中国科学技术大学
University of Science and Technology of China

人工智能/机器学习/数据挖掘

贝叶斯网络–从概率论看机器学习

April 14, 2022

Outline

随机变量

随机变量的定义

- 一个随机试验可能结果（称为基本事件）的全体组成一个基本空间 Ω。随机变量 X 是定义在基本空间 Ω 上的取值为实数的函数。

进一步理解

- 和传统变量定义一样，随机变量 X 首先定义在一个集合上，该集合称为随机变量的”值域”，可以解释为 X 表示将来会发生的某个事件，而值域就是所有可能事件的集合。
- 与传统变量定义不一样，随机变量 X 在其每个可能取值上都附加了一个“特征”：取到该值的可能性，可以理解为：该值对应的事件重复发生了很多次，该值出频率或比例。

完整描述一个随机变量

- 需要用两个集合来描述:
- 取值范围/值域：$X \in A = \{a_1, a_2, \ldots\}$
- 特征：$p(X) = \{p(a_1), p(a_2), \ldots\}$

连续型和离散型随机变量

连续型随机变量

- 连续型随机变量，若值域，即集合 A，由不可数的集合，如一段连续的区间组成。

离散型随机变量

- 离散型随机变量，若值域，即集合 A，由有限或可列的元素组成。通常离散随机变量的值域可以映射到自然数集合或自然数的某个有限子集。

随机变量的“特征”：概率

- 离散型：概率质量函数/Probability Mass Function/PMF, 每个值对应的概率;
- 连续型：每个取值对应的概率都是 0，但是我们还是给出一个非 0 的概率，表示在该值附近的一个小区间内取值的概率。概率密度函数/Probability Density Function/PDF, 每个值对应的概率（虽然无法一一列出）;
- 在不引发二义时，因计算机本质上是处理离散信息，我们的讨论将概率密度函数和概率质量函数都统一称为“概率质量函数”。

概率质量函数的表示

函数与表格

- 给定函数 $y = f(x), x \in \{v_1, v_2, \ldots, v_n\}$，可用如下表格来完全描述该函数

x	$y = f(x)$
v_1	$f(v_1)$
v_2	$f(v_2)$
$\ldots$	$\ldots$
v_n	$f(v_n)$

Table: 用表格表示任意函数

函数与表格

- 给定随机变量 X，及其概率质量函数 $p(X)$，可用如下表格来完全描述随机变量：

X	$p(X)$
a_1	$p_1 = p(a_1)$
a_2	$p_2 = p(a_2)$
$\ldots$	$\ldots$
a_n	$p_n = p(a_n)$

Table: 用表格表示概率质量函数

随机变量 X 的概率质量函数可以用表格右侧列来简化表示（假定表格左边的列序是固定的），即“概率向量”：$(p_1, p_2, \ldots, p_n)$

知识提升：随机变量的压缩描述

问题

- 用两个集合/向量来描述一个随机变量太复杂（描述长度），可以简化/压缩随机变量的描述吗？

例子

- 期望：$\sum_i p(X = a_i)a_i$ 两个向量（值域和每个值对应的概率特征）被综合成一个值。
- 类似期望，还有有方差，高阶矩等。
- 均值：不考虑概率向量，仅仅关注值域 $\sum_i a_i/n$；
- 信息熵：$-\sum_i p(X = a_i)\log_2 p(X = a_i)$，概率特征向量被综合成一个值，不考虑值域。

随机向量

随机向量的定义

- 随机向量: 多个相关或不相关的随机变量构成的向量。

理解

- 联合概率质量/密度函数：随机向量的任何一个取值（每个随机分量取一个值构成一个值向量）都对应一个概率，离散时称联合概率质量函数，连续时称联合概率密度函数。

工程实践问题

- 长度为 n 的随机向量，每个分量有 2 个取值，这样的联合概率质量函数（值域）有多大？或者说描述该联合概率质量函数的概率向量有多少维？
- 2^n，当 n 很大时，该联合概率质量函数在计算机中的“穷举”式表示存在问题，存储代价和遍历处理时间开销变得不可行。

