Statistical Modeling SDS 383D: Final Project

Latent Dirichlet Allocation

Khai Nguyen - kbn397

May 4, 2022

Abstract

Topic modeling is one of important tasks in natural language processing. Among topic models, Latent Dirichlet Allocation (LDA) is one of the most famous and fundamental models. LDA is a three-level probabilistic hierarchical model. In the context of text data, each document is represented by a mixture over topics and each topic is represented by a mixture of words in LDA. The model is estimated by doing approximate maximum log-likelihood via the expectation-maximization algorithm. Moreover, variational inference is used in LDA for fast Bayesian inference in LDA. In this report, LDA is fitted on a Web snippet dataset for exploring the topics in that corpus. Also, the latent representation of documents are used in down-stream tasks as compressed data.

1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is introduced in the first time in [1]. LDA is a generative probabilistic model of a corpus. The idea is of LDA is to represent documents as random mixtures over latent topics, where each topic is characterized by a distribution over words. In this section, I will review the graphical model and the generative process of LDA. Then, I will review the parameter estimation method of LDA.

1.1 Generative Process and Graphical Model

I first define some terminologies. A word is a basic unit in a vocabulary of size V, indexed by $\{1, \dots, V\}$. A corpus is a collection of M documents denoted by $D = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$. A document $\mathbf{w}_d$ is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, \dots, w_N)$. The generative process of LDA is as follow:

1. Given $k > 0$, for each topic $i \in \{1, \dots, k\}$:
   (a) Sample $\beta_i \sim Dir(\eta)$.
2. For For each document $d \in \{1, \dots, M\}$:
   (a) Sample $\theta_d \sim Dir(\alpha)$.
   (b) For each word $n \in \{1, \dots, N\}$
      i. Sample $z_n \sim Multinomial(\theta_d)$
      ii. Sample $w_n \sim Multinomial(\beta_{z_n})$

The graphical model of LDA is given in Figure 1. From the figure, it is clear that LDA has three layers. The parameters α, η, and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ are document-level variables, sampled once per document.. Finally, the variables z and w are word-level variables and are sampled once for each word in each document.

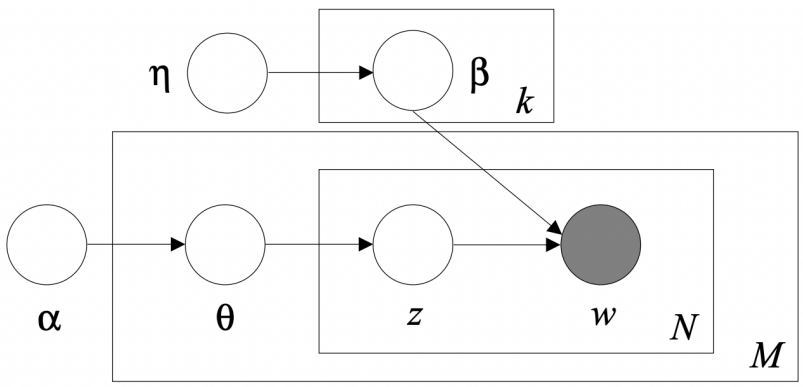

Figure 1: Graphical model of LDA.

1.2 Inference and Parameter Estimation

In this section, I will review the inference and parameter estimation of LDA. First, I will write the log-likelihood of LDA. Then, I present the configuration of the variational distributions for the variational Bayes method. The flow of derivation might be slightly different from the original paper. Now, the likelihood function with $D = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M)$ is:

$$\begin{aligned} p(D \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) &= \int\int \sum_{\boldsymbol{z}_1, \ldots \boldsymbol{z}_M} p(D, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{z}_1, \ldots, \boldsymbol{z_n} | \boldsymbol{\alpha}, \boldsymbol{\eta}) d\boldsymbol{\theta} d\boldsymbol{\beta} \\ &= \int\int \sum_{\boldsymbol{z}_1, \ldots \boldsymbol{z}_M} \prod_{k=1}^{K} p(\beta_k | \boldsymbol{\eta}) \prod_{d=1}^{M} p\left(\theta_d \mid \boldsymbol{\alpha}\right) \prod_{n=1}^{N} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \boldsymbol{\beta}) d\boldsymbol{\theta} d\boldsymbol{\beta} \end{aligned}$$

