扔一扔贝叶斯的骰子：采样（近似推断）——从蒙特卡洛方法说起

Throwing the Bayesian dice: Sampling (approximate inference) — Starting from the Monte Carlo method

 璟明 Jingming 2025-03-30 06:37:35 美国 2025-03-30 06:37:35 United States

前言 Preface

我常常有这样的体会：一些很容易解释清楚的事，人们却总喜欢冠以很高级的名字，让外人望而却步。典型的例子是：Euclidean Distance（欧几里得距离）和Manhattan Distance（曼哈顿距离）。

I often have this experience: some things that are easy to explain are always given very high-level names to discourage outsiders. Typical examples are: Euclidean Distance and Manhattan Distance.

以二维平面为例，前者其实就是直线距离 *d* = sqrt((*x*1−*x2*)^2 + (y*1*−y2*)*^2)，后者就是各坐标轴方向上的差的绝对值之和 d=|*x1*−x2/+|*y1*−y2|，*这是初中生*都会的计算。

Taking a two-dimensional plane as an example, the former is actually the straight-line distance *d* = sqrt((*x* 1− *x2*)^2 + (y *1*− y2 *)*^ 2), and the latter is the sum of the absolute values of the differences in the directions of each coordinate axis d=| *x 1* −x *2*| +| *y1* −y2| *, which* is a calculation that *even junior high school students* can do.

若说前者的名字出自欧几里得尚可理解；后者其实是想说在曼哈顿街区行走，你只能沿着横平竖直的街道从A点到B点，因此你走过的距离就是曼哈顿距离——要按这个逻辑，是不是也能叫东京距离（Tokyo Distance）？唐代长安城早是如此，那时候曼哈顿又在哪儿呢...怎么不叫长安距离（Chang'an Distance）？

If the former is named after Euclid, it is understandable; the latter actually means that when walking in Manhattan, you can only follow the straight and horizontal streets from point A to point B, so the distance you walk is the Manhattan distance. According to this logic, can it also be called Tokyo Distance? Chang'an City in the Tang Dynasty was like this, so where was Manhattan at that time... Why not call it Chang'an Distance?

我认为，为了让知识最小阻力地传播，要极力避免预设与问题无关的先验知识——最好的哲学，是小朋友也能听懂的哲学。最好的物理/数学/...也是如此。但既然世界已经是这样了，至少让我们更好地理解它。

I think that in order to spread knowledge with the least resistance, we should try our best to avoid presupposing a priori knowledge that is irrelevant to the problem. The best philosophy is the philosophy that children can understand. The same is true for the best physics/mathematics/... But since the world is already like this, let us at least understand it better.

这篇文章将从蒙特卡洛方法（Monte Carlo Method）说起，进而介绍在贝叶斯网络中进行采样(Sampling）的几种蒙特卡洛方法变体。

This article will start with the Monte Carlo Method and then introduce several Monte Carlo method variants for sampling in Bayesian networks.

一，蒙特卡洛方法是什么？ 1. What is the Monte Carlo method?

别被名字吓到了！即便你不认识它，直觉上也必定应用过无数次了。

Don’t let the name scare you! Even if you don’t recognize it, you’ve intuitively used it countless times.

蒙特卡洛方法是指：通过生成大量随机样本，用样本的统计特性（如频率、均值）逼近目标概率或期望值。

The Monte Carlo method refers to: by generating a large number of random samples, using the statistical characteristics of the samples (such as frequency, mean) to approximate the target probability or expected value.

虽然它有很多变种，但最普通的蒙特卡罗方法就是：**P(A|B) ≈ count(A,B) / count (B)**。

Although there are many variations, the most common Monte Carlo method is: **P(A|B) ≈ count(A,B) / count (B)** .

以扔硬币为例子：我们不知道正面朝上的概率是多少，但我们可以扔**n**次，数一数正面朝上的次数**n_正**，然后估计正面朝上的概率为 **P(正) = n_正 / n**。只要n足够大，这个估计的概率就可以足够逼近目标概率。

Take tossing a coin as an example: we don't know the probability of heads, but we can toss it **n** times, count the number of heads **n_positive** , and then estimate the probability of heads as **P(positive) = n_positive/n** . As long as n is large enough, this estimated probability can be close enough to the target probability.

二，贝叶斯网络的推理：计算（精确推断）——是否总是可行?

2. Reasoning in Bayesian Networks: Computation (Exact Inference) — Is It Always Possible?

我们以这个贝叶斯网络为例：**A->B->C，每个节点的取值是0或1**

Let's take this Bayesian network as an example: **A->B->C, the value of each node is 0 or 1**

已知：**P(A), P(B|A), P(C|B)**

Given: **P(A), P(B|A), P(C|B)**

问题：**P(A=1|C=1)?**

Question: **P(A=1|C=1)?**

计算 calculate

P(A=1|C=1) = P(A=1,C=1) / P(C=1) ——这是条件概率公式，基于链式法则推导就能得到。

P(A=1|C=1) = P(A=1,C=1) / P(C=1) ——This is the conditional probability formula, which can be derived based on the chain rule.

