Latent Dirichlet Allocation

Jérôme DOCKÈS (jerome{at}dockes.org), Pascal LU (pascal.lu{at}centraliens.net)

École Normale Supérieure de Cachan — December 30, 2015

Objectives

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 [1].

Notations

- $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ is a corpus.
- $\mathcal{V}$ is the vocabulary of size V.
- k is the number of topics.

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

- $d = (w_1^{(d)}, \ldots, w_{N_d}^{(d)})$ represents the document d, where the $w_i^{(d)}$ are all distinct. N_d is the number of **distinct** words in the document d.
- $w^{(d)}$ (`word_incidences`) is a matrix containing the number of times each word in the vocabulary appears in the document. Size of $w^{(d)} = N_d \times V$.
- $\theta^{(d)}$ 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. Size of $z^{(d)} = N_d \times k$.

Presentation of the model

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus.

- Documents = random mixtures over latent topics,
- Topic = a distribution over words.

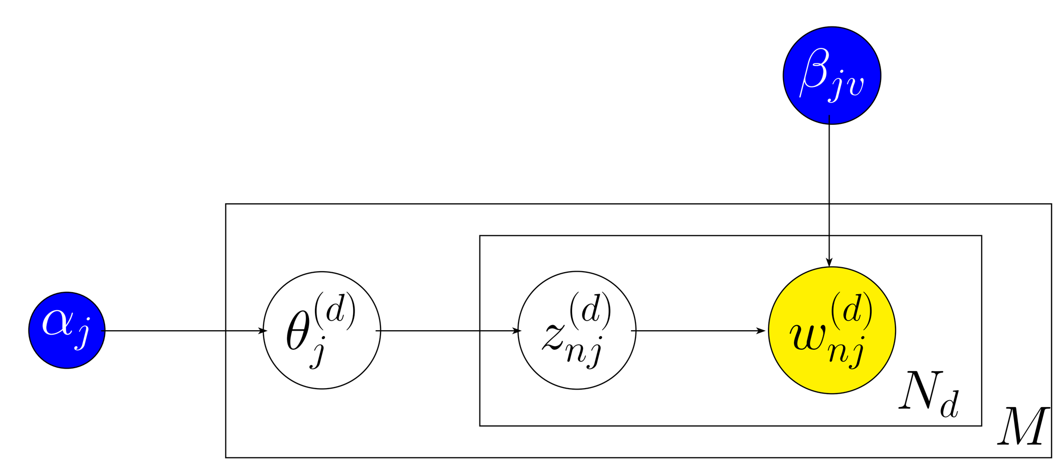
- **Input**: corpus $\mathcal{D}$

For *each document* $d \in \mathcal{D}$
  Choose $N \sim \text{Poisson}(\xi)$
  Choose $\theta^{(d)} \sim \text{Dir}(\alpha)$
  For *each of the* N *words* $w_n^{(d)}$
    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 $z_n^{(d)}$.

Generative model

The **goal** is to determine:

- α = estimate of the parameter of the Dirichlet distribution which generates the parameter for the (multinomial) probability distribution over topics in the document. Size of $\alpha = k$.
- β is a matrix of size $k \times V$ which gives the estimated probability that a given topic will generate a certain word: $\beta_{ij} = p(w^j = 1 | z^i = 1)$.

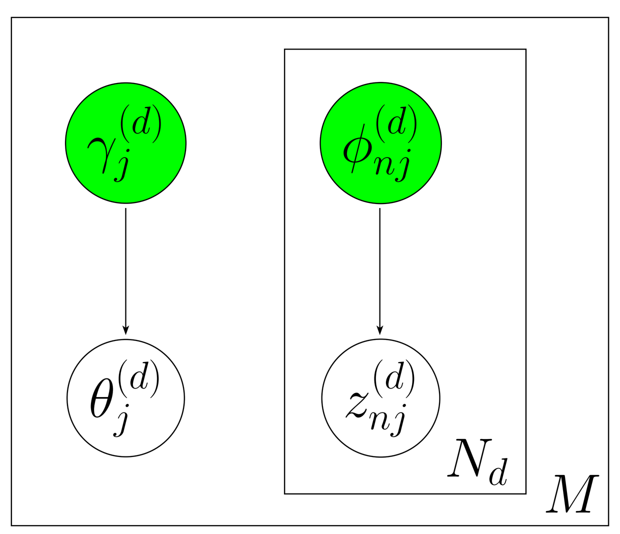


Variational inference

⇒ Use Jensen's inequality to obtain a lower bound on the log likelihood.
For a document $d \in \mathcal{D}$:

- $\gamma^{(d)}$ is the variational parameter for the dirichlet distribution. Size of $\gamma^{(d)} = k$.
- $\phi^{(d)}$ is the variational parameter for the multinomial distribution. Size of $\phi^{(d)} = N_d \times k$. $\phi_{ni}^{(d)}$ depends on the relation between the word in position n of the document and the topic i of the list of topics.

⇒ Estimate $\gamma^{(d)}, \phi_n^{(d)}$ instead of $\theta^{(d)}$ and $z_n^{(d)}$.



