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

\[ p(x, z) \]

of observed variables \( x \) corresponding to data, and latent variables \( z \) that provide the hidden structure to generate from \( x \).

The likelihood

\[ p(x|z) \]

is a probability that describes how any data \( x \) is likely given a particular hidden pattern described by \( z \).

The prior

\[ p(z) \]

posits a generating process of the hidden structure.

Inference amounts to conditioning on data and computing the posterior

\[ p(z|x) = \frac{p(x,z)}{\int p(x,z)dz}. \]

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 \( z \) with \( p(z|x) \) as stationary distribution. Another approach, Variational Inference (VI) approximates the posterior \( p(z|x) \) thought 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 \( q(z) \in D \)
• match \( q(z) \) to the posterior \( p(z|x) \)

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

\[
q^*(z) = \arg\min_{q(z) \in D} \text{divergence}(q(z), p(z|x)).
\]  

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

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 \( q(z) \) to \( p(z|x) \) uses the Kullback-Leibler divergence. We now minimize:

\[
\text{KL}(q(z)||p(z|x)) = \mathbb{E}^q[\log q(z)] - \mathbb{E}^q[\log p(z|x)]
= \mathbb{E}^q[\log q(z)] - \mathbb{E}^q[\log p(x, z)] + \log p(x)
\]

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

\[
\text{ELBO}(q) = \mathbb{E}^q[\log p(x, z)] - \mathbb{E}^q[\log q(z)].
\]  

Maximizing the ELBO function is equivalent to minimize the Kullback-Leibler divergence above. The name \text{ELBO} stands for \text{Evidence Lower BOund} because of the following inequality:

\[
\log p(x) = \text{KL}(q(z)||p(z|x)) + \text{ELBO}(q) 
\geq \text{ELBO}(q),
\]

since Kullback-Leibler divergence always takes positive values.

### 2.3 Mean-field variational family

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

\[
q(z) = \prod_{j=1}^{m} q_j(z_j)
\]  

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:

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

$$q^*_j(z_j) \propto \exp\{E_{-j}[\log p(z_j|x, z_{-j})]\}$$

(4)

3 Application: Bayesian Mixture of Gaussians

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

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

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

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

References