Variational Inference: an introduction

Massil Achab

May 18, 2017

This presentation is inspired by the very insightful review from Blei et al. [2017].

1 Inference of probabilistic models

A probabilistic models asserts how observation arise from a natural phenomenon.

We design the model via a *joint distribution*

$$p(\boldsymbol{x}, \boldsymbol{z})$$

of observed variables $\boldsymbol{x}$ corresponding to data, and latent variables $\boldsymbol{z}$ that provide the hidden structure to generate from $\boldsymbol{x}$.

The *likelihood*

$$p(\boldsymbol{x}|\boldsymbol{z})$$

is a probability that describes how any data $\boldsymbol{x}$ is likely given a particular hidden pattern described by $\boldsymbol{z}$.

The *prior*

$$p(\boldsymbol{z})$$

posits a generating process of the hidden structure.

Inference amounts to conditioning on data and computing the *posterior*

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{\int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}.$$

This last distribution is difficult to compute because of the normalizing constant. Such computation often requires *approximate inference*. For decades, MCMC was the method of choice to do approximate inference. MCMC methods construct an ergodic Markov chain on $\boldsymbol{z}$ with $p(\boldsymbol{z}|\boldsymbol{x})$ as stationary distribution. Another approach, Variational Inference (VI) approximates the posterior $p(\boldsymbol{z}|\boldsymbol{x})$ though optimization and tends to be faster than MCMC sampling.

2 Variational Inference

2.1 Idea behind Variational Inference

Core idea of VI:

- posit a family of distribution $q(\boldsymbol{z}) \in \mathcal{D}$

- match $q(\boldsymbol{z})$ to the posterior $p(\boldsymbol{z}|\boldsymbol{x})$

This strategy converts the problem of computing the posterior $p(\boldsymbol{z}|\boldsymbol{x})$ into an optimization problem of the form:

$$q^*(\boldsymbol{z}) = \arg\min_{q(\boldsymbol{z})\in\mathcal{D}} \text{divergence}(q(\boldsymbol{z}), p(\boldsymbol{z}|\boldsymbol{x})). \tag{1}$$

The optimized $q^*(\boldsymbol{z})$ is then used as a proxy of $p(\boldsymbol{z}|\boldsymbol{x})$.

MCMC and VI are different approaches for solving the same problem. Comparison:

- MCMC asymptotically provides exact samples from the posterior distribution.
- VI faster and can be parallelized.
- VI performs well on mixture models.

2.2 Evidence Lower BOund

The usual criterion to match $q(\boldsymbol{z})$ to $p(\boldsymbol{z}|\boldsymbol{x})$ uses the Kullback-Leibler divergence. We now minimize:

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

The *evidence* is defined as $p(\boldsymbol{x}) = \int p(\boldsymbol{x},\boldsymbol{z})d\boldsymbol{z}$. Since the evidence does not depend on $q(z)$, we optimize an alternative objective:

$$\mathrm{ELBO}(q) = \mathbb{E}^q[\log p(\boldsymbol{x},\boldsymbol{z})] - \mathbb{E}^q[\log q(\boldsymbol{z})]. \tag{2}$$

Maximizing the ELBO function is equivalent to minimize the Kullback-Leibler divergence above. The name **ELBO** stands for **E**vidence **L**ower **BO**und because of the following inequality:

$$\begin{aligned}\log p(\boldsymbol{x}) &= \mathrm{KL}(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) + \mathrm{ELBO}(q) \\ &\geq \mathrm{ELBO}(q),\end{aligned}$$

since Kullback-Leibler divergence always takes positive values.

2.3 Mean-field variational family

A popular choice of family is the *mean-field variational family* where the latent variables are mutually independent. A generic member writes:

$$q(\boldsymbol{z}) = \prod_{j=1}^{m} q_j(z_j) \tag{3}$$

Let emphasize that the variational family is not a model of the observed data.

The mutual independence of the latent variables triggers the separability of the ELBO's second term:

$$\mathbb{E}^q[\log q(\boldsymbol{z})] = \sum_{j=1}^{m} \mathbb{E}^q[\log q_j(z_j)]$$

2.4 CAVI: Coordinate Ascent mean-field Variational Inference

We have now cast the approximate conditional inference as an optimization problem. The previous remark on the separability incites the use of a coordinate ascent algorithm. Rewriting ELBO as a function of q_j, we prove that the optimal density q_j^* satisfies

$$q_j^*(z_j) \propto \exp\{\mathbb{E}_{-j}^q[\log p(z_j|\boldsymbol{z}_{-j}, \boldsymbol{x})]\} \tag{4}$$

3 Application: Bayesian Mixture of Gaussians

We consider K mixture components with means $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$ drawn from $\mathcal{N}(0, \sigma^2)$. The full hierarchical model writes

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2) && k = 1, \ldots, K, \\ c_i &\sim \text{Cat}(1/K, \ldots, 1/K) && i = 1, \ldots, n, \\ x_i|c_i, \boldsymbol{\mu} &\sim \mathcal{N}(c_i^\top \boldsymbol{\mu}, 1) && i = 1, \ldots, n, \end{aligned}$$

where $c_i \in \mathbb{R}^K$ is a one-hot encoded vector that assigns the latent class to x_i.

In this application, $\boldsymbol{z} = (\boldsymbol{c}, \boldsymbol{\mu})$. Following the previously detailed steps, we efficiently obtain an approximation of the posterior density $p(\boldsymbol{z}|\boldsymbol{x})$. See Section 3 in Blei et al. [2017] for more details.

References

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted), 2017.