where boldsymbols represents vectors of parameters. Since the likelihood is intractable, doing maximum log-likelihood directly is impossible. To overcome this issue, a variational Bayes approach is used. We define the variational distribution:

$$\begin{aligned} q(\boldsymbol{\theta}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_M, \boldsymbol{\beta} | \boldsymbol{\gamma}, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_M, \boldsymbol{\lambda}) &= q(\boldsymbol{\theta} | \boldsymbol{\gamma}) q(\boldsymbol{\beta} | \boldsymbol{\lambda}) \prod_{d=1}^{M} p(\boldsymbol{z}_d | \boldsymbol{\phi}_d) \\ &= \prod_{i=1}^{k} q(\beta_i | \lambda_i) \prod_{d=1}^{M} p(\theta_d | \gamma_d) \prod_{n=1}^{N} p(z_{dn} | \phi_{dn}) \end{aligned}$$

The above setting of varitational distributions is called mean-field approximation since all random variable are independent.

Now, we derive the lowerbound of the log-likelihood:

$$\begin{aligned}
\log p(D|\boldsymbol{\alpha},\boldsymbol{\eta}) &= \log \int\int \sum_{\boldsymbol{z}_1,\ldots\boldsymbol{z}_M} p(D,\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n}|\boldsymbol{\alpha},\boldsymbol{\eta}) d\boldsymbol{\theta} d\boldsymbol{\beta} \\
&= \log \int\int \sum_{\boldsymbol{z}_1,\ldots\boldsymbol{z}_M} p(D,\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n}|\boldsymbol{\alpha},\boldsymbol{\eta}) \frac{q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\gamma_1,\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})}{q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\gamma_1,\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})} d\boldsymbol{\theta} d\boldsymbol{\beta} \\
&= \log \mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\gamma_1,\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})} \left[\frac{p(D,\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n}|\boldsymbol{\alpha},\boldsymbol{\eta})}{q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\gamma_1,\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})} \right]
\end{aligned}$$

Using Jensen inequality, we have

$$\begin{aligned}
\log p(D|\boldsymbol{\alpha},\boldsymbol{\eta}) &\geq \mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})} \left[\log \frac{p(D,\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n}|\boldsymbol{\alpha},\boldsymbol{\eta})}{q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\gamma_1,\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})} \left[\log p(D,\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n}|\boldsymbol{\alpha},\boldsymbol{\eta}) \right] \\
&\qquad - \mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})} \left[\log q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\gamma_1,\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda}) \right] \\
&= \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\gamma})}[\log p(\theta|\boldsymbol{\alpha})] + \mathbb{E}_{q(\boldsymbol{\beta}|\boldsymbol{\lambda})}[\log p(\boldsymbol{\beta}|\boldsymbol{\eta})] \\
&\qquad + \sum_{d=1}^{M} \left(\mathbb{E}_{q(\boldsymbol{z}_d|\boldsymbol{\phi}_d)q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)}[\log p(\boldsymbol{z}_d|\theta_d)] + \mathbb{E}_{q(\boldsymbol{z}_d|\boldsymbol{\phi}_d)q(\boldsymbol{\beta}|\boldsymbol{\lambda}))}[\log p(\boldsymbol{w}_d|\boldsymbol{z}_d,\boldsymbol{\beta}] \right) \\
&\qquad - \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\gamma})}[\log q(\boldsymbol{\theta}|\boldsymbol{\gamma})] - \mathbb{E}_{q(\boldsymbol{\beta}|\boldsymbol{\lambda})}[\log q(\boldsymbol{\beta}|\boldsymbol{\lambda})] - \sum_{d=1}^{M} \mathbb{E}_{q(\boldsymbol{z}_d|\boldsymbol{\phi}_d)}[\log q(\boldsymbol{z}_d|\boldsymbol{\phi}_d)] \\
&= \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n};\boldsymbol{\alpha},\boldsymbol{\eta})
\end{aligned}$$

