Semi-Supervised Prediction-Constrained Topic Models

Michael C. Hughes[1], Gabriel Hope[2], Leah Weiner[3], Thomas H. McCoy, Jr.[4], Roy H. Perlis[4], Erik Sudderth[2,3], and Finale Doshi-Velez[1]

[1]School of Engineering and Applied Sciences, Harvard University
[2]School of Information & Computer Sciences, Univ. of California, Irvine
[3]Department of Computer Science, Brown University
[4]Massachusetts General Hospital & Harvard Medical School

1 Introduction

Discrete count data are common: news articles can be represented as word counts, patient records as diagnosis counts, and images as visual descriptor counts. Topic models such as *latent Dirichlet allocation* (LDA, Blei et al. (2003)) are popular for finding cooccurance structure in datasets of count vectors, producing a small set of learned *topics* which help users understand the core themes of a corpus too large to comprehend manually.

The low-dimensional feature space produced by LDA can be useful as the input to some predictive task, where the user seeks to predict *labels* associated with each count vector. *Supervised topic models* extend standard topic models to model document topics and labels jointly, leading to better label predictions and more informative topic representations.

In this work, we correct a key deficiency in previous formulations of supervised topic models. Our learning objective directly encourages low-dimensional data representations to produce accurate predictions by deliberately encoding the *asymmetry* of prediction tasks: we want to predict labels from text, not text from labels. Approaches like *supervised LDA* (sLDA, McAuliffe and Blei (2008)) that optimize the *joint* likelihood of labels and words ignore this crucial asymmetry.

2 Background

Standard LDA. The LDA topic model finds structure in a collection of D documents, or more generally, D examples of count vectors. Each document d is represented by a count vector x_d of V discrete words or features: $x_d \in \mathbb{Z}_+^V$. The LDA model generates these counts via a document-specific mixture of K topics:

$$\begin{aligned} \pi_d|\alpha &\sim \mathrm{Dir}(\pi_d \mid \alpha), \\ x_d|\pi_d, \phi &\sim \mathrm{Mult}(x_d \mid \textstyle\sum_{k=1}^K \pi_{dk}\phi_k, N_d). \end{aligned} \tag{1}$$

The random variable π_d is a document-topic probability vector, where π_{dk} is the probability of topic k in document d and $\sum_{k=1}^K \pi_{dk} = 1$. The vector ϕ_k is a topic-word probability vector, where ϕ_{kv} gives the probability of word v in topic k and $\sum_{v=1}^V \phi_{kv} = 1$. N_d is the (observed) size of document d: $N_d = \sum_v x_{dv}$. LDA assumes π_d and ϕ_k have symmetric Dirichlet priors, with hyperparameters $\alpha > 0$ and $\tau > 0$.

Topic-based Prediction of Binary Labels. Suppose document d also has a binary label $y_d \in \{0, 1\}$. Standard supervised topic models assume labels and word counts are conditionally independent given document-topic probabilities π_d:

$$y_d|\pi_d, \eta \sim \mathrm{Bern}(y_d \mid \sigma(\textstyle\sum_{k=1}^K \pi_{dk}\eta_k)), \tag{2}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ is the logit function, and $\eta \in \mathbb{R}^K$ is a vector of real-valued regression weights with a vague prior $\eta_k \sim \mathcal{N}(0, \sigma_\eta^2)$.

3 Prediction-Constrained sLDA

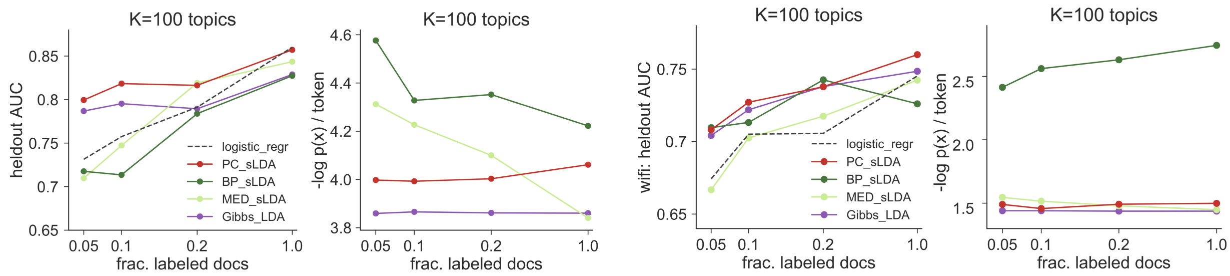
We propose a novel, *prediction-constrained* (PC) objective that finds the best generative model for words x, while satisfying the *constraint* that topics ϕ must yield accurate predictions about labels y given x alone:

$$\min_{\phi,\eta} \; -\Big[\sum_{d=1}^D \log p(x_d \mid \phi, \alpha)\Big] - \log p(\phi, \eta) \tag{3}$$

$$\text{subject to} \quad -\textstyle\sum_{d=1}^D \log p(y_d \mid x_d, \phi, \eta, \alpha) \leq \epsilon.$$

The scalar ϵ is the highest aggregate loss we are willing to tolerate, and $p(\phi, \eta) = p(\phi)p(\eta)$ are independent priors used for regularization. The structure of Eq. (3) matches the goals of a domain expert who wishes to explain as much of the data x as possible, while still making sufficiently accurate predictions of y.