= ∑_B P(A=1,B,C=1) / ∑_A ∑_B P(A,B,C=1) —— 左边对B的所有可能取值求和，消去B变量；右边消去B和A变量。

= ∑_B P(A=1,B,C=1) / ∑_A ∑_B P(A,B,C=1) —— The left side sums all possible values of B and eliminates the B variable; the right side eliminates the B and A variables.

= ∑_B P(A=1) * P(B|A=1) * P(C=1|A,B) / ∑_A ∑_B P(A) * P(B|A) * P(C=1|A,B) ——这也是链式法则。

= ∑_B P(A=1) * P(B|A=1) * P(C=1|A,B) / ∑_A ∑_B P(A) * P(B|A) * P(C=1|A,B) ——This is also the chain rule.

= ∑_B P(A=1) * P(B|A=1) * P(C=1|B) / ∑_A ∑_B P(A) * P(B|A) * P(C=1|B) ——为什么？因为给定B时（B作为条件），C与A条件独立，即C⊥A|B。

= ∑_B P(A=1) * P(B|A=1) * P(C=1|B) / ∑_A ∑_B P(A) * P(B|A) * P(C=1|B) ——Why? Because given B (B as a condition), C is conditionally independent of A, that is, C⊥A|B.

计算是可以得到答案，但是让我们来看看**时间复杂度**：

The answer can be obtained by calculation, but let's look at **the time complexity** :

对于这个例子，每次sum的时候要算3次乘法的2种取值，分子sum的次数是3-2=1次，分母sum的次数是3-1=2次：

For this example, each time we sum, we need to calculate 3 multiplications for 2 possible values. The number of times we sum the numerator is 3-2=1, and the number of times we sum the denominator is 3-1=2:

3*2^1 + 3*2^2

对于k种取值与n个节点，每次sum要算n次乘法的k种取值，分子sum的次数是n-2次，分母是n-1次：

For k values and n nodes, each sum needs to calculate n multiplications of k values, the number of sums of the numerator is n-2 times, and the number of sums of the denominator is n-1 times:

n*k^(n-1) + n*k^(n-2) = **O(k^n)**

计算的方法，随着节点数n的增加，时间复杂度指数增加——指数爆炸！

The calculation method, as the number of nodes n increases, the time complexity increases exponentially - exponential explosion!

如果贝叶斯网络的拓扑结构更复杂，多路径、环状结构，计算可能会非常困难。

If the topology of the Bayesian network is more complex, with multiple paths and a ring structure, the calculation may be very difficult.

三，贝叶斯网络的推理：采样（近似推断）——更容易 3. Reasoning of Bayesian Networks: Sampling (Approximate Inference) — Easier

由于计算方法的指数爆炸、难以计算等问题，对于复杂结构的贝叶斯网络，通过采样方法来近似推断容易得多。我们将介绍四种采样方法：**先验采样(Prior Sampling)**，**拒绝采样(Rejection Sampling)**, **似然加权采样(Likelihood Weighting Sampling)**, **Gibbs采样(Gibbs Sampling)。**因为它们都试图用样本的统计特性逼近目标概率，所以都属于蒙特卡洛方法的不同变种。

Due to the exponential explosion of computational methods and the difficulty of calculation, it is much easier to approximate inference through sampling methods for Bayesian networks with complex structures. We will introduce four sampling methods: **Prior Sampling** , **Rejection Sampling** , **Likelihood Weighting Sampling** , and **Gibbs Sampling.** Because they all try to use the statistical properties of samples to approximate the target probability, they are all different variants of the Monte Carlo method.

对于同样的例子： For the same example:

A->B->C，每个节点的取值是0或1 A->B->C, the value of each node is 0 or 1

已知：**P(A), P(B|A), P(C|B)**

Given: **P(A), P(B|A), P(C|B)**

为了更清楚解释，问题改成：**P(A=1|B=1)?**

For a clearer explanation, the question is changed to: **P(A=1|B=1)?**

在采样方法中，问题 P(A=1|B=1) 的条件B=1被称为证据Evidence，后面会看到它有何作用。

In the sampling method, the condition B=1 of the problem P(A=1|B=1) is called evidence. We will see its role later.

