Probabilistic graphical models, 2015-2016 Master MVA, ENS Cachan

Project progress report

Jérôme DOCKES, jdockes@ens-cachan.fr
Pascal LU, pascal.lu@student.ecp.fr

We are mainly based on the research paper *Latent Dirichlet allocation* written by D. Blei, A. Ng, and M. Jordan and published in *Journal of Machine Learning Research*, 3:993-1022, January 2003.

Presentation of the model

We consider a corpus D composed of $|D|$ documents, k topics and a vocabulary list V.
Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.
LDA is based on the computation of the parameters (α, β) where α is the parameter of the Dirichlet distribution which generates the parameter θ for the multinomial probability distribution over topics in the document and β gives the probability that a given topic will generate a certain word: $\beta_{ij} = p(w^j = 1 | z^i = 1)$.
To compute (α, β), the idea is to introduce, for each document d, $\gamma^{(d)}$ (the variational parameter for the Dirichlet distribution), $\phi^{(d)}$ (the variational parameter for the multinomial distribution, matrix of size number of words in document $d \times$ number of topics) and $w^{(d)}$ contains the number of times each word in the vocabulary appears in the document.

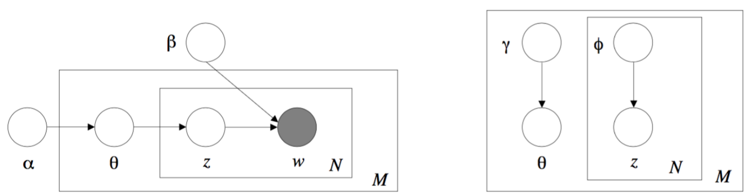

Figure 1: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA

Main principle of the algorithm

The algorithm is based on a variational expectation-maximization algorithm.

Algorithm 1: E-step for a document d

Data: `word_incidences` ($w^{(d)}$), `dirich_param` (α), `word_prob_given_topic` (β)
Result: `var_dirich` ($\gamma^{(d)}$), `var_multinom` ($\phi^{(d)}$)

Algorithm 2: M-step

Data: {`word_incidences` ($w^{(d)}$), `var_dirich` ($\gamma^{(d)}$), `var_multinom` ($\phi^{(d)}$), $d \in D$}
Result: `dirich_param` (α), `word_prob_given_topic` (β)

Implementation and problems

The preprocessing step (which reads the documents) has been implemented. However, the algorithm described earlier has been implemented but contains some bugs.

- We need to initialize of α and β before starting the EM-algorithm. The initialization step still remains a problem.
- The computation of β was simplified. It is essentially based on the sum of $(\phi^{(d)})^\top w^{(d)}$ for $d \in D$.
- The computation of α uses a Newton-Raphson algorithm "$\alpha \leftarrow \alpha - H(\alpha)^{-1} g(\alpha)$". However, the convergence of α depends on the initialization of α.
- A basic stopping criterion of E-step and M-step (error made by α) has been chosen but a more complex stopping criterion must be implemented.