Topic Modeling : Latent Dirichlet Allocation with VI

Caterina De Bacco and Isabel Valera

1 Topic modeling: the idea

Consider a collection of text corpora, for instance a set of documents.
Goal: find statistical pattern behind the data, so that we can parametrize the members of the collection by a short description. In other words, we want to find a low dimensional representation of the data.

Idea: documents are mixture of K latent variables called *topics*, and *topics* are mixtures of words. Formally, we have 3 types of latent variables:

- β_k: a distribution of words, needed to specify the topic;
- θ_d: a vector of topic proportion, needed to specify a document;
- z_{dn}: a topic assignment, needed to specify what words are seen in each document.

The data are words w, divided into documents.

The goal is to then estimate the posterior $p(\beta, \theta, z|w)$. This can then be used to perform various tasks, like classification, novelty detection, similarity or relevance judgement.

"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 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Figure 1: Example of Topic Modeling. Colors represent topics, and on top we have the words' mixture denoting each topic. Figure taken from Blei *et al.* (2003).

In Figure 1 you see an example of the results that we will obtain.

2 Topic modeling: the LDA model

We will follow the Latent Dirichlet Allocation (LDA) model from Blei *et al.* (2003), which is a *conditionally conjugate* topic model.

Formally, we define:

- *words*: w, basic unit of discrete data, where V is the length of the vocabulary;
- *documents*: $\mathbf{w} = (w_1, \dots, w_N)$ is a sequence of word assignments, where N is the length of the document;
- *corpus*: $D = (\mathbf{w}_1, \dots, \mathbf{w}_M)$ is a collection of documents, where M is the length of the corpus.

Words are represented as indicator vectors:

- $w_{dn}^v = 1$ indicates that the n-th word of document d is the v-th word of the vocabulary.

Idea: documents are random mixtures of K *topics*, and topics are a random mixture of words.

Therefore, we introduce latent variables:

- z_{dn}: is an indicator K-vector, *topic assignment*, where $z_{dn}^k = 1$ indicates that n-th word of document d is assigned to the k-th topic.
- θ_d is a K-vector of *topic proportion*, there is one for each document d.
- β_k is a V-vector of *word proportion*, there is one for each topic k.

2.1 LDA generative model

LDA is based on the following generative process:

1. For each topic $k = 1, \dots, K$ draw a distribution over words:
$$\beta_k \sim \text{Dir}_V(\eta) \quad . \tag{1}$$
2. For each document $d = 1, \dots, D$:
   (a) draw a vector of topic proportions:
   $$\theta_d \sim \text{Dir}_K(\alpha) \quad . \tag{2}$$
   (b) For each word in $n = 1, \dots, N$:
      (i) draw a topic assignment:
      $$z_{dn} \sim \text{Mult}(\theta_d) \quad , \tag{3}$$
      (ii) draw a word:
      $$w_{dn} \sim \text{Mult}(\beta_{z_{dn}}) \quad . \tag{4}$$

We have the hyper-parameters $\eta \in \mathbb{R}_{>0}$ and $\alpha \in \mathbb{R}_{>0}^K$ for the prior over the two Dirichlet distributions.[1] Similarly to the GMM example, it is convenient to consider categorical variables as indicator vectors, i.e. $z_{dn}^k = 1$ if word n of document d belongs to topic k; $w_{dn}^v = 1$ if word n of document d is the v-th element of the vocabulary. Sometimes we can abuse the notation using $\beta_{z_{dn}}$ when $z_{dn}^k = 1$ means that we consider the k-th topic, i.e. β_k.

[1] Recall that a k-dimensional Dirichlet distributed variable can take values over the $(k-1)$ simplex, i.e. a k-vector θ satisfies the constraint $\sum_{i=1}^k \theta_i = 1$. The Dirichlet distribution is a convenient distribution for such variables, it is a member of the exponential family and it is the conjugate prior of the Multinomial; it has the following probability density:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_i \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1} \tag{5}$$

.

