Undirected Topic Models

Ruslan Salakhutdinov
(joint work with Geoffrey Hinton)
Brain and Cognitive Sciences and CSAIL
Massachusetts Institute of Technology
rsalakhu@mit.edu

1 Introduction

Probabilistic topic models [1, 4, 3] are often used to analyze and extract semantic topics from large text collections. Many of the existing topic models are based on the assumption that each document is represented as a mixture of topics, where each topic defines a probability distribution over words. The mixing proportions of the topics are document specific, but the probability distribution over words, defined by each topic, is the same across all documents.

All these models can be viewed as graphical models in which latent topic variables have directed connections to observed variables that represent words in a document. One major drawback is that exact inference in these models is intractable, so one has to resort to slow or inaccurate approximations to compute the posterior distribution over topics. A second major drawback, that is shared by all mixture models, is that these models can never make predictions for words that are sharper than the distributions predicted by any of the individual topics. They are unable to capture the essential idea of distributed representations which is that the distributions predicted by individual active features get multiplied together (and renormalized) to give the distribution predicted by many active features. This allows individual features to be fairly general but their intersection to be much more precise. For example, distributed representations allow the topics "government", "mafia" and "playboy" to combine to give very high probability to a word "Berlusconi" that is not predicted nearly as strongly by each topic alone.

To date, there has been very little work on developing topic models using undirected graphical models. Several authors [2, 10] used two-layer undirected graphical models, called Restricted Boltzmann Machines (RBMs), in which word-count vectors are modeled as a Poisson distribution. While these models are able to produce distributed representations of the input and perform well in terms of retrieval accuracy, they are unable to properly deal with documents of different lengths, which makes learning very unstable and hard. This is perhaps the main reason why these potentially powerful models have not found their application in practice. Directed models, on the other hand, can easily handle unobserved words (by simply ignoring them), which allows them to easily deal with different-sized documents. For undirected models marginalizing over unobserved variables is generally a non-trivial operation, which makes learning far more difficult. Recently, [6] attempted to fix this problem by proposing a Constrained Poisson model that would ensure that the mean Poisson rates across all words sum up to the length of the document. While the parameter learning has been shown to be stable, the introduced model is not a proper generative model of documents.

We introduce a "Replicated Softmax" model. The model can be efficiently trained using Contrastive Divergence, it has a better way of dealing with documents of different lengths, and computing the posterior distribution over the latent topic values is easy. We also demonstrate that the proposed model is able to generalize much better compared to a popular topic model, Latent Dirichlet Allocation (LDA) [1], in terms of both the log-probability on previously unseen documents and the retrieval accuracy, which indicates that the model is able to learn low-dimensional topic features that contain a lot of semantic information.

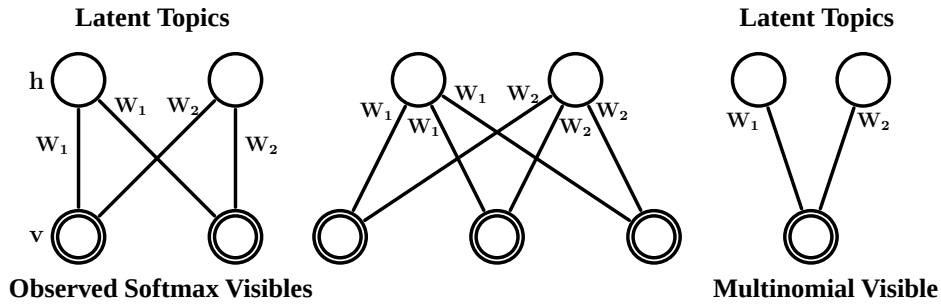

Figure 1: Replicated Softmax model. The top layer represents a vector $\mathbf{h}$ of stochastic, binary topic features and and the bottom layer represents softmax visible units $\mathbf{v}$. All visible units share the same set of weights, connecting them to binary hidden units. **Left:** The model for a document containing two and three words. **Right:** A different interpretation of the Replicated Softmax model, in which D softmax units with identical weights are replaced by a single multinomial unit which is sampled D times.

2 Replicated Softmax Model

We model discrete visible units $\mathbf{v}$ using a restricted Boltzmann machine, that has a two-layer architecture. Let $\mathbf{v} \in \{1, ..., K\}^D$, where K is the dictionary size and D is the document size, and let $\mathbf{h} \in \{0, 1\}^F$ be binary stochastic hidden topic features. Now suppose that for each document we create a separate RBM with as many softmax units as there are words in the document. Assuming we can ignore the order of the words, all of these softmax units can share the same set of weights, connecting them to binary hidden units, as shown in Fig. 1. We call this the "Replicated Softmax" model. A pleasing property of this model is that computing the approximate gradients of the CD objective for a document that contains 100 words is computationally not much more expensive than computing the gradients for a document that contains only one word. A key observation is that using D softmax units with identical weights is equivalent to having a single multinomial unit which is sampled D times, as shown in Fig. 1, right panel. Learning can be carried out efficiently using Contrastive Divergence.

