INTERPRETING THE PREDICTION PROCESS OF A DEEP NETWORK CONSTRUCTED FROM SUPERVISED TOPIC MODELS

Jianshu Chen, Ji He*†*, Xiaodong He*, Lin Xiao*, Jianfeng Gao*, and Li Deng**

*Microsoft Research, Redmond, WA 98052, USA
†Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA

ABSTRACT

In this paper, we propose an approach to interpret the prediction process of the BP-sLDA model, which is a supervised Latent Dirichlet Allocation model trained by Back Propagation over a deep architecture. The model is shown to achieve state-of-the-art prediction performance on several large-scale text analysis tasks. To interpret the prediction process of the model, often demanded by business data analytics applications, we perform evidence analysis on each pair-wise decision boundary over the topic distribution space, which is decomposed into a positive and a negative components. Then, for each element in the current document, a novel evidence score is defined by exploiting this topic decomposition and the generative nature of LDA. Then the score is used to rank the relative evidence of each element for the effectiveness of model prediction. We demonstrate the effectiveness of the method on a large-scale binary classification task on a corporate proprietary dataset with business-centric applications.

Index Terms— Topic model, BP-sLDA, mirror descent, back propagation, deep architecture

1. INTRODUCTION

Probabilistic topic models such as Latent Dirichlet Allocation (LDA) [1, 2] has been successfully applied to diverse tasks of text modeling and analysis. Supervised topic models [3–7], which use the additional label information to help with the modeling, have been shown to have the improved modeling ability and better prediction performance than the vanilla LDA. Recently, mirror-descent back propagation has been successfully applied to perform end-to-end discriminative learning of the supervised topic model (aka BP-sLDA). It has achieved much better performance than the prior-art supervised topic models, the traditional discriminative models such as logistic/linear regression, neural network, and is even on par with highly successful deep neural network (DNN) [8,9] on several large-scale text classification/regression tasks [10]. Moreover, different from DNNs [8, 11], the BP-sLDA model, as a probabilistic generative model, characterizes the internal dependency of the latent and observed variables under a probabilistic framework. This allows the model to be more interpretable than pure black-box models such as an DNN, while retaining good prediction performance [12]. This interpretability is a highly desirable property often benefiting applications in business data analytics.

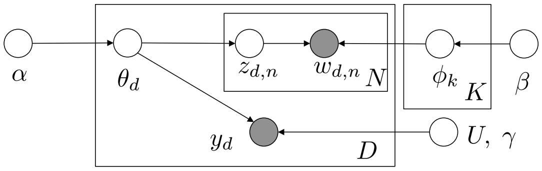

Fig. 1. Graphical representation of the supervised LDA model. Shaded nodes are observables.

In this paper, we develop an approach to interpret the "evidences" that BP-sLDA uses to perform its prediction, mainly in the classification scenario. We perform evidence analysis of each pair-wise decision boundary over the topic distribution space, which is decomposed into a positive and a negative components. Then, a novel metric is defined by the generative nature of the topic model to rank different evidences of prediction inside each document. The developed method is then applied to a large-scale text classification task using a proprietary corporate dataset.

2. AN OVERVIEW OF THE BP-SLDA

In this section, we briefly review the BP-sLDA model. Consider the graphical model in Fig. 1. Let K be the number of topics, N be the number of words in each document, V be the vocabulary size, and D be the number of documents in the corpus. The generative process of the model in Fig. 1 can be described as:

1. For each document d, choose the topic proportions according to a Dirichlet distribution: $\theta_d \sim p(\theta_d|\alpha) = \mathrm{Dir}(\alpha)$, where α is a $K \times 1$ vector consisting of nonnegative components.

