Probabilisitic Graphical Models IFT6269 Project
Latent Dirichlet Allocation

Patrice Béchard [*1] **Théophile Gervet** [*2]

Abstract

TODO

1. Introduction and Related Works

1.1. Latent Dirichlet Allocation

1.1.1. Generative Process

Latent Dirichlet allocation (LDA) is a generative probabilisitic model of a corpus of D documents $\{\mathbf{w}_1, ..., \mathbf{w}_D\}$. Each document $\mathbf{w}_d = (w_{d,1}, ..., w_{d,N_d})$ is a sequence of N_d words from a vocabulary indexed by $\{1, ..., V\}$.

The basic idea behind LDA is that each document is represented as a random mixture over K latent topics, where each topic is characterized by a multinomial distribution over words, parametrized by a vector $\boldsymbol{\beta}_k$, $k \in 1, ..., K$.

Formally, LDA assumes the following generative process for each document $\mathbf{w}_d$:

1. Sample a distribution over topics $\boldsymbol{\theta_d} \sim Dir(\boldsymbol{\alpha})$
2. For each of the N_d words $w_{d,n}$ independently:
   (a) Draw a topic from the distribution over topics $z_{d,n} \sim Mult(\boldsymbol{\theta}_d)$
   (b) Draw a word from this topic $w_{d,n} \sim Mult(\boldsymbol{\beta}_{z_{d,n}})$

This generative process is illustrated in the directed graphical model of figure 1.

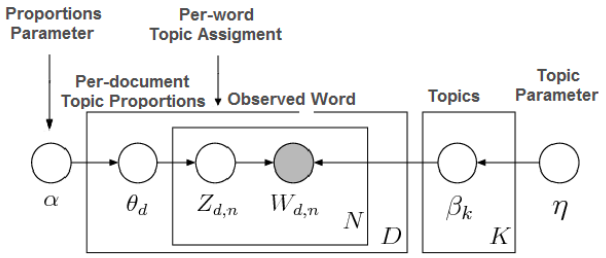

Figure 1. Graphical model corresponding to the generative process for LDA.

The joint distribution associated with this model is:

$$\prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\boldsymbol{\eta}) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_{n=1}^{N_d} p(z_{d,n}|\boldsymbol{\theta}_d) p(w_{d,n}|z_{d,n}, \boldsymbol{\beta}_{1:K}) \quad (1)$$

where
$z_{d,n} \sim Mult(\boldsymbol{\theta}_d)$, i.e $p(z_{d,n}|\boldsymbol{\theta}_d) = \theta_{d,z_{d,n}}$
$w_{d,n} \sim Mult(\boldsymbol{\beta}_{z_{d,n}})$, i.e $p(w_{d,n}|z_{d,n}, \boldsymbol{\beta}_{1:K}) = \beta_{z_{d,n},w_{d,n}}$
$\boldsymbol{\beta}_k \sim Dir(\boldsymbol{\eta})$
$\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$

This generative process is just a model of the structure we assume to be in our data. In reality, we only observe documents $\mathbf{w}_{1:D}$ and our goal is to infer the underlying topic structure: the topics $\boldsymbol{\beta}_{1:K}$, the per document distributions over topics $\boldsymbol{\theta}_d$, and per document per word topic assignments $z_{d,n}$.

Since the generative process assumes independence between documents, the probability of the whole corpus decomposes as a product of terms for individual documents. In order to clarify the notation, from now on let us consider equations for a single document $\mathbf{w}$ with N words.

Assuming the Dirichlet hyper-parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are

[*]Equal contribution [1]Université de Montréal, Montréal, Canada [2]McGill University, Montréal, Canada. Correspondence to: Patrice Béchard <patrice.bechard@umontreal.ca>, Théophile Gervet <theophile.gervet@umontreal.ca>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

fixed, we want to infer the posterior distribution:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})} \tag{2}$$

Unfortunately, this distribution is intractable to compute because the normalization factor $p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})$ cannot be computed exactly. We must use approximate inference to estimate the posterior over latent variables $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\mathbf{z}$. We will explore both variational inference and Gibbs sampling to solve this problem.

1.2. Related Works

TODO

2. Methods

2.1. Variational EM

To make our life easier, we derive a variational EM algorithm for a slightly simpler graphical model for LDA (the one presented in the original paper). We remove the Dirichlet prior over topics, this corresponds to removing the dependence of the $\boldsymbol{\beta}$ plate on the $\boldsymbol{\eta}$ random variable (see Figure 2, left). By doing so, we lose our ability to smooth topics through the $\boldsymbol{\eta}$ parameter.