For each value of $(\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})$, we have a variational distribution $q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})$ which gives a lowerbound on the log-likelihood. Since we want to seek the MLE of $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, it is better to maximize the tightest lowerbound. Therefore, we want to find $(\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})^*$ such that

$$(\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})^* = \underset{\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda}}{\arg\max}\, \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n};\boldsymbol{\alpha},\boldsymbol{\eta})$$

The gap of the bound is:

$$\log p(D|\boldsymbol{\alpha},\boldsymbol{\eta}) - \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n};\boldsymbol{\alpha},\boldsymbol{\eta}) = \mathbb{KL}(q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda}), p(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n}|D,\boldsymbol{\alpha},\boldsymbol{\eta}))$$

Given a fixed value of $(\boldsymbol{\alpha},\boldsymbol{\eta})$, the log-likelihood $p(D|\boldsymbol{\alpha},\boldsymbol{\eta})$ is a known value. Therefore, maximizing the lowerbound $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n};\boldsymbol{\alpha},\boldsymbol{\eta})$ w.r.t to $(\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})$ is equivalent to minimizing the $\mathbb{KL}$ divergence. Hence, $q(\boldsymbol{\theta},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M,\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M,\boldsymbol{\lambda})$ can be seen as a variational approximation of the true posterior $p(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z}_1,\ldots,\boldsymbol{z_n}|D,\boldsymbol{\alpha},\boldsymbol{\eta})$.

Variational Dirichlet: Collecting the terms that has γ_d for $d = 1,\ldots,M$, we have

$$\begin{aligned}
\mathcal{L}(\gamma_d) &= \mathbb{E}_{q(\theta_d|\gamma_d)}[\log p(\theta_d \mid \boldsymbol{\alpha})] - \mathbb{E}_{q(\theta_d|\gamma_d)}[\log q(\theta_d \mid \gamma_d)] + \mathbb{E}_{q(\boldsymbol{z}_d|\boldsymbol{\phi}_d)q(\theta_d|\gamma_d)}[\log p(\boldsymbol{z}_d|\theta_d)] \\
&= \mathbb{E}_{q(\theta_d|\gamma_d)}[\log p(\theta_d \mid \boldsymbol{\alpha})] - \mathbb{E}_{q(\theta_d|\gamma_d)}[\log q(\theta_d \mid \gamma_d)] + \sum_{n=1}^{N}\sum_{i=1}^{k} \mathbb{E}_{q(z_{dni}|\phi_{dni})q(\theta_d|\gamma_d)}[\log p(z_{dni}|\theta_{di})]
\end{aligned}$$

Now, I recall some property of the Dirichlet distribution. First, The k-dimensional Dirichlet random variable θ_d ($\sum_{i=1}^{k} \theta_{di} = 1$) has the pdf:

$$p(\theta_d \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_{di}^{\alpha_i - 1}$$

Using the property of the exponential family, namely, the derivative of the log normalization factor with respect to the natural parameter is equal to the expectation of the sufficient statistic. We can have

$$\mathbb{E}[\log \theta_{di} \mid \boldsymbol{\alpha}] = \Psi(\alpha_i) - \Psi\left(\sum_{j=1}^{k} \alpha_j\right)$$

where Ψ is digamma function. Therefore, we have

$$\mathbb{E}_{q(\theta_d|\gamma_d)}[\log p(\theta_d \mid \boldsymbol{\alpha})] = const + \sum_{i=1}^{k} (\alpha_i - 1)\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right)$$

$$\mathbb{E}_{q(\theta_d|\gamma_d)}[\log q(\theta_d \mid \gamma_d)] = \log \Gamma\left(\sum_{j=1}^{k} \gamma_{dj}\right) - \log \Gamma(\gamma_{di}) + \sum_{i=1}^{k} (\gamma_{di} - 1)\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right)$$

And

$$\begin{aligned}\mathbb{E}_{q(z_{dni}|\phi_{dni})q(\theta_d|\gamma_d)}[\log p(z_{dn}|\theta_{di})] &= \mathbb{E}_{q(z_{dni}|\phi_{dni})q(\theta_d|\gamma_d)}[z_{dni} \log \theta_{di}] \\ &= \phi_{dni}\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right)\end{aligned}$$

