Matthew N. Bernstein
Computational Biologist at Immunitas Therapeutics

Follow

Variational inference

5 minute read

Published: May 31, 2021

In this post, I will present a high-level explanation of variational inference: a paradigm for estimating a posterior distribution when computing it explicitly is intractable. Variational inference finds an approximate posterior by solving a specific optimization problem that seeks to minimize the disparity between the true posterior and the approximate posterior.

Introduction

Variational inference is a high-level paradigm for estimating a posterior distribution when computing it explicitly is intractable. More specifically, variational inference is used in situations in which we have a model that involves hidden random variables Z, observed data X, and some posited probabilistic model over the hidden and observed random variables $P(Z, X)$. Our goal is to compute the posterior distribution $P(Z \mid X)$. Under an ideal situation, we would do so by using Bayes theorem:

$$p(z \mid x) = \frac{p(x \mid z)p(z)}{p(x)}$$

where z and x are realizations of Z and X respectively and $p(.)$ are probability mass/density functions for the distributions implied by their arguments.

In practice, it is often difficult to compute $p(z \mid x)$ via Bayes theorem because the denominator $p(x)$ does not have a closed form. Usually, the denominator $p(x)$ can be only be expressed as an integral that marginalizes over z: $p(x) = \int p(x, z)\, dz$. In such scenarios, we're often forced to approximate $p(z \mid x)$ rather than compute it directly. Variational inference is one such approximation technique.

Intuition

Instead of computing $p(z \mid x)$ exactly via Bayes theorem, variational inference attempts to find another distribution $q(z)$ that is ``close" to $p(z \mid x)$ (how we define "closeness" between distributions will be addressed later in this post). Ideally, $q(z)$ is easier to evaluate than $p(z \mid x)$, and, if $p(z \mid x)$ and $q(z)$ are similar, then we can use $q(z)$ as a replacement for $p(z \mid x)$ for any relevant downstream tasks.

We restrict our search for $q(z)$ to a family of surrogate distributions over Z, called the **variational distribution family**, denoted by the set of distributions $\mathcal{Q}$. Our goal then is to find the distribution $q \in \mathcal{Q}$ that makes $q(z)$ as ``close" to $p(z \mid x)$ as possible. When, each member of $\mathcal{Q}$ is characterized by the values of a set of parameters ϕ, we call ϕ the **variational parameters**. Our goal is then to find the value for $\hat{\phi}$ that makes $q(z \mid \phi)$ as close to $p(z \mid x)$ as possible and return $q(z \mid \hat{\phi})$ as our approximation of the true posterior.

Details

Variational inference uses the KL-divergence from $p(z \mid x)$ to $q(z)$ as a measure of ``closeness" between these two distributions:

$$KL(q(z) \mid\mid p(z \mid x)) := E_{Z \sim q}\left[\log \frac{q(Z)}{p(Z \mid x)}\right]$$

Thus, variational inference attempts to find

$$\hat{q} := \text{argmin}_q \, KL(q(z) \mid\mid p(z \mid x))$$

and then returns $\hat{q}(z)$ as the approximation to the posterior.

Variational inference minimizes the KL-divergence by maximizing a surrogate quantity called the **evidence lower bound (ELBO)** (For a more in-depth discussion of the evidence lower bound, you can check out my previous blog post):

$$\text{ELBO}(q) := E_{Z \sim q}\left[\log p(x, Z)\right] - E_{Z \sim q}\left[\log q(Z)\right]$$

That is, we can formulate an optimization problem that seeks to maximize the ELBO:

$$\hat{q} := \text{argmax}_q \, \text{ELBO}(q)$$

The solution to this optimization problem is equivalent to the solution that minimizes the KL-divergence between $q(z)$ and $p(z \mid x)$. To see why this works, we can show that the KL-divergence can be formulated as the difference between the marginal log-likelihood of the observed data, $\log p(x)$ (called the *evidence*) and the ELBO:

$$\begin{aligned}
KL(q(z) \mid\mid p(z \mid x)) &= E_{Z \sim q}\left[\log \frac{q(Z)}{p(Z \mid x)}\right] \\
&= E_{Z \sim q}\left[\log q(Z)\right] - E_{Z \sim q}\left[\log p(Z \mid x)\right] \\
&= E_{Z \sim q}\left[\log q(Z)\right] - E_{Z \sim q}\left[\log \frac{p(Z, x)}{p(x)}\right] \\
&= E_{Z \sim q}\left[\log q(Z)\right] - E_{Z \sim q}\left[\log p(Z, x)\right] + E_{Z \sim q}\left[\log p(x)\right] \\
&= \log p(x) - \left(E_{Z \sim q}\left[\log p(x, Z)\right] - E_{Z \sim q}\left[\log q(Z)\right]\right) \\
&= \log p(x) - \text{ELBO}(q)
\end{aligned}$$

Because $\log p(x)$ does not depend on q, one can treat the ELBO as a function of q and maximize the ELBO.

Conceptually, variational inference allows us to formulate our approximate Bayesian inference problem as an optimization problem. By formulating the problem as such, we can approach this optimization problem using the full toolkit available to us from the field of mathematical optimization!

Why is this method called “variational” inference?

The term “variational” in “variational inference” comes from the mathematical area of the calculus of variations. The calculus of variations is all about optimization problems that optimize *functions of functions*, called functionals.

More specifically, let’s say we have some set of functions $\mathcal{F}$ where each $f \in \mathcal{F}$ maps items from some set A to some set B. That is,

$$f : A \to B$$

Let’s say we have some function g that maps functions in $\mathcal{F}$ to real numbers $\mathbb{R}$. That is,

$$g : \mathcal{F} \to \mathbb{R}$$

Then, we may wish to solve an optimization problem of the form:

$$\arg\max_{f \in \mathcal{F}} g(f)$$

This is precisely the problem addressed in the calculus of variations. In the case of variational inference, the functional, g, that we are optimzing is the ELBO. The set of functions, $\mathcal{F}$, that we are searching over is the set of measureable functions in the variational family, $\mathcal{Q}$.

Tags: machine learning | probability | statistics | tutorial

Previous | Next