联合概率质量函数的表格表示

随机向量的联合概率质量函数

- $X = (X_1, X_2, \ldots, X_n)$，第 i 个随机分量的值域大小记为 $|X_i|$，其第 j 个取值为 v_{ij} ，则有如下表格表示联合概率质量函数：

X_1	X_2	$\ldots$	X_n	$p(X_1, X_2, \ldots, X_n)$
v_{11}	v_{21}	$\ldots$	v_{n1}	0.00003
v_{11}	v_{21}	$\ldots$	v_{n2}	0.000001
$\ldots$	$\ldots$	$\ldots$	$\ldots$	...
$v_{1\|X_1\|}$	$v_{2\|X_2\|}$	$\ldots$	$v_{n\|X_n\|}$	0.000002

Table: 联合概率质量函数的表格表示

分析与思考

- 当 X_i 表示事物的“因”或“果”的时候，我们可以用来进行与事物相关的一些“推理”活动；
- 进一步，这种推理的价值和困难？表格有多少行？

条件概率

复杂的推理问题

- 如上一页的联合概率质量函数表格，假设已知 $X_2 = v_{21}$，而其他的 $X_j, j > 2$ 都未知，能否推断出 $X_1 =?$
- 这就引入了所谓的条件概率。

条件概率

- 定义：两个事件 A, B，在事件 A 发生的条件下，事件 B 发生的概率，记为 $p(B|A) = \frac{p(AB)}{p(A)}$.(当事件 A 发生的概率不为 0 时，上述定义有效)

带条件概率的推理问题

- $X^* = \arg max_a P(X_1 = a | X_2 = v_{21})$
- 如何用代码实现这个公式？从表中查找出所有 $X_2 = v_{21}$ 行，将这些行按 X_1 的不同取值分组，求每组内的概率和，求概率和的最大值对应的组对应的 X_1 的值。
- 对本 slides 中的任何一个概率/概率题，尝试编写代码实现。

核心问题

随机向量联合概率质量函数的表格表示

X_1	X_2	$\ldots$	X_n	$p(X_1, X_2, \ldots, X_n)$
v_{11}	v_{21}	$\ldots$	v_{n1}	0.00003
v_{11}	v_{21}	$\ldots$	v_{n2}	0.000001
$\ldots$	$\ldots$	$\ldots$	$\ldots$	...
$v_{1\|X_1\|}$	$v_{2\|X_2\|}$	$\ldots$	$v_{n\|X_n\|}$	0.000002

Table: 联合概率质量函数的表格表示

函数 f 遇上概率论: f 多了一列概率值

- 函数 f 如何获得？即学习问题，具体来说，函数 f 就是完整联合概率质量函数/表格，其行数是随机分量数目的指数函数，存储该表格的空间复杂度问题，获得该表格的方法问题等。
- 函数 f 如何使用？即搜索解决问题的方法，具体来说，就是基于函数 f，即完整联合概率质量函数/表格，实现概率推理计算/算法，以及算法的时间复杂度等方面的考虑。

f 的复杂性

随机向量联合概率质量函数的表格表示

X_1	X_2	$\ldots$	X_n	$p(X_1, X_2, \ldots, X_n)$
v_{11}	v_{21}	$\ldots$	v_{n1}	0.00003
v_{11}	v_{21}	$\ldots$	v_{n2}	0.000001
$\ldots$	$\ldots$	$\ldots$	$\ldots$	...
$v_{1\|X_1\|}$	$v_{2\|X_2\|}$	$\ldots$	$v_{n\|X_n\|}$	0.000002

Table: 联合概率质量函数的表格表示

存储需求/空间复杂度

- 表格共有 $|X_1| \cdot |X_2| \cdot \ldots \cdot |X_n|$ 行，是列数 n 的指数函数
- 现实应用中，通常没有上述完整的表格，该怎么办？上述表格不全（包括行不全、列不全）怎么办？事物的复杂性，相关因果成百上千，上述表格无法保存和精确制作出来。
- 所谓“概率论下的机器学习”，其目标就是找到方法将上述表格表示和制作出来。