2. Draw each column ϕ_k of a $V \times K$ matrix Φ independently from an exchangeable Dirichlet distribution:

Email: {jianshuc, lin.xiao, xiaohe, jfgao, deng}@microsoft.com, jvking@uw.edu.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

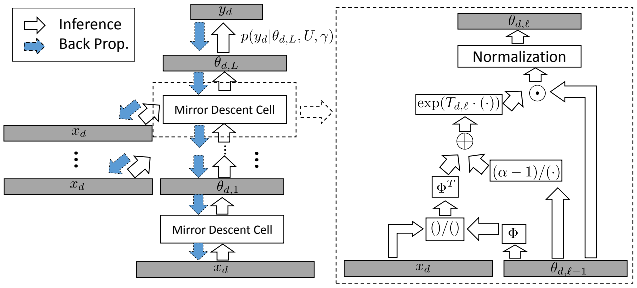

Fig. 2. BP-sLDA performs discriminative learning by back propagating over a deep mirror-descent architecture.

$\phi_k \sim \text{Dir}(\beta)$ (i.e., $\Phi \sim p(\Phi|\beta)$), where $\beta > 0$ is the smoothing parameter.

3. To generate each word $w_{d,n}$, first, choose a topic $z_{d,n} \sim p(z_{d,n}|\theta_d) = \text{Multi}(\theta_d)$, where $\text{Multi}(\cdot)$ denotes a multinomial distribution. Then, choose a word $w_{d,n} \sim p(w_{d,n}|z_{d,n}, \Phi) = \text{Multi}(\phi_{z_{d,n}})$.

4. Choose the $C \times 1$ response vector: $y_d \sim p(y_d|\theta, U, \gamma)$. In the classification scenario considered in this paper, $p(y_d|\theta_d, U, \gamma) = \text{Mult}\big(\text{Softmax}(\gamma U \theta_d)\big)$, where $\text{Softmax}(x)_c = \frac{e^{x_c}}{\sum_{c'=1}^{C} e^{x_{c'}}}$, $c = 1, \ldots, C$.

The model in Fig. 1 is slightly different from the one in [3], where the response variable y_d in Fig. 1 is coupled with θ_d instead of $z_{d,1:N}$ as in [3]. This modification leads to a differentiable end-to-end cost trainable by back propagation with superior prediction performance. Specifically, BP-sLDA is trained by maximizing the posterior probability $\prod_{d=1}^{D} p(y_d|w_{d,1:N}, \Phi, U)$ [10], and it consists of a deep mirror-descent [13–16] architecture for computing the maximum-a-posterior estimate of θ_d to sample the posterior probability $p(y_d|w_{d,1:N}, \Phi, U)$, and a back propagation process over the same architecture for computing the stochastic gradient (see Fig. 2). The model is then updated via stochastic mirror descent (SMD) (see [10] for the details). After the training, the feed forward architecture is used to predict the output variable y_d given input $w_{d,1:N}$. With such end-to-end discriminative training, it was shown in [10] that the prediction performance of BP-sLDA is on par or even better than that of deep neural networks (DNN), and significantly outperforms traditional supervised topic models and other discriminative models (e.g., logistic regresssion, neural network, etc.). Moreover, unlike DNN, which is less interpretable, the BP-sLDA model retains special structures designed from the probabilistic generative model, which can be used to interpret the prediction process, as we proceed to explain in sequel.

3. INTERPRETING THE PREDICTION PROCESS

In this section, we develop an approach to interpret the prediction process of the BP-sLDA model.

3.1. Evidence analysis over topic distributions

In BP-sLDA, the feed forward architecture in Fig. 2 computes the MAP estimate of $p(\theta_d|w_{d,1:N}, \Phi, \alpha)$:

$$\theta_{d,L} \approx \hat{\theta}_{d|w_{d,1:N}} = \arg\max_{\theta_d \in \mathcal{P}_K} p(\theta_d|w_{d,1:N}, \Phi, \alpha) \qquad (1)$$

where $\hat{\theta}_{d|w_{d,1:N}}$ denotes the MAP estimate of θ_d given the d-th document consisting of words $w_{d,1:N}$, and the approximation is due to using a finite number of mirror descent layers. Then, the posterior probability of y_d given input $w_{d,1:N}$ is computed according to