Hence

$$\begin{aligned}\mathcal{L}(\gamma_d) = &\sum_{i=1}^{k} (\alpha_i - 1)\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right) + \sum_{n=1}^{N}\sum_{i=1}^{k} \phi_{dni}\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right) \\ &- \log \Gamma\left(\sum_{j=1}^{k} \gamma_{dj}\right) + \log \Gamma(\gamma_{di}) - \sum_{i=1}^{k} (\gamma_{di} - 1)\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right) \\ = &\sum_{i=1}^{k}\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right)\left(\alpha_i + \sum_{n=1}^{N} \phi_{dni} - \gamma_{di}\right) - \log \Gamma\left(\sum_{j=1}^{k} \gamma_{dj}\right) + \log \Gamma(\gamma_{di})\end{aligned}$$

The derivative is:

$$\frac{\partial \mathcal{L}}{\partial \gamma_{di}} = \Psi'(\gamma_{di})\left(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_{di}\right) - \Psi'\left(\sum_{j=1}^{k} \gamma_{dj}\right)\sum_{j=1}^{k}\left(\alpha_j + \sum_{n=1}^{N} \phi_{dnj} - \gamma_{dj}\right)$$

Setting the equation to 0, we have

$$\gamma_{di} = \alpha_i + \sum_{n=1}^{N} \phi_{dni} \tag{1}$$

By similar derivation, we can obtain the update equation for the variational parameter λ_{ij} for $i = 1, \ldots, k$ and $j = 1, \ldots, V$:

$$\lambda_{ij} = \eta_i + \sum_{d=1}^{M} \sum_{n=1}^{N} \phi_{dni} w_{dnj} \tag{2}$$

Variational Multinomial: Collecting the terms that has ϕ_{dni} for $d = 1, \ldots, M$, $n = 1, \ldots, N$, and $i = 1, \ldots, k$, we have

$$\begin{aligned}\mathcal{L}(\phi_{dni}) = \mathbb{E}_{q(\boldsymbol{z}_{dni}|\phi_{dni})q(\theta_d|\gamma_d)} \left[\log p\left(z_{dni} \mid \theta_d\right)\right] + \sum_{j=1}^{V} \mathbb{E}_{q(\boldsymbol{z}_{dni}|\phi_{dni})q(\beta_i|\lambda_i)} \left[\log p\left(w_{dnj} \mid z_{dni}, \beta_{ij}\right]\right. \\ \left. - \mathbb{E}_{q(z_{dni}|\phi_{dni})} \left[\log q\left(z_{dni} \mid \phi_{dni}\right)\right]\right.\end{aligned}$$

From the previous part, we have

$$\mathbb{E}_{q(z_{dni}|\phi_{dni})q(\theta_d|\gamma_d)} \left[\log p\left(z_{dn} \mid \theta_d\right)\right] = \phi_{dni} \left(\Psi\left(\gamma_{di}\right) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right)$$

Now we derive

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{z}_{dni}|\phi_{dni})q(\beta_i|\lambda_i)} \left[\log p\left(w_{dnj} \mid z_{dni}, \beta_{ij}\right]\right. &= \mathbb{E}_{q(\boldsymbol{z}_{dni}|\phi_{dni})q(\beta_{ij}|\lambda_{ij})} \left[z_{dni} w_{dnj} \log \beta_{ij}\right] \\ &= \phi_{dni} w_{dnj} \left(\Psi\left(\lambda_{ij}\right) - \Psi\left(\sum_{v=1}^{V} \lambda_{iv}\right)\right)\end{aligned}$$

And

$$\mathbb{E}_{q(z_{dni}|\phi_{dni})} \left[\log q\left(z_{dni} \mid \phi_{dni}\right)\right] = \phi_{dni} \log \phi_{dni}$$

Since w_{dnj} has only one index v such that $w_{dnv} = 1$, adding Lagrange multiplier for the simplex constraint λ_{dn}, we have

$$\begin{aligned}\mathcal{L}(\phi_{dni}) = \phi_{dni} \left(\Psi\left(\gamma_{di}\right) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right) + \phi_{dni} \left(\Psi\left(\lambda_{ij}\right) - \Psi\left(\sum_{v=1}^{V} \lambda_{iv}\right)\right) \\ - \phi_{ni} \log \phi_{ni} + \lambda_{dn} \left(\sum_{i=1}^{k} \phi_{dni} - 1\right).\end{aligned}$$