为了方便讨论，我们预设一下已知的概率吧： For the sake of discussion, let's assume the known probabilities:

A	$P(A)$
0	0.6
1	0.4

A	B	$P(B \mid A)$
0	0	0.7
0	1	0.3
1	0	0.2
1	1	0.8

B	C	$P(C \mid B)$
0	0	0.9
0	1	0.1
1	0	0.4
1	1	0.6

已知概率 Known Probability

Prior Sampling

1. 随机生成一组[0,1]的随机数，看看落在节点的哪个概率区间，决定节点的采样结果。

1. Randomly generate a set of random numbers [0,1] to see which probability interval the node falls in and determine the sampling result of the node.

假设我们生成这样一组随机数：0.5，0.6，0.95，0.8，0.9，0.98, ...

Suppose we generate a set of random numbers like this: 0.5, 0.6, 0.95, 0.8, 0.9, 0.98, ...

2. 多组采样 2. Multiple sampling

第一组采样： The first set of samples:

用0.5决定A的采样，落在P(A=0)的概率区间，因此采样A=0

The sampling of A is determined by 0.5, which falls within the probability interval of P(A=0), so the sampling A=0

用0.6决定B的采样，落在P(B=0|A=0)的概率区间，因此采样B=0

The sampling of B is determined by 0.6, which falls within the probability interval of P(B=0|A=0), so the sampling B=0

用0.95决定C的采样，落在P(C=1|B=0)的概率区间，因此C=1

The sampling of C is determined by 0.95, which falls within the probability interval of P(C=1|B=0), so C=1

第一组结果：A=0, B=0, C=1 The first set of results: A=0, B=0, C=1

第二组采样： The second set of samples:

用0.8决定A的采样，落在P(A=1)的概率区间，因此采样A=1

The sampling of A is determined by 0.8, which falls within the probability interval of P(A=1), so the sampling A=1

用0.9决定B的采样，落在P(B=1|A=1)的概率区间，因此采样B=1

Use 0.9 to determine the sampling of B, which falls within the probability interval of P(B=1|A=1), so the sampling B=1

用0.98决定C的采样，落在P(C=1|B=1)的概率区间，因此C=1

The sampling of C is determined by 0.98, which falls within the probability interval of P(C=1|B=1), so C=1

第二组结果： A=1, B=1, C=1 The second set of results: A=1, B=1, C=1

.....重复多次 ...repeat multiple times

3. 问题 P(A=1|B=1) = count(A=1,B=1) / count(B=1)，搞定。

3. Problem P(A=1|B=1) = count(A=1,B=1) / count(B=1), done.

Rejection Sampling

刚刚在Prior Sampling中，我们需要的是仅仅是 **P(A=1|B=1)** 的相关采样，即条件B=1的采样，然后计算其中A=1的比例是多少。所以对于第一组采样，显然从B=0开始就知道它不是我们需要的，所以后面的节点就不用采样了，直接拒绝这一组采样。从头开始新的一组采样。这就是Rejection的含义，节省了很多工作。

Just now in Prior Sampling, what we need is only the relevant sampling of **P(A=1|B=1)** , that is, the sampling with condition B=1, and then calculate the proportion of A=1. Therefore, for the first set of samples, it is obvious from B=0 that it is not what we need, so there is no need to sample the following nodes, and this set of samples is directly rejected. A new set of samples is started from the beginning. This is what Rejection means, saving a lot of work.

Likelihood Weighting Sampling

上面的Rejection Sampling很方便，但问题是，我们会拒绝掉很多与问题不相关的采样机会。有没有一种办法，可以不浪费掉，让每一次采样都必定有用呢？这就是Likelihood Weighting Sampling。

The above Rejection Sampling is very convenient, but the problem is that we will reject many sampling opportunities that are not relevant to the problem. Is there a way to avoid wasting them and make every sampling useful? This is Likelihood Weighting Sampling.

它的不同之处是，在每次采样Evidence时，都直接取值为问题中Evidence需要的值，并为这组采样计算一个权重 **W= P(e_1 | parents(e_1)) * P(e_2 | parents(e_2)) * .....**。最后 **P = ∑weight的比值**，而不是count的比值。

The difference is that each time the evidence is sampled, the value required by the evidence in the problem is directly taken, and a weight is calculated for this set of samples. **W= P(e_1 | parents(e_1)) * P(e_2 | parents(e_2)) * ....** at last **P = ratio of ∑weight** , rather than the ratio of count.

例子中，每次采样直接令Evidence B=1，计算W_i =1 * P(B=1|A)。

In this example, each time sampling, Evidence B=1 is directly set, and Wi =1 * P(B=1|A) is calculated.

多次重复，每一组都有用（因为令Evidence B=1了），都有一个对应的W_i。

Repeat multiple times, each group is useful (because Evidence B=1), and each has a corresponding Wi.

最后 P(A=1|B=1) = ∑weight(A=1|B=1) / ∑weight(B=1) 。

Finally, P(A=1|B=1) = ∑weight(A=1|B=1) / ∑weight(B=1) .

