Latent Dirichlet allocation

Jérôme DOCKÈS (`jerome{at}dockes.org`)
Pascal LU (`pascal.lu{at}student.ecp.fr`)
École Normale Supérieure de Cachan

1 Latent Dirichlet allocation

1.1 Presentation of the model

We consider the problem of modeling text corpora. The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.

Notations

- $\mathcal{D} = \{d_1, d_2, \dots, d_M\}$ is a corpus (collection of $M = |\mathcal{D}|$ documents). We denote $|\mathcal{D}|$ (`num_docs`) the number of documents.
- $\mathcal{V}$ is the vocabulary. Its size is denoted V (`voc_size`).
- The number of topics is denoted k (`num_topics`).

For a document $d \in \mathcal{D}$,

- $d = (w_1^{(d)}, \dots, w_{N_d}^{(d)})$ represents the document d, where the $w_i^{(d)}$ are all distinct. N_d (`doc_size`) is the number of **distinct**[1] words in the document d.
- $w^{(d)}$ (`word_incidences`) is an matrix containing the number of times each word in the vocabulary appears in the document. The size of $w^{(d)}$ is the number of distinct words in document d × vocabulary size ($N_d \times V$).
- $\theta^{(d)}$ is an array of size k, representing a probability density.
- $z^{(d)}$ is the set of topics : $z_{ni}^{(d)} = 1$ if the word n is linked with the topic i. Hence, it is a matrix of size $N_d \times k$.

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The main idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

The following parameters are introduced:

- α (`dirich_param`) is an estimate of the parameter of the dirichlet distribution which generates the parameter for the (multinomial) probability distribution over topics in the document. The size of α is the number of topics, k. **We suppose that $\alpha = \alpha \mathbf{1}_k$[2] .**
- β (`word_prob_given_topic`) is a matrix of size (number of topics × vocabulary size $= k \times V$) which gives the (estimated) probability that a given topic will generate a certain word:
$$\beta_{ij} = p(w^j = 1 | z^i = 1)$$

[1] For implementation issues, this representation for a document is smaller than the representation proposed by [BNJ03].
[2] This assumption is suggested by the authors of [BNJ03].

Algorithm 1: Generative process

Data: corpus $\mathcal{D}$
begin
  for each document $d \in \mathcal{D}$ **do**
    Choose $N \sim \text{Poisson}(\xi)$;
    Choose $\theta^{(d)} \sim \text{Dir}(\alpha)$;
    for each of the N words $w_n^{(d)}$ **do**
      Choose a topic $z_n^{(d)} \sim \text{Multinomial}(\theta^{(d)})$;
      Choose a word w_n from $p(w_n | z_n^{(d)}, \beta)$, a multinomial probability conditioned on the topic $z_n^{(d)}$.

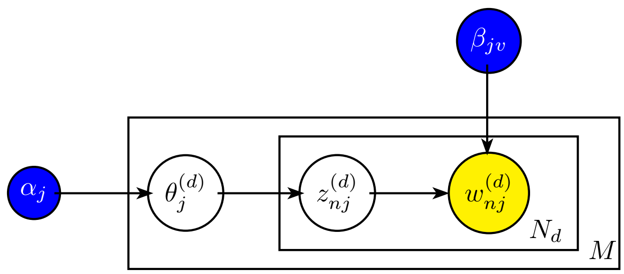


Figure 1: Generative model

LDA is based on the computation of the parameters (α, β), for instance by maximizing the log-likelihood. For a document d, the probability $p(d|\alpha, \beta)$ is given by:

$$
\begin{aligned}
p(d|\alpha,\beta) &= \int p(\theta^{(d)}|\alpha) \left(\prod_{n=1}^{N_d} p(w_n^{(d)}|\theta^{(d)}, \beta) \right) \mathrm{d}\theta \\
&= \int p(\theta^{(d)}|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_n^{(d)}} p(z_n|\theta^{(d)}) p(w_n^{(d)}|z_n^{(d)}, \beta) \right) \mathrm{d}\theta \\
&= \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} (\theta_i^{(d)})^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N_d} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i^{(d)} \beta_{ij})^{w_{nj}^{(d)}} \right) \mathrm{d}\theta
\end{aligned}
$$

1.2 Inference and parameter estimation

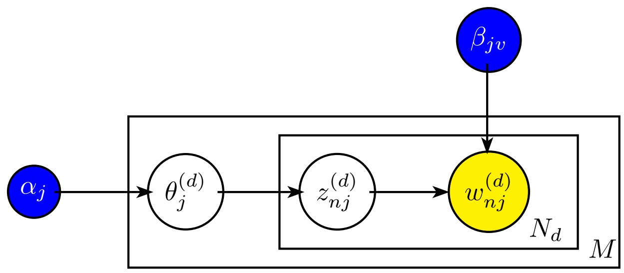
The basic idea of variational inference is to use Jensen's inequality to obtain an adjustable lower bound on the log likelihood.

