Topic Modeling : Latent Dirichlet Allocation

Team 11

Fatima-Zahra Banani
Mohamed Kleit
Robert Chamoun

Introduction

- Topic Modeling aims at discovering hidden themes in a collection of documents and annotating them according to these topics.
- Latent Dirichlet Allocation (LDA) is a generative probabilistic model of document collections that tries to capture the intution that documents exhibit multiple topics.
- It thus has a corresponding generative process from which each document is assumed to be generated.

Graphical Model

LDA is a three-level hierarchical Bayesian model as depicted below:

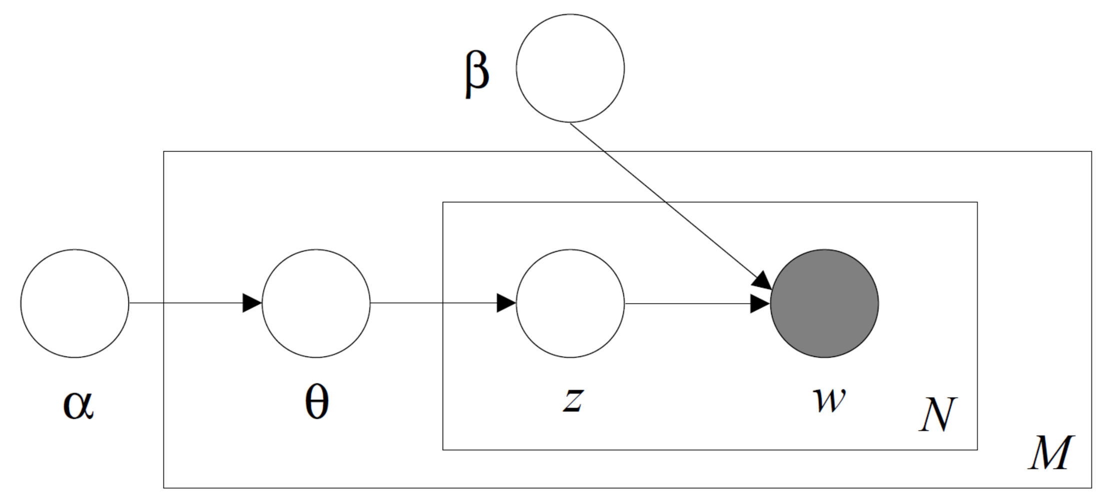

Figure 1:Graphical model representation of LDA

$$p(\theta_d, Z_d, W_d|\alpha, \beta) = p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_n|\theta_d)p(w_n|z_n, \beta)$$

α : topic proportions parameter
β : KxV matrix representing each topic k as a distribution over the vocabulary V, where $\beta_{k,v} = p(w^v = 1|z^k = 1)$
θ_d : topic proportions for document d
Z_d : topic assignments for document d
W_d : observed words for document d

Generative Process

Each document d is assumed to be generated by the following process:

- Draw distribution over topics $\theta_d \sim Dirichlet(\alpha)$
- For each word n:
  - Draw a topic from distribution over topics $z_{d,n} \sim Mult(\theta_d)$
  - Draw a word from the corresponding topic $w_{d,n} \sim Mult(\beta_{z_{d,n}})$

Goal

Find:

$$\alpha^*, \beta^* = argmax \log p(D|\alpha, \beta) = argmax \sum_{d=1}^{M} \log p(W_d|\alpha, \beta)$$

Parameter Estimation

- $p\,(\,\theta_d, Z_d\,|\,W_d, \alpha, \beta)$ is intractable $\Rightarrow$ Variational EM.

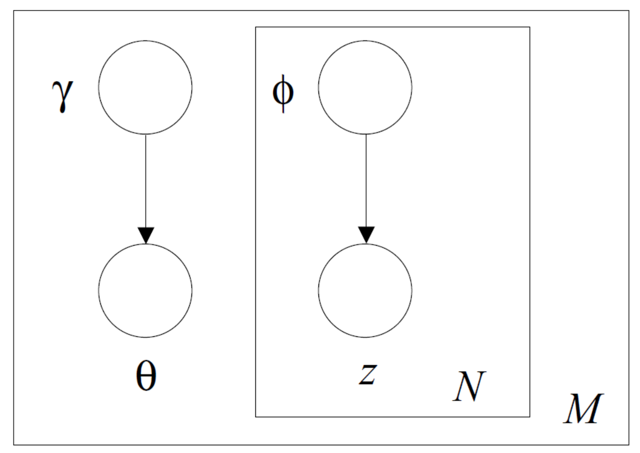

Figure 2:Graphical Model representation of variational distribution

$$q\,(\,\theta_d, Z_d\mid\gamma_d, \phi_d) = q(\theta_d\mid\gamma_d)\prod_{n=1}^{N}\; q\,(\,z_n\mid\phi_d^n)$$

$$\log p(W_d|\alpha, \beta) = L(\gamma_d, \phi_d, \alpha, \beta) + D(q(\theta_d, Z_d|\gamma_d, \phi_d)||p(\theta_d, Z_d|W_d, \alpha, \beta))$$

- **Variational E Step:**

For each document d:

$$\gamma_d^{*t+1}, \phi_d^{*t+1} = argmin\, D(q(\theta_d, Z_d|\gamma_d, \phi_d)||p(\theta_d, Z_d|W_d, \alpha^{*t}, \beta^{*t}))$$

- **Variational M Step:**

$$\alpha^{*^{t+1}}, \beta^{*^{t+1}} = argmax L(\gamma^{*^{t+1}}, \phi^{*^{t+1}}, \alpha, \beta) = argmax \sum_{d=1}^{M} L(\gamma_d^{*^{t+1}}, \phi_d^{*^{t+1}}, \alpha, \beta)$$

Implementation and Results

Reproduced results from paper with our own implementation:

- Documents: 2247
- Topics: 100

Topic Sample				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
club	new	child	ago	american
year	percent	people	africa	states
york	yen	having	leaders	union
old	economy	report	france	told
building	rate	education	people	political
business	rates	aids	african	leaders
years	said	children	french	conference
new	prices	said	south	president
city	market	care	police	said
said	dollar	health	said	soviet

Tested our implementations on scribe notes:

- Documents: 17
- Topics: 8

Topic Sample				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
inference	number	problem	constraints	sufficient
structure	possible	passing	uniform	moments
active	beta	running	hessian	divergence
parents	equal	vector	point	likelihood
treewidth	parameters	latent	constant	principle
elimination	information	following	lagrange	inference
eliminate	posteriori	inference	equality	optimization
graphs	priori	regression	constraint	proof
property	posterior	property	family	information
ordering	case	matrix	stationary	general

References

[1] Michael I. Jordan David M. Blei, Andrew Y. Ng.
Latent dirichlet allocation.
Journal of Machine Learning Research, (3):993--1022, 2003.