Obs1: notice that the same set of topics $\{\beta_1, \dots, \beta_K\}$ is used for all the documents, but each d has a different proportion θ_d. This an example of a *mixed membership* model, θ_d can be seen as a membership to various topics.
Obs2: recall that the standard SBM model from Lecture 12 was instead not a mixed membership, because each node i could belong to only one group at a time, i.e. *hard* membership.

Objective: perform *posterior inference* . We aim at finding topics β and documents' memberships θ_d. This allows to analyze the data, to browse it, understand, classify, etc...

Obs3: the fact that topics have an interpretation, e.g. "sport", "politics" is something that comes out *a posterirori*, this might not be the case in general.

With all these ingredients, we have the posterior (omitting explicit dependence on the hyper-parameters):

$$\begin{aligned} p(\beta, \theta, z|w) &\propto p(w|\beta, \theta, z)\, p(z|\theta)\, p(\theta)\, p(\beta) && (6) \\ &= \prod_{d,n} \text{Mult}(w_{dn}; \beta_{z_{dn}}) \text{Mult}(z_{dn}; \theta_d) \prod_d \text{Dir}_K(\theta_d; \alpha) \prod_k \text{Dir}_V(\beta_k; \eta) \quad . && (7) \end{aligned}$$

2.1.1 Bag of words assumption and exchangeability

Notice that the posterior and the likelihood do *not* depend on the *order* of the words. This is the *bag of words assumption*, probabilistically, this is the notion of *exchangeability* for the words in a document.

A finite set of random variables $\{z_1, \dots, z_K\}$ is said to be exchangeable if the joint distribution is invariant to permutations. If π is a permutation of integers from 1 to N:

$$p(z_1, \dots, z_K) = p(z_{\pi_1, \dots, \pi_K}) \quad , \tag{8}$$

this is valid for all permutations π. Moreover, an infinite sequence of random variables is *infinitely exchangeable* if every finite subsequence is exchangeable.

Question: why all these formal notions?
Because with them we can apply De Finetti's theorem!
This theorem says that, the joint distribution of an infinitely exchangeable sequence of random variables is as if we were drawing a latent parameter (from some distribution) and then each of the random variable in the sequence can be treated as *indipendent* and *identically* distributed *conditioned* on the latent variable.

LDA assumes that words are generated by topics and then topics are exchangeable within a document, i.e. their order do not matter. This means that, according to De Finetti's theorem, the joint between words and topics, given a latent parameter θ, topic proportions, is:

$$p(\mathbf{w}, \mathbf{z}) = \int d\theta\, p(\theta) \left(\prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n) \right) \quad . \tag{9}$$

Similarly, also the order of the documents in a corpus is assumed to be irrelevant.

3 CAVI updates for LDA

The posterior probability in the LHS of Eq. (6) cannot be analytically written, due to the untractable normalization. We thus apply the formalism of Variational Inference and propose a variational distribution for the latent parameters using the Mean-Field family:

$$q(\beta,\theta,z)=\prod_d\left(q(\theta_d;\gamma_d)\prod_{n=1}^{N}q(z_{dn};\phi_{dn})\right)\prod_{k=1}^{K}q(\beta_k;\lambda_k) \tag{10}$$

3.0.1 Topic proportion θ_d update.

Let's write the complete conditional, and consider α a K-dimensional vector with all entries equal to α (with an abuse of notation):

$$\begin{aligned}\mathbb{E}_{q_{\setminus\theta_d}}\left[\log p(\theta_d|\beta,\theta_{\setminus d},z,w)\right] &= \log\mathrm{Dir}(\theta_d;\alpha)+\sum_n\mathbb{E}_{q(z_{dn})}[\log\mathrm{Mult}(z_{dn};\theta_d)]+const &(11)\\ &= \log\mathrm{Dir}(\theta_d;\alpha)+\sum_{n,k}\mathbb{E}_{q(z_{dn})}\left[z_{dn}^k\log\theta_{dk}\right]+const &(12)\\ &= \log\mathrm{Dir}(\theta_d;\alpha)+\sum_{n,k}\phi_{dn}^k\log\theta_{dk}+const \quad . &(13)\end{aligned}$$

