机器学习白板推导系列笔记 (13.1~13.8)

Dexing Huang
dxhuang@bupt.edu.cn
Beijing University of Posts and Telecommunications

日期：2021 年 11 月 27 日

1 采样概述

首先说明一下采样的动机, 首先采样本身就是一个任务, 假设得到了某个概率分布 $p(x)$, 要获得样本, 就必须采样. 另外采样可以用于求积分, 特别是数学期望的估计. 假设 $f(x)$ 关于分布 $p(x)$ 的期望 $\mathbb{E}_{p(x)}[f(x)]$, 可以使用采样的方法求出

$$\mathbb{E}_{p(x)}[f(x)] = \frac{1}{L}\sum_{l=1}^{L} f(x^{(l)}) \tag{1}$$

根据大数定律, 当 $L \to \infty$ 时, 样本均值依概率收敛于期望.

知道采样的动机后, 对于采出来的样本, 如何判断是否是一个好样本呢. 首先样本要趋向高概率区域, 其次样本之间相互独立.

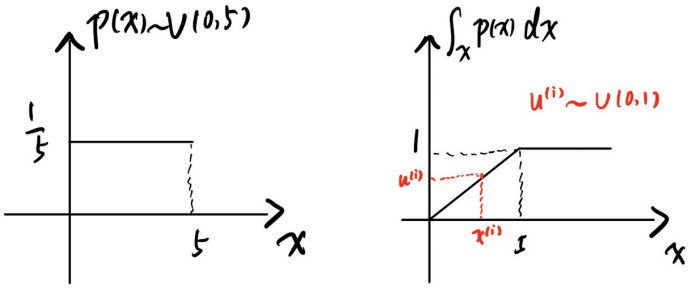
但是在实际操作中, 采样是十分困难的, 在这里主要介绍两方面. 首先Partition Function is Intractable, 在概率无向图模型中 $P(x) = \frac{1}{Z}\hat{P}(x)$, $\hat{P}(x)$ 通常是可以求的, 但是 $Z = \int \hat{P}(x)dx$ 往往是难以求解的, 受到维度等方面的制约. 其次高维使得采样难以满足“好样本”的假设, 之前说过好的样本其中一点是样本趋向高概率区, 想象这么一种情况, x 是 p 维离散随机变量, 每个维度都有 K 个状态, 那么所有的状态为 K^p, 你想知道哪些是该概率区域, 就得去遍历每个状态, 这显然是不可能的.

2 采样的三种方法

2.1 概率分布采样

假设知道随机变量 x 的概率密度函数 $p(x)$, 那么可以通过积分算出其累计分布函数 $(c.d.f)$, 即 $\int_{-\infty}^{x} p(x)dx$(假设这一步可以做到). 因为 $\int_{-\infty}^{x} p(x)dx \in [0,1]$, 那么可以使用均匀分布在 $[0,1]$ 区间上采样, $u^{(l)} \sim U(0,1)$, 那么 $x^{(l)} = c.d.f^{-1}(u^{(l)})$, 这就完成了采样.

但是这种方法有很大的局限性, 首先 $p(x)$ 就不一定是已知的, 其次 $c.d.f$ 一般也不是那么好求, 因此这种方法只适用于简单的 $p(x)$.



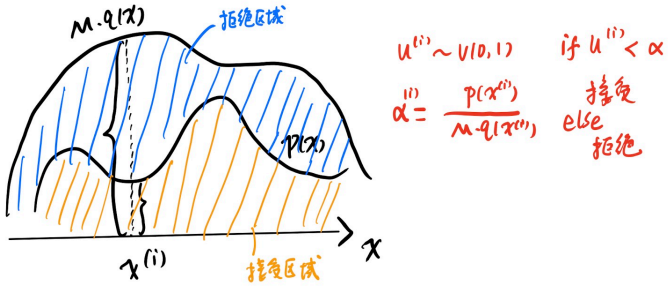
2.2 Rejection Sampling

假设对 $p(x)$ 采样十分困难, 那么可以对一个比较简单的分布 $q(x)$ 进行采样, $q(x)$ 分布是我们已知且可求的, 称为 (proposal distribution). 假设存在 $M > 0$ 使得对任意的 x 有

