Hierarchically Supervised Latent Dirichlet Allocation

Adler Perotte **Nicholas Bartlett** **Noémie Elhadad** **Frank Wood**
Columbia University, New York, NY 10027, USA
{ajp9009@dbmi,bartlett@stat,noemie@dbmi,fwood@stat}.columbia.edu

We introduce hierarchically supervised latent Dirichlet allocation (HSLDA), a model for hierarchically and multiply labeled bag-of-word data. Examples of such data include web pages and their placement in directories, product descriptions and associated categories from product hierarchies, and free-text clinical records and their assigned diagnosis codes. Out-of-sample label prediction is the primary goal of this work, but improved lower-dimensional representations of the bag-of-word data are also of interest.

Our work operates within the framework of topic modeling. Our approach learns topic models of the underlying data and labeling strategies in a joint model, while leveraging the hierarchical structure of the labels. For the sake of simplicity, we focus on is-a hierarchies, but the model can be applied to other structured label spaces. We extend supervised latent Dirichlet allocation (sLDA) [2] to take advantage of hierarchical supervision. We propose an efficient way to incorporate hierarchical information into the model. We hypothesize that the context of labels within the hierarchy provides valuable information about labeling. Other models that incorporate LDA and supervision include LabeledLDA[6] and DiscLDA[5]. None of these models, however, leverage dependency structure in the label space.

We demonstrate our model on large, real-world datasets in the clinical and web retail domains. We observe that hierarchical information is valuable when incorporated into the learning and improves our primary goal of multi-label classification. Our results show that a joint, hierarchical model outperforms a classification with unstructured labels as well as a disjoint model, where the topic model and the hierarchical classification are inferred independently of each other.

HSLDA is a model for hierarchically, multiply-labeled, bag-of-word data. We will refer to individual groups of bag-of-word data as documents. Let $w_{n,d} \in \Sigma$ be the nth observation in the dth document. Let $\mathbf{w}_d = \{w_{1,d}, \ldots, w_{1,N_d}\}$ be the set of N_d observations in document d. Let there be D such documents and let the size of the vocabulary be $V = |\Sigma|$. Let the set of labels be $\mathcal{L} = \left\{l_1, l_2, \ldots, l_{|\mathcal{L}|}\right\}$. Each label labels $l \in \mathcal{L}$, except root, has a parent $\mathrm{pa}(l) \in \mathcal{L}$ also in the set of labels. We will for exposition purposes assume that this label set has hard "is-a" parent-child constraints (explained later), although this assumption can be relaxed at the cost of more computationally complex inference. Such a label hierarchy forms a multiply rooted tree. Without loss of generality we will consider a tree with a single root $r \in \mathcal{L}$. Each document has a variable $y_{l,d} \in \{-1, 1\}$ for every label which indicates whether the label is applied to document d or not. In most cases $y_{i,d}$ will be unobserved, in some cases we will be able to fix its value because of constraints on the label hierarchy, and in the relatively minor remainder its value will be observed. In the applications we consider, only positive label applications are observed.

In HSLDA, documents are modeled using the LDA mixed-membership mixture model with global topic estimation. Label responses are generated using a conditional hierarchy of probit regressors[3]. The HSLDA graphical model is given in Figure 1. In the model, K is the number of LDA "topics" (distributions over the elements of Σ), $\boldsymbol{\phi}_k$ is a distribution over "words," $\boldsymbol{\theta}_d$ is a document-specific distribution over topics, $\boldsymbol{\beta}$ is a global distribution over topics, $\mathrm{Dir}_K(\cdot)$ is a K-dimensional Dirichlet distribution, $\mathcal{N}_K(\cdot)$ is the K-dimensional Normal distribution, $\mathbf{I}_K$ is the K dimensional identity matrix, $\mathbf{1}_d$ is the d-dimensional vector of all ones, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise.

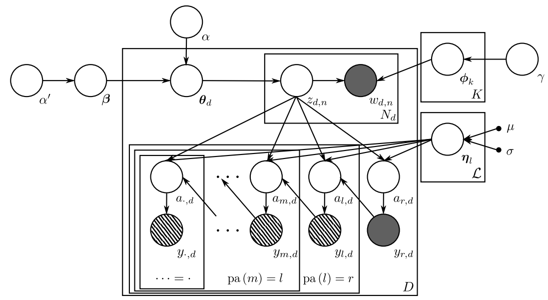
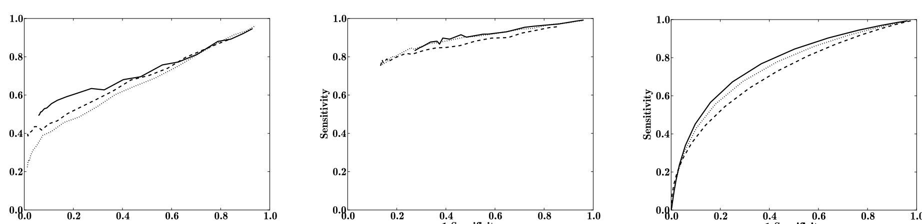


Figure 1: HSLDA graphical model

Posterior inference in HSLDA was performed using Gibbs sampling and Markov chain Monte Carlo. Note that, like in collapsed Gibbs samplers for LDA [4], we have analytically marginalized out the parameters $\boldsymbol{\phi}_{1:K}$ and $\boldsymbol{\theta}_{1:D}$. HSLDA also employs a hierarchical Dirichlet prior over topic assignments (i.e., $\boldsymbol{\beta}$ is estimated from data rather than assumed to be symmetric). This has been shown to improve the quality and stability of inferred topics [7]. The hyperparameters α, α', and γ are sampled using Metropolis-Hastings.

We applied HSLDA to data from two domains: predicting medical diagnosis codes from hospital discharge summaries and predicting product categories from Amazon.com product descriptions. The clinical dataset consists of 6,000 clinical notes along with associated billing codes that are used to document conditions that a particular patient was treated for. These billing codes (7298 distinct codes in our dataset) are organized in an is-a hierarchy. The retail dataset consists of product descriptions for DVDs from the Amazon.com product catalog. This data was partially obtained from the Stanford Network Analysis Platform (SNAP) dataset [1]. The comparison models included sLDA with independent regressors (hierarchical constraints on labels ignored) and HSLDA fit by first performing LDA then fitting tree-conditional regressions. The number of topics for all models was set to 50, the prior distributions of $p(\alpha)$, $p(\alpha')$, and $p(\gamma)$ were gamma distributed with a shape parameter of 1 and a scale parameters of 1000.



(a) (b) (c)

Figure 2: ROC curves for out-of-sample ICD-9 code prediction from patient free-text discharge records ((a),(c)). ROC curve for out-of-sample Amazon product category predictions from product free-text descriptions (b). Figures (a) and (b) are a function of the prior means of the regression parameters. Figure (c) is a function of auxiliary variable threshold. In all figures, solid is HSLDA, dashed are independent regressors + sLDA (hierarchical constraints on labels ignored), and dotted is HSLDA fit by running LDA first then running tree-conditional regressions.

The results in Figures 2(a) and 2(b) suggest that in most cases it is better to do full joint estimation of HSLDA. An alternative interpretation of the same results is that, if one is more sensitive to the performance gains that result from exploiting the structure of the labels, then one can, in an engineering sense, get nearly as much gain in label prediction performance by first fitting LDA and then fitting a hierarchical probit regression. There are applied settings in which this could be advantageous.

References

[1] Stanford network analysis platform. `http://snap.stanford.edu/`, 2004.

[2] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing*, 20: 121–128, 2008.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[4] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[5] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

[6] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[7] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.