概率推理/不确定性推理

随机向量联合概率质量函数的表格表示

X_1	X_2	$\ldots$	X_n	$p(X_1, X_2, \ldots, X_n)$
v_{11}	v_{21}	$\ldots$	v_{n1}	0.00003
v_{11}	v_{21}	$\ldots$	v_{n2}	0.000001
$\ldots$	$\ldots$	$\ldots$	$\ldots$	...
$v_{1\|X_1\|}$	$v_{2\|X_2\|}$	$\ldots$	$v_{n\|X_n\|}$	0.000002

Table: 联合概率质量函数的表格表示

时间复杂度

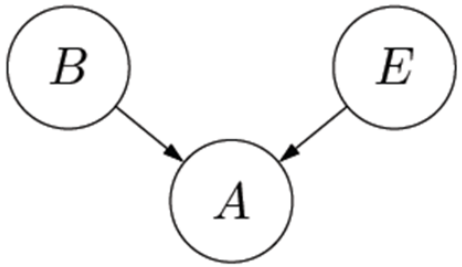
- 假设 $X_1 \in \{+1, -1\}$ 是类标号，任意给定一个 $(X_2, X_3, \ldots, X_n)$ 的值 $(v_2, v_3, \ldots, v_n)$，求 X_1 的值，这就是分类问题，你能从上述联合概率质量函数中给出分类结果吗？
- 分类非常简单，查询上述表格，比较 $p(+1, v_2, v_3, \ldots, v_n)$ 和 $p(-1, v_2, v_3, \ldots, v_n)$ 二者的大小即可。
- 假设 $X_2 = 3$，能得到 $X_1 =?$，这就是求条件概率问题。
- 典型的计算过程：遍历一遍表格，统计 $X_2 = 3$ 的那些行的概率值之和 p，同时完成对这些行的分组 (按 X_1 的取值不同分组)，计算每组的概率值之和 q_i，然后计算条件概率 $\frac{q_i}{p_i}$，然后比较条件概率值的大小，取最大条件概率值对应的组的 X_1 的取值为最终结果。整个计算时间代价在遍历函数 f/完整的联合概率质量函数表格，时间代价太大!

f 遇上概率论

出现了新情形

- 知道了联合概率质量函数 (PMF) 的所有信息，就可以进行任意的概率值计算，对应某一类不确定性推理活动；
- 实践难题 1：如何表示和存储 PMF？数学上，通常假设 PMF 有数学的解析表达式，我们该如何给出这种解析表达式？
- 实践难题 2：如何计算某个概率值？如果采用表格存储 PMF，时间代价？空间代价？

Bayes 网络的例子



网络结构

b	$p(b)$
1	ϵ
0	$1-\epsilon$

e	$p(e)$
1	ϵ
0	$1-\epsilon$

边缘分布

b	e	a	$p(a \mid b,e)$
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

条件概率质量函数

B=1: 发生入室盗窃，
否则 B=0

E=1: 发生地震，
否则 E=0

A=1: 警铃响了，
否则 A=0

b	e	a	$\mathbb{P}(B=b,E=e,A=a)$
0	0	0	$(1-\epsilon)^2$
0	0	1	0
0	1	0	0
0	1	1	$(1-\epsilon)\epsilon$
1	0	0	0
1	0	1	$\epsilon(1-\epsilon)$
1	1	0	0
1	1	1	ϵ^2

