Some Notes on Markov Chain Monte Carlo

田博宇

1 Introduction

在概率论课上, 我们学习了马尔科夫链的相关知识. 马尔科夫链的一个重要应用是马尔科夫蒙特卡罗方法. 这是一种利用马氏链从随机分布取样的算法. 我将整理有关马尔科夫蒙特卡罗的一些知识和思考, 并且介绍我之前利用马尔科夫蒙特卡罗方法做的一个推荐系统小项目.

2 问题引入

2.1 问题

Given an irreducible transition matrix P, there is a unique stationary distribution satisfying $\pi = \pi P$. Consider the inverse problem: given a probability distribution π on χ, can we find a transition matrix P for which π is its stationary distribution.

2.2 问题解释

A random sample from a finite set χ will mean a random uniform selection from χ, i.e., one such that each element has the same chance $\frac{1}{|\chi|}$ of being chosen.

Consider the q-coloring problem. suppose that X_t is a chain with state space χ and with stationary distribution uniform on χ. By the Convergence Theorem X_t is approximately uniformly distributed when t is large.

This method of sampling from a given probability distribution is called Markov chain Monte Carlo.

2.3 用途

The MCMC method is used to estimate the expected value of a fuinction f(x):

$$E(f) = \sum_x f(x)p(x)$$

If each x_i can take two or more values, then there are at least 2^d values for x, so an explicit summation requires exponential time. Instead, one could draw a set of samples, where each sample x is selected with probability p(x): Averaging f over these samples provides an estimate of the sum.

3 基本方法

To sample according to p(x), design a Markov Chain whose states correspond to the possible values of x and whose stationary probability distribution is p(x): There are two general techniques to design such a Markov Chain: the Metropolis-Hastings algorithm and Gibbs sampling.

3.1 Metropolis-Hastings

The transitions of the Markov chain are de

ned as follows. At state i select neighbor j with probability $\frac{1}{r}$. Since the degree of i may be less than r, with some probability no edge is selected and the walk remains at i. If a neighbor j is selected and $p_j \geq p_i$, go to j. If $p_j < p_i$, go to j with probability $p_j = p_i$ and stay at i with probability . Intuitively, this favors "heavier" states with higher p_i values. For i adjacent to j in G,

$$p_{ij} = \frac{1}{r} min(1, \frac{p_j}{p_i})$$

and

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}$$

3.2 Gibbs Sampling

To generate samples of x = $(x_1 \cdots x_d)$ with a target distribution p (x), the Gibbs sampling algorithm repeats the following steps. One of the

variables x_i is chosen to be updated. Its new value is chosen based on the marginal probability of x_i with the other variables fixed. There are two commonly used schemes to determine which x_i to update. One scheme is to choose x_i randomly, the other is to choose x_i by sequentially scanning from x_1 to x_d.

3.3 小项目 Item Recommendation

3.4 Introduction

Recommender systems (RS) or recommendation systems (sometimes replacing "system" with a synonym such as platform or engine) are a subclass of information filtering system that seek to predict the "rating" or "preference" that a user would give to an item.

RS has shown great power to alleviate the workload of users absorbing the massive information on the Internet. Many works are proposed to improve the performance of recommending items e.g. websites, musics, social posts, articles, etc.

In this task, my goal is to conduct a comprehensive RS to predict the preference score of the given user on the specific items.

3.5 Data

A given set of vector which has user-id, item-id, the features of items, and the score that users gave to items. I need to train a model using training data and predict the score that some specific users will give to an item.

3.6 MCMC 的用途

It is used to update the model on the each step. Optimality of model parameters is usually defined with a loss function l where the task is to minimize the sum of losses over the observed data S.

$$Opt(S) = argmin \sum l(y(\hat{x}|\theta), y)$$

MCMC generates the distribution of $\hat{y}$ by Gibbs sampling.

$$\theta|X, y, \Theta/\{\theta\}, \Theta_H \ N(\mu_\theta, \sigma_\theta^2)|$$

where

$$\sigma_\theta^2 = (\alpha \sum_{i=1}^{n} h_\theta(x_i)^2 + \lambda_\theta)^{-1}$$

$$\mu_\theta = \sigma_\theta^2(\alpha\theta \sum_{i=1}^{n} h_\theta(x_i)^2 + \alpha \sum_{i=1}^{n} h_\theta^2(x_i) + \mu_\theta\lambda_\theta)$$