Conditionally Trained Latent Dirichlet Allocation for Text Modeling and Categorization

Simon Lacoste-Julien[1]*, Fei Sha[2] and Michael I. Jordan[3]

[1] Computer Science Division, UC Berkeley, CA

[2] Yahoo! Research, 2821 Mission College Blvd., Santa Clara, CA

[3] Dept. of EECS and Statistics, UC Berkeley, CA

The hierarchical modeling paradigm of parametric and nonparametric Bayesian statistics have found a widespread application in modeling data in many areas, for instance, text and language processing, computer vision and bioinformatics [6, 2]. Most commonly, Bayesian models are fit *generatively*, ie, by maximum likelihood estimation on unsupervised data. In many cases, however, we have access to extra side information that is *discriminative* in nature. For example, scientific papers group very naturally according to their scientific domains that distinguish one type of documents from the others. How can we incorporate such discriminative information in our (clustering) models for these documents?

In this work, we investigate *conditional and discriminative training* techniques to yield models that can explain both features that are shared across domains and features that can differentiate them. We have proposed a general framework to fit the latent Dirichlet allocation (LDA) model for topic modeling of text corpus with discriminative information such as document categories. Specifically, we show how to estimate the model parameters by maximizing the *conditional likelihood* of the categories. Note that, we are interested *not only* in deriving classification rules to recognize categories *but also* modeling these documents jointly to reveal shared structures. Our goal is to come up with a compact representation of the documents — such as the topics — that are both discriminative and transferrable between domains [3].

Model In our setting, each document $\boldsymbol{w}_d$ in the corpus is associated with a categorical variable or class label $y_d \in \{1, 2, \ldots, C\}$, for instance, whether the document is published in the proceedings of NIPS conferences or in the journal *Psychological Review*. To model this labeling information, we introduce a simple extension to the standard LDA model [2]. The graphical model in Figure 1 shows the generative process. For each class label c, we associate it with a linear transformation matrix $\boldsymbol{T}_c : \Re^K \to \Re^L$, which transforms the $K-$dimensional Dirichlet variable $\boldsymbol{\theta}_d$ to a mixture of Dirichlet distributions $\boldsymbol{T}_c\boldsymbol{\theta}_d \in \Re^L$. To generate a word w_{dn}, we draw its topic z_{dn} from $\boldsymbol{T}_{y_d}\boldsymbol{\theta}_d$ instead of $\boldsymbol{\theta}_d$ as in the standard LDA model. Intuitively, while every document in the text corpus is represented through $\boldsymbol{\theta}_d$ as a point in the topic simplex $\{\boldsymbol{\theta} \mid \sum_k \theta_k = 1\}$, we hope the linear transformation $\{\boldsymbol{T}\}$ would be able to *reposition* these points such that similar documents — those with same class labels — are represented by points nearby to each other. Note that these points can *not* be arbitrarily moved around because all documents — similar or not — share the same parameters $\boldsymbol{\Phi}$ to generate words.

To estimate the parameters $\{\boldsymbol{T}_c\}$, we maximize the conditional likelihood $\sum_d \log p(y_d \mid \boldsymbol{w}_d; \{\boldsymbol{T}_c\}, \boldsymbol{\Phi}, \alpha, \beta)$ while holding $\boldsymbol{\Phi}$ fixed. This optimization requires computing the gradient of the likelihood with respect to $\{\boldsymbol{T}_c\}$, which is intractable (though can be estimated using Gibbs sampling). As a preliminary study, we simplified the optimization by restricting $\{\boldsymbol{T}_c\}$ to specific forms. A particular restriction for binary labels is given by the following convex combination of stochastic matrices whose columns sum to one:

$$\boldsymbol{T}_1 = \lambda_1 \begin{pmatrix} \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_K \end{pmatrix} + (1-\lambda_1) \begin{pmatrix} \boldsymbol{I}_K & \boldsymbol{I}_K \\ \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{T}_2 = \lambda_2 \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_K \end{pmatrix} + (1-\lambda_2) \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I}_K & \boldsymbol{I}_K \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \quad (1)$$

where $\boldsymbol{I}_K$ stands for the identity matrix with K rows and columns. These linear transformations have a nice interpretation of sharing the last K topics amongst classes while keeping the other two blocks distinct for each class. Because we only have two parameters λ_1 and λ_2, we can do a brute-force grid search to optimize over them. To estimate the