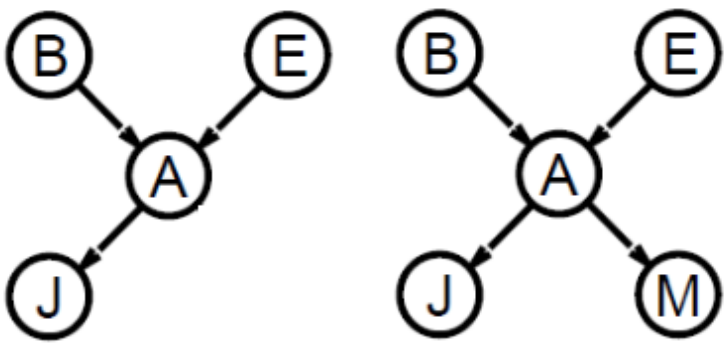
联合概率质量函数

约简：联合概率质量函数 = 网络结构 + 若干边缘分布 + 若干条件概率质量函数

例子的讨论

存在问题

- 为什么说二者是等价的？即：“联合概率质量函数 = 网络结构 + 若干边缘分布 + 若干条件概率质量函数”能证明吗？
- 如果等价，二者如何相互转换？
- 实现了约简吗？即降低了存储代价吗？(该例子没有！)，下图中的例子呢？



解释说明

- J=1: john 打电话通知我警铃响了，否则 J=0；M=1：mary 打电话通知我警铃响了，否则 M=0
- 前一个例子比较：联合概率质量函数 8 行，而约简的 Bayes 网络表示：12 行 + 一个网络/图
- 左边例子：联合概率质量函数 16 行，而约简的 bayes 网络表示：16 行 + 一个网络/图
- 右边例子：联合概率质量函数 32 行，而约简的 bayes 网络表示：20 行 + 一个网络/图

Bayes 网络转换为联合概率质量函数

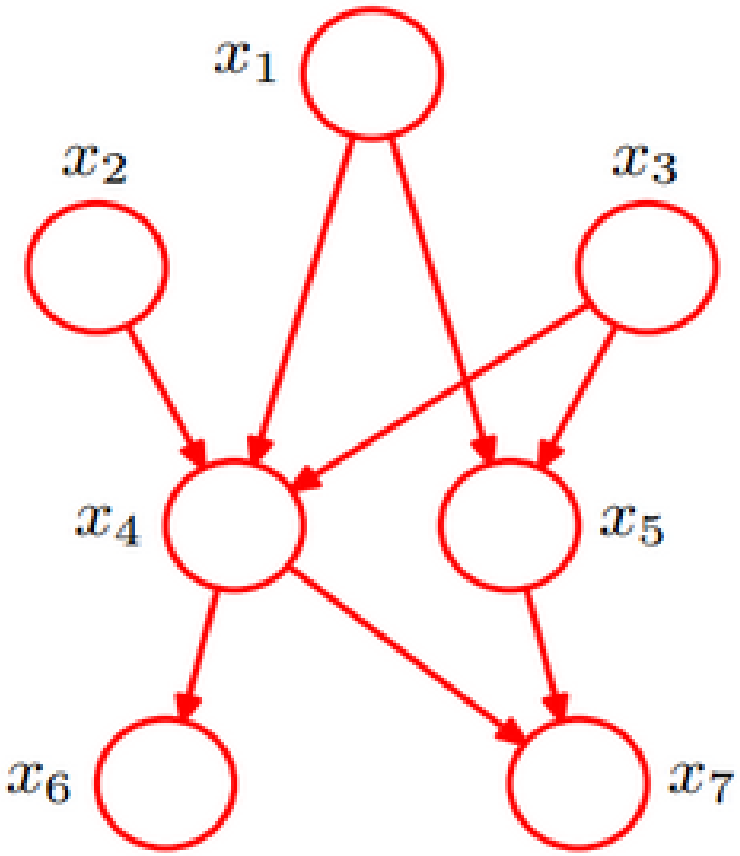
步骤与方法

- 网络结构 $\Longrightarrow$ 联合概率的计算公式，一个乘积表达式；
- 边缘概率质量函数和条件概率质量函数，给上述乘积表达式中每个乘积因子提供一个特定的值。

进一步解释说明

- 对任意联合概率（也就是联合概率质量函数的任何一行），我们都可以用上述方法计算出其概率值；因此，我们从 Bayes 网络可以恢复出整个联合熵质量函数；
- 核心：乘积表达式的合理性来自随机分量的独立性假设。$p(AB) = p(A)p(B)$，当 A, B 独立时成立。
- 网络结构描述了随机分量之间的独立性。
- 巧妙地用一些小的“表格/函数”相乘（SQL 中的连接操作，笛卡儿积），获得大的完整的联合概率质量函数/表格。

