采样——MCMC

Mia Feng

2018年4月20日

1 概述

MCMC: 粗暴的采样模拟方式,用于模拟直接计算困难的分布。用于采样,数值积分等等。

求解目标:用多次采样得到的频率分布近似原概率分布。即本来对复杂的 f(x) 做积分,但是因为 f(x) 比较复杂所以显式积分困难。迂回方法是构造统计量 $\frac{f(x)}{p(x)}$,通过对 $x \sim p(x)$ 进行采样,求取统计量 $\frac{f(x)}{p(x)}$ 的期望得到数值积分值。

$$\theta = \int_{a}^{b} f(x) dx = \int_{a}^{b} \frac{f(x)}{p(x)} p(x) \approx \frac{1}{n} \sum_{i=0}^{n-1} \frac{f(x_i)}{p(x_i)}$$
(1)

求解思路:微积分思想(recall:学习微积分的时候,用无数个划分的小矩形的面积来近似面积,但是此时小矩形是来自均匀分布的)。实际上,来自均匀分布的可能性很小,此时需要

求解方法: Markov Chain, 蒙特卡洛积分。

1.1 推导

推导 取 I 为示性函数。 a_l 表示 X 的第 l 个特征。样本有 n 个。类标 m 个,特征 s 个。

$$P(Y = c_k) = \frac{\sum_{i=1}^{n} I(y_i = c_k)}{n}, k = 1, 2, \dots, m$$
 (2)

2 算法实现 2

$$P(X_j = a_{jl}|Y = c_k) = \frac{\sum_{i=1}^n I(x_i^j = a_{jl}, y_i = c_k)}{\sum_{i=1}^n I(y_i = c_k)}$$
(3)

其中, $i=1,2,\cdots,n, l=1,2,\cdots,s, k=1,2,\cdots,m$

改进 为了避免分母为 0, 进行了拉普拉斯平滑, 即在分母上加了类数目。

$$P(X_j = a_{jl}|Y = c_k) = \frac{\sum_{i=1}^n I(x_i^j = a_{jl}, y_i = c_k) + 1}{\sum_{i=1}^n I(y_i = c_k) + m}$$
(4)

2 算法实现

注意实现时取了拉普拉斯平滑,见公式(4),且为了防止下溢取对概率值取了对数。[??]

3 Implementation

MCMC

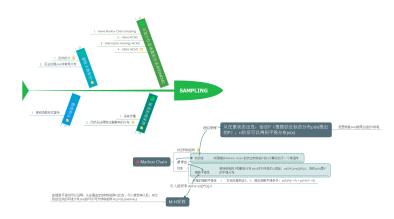


图 1: MCMC 思维导图