Semih Akbayrak

Probabilistic Topic Modeling

CMPE547 Final Project Report

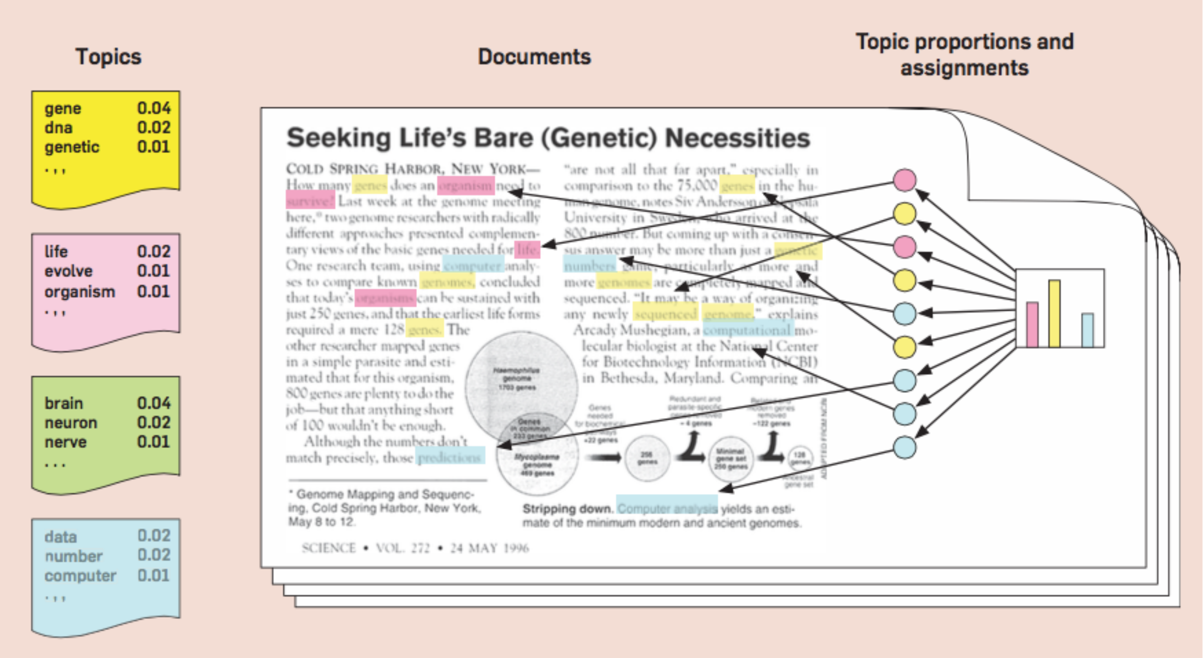
Abstract

In this project, I summarized my study on David M. Blei’s articles on Probabilistic Topic Models(2012) and Latent Dirichlet Allocation(2003). I applied LDA on a small text dataset as a simple example, as well.

Probablistic Topic Models

Today, in this digital age, tons of written and visual data are being produced every day and this makes harder to organize them by hand. It is must to create efficient methods and algorithms to do that. One way to do this is to propose topic models and estimate the model parameters. The goal in there is exploring the themes behind the documents.

