MATH6450E Project 1: Latent Dirichlet Allocation using Variational Inference

FAN Min *
Department of Mathematics
Hong Kong University of Science and Technology
mfanab@connect.ust.hk

Abstract

Latent Dirichlet allocation (LDA) is a famous topic modeling algorithm proposed at the beginning of the 21^{th} century, which is a three-level hierarchical Bayesian model. Variational inference is implemented in E-step of EM algorithm to approximate the posterior distribution of the latent variables. Under the framework of EM algorithm, the log-likelihood of observed variables is maximized and meanwhile the parameters in the model are learned from training data. In this report, we deduced LDA algorithm and did simulation on a small corpus.

1 Introduction

Latent Dirichlet allocation (LDA) is a three-level hierarchical generative probabilistic model, which generates text as 2 shows. N is the number of the documents. N_i is the number of words in document i, which follows Poisson distribution in generating process . K is the number of topics. M is the number of words in vocabulary. θ_i is a K-length vector as the parameter of Multinomial distribution. α is a K-length vector as the parameter of Dirichlet distribution as the prior of θ_i. φ_k is a M-length vector as the parameter of Multinomial distribution. β is a M-length vector as the parameter of Dirichlet distribution as the prior of φ_k. $z_{i,j}$ refers to topic index in topic list of word j in document i. $w_{i,j}$ refers to the word index in vocabulary of word j in document i.

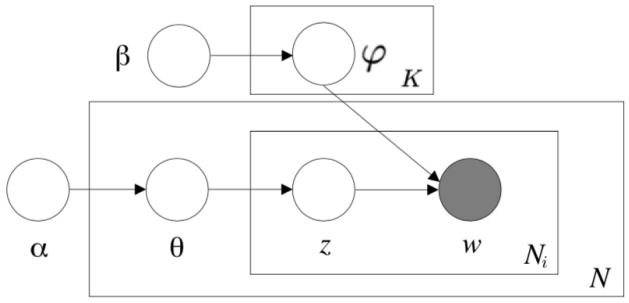

Figure 1: Latent Dirichlet Allocation

*https://github.com/ProteusFAN/MATH6450E/tree/master/Latent%20Dirichlet%20Allocation%20using%20Variational%20Inference

2 Method

The goal is to maximize the log-likelihood of observed variables with respect to parameter α and φ, that is,

$$\begin{aligned} p(w|\alpha,\varphi) &= \int p(\theta|\alpha)(\prod_{n=1}^{N_i}\sum_{k=1}^{K} p(z_n=k|\theta)p(w_n|z_n=k))d\theta \\ &= \int p(\theta|\alpha)(\prod_{n=1}^{N_i}\sum_{k=1}^{K}\prod_{j=1}^{M}(\theta_k\varphi_{kj})^{w_n^j})d\theta, \end{aligned}$$

which is not analytically tractable. EM algorithm is to maximize the log-likelihood of the observed variables.

2.1 E-step

The posterior distribution of latent variables

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

is hard to compute. Mean-field variational inference is implemented to approximate $p(\theta, z|w,\alpha,\varphi)$ by $q(\theta, z|\gamma,\phi)$, where

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

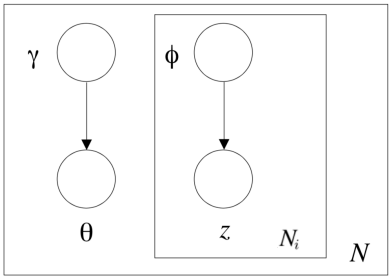

Figure 2: Mean-field variational approximation

The difference between two distributions is measured by KL-divergence. So the solution is given by

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

It is easy to show

$$\log(p(w|\alpha,\varphi)) = \mathcal{L}(\gamma,\phi;\alpha,\varphi) + KL(q(\theta,z|\gamma,\phi)||p(\theta,z|w,\alpha,\varphi),$$

where

$$\mathcal{L}(\gamma,\phi;\alpha,\varphi) = E_q[\log(p(\theta,z,w|\alpha,\varphi))] - E_q[\log(q(\theta,z|\gamma,\phi))].$$

So minimizing $KL(q(\theta,z|\gamma,\phi)||p(\theta,z|w,\alpha,\varphi)$ is equivalent to maximizing $\mathcal{L}(\gamma,\phi;\alpha,\varphi)$.

