Focused Topic Models

Sinead Williamson Cambridge University
Chong Wang Princeton University
Katherine Heller Cambridge University
David Blei Princeton University

Abstract

We present the *focused topic model* (FTM), a family of nonparametric Bayesian models for learning sparse topic mixture patterns. The FTM integrates desirable features from both the hierarchical Dirichlet process (HDP) and the Indian buffet process (IBP) – allowing an unbounded number of topics for the entire corpus, while each document maintains a sparse distribution over these topics. We observe that the HDP assumes correlation between the global and within-documant prevalences of a topic, and note that such a relationship may be undesirable. By using an IBP to select which topics contribute to a document, and an unnormalized Dirichlet Process to determine how much of the document is generated by that topic, the FTM decouples these probabilities, allowing for more flexible modeling. Experimental results on three text corpora demonstrate superior performance over the hierarchical Dirichlet process topic model.

1 Introduction

Probabilistic topic modelling has emerged as a powerful method of unsupervised analysis of document collections. Topic models uncover the salient patterns of a collection under the *mixed-membership assumption*: Each data point can exhibit multiple patterns to different degrees. When analysing text, these patterns are represented as distributions over words and often place high probability on semantically meaningful sets. Thus, they are called "topics."

A central problem in many topic models is that the number of topics must be specified in advance. In many settings we are unlikely to have good intuitions regarding the "correct" number – if such a concept is even meaningful. Moreover, the number of topics is likely to grow with the number of observed documents. Thus, we would like a flexible model where the observed data (and model prior) determine this number for us. This flexibility is provided by nonparametric Bayesian methods. Specifically, the hierarchical Dirichlet process (HDP) [4] can be used as a topic model with an unbounded number of topics.

While we do not expect to be able to model all documents with a fixed finite number of topics, it is reasonable to assume that a given document will only contain a small subset of possible topics. The HDP allows sparsity over possible topics in a document, but only when the distribution over those topics is peaky. Obtaining greater flexibility in the distribution over topics in a document involves increasing the prior expectation of the number of topics. Thus there is an inherent trade-off between sparsity and flexibility in the distribution over topics in the HDP.

Furthermore, topic proportions in a particular document are tied to the topic proportions in the corpus overall. There is no way for a document to pick and choose, or to focus in on its own major topics without being highly influenced by the prevalence of those topics across the entire corpus. In other words, the HDP assumes a positive correlation between the global prevalance of a topic and the prevalence of that topic within a given document. This might not be the case in the datasets we are modeling - a topic may occur infrequently throughout a corpus, but dominate the few documents in which it does occur.

We present the *focused topic model*, a nonparametric Bayesian topic model that decouples the probability of a topic appearing within a document from the importance of that topic within the document. This model combines the flexible distributions of the HDP with the sparsity of the IBP, to give a flexible, compact representation. The name "focused" reflects the fact that, while the total number of topics represented by the model is unbounded, the latent topic distribution associated with a specific document is focused on a sparse subset of these topics.

2 Model

The FTM is based on a distribution over infinite matrices of integers, which is constructed using the Indian buffet process (IBP) [1]. The IBP can be described in terms of a series of customers choosing from a buffet with an infinite number of dishes. A customer chooses a finite number of dishes according to $\text{Poisson}(\alpha)$. He then chooses the specific dishes according to their popularity with previous customers. In our variant, each dish is associated with a parameter ϕ_k, from which the customer draws an integer $n \sim \text{Poisson-gamma}(\phi_k, 1)$. This integer determines the size of the portion the customer selects from that dish.

As successive customers make their way through the buffet, we build a random matrix with customers in rows, dishes in columns, and integers in cells. In our topic model, the customers represent documents, and the dishes represent topics. The integer associated with a specific document-topic pair gives the number of words in that document that have been generated by that topic.