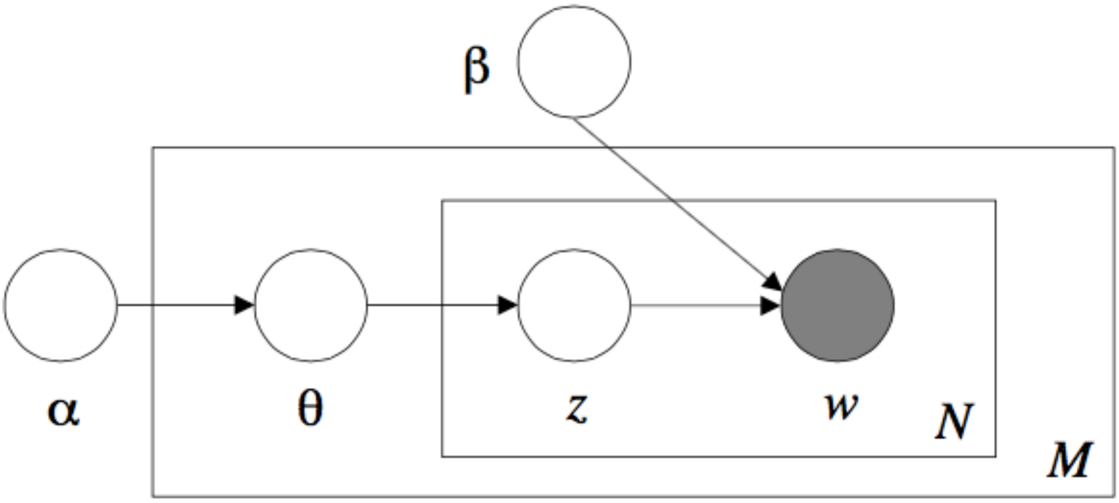
Latent Dirichlet Allocation



The idea behind the LDA is that a document is a composition of different topics with different proportions and the appearance rate of words in topics are different, as well.

Above figure is a good illustration of how generative model works for LDA. This document is mostly about three topics which are data science, evolutionary biology, and genetics. The first thing specified is the topic rates for the document and in the right of the figure, you can see these topic ratios. In order to produce a word, generative model first choose a topic, then choose the word according to their appearance probabilities in this topic.

Graphical Model



This is the graphical model of what I described above. And here are the small definitions for parameters.

$$\theta_d : k\ dimensional\ vector.\ Topic\ distributions\ for\ one\ document.$$
$$z_{dn} : k\ dimensional\ vector. Topic\ assignment\ to\ n^{th}\ word\ of\ document\ d$$
$$w_d : N_d\ dimensional\ vector\ represents\ d^{th}\ document$$
$$\beta : k \times V\ dimensional\ matrix\ where\ \beta_{ij} = p(w^j = 1|z^i = 1)$$

Here k is number of topics, N_d is number of words in document d and V is the number of words in vocabulary.
We draw topic distributions from Dirichlet distribution.

$$(\theta_{d1}, \dots, \theta_{dk}) \sim Dir(\alpha_1, \dots, \alpha_k)$$
$$p(\theta_d|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_{di}^{\alpha_i - 1}$$

z_{dn} coming from multinomial distribution.

$$z_{dn} \sim Multinomial(\theta_d)$$

$$p(z_{dn}|\theta_d) = \prod_{i=1}^{k} \theta_{di}^{[z_{dn}==1]}$$

It has been assumed that words in a document are independent, so we can write

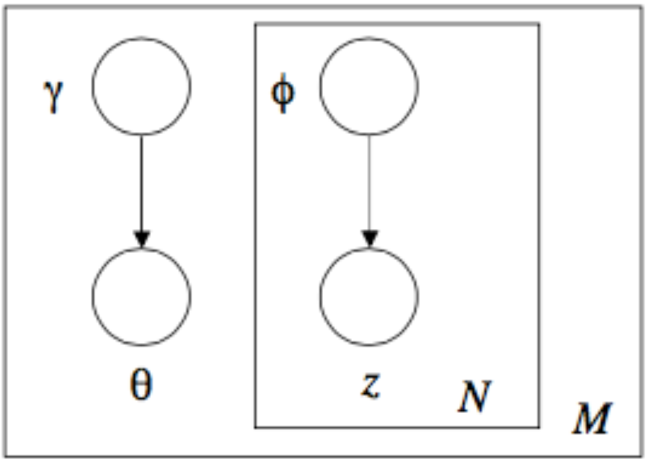
$$p(w_d|z,\beta) = p(w_{d1},\ldots w_{dN_d}|z,\beta) = p(w_{d1}|z_{d1},\beta)\ldots p(w_{dN_d}|z_{dN_d},\beta)$$
$$= \prod_{n=1}^{N_d} \beta_{z_{dn}w_{dn}}$$

Our aim is to find the posterior for hidden variables.

