hca: Tool for Multicore Non-parametric Topic Models

Wray Buntine   WRAY.BUNTINE@MONASH.EDU
Faculty of Information Technology
Monash University
Clayton, VIC, 3800, Australia

Editor: Enid Blighton

Abstract

This paper describes `hca`, an open source command-line tool to train and test topic models. The tool implements a variety of Gibbs samplers for non-parametric models using an efficient implementation of hierarchical Pitman-Yor processes. These are used for both the document-topic component and the topic-word component, and to model burstiness of words in topics. Various diagnostics, document completion testing and coherence measurements with PMI are also supported. This is also the first efficient multicore non-parametric topic model to be distributed. The package consists in a main command-line tool and a set of support utilities. The documentation includes a user's guide with a mini tutorial.

Keywords: topic model, hierarchical Pitman-Yor process, non-parametric Bayesian model, Gibbs sampling

1. Introduction

Topic models are a form of non-negative matrix factorisation and some versions also correspond to normalised independent (multinomial) components. They were originally developed as a Bayesian variant of probabilistic latent semantic analysis. The early model, Latent Dirichlet Allocation (LDA), used a simple symmetric Dirichlet prior for columns of the document-topic component and rows of the topic-word component, but researchers soon realised non-symmetric priors could improve modelling performance. Using a Dirichlet process (DP) as a prior for columns of the document-topic component was proposed by Teh et al. (2006), and Teh distributed a relatively robust Gibbs sampler for the task Teh (2004). While this purportedly allows the model to "pick" the right number of topics, more importantly, it allows some topics to be rarer than others. In simple symmetric models, all topics are equally likely *apriori*.

Over the subsequent decade many research efforts have attempted to improve on Teh's algorithm, using methods including collapsed Gibbs samplers, approximate fitting and variational methods. Others have developed alternative extensions to LDA. Sato and Nakagawa (2010) have also placed a Pitman-Yor process (PYP) as a prior for the rows of the topic-word component. The common prior acts like a "background" topic: it alleviates the need for non-topical words (like stop words) to be modelled by a single topic because they are shared across all topics. Doyle and Elkan (2009) have modelled rows of the topic-word component using a Dirichlet Compound multinomial, rather than a multinomial, in order to introduce "burstiness," whereby some words in a document are encouraged to appear multiple times

(in a "burst"). This tends to produce models with less topics, and dramatically improves perplexity. We implement this with a PYP, yielding a small factor overhead in space and time for our standard collapsed Gibbs sampler.

While there are numerous other improvements made to vanilla topic modelling, for instance using an Indian buffet process on both component matrices to sparsify them, from Archambeau et al. (2015), these above are the ones we reproduce in our tool `hca`:

- DPs and PYPs on the document-topic component and the topic-word component,

- burstiness via an additional document specific PYP on topic-word component, and

- hyperparameter sampling (for the nonparametric models) by default using adaptive rejection sampling of Gilks and Wild (1992).

Results using the tool were presented at KDD 2014 by Buntine and Mishra (2014). The implementation uses the table indicator sampling method of Chen et al. (2011) which is a collapsed version of Teh's sampler but requires just a small factor in space and time overhead over the traditional collapsed sampler for LDA of Griffiths and Steyvers (2004). Most importantly, it uses no dynamic memory like the Chinese restaurant samplers.

2. The `hca` Commandline Tool

The `hca` tool is distributed on the MPL 2.0 license and is implemented as a Unix style command line tool with multiple options for tasks including specifying a variety of different non-parametric topic models, training and testing models, reporting diagnostics, initialising and controlling fitting of hyperparameters, restarts and checkpointing. These options are all documented in a standard Unix style man page with duplicate PDF version. Moreover, further examples of using the tool are given at the end of the man page.

The tool itself is implemented in C. A multicore version can be compiled which uses threads and atomic operations available in C11 to keep data consistent. This gets, for instance, about 5 times speed-up with 8 cores, and uses little memory duplication so can work with quite large data sets (gigabytes of text). In our experience, however, it does not work well with more cores. The code is self-contained, with no library dependencies. Source code for tasks such as adaptive rejection sampling, and Gamma variable sampling, have been included from existing open source libraries to eliminate dependencies. Compilation and testing has been successful on Mac OSX and a variety of Unices, and single-core compilation has been done under Cygwin.

While the development repository is in Github[1], the best way to obtain the code is via MLOSS[2] because this combines the two components.

3. Using the System

While the tool accepts many different data formats, the most common is a format sometimes referred to as the LDA-C format. This is a sparse matrix format where each line represents a document, and the line contains a list of "index:count" pairs giving the non-zero

---

1. `https://github.com/wbuntine/topic-models` and `https://github.com/wbuntine/libstb`
2. `http://www.mloss.org/software/view/527/`

entries of the document by word matrix. An example dataset ("ch.ldac" and its companion "ch.tokens" listing the words corresponding to the indices) is included in the release. This is composed of 400 news articles about church, "bagged up" as a data product from Lewis et al. (2004). A typical sequence of commands is:

```
hca -v -C200 -K40 -q6 data/ch /tmp/C1
hca -v -v -V -r0 -C0 -orat,10 -e data/ch /tmp/C1
```

The first line builds a 40-topic model with 300 Gibbs cycles using the default configuration for hca, run on 6 cores in 12 seconds. The default configuration has a DP on the document-word component, a PYP on the topic-word component, both with a GEM at their root, and does no burstiness modelling. The second lines reports on the learnt model and parameters giving the top ten words per topic. For instance, the PYP on the topic-word component settled on a discount of 0.31 and concentration of 209, and the DP on the document-word component settled on a concentration of 0.8. The three most populous topics have significant words that show the era of the articles (1986-1987):

```
teresa,missionaries,nirmala,calcutta,nursing,respirator
diana,parker,bowles,camilla,charles,divorce
appendix,crucitti,tumour,parkinson,trembling,pope
```

More examples of the system, for instance document completion testing Buntine (2009), and MCMC estimates of the component matrices are given in the man page. Note several versions of document completion are implemented in the paper as different authors use them.

4. Related Software

Acknowledgments

References

C. Archambeau, B. Lakshminarayanan, and G. Bouchard. Latent IBP compound Dirichlet allocation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2): 321–333, 2015.

W. Buntine. Estimating likelihoods for topic models. In *Proceedings of the 1st Asian Conference on Machine Learning*, Nanjing, China, 2009.

W.L. Buntine and S. Mishra. Experiments with non-parametric topic models. In *20th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, page 881890, 2014.

C. Chen, L. Du, and W. Buntine. Sampling table configurations for the hierarchical Poisson-Dirichlet process. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD*, pages 296–311. Springer, 2011.

G. Doyle and C. Elkan. Accounting for burstiness in topic models. In *Proc. of the 26th Annual Int. Conf. on Machine Learning*, ICML '09, pages 281–288, 2009.

Walter R Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

D.D. Lewis, Y. Yang, T.G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5, December 2004.

I. Sato and H. Nakagawa. Topic models with power-law using Pitman-Yor process. In *16th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, pages 673–682. ACM, 2010.

Y. W. Teh. Nonparametric Bayesian mixture models - release 2.1. `http://www.stats.ox.ac.uk/~teh/software.html`, 2004. MATLAB and C code [Online; accessed 18 March 2016].

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