$$\begin{aligned} &p(y_d|w_{d,1:N}, \Phi, U, \alpha, \gamma) \\ &\quad= \int p(y_d|\theta_d, U, \gamma) p(\theta_d|w_{d,1:N}, \Phi, \alpha) d\theta_d \\ &\quad= \mathbb{E}_{\theta_d|w_{d,1:N}} \left[p(y_d|\theta_d, U, \gamma)\right] \approx p(y_d|\theta_{d,L}, U, \gamma) \qquad (2) \end{aligned}$$

where in the last step the expectation is sampled by the MAP estimate, and $p(y_d|\theta_d, U, \gamma)$ assumes the following form for classification case:

$$p(y_d = c|\theta_{d,L}, U, \gamma) = \frac{\exp(\gamma u_c \theta_{d,L})}{\sum_{c'=1}^{C} \exp(\gamma u_{c'} \theta_{d,L})} \qquad (3)$$

where $c = 1, \ldots, C$ and C is the total number of classes[1], and u_c denotes the c-th row of the matrix U. Expressions (1)–(2) imply that the deep mirror-descent architecture in Fig.2 is indeed extracting the topic distribution as the high-level features from the d-th document, and then feeds them into a multi-class logistic regression for classification.

To interpret the entire prediction process, we first analyze the decision making from $\theta_{d,L}$ to y_d in (3). After the posterior probability $p(y_d|w_{d,1:N}, U, \gamma)$ is computed according to (2), the d-th input document will be classified according to

$$c = \arg\max_{c'=1,\ldots,C} p(y_d = c'|w_{d,1:N}, U, \gamma) \qquad (4)$$

i.e., we use (4) to generate the predicted class. We now proceed to analyze how the model chooses to believe that c is the class associated with the d-th document. First, note that expression (4) is also equivalent to

$$\ln \frac{p(y_d = c|w_{d,1:N}, U, \gamma)}{p(y_d = c'|w_{d,1:N}, U, \gamma)} > 0, \quad \forall c' \neq c \qquad (5)$$

which is further equivalent to the following pair-wise decision rule after substituting (3) into (5):

$$u_{cc'} \theta_{d,L} > 0, \quad \forall c' \neq c \qquad (6)$$

[1] We use 1-hot coding to represent each class in our implementation.

where $u_{cc'} \triangleq u_c - u_{c'}$. Let $u_{cc',j}$ and $\theta_{d,L,j}$ denote the j-th element of the vectors $u_{cc'}$ and $\theta_{d,L}$, respectively. Then, expression (6) can be rewritten as

$$\sum_{j=1}^{K} u_{cc',j}\theta_{d,L,j} > 0, \quad \forall c' \neq c \tag{7}$$

where the left-hand side is a sum of K terms. Eq. (7) defines a pairwise decision boundary between the two classes, c and c'. As long as the inequality (7) holds for all $c' \neq c$, the class c is the MAP decision according to (4). Therefore, to interpret how the BP-sLDA model predicts c for the d-th document, it suffices to examine each of the $C - 1$ pair-wise decision boundaries.

Note that $\theta_{d,L,j}$ is always nonnegative as it represents the probability of the j-th topic in the d-th document. Therefore, we can partition all the $u_{cc',j}$, $j = 1, \ldots, K$ into two subsets consisting of the positive $u_{cc',j}$ and the negative $u_{cc',j}$, respectively. Then the positive $u_{cc',j}$ would make the classifier (7) prefer class c over c', while the negative $u_{cc',j}$ makes the classifier prefer c', and the final decision is based on whether the weighted sum (of all the evidences) is positive. For this reason, $u_{cc',j}$ is defined as the *weight of evidence* (WOE), and the entire sum is called the evidence of the decision. Furthermore, for each particular pair-wise decision boundary (7), the topics can be decomposed into two categories: the ones associated with positive $u_{cc',j}$ and the ones associated with negative $u_{cc',j}$, which are the positive and negative "evidences", respectively. More formally, let $\mathcal{J}_{cc',+}$ and $\mathcal{J}_{cc',-}$ denote the two sets consisting of the topics associated with the positive and negative WOEs, respectively. Introduce

