Probabilistic Graphical Models, Master MVA, 2017/2018

Latent Dirichlet Allocation: Final Report

Karl Hajjar, Léonard Hussenot-Desenonges, Charles Maussion

firstname.name@polytechnique.edu

1 Presentation of the model

1.1 Introduction

The Latent Dirichlet Allocation (LDA) is a generative probabilistic model whose purpose is to automatically extract content and information out of unstructured and unannotated textual data. This model is used to extract topics or themes out of a corpus of documents in an unsupervised setting : that is, to be able to know what a corpus of documents is talking about without previous annotation. This can then be used for document classification, summarization, annotating a new arriving document...

1.2 The generative model

The LDA gives a mathematical model (which is, of course not realistic, it is just a model) to understand how a corpus of documents is generated. The setting we consider is the following:

- A corpus $\mathcal{D}$ of documents is a set of $\{d_1, d_2, ..., d_D\}$ documents where $D = |\mathcal{D}|$ is the number of documents
- Each document d is itself a set of words (thus unordered, which is a pretty strong assumption) $\{w_1, ..., w_{N_d}\}$ where N_d is the number of words in document d.
- Words are taken from a common vocabulary for the corpus (or even for different corpora) $\mathcal{V} = \{w_1, ..., w_V\}$ where $V = |\mathcal{V}|$ is the number of words in the vocabulary. In practice, each word is associated with its index in the vocabulary or even with the corresponding one-hot encoding, so that, for $i \in \llbracket 1, V \rrbracket$, a word $w \in \mathcal{V}$ is the i^{th} world in the vocabulary iff $w = i$, or $w = (\mathbb{1}_{\{k=i\}})_{k \in \llbracket 1, V \rrbracket}$ (we will use both representations in the computations of part 2).

The model proposed for generating documents is the following : we suppose there are K different topics in the corpus (K is actually a hyperparameter of the model that is fixed in advance). Given that, we generate the documents as follows :

(i) choose a distribution of words for each topic, that is to say a matrix $\beta = (\beta_{k,w})_{(k,w) \in \llbracket 1, K \rrbracket \times \llbracket 1, V \rrbracket}$, where, for each $k \in \llbracket 1, K \rrbracket$, $\beta_{k,w}$ is the probability of word w in topic k (remember that w is not actually the word itself but simply its index in the vocabulary)

(ii) for each document d in $\mathcal{D}$, choose a distribution of topics (i.e topic proportions in document d) $\theta_d = (\theta_{d,1}, ..., \theta_{d,K})$ in document d

(iii) for each word in document d,

first choose a topic k for this word by sampling from the discrete distribution $(\theta_{d,1}, ..., \theta_{d,K})$

then choose a word w from the distribution of words in the given topic $(\beta_{k,1}, ..., \beta_{k,V})$

1.3 Objective of the model

In the model, there are two parameters and one latent variable that we wish to estimate : the distributions θ and β and the latent variable z giving the topics assigned to each word of each document.

The variables we consider in the model are :

- the list of all words observed $w = (w_1, ..., w_{N_d})_{d \in \mathcal{D}}$
- the latent variable $z = (z_1, ..., z_{N_d})_{d \in \mathcal{D}}$ representing the topics (indexed by $[\![1,K]\!]$) assigned to all the words in the corpus

Finally, an additional assumption to the model is that we have prior information on how the distribution over topics is generated : namely from a **Dirichlet** law of parameter α (which is thus a new hyperparameter of the model). The choice of a Dirichlet distribution is not an accident, and results from the fact that the Dirichlet distribution is conjugate to the multinomial distribution, and thus makes the calculations a lot simpler, i.e. the posterior distribution on the latent variable has the same form as the prior distribution but with different parameter values. Doing this thus yields three variables for the models, two of which are **latent** or **unobserved**, namely z and θ, and one observed variable w, and two parameters α and β.

The resulting model is considered as (partially) **bayesian** as it considers the parameter θ as a random variable whose distribution is to be estimated. In that sense, we will not try to learn a single value for the latent variables θ and z but rather a distribution over the different possible values taken by those variables. In that respect, one of the goals of the LDA can be seen as to infer the posterior $p(\theta, z|w, \alpha, \beta)$.

Unfortunately, this posterior probability is intractable because, when using Bayes' rule via the classical formulation $posterior = \frac{prior \times likelihood}{evidence}$ the denominator $p(w|\alpha, \beta)$ is itself intractable because of the marginalization over the latent variables.