For each document $d \in \mathcal{D}$, the following latent variables are introduced:

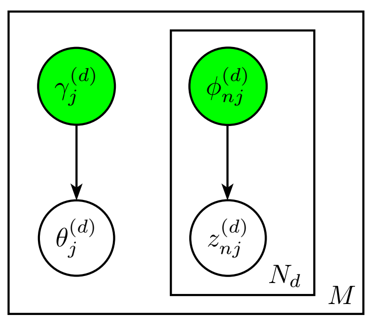
- $\gamma^{(d)}$ (`var_dirich`) the variational parameter for the dirichlet distribution. The size of $\gamma^{(d)}$ is the number of topics, k.
- $\phi^{(d)}$ (`var_multinom`) the variational parameter for the multinomial distribution The size of $\phi^{(d)}$ is (number of distinct words in document $d \times$ number of topics), $N_d \times k$. $\phi_{ni}^{(d)}$ depends on the relation between the word in position n of the document and the topic i of the list of topics.

and to try to estimate them instead of $\theta^{(d)}$ and $z_n^{(d)}$. The conditional probability is:

$$q(\theta^{(d)}, z^{(d)}|\gamma^{(d)}, \delta^{(d)}) = q(\theta^{(d)}|\gamma^{(d)}) \prod_{n=1}^{N_d} q(z_n^{(d)}|\phi_n^{(d)})$$



(a) Generative model



(b) Variational model

Figure 2: Graphical model

EM algorithm

Algorithm 2: EM algorithm

Data: Corpus $\mathcal{D}$ of documents, number of topics
Result: `dirich_param` (α), `word_prob_given_topic` (β)
begin
    for each $d \in \mathcal{D}$ **do**
        Compute `word_incidences` ($w^{(d)}$).
    Initialize `dirich_param` (α) with a uniform vector of size "number of topics";
    Initialize `word_prob_given_topic` (β) ;

    while *not converged* **do**
        for each $d \in \mathcal{D}$ **do**
            `var_dirich` ($\gamma^{(d)}$), `var_multinom` ($\phi^{(d)}$) = apply **E-step** to each document d given `word_incidences` ($w^{(d)}$), `dirich_param` (α), `word_prob_given_topic` (β)
        `dirich_param` (α), `word_prob_given_topic` (β) = Apply **M-step** given $\{$`word_incidences` ($w^{(d)}$), `var_dirich` ($\gamma^{(d)}$), `var_multinom` ($\phi^{(d)}$), $d \in \mathcal{D}\}$.

The expected log-likelihood, for a document d, is:

$$
\begin{aligned}
L(\gamma^{(d)}, \phi^{(d)}, \alpha, \beta) = \log \Gamma\left(k\alpha\right) - k \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{k} \left(\Psi(\gamma_i^{(d)}) - \Psi\left(\sum_{j=1}^{k} \gamma_j^{(d)} \right) \right) \\
+ \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{ni}^{(d)} \left(\Psi(\gamma_i^{(d)}) - \Psi\left(\sum_{j=1}^{k} \gamma_j^{(d)} \right) \right) \\
+ \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{ni}^{(d)} w_{nj}^{(d)} \log \beta_{ij} \\
- \log \Gamma\left(\sum_{j=1}^{k} \gamma_j^{(d)} \right) + \sum_{i=1}^{k} \log \Gamma(\gamma_i^{(d)}) - \sum_{i=1}^{k} (\gamma_i^{(d)} - 1) \left(\Psi(\gamma_i^{(d)}) - \Psi\left(\sum_{j=1}^{k} \gamma_j^{(d)} \right) \right) \\
- \sum_{n=1}^{N_d} \sum_{i=1}^{k} \phi_{ni}^{(d)} \log \phi_{ni}^{(d)}
\end{aligned}
$$

Algorithm 3: E-step for a document d (Variational Inference Procedure)

Data: `word_incidences` ($w^{(d)}$), `dirich_param` (α), `word_prob_given_topic` (β)
Result: `var_dirich` ($\gamma^{(d)}$), `var_multinom` ($\phi^{(d)}$)
begin
- Initialize $\phi_{ni}^{(d)} = \frac{1}{k}$ for all i and n;
- Initialize $\gamma_i^{(d)} = \alpha + \frac{N_d}{k}$ for all i ;
- **while** *not converged* **do**
  - **for** $n = 1 \dots N_d$ **do**
    - **for** $i = 1 \dots k$ **do**
      - $\phi_{ni}^{(d)} = \beta_{iw_n^{(d)}} \exp(\Psi(\gamma_i^{(d)}))$
    - normalize $\phi_n^{(d)}$ to sum to 1.
  - $\gamma^{(d)} = \alpha + \sum_{n=1}^{N_d} \phi_n^{(d)}$

Algorithm 4: M-step

Data: $\{$`word_incidences` ($w^{(d)}$), `var_dirich` ($\gamma^{(d)}$), `var_multinom` ($\phi^{(d)}$), $d \in \mathcal{D}\}$
Result: `dirich_param` (α), `word_prob_given_topic` (β)
begin
- $\beta \propto \sum_{d \in \mathcal{D}} (\phi^{(d)})^\top w^{(d)}$ (which corresponds to $\beta_{ij} \propto \sum_{d \in \mathcal{D}} \sum_{n=1}^{N_d} \phi_{ni}^{(d)} w_{nj}^{(d)}$)
- **while** *not converged* **do**
  - $\alpha \leftarrow \alpha - \frac{L'(\alpha)}{L''(\alpha)}$ where

$$
\begin{cases}
L'(\alpha) = |\mathcal{D}| k \left[\Psi\left(k\alpha\right) - \Psi(\alpha)\right] + \sum_{d \in \mathcal{D}} \left[\sum_{i=1}^{k} \Psi(\gamma_i^{(d)}) - \Psi\left(\sum_{j=1}^{k} \gamma_j^{(d)} \right) \right] \\
L''(\alpha) = |\mathcal{D}| k [k \Psi'(\alpha) - \Psi'\left(\alpha\right)]
\end{cases}
$$

2 Implementation and results

2.1 Implementation issues

2.2 Results

The database we used may be found at the address http://www.daviddlewis.com/resources/testcollections/reuters21578/.

3 Conclusion

References

[BNJ03] David M. Blei, Andrew Y. Ng and Michael I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.