贝叶斯网（k2）

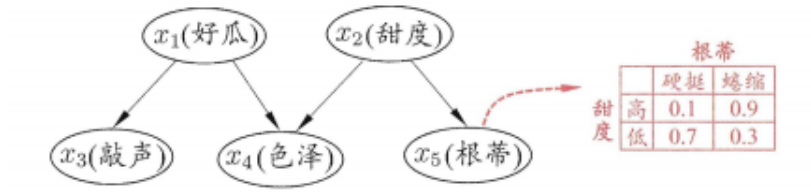
谢欣然

2021 年 3 月 24 日

1 原理

贝叶斯网络 (Bayesian network)，又称信念网络 (Belief Network)，它借住有向无环图（DAG）来刻画属性之间的依赖关系，并使用条件概率表（CPT）来描述属性的联合概率分布。

具体来说，一个贝叶斯网 B 由结构 G 和参数 θ 两部分构成，即 $B =< G, \theta >$, 网络结构 G 是一个有向无环图，其中每个节点对应一个属性，若两个属性之间有直接的依赖关系，则它们由一条边连接起来；参数 θ 定量描述了这种依赖关系，假设属性 x_i 在 G 中的父节点集为 π_i, 则 θ 包含了每个属性的条件概率表 $\theta_{x_i|\pi_i} = P_B(x_i|\pi_i)$



1.1 结构

叶斯网结构有效地表达了属性间的条件独立性，给定父结点集，贝叶斯网假设每个属性与它后裔属性独立，于是 $B =< G, \theta >=$ 将属性 $x_1, x_2 ... x_d$ 的联合概率分布定义为：

$$P_B(x_1, x_2, ...x_d) = \prod_{i=1}^{d} P_B(x_i|\pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i} \tag{1}$$

以上图为例，联合概率分布可定义为：

$$P_B(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2) \qquad (2)$$

1.2 学习

若网络结构已知，即属性间的依赖关系已知，则贝叶斯网的学习过程相对简单，只需通过对训练样本” 计数”，估计出每个结点的条件概率表即可. 但在现实应用中我们往往并不知晓网络结构。于是，贝叶斯网学习的首要任务就是根据训练数据集来找出结构最” 恰当” 的贝叶斯网。” 评分搜索” 是求解这一问题的常用办法. 具体来说，我们先定义一个评分函数 (score function) ，以此来评估贝叶斯网与训练数据的契合程度，然后基于这个评分函数来寻找结构最优的贝叶斯网. 显然，评分函数引入了关于我们希望获得什么样的贝叶斯网的归归纳偏好。

k2 算法:因为 n 个节点的有向无环图有 $G(n) = \sum_n^{k=1}(-1)^{k+1}\binom{n}{k}2^{k(n-k)}G(n-k)$, 但是得通过遍历所有情况才能得到最优解，这几乎是不可能的。所以我们采用拓扑序列，这样只需要遍历 $G(n) = n!$ 个有向无环图即可，可以得到大大的化简。虽然一个属于拓扑排序的 DAG 不是唯一的，但也许我们可以用启发式或评分准则在局部搜索。此处我们采用的评分函数是 k2 评分。

$$\sum_{i=1}^{n}\sum_{j=1}^{q_i}\{\log\frac{(r_i-1)!}{(m_{ij*}+r_i-1)!}\sum_{k=1}^{r_i}\log(m_{ijk}!)\}$$

D：数据集

n：样本变量的数目

qi：样本父节点的取值数目

mijk：属性为 i，父节点为 j，i 属性取值为 k 的时候样本的个数。

给定一个拓扑排序，我们可以迭代地检查每个节点，看看添加父节点的边是否会增加 k2 的分数. 最终得分是各节点 log 分数的和.

1.3 推断

贝叶斯网训练好之后就能用来回答” 查询” (query)，即通过一些属性变量的观测值来推测其他属性变量的取值. 例如在西瓜问题中，若我们观测到

西瓜色泽青绿、敲声烛响、根蒂蜷缩，想知道它是否成熟、甜度如何. 这样通过已知变量观测值来推测待查询变量的过程称为” 推断” (inference) ，已知变量观测值称为” 证据” (evidence).

最理想的是直接根据贝叶斯网定义的联合概率分布来精确计算后验概率，不幸的是，这样的” 精确推断” 已被证明是 NP 难的。当网络结点较多、连接稠密时难以进行精确推断，此时需借助” 近似推断”，通过降低精度要求，在有限时间内求得近似解. . 在现实应用中，贝叶斯网的近似常使用吉布斯采样，这是一种随机采样方法.

如图 7.5 所示，吉布斯采样算法先随机产生一个与证据 E=e 一致的样本,q_0 作为初始点，然后每步从当前样本出发产生下一个样本. 具体来说，在第 t 次来样中，算法先假设 $q_t = q_{t-1}$，然后对非证据变量逐个进行采样改变其取值，采样概率根据贝叶斯网 B 和其他变量的当前取值 (即 Z= z) 计算获得. 假定经 T 次采样得到的与 q 一致的样本共有 n_q 个，则可近似估算出后验概率:

$$P(Q = q|E = e) \simeq \frac{n_q}{T} \tag{3}$$

输入: 贝叶斯网 $B = \langle G, \Theta \rangle$;
采样次数 T;
证据变量 $\mathbf{E}$ 及其取值 $\mathbf{e}$;
待查询变量 $\mathbf{Q}$ 及其取值 $\mathbf{q}$.

