

计算机科学188：人工智能导论

2024年春季

Note 13

作者（其他所有注释）：尼基尔·夏尔马

作者（贝叶斯网络注释）：乔希·胡格和杰基·梁，由王瑞吉编辑

作者（逻辑注释）：亨利·朱，由考佩林编辑

学分（机器学习和逻辑注释）：部分章节改编自教材《人工智能：一种现代方法》。

最后更新时间：2023年8月26日

贝叶斯网络中的近似推断：采样

概率推理的另一种方法是通过简单地对样本进行计数来隐式计算查询的概率。这不会像在IBE或变量消去法中那样得到精确解，但这种近似推断通常就足够好了，尤其是考虑到在计算上能大幅节省。


例如，假设我们想计算 $P(+t \mid +e)$ 。如果我们有一台神奇的机器可以从我们的分布中生成样本，我们可以收集所有满足 $E = +e$ 的样本，然后计算其中满足 $T = +t$ 的样本所占的比例。通过查看样本，我们就能轻松计算出任何我们想要的推断。让我们看看一些不同的生成样本的方法。

先验采样

给定一个贝叶斯网络模型，我们可以轻松编写一个模拟器。例如，考虑下面给出的仅包含两个变量 T 和 C 的简化模型的条件概率表。

T	P(T)
+t	0.99
-t	0.01

T	C	P(C T)
+t	+c	0.95
+t	-c	0.05
-t	+c	0.0
-t	-c	1.0



The diagram shows a Bayesian network with two nodes, T and C, represented as circles. A directed edge points from T to C, indicating that C is conditionally dependent on T.

用Python编写的一个简单模拟器如下所示：

```
import random

def get_t():
    if random.random() < 0.99:
        return True
    return False

def get_c(t):
    if t and random.random() < 0.95:
        return True
    return False

def get_sample():
    t = get_t()
    c = get_c(t)
    return [t, c]
```

我们将这种简单方法称为先验采样。这种方法的缺点是，为了对不太可能出现的情况进行分析，可能需要生成大量样本。如果我们想计算 $P(C \mid -t)$ ，就不得不舍弃99%的样本。

拒绝采样

缓解上述问题的一种方法是修改我们的过程，以便尽早拒绝任何与我们的证据不一致的样本。例如，对于查询 $P(C \mid -t)$ ，除非 t 为假，否则我们会避免为 C 生成一个值。这仍然意味着我们必须舍弃大部分样本，但至少我们生成的坏样本创建所需的时间更少。我们将这种方法称为拒绝采样。

这两种方法的原理相同：任何有效样本出现的概率都与联合概率密度函数中指定的概率相同。

似然加权

一种更奇特的方法是似然加权，它确保我们永远不会生成一个不好的样本。在这种方法中，我们手动将所有变量设置为与查询中的证据相等。例如，如果我们想计算 $P(C \mid -t)$ ，我们只需声明 t 为假。这里的问题是，这可能会产生与正确分布不一致的样本。

如果我们只是简单地强制某些变量等于证据，那么我们的样本出现的概率仅等于非证据变量的条件概率表（CPT）的乘积。这意味着联合概率密度函数（PDF）不能保证是正确的（尽管在某些情况下可能是正确的，比如我们的双变量贝叶斯网络）。相反，如果我们已经对从 Z_1 到 Z_p 的变量进行了采样，并固定了从 E_1 到 E_m 的证据变量，那么一个样本由概率 $P(Z_1 \dots Z_p, E_1 \dots E_m) = \prod_i^p P(Z_i \mid \text{Parents}(Z_i))$ 给出。所缺少的是，一个样本的概率并不包括 $P(E_i \mid \text{Parents}(E_i))$ 的所有概率，即并非每个条件概率表都参与其中。

似然加权通过为每个样本使用一个权重来解决这个问题，该权重是给定采样变量时证据变量的概率。也就是说，我们不是平等地计算所有样本，而是可以为样本 j 定义一个权重 w_j ，该权重反映了在给定采样值的情况下，证据变量的观察值出现的可能性。

通过这种方式，我们确保每个CPT都参与进来。为此，我们遍历贝叶斯网络中的每个变量（就像我们进行普通采样时那样），如果变量不是证据变量，则对其进行采样，如果变量是证据变量，则更改样本的权重。

例如，假设我们要计算 $P(T \mid +c, +e)$ 。对于第 j 个样本，我们将执行以下算法：

- 将 w_j 设置为1.0，将 $c =$ 设置为真，将 $e =$ 设置为真。
- 对于 T ：这不是一个证据变量，所以我们从 $P(T)$ 中对 t_j 进行采样。
- 对于 C ：这是一个证据变量，所以我们将样本权重乘以 $P(+c \mid t_j)$ ，即 $w_j = w_j \cdot P(+c \mid t_j)$ 。
- 对于 S ：来自 $P(S \mid t_j)$ 的样本 s_j 。
- 对于 E ：将样本权重乘以 $P(+e \mid +c, s_j)$ ，即 $w_j = w_j \cdot P(+e \mid +c, s_j)$ 。