2.1.1 Computing $\mathcal{L}(\gamma,\phi;\alpha,\varphi)$

It is easy to show

$$\begin{aligned} \mathcal{L}(\gamma,\phi;\alpha,\varphi) &= E_q[\log(p(\theta,z,w|\alpha,\varphi))] - E_q[\log(q(\theta,z|\gamma,\phi))] \\ &= E_q[\log(p(\theta|\alpha))] + E_q[\log(p(z|\theta))] + E_q[\log(p(w|z,\varphi))] \\ &\quad - E_q[\log(q(\theta|\gamma))] - E_q[\log(q(z|\phi))]. \end{aligned}$$

In order to compute $\mathcal{L}(\gamma, \phi; \alpha, \varphi)$, we just need to compute these five terms.

As for the first term $E_q[\log(p(\theta|\alpha))]$,

$$\begin{aligned} E_q[\log(p(\theta|\alpha))] &= \sum_{i=1}^{K}(\alpha_i - 1)E_q[\log\theta_i] + \log\Gamma(\sum_{i=1}^{K}\alpha_i) - \sum_{i=1}^{K}\log\Gamma(\alpha_i) \\ &= \sum_{i=1}^{K}(\alpha_i - 1)(\psi(\gamma_i) - \psi(\sum_{j=1}^{K}\gamma_j)) + \log\Gamma(\sum_{i=1}^{K}\alpha_i) - \sum_{i=1}^{K}\log\Gamma(\alpha_i), \end{aligned}$$

where θ is generated from $Dir(\theta|\gamma)$ and ψ is digamma function.
As for the second term $E_q[\log(p(z|\theta))]$,

$$\begin{aligned} E_q[\log(p(z|\theta))] &= E_q[\sum_{n=1}^{N_d}\sum_{i=1}^{K} z_{ni}\log\theta_i] \\ &= \sum_{n=1}^{N_d}\sum_{i=1}^{K} E_q[z_{ni}]E_q[\log\ theta_i] \\ &= \sum_{n=1}^{N_d}\sum_{i=1}^{K}\phi_{ni}(\psi(\gamma_i) - \psi(\sum_{j=1}^{K}\gamma_j)), \end{aligned}$$

where z is generated from $\mathrm{Multinomial}(z|\phi)$ and N_d is the number of words in document d.
As for the third term $E_q[\log(p(w|z,\varphi))]$,

$$\begin{aligned} E_q[\log(p(w|z,\varphi))] &= E_q[\sum_{n=1}^{N_d}\sum_{i=1}^{K}\sum_{j=1}^{M} z_{ni}w_n^j\log\varphi_{ij}] \\ &= \sum_{n=1}^{N_d}\sum_{i=1}^{K}\sum_{j=1}^{M} E_q[z_{ni}]w_n^j\log\varphi_{ij} \\ &= \sum_{n=1}^{N_d}\sum_{i=1}^{K}\sum_{j=1}^{M}\phi_{ni}w_n^j\log\varphi_{ij}. \end{aligned}$$

As for the fourth term $E_q[\log(q(\theta|\gamma))]$,

$$\begin{aligned} E_q[\log(q(\theta|\gamma))] &= \sum_{i=1}^{K}(\gamma_i - 1)E_q[\log\theta_i] + \log\Gamma(\sum_{i=1}^{K}\gamma_i) - \sum_{i=1}^{K} log\Gamma(\gamma_i) \\ &= \sum_{i=1}^{K}(\gamma_i - 1)(\psi(\gamma_i) - \psi(\sum_{j=1}^{K}\gamma_j)) + \log\Gamma(\sum_{i=1}^{K}\gamma_i) - \sum_{i=1}^{K} log\Gamma(\gamma_i). \end{aligned}$$

