[Paper review 39]

Autoencoding Variational Inference for Topic Models

(Srivastava, Sutton, 2017)

[Contents]

1. Abstract

Topic Models

- popular method for **learning representations of text**
- challenges : change to topic model... require new inference algorithm

  $\rightarrow$ **AEVB** : to address the problem... but impractical!

2. Background

2.1 LDA

- most popular topic model
- each document = mixture of topics
  - each topic (β_k)= probability distribution over vocabulary

    (matrix $\beta = (\beta_1, \ldots, \beta_K)$)
- marginal likelihood of a document $\mathbf{w}$:

  $p(\mathbf{w} \mid \alpha, \beta) = \int_\theta \left(\prod_{n=1}^N \sum_{z_n=1}^k p\left(w_n \mid z_n, \beta\right) p\left(z_n \mid \theta\right)\right) p(\theta \mid \alpha) d\theta.$
- posterior inference over θ and z is intractable.....

  (due to coupling between those two, under multinomial assumption)
- summary

```
for each document w do
    Draw topic distribution θ ~ Dirichlet(α);
    for each word at position n do
        Sample topic z_n ~ Multinomial(1, θ);
        Sample word w_n ~ Multinomial(1, β_{z_n});
    end
end
```

2.2 Mean Field and AEVB

Mean Field

- assumption : independency between latent variables
- break coupling between θ and z

  by introducing **free variational parameters** : γ over θ, ϕ over z
- $q(\theta, z \mid \gamma, \phi) = q_\gamma(\theta) \prod_n q_\phi\left(z_n\right).$

- minimize (negative) ELBO :

  $L(\gamma, \phi \mid \alpha, \beta)=D_{KL}[q(\theta, z \mid \gamma, \phi) \| p(\theta, z \mid \mathbf{w}, \alpha, \beta)]-\log p(\mathbf{w} \mid \alpha, \beta)$
- will call DMFVI (Decoupled Mean-Field Variational inference)

under MF...

- pros) have closed form! (∵ conjugacy of Dirichlet & Multinomial distn)
- cons) limits its flexibility

AEVB

(Auto Encoding Variational Bayes)

- **black box** inference models
- write ELBO as...

  $L(\gamma, \phi \mid \alpha, \beta)=-D_{KL}[q(\theta, z \mid \gamma, \phi) \| p(\theta, z \mid \alpha)]+\mathbb{E}_{q(\theta, z \mid \gamma, \phi)}[\log p(\mathbf{w} \mid z, \theta, \alpha, \beta)].$
  - 1st term) match variational posterior to the prior
  - 2nd term) reconstruction term
- variational parameters are computed using NN, called **inference network**

  ex) choose Gaussian variational distribution $q_\gamma(\theta)=N(\theta ; \mu(\mathbf{w}), \operatorname{diag}(\mathbf{v}(\mathbf{w})))$
- unlike DMFVI, have coupled the variational parameters!

  (since they are computed from same NN)
- use reparameterization trick

3. AEVB in LDA

practical challenges in applying AEVB to topic models

3-1. Collapsing z's

z can be summed out

$p(\mathbf{w} \mid \alpha, \beta)=\int_\theta\left(\prod_{n=1}^N p\left(w_n \mid \beta, \theta\right)\right) p(\theta \mid \alpha) d \theta.$

- $w_n \mid \beta, \theta$ is Multinomial $(1, \beta \theta)$
- β : matrix of all topic-word probability vectors

3-2. Working with Dirichlet Beliefs : Laplace Approximation

topic proportion θ : **Dirichlet prior**

- difficult to handle Dirichlet in AEVB

  (∵ reparam trick works for Gaussian distn)
- ∴ use Laplace Approximation to the Dirichlet prior

Laplace approximation

- do it in "softmax basis" instead of simplex
- $P(\theta(\mathbf{h}) \mid \alpha)=\frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma\left(\alpha_k\right)} \prod_k \theta_k^{\alpha_k} g\left(\mathbf{1}^T \mathbf{h}\right)$
  - $\theta=\sigma(\mathbf{h})$ where $\sigma(.)$ is softmax function
- results in a distribution over softmax variables $\mathbf{h}$ with..