Assuming the $\boldsymbol{\alpha}$ parameter and topics $\boldsymbol{\beta}$ are fixed, the posterior for a single document $\mathbf{w}$ now takes the form:

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})} \tag{3}$$

This expression is still intractable because of the denominator $p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Let us motivate a variational EM algorithm by looking at how this intractable posterior comes back to bite us when we try to maximize the likelihood of the observed data. Since the probability of the whole corpus decomposes as a product of terms for individual documents, the log likelihood of the whole corpus decomposes as a sum of terms for individual documents. The log likelihood of a single document $\mathbf{w}$ takes the form:

$$logp(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = log \int \sum_{\mathbf{z}} p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} \tag{4}$$

Because of the integral over $\boldsymbol{\theta}$ and the sum over $\mathbf{z}$ inside the log, we cannot maximize this objective directly. One common way to solve this problem is the Expectation Maximization (EM) algorithm. We make the following observation:

$$logp(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = log \int \sum_{\mathbf{z}} \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) q(\boldsymbol{\theta}, \mathbf{z})}{q(\boldsymbol{\theta}, \mathbf{z})} d\boldsymbol{\theta} \tag{5}$$

$$= \mathbf{E}_q[log \int \sum_{\mathbf{z}} \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{q(\boldsymbol{\theta}, \mathbf{z})} d\boldsymbol{\theta}] \tag{6}$$

$$\geq \mathbf{E}_q[logp(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})] + \mathbf{E}_q[logq(\boldsymbol{\theta}, \mathbf{z})] \tag{7}$$

$$= \mathcal{L}(q, \boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{8}$$

where we have introduced an arbitrary distribution $q(\boldsymbol{\theta}, \mathbf{z})$ over the problematic latent variables $\boldsymbol{\theta}$ and $\mathbf{z}$, and the last step follows from Jensen's inequality.

The EM algorithm consists in alternatively maximizing the lower bound $\mathcal{L}$ on the log likelihood with respect to the distribution q (the E-step), and with respect to parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (the M-step).

One can easily show that maximizing $\mathcal{L}$ with respect to q is equivalent to minimizing $KL(q(\boldsymbol{\theta}, \mathbf{z})||p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}))$. A simple solution is thus to set $q(\boldsymbol{\theta}, \mathbf{z})$ to be $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, the true posterior over latent variables under the model.

In our case, things are not so simple: as we have seen, the posterior $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is intractable. This is were variational inference comes into play. The basic idea is to introduce a parametrized family of distributions over latent variables and phrase inference as an optimization problem.

We parametrize q as follows, according to the mean field independence assumption:

$$q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{n=1}^{N} q(z_n|\boldsymbol{\phi}_n) \tag{9}$$

where the dirichlet parameters $\boldsymbol{\gamma}$ and the multinomial parameters $\boldsymbol{\phi}_n$, $n \in \{1, ..., N\}$ are free variational parameters. The graphical model corresponding to this parameterization is illustrated in Figure 2, right.

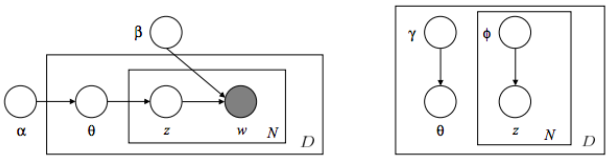

Figure 2. Graphical model corresponding to the generative process for LDA.

Now variational inference consists in solving the following

optimization problem:

$$(\boldsymbol{\gamma}^*, \boldsymbol{\phi}^*) = \arg\min_{\boldsymbol{\gamma},\boldsymbol{\phi}} KL(q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})) \tag{10}$$

Note that $q(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\gamma}^*, \boldsymbol{\phi}^*)$ is actually a conditional distribution varying as a function of $\mathbf{w}$ because the optimization objective depends on $\mathbf{w}$. It can thus be seen as an approximation to the posterior $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

With this tool at our disposal, we can come back and complete the EM algorithm described in section 1.1.2. We can expand our variational lower bound on the per-document log likelihood as follows:

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{E}_q[logp(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \mathbf{E}_q[logp(\mathbf{z}|\boldsymbol{\theta})] + \mathbf{E}_q[logp(\mathbf{w}|\mathbf{z}, \boldsymbol{\beta})] - \mathbf{E}_q[logq(\boldsymbol{\theta}|\boldsymbol{\gamma})] - \mathbf{E}_q[logq(\mathbf{z}|\boldsymbol{\phi})] \tag{11}$$

where the first three terms correspond to the expected complete log likelihood under q and the last two terms to the entropy of q.

