Probability Book

Prior and data gives the posterior.

Author: occupymars

Date: June. 23, 2025

Version: 0.1

Contents

Chapter 1 Proofs of some basic things

1.1 Basic Probabilistic Theorem

Theorem 1.1 (Law of total expectation and variance)

Let X and Y be random variables defined on the same probability space and assume that the variance of Y is finite. Then

$$\mathbb{E}[X] = \mathbb{E}_Y\left[\mathbb{E}_{X|Y}[X|Y]\right] \tag{1.1}$$

$$\mathrm{Var}_X = \mathbb{E}_Y\left[\mathrm{Var}_{X|Y}[X|Y]\right] + \mathrm{Var}_Y\left[\mathbb{E}_{X|Y}[X|Y]\right] \tag{1.2}$$

♡

1.2 Gaussian

Theorem 1.2 (Product of two guassian density function)

Given two random variable $x \sim \mathcal{N}(\mu_x, \Sigma_x)$ *and* $y \sim \mathcal{N}(\mu_y, \Sigma_y)$*, then*

$$\int_x \mathcal{N}(\mu_a, \Sigma_a)\mathcal{N}(\mu_b, \Sigma_b)dx = \mathcal{N}(\mu, \Sigma)$$

where

$$\mu = \frac{\Sigma_b\mu_a + \Sigma_a\mu_b}{\Sigma_a + \Sigma_b}$$

and

$$\Sigma = |\Sigma_a^{-1} + \Sigma_b^{-1}|^{-1} = \frac{\Sigma_a\Sigma_b}{\Sigma_a + \Sigma_b}$$

♡

Proof

$$\begin{aligned}\int_x \mathcal{N}(\mu_a, \Sigma_a)\mathcal{N}(\mu_b, \Sigma_b)dx = \int_x (2\pi)^{-\frac{n}{2}}\Sigma_a^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu_a)^T\Sigma_a^{-1}(x-\mu_a)\right)\\ \times (2\pi)^{-\frac{n}{2}}\Sigma_b^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu_b)^T\Sigma_b^{-1}(x-\mu_b)\right)dx\end{aligned}$$

Theorem 1.3 (Linear Gaussian Systems)

Suppose the two Gaussian densities

$$p(u) = \mathcal{N}(u|\mu_0, \Sigma_0) \quad \textit{and} \quad p(y|u) = \mathcal{N}\big(y\big|Hu, V\big),$$

where $H \in \mathbb{R}^{N\times M}$ *and* $V \in \mathbb{R}^{N\times N}$*. Then we can compute*

$$\begin{aligned}p(y) &= \int p(y|u)p(u)du\\ &= \mathcal{N}\big(y\big|H\mu_0, V + H\Sigma_0H^T\big)\end{aligned}$$

And

$$p(u|y) = \mathcal{N}\big(u\big|\Sigma\left(H^TV^{-1}y + \Sigma_0^{-1}\mu_0\right), \Sigma\big)$$

where

$$\Sigma = \left(\Sigma_0 + H^TV^{-1}H\right)^{-1}.$$

♡

Chapter 2 Variational Inference

2.1 Detailed Explanation of Variational Inference

Variational Inference (VI) is an approximate inference method in Bayesian statistics used to approximate complex posterior distributions, particularly when the posterior is difficult to compute directly. Below is a detailed explanation of variational inference using mathematical language, covering its basic ideas, mathematical derivations, and application scenarios.

2.1.1 Background and Objective

In Bayesian inference, we aim to compute the posterior distribution $p(\mathbf{z}|\mathbf{x})$ given observed data $\mathbf{x}$ and model parameters $\mathbf{z}$. According to Bayes' theorem:

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

where:

- $p(\mathbf{x}|\mathbf{z})$ is the likelihood function,
- $p(\mathbf{z})$ is the prior distribution,
- $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z})\, d\mathbf{z}$ is the marginal likelihood (evidence), which is often difficult to compute analytically.

When the integral for $p(\mathbf{x})$ is complex or high-dimensional, computing the posterior $p(\mathbf{z}|\mathbf{x})$ directly becomes infeasible. Variational inference introduces an approximate distribution $q(\mathbf{z})$ to approximate the posterior $p(\mathbf{z}|\mathbf{x})$, optimizing it to be as close as possible to the true posterior.