$$M \cdot q(x) \geq p(x) \tag{2}$$

定义一个接受率 α

$$\alpha = \frac{p(x^{(i)})}{M \cdot q(x^{(i)})} \tag{3}$$



总体步骤如下

- 在 $q(x)$ 上采样得到 $q(x^{(l)})$
- 在 $U(0,1)$ 上采样得到 $u^{(l)}$
- 如果 $u^{(l)} \leq \frac{p(x^{(l)})}{M \cdot q(x^{(l)})}$ 那么 $x^{(l)}$ 有效, 否则无效

该方法的性能比较依赖 $q(x)$ 的选择, 需要选择合适的 $q(x)$.

2.3 Importance Sampling

重要性采样在强化学习中应用比较多, 比如 off-policy 中有使用. 它不是直接对概率分布进行采样, 而是对概率分布的期望进行采样.

$$\begin{aligned}\mathbb{E}_{p(x)}[f(x)] &= \int_x p(x)f(x)dx \\ &= \int_x \frac{p(x)}{p'(x)}f(x)p'(x)dx \\ &= \mathbb{E}_{p'(x)}[\frac{p(x)}{p'(x)}f(x)] \\ &= \frac{1}{L}\sum_{l=1}^{L}\frac{p(x^{(l)})}{p'(x^{(l)})}f(x^{(l)})\end{aligned} \tag{4}$$

其中 $\frac{p(x^{(l)})}{p'(x^{(l)})}$ 称为重要性采样比, 很明显可以看出, 如果 $p'(x)$ 选择不当的话, 可能导致采样不均匀 (经常采到低概率区域). 为了解决这一问题提出了 Sampling Importance Resampling. 主要步骤是, 先进行重要性采样, 然后在获得的样本中, 按权重的大小再采样一次, 得到最终样本.

3 Markov Chain

马尔可夫链在随机过程以及信息论中都学习过, 比较熟悉, 对其进行简单总结, 主要考虑的是离散状态的一阶马氏链.

假设一随机过程序列 $\{X_0, X_1, \cdots, X_t, \cdots\}$, 满足马尔科夫性

$$P(X_t|X_0, X_1, \cdots, X_{t-1}) = P(X_t|X_{t-1}) \tag{5}$$

则称上述序列为马尔可夫链, 如果转移概率与时间无关

$$P(X_{t+s}|X_{t+s-1}) = P(X_t|X_{t-1}) \tag{6}$$

那么称其为齐次马氏链, 以下所有内容都是基于一阶齐次马尔可夫链讨论的.

3.1 状态转移概率矩阵/状态分布/平稳分布

状态转移概率矩阵 $P_{n\times n}$

$$p_{ij} = P(X_t = j|X_{t-1} = i) \tag{7}$$

t 时刻的状态分布

$$\pi(t) = \begin{pmatrix}\pi_1(t) & \pi_2(t) & \cdots & \pi_n(t)\end{pmatrix} \tag{8}$$

$t+1$ 时刻的状态分布

$$\pi(t+1) = \pi(t)P \tag{9}$$

这里进行简单的证明

$$\begin{aligned} P(X_{t+1}=i) &= \sum_m P(X_{t+1}=i, X_t=m) \\ &= \sum_m P(X_{t+1}=i|X_t=m)P(X_t=m) \\ &= \sum_m p_{mi}\pi_m(t) \end{aligned}$$

如果 $\pi(t)$ 与时间无关, 记为 π, 如果满足

$$\pi = \pi P \tag{10}$$

那么称其满足平稳分布, 有限状态的马尔可夫链平稳分布一定存在, 但不一定是唯一的.