We have the derivative

$$\frac{\partial L}{\partial \phi_{dni}} = \Psi\left(\gamma_{di}\right) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right) + \Psi\left(\lambda_{ij}\right) - \Psi\left(\sum_{v=1}^{V} \lambda_{iv}\right) - \log \phi_{dni} - 1 + \lambda_{dn}$$

Setting the derivative to 0 (also setting the derivative w.r.t λ_{dn} to 0), we have

$$\phi_{dni} \propto \exp\left(\Psi\left(\gamma_{di}\right) - \Psi\left(\sum_{j=1}^{k}\gamma_{dj}\right) + \Psi\left(\lambda_{ij}\right) - \Psi\left(\sum_{v=1}^{V}\lambda_{iv}\right)\right) \tag{3}$$

Parameter Estimation: For doing maximum log-likelihood, we need to maximize the lowerbound of log-likelihood:

$$(\alpha, \eta)^* = \arg\max_{\alpha,\eta} \mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{z}_1, \ldots, \boldsymbol{z_n}; \boldsymbol{\alpha}, \boldsymbol{\eta}\right)$$

Since it is intractable to have the closed form update of α and η, we use gradient-based methods to update them e.g., Newton method.
We have

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{d=1}^{M}\left(\log\Gamma\left(\sum_{j=1}^{k}\alpha_j\right) - \sum_{i=1}^{k}\log\Gamma\left(\alpha_i\right) + \sum_{i=1}^{k}\left(\left(\alpha_i - 1\right)\left(\Psi\left(\gamma_{di}\right) - \Psi\left(\sum_{j=1}^{k}\gamma_{d_j}\right)\right)\right)\right)$$

Then the gradient and the Hessian w.r.t λ_i are :

$$\frac{\partial L}{\partial \alpha_i} = M\left(\Psi\left(\sum_{j=1}^{k}\alpha_j\right) - \Psi\left(\alpha_i\right)\right) + \sum_{d=1}^{M}\left(\Psi\left(\gamma_{di}\right) - \Psi\left(\sum_{j=1}^{k}\gamma_{d_j}\right)\right)$$

$$\frac{\partial L}{\partial \alpha_i \alpha_j} = \delta(i,j) M \Psi'\left(\alpha_i\right) - \Psi'\left(\sum_{j=1}^{k}\alpha_j\right)$$

Similarly, we derive the gradient and Hessian of $\boldsymbol{\eta}$.

Note: Solving only varitational inference can be seen as a Bayesian approach for LDA. In this case, we need to tune the prior $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$.

1.3 The algorithm of LDA

I summarize the algorithm of LDA in Algorithm 1 which will be used in training LDA experiments.

2 Experiments

In the section, I will conduct experiments on Web Snippets dataset [2] which contains pieces of text to summarize a web page on a search engine results page. In the dataset, 10060 training snippets are labeled into 8 classes including Business (1200), Computers (1200), Culture-Arts-Ent (1880), Education-Science (2360), Engineering (220), Health (880), Politics-Society (1200), and Sports (1120). Moreover, a hold-out test set with 2280 snippets of similar classes of size 300, 300, 330, 300, 150, 300, 300, and 300 in turn.

Algorithm 1 Latent Dirichlet Allocation Algorithm

Initialize $\phi_{dni} = 1/k$ for all d, n, i.
Initialize λ_{ij} for all i, j.
Initialize $\gamma_{d,i}$ for all d, i.
Initialize $\alpha_i = 1/k$ for all i, $\eta_j = 1/V$ for all j.
while $\phi, \lambda, \gamma, \alpha, \eta$ are not converged **do**
  for $d = 1$ to M **do**
    for $n = 1$ to N **do**
      for $i = 1$ to k **do**
        $\phi_{dni} \propto \exp\left(\Psi\left(\gamma_{di}\right) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right) + \Psi\left(\lambda_{ij}\right) - \Psi\left(\sum_{v=1}^{V} \lambda_{iv}\right)\right)$.
      end for
      Normalize ϕ_{dn} to sum to 1.
    end for
    for $i = 1$ to k **do**
      $\gamma_{di} = \alpha_i + \sum_{n=1}^{N} \phi_{dni}$.
    end for
  end for
  for $i = 1$ to k **do**
    for $j = 1$ to V **do**
      $\lambda_{ij} = \eta_i + \sum_{d=1}^{M} \sum_{n=1}^{N} \phi_{dni} w_{dnj}$
    end for
  end for
  Update α, η via gradient-based methods.
end while
Return: $\phi, \lambda, \gamma, \alpha, \eta$