As for the fifth term $E_q[\log(q(z|\phi))]$,

$$\begin{aligned} E_q[\log(q(z|\phi))] &= E_q[\sum_{n=1}^{N_d}\sum_{i=1}^{K} z_{ni}\log\phi_{ni}] \\ &= \sum_{n=1}^{N_d}\sum_{i=1}^{K} E_q[z_{ni}]\log\phi_{ni} \\ &= \sum_{n=1}^{N_d}\sum_{i=1}^{K}\phi_{ni}\log\phi_{ni}. \end{aligned}$$

Finally, we have

$$\begin{aligned}\mathcal{L}(\gamma,\phi;\alpha,\varphi) &= \sum_{i=1}^{K}(\alpha_i-1)(\psi(\gamma_i)-\psi(\sum_{j=1}^{K}\gamma_j))+\log\Gamma(\sum_{i=1}^{K}\alpha_i)-\sum_{i=1}^{K}\log\Gamma(\alpha_i)\\ &+\sum_{n=1}^{N_d}\sum_{i=1}^{K}\phi_{ni}(\psi(\gamma_i)-\psi(\sum_{j=1}^{K}\gamma_j))\\ &+\sum_{n=1}^{N_d}\sum_{i=1}^{K}\sum_{j=1}^{M}\phi_{ni}w_n^j\log\varphi_{ij}\\ &-\sum_{i=1}^{K}(\gamma_i-1)(\psi(\gamma_i)-\psi(\sum_{j=1}^{K}\gamma_j))+\log\Gamma(\sum_{i=1}^{K}\gamma_i)-\sum_{i=1}^{K}log\Gamma(\gamma_i)\\ &-\sum_{n=1}^{N_d}\sum_{i=1}^{K}\phi_{ni}\log\phi_{ni}\end{aligned}$$

2.1.2 Updating γ, ϕ

We update ϕ first. Because $\sum_{j=1}^{K}\phi_{ni}=1$, we implement Lagrange Multiplier and have

$$\mathcal{L}_{\phi_{ni}}=\phi_{ni}(\psi(\gamma_i)-\psi(\sum_{j=1}^{K}\gamma_j))+\phi_{ni}\log\varphi_{iw}-\phi_{ni}\log\phi_{ni}+\lambda(\sum_{j=1}^{K}\phi_{ni}-1),$$

and take derivative

$$\frac{\partial\mathcal{L}}{\partial\phi_{ni}}=(\psi(\gamma_i)-\psi(\sum_{j=1}^{K}\gamma_j))+\log\varphi_{iw}-\log\phi_{ni}-1+\lambda,$$

and let it be 0 whereby we have

$$\phi_{ni}\propto\varphi_{iw}\exp(\psi(\gamma_i)-\psi(\sum_{j=1}^{K}\gamma_j)).$$

We update γ next.

$$\mathcal{L}_{\gamma}=\sum_{i=1}^{K}(\psi(\gamma_i)-\psi(\sum_{j=1}^{K}\gamma_j))(\alpha_i+\sum_{n=1}^{N_d}\phi_{ni}-\gamma_i)-\log\Gamma(\sum_{i=1}^{K}\gamma_i)+\sum_{i=1}^{K}log\Gamma(\gamma_i)),$$

and take derivative

$$\frac{\partial\mathcal{L}}{\partial\gamma_i}=\psi(\gamma_i)(\alpha_i+\sum_{n=1}^{N_d}\phi_{ni}-\gamma_i)-\psi(\sum_{j=1}^{K}\gamma_j)\sum_{j=1}^{K}(\alpha_j+\sum_{n=1}^{N_d}\phi_{nj}-\gamma_j),$$

and let it be 0 whereby we have

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

We initialize ϕ_{ni}^0 to be $\frac{1}{K}$ for all i,n and γ_i to be $\alpha_i+\frac{N_d}{K}$, then update ϕ and γ alternatively until convergence.

