M}_l(w_d) \right),$$

where $\alpha_1, \ldots, \alpha_L$ are computed by AdaBoost. Algorithm works as follows, in first iteration $l = 1$, all observations are given weights $r_d^{(l)} \in \mathbb{R},\ d = 1, \ldots, M$, that control, how much attention is classification algorithm giving to those observations. Then, classifier is trained on those weighted observations. In next step, weights are recalculated in such a way, that misclassified observations have their weights increased. New classifier is trained on those new weights $r_d^{(l+1)} \in \mathbb{R},\ d = 1, \ldots, M$. Described steps are repeated many times. For further details and analysis, see for example [4] or [7].

In this paper we present a combination of Multi-class sLDA and AdaBoost algorithms. For those two algorithms to work together, we have to modify corpus level ELBO in such a way, that it take into account observation weights. No changes to inference step are required, because it is done separately for each document.

4.3 Changes to Estimation of Parameters

Corpus level ELBO is a sum of document level ELBOs. Therefore, we can apply observation weights $r_d^{(l)},\ d = 1, \ldots, M$, to elements of this sum. Equation 5 changes to

$$\mathcal{L}(\mathcal{T}|\mathcal{M}_l) = \sum_{d=1}^{M} r_d^{(l)} \log p(w_d, c_d|\mathcal{M}_l).$$

Update rules for β were obtained by differentiating 5 with respect to $\beta_{k,i}$. Simple calculation shows, that Eq. 6 changes to

$$\beta_{k,i}^{(l)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} r_d^{(l)} w_{d,n}^i \phi_{d,n,i}^{(l)} \tag{9}$$

Optimization for η is based on conjugate gradient that requires first derivative of optimized function. Equation 7 changes to

$$\frac{\partial \mathcal{L}_{[\eta]}(\mathcal{T}|\mathcal{M}_l)}{\partial \eta_{ci}^{(l)}} = \sum_{d=1}^{M} r_d^{(l)} \mathbb{1}[c_d = c] \bar{\phi}_{di}^{(l)} - \sum_{d=1}^{M} r_d^{(l)} \frac{1}{u} \frac{\partial u}{\partial \eta_{ci}^{(l)}}. \tag{10}$$

Final version is presented in Algorithm 1.

5 Empirical Study

In this section, we provide an example of the use of a Ensemble Multi-class sLDA model on real data. For this purpose, we use two datasets: first consists of 942 (742 train, 200 test), documents containing a set of SMS labeled messages that have been extracted from SMS Spam Collection Data Set available on UCI Machine Learning Repository site and first introduced in [14]. The second dataset, is a collection of 773 (573 train, 200 test) political blogs available as part of the LDA R package [16]. Both datasets were split to train and test datasets. Train datasets were used for training models. Final accuracy was measured on test datasets.

Algorithm 1. Boost Multi-class sLDA

```
1. Initialize weights for first model $r^{(1)} \in \mathbb{R}^M$, $r_d^{(1)} = 1/M$, $d = 1, 2, \ldots, M$
2. For $l = 1$ to $L$:
    (a) Fit Multi-class sLDA model $\mathcal{M}_l$ to the training data using weights $r^{(l)}$, using VEM algorithm with the following modifications
        i. Inference part is done exactly as described in Sect. 3.2
        ii. Estimate parameters as described in Sect. 4.3
    (b) Compute
        $$\mathrm{err}_l = \frac{\sum_{d=1}^{M} r_d^{(l)} I[c_d \neq \mathcal{M}_l(w_d)]}{\sum_{d=1}^{M} r_d}.$$
    (c) Compute $\alpha_l = \log((1 - \mathrm{err}_l)/\mathrm{err}_l)$.
    (d) Set $r_d^{(l+1)} \leftarrow r_d^{(l)} \cdot \exp[\alpha_l \cdot I[c_d \neq \mathcal{M}_l(w_d)]])$, $d = 1, 2, \ldots, M$.
3. Output
    $$\mathcal{M}(w_d) = \mathrm{sign}\left(\sum_{l=1}^{L} \alpha_l \mathcal{M}_l(w_d)\right),$$
```

5.1 Multi-class sLDA for Various Number of Topics

In this experiment, we have compared Multi-class sLDA models with different number of topics. We have chosen $K = 5, 10, 15, 20, 30, 40, 50, 75, 100$, and for each value we have trained the model and reported its error rate on test as well as on train datasets. Results are presented in Fig. 1. Left picture (A) contains error rates for SMSSpam dataset. As shown, the model overfitted a little bit for 15 topics. Overall, error on test dataset oscillates somewhere between 18% and 29%. This discrepancy is a clear indication, that the choice of K, has great influence on the accuracy of the algorithm. Right picture (B), contains results for Poliblog dataset. As we can see, this model suffers from severe overfitting. While error on train dataset is around 22%, error on test dataset equals 29% at best.