More formally, we can describe the generative process as follows:

1. Draw the infinite binary matrix $\mathbf{B} \sim \text{IBP}(\alpha)$.
2. For each topic k, draw a topic proportion parameter $\phi_k \sim \text{Gamma}(\gamma, 1)$, and draw the associated distribution over words $\boldsymbol{\beta}_k \sim \text{Dirichlet}(\eta)$.
3. For each document $m = 1, \ldots, M$:
   - For each topic $k = 1, 2, \ldots$:
     (a) Draw the number of words from that topic, $n_k^{(m)} \sim \text{Poisson-gamma}(b_{mk}\phi_k, 1)$
     (b) For each word from that topic, $w_{mki} : i = 1, \ldots, n_k^{(m)}$:
         – Draw $w_{mki} \sim \text{Discrete}(\boldsymbol{\beta}_k)$

Noting that the total number of words in the m^{th} document is distributed according to $n_{\cdot}^{(m)} \sim \text{Poisson-gamma}(\sum_k b_{mk}\phi_k, 1)$, we can write an equivalent representation of the generation of the m^{th} document under the FTM in the following Dirichlet-multinomial format:

1. Draw the total number of words for the document, $n_{\cdot}^{(m)} \sim \text{Poisson-gamma}(\sum_k b_{mk}\phi_k, 1)$.
2. If $n_{\cdot}^{(m)} > 0$
   (a) Draw $\boldsymbol{\theta}_m \sim \text{Dirichlet}(\mathbf{b}_m \cdot \boldsymbol{\phi})$, where $\mathbf{a} \cdot \mathbf{b}$ is the Hadamard product of $\mathbf{a}$ and $\mathbf{b}$.
   (b) For each word $w_{mi} : i = 1, \ldots, n_{\cdot}^{(m)}$
      i. Draw the topic allocation $z_{mi} \sim \text{Discrete}(\boldsymbol{\theta}_m)$.
      ii. Draw $w_{mi} \sim \text{Discrete}(\beta_{z_{mi}})$.

We use a stick-breaking representation of the Indian buffet process, described in [3]. In this representation, we introduce a stick-length π_k for each column, such that, for a strictly decreasing stick ordering:

$$\mu_k \sim \text{Beta}(\alpha, 1) \qquad \pi_k = \textstyle\prod_{j=1}^{k} \mu_j \qquad b_{mk} \sim \text{Bernoulli}(\pi_k) \tag{1}$$

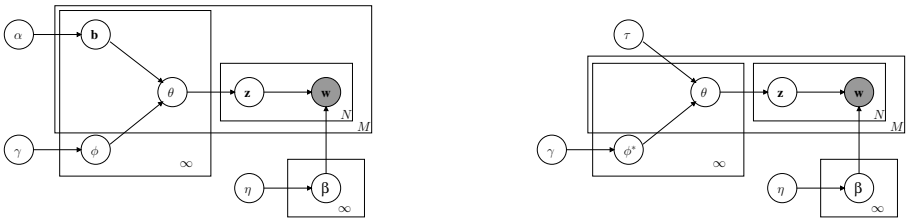
2.1 Relationship to the Hierarchical Dirichlet Process

We note that the Dirichlet distribution can be constructed by normalisation of a set of Gamma random variables. Letting $\boldsymbol{\phi}^* = \frac{\boldsymbol{\phi}}{\tau}$, where $\tau = \sum_{k=1}^{\infty} \phi_k$, we can rewrite the Dirichlet-multinomial

representation of the FTM as:

$$\begin{aligned}
\boldsymbol{\phi}^* | \gamma &\sim \text{stick}(\gamma) \\
\boldsymbol{\theta}_m &\sim \text{DP}\left(\tau \sum_k b_{mk}\phi_k^*, \frac{\mathbf{b} \cdot \boldsymbol{\phi}^*}{\sum_k b_{mk}\phi_k^*}\right) \\
\mathbf{z}_m | \boldsymbol{\theta}_m, n_\cdot^{(m)} &\sim \text{Multinomial}(\boldsymbol{\theta}_m, n_\cdot^{(m)})
\end{aligned}$$

where 'stick' refers to the stick-breaking construction for the Dirichlet Process [2]. The above representation is similar to the formulation of the HDP-based topic model [4]. In both cases, $\boldsymbol{\theta}_m$ is drawn from a DP; in the HDP, the base distribution for this DP is another DP; in the FTM, the base distribution is an unnormalized DP which is 'masked' by $\mathbf{b}_m$ and then normalized. Figure 1 shows the graphical models for the two models.



(a) FTM (b) HDP

Figure 1: Graphical models for FTM and HDP

3 Inference

We use a Gibbs sampling scheme to infer the posterior distribution over latent variables, cyclically sampling the value for a single variable from its probability distribution conditioned on the current values of the remaining variables. We integrate out the topic-specific word distributions $\boldsymbol{\beta}$ and the sparsity pattern $\mathbf{B}$, sampling only the global topic proportion variables $\boldsymbol{\phi}$, the global topic sparsity probability variables $\boldsymbol{\pi}$, and the topic assignments $\mathbf{z}$.