从网络结构到乘积公式的例子



解释说明

- 开始之前，先思考存储代价降低了多少。
- 看左图，写乘积公式的规律/方法是什么？
- 原理是什么？
  - （条件）概率计算的链式法则
  - （条件）独立性

$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

链式法则

链式法则：用条件概率和边缘概率计算联合概率的通用方法

$$
\begin{aligned}
& p(x_1, \ldots, x_n) = p(x_n|x_1, \ldots, x_{n-1}) \cdot p(x_1, \ldots, x_{n-1}) \\
&= p(x_n|x_1, \ldots, x_{n-1}) \cdot p(x_{n-1}|x_1, \ldots, x_{n-2}) \cdot p(x_1, \ldots, x_{n-2}) \\
&\ldots \\
&= p(x_n|x_1, \ldots, x_{n-1}) \cdot p(x_{n-1}|x_1, \ldots, x_{n-2}) \cdot \ldots \cdot p(x_1)
\end{aligned}
$$

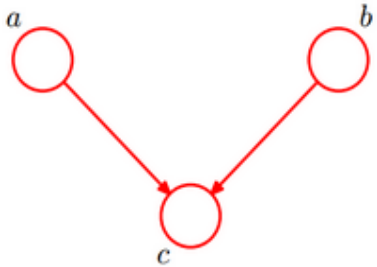
链式法则：应用的例子

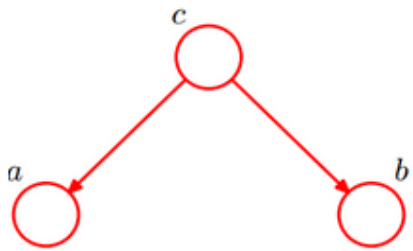
$$
\begin{aligned}
& p(x_1, \ldots, x_7) = p(x_7|x_1, \ldots, x_6) \cdot p(x_1, \ldots, x_6) \\
&= p(x_7|x_1, \ldots, x_6) \cdot p(x_6|x_1, \ldots, x_5) \cdot p(x_1, \ldots, x_5) \\
&\ldots \\
&= p(x_7|x_1, \ldots, x_6) \cdot p(x_6|x_1, \ldots, x_5) \cdot \ldots \cdot p(x_1)
\end{aligned}
$$

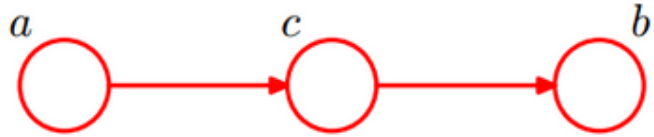
链式法则应用后，如何得到下述的公式？

- $p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$
- 结合网络结构图，应用独立性，去掉条件中不相关的随机分量。

条件独立性







条件独立

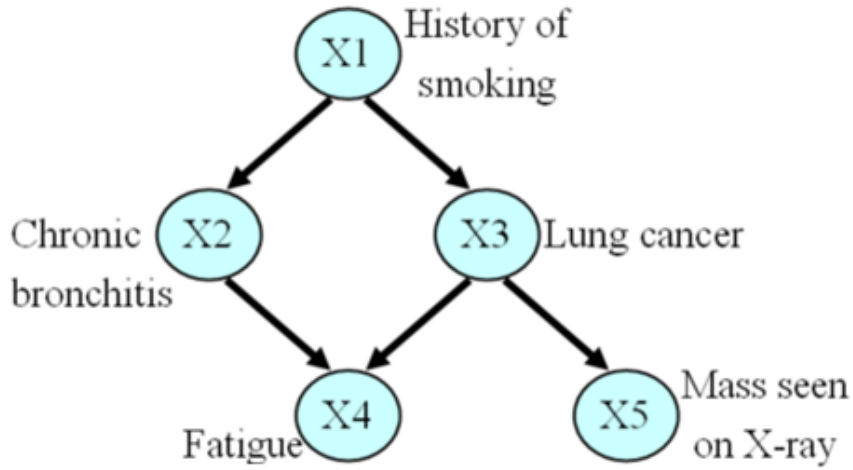
- 先写出各个图形表示的联合概率质量函数的乘积表达式，然后用后面的条件独立公式来证明/约简链式法则的结果。
- 情形 1：$p(a,b,c) = p(a)p(b)p(c|a,b)$，因为 a,b 在 c 未知的条件下独立：$p(a,b) = p(a)p(b)$，绝对独立。
- 情形 2：$p(a,b,c) = p(c)p(a|c)p(b|c)$，因为 a,b 在 c 已知/给定的条件下独立：$p(a,b|c) = p(a|c)p(b|c)$
- 情形 3：$p(a,b,c) = p(a)p(c|a)p(b|c)$，因为 a,b 在 c 已知/给定的条件下独立：$p(a,b|c) = p(a|c)p(b|c)$
- 一句话：网络结构图中，不直接连边的节点之间条件独立

