Discussion of "The Discrete Infinite Logistic Normal Distribution for Mixed-Membership Modeling"

Frank Wood
Columbia University

1 Introduction

Mixed-membership models (e.g. "topic models") are inarguably popular; especially latent Dirichlet allocation (LDA) [Blei et al., 2003] and its variants. Such models have become a fundamental tool in the analysis and exploration of many types of data. Originally designed to model text documents as per-word draws from a document-specific weighting of a finite collection of "topics" (distributions over words), mixed-membership models now are applied very broadly. Example usage includes applications in information retrieval, image processing, audio classification, and more. Because of the wide applicability of mixed-membership modeling, improving models of this type has the potential to have significant impact. The discrete infinite logistic normal distribution (DILN) for mixed-membership modeling is a significant advance in mixed-membership modeling.

A key to the success of mixed-membership models is simplicity. They are easy to describe mathematically and easy to explain informally. The latent variables defined by such models are often readily interpretable by lay practitioners and often are visibly fascinating.

This kind of simplicity can also be an Achilles heel of sorts. To keep such models relatively simple, unreasonable assumptions about the nature of the true generative process must be made. This kind of trade-off is very common when designing or choosing a statistical model. The trick to designing good new models is to do so in such a way as to address the true shortcomings of the models being built upon while retaining desirable traits like interpretability, simplicity, and elegance. DILN is arguably such a contribution to mixed-membership modeling. DILN directly addresses one of the more pressing problems in mixed-membership modeling, namely, how to model correlations between latent features.

Mixed membership models are characterized by grouped observations (think of a bag-of-words representation of a document or a bag-of-features representation of an image) generated by a mixture of latent distributions over the observation space ("features"). In the canonical document modeling example a feature ("topic") is a distribution over words and a document is a collection of draws from a per-document mixture of topics. The fact that certain words tend to co-occur in the same document can be used to infer both what topics occur in a corpus and what proportions of each topic are found in each document.

Original work on mixed-membership modeling (notably LDA for document modeling) assumed both a fixed and finite number of statistically independent topics. In other words, the presence of one topic in a document was nearly independent of the presence of other topics, and the number of topics was fixed a priori to a pre-determined finite value. A great deal of subsequent work went into defining topic models of unbounded topic cardinality including, notably, the hierarchical Dirchlet process (HDP) as applied to mixed-membership modeling (HDP-LDA) [Teh et al., 2006]. Realistically though, specifying a mixed-membership model with a large number of latent features results in a model that is very nearly like that of a model with an unbounded number of features—Bayesian priors encourage sharing and discourage overfitting whether the number of topics is unbounded or large and fixed.

Less work has gone into learning mixed-membership models in which the latent factors are correlated (e.g. work by [Blei and Lafferty, 2006, Li et al., 2007, Doshi-Velez and Ghahramani, 2009, Rai and Daumé III, 2009]) though arguably this is the more important and less easy to remedy shortcoming of first generation mixed-membership models. To illustrate what is meant by correlations between latent factors we turn to an illuminating visual scene modeling example from the introduction to [Doshi-Velez and Ghahramani, 2009]. When modeling a visual scene using a mixed-membership model we can think of an "image"

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

as being a collection of observations of a world containing some number of latent features. Latent features here can be thought of as being in correspondence with physical objects in the world like lamps, desks, elephants, and so forth. Clearly a model that accounts for correlations in the presence and/or absence of features (desks and lamps often occur together, chairs and elephants less so) should outperform one that does not. A similar story can be told when modeling document collections using a topic representation, namely that some topics tend to occur together. DILN is able to account for correlations between feature presence rates for mixed-membership models of "bag-of-words" (order-invariant, discrete) grouped observations.

2 Key Idea

Conceptually, the key idea behind DILN is to imbue each latent feature (topic) with an arbitrary, free, latent "location" vector. The purpose for doing this is not to place each feature at some point in space, but instead is to let these vectors self-configure in way such that "close" features tend to co-occur more often than those that are "far" apart. This latent space embedding is similar to ideas that have found purchase in other domains such as models of data generated on graphs (e.g. Hoff et al. [2002]). Closeness in this space is computed using a distance (kernel) function that can be specified as part of the model definition, or, as is the case in this paper, learned implicitly.

3 Questions and Future Work

While DILN is apparently able to represent correlations between topics, the specific mechanism by which it accomplishes this has what might be in some applications undesirable characteristics. In particular, the topic prevalence covariance given by Eqn. 12 suggests that marginally prevalent topics are likely to more strongly co-vary with other marginally prevalent topics than with marginally less prevalent topics. It seems strange that the covariance of two topics should be coupled to their marginal prevalences. Additionally, the same covariance equation suggests that strong positive correlations are easy to achieve but strong negative correlations are hard to achieve. It would seem to be desirable to both be able to easily represent strong positive and strong negative correlations.

Despite the elegance of the combination of a Gamma process construction of the HDP and the latent space Gaussian process, the specific representations used in inference and the specific choice of mean-field variational inference algorithm for a truncated (finite topic cardinality) model suggest room for future improvement. That the GP kernel matrix is explicitly represented is interesting, in that no specific choice of kernel function must be made, but doing so directly precludes straightforward scaling to variational inference with large topic cardinality truncations. Likewise, sampling in a non-truncated representation would be computationally infeasible in the model as currently described.

More generally, representing correlations between latent features through distance in kernel space rather than explicitly through, for instance, more levels in a hierarchical probabilistic model, precludes sophisticated inference about latent feature similarities (no similarities between latent features beyond pairwise can be represented by this model). This is a probable direction for future work.

4 Conclusion

Accounting for covariance between high-level latent features is difficult but central in the development of next-generation latent variable models. DILN is a significant step in this direction, topic modeling is an excellent framework to "play with" the kind of ideas DILN embodies, and empirical evidence suggests the DILN approach to topic modeling is fruitful.

4.1 Acknowledgements

Thanks to S. A. Williamson for helpful comments and suggestions.

References

D. Blei and J. Lafferty. Correlated topic models. *Advances in Neural Information Processing Systems*, 18:147, 2006. ISSN 1049-5258.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003.

F. Doshi-Velez and Z. Ghahramani. Correlated nonparametric latent feature models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 143–150. AUAI Press, 2009.

P.D. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002. ISSN 0162-1459.

W. Li, D. Blei, and A. McCallum. Nonparametric Bayes pachinko allocation. In *Proceedings of the 23th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

P. Rai and H. Daumé III. The infinite hierarchical factor regression model. In *Advances in Neural Information Processing Systems 21*, pages 1321–3128, 2009.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.