Applying the Karush-Kuhn-Tucker conditions, we transform the inequality constrained objective



(a) Movie reviews

(b) Yelp reviews

Figure 1: Movie and Yelp tasks: Performance metrics vs. fraction of labeled training documents used for 100 topics. On both tasks, we artificially include only a small fraction (0.05, 0.10, or 0.20) of available training labels, chosen at random. Fully supervised methods (e.g. BP-sLDA, MED-sLDA) are *only given* documents (x_d, y_d) from this subset, because third-party code does not allow using unlabeled data at training. *Left:* Heldout discriminative performance (AUC, higher is better). Note that improvements over supervised learning algorithms, including logistic regression, are particularly large when the fraction of labeled documents is small. *Right:* Heldout generative performance (negative likelihood, lower is better).

in Eq. (3) to an equivalent unconstrained optimization problem:

$$\min_{\phi,\eta} - \sum_{d=1}^{D} \Big[\log p(x_d|\phi) + \lambda_\epsilon \log p(y_d|x_d, \phi, \eta) \Big] - \log p(\phi, \eta). \qquad (4)$$

For any prediction tolerance ϵ, there exists a scalar multiplier $\lambda_\epsilon > 0$ such that the optimum of Eq. (3) is a minimizer of Eq. (4). The relationship between λ_ϵ and ϵ is monotonic, but does not have an analytic form, so we search over the 1-D space of penalties λ_ϵ for an appropriate value.

Computing this objective directly is not feasible as computing both $p(x_d|\phi)$ and $p(y_d|x_d, \phi, \eta)$ requires taking an intractable integral over the latent variable π_d. For learning, we approximate this objective using a point estimate of π_d, where π_d is fixed to its MAP estimate given x_d: $\arg\max_{\pi_d} p(\pi_d|x_d, \phi, \alpha)$. Using this tractable objective, we can fit the model parameters (ϕ, η) with stochastic gradient descent.

If documents are partially labeled, the objective of Eq. (3) can be naturally generalized to only include prediction constraints for observed labels. This allows our approach to be easily applied to semi-supervised settings.

4 Experimental Results

We evaluate our approach on two real-world bag-of-words prediction tasks, comparing with a number of discriminative baselines including: logistic regression, the fully supervised BP-sLDA algorithm of Chen et al. (2015), the unsupervised Gibbs sampler for LDA (Griffiths and Steyvers, 2004) from the Mallet toolbox (McCallum, 2002), and the supervised MED-sLDA Gibbs sampler (Zhu et al., 2013) which is reported to improve on an earlier variational method (Zhu et al., 2012). In our experiments we partition documents into three sets (training/validation/test). For all methods shown, we tune any hyperparameters on the validation set and report results on the test set.

Movie task. Each of the 4004/500/501 documents is a movie review by a professional critic (Pang and Lee, 2005), with $V = 5338$ terms. Each review has a binary label, where $y_d = 1$ means the critic gave the film more than 2 stars.

Yelp task. Each of the 23159/2895/2895 documents (Yelp Dataset Challenge, 2016) aggregates all text reviews for a single restaurant, using $V = 10,000$ vocabulary terms. Each document also has a label indicating the availability of wifi.

Discussion. Fig. 1 shows that PC-sLDA is consistently competitive with purely discriminative methods like logistic regression (LR) or BP-sLDA when datasets are fully labeled. When the number of available labels is limited, PC-sLDA still performs well. For the Movie task in Fig. 1(a), PC-sLDA dominates the AUC metric for small fractions of labels (0.05, 0.1), beating even LR. In this regime, unsupervised Gibbs-LDA has better AUC than BP-sLDA and MED-sLDA, demonstrating the value of unlabeled data for prediction. On Yelp, PC-sLDA predictions at small fractions are better than all but BP-sLDA.

Fig. 1 also shows trends in heldout data negative log likelihood (lower is better). As expected, unsupervised Gibbs-LDA consistently achieves the best scores, as explaining data is its sole objective. However, PC-sLDA is able to achieve competitive likelihood scores, while achieving better AUC scores when compared to all other methods.

Acknowledgements

An extended version of this abstract will appear at the 2018 International Conference on Artificial Intelligence and Statistics

This research supported in part by NSF CAREER Award No. IIS-1349774, and by the Office of the Director of National Intelligence (ODNI) via IARPA contract 2016-16041100004. MCH is supported by Oracle Labs. We thank Victor Castro for preparing deidentified EHR data. We thank Harvard FAS Research Computing, Partners Research Computing, and Brown Computer Science for computational resources.

References

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

Jianshu Chen, Ji He, Yelong Shen, Lin Xiao, Xiaodong He, Jianfeng Gao, Xinying Song, and Li Deng. 2015. End-to-end learning of LDA by mirror-descent back propagation over a deep architecture. In *Neural Information Processing Systems*.

Yelp Dataset Challenge. 2016. Yelp dataset challenge. https://www.yelp.com/dataset_challenge. Accessed: 2016-03.

Tom L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*.

Jon D McAuliffe and David M Blei. 2008. Supervised topic models. In *Neural Information Processing Systems*, pages 121–128.

Andrew Kachites McCallum. 2002. MALLET: Machine learning for language toolkit. mallet.cs.umass.edu.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Jun Zhu, Amr Ahmed, and Eric P Xing. 2012. MedLDA: maximum margin supervised topic models. *The Journal of Machine Learning Research*, 13(1):2237–2278.

Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. 2013. Gibbs max-margin topic models with fast sampling algorithms. In *International Conference on Machine Learning*.