A fully Bayesian view of Latent Dirichlet Allocation

Anonymous Author 1
Unknown Institution 1

Anonymous Author 2
Unknown Institution 2

Anonymous Author 3
Unknown Institution 3

Abstract

Abstract...

1 Introduction

2 Related work

essai git

3 A conjugate prior for the Dirichlet distribution

It is well known that the exponential family of distributions satisfies important stability properties, which lead to elegant, analytical expressions in Bayesian inference, hence its massive use in Bayesian models. One property is that any distribution in that family has a natural conjugate also in that family, and there is a systematic procedure to compute it. For example, the Dirichlet distribution can be derived in this way as a conjugate to the multinomial distribution, and they are both in the exponential family. In the sequel, we are interested in a conjugate to the Dirichlet itself, which is easy to derive by application of the same systematic procedure. We name it here the **Boojum** distribution[1], for lack of a better name. It is a distribution over the positive orthant $\mathbb{R}_+^N$ (the parameter space of the N-dimensional Dirichlet), and has two parameters $m, \boldsymbol{\tau}$ where $m \in \mathbb{R}$ is a scalar and $\boldsymbol{\tau} \in \mathbb{R}^N$ is a vector (as a general rule, the parameter space of the conjugate has dimension one plus the dimension of the parameter space of the original distribution). It is defined, for $\boldsymbol{x} \in \mathbb{R}_+^N$ by

$$\textbf{Boojum}(\boldsymbol{x}; m, \boldsymbol{\tau}) \triangleq \frac{1}{Z(m, \boldsymbol{\tau})} \mathcal{B}(\boldsymbol{x})^m \exp -\boldsymbol{\tau}\boldsymbol{x} \quad (1)$$

where $\mathcal{B}$ denotes the multivariate beta function $\mathcal{B}(\boldsymbol{x}) \triangleq \frac{\prod_n \Gamma(x_n)}{\Gamma(\sum_n x_n)}$ which is also the normalising constant of the Dirichlet distribution with parameter $\boldsymbol{x}$, and $Z(m, \boldsymbol{\tau})$ is the normalising constant of the **Boojum** distribution itself, defined by

$$Z(m, \boldsymbol{\tau}) \triangleq \int_{\boldsymbol{x} \in \mathbb{R}_+^N} \mathcal{B}(\boldsymbol{x})^m \exp -\boldsymbol{\tau}\boldsymbol{x} \, \mathbf{d}\boldsymbol{x}$$

The expression $\boldsymbol{\tau}\boldsymbol{x}$ in the definition denotes the scalar product of the two vectors. Of course, for the distribution to be proper, the normalising constant must be finite, which is not always the case. Although no analytical formula is known for $Z(m, \boldsymbol{\tau})$, one exists for its finiteness:

Proposition 1. *The distribution* $\textbf{Boojum}(m, \boldsymbol{\tau})$ *is proper, i.e.* $Z(m, \boldsymbol{\tau}) < \infty$*, if and only if*[2]

$$\begin{array}{l} \forall n \in N \; \tau_n > 0 \quad \textit{and} \quad m < 1 \quad \textit{and} \\ (m \geq 0 \quad \textit{or} \quad \sum_{n \in N} \exp -\frac{\tau_n}{|m|} < 1) \end{array}$$

Now, conjugacy with the Dirichlet distribution is expressed by the following property.

Proposition 2. *Let* $(\boldsymbol{y}_p)_{p \in P}$ *be a family of random variables over the simplex of* $\mathbb{R}_+^N$*, which are mutually independent given a random variable* $\boldsymbol{x}$ *over* $\mathbb{R}_+^N$*. We have*

$$\begin{array}{ll} \textit{Prior:} & \boldsymbol{x} \sim \textbf{Boojum}(m, \boldsymbol{\tau}) \\ \textit{Observation:} & \forall p \in P \; \boldsymbol{y}_p | \boldsymbol{x} \sim \textbf{Dirichlet}(\boldsymbol{x}) \\ \implies \textit{Posterior:} & \boldsymbol{x} | (\boldsymbol{y}_p)_{p \in P} \sim \textbf{Boojum}(m', \boldsymbol{\tau}') \end{array}$$

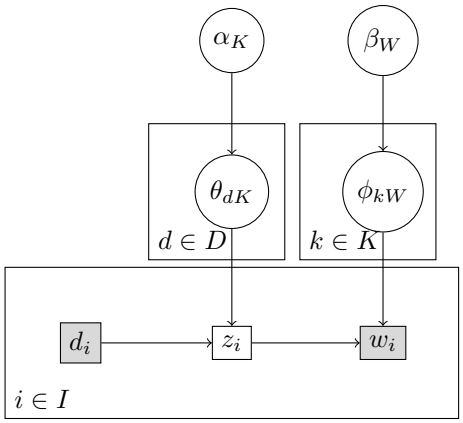
where $m' = m - |P| \;\; \boldsymbol{\tau}' = \boldsymbol{\tau} - \sum_{p \in P} \log \boldsymbol{y}_p$

This holds whenever the prior is proper, in which case so is the posterior.