Integrating out $\mathbf{B}$ exactly is intractable due to the infinite sum involved. We use an approximation to the exact integral in our inference.

4 Experiments

We compared the FTM to the HDP topic model on three datasets:

- A collection of 1766 abstracts from the Proceedings of the National Academy of Sciences (PNAS) from the years 1991 to 2001, with a vocabulary of 2452 words.
- A collection of 1000 randomly selected articles from the 20 newsgroups dataset[1], with a vocabulary of 1407 words.
- A collection of 2000 randomly selected documents from the Reuters-21578 dataset[2], with a vocabulary of 1472 words.

In each case,the vocabulary excluded stop-words and words occurring in fewer than 5 documents.

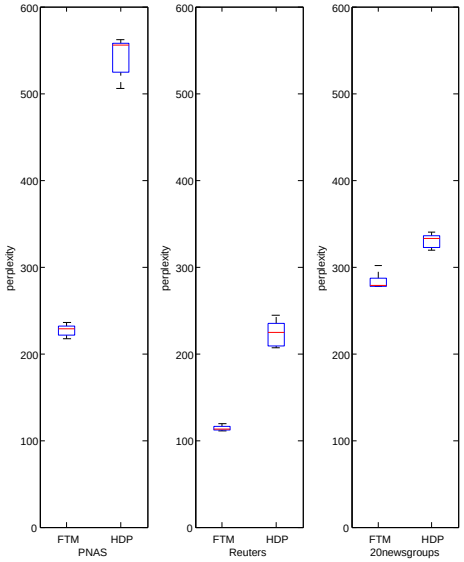
Figure 2(a) shows the perplexities obtained on these datasets using the two models, evaluated using 5-fold cross validation. In each case, the FTM achieves lower perplexity on the held out data.

In motivating the Focused Topic Model, we wished to increase the flexibility to vary the probability of a topic being active within a document, and the proportion of the document attributed to that

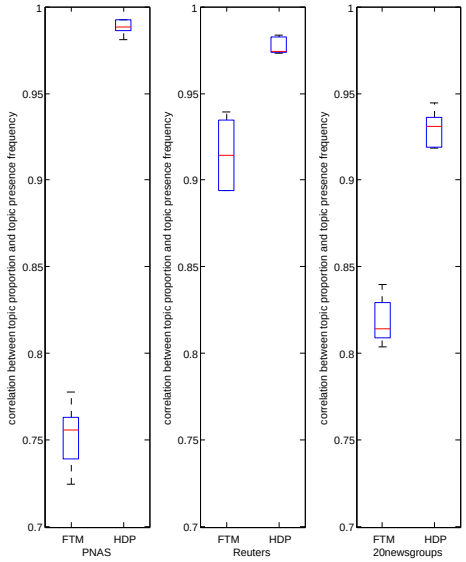
[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

[2] http://kdd.ics.uci.edu/databases/reuters21578/

topic, avoiding the correlation bias inherent in the HDP. To consider whether this is observed in the models learnt, we compared two statistics for each topic found. We first define the *topic presence frequency* for a given topic as the fraction of the documents within the corpus that contain at least one incidence of that topic. We then define the *topic proportion* for a topic as the fraction of the words within the corpus attributed to that topic. The correlation between topic presence frequencies and topic proportions is shown in the box plots of figure 2(b). In each case we see lower correlation for the FTM. We note that the dataset with the greatest decrease in correlation under the FTM – PNAS – is also the dataset with the greatest improvement in perplexity under the FTM.



(a) Test set perplexities across three datasets



(b) Correlation between topic presence frequency and topic proportion, across three datasets

5 Conclusion

We have presented the *focused topic model* (FTM), a non-parametric topic model that models corpora using an unbounded number of latent topics, while describing individual documents within a given corpus using a finite subset of these topics. We use a sparse binary matrix drawn from an IBP to enforce sparsity in the latent document representations. This binary matrix acts to decouple the probability of a topic being present in a document, and the relative importance of a topic within a document. Our experiments demonstrate that the FTM achieves lower test set perplexity than the HDP topic model across a range of corpora. To perform inference in the FTM, we have developed an efficient Gibbs sampling strategy which avoids sampling from infinite combinatorial space.

References

[1] Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2005.

[2] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[3] Y. W Teh, D. Gorur, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *AIStats*, 2007.

[4] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *JASA*, 101(476):1566–1581, 2006.