$$\theta^{+}_{d,cc',j} \triangleq \begin{cases} \frac{\theta_{d,L,j}}{\sum_{k \in \mathcal{J}_{cc',+}} \theta_{d,L,k}} & \text{if } u_{cc',j} \geq 0 \\ 0 & \text{if } u_{cc',j} < 0 \end{cases} \tag{8}$$

$$\theta^{-}_{d,cc',j} \triangleq \begin{cases} 0 & \text{if } u_{cc',j} \geq 0 \\ \frac{\theta_{d,L,j}}{\sum_{k \in \mathcal{J}_{cc',-}} \theta_{d,L,k}} & \text{if } u_{cc',j} < 0 \end{cases} \tag{9}$$

and let $\theta^{+}_{d,cc}$ and $\theta^{-}_{d,cc'}$ be the vectors that collect $\theta^{+}_{d,cc',j}$ ($j = 1, \ldots, K$) and $\theta^{-}_{d,cc',j}$ ($j = 1, \ldots, K$), respectively. Note that both $\theta^{+}_{d,cc}$ and $\theta^{-}_{d,cc'}$ are normalized to add up to one.

3.2. Interpreting the positive and negative topics

So far we have discussed how the multi-class logistic regression accumulates the positive and negative evidences for each pairwise decision, and how it partitions the topic space into positive and negative parts. We now proceed to develop an approach to propagate the analysis into the input space.

It was shown in [10] that the probability of the d-th input document given the topic distribution θ_d can be expressed as

$$p(w_{d,1:N}|\theta_d, \Phi) = p(x_d|\theta_d, \Phi) = \prod_{v=1}^{V}\Bigg(\sum_{j=1}^{K}\theta_{d,j}\Phi_{vj}\Bigg)^{x_{d,v}} \tag{10}$$

where $x_{d,v}$ denotes the term frequency of the v-th word (in vocabulary) inside the d-th document, and x_d denotes the V-dimensional bag-of-words vector of the d-th document. Note from (8)–(9) that $\theta^{+}_{d,cc}$ and $\theta^{-}_{d,cc'}$ are the positive and the negative topic distributions of the d-th document, respectively, with respect to the current decision boundary between c and c'. One useful property of such a probabilistic generative model is that we can sample the input documents for $\theta^{+}_{d,cc'}$ and $\theta^{-}_{d,cc'}$ according to $p(w_{d,1:N}|\theta^{+}_{d,cc'}, \Phi)$ and $p(w_{d,1:N}|\theta^{-}_{d,cc'}, \Phi)$, respectively. In this paper, however, we do not perform such sampling for $\theta^{+}_{d,cc'}$ and $\theta^{-}_{d,cc'}$, since our objective is to identify in the existing d-th document the clues for making the positive (negative) decision. Consider the following log-likelihood ratio (LLR) of the d-th document between the positive and the negative topic distributions:

$$\ln\left(\frac{p(w_{d,1:N}|\theta^{+}_{d,cc'}, \Phi)}{p(w_{d,1:N}|\theta^{-}_{d,cc'}, \Phi)}\right) = \sum_{v=1}^{V} x_{d,v}\eta_{d,cc',v} \tag{11}$$

where $\eta_{d,cc',v}$ is the score of evidence defined as

$$\eta_{d,cc',v} \triangleq x_{d,v} \ln\left(\frac{\sum_{j=1}^{K}\theta^{+}_{d,cc',j}\Phi_{vj}}{\sum_{j=1}^{K}\theta^{-}_{d,cc',j}\Phi_{vj}}\right) \tag{12}$$

Observe that the log-likelihood ratio (11) is a sum of V terms, where each term characterizes how the v-th term contributes to the total LLR. If the BP-sLDA model chooses c as the predicted class for the d-th document (according to (4)), then for each element in the d-th document, we can use $\eta_{d,cc',v}$ to rank its evidence for preferring class c over any other class c'.