3.2 马尔可夫链的性质

- **不可约**

对于任意两个状态 $i, j \in \mathcal{S}$, 如果存在一个时刻 t, 使得

$$P(X_t = j|X_0 = i) > 0 \tag{11}$$

那么称马尔可夫链是不可约的, 即 P^t 无零元.

- **非周期**

如果从状态 i 出发, 经过 t 时间返回, 所有时间的最大公约数是 1, 那么马尔可夫链是非周期的.

- **正常返**

即从某个状态出发经过有限的时间后会回到起始状态.

对于不可约, 非周期且正常返的马尔可夫链, 有唯一的平稳分布存在, 并且转移概率矩阵的极限分布试马尔可夫链的平稳分布, 即

$$\lim_{t\leftarrow\infty} P^t = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_n \\ \pi_1 & \pi_2 & \cdots & \pi_n \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_n \end{pmatrix} \tag{12}$$

- **可逆马尔可夫链**

对于马尔可夫链, 如果任意时刻 t 满足

$$P(X_t = i|X_{t-1} = j)\pi_j = P(X_{t-1} = j|X_t = i)\pi_i \tag{13}$$

$$p_{ji}\pi_j = p_{ij}\pi_i \quad \text{(细致平衡方程)} \tag{14}$$

满足细致平衡方程的状态分布就是马尔可夫链的平稳分布, 且平稳分布是唯一的, 对其进行简单的证明.

$$\begin{aligned} (\pi P)_i &= \sum_m \pi_m p_{mi} \\ &= \sum_m \pi_i p_{im} \\ &= \pi_i \end{aligned}$$

4 MCMC

MCMC 的基本想法是在随机变量 x 的状态空间上定义一个满足遍历定理的马尔可夫链, 使得其平稳分布就是目标分布 $p(x)$. 等到平稳后在马尔可夫链上随机游走, 得到的样本集合就是目标概率分布抽样的结果, 那么就可以近似数学期望值, 在未到达平稳分布的那段时间称为 bum-in, 这段时间的样本需要舍弃, 常用的 MCMC 方法有 Metropolis-Hastings 算法, Gibbs 抽样.

4.1 MH 算法

因为最终的想法是构造出一个平稳分布近似目标分布 $p(x)$, 因此如何构造转移概率矩阵成了问题的关键. 在上一节中证明了, 只要满足细致平衡方程, 那么分布就是平稳分布. 假设一个任意的转移概率矩阵 $q(x, x')$, 显然

$$p(x)q(x, x') \neq p(x')q(x', x) \tag{15}$$

假设引入一个接受分布 (acceptance distribution)$\alpha(x, x')$, 使得

$$p(x)q(x, x')\alpha(x, x') = p(x')q(x', x)\alpha(x', x) \tag{16}$$

那么 $p(x)$ 就成了对应 $q(x, x')\alpha(x, x')$ 马氏链的平稳分布, 其中 $q(x, x')$ 是另一个马尔可夫链的概率转移矩阵, 并且 $q(x, x')$ 是不可约的, 称为建议分布 (proposal distribution)同时因为是我们自己提出的, 所以很容易进行抽样.

$$p(x, x') = q(x, x')\alpha(x, x') \tag{17}$$

$\alpha(x, x')$ 满足以下情况能使得式 (16) 成立

$$\alpha(x, x') = \min\{1, \frac{p(x')q(x', x)}{p(x)q(x, x')}\} \tag{18}$$

下面进行证明

$$\begin{aligned} p(x)p(x, x') &= p(x)q(x, x')\alpha(x, x') \\ &= p(x)q(x, x') \min\{1, \frac{p(x')q(x', x)}{p(x)q(x, x')}\} \\ &= \min\{p(x)q(x, x'), p(x')q(x', x)\} \\ &= p(x')q(x', x) \min\{\frac{p(x)q(x, x')}{p(x')q(x', x)}, 1\} \\ &= p(x')p(x', x) \end{aligned}$$