Thus we are reduced to using approximate inference techniques such as **MCMC** or **variational inference** (or even Gibbs Sampling) to approximate this posterior conditional distribution, as well as the **Expectation-Maximization** algorithm to estimate the parameters.

With the variational inference procedure, this actually gives a lower bound on the true log-likelihood of the posterior, depending only on the value of α, β and the optimal variational parameters learned, which can then be maximized with respect to α and β, thus elevating the value of $p(\theta, z|w, \alpha, \beta)$.

2 Inference and Parameter Estimation

2.1 Inference

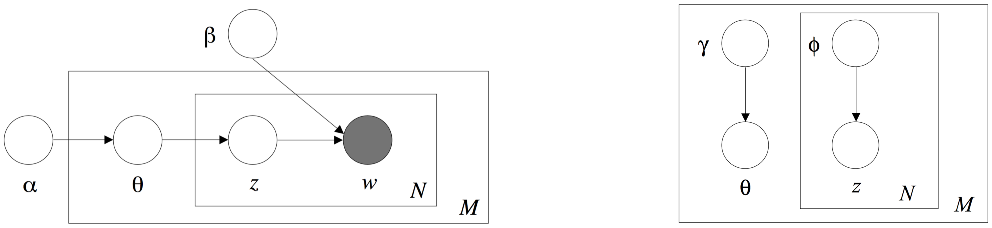
As we have seen in the previous part, the posterior probability is intractable to compute in general. Indeed, the normalization term is coupled in θ and β and thus impossible to compute :

$$p(w|\alpha,\beta) = \frac{\Gamma(\Sigma_i \alpha_i)}{\Pi_i \Gamma(\alpha_i)} \int \Big(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \Big) \Big(\prod_{n=1}^{N} \sum_{k=1}^{K} \prod_{j=1}^{V} \theta_{nk} \beta_{kj}^{w_n^{(j)}} \Big) d\theta$$

That's why we will use **variational inference** : thanks to Jensen's inequality we will consider a family of lower bounds on the log likelihood, indexed by a set of variational parameters. A way to obtain a tractable family of such lower bounds is to slightly modify the graphical model by dropping some edges or nodes that create intractable couplings, like between θ and β. We thus obtain the variational model in Figure 1. This family, described by the two variational parameters γ and ϕ, is characterized by the following distribution making θ and z independent (decoupled) :

$$q(\theta, z|\gamma,\phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n)$$

where $q(\theta|\gamma)$ is a Dirichlet distribution of parameter γ and $q(z_n|\phi_n)$ are multinomial distributions of parameters $(\phi_1, ..., \phi_N)$. γ and $\phi = (\phi_1, ..., \phi_N)$ are thus the free variational parameters we want to estimate.



(a) Initial model

(b) Variational model approximating posterior

Figure 1: LDA and Variational Graphical Models

The goal here is thus to find the distribution q (and in fact only the parameters γ and ϕ) which minimizes the Kullback-Leibler divergence $D(q(\theta, z|\gamma,\phi) \mid\mid p(\theta, z|w, \alpha, \beta))$:

$$(\gamma^*, \phi^*) = \text{argmin}_{\gamma,\phi} D(q(\theta, z|\gamma,\phi)||p(\theta, z|w,\alpha,\beta))$$

Given the formula below :

$$\log p(w|\alpha,\beta) = L(\gamma,\phi;\alpha,\beta) + D(q(\theta,z|\gamma,\phi)||p(\theta,z|w,\alpha,\beta))$$

we can simply maximize the lower bound L on the true posterior likelihood $\log p(\theta, z|w, \alpha, \beta)$.

By computing the derivatives of L (actually of the corresponding Lagrangian since ϕ_n has to sum to 1 for every n) and setting them equal to zero, we obtain the following pair of update equations:

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$

$$\phi_{ni} \propto \beta_{iwn} \exp\{\mathbb{E}_q[log(\theta_i)|\gamma]\}$$

We summarize the variational inference procedure in the following algorithm, with appropriate starting points for γ and ϕ_n.

```
initialize $\forall i$ and $n, \phi_{ni}^0 = \frac{1}{n}$;
initialize $\forall i, \gamma_i^0 = \alpha_i + \frac{N}{K}$;
while not converged do
    for $n = 1$ to $N$ do
        for $i = 1$ to $K$ do
            $\phi_{ni}^{t+1} = \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
        end
    end
    $\gamma^{t+1} = \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
end
```

Algorithm 1: Variational inference algorithm

2.2 Parameter estimation

Now that we have seen how to minimize the Kullback-Leibler divergence between the true posterior and the variational distribution, or equivalently getting a maximal lower bound on the true posterior using a variational distribution decoupling θ and z, we can explain how we can learn values of the parameters α and β making the true posterior likelihood $\log p(\theta, z|w, \alpha, \beta)$ go up via an **EM** algorithm.