过程:

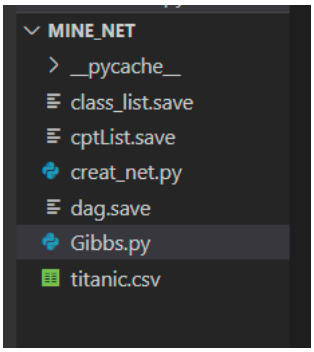
$n_q = 0$
$\mathbf{q}^0 =$ 对 $\mathbf{Q}$ 随机赋初值
for $t = 1, 2, \ldots, T$ **do**
  for $Q_i \in \mathbf{Q}$ **do**
    $\mathbf{Z} = \mathbf{E} \cup \mathbf{Q} \setminus \{Q_i\}$;
    $\mathbf{z} = \mathbf{e} \cup \mathbf{q}^{t-1} \setminus \{q_i^{t-1}\}$;
    根据 B 计算分布 $P_B(Q_i \mid \mathbf{Z} = \mathbf{z})$;
    $q_i^t =$ 根据 $P_B(Q_i \mid \mathbf{Z} = \mathbf{z})$ 采样所获 Q_i 取值;
    $\mathbf{q}^t =$ 将 $\mathbf{q}^{t-1}$ 中的 q_i^{t-1} 用 q_i^t 替换
  end for
  if $\mathbf{q}^t = \mathbf{q}$ **then**
    $n_q = n_q + 1$
  end if
end for

输出: $P(\mathbf{Q} = \mathbf{q} \mid \mathbf{E} = \mathbf{e}) \simeq \frac{n_q}{T}$

图 7.5 吉布斯采样算法

2 代码

2.1 代码结构/运行方式



其中运行 creat-net.py 文件可以依据 k2 算法对数据集 titanic 进行贝叶斯网的构建，并且将生成的 dag 矩阵保存至 dag.save 文件以方便调用。同理生成并保存 cptList.save 和 class-list.save 文件。Gibbs.py 文件是一个单独的文件，实现了 Gibbs 采样（推断）的过程。

2.2 实验数据-titanic

对 titanic 的数据集进行处理, 处理成以下属性和取值情况：

”age”:[1,2,3]=> 年龄插补后分 0,1,2 代表幼年（0-15）成年（15-55）老年（55-）

”portembarked”:[1,2,3]

”fare”:[1,2,3] => 船票价格经聚类变 0 1 2 代表少多很多”numparentschildren”:[1,2,3]

”passengerclass”:[1,2,3] => 客舱等级 (0/1/2 等舱位)

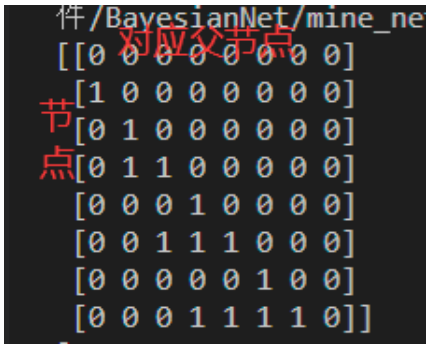
”sex”:[1,2] => 性别 male=1 female=0

”numsiblings”:[1,2,3]

”survived”:[1,2]

2.3 贝叶斯网络–k2 搭建 net

2.3.1 DAG 图

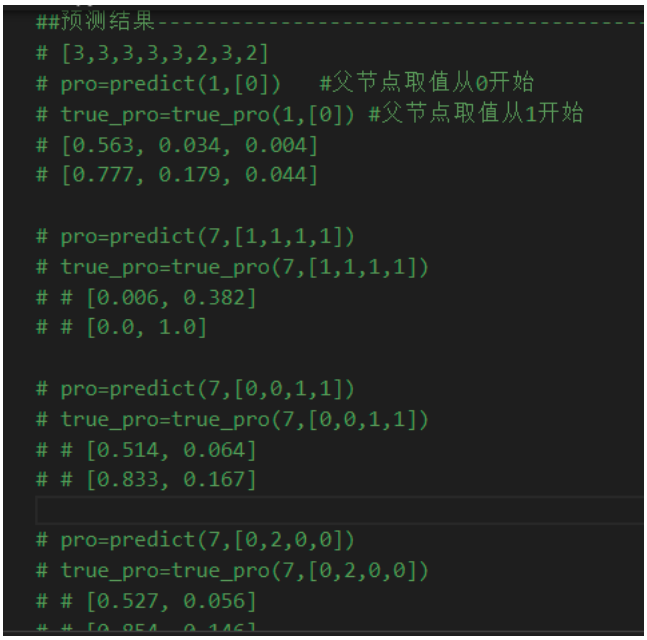


2.3.2 条件概率

根据数据集和计算出的 DAG 关系图，计算每个类别的条件概率以及其类概率。保存到 cptList 里面。

2.4 贝叶斯网络–推断

这里采用每个类别的先验概率以及其类概率，计算出给定情况下的预测值。



参考链接：

https://github.com/AaronHavens/structure_learning

https://github.com/FyuNaru/HIT-artificial-intelligence/tree/master/ai_lab3

https://blog.csdn.net/qq_36739040/article/details/105465486

https://blog.csdn.net/qq_43497779/article/details/102985820

https://www.cnblogs.com/pinard/p/6645766.html

https://github.com/luabras/Bayesian-network-weighted-and-Gibbs-sampling-