因此 MH 算法的描述如下

- 随机选取初始值 x_0
- 对 $i = 1, 2, \cdots, n$ 循环执行
  - $x_{i-1} = x$, 按照建议分布 $q(x, x')$ 随机选取状态 x'
  - 计算接受率
  - 在 $U(0, 1)$ 中取一个数 u_i

    如果 $u \leq \alpha(x, x')$ 那么 $x_i = x'$, 否则 $x_i = x$
- 得到样本集合 $\{x_i\}_{i=m+1}^{n}$

  计算 $f_{mn} = \frac{1}{n-m}\sum_{i=m+1}^{n} f(x_i)$

4.2 Gibbs 抽样

Gibbs 抽样用于多元变量联合分布的抽样和估计, 从联合概率密度中定义满条件概率分布进行抽样, 得到样本序列, 可以在理论上证明这样的抽样过程是在一个马尔可夫链上的随机游走.

定义多元联合分布为 $p(x_1, x_2, \cdots, x_k)$, 每次对其中的第 j 个维度抽样 x_j, 固定其他维度. 重复 k 次得到一个样本 $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)}\}$, 再重复 N 次得到完全样本 $\{x^{(i)}\}_{i=1}^N$.

即然抽样过程是一个马尔可夫链上的随机游走, 那么一定满足细致均衡方程, 借用 MH 算法中推导, 可以知道在 Gibbs 采样中, proposal distribution $q(x, x') = p(x'|x)$, 下面对其进行分析, 假设正在采样第 i 个样本的第 j 个分量

$$\begin{aligned} p(x'|x) &= p(x_j^{(i)} \underbrace{x_1^{(i)} \cdots x_{j-1}^{(i)}}_{x_{<j}^{(i)}} \underbrace{x_{j+1}^{(i-1)} \cdots x_k^{(i-1)}}_{x_{>j}^{(i-1)}} | x_j^{(i-1)} \underbrace{x_1^{(i)} \cdots x_{j-1}^{(i)}}_{x_{<j}^{(i)}} \underbrace{x_{j+1}^{(i-1)} \cdots x_k^{(i-1)}}_{x_{>j}^{(i-1)}}) \\ &= p(x_j^{(i)} x_{<j}^{(i)} x_{>j}^{(i-1)} | x_j^{(i-1)} x_{<j}^{(i)} x_{>j}^{(i-1)}) \\ &= p(x_j^{(i)} | x_{<j}^{(i)} x_{>j}^{(i-1)}) \end{aligned} \tag{19}$$

淡化样本的概念可以将上式写为

$$p(x'|x) = p(x'_j|x_{-j}) \tag{20}$$

下面求一下 acceptance distribution, 这里先提前做一下说明, 在式 (19)(20) 的推导中, 可以看到

$$x'_{-j} = x_{-j} \tag{21}$$

这将极大简化接下来的计算

$$\begin{aligned} \alpha(x, x') &= \min\{1, \frac{p(x')q(x', x)}{p(x)q(x, x')}\} \\ &= \min\{1, \frac{p(x'_j|x'_{-j})p(x'_{-j})p(x_j|x'_{-j})}{p(x_j|x_{-j})p(x_{-j})p(x'_j|x_{-j})}\} \\ &= \min\{1, 1\} \\ &= 1 \end{aligned} \tag{22}$$

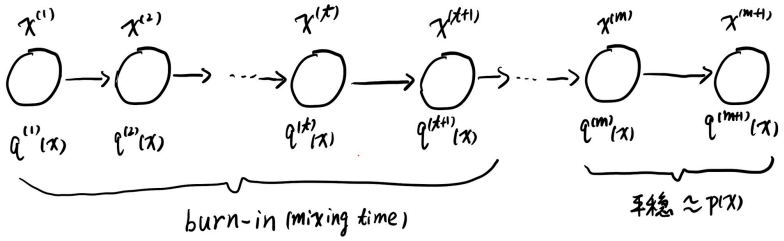
可以看到acceptance distribution 恒等于 1, Gibbs 抽样对每次抽样结果都接受, 没有拒绝.

