Variational Inference

Nipun Batra

November 11, 2023

IIT Gandhinagar

Introduction

Bayesian ML: Recap

- We assume a prior distribution over the parameters of the model given as $P(\theta)$
- We assume a likelihood function $P(D|\theta)$
- We use Bayes' rule to find the posterior distribution of the parameters given the data: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$
- Typically, we can not compute the posterior distribution analytically as the denominator is intractable

Bayesian ML: Methods

Laplace Approximation

Approximates the posterior with a Gaussian distribution parameterized by $\Psi = (\mu, \Sigma)$.

$$q_\Psi(\theta) = \mathcal{N}(\mu, \Sigma)$$

where μ is the mode of the posterior and Σ is the negative inverse Hessian of the log joint distribution evaluated at θ_{MAP}.

MCMC (Markov Chain Monte Carlo)

Generates samples from the posterior distribution by constructing a Markov chain.

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Variational Inference

Poses posterior inference as an optimization problem. The approximating distribution is parameterized by Ψ.

$$\Psi^* = \arg\min_\Psi \text{KL}(q_\Psi(\theta)||P(\theta|D))$$

KL Divergence

- KL divergence is a measure of dissimilarity between two distributions.
- It is defined as: $\mathrm{KL}(q||p) = \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta$

Exercise

Compute the KL divergence between two Gaussian distributions $q(\theta) = \mathcal{N}(\mu_1, \sigma_1^2)$ and $p(\theta) = \mathcal{N}(\mu_2, \sigma_2^2)$.

The answer is: $\frac{1}{2}\left(\log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{\sigma_2^2} - 1\right)$

Notebook demo

Optimizing

Notebook demo

Monte Carlo Sampling

Notebook demo

Repameterization Trick

Notebook demo

ELBO

Worked out example: Coin Toss

Worked out example: Linear Regression

Worked out example: Neural Networks