2.2 M-step

2.2.1 Updating φ

Because $\sum_{j=1}^{M}\varphi_{ij}=1$ for all i, we implement Lagrange Multiplier and have

$$\mathcal{L}_{\varphi}=\sum_{d=1}^{N}\sum_{n=1}^{N_d}\sum_{i=1}^{K}\sum_{j=1}^{M}\phi_{dni}w_{dn}^j\log\varphi_{ij}+\sum_{i=1}^{K}\lambda_i(\sum_{j=1}^{M}\varphi_{ij}-1),$$

and take derivative and set it to be 0 whereby we have

$$\varphi_{ij} \propto \sum_{d=1}^{N} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}$$

2.2.2 Updating α

We have

$$\mathcal{L}_\alpha = \sum_{d=1}^{N} (\log \Gamma(\sum_{j=1}^{K} \alpha_j) - \sum_{i=1}^{K} \log \Gamma(\alpha_i)) + \sum_{i=1}^{K} ((\alpha_i - 1)(\psi(\gamma_{di}) - \psi(\sum_{j=1}^{K} \gamma_{dj}))),$$

and take derivative

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = N(\psi(\sum_{j=1}^{K} \alpha_j) - \psi\alpha_i) + \sum_{d=1}^{N} (\psi(\gamma_{di} - \psi(\sum_{j=1}^{K} \gamma_{dj})),$$

and Hessian matrix is

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_i \partial \alpha_j} = N(\psi(\sum_{j=1}^{K}) - \delta(i,j)\psi(\alpha_i)).$$

We optimize $\mathcal{L}_\alpha$ w.r.t. α using linear-time Newton-Raphson algorithm.

3 Experiment and Result

In experiments, we follow EM algorithm framework and updating strategies deduced in previous section, except that hyperparameter α is set to be initial value and does not update for convenience.

We implement it on a small corpus. Training data contains 10 documents about one anime in Wikipedia. The top 10 topics of each document are showed in 1.

Table 1: Top 10 topics of each document

Document index	Topics
0	manga, baroque, alabasta, series, tony, crocodile, book, copies, volumes, oda
1	island, pirates, fishman, straw, hats, crew, alliance, half, emperors, hazard
2	luffy, crew, ace, navy, pirate, roger, sabaody, archipelago, whitebeard, rayleigh
3	haki, color, treasure, king, body, shoku, possesses, properties, wan, kaizoku
4	crew, franky, government, straw, robin, pirates, war, hat, battle, joins
5	piece, animals, den, set, grand, wind, snails, daiaru, anachronisms, applications
6	grand, sea, su, called, blue, red, bur, mountain, pose, belts
7	devil, fruit, user, fruits, sea, water, series, powers, north, transform
8	luffy, pirates, captain, monkey, named, roger, head, nami, crew, navy
9	luffy, dressrosa, flame, zou, pirates, sanji, mom, alliance, competition, rescue

The topics of two held-out sentences are predicted using LDA model as 2 and prediction is quite accurate based from a human perspective.

Table 2: Topic prediction

Topic index	Sentences
7	If someone eats devil fruit, he will be hated by sea and will lose his strength if he is submerged in sea.
6	The weather on the Grand Line's open sea are extremely unpredictable.

4 Conclusion

Latent Dirichlet Allocation is a powerful topic model under probabilistic framework. Its goal is clear, to maximize log-likelihood of observed variables(words in documents). Mean-field variance inference is introduced to approximate the posterior distribution of latent variables which is analytically intractable. From simulation on a small corpus, we can see Latent Dirichlet Allocation is effective and accurate.