2.1.2 Basic Idea of Variational Inference

Variational inference transforms the posterior inference problem into an optimization problem. Suppose we choose a variational distribution $q(\mathbf{z}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are variational parameters. We aim to adjust $\boldsymbol{\theta}$ to make $q(\mathbf{z}; \boldsymbol{\theta})$ as close as possible to the true posterior $p(\mathbf{z}|\mathbf{x})$ under some measure. Typically, the **Kullback-Leibler (KL) divergence** is used to quantify the difference between the two distributions:

$$\mathrm{KL}(q(\mathbf{z}; \boldsymbol{\theta}) \| p(\mathbf{z}|\mathbf{x})) = \int q(\mathbf{z}; \boldsymbol{\theta}) \log \frac{q(\mathbf{z}; \boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{x})}\, d\mathbf{z}$$

The goal is to minimize the KL divergence:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}(q(\mathbf{z}; \boldsymbol{\theta}) \| p(\mathbf{z}|\mathbf{x}))$$

However, computing the KL divergence directly requires knowledge of $p(\mathbf{z}|\mathbf{x})$, which is the very quantity we are trying to approximate. Therefore, variational inference optimizes an equivalent objective function—the Evidence Lower Bound (ELBO).

2.1.3 Evidence Lower Bound (ELBO)

To derive the ELBO, we start with the log marginal likelihood $p(\mathbf{x})$:

$$\log p(\mathbf{x}) = \log \int p(\mathbf{x}, \mathbf{z})\, d\mathbf{z}$$

Introducing the variational distribution $q(\mathbf{z}; \boldsymbol{\theta})$, we can rewrite it as:

$$\log p(\mathbf{x}) = \log \int p(\mathbf{x}, \mathbf{z}) \frac{q(\mathbf{z}; \boldsymbol{\theta})}{q(\mathbf{z}; \boldsymbol{\theta})} \, d\mathbf{z} = \log \mathbb{E}_q \left[\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\theta})} \right]$$

Using Jensen's inequality (for the concave function $\log \cdot$):

$$\log \mathbb{E}_q \left[\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\theta})} \right] \geq \mathbb{E}_q \left[\log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\theta})} \right]$$

We define the Evidence Lower Bound (ELBO) as:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_q \left[\log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\theta})} \right] = \int q(\mathbf{z}; \boldsymbol{\theta}) \log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}; \boldsymbol{\theta})} \, d\mathbf{z}$$

Expanding the ELBO:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_q[\log q(\mathbf{z}; \boldsymbol{\theta})]$$

Decomposing the joint distribution $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, we get:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z})] + \mathbb{E}_q[\log p(\mathbf{z})] - \mathbb{E}_q[\log q(\mathbf{z}; \boldsymbol{\theta})]$$

Noticing that:

$$\mathbb{E}_q[\log p(\mathbf{z})] - \mathbb{E}_q[\log q(\mathbf{z}; \boldsymbol{\theta})] = -\mathrm{KL}(q(\mathbf{z}; \boldsymbol{\theta}) \| p(\mathbf{z}))$$

Thus, the ELBO can be written as:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q(\mathbf{z}; \boldsymbol{\theta}) \| p(\mathbf{z}))$$

The significance of the ELBO lies in:

$$\log p(\mathbf{x}) = \mathcal{L}(\boldsymbol{\theta}) + \mathrm{KL}(q(\mathbf{z}; \boldsymbol{\theta}) \| p(\mathbf{z}|\mathbf{x}))$$

Since the KL divergence is non-negative, maximizing the ELBO is equivalent to minimizing $\mathrm{KL}(q(\mathbf{z}; \boldsymbol{\theta}) \| p(\mathbf{z}|\mathbf{x}))$. Thus, the goal of variational inference is:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$$

2.1.4 Choice of Variational Distribution

To make optimization feasible, structural assumptions are often imposed on the variational distribution $q(\mathbf{z}; \boldsymbol{\theta})$. The most common is **Mean-Field Variational Inference**, which assumes that the components of $\mathbf{z} = \{z_1, z_2, \ldots, z_n\}$ are independent:

$$q(\mathbf{z}; \boldsymbol{\theta}) = \prod_{i=1}^{n} q_i(z_i; \theta_i)$$

Each $q_i(z_i; \theta_i)$ is a distribution controlled by parameters θ_i (e.g., a Gaussian distribution). The mean-field assumption simplifies the optimization problem but may fail to capture complex dependencies between variables.

