STAT201B Lecture 15 Sampling from Posterior

1 Sampling from Posterior Distribution

Logic

通常情况下, 我们难以得到 posterior distribution 的 closed form (由于分母上的 normalizer 难以积分求出)

为了进行后续的 inference, 我们考虑 sampling from posterior distribution (即便没有 closed form 也能进行 sampling)

1.1 Sampling from posterior 的方法概述

两种常见的 sampling from the posterior 的方法为:

- Rejection sampling: 可以得到 exact 且 iid 的 samples, 但是难以根据特定问题调整
- Importance sampling: 更 general, 但只能对 distribution 进行近似

Remark

Rejection sampling 一般只适合低维单峰问题, 且 proposed distribution 和 true distribution 差异很大时效率会很低

Remark

另一种常见的方法是 Markov Chain Monte Carlo (MCMC) methods, 其大致思路为:

构建一个 Markov chain, 使得其 stationary distribution 恰好为 posterior distribution

本课程不详细介绍该方法

Logic

关于 MCMC methods 的详细论述, 见 STA4102 Lecture 20

Logic

在考虑具体的 sampling method 之前, 我们不妨先考虑, 如果 posterior distribution 的形式较为简单或可以对其进行 sampling, 那我们可以做到什么?

一个显而易见的想法是: 可以使用 Monte Carlo approximation 来估计 posterior mean, 我们将在下面简单介绍这种想法

除此之外, 其他想法包括但不限于:

- 构建 ECDF 来估计 posterior CDF
- 使用 histogram 或 kernel density estimation 来估计 posterior PDF

1.2 Sampling 后的应用: Monte Carlo approximation

若可以对 posterior distribution 进行采样: $\theta_1, \dots, \theta_B \overset{i.i.d.}{\sim} f(\theta|x^n)$, 则可以使用 **Monte Carlo approximation** 来近似任意 function $q(\theta)$ 的 posterior mean:

$$\begin{aligned}\mathbb{E}[q(\theta|x^n)] &= \int q(\theta) f(\theta|x^n)\, d\theta \\ &\approx \frac{1}{B}\sum_{i=1}^{B} q(\theta_i)\end{aligned}$$

> ⚠ **Remark**
>
> 一个特殊的用法是, 令 q 为 indicator function, 则可以用于近似 posterior probability of any event

---

1.3 Sampling 的方法: rejection sampling

> **Logic**
>
> 由于 posterior distribution 的 normalizer 通常无法求出, 我们无法从 posterior distribution 直接进行 sample
>
> 为了解决这个问题, 我们首先考虑使用 rejection sampling, 在 normalizer 未知的情况下仍能从 posterior distribution 采样出 exact iid samples

> **Logic**
>
> 关于 rejection sampling 的详细论述, 见 STA4042 Lecture 12 DDA2001 Lecture 9-10

> ⚠ **Remark: 回顾: Rejection sampling**
>
> ***情景:***
> 我们希望从一个 density $f \propto \tilde{f}$ 进行采样, 但我们只知道 $\tilde{f}$ 的形式
>
> ***算法:***
> 首先选择一个 (易于采样的) density $q : \mathbb{R} \to \mathbb{R}$, 其满足: 存在一个 constant $C > 0$, 使得
>
> $$\tilde{f}(x) \le C q(x), \quad \forall x \in \mathbb{R}$$
>
> 即确保函数 $Cq(x)$ 永远不低于 density $\tilde{f}(x)$
>
> 随后我们重复进行以下步骤:
>
> 1. 从选取的 density q 进行 sampling, 得到 x
> 2. 从均匀分布 $\sim \mathrm{Unif}([0, Cq(x)])$ 进行 sampling, 得到 u
> 3. 进行以下判断, 选择是否将 x 添加进 final sample dataset:
>   - 若 $u \le \tilde{f}(x)$, 则 accept x
>   - 若 $u > \tilde{f}(x)$, 则 reject x
>
> 由此得到的 final sample dataset 可视为 iid sample from density f

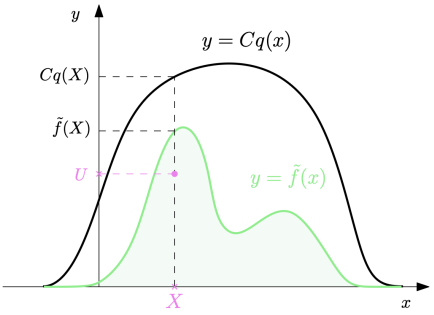

Figure 5.4: Two possible samples of U and X. $U \leq \tilde{f}(X)$ so we accept X.