Topic discovery: I first set the number of topic $k = 8$, then fit the LDA model. I show the word-cloud of each topic in Figure 2, namely, I plot words in the vocabulary with the size is proportional to the probability of $E_{q(\beta_i|\lambda_i)}[\beta_i]$ for $i = 1, \ldots, k$. From the figure, we can interpret the topic 1 is Business, topic 2 is Sports, topic 3 is Culture-Arts-Ent, topic 4 is Politics-Society, topic 5 is Engineering/Culture-Arts-Ent , topic 6 is Business/Computers, topic 7 is Engineering, and topic 8 is Health/Education-Science. With the LDA, we can know which words are the most important in each topic that is very useful.

Perplexity and Log-likelihood: The perplexity is computed as:

$$\text{perplexity}\,(D) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p\left(\mathbf{w}_d\right)}{\sum_{d=1}^{M} N_d}\right\}.$$

For perplexity, lower is better. I show the perplexity and the log-likelihood of LDA on the train set and the test set across various values of number of topics $k \in \{4, 8, 10, 16, 20, 100, 200, 400, 1000\}$ in Figure 3. From the figure, we can see that a higher value of k might leads to a better log-likelihood on test set. In contrast, the perplexity might be worse when k increases.

Figure 2: Word-cloud of topics in Web Snippets dataset.

Topic proportion as compressed data: I can also used $E_{q(\theta_d|\gamma_d)}[\theta_d]$ as the compressed data of the original bag of words representation. I fit the logistic regression (one versus rest strategy) with the original train data and with the compressed train data ($k \in \{4, 8, 10, 16, 20, 100, 200, 400, 1000\}$ then measure the classification performance on the test set. For the metrics, I use classification accuracy, macro (micro) precision (recall, f1). The result is given in Figure 4. From the figures, we can observe that the topic model can reduce the dimension from about 10000 to 1000 while only losing about 10% in accuracy and other metrics.

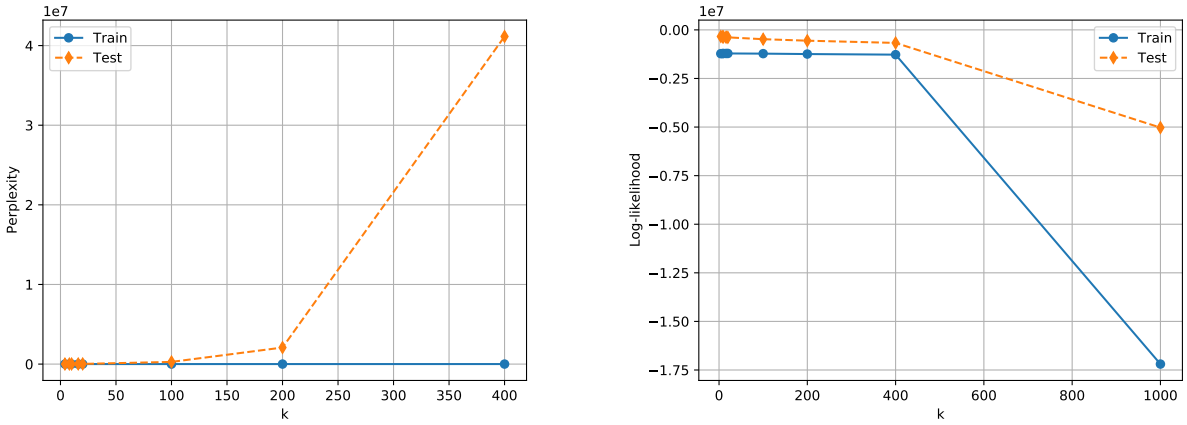

Figure 3: Perplexity and log-likelihood across k.