- mean μ_1 : $\mu_{1k}=\log \alpha_k-\frac{1}{K} \sum_i \log \alpha_i$
- cov Σ_1 : $\Sigma_{1kk}=\frac{1}{\alpha_k}\left(1-\frac{2}{K}\right)+\frac{1}{K^2} \sum_i \frac{1}{\alpha_k}$.

- approximate $p(\theta \mid \alpha)$ in the simplex basis, with $\hat{p}\left(\theta \mid \mu_1, \Sigma_1\right)=\mathcal{LN}\left(\theta \mid \mu_1, \Sigma_1\right)$
  - $\mathcal{LN}$: logistic normal distribution, with params μ_1 and Σ_1
  - diagonal covariance matrix

3-3. Variational Objective

2 inference networks : f_μ and f_Σ with parameters δ

- output of each network is a vector in $\mathbb{R}^K$

for document $\mathbf{w}$, define $q(\theta)$ to be $\mathcal{LN}$

- mean : $\mu_0=f_\mu(\mathbf{w}, \boldsymbol{\delta})$
- diag cov : $\Sigma_0=\operatorname{diag}\left(f_{\Sigma}(\mathbf{w}, \boldsymbol{\delta})\right)$

can sample using **reparam trick + laplace approximation**

- $\theta=\sigma\left(\boldsymbol{\mu}_0+\boldsymbol{\Sigma}_0^{1/2} \boldsymbol{\epsilon}\right)$, where $\epsilon \sim \mathcal{N}(0, I)$

ELBO :

$$L(\boldsymbol{\Theta})=\sum_{d=1}^D\left[-\left(\frac{1}{2}\left\{\operatorname{tr}\left(\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_0\right)+\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0\right)-K+\log \frac{\left|\boldsymbol{\Sigma}_1\right|}{\left|\boldsymbol{\Sigma}_0\right|}\right\}\right)+\mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)}\left[\mathbf{w}_d^{\top} \log \left(\sigma(\boldsymbol{\beta}) \sigma\left(\boldsymbol{\mu}_0+\boldsymbol{\Sigma}_0^{1/2} \boldsymbol{\epsilon}\right)\right)\right]\right]$$

- $\boldsymbol{\Theta}$: set of all the model and variational parameters
- $\mathbf{w_1} \ldots \mathbf{w_D}$: documents in corpus
- 1st term) KL divergence between two $\mathcal{LN}$

  2nd term) reconstruction error

3-4. Training and Practical Considerations : Dealing with Component Collapsing

AEVB is prone to component collapsing

- reason) as latent dim increases, KL reg (1st term in ELBO) dominates ELBO

  (decoder weights collapse, close to prior & do not show posterior divergence)
- solve) Adam + high moment weight (=β_1) and learning rate η

  ($\rightarrow$ early peaks in functional spaces can be easily avoided)
- other solutions) batch norm, drop out, down-weight the KL term...

4. ProdLDA : LDA with Products of Experts

$p(\mathbf{w} \mid \theta, \beta)$: mixture of multinomials

- problem : no sharper than the components being mixed!

  how to solve...?

Solution : replace **word-level mixture** with **weighted product of experts**

$\rightarrow$ drastic improvement in topic coherence

4-1. Model

ProdLDA

- LDA where word-level mixture over topics

  (that is, topic matrix is not constrained to exist in multinomial simplex prior to mixing)
- LDA +
  - (1) β is unnormalized
  - (2) w_n is defined as $w_n \mid \beta, \theta \sim$ Multinomial $(1, \sigma(\beta\sigma))$

The connection to a product of experts is straightforward, as for the multinomial, a mixture of natural parameters corresponds to a weighted geometric average of the mean parameters. That is, consider two N dimensional multinomials parametrized by mean vectors $\mathbf{p}$ and $\mathbf{q}$.

$P(\mathbf{x} \mid \delta \boldsymbol{r} + (1-\delta)\boldsymbol{s}) \propto \prod_{i=1}^{N} \sigma(\delta r_i + (1-\delta)s_i)^{x_i} \propto \prod_{i=1}^{N} \left[r_i^{\delta} \cdot s_i^{(1-\delta)} \right]^{x_i}$.

- $\mathbf{p} = \sigma(\boldsymbol{r})$.
- $\mathbf{q} = \sigma(\boldsymbol{s})$.

5. Discussion and Future Work

Problem of AEVB + LDA … difficult to train because of

- 1) **Dirichlet prior**
- 2) **component collapsing problem**

Present **blackbox inference** method for topic models