然后，当我们执行常规计数过程时，我们用 w_j 而非1对样本 j 进行加权，其中 $0 \leq w_j \leq 1$ 。这种方法可行是因为在概率的最终计算中，权重有效地用于替代缺失的条件概率表。实际上，我们确保每个样本的加权概率由 $P(z_1 \dots z_p, e_1 \dots e_m) = \left[\prod_i^p P(z_i \mid \text{Parents}(z_i)) \right] \cdot \left[\prod_i^m P(e_i \mid \text{Parents}(e_i)) \right]$ 给出。下面给出了似然加权的伪代码。

函数“似然加权法 (LIKELIHOOD - WEIGHTING)”(X, e, bn, N)返回对 $P(X \mid e)$ 的估计值

输入： X ，查询变量

e ，变量 E 的观测值

bn ，一个指定联合分布 $P(X_1, \dots, X_n)$ 的贝叶斯网络

N ，要生成的样本总数

局部变量： \mathbf{W} ，一个针对 X 每个值的加权计数向量，初始值为零

从 $j = 1$ 到 N 进行循环

$\mathbf{x}, w \leftarrow \text{加权采样}(\text{bn}, e)$

$\mathbf{W}[x] \leftarrow \mathbf{W}[x] + w$ 其中 x 是 \mathbf{x} 中 X 的值

返回归一化后的 \mathbf{W}

函数加权采样(bn, e)返回一个事件和一个权重

$w \leftarrow 1; \mathbf{x} \leftarrow$ 一个具有从 e 初始化的 n 个元素的事件

对于 X_i 中每个变量 X_i ，执行以下操作

如果 X_i 是一个证据变量，在 e 中的值为 x_i

那么 $w \leftarrow w \times P(X_i = x_i \mid \text{parents}(X_i))$

否则 $\mathbf{x}[i] \leftarrow$ 是从 $P(X_i \mid \text{parents}(X_i))$ 中抽取的一个随机样本

返回 \mathbf{x}, w

图14.15 贝叶斯网络中用于推理的似然加权算法。在WEIGHTED - SAMPLE中，每个非证据变量根据给定其父母节点已采样值的条件分布进行采样，同时根据每个证据变量的似然性累积权重。

对于我们的三种采样方法（先验采样、拒绝采样和似然加权），通过生成更多样本，我们可以获得更高的精度。然而，在这三种方法中，似然加权在计算上是最有效的，原因超出了本课程的范围。

吉布斯采样

吉布斯采样是第四种采样方法。在这种方法中，我们首先将所有变量设置为某个完全随机的值（不考虑任何条件概率表）。然后，我们一次重复选择一个变量，清除其值，并根据当前分配给所有其他变量的值对其重新采样。

对于上述 T, C, S, E 示例，我们可以将 $t =$ 赋值为真， $c =$ 赋值为真， $s =$ 赋值为假， $e =$ 赋值为真。然后我们从四个变量中选择一个进行重采样，比如 S ，并将其清空。接着我们从分布 $P(S \mid +t, +c, +e)$ 中选择一个新变量。这需要我们知道这个条件分布。事实证明，给定所有其他变量，我们可以轻松计算任何单个变量的分布。更具体地说， $P(S \mid T, C, E)$ 仅使用将 S 与其邻居相连的条件概率表即可计算得出。因此，在典型的贝叶斯网络中，大多数变量只有少量邻居，我们可以在线性时间内预先计算出给定其所有邻居时每个变量的条件分布。

我们不会证明这一点，但如果我们足够多次地重复这个过程，即使我们可能从一个低概率的赋值开始，我们后来的样本最终也会收敛到正确的分布。如果你感兴趣，在维基百科文章关于吉布斯采样的“失败模式”部分，有一些超出本课程范围的注意事项可供你阅读。

下面给出了吉布斯采样的伪代码。

```
函数GIBBS - ASK( $X, e, bn, N$ )返回对  $\mathbf{P}(X \mid e)$  的估计值
  局部变量:  $\mathbf{N}$  , 一个针对  $X$  的每个值的计数向量, 初始值为零
              $\mathbf{Z}$  ,  $bn$  中的非证据变量
              $\mathbf{x}$  , 网络的当前状态, 初始时从  $e$  复制

  用 $\langle b1 \rangle$ 中变量的随机值初始化  $\mathbf{x}$ 
  从 $\langle b0 \rangle$ 到 $\langle b1 \rangle$ 进行循环
    对 $\langle b1 \rangle$ 中的每个 $\langle b0 \rangle$ 执行以下操作
      通过从 $\langle b2 \rangle$ 中采样来设置 $\langle b1 \rangle$ 中 $\langle b0 \rangle$ 的值
       $\langle b0 \rangle$ , 其中 $\langle b1 \rangle$ 是 $\langle b3 \rangle$ 中 $\langle b2 \rangle$ 的值
  返回NORMALIZE( $N$ )
```

图14.16 用于贝叶斯网络近似推理的吉布斯采样算法；此版本按变量循环，但随机选择变量也可行。

结论

总之，贝叶斯网络是联合概率分布的一种强大表示。其拓扑结构编码了独立性和条件独立性关系，我们可以用它来对任意分布进行建模，以执行推理和采样。

在本笔记中，我们介绍了两种概率推理方法：精确推理和概率推理（采样）。在精确推理中，我们能保证得到精确正确的概率，但计算量可能大得令人望而却步。

所涵盖的精确推理算法如下：

- 枚举推理
- 变量消除

我们可以转向采样来近似求解，同时减少计算量。

所涵盖的采样算法如下：

- 先验采样
- 拒绝采样
- 重要性加权
- 吉布斯采样