Both pictures, also contain results on test set for ensemble. As we will see in the next section, boosting improves the accuracy and does not require the choice of K.

5.2 Boost Multi-class sLDA for Varying Number of Topics

In the second experiment, we have researched the ability of Boost Multi-class sLDA to improve classification accuracy, when the number of topics is unknown. We have chosen 5 models with K equal to 5, 10, 15, 20 and 30. First model was trained using $K = 5$, second model used $K = 10$ and so on, up to $K = 30$. After training 5 models, we started next iteration for $K = 5$ again, then $K = 10$ and so forth. We stopped, when the accuracy on a test was not improving. Results for SmsSpam and Poliblog datasets, were presented in Fig. 2. Left picture (A), contains error rate for successive boosting iterations for SmsSpam test dataset. As we can see, Boost Multi-class sLDA kept improving up to 38th iteration,

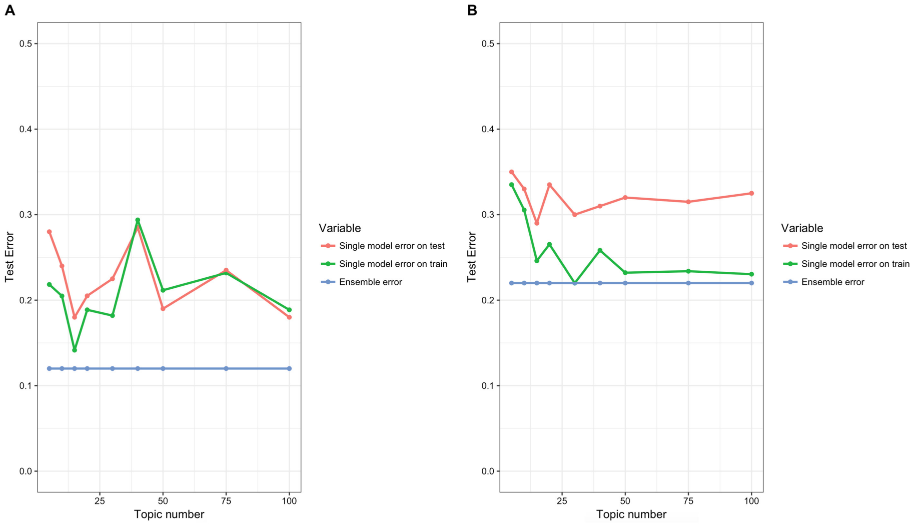

Fig. 1. Multi-class sLDA error rate for different number of topics compared with ensemble error rate, applied to SmsSpam (A) and Poliblog (B) dataset

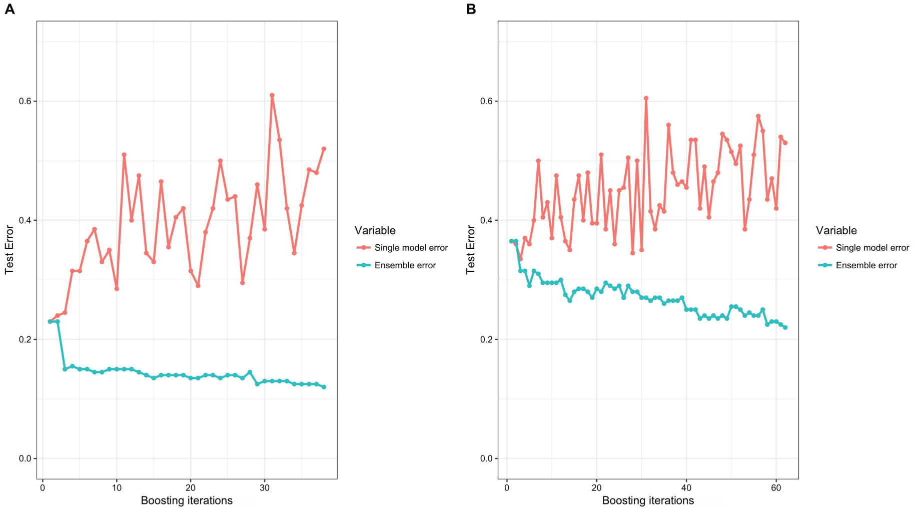

Fig. 2. Boost Multi-class sLDA for 5 LDA models with 5, 10, 15, 20 and 30 number of topics applied to SmsSpam (A) and Poliblog (B) datasets