证明:

令 random variable Y 服从和 final sample dataset 相同的分布, 我们希望证明 $Y = X|(\ \leq \tilde{f}(X)) \sim f$, 即 $p_Y(y) = f(y)$

我们首先对 $p_Y(y)$ 进行处理:

$$p_Y(y) = p(X = y| \leq \tilde{f}(X)) = \frac{p(\ \leq \tilde{f}(X)|X = y)\, p_X(y)}{\mathbb{P}(\ \leq \tilde{f}(X))}$$

我们逐项进行分析:

- 由于 X 是从 q 中 sample 的, 有 $p_X(y) = q(y)$,
- 对于分子中的另一项, 由于 $\ \sim \mathrm{Unif}([0, Cq(x)])$, 有

$$p(\ \leq \tilde{f}(X)|X = y) = p(\ \leq \tilde{f}(y)|X = y) = \frac{1}{Cq(y)} \int_0^{\tilde{f}(y)} dt = \frac{\tilde{f}(y)}{Cq(y)}$$

- 对于分母:

$$\mathbb{P}(\ \leq \tilde{f}(X)) = \int_{\mathbb{R}} p(\ \leq \tilde{f}(X)|X = x) p_X(x)\, dx = \int_{\mathbb{R}} \frac{\tilde{f}(x)}{Cq(x)} q(x)\, dx = \frac{\int_{\mathbb{R}} \tilde{f}(x)\, dx}{C}$$

因此,

$$p_Y(y) = \frac{\tilde{f}(y)}{\int_{\mathbb{R}} \tilde{f}(x)\, dx} = f(y)$$

目标:

Sample from the posterior $f(\theta|x^n) \propto f(x^n|\theta) f(\theta)$, 其中 $f(x^n|\theta) f(\theta) =: \tilde{f}(\theta)$ 已知

$q(x)$ 的选取:

为了便于计算, 我们考虑首先 sample from the prior $f(\theta)$

C 的选取:

令 $\hat{\theta}_n$ 为 θ 的 MLE, 则有 $\tilde{f}(\theta)/q(x) = \frac{f(x^n|\theta) f(\theta)}{f(\theta)} = f(x^n|\theta) \leq f(x^n|\hat{\theta}_n) =: M$, 因此 M 是一个合理 (且高效) 的选取

算法:

首先计算:

- θ 的 MLE $\hat{\theta}_n$
- $M := f(x^n|\hat{\theta}_n)$

随后重复以下步骤直到 B 个 θ^{cand} 被 accepted:

1. Sample $\theta^{cand} \sim f(\theta)$

2. Sample $u \sim nif(0,1)$
3. 若 $u \leq f(x^n|\theta^{cand})/M$, 则 accept θ^{cand}, 否则 reject θ^{cand}

> ⚠ **Remark**
>
> 在选取 C 时, 只需要满足 $M \geq f(x^n|\hat{\theta}_n)$ 即可, 但其他选择会导致 sampling 的效率变低

> **Logic**
>
> 我们继续考虑无法从 posterior 直接进行 sampling 的问题
>
> 另一种考虑的角度是避免 sampling from posterior, 由于我们最终的目的是利用 samples 和 Monte Carlo integration 近似 posterior mean $\mathbb{E}[q(\theta|x^n)]$, 我们可以想办法绕开 posterior distribution, 并从另一个易于 sample 的 "important distribution" g 进行 sample

1.4 Sampling 的方法: importance sampling

> **Logic**
>
> 关于 importance sampling 的详细论述, 见 STAT201B Lecture 3 Bootstrap (1)

> ⚠ **Remark: 回顾: importance sampling**
>
> 若 target density 为 h, 则 Importance sampling 的原理为:
>
> $$\begin{aligned}\mathbb{E}_h[q(\theta)] &= \int q(\theta)h(\theta)d\theta \\ &= \int q(\theta)\frac{h(\theta)}{g(\theta)}g(\theta)d\theta \\ &\approx \frac{1}{B}\sum_{i=1}^{B} q(\theta_i)\frac{h(\theta_i)}{g(\theta_i)}\end{aligned}$$
>
> 其中 $\theta_1, \dots, \theta_B \overset{i.i.d.}{\sim} g(\theta)$

目标:

直接近似 posterior mean $\mathbb{E}[q(\theta|x^n)]$

Important distribution 的选取:

为了便于计算, 我们选择 prior distribution $f(\theta)$ 作为 important distribution

此时, $\frac{h(\theta_i)}{g(\theta_i)}$ 这一项可以被化简 (并近似) 为:

$$\frac{f(\theta_i|x_1,\dots,x_n)}{f(\theta_i)} = \frac{f(x_1,\dots,x_n|\theta_i)\cdot f(\theta_i)}{f(\theta_i)\cdot f(x_1,\dots,x_n)} = \frac{f(x_1,\dots,x_n|\theta_i)}{\int f(x_1,\dots,x_n|\theta_i)f(\theta_i)\,dx} \approx \frac{\mathcal{L}(\theta_i)}{\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}(\theta_i)}$$

最后一步我们使用了 Monte Carlo integration, 为了避免再次进行 sampling, 我们这里直接使用之前的 samples: $\theta_1, \dots, \theta_B$

计算流程:

1. 从 prior distribution 进行 sampling: $\theta_1, \dots, \theta_B \overset{i.i.d.}{\sim} f(\theta)$
2. 对于 $i = 1, \dots, B$, 计算:

$$w_i = \frac{\mathcal{L}_n(\theta_i)}{\sum_{i=1}^B \mathcal{L}_n(\theta_i)}$$

3. 将 posterior mean 近似为:

$$\begin{aligned}\mathbb{E}[q(\theta)|x^n] &\approx \frac{1}{B}\sum_{i=1}^B q(\theta_i)\frac{f(\theta_i|x_1,\dots,x_n)}{f(\theta_i)} \\ &\approx \frac{1}{B}\sum_{i=1}^B q(\theta_i)\frac{\mathcal{L}(\theta_i)}{\frac{1}{B}\sum_{i=1}^B \mathcal{L}(\theta_i)} \\ &= \sum_{i=1}^B q(\theta_i)w_i\end{aligned}$$

> ⚠ **Remark: 考虑 general g 的情况**
>
> 若我们考虑 general $g(\theta)$ (而不是选取 $g(\theta) = f(\theta)$), 则使用 importance sampling 后我们会得到:
>
> $$\mathbb{E}[q(\theta)|x_1,\dots,x_n] \approx \frac{1}{B}\sum_{i=1}^B \frac{q(\theta_i)\cdot f(\theta_i|x_1,\dots,x_n)}{g(\theta_i)}$$
>
> 其中 samples $\theta_1,\dots,\theta_n \overset{i.i.d.}{\sim} g(\theta)$
>
> ---
>
> 接下来考虑如何近似 posterior $f(\theta|x_1,\dots,x_n)$:
>
> $$\begin{aligned}f(\theta|x_1,\dots,x_n) &= \frac{f(x_1,\dots,x_n|\theta)f(\theta)}{\int f(x_1,\dots,x_n|\theta)f(\theta)\,d\theta} \\ &= \frac{f(x_1,\dots,x_n|\theta)f(\theta)}{\int f(x_1,\dots,x_n|\theta)\frac{f(\theta)}{g(\theta)}g(\theta)\,d\theta} \quad \text{(Importance sampling)} \\ &\approx \frac{\mathcal{L}_n(\theta)\cdot f(\theta)}{\frac{1}{B}\sum_{i=1}^B \frac{\mathcal{L}_n(\theta_i)\cdot f(\theta_i)}{g(\theta_i)}} \quad (\theta_1,\dots,\theta_n \overset{i.i.d.}{\sim} g(\theta))\end{aligned}$$
>
> 这一步操作和之前的类似, 只是使用了 importance sampling 而不是 Monte Carlo integration
>
> 类似的, 为了避免再次进行 sampling, 我们这里直接使用之前的 samples: $\theta_1,\dots,\theta_B$
>
> ---
>
> 代回到之前的式子, 有:
>
> $$\begin{aligned}\mathbb{E}[q(\theta)|x_1,\dots,x_n] &\approx \frac{1}{B}\sum_{i=1}^B \frac{q(\theta_i)\cdot f(\theta_i|x_1,\dots,x_n)}{g(\theta_i)} \\ &\approx \frac{1}{B}\sum_{i=1}^B \frac{q(\theta_i)}{g(\theta_i)}\cdot\frac{\mathcal{L}_n(\theta_i)\cdot f(\theta_i)}{\frac{1}{B}\sum_{j=1}^B \frac{\mathcal{L}_n(\theta_j)\cdot f(\theta_j)}{g(\theta_j)}} \\ &= \sum_{i=1}^B \frac{q(\theta_i)}{g(\theta_i)}\cdot\frac{\mathcal{L}_n(\theta_i)\cdot f(\theta_i)}{\sum_{j=1}^B \frac{\mathcal{L}_n(\theta_j)\cdot f(\theta_j)}{g(\theta_j)}}\end{aligned}$$