POSTECH
INDUSTRIAL AND MANAGEMENT ENGINEERING, POSTECH

Variational inference: A review for statisticians

포항공과대학교 산업경영공학과

Stochastic Systems Lab

Jongwon Kim

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." *Journal of the American statistical Association* 112.518 (2017): 859-877.

STOCHASTIC SYSTEMS LAB.

Contents

1. Problem statement
2. Variational inference and MCMC
3. Application
    1. Scalable Variational Gaussian Process

STOCHASTIC SYSTEMS LAB.

Problem statement

- **Notation**

$x \in \boldsymbol{\Omega}$: Input

$z \in \mathcal{L}$: Latent input

$y \in \mathbb{R}$: Output

$D = \{(x_i, y_i) | i = 1, \dots, n\}$: Observation set

$\theta \in \boldsymbol{\Theta}$: Model parameters

Vectors are column vectors by default

Capital letters refer to matrix

- $\mathrm{X} = (x_1, x_2, \dots, x_n)^T \ where \ x_i \in \boldsymbol{\Omega}$

Problem statement

Inference problem

The inference problem is to **compute the conditional density (posterior distribution) $\boldsymbol{p(\theta|D)}$** of the model parameters given the observation.

We can write the conditional density as

$$\boldsymbol{p(\theta|D)} = \frac{p(\theta, D)}{p(D)} = \frac{p(D|\theta)p(\theta)}{p(D)}$$

The denominator contains the marginal density of the observation, also called the evidence.

We calculate it by marginalizing out the latent variables from the joint density.

The **evidence $\boldsymbol{p(D)}$ is what we need to compute**.

$$p(D) = \int p(\theta, D) d\theta$$

Problem statement

Inference problem

In deterministic model, we should get the fixed model parameters θ.

So in this case, we estimate the model parameters $\boldsymbol{\hat{\theta}}$ from the MLE, and the predictive model is given as follows:

$$\boldsymbol{y'} = f_{\boldsymbol{\hat{\theta}}}(x') \; where \; \boldsymbol{\hat{\theta}} \; is \; estimated \; by \; MLE.$$

In probabilistic model (Bayesian model)

$$\boldsymbol{p}(\boldsymbol{y'}|\boldsymbol{x'}, \boldsymbol{D}) = \int p(y'|x', D, \theta) \; \boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{D}) \; d\theta$$

$$= \int p(y'|x', \theta) \; \boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{D}) \; d\theta \; (\because D \text{ tells no more than what } \theta \text{ does})$$

To get the predictive distribution, we should calculate the posterior distribution or generate samples.

Variational inference and MCMC

MCMC(Markov chain Monte Carlo) algorithms approximate the posterior **with samples** from **the chain.**

MCMC generates $\theta_1, \theta_2, \ldots, \theta_n \sim r.v\ with\ p(\theta|D)$. (with abuse of notation)

$$p(y'|x', D) = \int p(y'|x', \theta)\ \boldsymbol{p(\theta|D)}\ d\theta = \frac{1}{n}\sum_{i=1}^{n} p(y'|x', \theta_i)$$

Variational inference algorithms solve an **optimization** problem to get the tractable proxy model.

We utilizes KL divergence to obtain an similar proxy distribution q_τ fully defined by τ from some function space $\mathcal{F}$.

$$q_\tau^* = \min_{q_\tau \in \mathcal{F}} KL(q_\tau(\theta), p(\theta|D))$$

$$p(y'|x', D) = \int p(y'|x', \theta)\ \boldsymbol{q_\tau^*(\theta)}\ d\theta$$

Variational inference and MCMC

Variational inference

One of the key ideas behind variational inference is to define a family of proxy distribution $\mathcal{F}$ to be flexible enough to capture a density close.

To $p(\theta|D)$, but simple enough for efficient optimization.

In usual, instead of minimization of KL divergence, we solve the maximization problem of ELBO (Evidence lower bound)

$$q_\tau^* = \min_{q_\tau \in \mathcal{F}} KL(q_\tau(\theta), p(\theta|D)) = \min_{q_\tau \in \mathcal{F}} \log p(D) - \mathrm{E}_{q_\tau}[\log p(\theta, D)] + \mathrm{E}_{q_\tau}[\log q(\theta)]$$
$$= \max_{q_\tau \in \mathcal{F}} \mathrm{E}_{q_\tau}[\log p(\theta, D)] - \mathrm{E}_{q_\tau}[\log q(\theta)] = \max_{q_\tau \in \mathcal{F}} ELBO(q_\tau)$$

We can decompose the prior distribution into as follows:

$$p(D) = ELBO(q_\tau) + KL(q_\tau(\theta), p(\theta|D))$$

Variational inference and MCMC

Theoretical guarantee

Since variational inference easily takes advantage of methods like stochastic optimization, distributed optimization (though some MCMC can also exploit these innovations)

MCMC is suited to smaller data sets and scenarios where we happily pay a heavier computational cost for more precise samples.

Moreover, traditionally, MCMC has guarantee for convergence but variational inferenence does not. In these days, there is some research to show the guarantee of variational inference for convergence.

Yang, Yun, Debdeep Pati, and Anirban Bhattacharya. "$\alpha $-variational inference with statistical guarantees." *The Annals of Statistics* 48.2 (2020): 886-905.
Zhang, Fengshuo, and Chao Gao. "Convergence rates of variational posterior distributions." *The Annals of Statistics* 48.4 (2020): 2180-2207.

Scalable Variational Gaussian Process

Scalable Variational Gaussian Process regression

1. Using $\boldsymbol{m < n}$ **inducing point**, U, we replace the exact kernel, K by Nyström approximation.

$$(\boldsymbol{K_{XX}}) \approx K_{XU} K_{UU} K_{UX}$$

2. In inducing point method, we need to **approximate posterior distribution**

$$\boldsymbol{p(y|u)} = \log\left(E_{p(f|u)}[p(y|f)]\right) \geq E_{p(f|u)}\left[\log\left(p(y|f)\right)\right] \approx \boldsymbol{\tilde{p}(y|u)}$$

3. They introduce a **variational distribution** $\boldsymbol{q}$ to approximate this distribution.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In A. Nicholson and P. Smyth, editors, Uncertainty in Artificial Intelligence, volume 29. AUAI Press, 2013r

Scalable Variational Gaussian Process

Gaussian process classification

In the regression setting, when the **likelihood is Gaussian**, inference can be performed in closed-form using linear algebra.

$$y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma_n^2 I)$$

$$p(y|x,\theta) = Likelihood = p(y|f(x)) \sim N(f(x), \sigma_n^2 I)$$

When the **likelihood is non-Gaussian**, such as in GP classification, the **posterior and marginal likelihood must be approximated.**

$$p(y|f(x)) \sim Bernoulli\ distribution$$

In this case, approximation step is replaced into variational inference.

Thank you