finally reaching error level equal 12%. It is improvement in comparison to the best Multi-class sLDA model with $K = 15$, that achieved 14.1% on test dataset (see Table 1). Right picture (B), contains results of similar experiment for Poliblog dataset. This time, boosting achieved the best result equal to 22% error rate, after 63 iterations. In comparison, best Multi-class sLDA model for $K = 30$ reached only 29% error rate.

Table 1. Comparison of best Multi-class sLDA and Boost Multi-class sLDA

Model	SmsSpam	Poliblog
Multi-class sLDA on train dataset (best)	0.141 (15 topics)	**0.22** (30 topics)
Multi-class sLDA on test dataset (best)	0.18 (15 topics)	0.29 (15 topics)
Boost Multi-class sLDA on test dataset (5, 10, 15, 20, 30 topics)	**0.12** (38 iterations)	**0.22** (63 iterations)
Boost Multi-class sLDA on test dataset (15 topics - best)	**0.12** (35 iterations)	0.23 (92 iterations)

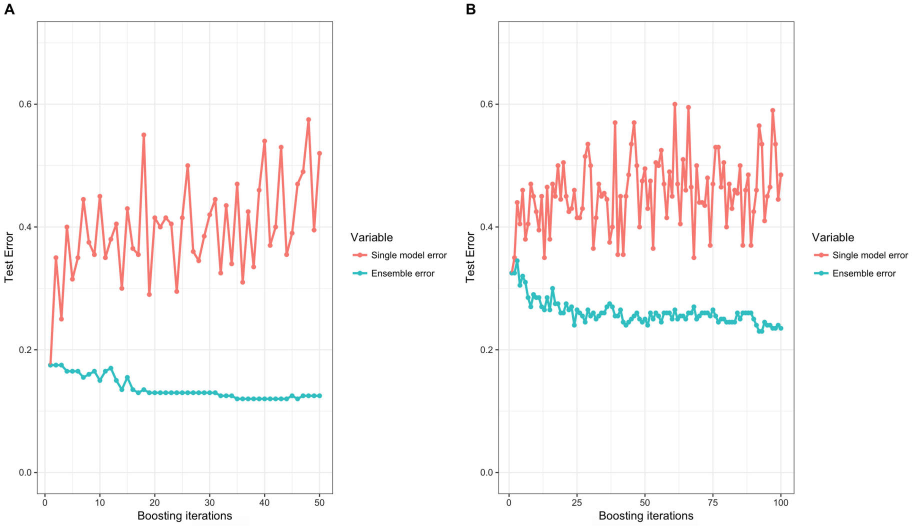

Fig. 3. Boost Multi-class sLDA applied to SmsSpam 15 topics (A) and Poliblog 15 topics (B)

5.3 Boost Multi-class sLDA for Best Number of Topics

Knowing the best number of topics for Multi-class sLDA, we have checked how Boost Multi-class sLDA can improve classification accuracy, if applied to a single model with optimal value of K. In Fig. 3, we present results for SmsSpam and Poliblog datasets. Left picture (A) contains 50 iterations of boosting applied to Multi-class sLDA model with $K = 15$, which is the optimal choice according to previous results. The lowest error equal to 12%, was achieved after 35 iterations. Recall, that this is exactly the same error as for Boost Multi-class sLDA with varying number of topics. This shows, that for SmsSpam dataset, the choice of K is not that important. We can choose many values, run Boost Multi-class sLDA and expect optimal result. Right picture (B), contains summary of similar