为什么这样可行？ Why does this work?

因为我们采样虽然使用随机值，但对Evience的采样结果符合其概率分布。既然如此，我们直接令Evidence等于我们想要的值，并把权重设为对应的概率分布。最后用权重的比值代替样本的比值计算P就好了。

Because although we use random values for sampling, the sampling results of Evience conform to its probability distribution. In this case, we directly set Evidence equal to the value we want and set the weight to the corresponding probability distribution. Finally, we use the ratio of weights instead of the ratio of samples to calculate P.

Gibbs Sampling

上面的 Likelihood Weighting Sampling 很好，但如果证据变量很多（比如同时固定 B=1, C=1），似然加权采样的权重可能变得非常小或不均匀（权重方差大），导致结果不稳定。且在网络结构复杂（比如有环）时，似然加权采样就难以处理。

The Likelihood Weighting Sampling above is very good, but if there are many evidence variables (for example, B=1, C=1 are fixed at the same time), the weights of the likelihood weighted sampling may become very small or uneven (large weight variance), resulting in unstable results. And when the network structure is complex (for example, there are loops), likelihood weighted sampling is difficult to handle.

Gibbs Sampling，它是**马尔可夫链蒙特卡洛（MCMC）**方法，适合处理多变量和复杂网络结构的情况。它的思路是：从一个初始样本开始，保持Evidence永恒不变，每次采样只更新一个变量的值，其他变量保持不变。通过不断迭代，生成一系列样本，最终逼近目标分布。

Gibbs Sampling , which is **a Markov Chain Monte Carlo (MCMC)** method, is suitable for dealing with multivariate and complex network structures. The idea is: start with an initial sample, keep the evidence constant, update the value of only one variable each time you sample, and keep the other variables constant. Through continuous iteration, a series of samples are generated, and finally the target distribution is approached.

还是例子 A->B->C，求 P(A=1|B=1)。

Still using the example A->B->C, find P(A=1|B=1).

随机初始化一组样本，比如 [A=0, B=1, C=0]（注意 B=1 是固定的证据）。每次迭代，选一个非证据变量（A 或 C）更新：

Initialize a set of samples randomly, such as [A=0, B=1, C=0] (note that B=1 is a fixed evidence). In each iteration, select a non-evidence variable (A or C) to update:

1. 更新 A：根据 P(A|B=1, C=0) 采样新值，假设得到 A=1；

1. Update A: Sample new values according to P(A|B=1, C=0), assuming that A=1;

新样本：[A=1, B=1, C=0]； New sample: [A=1, B=1, C=0];

2。 更新 C：根据 P(C|A=1, B=1) 采样新值，假设得到 C=1；

2. Update C: Sample new values according to P(C|A=1, B=1), assuming that C=1;

新样本：[A=1, B=1, C=1]； New sample: [A=1, B=1, C=1];

3. 重复多次，记录所有样本； 3. Repeat several times and record all samples;

4. 结果P(A=1|B=1) ≈ 样本中 A=1 的比例。

4. The result P(A=1|B=1) ≈ the proportion of A=1 in the sample.

注意：Gibbs 采样需要动态计算新的条件概率分布、迭代足够多次，且初始样本可能会影响早期结果，因此通常会丢弃前几步（称为“burn-in”阶段）。

Note: Gibbs sampling requires dynamically computing new conditional probability distributions, iterating enough times, and the initial samples may affect early results, so the first few steps are usually discarded (called the "burn-in" phase).

由于 Gibbs 采样需要动态计算新的条件概率分布，更适合处理多变量和复杂网络结构；而简单网络和证据较少时， Likelihood Weighting 采样可以直接用已知的条件概率分布，计算会更容易。

Since Gibbs sampling requires dynamic calculation of new conditional probability distributions, it is more suitable for dealing with multiple variables and complex network structures. When there is a simple network and less evidence, Likelihood Weighting sampling can directly use the known conditional probability distribution, which makes calculation easier.

总结 Summarize

如果整篇文章只能记住一句话.....Euclidean Distance 就是直线距离，Manhattan Distance 应该叫长安距离，Monte Carlo 就是用统计算概率，贝叶斯网络计算不了就采样。Prior就是随便扔，Rejection就是挑着用，Likelihood就是加权重，Gibbs就是不断迭代.....

If you can only remember one sentence from the entire article... Euclidean distance is the straight-line distance, Manhattan distance should be called Chang'an distance, Monte Carlo is to calculate the probability by statistics, and sampling is used when Bayesian network cannot calculate. Prior is to throw it randomly, Rejection is to use it selectively, Likelihood is to add weights, Gibbs is to iterate continuously...

完

投诉 complaint