2.1.1. E-STEP

In the E-step, we maximize $\mathcal{L}$ with respect to the variational parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$. This corresponds to applying the following update rules iteratively until convergence for each document:

$$\boldsymbol{\phi}_{n,k} \propto \boldsymbol{\beta}_{k,w_n} exp\{\Psi(\boldsymbol{\gamma}_k)\} \tag{12}$$

$$f\boldsymbol{\gamma}_k = \boldsymbol{\alpha}_k + \sum_{n=1}^{N} \boldsymbol{\phi}_{n,k} \tag{13}$$

where $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function.

Theses updates make sense intuitively. Equation 12 tells us that $\boldsymbol{\phi}_{n,k}$, the probability that word w_n is generated by latent topic k, is proportional to the probability of drawing w_n from topic k times a quantity proportional to how much the document likes topic k (the dirichlet parameter $\boldsymbol{\gamma}_k$). Equation 13 tells us that the dirichlet parameter $\boldsymbol{\gamma}_k$ is equal to the pseudo-count $\boldsymbol{\alpha}_k$ given by the dirichlet prior plus the expected count of words picking topic k.

2.1.2. M-STEP

For the M-step, we maximize the overall variational lower bound with respect to parameter $\boldsymbol{\beta}$. We could also maximize with respect to $\boldsymbol{\alpha}$ but we chose to keep it fixed to make things simpler. As said before, the overall log likelihood of the corpus is the sum of the log likelihoods for the individual documents. Likewise, the overall variational lower bound is the sum of the individual variational bounds.

Algorithm 1 Variational EM

Input: documents $\mathbf{w}_1, ..., \mathbf{w}_D$
Set $\boldsymbol{\alpha}_k = 1/K$ for all k
Randomly initialize variational parameters $\boldsymbol{\gamma}$, $\boldsymbol{\phi}$
repeat
  for $d = 1$ **to** D **do**
    for $k = 1$ **to** K **do**
      for $n = 1$ **to** N_d **do**
        $\boldsymbol{\phi}_{d,n,k} = \boldsymbol{\beta}_{k,w_n} exp\{\Psi(\boldsymbol{\gamma}_{d,k})\}$
      end for
      normalize $\boldsymbol{\phi}_{d,n}$ to sum to 1
      $\boldsymbol{\gamma}_{d,k} = \boldsymbol{\alpha}_k + \sum_{n=1}^{N_d} \boldsymbol{\phi}_{d,n,k}$
    end for
  end for
  for $k = 1$ **to** K **do**
    for $j = 1$ **to** V **do**
      $\boldsymbol{\beta}_{k,j} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \boldsymbol{\phi}_{d,n,k} w_{d,n}^{(j)}$
    end for
    normalize $\boldsymbol{\beta}_k$ to sum to 1
  end for

until objective $\mathcal{L}$ doesn't improve anymore

Maximizing with respect to $\boldsymbol{\beta}$ gives the following update:

$$\boldsymbol{\beta}_{k,j} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \boldsymbol{\phi}_{d,n,k} w_{d,n}^{(j)} \tag{14}$$

where we abuse notation and index word $w_{d,n}$ which was a scalar until here as a one hot vector.

Equation 14 also has an intuitive meaning: the probability of word j under topic k is proportional to the expected count under the variational distribution of the number of times word j is assigned topic k across the whole corpus.

2.1.3. OVERALL ALGORITHM

We implemented a variational "EM flavored" algorithm which we described in detail in the algorithm below. Note that to be strictly an EM algorithm, we should maximize $\mathcal{L}$ with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ all the way at each E-step. Instead, we alternate a single step of improvement with respect to the variational parameters with one step of improvement with respect to $\boldsymbol{\beta}$. Our algorithm has the same guarantee to increase the variational lower bound at each iteration, it is just more convenient to check for convergence only for the overall algorithm instead of for each E-step.

2.2. Gibbs Sampling

TODO

3. Experiments

TODO

4. Results

TODO