2.1.5 Optimizing the ELBO

The ELBO is typically optimized using the following methods:

2.1.5.1 Coordinate Ascent Variational Inference (CAVI)

Under the mean-field assumption, each $q_i(z_i;\theta_i)$ is optimized sequentially while keeping the others fixed. Fixing the other distributions $q_{-i} = \prod_{j\neq i} q_j(z_j;\theta_j)$, the goal is to maximize:

$$\mathcal{L}(\theta_i) = \mathbb{E}_{q_i}\mathbb{E}_{q_{-i}}[\log p(\mathbf{x},\mathbf{z})] - \mathbb{E}_{q_i}[\log q_i(z_i;\theta_i)] - \text{const}$$

Using variational calculus, the optimal $q_i(z_i)$ satisfies:

$$q_i^*(z_i) \propto \exp\left(\mathbb{E}_{q_{-i}}[\log p(\mathbf{x},\mathbf{z})]\right)$$

where $\mathbb{E}_{q_{-i}}[\log p(\mathbf{x},\mathbf{z})]$ is the expectation over all variables except z_i. This is iterated for each i until convergence.

2.1.5.2 Gradient-Based Methods

When the variational distribution is complex or analytical solutions are infeasible, gradient descent can be used to optimize the ELBO. The gradient of the ELBO with respect to $\boldsymbol{\theta}$ is:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbb{E}_q[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z};\boldsymbol{\theta})]$$

To avoid computing high-dimensional integrals directly, the **Reparameterization Trick** is often used. Assume $q(\mathbf{z};\boldsymbol{\theta})$ can be expressed as $\mathbf{z} = g(\boldsymbol{\epsilon},\boldsymbol{\theta})$, where $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ is a standard distribution (e.g., standard normal). Then:

$$\mathbb{E}_q[f(\mathbf{z})] = \mathbb{E}_{p(\boldsymbol{\epsilon})}[f(g(\boldsymbol{\epsilon},\boldsymbol{\theta}))]$$

This allows the gradient to be moved inside the expectation, facilitating Monte Carlo estimation:

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_q[f(\mathbf{z})] = \mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} f(g(\boldsymbol{\epsilon},\boldsymbol{\theta}))]$$

2.1.6 Advantages and Disadvantages of Variational Inference

Advantages:

- Computationally efficient, suitable for large-scale datasets and high-dimensional models.
- Systematically approximates the posterior by optimizing the ELBO.
- Highly flexible, compatible with various variational distributions and optimization algorithms.

Disadvantages:

- The mean-field assumption may underestimate the posterior covariance.
- ELBO optimization may converge to local optima.
- The approximate distribution may not fully capture the complexity of the true posterior.

2.1.7 Application Scenarios

Variational inference is widely used in:

- **Topic Models** (e.g., Latent Dirichlet Allocation, LDA): Approximating document topic distributions.
- **Variational Autoencoders (VAE)**: Learning latent variable distributions for generative models.
- **Bayesian Neural Networks**: Approximating posterior distributions over weights.
- **Gaussian Processes**: Approximate inference for large-scale datasets.

2.1.8 Summary

Variational inference transforms posterior inference into an optimization problem, using a variational distribution $q(\mathbf{z};\boldsymbol{\theta})$ to approximate the true posterior $p(\mathbf{z}|\mathbf{x})$, and optimizing the parameters $\boldsymbol{\theta}$ by maximizing the ELBO. Its mathematical core lies in the KL divergence and Jensen's inequality, with optimization methods including coordinate ascent

and gradient descent. Variational inference is theoretically elegant and computationally efficient, making it a key tool in modern Bayesian inference.