解释了为什么从链式法则，可得到最终的乘积公式。

又一个 Bayes 网络的例子

把所有的边缘概率值，条件概率值都列出来

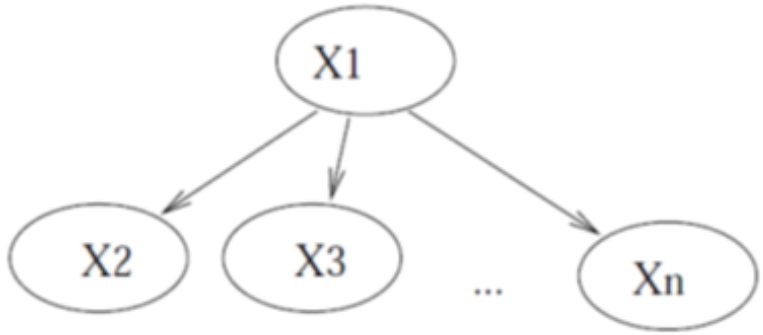
网络结构和这些概率值就构成了联合概率质量函数的约简表示



P(X1=no)=0.8
P(X2=absent | X1=no)=0.95
P(X2=absent | X1=yes)=0.75
P(X3=absent | X1=no)=0.99995
P(X3=absent | X1=yes)=0.997
P(X4=absent | X2=absent, X3=absent) =0.95
P(X4=absent | X2=absent, X3=present) =0.5
P(X4=absent | X2=present, X3=absent) =0.9
P(X4=absent | X2=present, X3=present) =0.25
P(X5=absent | X3=absent)=0.98
P(X5=absent | X3=present)=0.4

P(X1 = yes)=0.2
P(X2=present | X1=no)=0.05
P(X2=present | X1=yes)=0.25
P(X3=absent | X1=no)=0.00005
P(X3=absent | X1=yes)=0.003
P(X4=present | X2=absent, X3=absent) =0.05
P(X4=present | X2=absent, X3=present) =0.5
P(X4=present | X2=present, X3=absent) =0.1
P(X4=present | X2=present, X3=present) =0.75
P(X5=present | X3=absent)=0.02
P(X5=present | X3=present)=0.6

