INTRODUCTION TO VARIATIONAL INFERENCE

WON I. LEE

1. Introduction

This article is dedicated to providing an introduction to **variational inference (VI)**. It will closely follow the presentation of [1], but seeks to provide a more condensed, to-the-point exposition for those interested in quickly diving into the subject area.

Variational inference is concerned with approximating the conditional density of latent variables $z = z_{1:m}$ given the observations $x = x_{1:n}$. For the purposes of illustration, we focus on the case of Bayesian analysis, in which $z = \theta$ are the parameters of interest. This conditional density is given by:

$$p(z|x) = \frac{p(z,x)}{p(x)} = \frac{p(x|z)p(z)}{p(x)}$$

which is the familiar Bayes' rule for computing the posterior of the parameters given the observations.

Often, computing the exact posterior density is intractable, since computing the marginal (or evidence) $p(x) = \int p(z,x)dz$ is computational prohibitive. Consequently, we often resort to techniques such as Markov chain Monte Carlo (MCMC) to use samples to approximate the posterior. VI is an alternative methodology.

2. Problem Formulation

To set up the problem, we consider a family of densities over the latent variables:

$$\mathcal{D} = \{q \in P(z)\}$$

where $P(z)$ is the set of all probability distributions over z. Note that $p(z|x)$ is a particular member of $P(z)$, since it is a distribution over the latent variables. The problem is then to find the "optimal" candidate $q^*(z) \in \mathcal{D}$ in the sense of KL divergence:

$$q^*(z) \equiv \arg\min_{q \in \mathcal{D}} KL(q(z) \| p(z|x))$$

However, this problem is intractable for the same reason that the original conditional is not computable; by definition of the KL divergence, the optimization problem becomes:

$$KL(q(z)\|p(z|x)) = E[\log q(z)] - E[\log p(z,x)] + \log p(x)$$

where all expectations are taken with respect to z. But we immediately note that $\log p(x)$ is constant with respect to z, so that for the purposes of finding the optimal density over the latent variables, we can simply ignore this term. To minimize KL divergence, we equivalently maximize the negative of divergence, so ignoring the marginal term, we obtain the **evidence lower bound (ELBO)**:

$$ELBO(q) \equiv E[\log p(z,x)] - E[\log q(z)]$$

Note that we can gain intuition for the optimization problem by considering the following decomposition of the ELBO:

$$ELBO(q) = E[\log p(z)] + E[\log p(x|z)] - E[\log q(z)] = E[\log p(x|z)] - KL(q(z)\|p(z))$$

Thus, in order to maximize the ELBO, we seek densities over z that place significant mass on values of z that "best explain" the observed data x (the first term) while simultaneously being relatively close to the prior over z (the second term).

3. Mean-Field Variational Family

In order to complete the VI problem specification, we must designate a family of densities $\mathcal{D}$ that we are optimizing over. One particular but popular family of densities is the **mean-field variational family**, in which each latent variable is treated as mutually independent random variables. Thus, each member of this family of densities allows the decomposition:

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

Note that this family of densities is a considerable simplification of any reasonable distribution on the latent variables, as it does not allow any covariance or interaction between the latent variables.

With this specification of the family of approximate distributions, we now have the complete problem formulation:

$$\max_{q \in \mathcal{D}} ELBO(q)$$

$$\text{where } \mathcal{D} = \{q = \prod_{j=1}^{m} q_j(z_j) \in P(z)\}$$

While we have been writing the optimization as over $q \in \mathcal{D}$, in most cases the family $\mathcal{D}$ will be parametrized by a few variables τ, so it would really be proper to write the optimization over τ.

For example, in the canonical Normal-Normal conjugate Bayesian model, we are trying to approximate the posterior $p(z|x)$ in which $x|z \sim \mathcal{N}(z, \sigma_x^2)$ and $z \sim \mathcal{N}(\mu_0, \sigma_0^2)$, where we assume that σ_0^2, σ_x^2 are known. The posterior in this case has a known closed-form Normal density, but we can consider using VI in this example with:

$$\tau \equiv (\mu, \sigma^2)$$

being the posterior mean and variance of the parameter z, with the mean-field parametrization:

$$q(z) \equiv \mathcal{N}(z|\mu, \sigma^2)$$

Note that this is a single density factor since the latent variable z is a scalar. Thus, in this case the problem is:

$$\max_{\mu, \sigma^2} ELBO(q)$$

$$\text{where } q(z) \equiv \mathcal{N}(z|\mu, \sigma^2)$$

4. Coordinate-Ascent Variational Inference

In order to solve the above problem, one simple but common algorithm is **coordinate-ascent variational inference (CAVI)**. The algorithm is the following:

- Initialize each variational density factor $q_j(z_j)$ arbitrarily
- While the *ELBO* has not converged:
$$q_j(z_j) \propto \exp\{E_{-j}[\log p(z_j|z_{-j}, x)]\}$$
for each $j = 1, \dots, m$ in order.
- The expectation E_{-j} is taken with respect to $\prod_{l \neq j} q_l(z_l)$, i.e. the current set of variational density factors other than q_j.

This algorithm follows from a simple observation at the ELBO. Suppose that, as the name suggests, we climb the ELBO "coordinate-wise", so that we maximize the ELBO along one of the latent variables z_j at a time (actually along the q_j). Then we are only considering z_j as the variable of interest, so we can write the ELBO as follows:

$$ELBO(q) = E_j[E_{-j}[\log p(z_j, z_{-j}, x)]] - E_j[\log q_j(z_j)] + G(z_{-j})$$

where we have collapsed all constant terms not containing z_j into the $G(z_{-j})$ term. We note moreover that:

$$E_j[E_{-j}[\log p(z_j, z_{-j}, x)]] - E_j[\log q_j(z_j)] \equiv -KL(q_j(z_j) \| \exp\{E_{-j}[\log p(z_j|z_{-j}, x)]\})$$

Thus, if we perform the following:

$$\max_{q_j} ELBO(q_j, q_{-j})$$

given the above decomposition, we note that the maximum occurs precisely when the KL divergence is minimized; i.e., when we have:

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

as given in the algorithm.

We note that despite the seemingly simple nature of the algorithm, deriving the complete conditional $E_{-j}[\log p(z_j|z_{-j}, x)]$ is often non-trivial for many problems, and involves the majority of the work in formulating the VI method for a given problem.

References

[1] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," arXiv: 1601.0067.

[2] A. Grothendieck, Resume de la theorie metrique des produits tensoriels topologiques (French), Bol. Soc. Mat. Sao Paulo 8, 179 (1953).

[3] G. Pisier, "Grothendieck's theorem, past and present", arXiv.