[1] a.k.a. the Snark distribution, because the Snark is a Boojum, you see.

Preliminary work. Under review by AISTATS 2016. Do not distribute.

[2] For any whole number N, we use the shorthand $n \in N$ to mean $n \in \{1 \dots N\}$



$$
\begin{aligned}
&q(\alpha_K \beta_W \theta_{DK} \phi_{KW} z_I w_I | d_I) = \\
&\quad q(\alpha_K) q(\beta_W) \\
&\quad \textstyle\prod_{d \in D} q(\theta_{dK} | \alpha_K) \prod_{k \in K} q(\theta_{kW} | \beta_W) \\
&\quad \textstyle\prod_{i \in I} q(z_i | d_i \theta_{DK}) q(w_i | z_i \phi_{KW})
\end{aligned}
$$

$$
\begin{array}{rrrcl}
\alpha_K & \sim & q^{(K)} & \sim & q^{(W)} \\
\forall d \in D & & \theta_{dK} | \alpha_K & \sim & \mathbf{Dirichlet}(\alpha_K) \\
\forall k \in K & & \phi_{kW} | \beta_W & \sim & \mathbf{Dirichlet}(\beta_W) \\
\forall i \in I & & z_i | d_i \theta_{DK} & \sim & \mathbf{Cat}(\theta_{d_i K}) \\
\forall i \in I & & w_i | z_i \phi_{KW} & \sim & \mathbf{Cat}(\phi_{z_i W})
\end{array}
$$

$$
\begin{aligned}
\log q(\alpha_K \beta_W \theta_{DK} \phi_{KW} z_I | w_I d_I) &= \log q^{(K)}(\alpha_K) + \log q^{(W)}(\beta_W) - D \log \mathcal{B}(\alpha_K) - K \log \mathcal{B}(\beta_W) \\
&+ \sum_{dk} (\alpha_k - 1) \log \theta_{dk} + \sum_{kw} (\beta_w - 1) \log \phi_{kw} + \sum_{ik} \mathbb{I}[z_i = k] (\log \theta_{d_i k} + \log \phi_{k w_i})
\end{aligned} \tag{2}
$$

Figure 1: The full Bayesian network of the LDA model and the decomposition of its joint distribution q.

4 The LDA model revisited

Given a number D of documents, W of words (or terms), K of topics and I of occurrences, a realisation consists of the following random variables[3]:

- a tuple $(d_i, z_i, w_i)_{i \in I}$, where for each $i \in I$, the discrete objects $d_i \in D, z_i \in K, w_i \in W$ denote, respectively, the document, the topic and the word associated with occurrence i;
- two stochastic matrices θ_{DK} (of dimension $D \times K$) and ϕ_{KW} (of dimension $K \times W$) which characterise each document d as the distribution of topics θ_{dK} and each topic k as the distribution of words ϕ_{kW};
- two positive vectors α_K and β_W which characterise the topics and words of the whole collection.

The space of realisations is that of complete collections of documents: this explains why the collection-level variables α_K and β_W are random, and not parameters. Besides the known size parameters D, W, K, I, the model has two possibly unknown parameters $q^{(K)}$ and $q^{(W)}$, which are distributions for α_K and β_W, respectively. For the sake of symmetry, the occurrence-document vector d_I is considered a random variable, but it is assumed always observed and independent of all the rest, so it could as well have been treated as a known parameter. All probability expressions are conditioned upon it. The full graphical representation of the LDA model is given in Figure 1. Conditioned to the observation of the occurrence-word vector w_I, the log probability is given, up to some additive constant depending only on d_I, w_I, by

Note that, at this point, we make no assumption on the distributions $q^{(K)}$ and $q^{(W)}$. Our main contribution is precisely in studying the impact of the choice of these distributions. We show that an appropriate choice leads to a new formulation of the variational approximation of LDA, which we investigate.

[3]In the sequel, we take the convention of not distinguishing the vectors by typesetting them in boldface, but rather by systematically recalling their index ranges as subscripts.

5 Fully variational Bayes for Latent Dirichlet Allocation

6 Experiments

7 Discussion

Acknowledgements

Use unnumbered third level headings for the acknowledgements. All acknowledgements go at the end of the paper. Be sure to omit any identifying information in the initial double-blind submission!

References

J. Alspector, B. Gupta, and R. B. Allen (1989). Performance of a stochastic learning microchip. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 1*, 748-760. San Mateo, Calif.: Mor-

gan Kaufmann.

F. Rosenblatt (1962). *Principles of Neurodynamics.* Washington, D.C.: Spartan Books.

G. Tesauro (1989). Neurogammon wins computer Olympiad. *Neural Computation* **1**(3):321-323.