4. EXPERIMENTS

4.1. Prediction Performance

To examine the effectiveness of the analysis method developed in this paper, we consider a binary classification task on a corporate proprietary dataset with application to business-centric applications. The dataset consisting of a training set of 1.2 million documents, a development set of 149K documents and a test set of 149K documents, with the vocabulary size being 128K. A BP-sLDA model with $K = 200$ topics and $L = 10$ mirror-descent layers is trained on the dataset, and the hyper parameters are tuned on the development set. Moreover, we also trained a logistic regression and a neural network (200 tanh units) as the baselines. The AUC (area-under-the-curve) and the accuracy on the test set are shown

Table 1. Prediction performance. LR stands for logistic regression, and NN stands for neural network.

Model	BP-sLDA	LR	NN
AUC	**93.49%**	90.56%	91.99%
Accuracy	**86.00%**	82.95%	84.67%

Table 2. Examples of BP-sLDA prediction and the top evidences.

Truth	Prediction	Top three terms with highest evidences and the score $\eta_{d,cc',v}$		
0	0.001	Country/Turkey: 0.27	CustomerCode/xxx1: 0.24	EngagementName/sam: 0.10
0	0.014	RecommendationCode/xx3: 0.58	Country/Brazil: 0.20	City/Rio: 0.183
0	0.001	RecommendationCode/xx3: 0.69	Country/Thailand: 0.42	Currency/baht: 0.27
0	0.348	RecommendationCode/xx3: 0.46	Country/Ukraine: 0.28	ProductFamily/server: 0.21
0	0.025	RecommendationCode/xx3: 0.74	Country/Paraguay: 0.20	Program/assurance: 0.12
1	0.999	Country/China: 0.79	Currency/USD: 0.39	RecommendationCode/xx0: 0.42
1	0.907	Country/United States: 0.50	RecommendCode/xx3: 0.41	EngagementName/onsite: 0.33
1	0.999	Currency/AUD: 0.94	Engagement/1: 0.51	RecommendationCode/xx0: 0.11
1	0.992	RecommendationCode/xx0:1.08	Country/Japan:0.59	CurrencyName/JPY: 0.44
1	0.933	ProductName/Tablet A: 0.60	Currency/USD: 0.32	ServicesEngagement/1: 0.26

in Table 2, which clearly shows that BP-sLDA outperforms other methods.

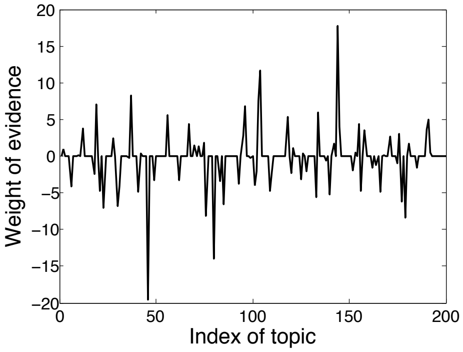

Fig. 3. Example of $u_{cc',j}$ (weight of evidence). The result is obtained from a binary classification task described in Sec. 4.

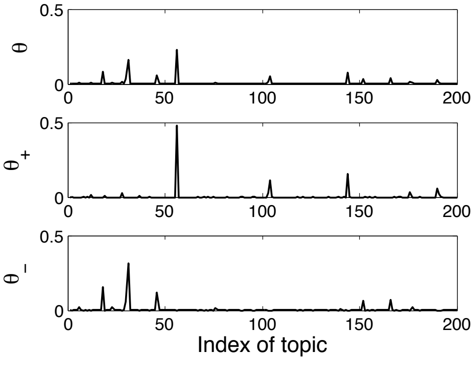

Fig. 4. Example of decomposing the topic distribution into a positive part and a negative part.

