Projection
Var down
Point estimation은 가능
credible interval X

Variational Bayes

Gwangsu Kim

JBNU

Second semester of 2024

Approximation of Posterior I

- Basically, the posterior sampling requires heavy computations, especially for the big data
- To alleviate this problem, the variational Bayes arises
  1. We let $p(z_1, \ldots, z_m \mid x) \approx q(\boldsymbol{z}) = \prod_{i=1}^{m} q_i(z_i)$.
  2. Finding the best q minimizing

$$\mathrm{KL}(q(\boldsymbol{z}) \| p(\boldsymbol{z} \mid \boldsymbol{x}))$$

  where $\boldsymbol{z} = (z_1, \ldots, z_m)$.

ELBO I

- (ELBO) Evidence of Lower BOund

- KL (Kullback-Leibler) divergence

$$\mathrm{KL}(f\|g) = \int \log \frac{f(z)}{g(z)} f(z) d\mu(z)$$

- Properties

1. For any f and g, $\mathrm{KL}(f\|g) \geq 0$
2. $\mathrm{KL}(f\|g) \neq \mathrm{KL}(g\|f)$
3. $\mathrm{KL}(f\|g) = 0 \Longleftrightarrow f = g \ (a.e.\ \mu)$

ELBO II

- Approximating q

$$q^*(z) = \text{argmin}_{q \in \mathcal{F}} \text{KL}\left(q(z) \| p(z \mid \boldsymbol{x})\right)$$

- We have the follows:

$$\begin{aligned}
\text{KL}(q(z) \| p(z \mid \boldsymbol{x})) &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z \mid \boldsymbol{x})] \\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z, \boldsymbol{x})] + \log p(\boldsymbol{x}) \\
\log p(\boldsymbol{x}) &= \mathbb{E}_q[\log p(z, \boldsymbol{x})] - \mathbb{E}_q[\log q(z)] \\
&\quad + \text{KL}(q(z) \| p(z \mid \boldsymbol{x}))
\end{aligned}$$

ELBO III

- Here, the $p(\boldsymbol{x})$ does not depend on q, and KL divergence decrease when the $\mathbb{E}_q[\log p(z, \boldsymbol{x})] - \mathbb{E}_q[\log q(z)]$ increases.
- Rearrange for ELBO

$$\log p(\boldsymbol{x}) \geq \text{ELBO} = \mathbb{E}_q[\log p(z, \boldsymbol{x})] - \mathbb{E}_q[\log q(z)]$$

  - It is the lower bound of $\log p(\boldsymbol{x})$.

ELBO IV

- Final formula for the ELBO

$$\text{ELBO} = \mathbb{E}_q[\log p(\boldsymbol{x} \mid z)] - \text{KL}(q(z) \| p(z))$$

Mean-field Approximation and Bayesian Mixtures I

- The q for the Bayesian mixtures

$$q(\boldsymbol{\mu}, \boldsymbol{c}) = \prod_{i=1}^{K} q(\mu_k \mid m_k, s_k^2) \prod_{i=1}^{n} q(c_i \mid \xi_i)$$

where c_i is the configuration assigning to components of mixture.

Mean-field Approximation and Bayesian Mixtures II

- Mean-field approximation: given the variation density $q(\boldsymbol{z}) = \prod_{j=1}^{m} q_j(z_j)$

$$q_j(z_j) \sim \exp\left(\mathbb{E}_{-j}\left[\log p(z_j \mid \boldsymbol{z}_{-j}, \boldsymbol{x})\right]\right)$$

- Derivation

$$\text{ELBO}(q_j) = \mathbb{E}_j \mathbb{E}_{-j} \log p(z_j, \boldsymbol{z}_{-j}, \boldsymbol{x}) - \mathbb{E}_j \log q_j(z_j) \tag{1}$$

Mean-field Approximation and Bayesian Mixtures III

- By careful observation, we can validate the equation (1) is the negative KL divergence between $C\exp\left(\mathbb{E}_{-j}\log p(z_j, \boldsymbol{z}_{-j}, \boldsymbol{x})\right)$ and q_j.
- It implies that $q_j^*(z_j) \propto \exp\left(\mathbb{E}_{-j}\log p(z_j, \boldsymbol{z}_{-j}, \boldsymbol{x})\right) \propto \exp\left(\mathbb{E}_{-j}\log p(z_j \mid \boldsymbol{z}_{-j}, \boldsymbol{x})\right)$.

Mean-field Approximation and Bayesian Mixtures IV

- Application to Bayesian mixture

$$q^*(c_i \mid \xi_i) \quad \propto \quad \exp\left\{\log p(c_i) + \mathbb{E}[\log p(x_i \mid c_i, \boldsymbol{\mu}; \boldsymbol{m}, \boldsymbol{s}^2)]\right\}$$

Note that $p(x_i \mid c_i, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(x_i \mid \mu_k)^{c_{ik}}$ where c_i is the one-hot vector.