Table 2. Notation used in paper

$K \in \mathbb{N}$	Number of topics
$V \in \mathbb{N}$	Number of words in dictionary
$M \in \mathbb{N}$	Number of documents
$C \in \mathbb{N}$	Number of classes
$c_d \in \{1, \ldots, C\}$	Class label for document d
$N_d \in \mathbb{N}, \quad d = 1, \ldots, M$	Number of words in document d
$w_d \in \{0,1\}^{N_d \times V}$	Single document
$w_{d,n} \in \{0,1\}^V$	Indicator vector that denotes a single word
$w^i_{d,n} \in \{0,1\}$	i-th coordinate of indicator vector
$\mathcal{T} = \{w_d, c_d\}_{d=1}^M$	Corpus
$\alpha \in \mathbb{R}^K$	Dirichlet prior for per document topic proportions
$\theta_{d,k} \in \mathbb{R}$	Proportion of topic k in document d
$\theta_d \in \mathbb{R}^K$	Topic proportions in document d
$\theta \in \mathbb{R}^{M \times K}$	Topic proportions in corpus
$z_{d,n} \in \{0,1\}^K$	Topic assignment for single word (one-hot)
$\bar{z_d} = \frac{1}{N_d} \sum_{n=1}^{N_d} z_{d,n} \in \mathbb{R}^K$	Empirical topic frequencies for single document d
$z_d \in \{0,1\}^{N_d \times K}$	Topic assignments for single document d
$\beta \in \mathbb{R}^{K \times V}$	Topics - distributions over vocabulary
$\beta_k \in \mathbb{R}^V$	Single topic - distribution over vocabulary
$\eta = (\eta_1, \ldots, \eta_C) \in \mathbb{R}^{C \times K}$	Parameters of multinomial logistic regression model
$\eta_c \in \mathbb{R}^K$	c-th parameter of multinomial logistic regression model
$\phi_{d,n} \in \mathbb{R}^K$	Free variational parameters of multinomial distribution of topic assignments for $z_{d,n}$
$\phi_d \in \mathbb{R}^{N_d \times K}$	Free variational parameters of multinomial distribution of topic assignments for d-th document
$\phi \in \mathbb{R}^{M \times N_d \times K}$	Free variational parameters of multinomial distribution of topic assignments
$\gamma_d \in \mathbb{R}^K$	Free variational parameters of Dirichlet prior for d-th document
$\gamma \in \mathbb{R}^{M \times K}$	Free variational parameters of Dirichlet prior

experiment carried out for Poliblog dataset. Boost Multi-class sLDA reached 23% error rate after 92 iterations. It is slightly worse than ensemble for varying number of topics which equals 22%.

6 Summary

We have developed Boost Multi-class sLDA model, for text classification that is an ensemble of Multi-class sLDA models. We have demonstrated, that our model outperforms Multi-class sLDA, reducing error rate by 6% (SmsSpam) and 7% (Poliblog). We have also shown, that the model does not require a previous choice of hyperparameter K, that was required in Multi-class sLDA.

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodol.) **39**(1), 1–38 (1977)
2. Freund, Y., Schapire, R.: A decision-theoretic generalization of online learning and an application to boosting. J. Comput. Syst. Sci. **55**, 119–139 (1997)
3. Rubin, D.: Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. **12**(4), 1151–1172 (1984)
4. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001). https://doi.org/10.1007/978-0-387-84858-7
5. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Griffiths, T., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**, 5228–5235 (2004). https://doi.org/10.1073/pnas.0307752101
7. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, New York (2006)
8. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. J. Am. Stat. Assoc. **101**(476), 1566–1581 (2006)
9. Mcauliffe, J.D., Blei, D.M.: Supervised topic models. In: Advances in Neural Information Processing Systems (2008)
10. Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S.: A density-based method for adaptive LDA model selection. Neurocomputing **72**(7–9), 1775–1781 (2008). 16th European Symposium on Artificial Neural Networks
11. Wang, C., Blei, D., Fei-Fei, L.: Simultaneous image classification and annotation. In: Computer Vision and Pattern Recognition (2009)
12. Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N.: On finding the natural number of topics with latent Dirichlet allocation: some observations. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010. LNCS (LNAI), vol. 6118, pp. 391–402. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13657-3_43
13. Mimno, D., Blei, D.: Bayesian checking for topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2011)
14. Almeida, T.A., Gomez Hidalgo, J.M., Yamakami, A.: Contributions to the study of SMS spam filtering: new collection and results. In: Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG 2011), Mountain View, CA, USA (2011)
15. Deveaud, R., Sanjuan, E., Bellot, P.: Accurate and effective latent concept modeling for ad hoc information retrieval. Revue des Sciences et Technologies de l'Information - Série Document Numérique, Lavoisier, 61–84 (2014)
16. Chang, J.: LDA: Collapsed Gibbs Sampling Methods for Topic Models. R package version 1.4.2 (2015). https://CRAN.R-project.org/package=lda
17. Blei, D., Kucukelbir, A., McAuliffe, J.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017)