Let us thus recap what we have so far : given values of α and β, at the end of the variational inference procedure we are left with values γ^* and ϕ^* which maximize a lower bound $L(\gamma, \phi; \alpha, \beta)$ (L implicitly depends on the observed data w) on the log of the true posterior for a given document d:

$$\forall \gamma, \phi : \log p(\theta, z|w, \alpha, \beta) \geq L(\gamma^*, \phi^*; \alpha, \beta) \geq L(\gamma, \phi; \alpha, \beta)$$

Since the probability of the whole corpus decomposes into the product of the probability of all the documents, a lower bound on the complete log-likelihood of the true posterior is just the sum of all of these optimal lower bounds for all documents, which we can get (learn) independently.

This maximization procedure (maximizing all the lower bounds independently) is actually the **E-step** of our EM procedure since the lower bound comes from an expectation under the variational distribution which we then maximize with respect to γ and ϕ.

As for the **M-step**, we maximize $L(\gamma^*, \phi^*; \alpha, \beta)$ with respect to α and β, hoping that, if our lower bound is tight enough, this will make the value of $p(\theta, z|w, \alpha, \beta)$ go up. Thus, the M-step yields optimal values α^* and β^* for the parameters, and along with them, a new true posterior distribution $p(\theta, z|w, \alpha^*, \beta^*)$, which we can again try to estimate using the variational inference in the E-step with the new values learned of α and β, and thus repeat the EM procedure.

This EM algorithm alternatively maximizes L with respect to (γ, ϕ) and then (α, β) and can thus be seen as a coordinate ascent algorithm on L.

2.3 Taking the fully bayesian approach

We now consider β not to be a hyperparameter but a **latent** (random) variable we wish to estimate. We treat β as a random variable where we assume that each row is independently drawn from a Dirichlet distribution of parameter η_i, as shown in Figure 2

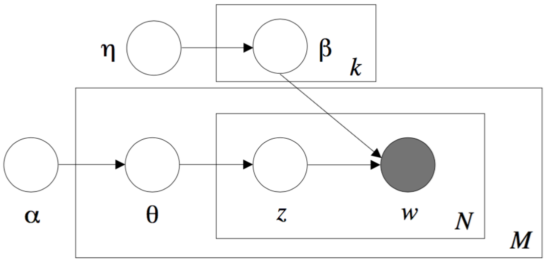


Figure 2: New graphical model of the fully bayesian approach

The variational approach leads us to a distribution q of the form:

$$q(\beta, \theta, z|\lambda, \gamma, \phi) = \prod_{i=1}^{k} Dirichlet(\beta_i|\lambda_i) \prod_{d=1}^{M} q_d(\theta_d, z_d|\gamma_d, \phi_d)$$

Where q_d is the variational distribution computed in the previous part and λ is the new variational parameter associated to β. In that setting, among our 3 free variational parameters, γ and ϕ are updated the same way they were in the previous part (because of the decoupling between β, θ and z under q), and the additional one, λ is updated as :

$$\lambda_{ij} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^{(j)}_{dn}$$

This formula comes when setting to zero the derivative of the Lagrangian associated to the optimization problem of maximizing the lower bound L with respect to λ.

We can then include this in the previous E-step, and compute a new corresponding M-step which is $\max_{\alpha,\eta} L(\lambda^*, \gamma^*, \phi^*; \alpha, \eta)$.

2.4 Applications

After learning those parameters, we can retrieve the top words from some of the resulting multinomial distributions $p(w|z)$ as illustrated in Figure 3 (top). These distributions seem to capture some of the underlying topics in the corpus (that are then manually named after these topics). We can also look at the distributions $p(z_n|w)$, that are approximated by our ϕ_n parameters. These distributions generally peak towards one of the k possible topic values. In Figure 3, the words are color coded according to these values (i.e., the i_{th} color is used if $q_n(z_{in} = 1) > 0.9$).

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 3: Inference on the text from a corpus