This means that the optimal variational distribution for θ_d is:

$$q^*(\theta_d)\propto\prod_{k=1}^{K}\left(\theta_{dk}^{\alpha-1}\prod_{n=1}^{N}\theta_{dk}^{\phi_{dn}^k}\right)=\prod_{k=1}^{K}\theta_{dk}^{\alpha+\sum_n\phi_{dn}^k-1} \quad . \tag{14}$$

This means that the optimal variational distribution is also Dirichlet (as the prior), with variational parameter:

$$\gamma_d^*=\alpha+\sum_n\phi_{dn} \quad . \tag{15}$$

3.0.2 Topic assignment z_{dn} update.

Let's write the complete conditional:

$$\begin{aligned}\mathbb{E}_{q_{\setminus z_{dn}}}\left[\log p(z_{dn}=k|\beta,z_{\setminus dn},\theta,w)\right] &= \mathbb{E}_{q_{\setminus z_{dn}}}[\log\mathrm{Mult}(z_{dn}=k;\theta_d)]+\mathbb{E}_{q(\beta)}[\log\mathrm{Mult}(w_{dn};\beta_k)]+const\\ &= \mathbb{E}_{q(\theta_d)}[\log\theta_{dk}]+\mathbb{E}_{q(\beta)}\left[\log\beta_{kw_{dn}}\right]+const &(16)\\ &= \Psi(\lambda_{dk})-\Psi\left(\sum_k\lambda_{dk}\right)+\Psi(\lambda_{kw_{dn}})-\Psi\left(\sum_v\lambda_{kv}\right)+const \quad ,&(17)\end{aligned}$$

where $\Psi(x)$ is the digamma function.[2] The second term does not depend on k, so we can neglect it as it will be included in the normalization. This means that:

$$q^*(z_{dn}=k;\phi_{dn})\propto\exp\left\{\Psi(\lambda_{dk})+\Psi(\lambda_{kw_{dn}})-\Psi\left(\sum_v\lambda_{kv}\right)\right\}\propto\phi_{dn}^k \quad , \tag{18}$$

i.e. this is also a Multinomial, with optimal parameter:

$$\phi_{dn}^k\propto\exp\left\{\Psi(\lambda_{dk})+\Psi(\lambda_{kw_{dn}})-\Psi\left(\sum_v\lambda_{kv}\right)\right\} \quad . \tag{19}$$

[2] For a Dirchlet distributed variable θ with parameter λ, we have that $\mathbb{E}[\log\theta_i]=\Psi(\lambda_i)-\Psi\left(\sum_i\lambda_i\right)$.

3.0.3 Topic β_k update.

Let's write the complete conditional:

$$\begin{aligned}
\mathbb{E}_{q_{\setminus\beta}}[\log p(\beta|\theta,z,w)] &= \log \mathrm{Dir}_V(\beta_k;\eta) + \sum_{d,n} \mathbb{E}_{q_{\setminus\beta}}\left[\log \mathrm{Mult}(w_{dn};\beta_{z_{dn}})\right] + const && (20)\\
&= \log \mathrm{Dir}_V(\beta_k;\eta) + \sum_{d,n,v} \mathbb{E}_{q_{\setminus\beta}}\left[z^k_{dn}\, w^v_{dn}\right] \log \beta_{kv} + const && (21)\\
&= \log \mathrm{Dir}_V(\beta_k;\eta) + \sum_{v} \log \beta_{kv} \sum_{d,n} w^v_{dn} \phi^k_{dn} + const \quad . && (22)
\end{aligned}$$

This means that:

$$q^*(\beta_k;\lambda_k) \propto \prod_v \beta_{kv}^{\eta-1}\, \beta_{kv}^{\sum_{d,n}\phi^k_{dn} w^v_{dn}} = \prod_v \beta_{kv}^{\eta+\sum_{d,n}\phi^k_{dn} w^v_{dn}-1} \quad . \tag{23}$$

This means that $q^*(\beta_k;\lambda_k)$ is also Dirichlet with parameter:

$$\lambda^*_k = \eta + \sum_{d,n} \phi^k_{dn} w_{dn} \quad . \tag{24}$$

Finally, putting all together, Equations (15), (19) and (24) are the optimal variational parameter updates for the CAVI algorithm. An pseudocode for it is given in Algorithm 1.