2.1 Assessing Topic Models as Generative Models

We considered three datasets: The NIPS proceedings papers[1], the 20-newsgroups corpus, and The Reuters Corpus Volume I, which contains 804,414 documents. For each of the three datasets, we estimated the log-probability for 50 held-out documents. To estimate the log-probabilities of held-out documents we used Annealed Importance Sampling [5] for both LDA and Replicated Softmax models (for details see [7] and [9]). The average test perplexity per word was estimated as $\exp\left(-1/N \sum_{n=1}^{N} 1/D_n \log p(\mathbf{v}_n)\right)$, where N is the total number of documents, D_n and $\mathbf{v}_n$ are the total number of words and the observed word-count vector for a document n.

Table 1 shows that for all three datasets the 50-dimensional Replicated Softmax consistently outperforms the LDA with 50-topics. For the NIPS dataset, the undirected model achieves the average test perplexity of 3405, improving upon LDA's perplexity of 3576. The LDA with 200 topics performed much better on this dataset compared to the LDA-50, but its performance only slightly improved upon the 50-dimensional Replicated Softmax model. For the 20-newsgroups dataset, even with 200 topics, the LDA could not match the perplexity of the Replicated Softmax model with 50 topic units. The difference in performance is particularly striking for the large Reuters dataset, whose vocabulary size is 10,000. LDA achieves an average test perplexity of 1437, substantially reducing it from 2208, achieved by a simple smoothed unigram model. The Replicated Softmax further reduces the perplexity down to 986, which is comparable in magnitude to the improvement produced by the LDA over the unigram model. LDA with 200 topics does improve upon LDA-50, achieving a perplexity of 1142. However, its performance is still considerably worse than that of the Replicated Softmax model.

Figure 2 further shows three scatter plots of the average test perplexity per document. Observe that for almost all test documents, the Replicated Softmax achieves a better perplexity compared to the corresponding LDA model. For the Reuters dataset there are many documents that are mod-

[1] Available at http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.

Data set	Number of docs		K	$\bar{D}$	St. Dev.	Avg. Test perplexity per word (in nats)			
	Train	Test				LDA-50	LDA-200	R. Soft-50	Unigram
NIPS	1,690	50	13,649	98.0	245.3	3576	3391	3405	4385
20-news	11,314	7,531	2,000	51.8	70.8	1091	1058	953	1335
Reuters	794,414	10,000	10,000	94.6	69.3	1437	1142	988	2208

Table 1: Results for LDA using 50 and 200 topics, and Replaced Softmax model that uses 50 topics. K is the vocabulary size, $\bar{D}$ is the mean document length, St. Dev. is the estimated standard deviation in document length.

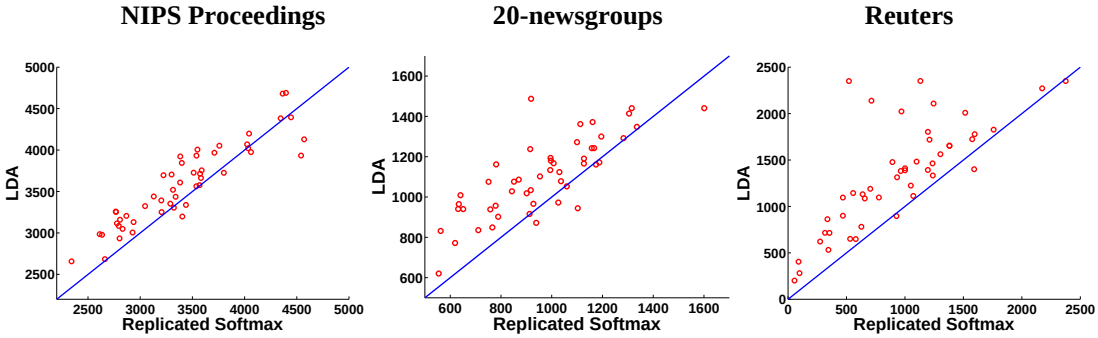

Figure 2: The average test perplexity scores for each of the 50 held-out documents under the learned 50-dimensional Replicated Softmax and LDA that uses 50 topics.