Gibbs 抽样的步骤如下

- 给出初始化样本 $x^{(0)}$
- 对 $i = 1, 2, \cdots, n$ 循环执行
  - 对第 j 维进行采样, 固定其他维度 ($j = 1, 2, \cdots, k$)
  - 最终得到 $x^{(i)}$
- 得到样本集合 $\{x^{(i)}\}_{i=m+1}^n$
- 计算 $f_{mn} = \frac{1}{n-m} \sum_{i=m+1}^n f(x^{(i)})$

5 存在的问题

先来说一些基本概念, MCMC 的想法就是构造一个马尔可夫链, 使得最后收敛的平稳分布趋向于目标分布 $p(x)$, 相当于隐式的求 $p(x)$, 在收敛前的过程称为 burn-in, 中文名为燃烧期, 采样的有效数据是在平稳后的.

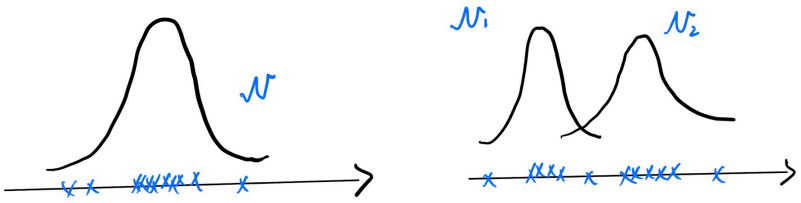


MCMC 方法十分巧妙简便, 但是仍存在一些问题. 首先是收敛时间未知, 在理论上只保证了构造的马尔可夫链的收敛性, 但并不知道什么时候可以收敛, 而只有当收敛后采样的分布才能近似等于原分布. 最简单粗暴的解决办法就是每隔一段时间进行统计, 查看是否收敛.

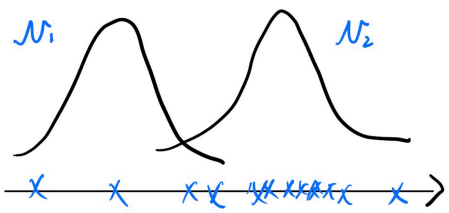
第二个问题是 mixing time 过长, 就算知道收敛的时间, 但是如果很长的话跟不收敛也没什么区别, 这主要是由于 $p(x)$ 维度过高, 以及维度之间的相关性过高.

最后是即便前两个问题都解决, MCMC 模型本身就有问题, 我们在" 什么是好的采样中" 讲到, 样本之间要独立同分布, 这对于 MCMC 来说显然不可能, 因为状态转移之间的马尔可夫性质, 解决的办法之一是取较大的时间间隔采样.

下面举一个简单的例子, 如高斯分布和高斯混合模型, 对于“好的采样”, 样本分布应该如下所示



这点在高斯分布上很容易实现, 但是 GMM 模型要想采出这样的点是很困难的, 因为它存在" 多峰", 实际采出来的样本往往是如下所示, 那么样本间就有了比较强的相关性.



样本会聚集在一个峰的附近很难到达另外的峰. 为什么会出现这种情况呢, 可以用能量的角

度解释这个问题, 再无向图中概率密度函数表示为如下形式

$$\begin{aligned} p(x) &= \frac{1}{Z}\hat{p}(x) \\ &= \frac{1}{Z}\exp(-E(x)) \end{aligned} \tag{23}$$

可以看到能量和概率是成反比的, 能量越小概率越大, 越稳定, 就越难以跳跃到其他状态, 因此在高维情况下就很容易发生只在一个峰值附近采样的情况.