Mean-field Approximation and Bayesian Mixtures V

1) Calculations of cluster assignment

$$
\begin{aligned}
\mathbb{E}[\log p(x_i \mid c_i, \boldsymbol{\mu})] &= \sum_k c_{ik}\mathbb{E}[p(x_i \mid \mu_k; \boldsymbol{m}, \boldsymbol{s}^2)] \\
&= \sum_k c_{ik}\mathbb{E}[-(x_i - \mu_k)^2/2; \boldsymbol{m}, \boldsymbol{s}^2) + const. \\
&= \sum_k c_{ik}\left\{\mathbb{E}[\mu_k; \boldsymbol{m}, \boldsymbol{s}^2]x_i - \mathbb{E}[\mu_k^2/2; \boldsymbol{m}, \boldsymbol{s}^2]\right\} + const.
\end{aligned}
$$

where c_{ik} is an indicator vector

Mean-field Approximation and Bayesian Mixtures VI

- Therefore, we update

$$\xi_{ik} = \exp\left\{\mathbb{E}[\mu_k; \boldsymbol{m}, \boldsymbol{s}^2]x_i - \mathbb{E}[\mu_k^2/2; \boldsymbol{m}, \boldsymbol{s}^2]\right\}$$

2) Calculations of component mean

$$q(\mu_k) \propto \exp\left(\log p(\mu_k) + \sum_{i=1}^{n} \mathbb{E}[p(x_i \mid c_i, \boldsymbol{\mu}; \xi_i, \boldsymbol{m}_{-k}, \boldsymbol{s}_{-k}^2]\right)$$

Mean-field Approximation and Bayesian Mixtures VII

$$
\begin{aligned}
\log q(\mu_k) &= \log p(\mu_k) + \sum_{i=1}^{n} \mathbb{E}[p(x_i \mid c_i, \boldsymbol{\mu}; \xi_i, \boldsymbol{m}_{-k}, \boldsymbol{s}^2_{-k})] + const. \\
&= -\mu_k^2/2\sigma^2 + \sum_i \mathbb{E}[c_{ik} \mid \xi_i] \log p(x_i \mid \mu_k) + const. \\
&= -\mu_k^2/2\sigma^2 + \sum_i \mathbb{E}[c_{ik} \mid \xi_i] \log p(x_i \mid \mu_k) + const. \\
&= -\mu_k^2/2\sigma^2 + \sum_i \xi_{ik}(-(x_i - \mu_k)^2/2) + const. \\
&= \sum_i \xi_{ik} x_i \mu_k - \left(1/\sigma^2 + \sum_k \xi_{ik}\right) \mu_k^2/2 + const.
\end{aligned}
$$

Mean-field Approximation and Bayesian Mixtures VIII

- It implies that

$$m_k = \frac{\sum_[\xi_{ik} x_i}{1/\sigma^2 + \sum_i \xi_{ik}} \text{ and } s_k = \frac{1}{1/\sigma^2 + \sum_i \xi_{ik}}.$$

■ Update the q by the sequential procedure

Autoencoder I

- Challenging point: if we do not know the $p(\boldsymbol{x} \mid z)$ exactly or too complicated, then the posterior is difficult to be obtained
  1. when the mean of normal is determined by the deep neural networks
  2. The explicit form of density is not available
  3. A large dataset

Autoencoder II

- Key idea
  1. Parameterized ELBO

$$\ell(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)}) = \mathbb{E}_{q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)} \mid \boldsymbol{z})] - \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}) \| p_{\boldsymbol{\theta}}(\boldsymbol{z}))$$

  2. Gradient to the expectation with an approximation using samples

$$\nabla_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}}[f(\boldsymbol{z})] \approx \nabla_{\boldsymbol{\phi}} \frac{1}{L} \sum_{l=1}^{L} f(\boldsymbol{z}^{(l)}) \text{ where } \boldsymbol{z}^{(l)} = g(\boldsymbol{\phi}, \boldsymbol{\epsilon})$$

Autoencoder III

- The re-parameterization can be done by $g(\boldsymbol{\phi}, \boldsymbol{\epsilon}^{(l)}) = \boldsymbol{s} \odot \boldsymbol{\epsilon}^{(l)} + \boldsymbol{\mu}$ where $\boldsymbol{\phi} = (\boldsymbol{\mu}, \boldsymbol{s})$
- This implies the gradient to the expectation wrt $\boldsymbol{\phi}$ can be accomplished by the following:

$\mu + \sigma$

$$\frac{\partial f(\boldsymbol{z}^{(l)})}{\partial \boldsymbol{\mu}} = \frac{\partial f(\boldsymbol{z}^{(l)})}{\partial \boldsymbol{z}^{(l)}} \mathbf{1}$$

$$\frac{\partial f(\boldsymbol{z}^{(l)})}{\partial \boldsymbol{s}} = \frac{\partial f(\boldsymbol{z}^{(l)})}{\partial \boldsymbol{z}^{(l)}} \boldsymbol{\epsilon}^{(l)}$$

$\mu + \sigma \odot \epsilon$

Autoencoder IV

- Gradient w.r.t. $\boldsymbol{\theta}$ can be obtained by the conventional SGD

Autoencoder V

- What's the innovation?
  1. The explicit (partial) calculation of the expectation is not required q에 대해 크게 고민할 필요가 없다
  2. In fact, we can consider all cases of likelihood by the combining the sampling scheme
  3. The merit of this approach is the flexibility in using the auto encoder-decoder architectures

단점 : 정규분포, t를 빼고 못씀
Sharpness가 약함

요새는 diffusion 끝

Autoencoder VI

autoencoder 논문보기
Kingma, Maxwell

- What's the drawbacks?
  1. Using the r.v. can have effects of smoothing
  2. The normal distribution can be extended to t-distribution., but the above extension is not easy