Algorithm 1: CAVI for LDA

Input: Set of words in documents **w** and model parameters α, η, K

Output: Variational parameters λ, γ, ϕ

Initialize: Variational parameters λ, γ randomly

while *the* ELBO *has not converged* **do**

  repeat

    for *each document* d **do**

      for *each word* n **do**

        Set $\phi^k_{dn} \propto \left\{\Psi(\lambda_{dk}) + \Psi(\lambda_{kw_{dn}}) - \Psi\left(\sum_v \lambda_{kv}\right)\right\} \quad , \forall k = 1,\ldots,K$

      end

      Set $\gamma_d = \alpha + \sum_n \phi_{dn}$

    end

  until ϕ *and* γ *have converged*;

  for $k \in \{1,\ldots,K\}$ **do**

    Set $\lambda_k = \eta + \sum_{d,n} \phi^k_{dn} w_{dn}$

  end

  Compute ELBO(ϕ,γ,λ)

end

return $q(\phi,\gamma,\lambda)$

4 Inference and prediction task

Once you perform inference and obtained estimates for the variational distributions' parameters (ϕ, λ, γ), you can use these results to perform several inference and prediction tasks.

Document classification. Imagine you want to classify documents, i.e. assign labels to them. The standard scenario is to use the set of words in each document, potentially thousand of them, as your features. LDA allows you to reduce the dimensionality of these features by considering instead the parameters just estimated. This means that for each document d, you have the set of features γ_d, a K-dimensional vector which is much shorter than the, potentially huge, set of words. An example of such result is given in Figure 2.

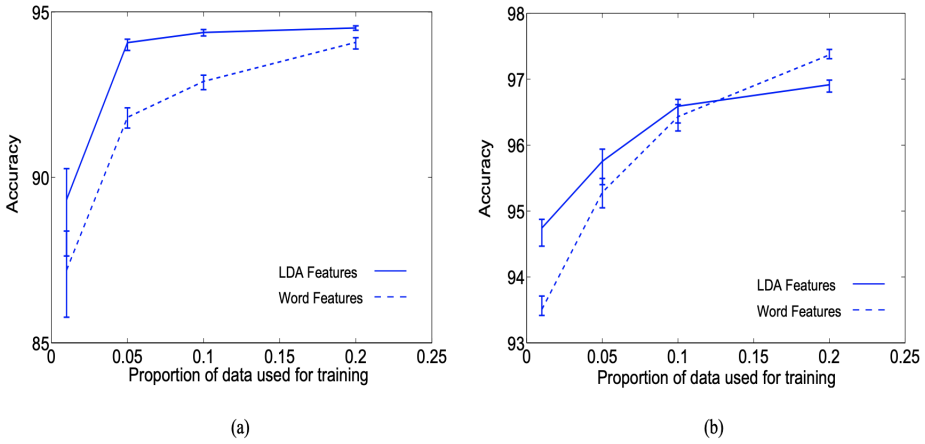

Figure 2: Example of document classification task. Each document can be classified according to 2 labels (binary classification). This dataset consists of 8000 documents and order of 15000 words. Figure taken from Blei *et al.* (2003).

Compare documents Given that each document can be represented (in a low dimensional space) by its membership (topic proportions) θ_d, we can measure similarities between documents by means of vector metrics, e.g. cosine similarity or L_1 distance.

Predictive tasks One can also predict words in an unseen document d_{new}, provided the topic distributions is estimated for that d. For this task, one has to first learn topics β_k from a corpus. This is then fixed for all future (unseen) documents. Then one either fixes $\theta_{d_{new}}$ from prior knowledge, or estimates it from a partial observation of the document $\mathbf{w}_{obs}$, i.e. observing only a subset of its words. Then one estimates:

$$p(w_{new}|\mathbf{w}_{obs}, \mathscr{D}) = \int\int d\beta d\theta \left(\sum_k \theta_k \beta_{k,w_{new}}\right) p(\theta|\mathbf{w}_{obs}, \beta) p(\beta|\mathscr{D}) \tag{25}$$