*Keywords:** Graphical model, learning algorithm **Preference:** Oral

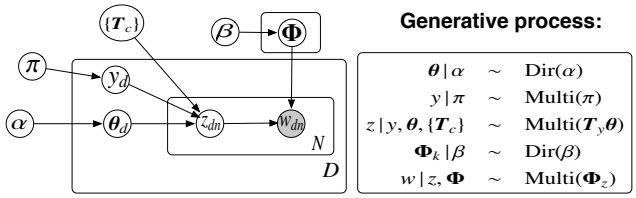

Figure 1: Extended LDA model for modeling text with discriminative labeling information.

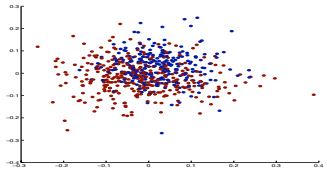

Figure 2: Embedding using topic assignment $\boldsymbol{\theta}$

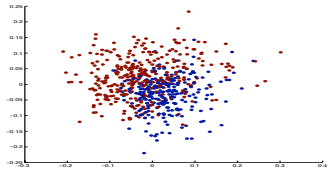

Figure 3: Embedding using topic assignment $\boldsymbol{T\theta}$

parameters $\boldsymbol{\Phi}$, we hold the transformation matrices fixed and maximize the likelihood of the model, as in standard LDA models.

Experimental results We tested our extended LDA model with two preliminary experiments which gave promising results. In the first task, we combine two sets of text corpus – the proceedings of NIPS and the abstracts of the journal *Psychological Review* [5] – to form a two-class corpus for classification. It appears that this classification is easy with near perfect classification for a Naive-Bayes classifier, though we found interesting to validate the output of our model by visualizing the compact representation of the document that is induced by the model. Specifically, we embed documents in a $2D$ space using multidimensional scaling with metrics computed from the symmetrized KL divergence between documents' topic assignments (see [5]). Figure 2 shows the embedding using the (untransformed) topic assignments $\boldsymbol{\theta}_d$ for the documents, with colors indicating the class, and which shows no clear boundary between the two corpora. On the other hand, figure 3 shows a clearer separation between the two corpora where we have used the transformed topic assignments $\boldsymbol{T}_y \boldsymbol{\theta}_d$ marginalized over the class label y. Since the transformation $\boldsymbol{T}$ is estimated to maximize the conditional likelihood of (target) class labels, we expect the topic assignment $\boldsymbol{\theta}_d$ is moved around in the topic simplex such that documents with same class labels are grouped much tighter.

In the second task, we tried to classify the difficult pair `alt.atheism` vs. `talk.religion.misc` in the Newsgroup dataset. A simple NB baseline obtained 84% accuracy. Our preliminary results showed that maximum CL on our two-parameters model yielded an improvement over NB and LDA.

Related work Recently there have been growing interest in topic modeling with supervised information [1]. The author-topic model by Rosen-ziv et al is somewhat closer to our model in both modeling philosophy and model structure [4], with one important distinction: our parameters are trained by maximizing conditional likelihood as opposed to their maximum likelihood estimation. Our future work includes a fully discriminative framework for estimating unrestricted transformation matrices.

References

[1] David Blei and Jon McAuliffe. Supervised topic models. In Yoram Singer John Platt, Daphne Koller and Sam Roweis, editors, *Advances in Neural Information Processing Systems 21*. MIT Press, 2008.

[2] David M. Blei, Andew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jenn Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems*, 2008.

[4] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of 20th Conference on Uncertainty in Artificial Intelligence*, 2004.

[5] Mark Steyvers and Thomas Griffiths. Matlab topic modeling toolbox. `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm`.

[6] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.