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

$$p(w|\alpha,\beta) = \int \sum_z p(\theta,z,w|\alpha,\beta)d\theta$$

But difficulty in marginalization of joint distribution directs us to use variational approximation.



Inference & Parameter Estimation

By utilizing from Jensen's inequality, we can write the below statements.

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

$$= \log E_q[\frac{p(\theta, z, w|\alpha, \beta)}{q(\theta, z)}] \geq E_q[\log p(\theta, z, w|\alpha, \beta)] - E_q[\log q(\theta, z)]$$

The right hand side of this statement is lower bound and the difference between right side and left side is KL divergence. Our goal is to minimize KL divergence or equivalently maximize lower bound. By looking at the graphical models we can write lower bound as

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

Now in the below formulas you can see how we find these expectations.

$$E_q[\log p(\theta|\alpha)] = E_q[\log(\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1})]$$

$$= E_q[\log(\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}) + \sum_i \log \theta_i^{\alpha_i - 1}]$$

$$= \log \Gamma(\sum_i \alpha_i) - \sum_i \log \Gamma(\alpha_i) + E_q[\sum_i (\alpha_i - 1) \log \theta_i]$$
$$= \log \Gamma(\sum_i \alpha_i) - \sum_i \log \Gamma(\alpha_i) + \sum_i (\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_j \gamma_j))$$

$$E_q[\log p(z|\theta)] = E_q[\log \prod_n \prod_i \theta_i^{[z_n == i]}]$$
$$= E_q[\sum_n \sum_i [z_n == i] \log \theta_i] = \sum_n \sum_i \phi_{ni} E_q[\log \theta_i]$$
$$= \sum_n \sum_i \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_j \gamma_j))$$

$$E_q[\log p(w|z,\beta)] = E_q[\log \beta_{z_{dn}w_{dn}}] = E_q[\log \prod_{v}^{V} \prod_{i}^{K} \beta_{iv}^{[w_{dn}==v, z_{dn}==i]}]$$

$$= \sum_{v} \sum_{i} E_q[w_{dn} == v, z_{dn} == i] \log \beta_{iv}$$

$$= \sum_{v} \sum_{i} \phi_{ni} w_{dn}^{v} \log \beta_{iv}$$

$$E_q[\log q(\gamma)] = \log \Gamma(\sum_{j} \gamma_j) - \sum_{i} \log \Gamma(\gamma_i)$$

$$+ \sum_{i} (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j} \gamma_j))$$

$$E_q[\log q(\phi_{dn})] = \sum_{i} \phi_{dni} \log \phi_{dni}$$

In order to find optimizing parameter values which make lower bound L maximum, we can take derivative of L for each parameter individually and set equal to 0.

$$\frac{\partial L}{\partial \phi_{ni}} = 0 \;\; => \phi_{ni} \propto \beta_{iv} exp(\Psi(\gamma_i) - \Psi(\sum_{j} \gamma_j))$$

$$\frac{\partial L}{\partial \gamma_i} = 0 \;\; => \gamma_i \propto \alpha_i + \sum_{n=1}^{N_d} \phi_{ni}$$

$$\frac{\partial L}{\partial \beta_{ij}} = 0 \;\; => \beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}$$

Algorithm

We’ve found update equations, now we can use them in EM algorithm.

```
$Initialize\ \phi,\ \gamma,\ \alpha,\ \beta\ randomly$
$//E - step$
$for\ d = 1\ to\ M$
    $repeat$
        $for\ n = 1\ to\ N_d$
            $for\ i = 1\ to\ k$
                $\phi_{dni}^{t+1} = \beta_{iw_n} exp(\Psi(\gamma_{di}^{t}))$
            $normalize\ \phi_{dn}^{t+1}$
    $\gamma_d^{t+1} = \alpha + \sum_{n=1}^{N_d} \phi_{dn}^{t+1}$
  $until\ convergence$
$//M - step$
$for\ i = 1\ to\ k$
  $for\ j = 1\ to\ V$
    $\beta_{ij} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}$
  $normalize\ \beta_i$
```

Practicle Example

I applied LDA by using NLTK's Brown corpus. To keep it simple, I chose 5 documents from politics and 5 from science-fiction. By arranging topic number k=2, I got the following words as the most probable words for 2 topics.