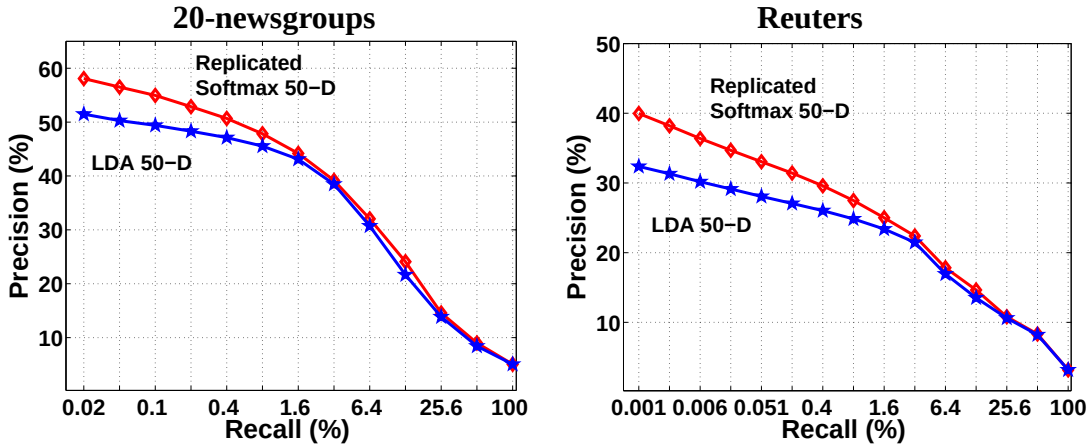

Figure 3: Precision-Recall curves for the 20-newsgroups and Reuters datasets, when a query document from the test set is used to retrieve similar documents from the training corpus. Results are averaged over all 7,531 (for 20-newsgroups) and 10,000 (for Reuters) possible queries. Both LDA and Replicated Softmax models used 50 topic units.

eled much better by the undirected model than an LDA. Clearly, the Replicated Softmax is able to generalize much better.

2.2 Document Retrieval

We used 20-newsgroup and Reuters datasets to evaluate model performance on a document retrieval task. To decide whether a retrieved document is relevant to the query document, we simply check if they have the same class label. This is the only time that the class labels are used. For the Replicated Softmax, the mapping from a word-count vector to the values of the latent topic features is fast, requiring only a single matrix multiplication followed by a componentwise sigmoid non-linearity.

1	neuronal, firing, spike, spikes, fire, stimulus, membrane, neuron, activity, period
2	connectionist, representations, hinton, task, human, training, tasks, rumelhart, network, performance
3	error, theorem, function, case, bounds, functions, learning, optimal, generalization, problem
4	chip, cmos, fabricated, transistor, transistors, capacitor, charge, analog, chips, vlsi
5	likelihood, bayesian, em, mixture, gaussian, estimation, probabilistic, distribution, distributions, variance
6	policy, reinforcement, singh, policies, sutton, discounted, mdp, bellman, barto, reward
7	spatial, surround, orientation, receptive, selectivity, retinal, visual, fields, tuned, selective
8	hopfield, basins, attractors, basin, attractor, bifurcation, attraction, stability, collective, memories

Table 2: Discovered topics in the NIPS proceedings dataset. Topics were extracted from the distributed representations learned by the Replicated Softmax model.

For the LDA, we used 1000 Gibbs sweeps per test document in order to get an approximate posterior over the topics. Figure 3 shows that when we use the cosine of the angle between two topic vectors to measure their similarity, the Replicated Softmax significantly outperforms LDA, particularly when retrieving the top few documents.

3 Conclusions

We have proposed a simple two-layer undirected topic model that be used to model and automatically extract distributed semantic representations from large collections of text corpora. The model can be viewed as a family of different-sized RBM's that share parameters. The proposed model have several key advantages: First, the learning is easy and stable, it can model documents of different lengths, and computing the posterior distribution over the latent topic values is easy. Second, using stochastic gradient descent, scaling up learning to billions of documents would not be particularly difficult. Finally, the proposed model is able to generalize much better than LDA in terms of both the log-probability on held-out documents and the retrieval accuracy.

In addition, we can use learned distributed representations as an input to the HDP model [8], which allows us to discover more interpretable topics (see table 2). The resulting model can viewed as a semi-parametric two-layer deep belief network. Clearly, the proposed model is able to both learn distributed representations, which can be used for fast and accurate information retrieval/document classification tasks, as well as discover global semantic topics, which could greatly facilitate exploratory data analysis of large document collections.

References

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

[2] P. Gehler, A. Holub, and M. Welling. The Rate Adapting Poisson (RAP) model for information retrieval and object recognition. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[3] Thomas Griffiths and Mark Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1), 2004.

[4] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in AI*, pages 289–296, San Fransisco, California, 1999. Morgan Kaufmann.

[5] R. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

[6] R. Salakhutdinov and G. Hinton. Semantic Hashing. In *SIGIR workshop on Information Retrieval and applications of Graphical Models*, 2007.

[7] R. Salakhutdinov and I. Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the International Conference on Machine Learning*, volume 25, pages 872 – 879, 2008.

[8] 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.

[9] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, 2009.

[10] E. Xing, R. Yan, and A. Hauptmann. Mining associated text and images with dual-wing harmoniums. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI-2005)*, 2005.