4.2. Weight of Evidence and Topic Decomposition

After the BP-sLDA model is trained, we analyze the weight of evidence, $u_{cc',j}$, over the topic distribution space according to (7), which is shown in Fig. 3. In this task, we only have two classes, i.e., positive and negative classes. The positive WOE in Fig. 3 implies that the corresponding topic is relevant to positive decision and vice versa. Moreover, in Fig. 4, we show an example of the topic distribution of a document, and its decomposition into positive and negative parts.

4.3. Interpreting Prediction Results

We now focus on applying (11)–(12) to interpret the prediction of BP-sLDA on this task. In Table 2, we show ten examples of the prediction and its interpretation analysis, including five positive examples and five negative examples. Each row represents one sample (document) in the test set. We list the ground truth class (in the first column) associated with each sample, the predicted probability of the sample being positive (in the second column), and the top three terms and their corresponding values of $\eta_{d,cc',v}$ (in the third to the fifth columns). For privacy reasons, we anonymize some of the information. Higher $\eta_{d,cc',v}$ scores implies the term is more relevant to making this particular prediction from BP-sLDA. The results show strong correlation between the label with some terms such as the countries, a certain Recommendation Code in the business process. Such result is useful in helping us understand how the model judges the input data and makes a particular prediction.

5. CONCLUSION

We have proposed an approach to interpret the prediction process of the BP-sLDA model by performing evidence analysis of the topic distribution space, which is decomposed into a positive and a negative components. A novel evidence score has been introduced for each element in the current document to rank its relative evidence for making a particular prediction. The effectiveness of the analysis method is demonstrated on a large-scale binary classification task on a corporate proprietary dataset with business-centric applications. The method is also directly applicable to multi-class cases. Quantitative evaluation of the interpretation results is left as our future work.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *JMLR*, vol. 3, pp. 993–1022, 2003.

[2] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proc. of the National Academy of Sciences*, pp. 5228–5235, 2004.

[3] D. M. Blei and J. D. Mcauliffe, “Supervised topic models,” in *Proc. NIPS*, 2007, pp. 121–128.

[4] S. Lacoste-Julien, F. Sha, and M. I. Jordan, “DiscLDA: Discriminative learning for dimensionality reduction and classification,” in *Proc. NIPS*, 2008, pp. 897–904.

[5] J. Zhu, A. Ahmed, and E. P. Xing, “MedLDA: maximum margin supervised topic models,” *JMLR*, vol. 13, no. 1, pp. 2237–2278, 2012.

[6] J. Zhu, N. Chen, H. Perkins, and B. Zhang, “Gibbs max-margin topic models with data augmentation,” *JMLR*, vol. 15, no. 1, pp. 1073–1110, 2014.

[7] Y. Wang and J. Zhu, “Spectral methods for supervised topic models,” in *Proc. NIPS*, 2014, pp. 1511–1519.

[8] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[9] George E Dahl, Dong Yu, Li Deng, and Alex Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[10] J. Chen, J. He, Y. Shen, L. Xiao, X. He, J. Gao, X. Song, and L. Deng, “End-to-end learning of LDA by mirror-descent back propagation over a deep architecture,” in *Proc. NIPS*, 2015.

[11] Li Deng and Dong Yu, *Deep Learning: Methods and Applications*, NOW Publishers, 2014.

[12] L. Deng and N. Jaitly, “Deep discriminative and generative models for pattern recognition,” in *Handbooks of Pattern Recognition and Computer Vision (Ed. C. H. Chen)*, pp. 27–52. Springer, 2015.

[13] D. B. Nemirovsky. A. S., Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley, New York, 1983.

[14] A. Beck and M. Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[15] P. Tseng, “On accelerated proximal gradient methods for convex-concave optimization,” *SIAM Journal on Optimization*, 2008.

[16] D. Sontag and D. Roy, “Complexity of inference in latent dirichlet allocation,” in *Proc. NIPS*, 2011, pp. 1008–1016.