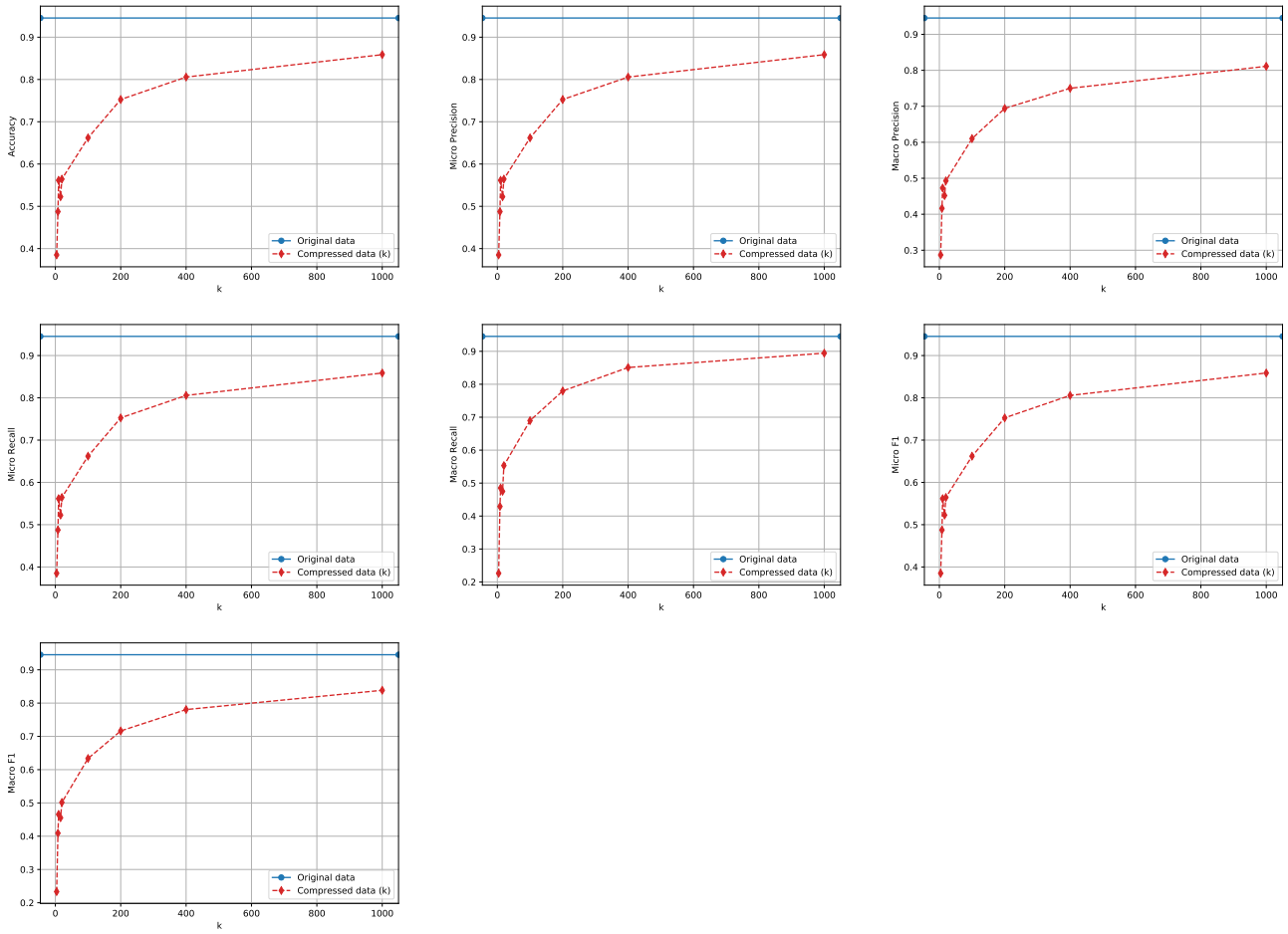

Figure 4: Classification performance of compressed data.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. (Cited on page 1.)

[2] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100, 2008. (Cited on page 6.)