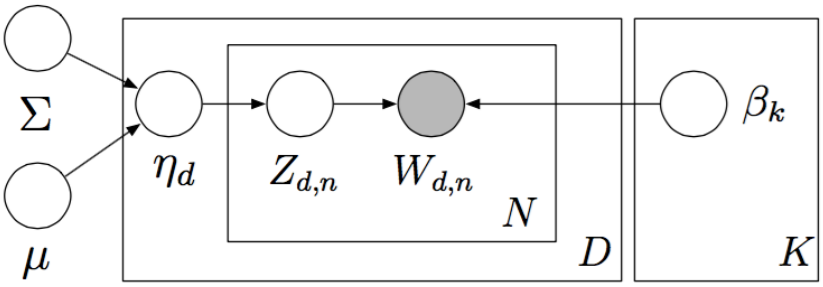
Topic1: man, time, people, long, know, ship, back, mind, felt, years

Topic2: state, development, vehicles, industrial, years, governor, government, countries, aid, tax

By looking at these words, we can label the Topic1 as science-fiction and Topic2 as politics.

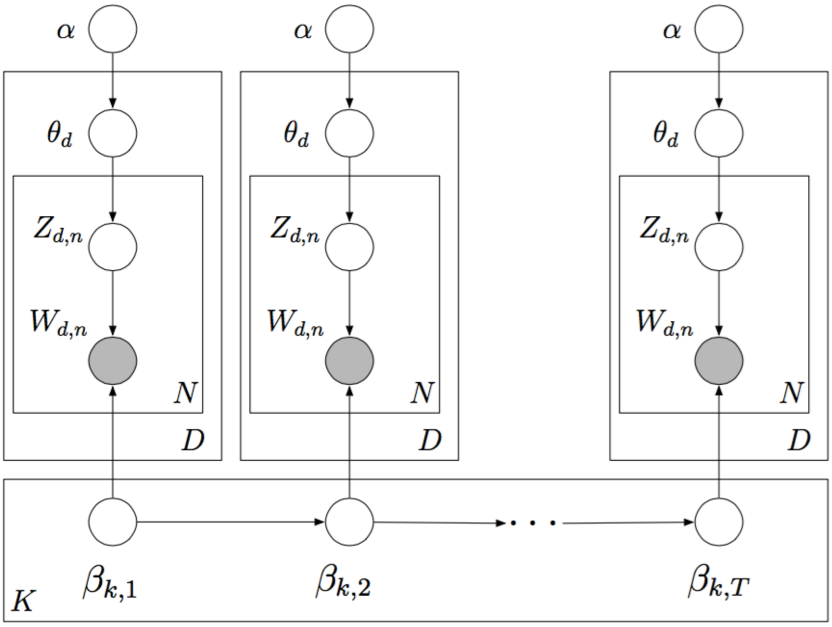
Beyond LDA

Correlated Topic Model(CTM)



In LDA, we used Dirichlet distribution for priors but this makes the topic distributions independent for a document. But we know that, an article about computer science is most likely to be about electrical engineering than about genetics. To set the correlations between topics, we can use covariance matrix and this directs us to draw topic assignments from multivariate normal distribution.

Dynamic Topic Models



When we work on documents which are published in a short period of time, the order of the documents may not be important. But if we want to work for the longer periods, the order will be important, because language and terms change in the long period. By setting a Markov model, these changes can be tracked and Dynamic Topic Models are used for these purposes.

References

1)Blei D., Ng A., Jordan M. Latent dirichlet allocation. The Journal of Machine Learnning Research, 2003.
2)Blei D. Probabilistic topic models. August 2011
3)Blei D. Probabilistic topic models. Communication of the ACM, 2012
4)Reed C. Latent dirichlet allocation:Towards a deeper understanding. January 2012.
5)Boyd-Grabber J. Variational inference, 2015. Retrieved from **https://www.youtube.com/watch?v=2pEkWk-LHmU**