Variational Inference

Karel Stryczek

Case Western Reserve University

May 17, 2025

Overview

- Notation and Definitions
- Problem Setup
- Remedies to the Proposed Problem
- KL Divergence and ELBO

Notation and Definitions

- x: The observed data
- z: Latent variables (Unobserved but can be inferred from observation)
- $P(z)$: Prior distribution (Beliefs prior to seeing any data)
- $P(x|z)$: Likelihood of data given z (Given the latent variable, what is the probability of observing x?)
- $P(z|x)$: Posterior distribution (This is what we want to solve for; we are inferring z based on evidence x)
- Recall: $P(x) = \int P(x|z)P(z)dz = \int P(x,z)dz$

Bayes' Rule and Derivation of Problem

Recall Bayes' Rule: $P(z|x) = \frac{P(x|z)P(z)}{P(x)}$

- $P(x|z)$ and $P(z)$ are typically defined by the model designer, so they are trivial to compute
- However, $P(x)$ is intractable, as we are dealing usually with high dimensional z and x
- Thus, most of the time in Variational Inference is spent finding clever ways to approximate this $P(z|x)$ posterior and its parameters

Finding a Surrogate Posterior

The approach begins with picking a family of distributions over the latent variables that has its own variational parameters (ν): $q(z_{1:m}, \nu)$

- The variational parameters is what we will be optimizing to make the distribution as close as possible to the posterior $P(z|x)$
- After this, we can use q distribution with the optimized parameters in place of the posterior

Kullback-Leibler (KL) Divergence

KL Divergence is an information theory method of finding the closeness of two distributions.

- Defined as: $KL(q||p) = \int_z q(z) \log(\frac{q(z)}{p(z|x)}) dz = \mathbb{E}_q[\log(\frac{q(z)}{p(z|x)})]$
- If this value is low, then this implies close distributions
- Intuitively, we want to minimize this KL divergence.
- We can do this using ELBO, which will be shown later, but there are some caveats to KL divergence.

Evidence Lower Bound (ELBO)

Instead of finding $P(z|x)$ via minimizing the KL divergence directly, we look at Evidence Lower Bound and try to maximize it.

- Recall: the real $P(z|x)$ is usually quite nasty looking, several peaks, high dimensional, etc.
- Denote this surrogate posterior $q(z)$
- We want to use ELBO to define this $q(z)$

$$q(z) = \arg\max(L(q))$$
$$L(q) = \mathbb{E}_{q(z)}[\log(\frac{P(x,z)}{q(z)})]$$

  Can be rewritten as: $\mathbb{E}_q[\log p(x,z)] - \mathbb{E}_q[\log p(z)]$
- The above expression is what we will be working with
- It can be derived that the KL divergence of two distributions is equal to the negative ELBO plus a constant

The Mean Field Implementation

- We begin by assuming a naive factorization of $q(z)$ as follows:
$$q(z_1, z_2, ..., z_m) = \prod_{i=1}^{m} q(z_i)$$
- Note: more likely, these will be factorized in groups
- Write the ELBO in terms of this factorization:
$$L = \mathbb{E}_q[\log(P(x, z))] - \mathbb{E}_q[\log(q(z))]$$
Substitute:
$$L = \log(P(x_{1:m})) + \sum_{i=1}^{m}(\mathbb{E}_q[\log(P(z_i|z_{1:(i-1)}, x_{1:n})] - \mathbb{E}_{q_i}[\log(q(z_i))]])$$
- Using this we can actually derive a gradient ascent algorithm using: $\arg\max(L)$

The Mean Field Implementation Cont.

- Skipping over some minor derivation, we will use Lagrange multipliers to optimize $q(z_j)$ and return:

$$q^*(z_j) \propto \exp\{\mathbb{E}_{q_{-j}}[\log(P(z_j, z_{-j}, x)]\}$$

- $q^*(z_j)$ is the gradient ascent update of $q(z_j)$

Concluding Remarks

- Core Concepts: Bayes' Rule, KL Divergence, ELBO
- VI is a Bayesian Computing method meant to find the distributions of parameters using optimization (opposed to something like sampling in MCMC)
- End goal is to approximate the posterior distributions of the parameters, so we can infer for unobserved data
- *Be wary of KL Divergence's shortfalls