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

- **Input**: a document d defined by its `word_incidences` $(w^{(d)})$, α, β
- **Output**: $\gamma^{(d)}$, $\phi^{(d)}$

Initialize $\phi_{ni}^{(d)} = \frac{1}{k}$ for all i and n.
Initialize $\gamma_i^{(d)} = \alpha + \frac{1}{k}\sum_{n=1}^{N_d} w_n^{(d)}$ for all i.
While *the expected log-likelihood* $L(\gamma^{(d)}, \phi^{(d)}, \alpha, \beta)$ *for the document* d *has not converged*
  For $n = 1 \ldots N_d$
    For $i = 1 \ldots k$
      $\phi_{ni}^{(d)} = \beta_{iw_n^{(d)}} \exp(\Psi(\gamma_i^{(d)}))$
  Normalize $\phi_n^{(d)}$ to sum to 1.
  $\gamma^{(d)} = \alpha + \sum_{n=1}^{N_d} w_n^{(d)} \phi_n^{(d)}$
  Update $L(\gamma^{(d)}, \phi^{(d)}, \alpha, \beta)$

EM-algorithm

- **Input**: Corpus $\mathcal{D}$, number of topics k
- **Output**: α, β

For each $d \in \mathcal{D}$, compute $w^{(d)}$ (`word_incidences`).
Initialize α, β and $\Sigma_\gamma = 0$.
While *the expected log-likelihood* $L(\alpha, \beta)$ *has not converged*:
  For *each* $d \in \mathcal{D}$
    $(\gamma^{(d)}, \phi^{(d)}) = \textbf{E-step}(w^{(d)}, \alpha, \beta)$
    Update $\beta \leftarrow \beta + (\phi^{(d)})^\top w^{(d)}$
    Update $\Sigma_\gamma \leftarrow \Sigma_\gamma + \sum_{i=1}^{k} \Psi(\gamma_i^{(d)}) - \Psi\left(\sum_{j=1}^{k} \gamma_j^{(d)}\right)$
    Update $L(\alpha, \beta) \leftarrow L(\alpha, \beta) + L(\gamma^{(d)}, \phi^{(d)}, \alpha, \beta)$
  Normalize β
  While α *has not converged*
    $\alpha \leftarrow \alpha - \frac{L'(\alpha)}{L''(\alpha)}$ where

$$\begin{cases} L'(\alpha) = |\mathcal{D}| k \left[\Psi(k\alpha) - \Psi(\alpha)\right] + \Sigma_\gamma \\ L''(\alpha) = |\mathcal{D}| k [k \Psi'(k\alpha) - \Psi'(\alpha)] \end{cases}$$

Implementation issues

- Preprocessing of the documents.
- Initialization of α et β.
- Optimization of the parameters.

Results

Tested on real data from the Reuters21578 database (`reut2-000.sgm`).

Topic 1	Topic 2	Topic 3	Topic 4
devices	prolonged	zestril	features
disk	council	annesthetic	shipping
megabyte	forum	hypertension	798
expandable	dissident	oth	998
megabytes	flying	statil	sells
equipped	sparks	diabetic	AppleWorld
monochrome	talks	complications	Conference
peripheral	outweighed	Barbara	899
color	accomplishments	definitive	science

Table 1: Results for 4 topics ($k = 10$)

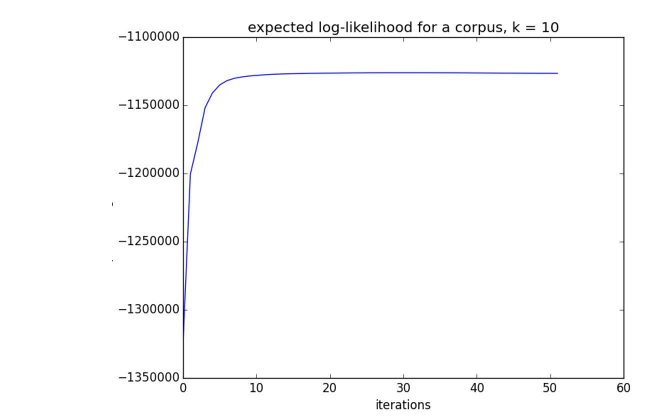


Figure 1: Expected log-likelihood for the corpus ($k = 10$)

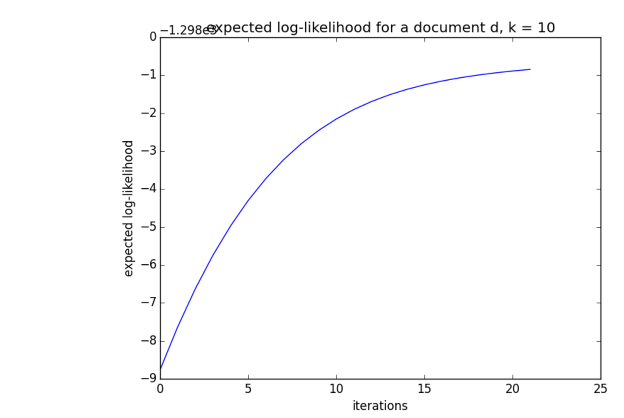


Figure 2: Expected log-likelihood for a document ($k = 10$)

Conclusion

LDA = a way to apply graphical models to information retrieval = group words in the same categories.
Key idea = variational inference.
Other interesting applications: biology (DNA sequence), content-based image retrieval...

References

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