$$\approx \int\int d\beta d\theta \left(\sum_k \theta_k \beta_{k,w_{new}}\right) q(\theta) q(\beta) \tag{26}$$

$$= \sum_k \mathbb{E}_{q(\theta)}[\theta_k]\, \mathbb{E}_{q(\beta)}\left[\beta_{k,w_{new}}\right] \quad , \tag{27}$$

where $q(\beta)$ depends on the training corpus $\mathscr{D}$ and $q(\theta)$ depends on β and the observed words $\mathbf{w}_{obs}$ in the new document. The first step uses $\sum_z p(w_{new}|z)p(z|\theta) = \left(\sum_k \theta_k \beta_{k,w_{new}}\right)$. Recall that $p(z_{dn} = k|\theta_d, \beta, w_{dn}) \propto \exp\left\{\log \theta_{dk} + \log \beta_{k,w_{dn}}\right\}$.

5 Beyond standard LDA

LDA relies on few assumptions.

1. Dirichlet distribution for topic proportions: topics are independent within a document. This means that it fails to directly model correlation between the occurrence of topics.
   Specifically, under a Dirichlet, the components of the proportions vector are nearly independent[3], which leads to the strong assumption that the presence of one topic is not correlated with the presence of another. The Correlated Topic Model of Blei *et al.* (2007) tackles this with a more structured distribution that accounts for covariance, it uses the logistic normal distribution, a multivariate distribution on the simplex, i.e. it preserve the constraint that $\sum_k \theta_{dk} = 1$.
   Question: what is the cost? Higher complexity in deriving the updates. Mean Field variational family is still considered, but there is not a closed-form CAVI update for the $q(\theta_d)$, one needs gradient updates instead.
2. Bag of words: no order in the words nor in the sequence of documents inside the corpus. This is ok if we want to find semantic meanings, but it fails on other tasks such as language learning or generating synthetic sentences. The model of Wallach (2006) is a topic model that relaxes the bag of words assumption by assuming that the topics generate words conditional on the previous word.
   The order of the document is often relevant: the topics of a document collection can evolve over time. The Dynamic Topic Model of Blei and Lafferty (2006) captures the dynamics in time by making the topics at time t depend on the topics at time $t-1$.
3. One has to input K or find it through model selection. One can consider Bayesian nonparametric as in Blei *et al.* (2010); Griffiths *et al.* (2004) a hierarchical topic modeling where the prior of the Dirichlet determine the effective number of topics K.
4. Mean Field approximation: recent works Fan *et al.* (2018) have considered a more flexible approximation using a TAP free energy optimized via Approximate Message Passing. They show that this makes improvements over the MF case in cases where the signal-to-noise ratio is low, i.e. the data do not contain enough signal to allow inference. In other words, there is a region in the signal-to-noise domain, where the data do not contain enough information but the MF solution shows a result that seems to make sense, but it is not indeed correlated to the ground truth.

Topic Model: summary

- We have studied LDA, a generative model for topic models. This is a mixed membership model where one document is made of a mixture of topics.
- We have applied Variational Inference to perform posterior inference.
- This is a nice testbed to derive CAVI updates and analyze the problem.

A main reference for this lecture is Blei *et al.* (2017).

References

D. M. Blei, A. Y. Ng, and M. I. Jordan, Journal of machine Learning research **3**, 993 (2003).

D. M. Blei, J. D. Lafferty, *et al.*, The Annals of Applied Statistics **1**, 17 (2007).

[3] "Nearly" because there is some small negative correlation caused by the constraint that they should sum to 1.

H. M. Wallach, in *Proceedings of the 23rd international conference on Machine learning* (2006) pp. 977–984.

D. M. Blei and J. D. Lafferty, in *Proceedings of the 23rd international conference on Machine learning* (2006) pp. 113–120.

D. M. Blei, T. L. Griffiths, and M. I. Jordan, Journal of the ACM (JACM) **57**, 1 (2010).

T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, and D. M. Blei, in *Advances in neural information processing systems* (2004) pp. 17–24.

Z. Fan, S. Mei, and A. Montanari, arXiv preprint arXiv:1808.07890 (2018).

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, Journal of the American Statistical Association **112**, 859 (2017).