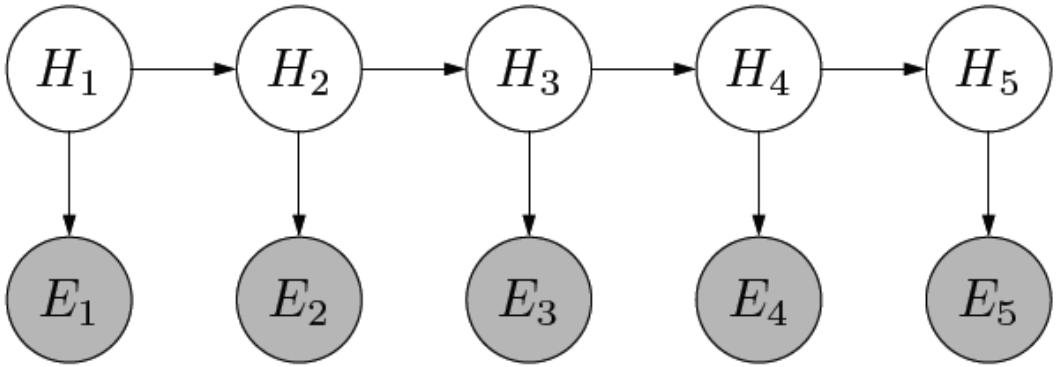
朴素 Bayes 模型



解释说明

- 如图所示的特殊网络结构，其中 X_1 一般称为“cause”或类标签，其他的 $X_i, i > 1$ 称为观测值，现象，效果等；
- 联合概率计算公式为：
  $p(X_1, X_2, \ldots, X_n) = p(X_1) \cdot p(X_2|X_1) \cdot \ldots \cdot p(X_n|X_1)$
- 朴素 Bayes 模型的应用：分类或推理，就是已知 $X_2, X_3, \ldots, X_n$，求 X_1；在概率论框架下，我们转换为求 $p(X_1|X_2, \ldots, X_n)$，计算概率值！

隐马尔可夫模型



解释说明

- 如图，$H = (H_1, H_2, H_3, H_4, H_5)$，$E = (E_1, E_2, E_3, E_4, E_5)$，$H$ 是隐藏的随机分量，看不见/无法观测的 "causes"，E 是可观测的随机分量，在某个时刻，也许只能观测到一部分，比如前 k 个 E_i
- 联合概率计算公式：

$$p(H = h, E = e) = p(h_1)\Pi_{i=2}^{n} p(h_i|h_{i-1})\Pi_{i=1}^{n} p(e_i|h_i)$$

Bayes 网络的学习

问题描述

- 已知条件：给定数据集 $\mathcal{D}$
- 目标：Bayes 网络：包括网络结构和各种边缘/条件概率质量函数/表

基本思想

- 分别解决两个子问题：
  - 网络结构的确定。
  - 概率质量函数的确定。

网络结构的确定

问题描述

- 给定数据集 $\mathcal{D}$（有时有专家/知识库存在），求网络结构（变量之间的依赖关系）。

方法分类

- 依靠专家建模，手工给出网络结构
- 从数据集中自动学习网络结构
  - 利用卡方/互信息等做相关性测试。如用数据集验证变量间是否独立/条件独立，从完全图开始，逐步删除独立/条件独立变量之间的边；要验证的次数随变量个数指数增加；
  - 基于搜索-评分的方法。一般从无边图开始（变量两两独立的假设模型），逐步添加变量之间的边（每次得到一个新假设模型），然后对每个新模型评分（预设的评分函数，用来验证模型的优劣），依据分数选择一个最好的模型，这个步骤多次迭代。
- 结合二者的混合方法

概率质量函数的确定

问题描述

- 给定数据集 $\mathcal{D}$ 和网络结构（图 G），求各边缘/条件概率质量函数/表。

方法思想

- 利用蒙特卡洛逼近。将数据集视为“粒子”集，然后计算网络结构需要的各种边缘/条件概率质量函数/表。

求概率质量函数例子—单变量

问题描述

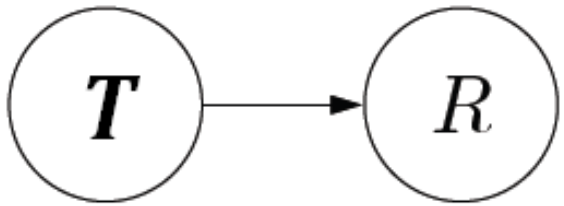
- 数据集：$D = \{1, 5, 4, 4, 3, 5, 5, 4, 4, 4\}$，表示对一个餐馆的打分；
- 随机变量 $R \in \{1, 2, 3, 4, 5\}$ 表示餐馆的分数，求 $(p_1, p_2, p_3, p_4, p_5)$

求解过程 蒙特卡罗逼近：

- 分别统计每个分数出现的次数 c_i 以及数据总数 c
- 得到 $(p_1, p_2, p_3, p_4, p_5) = (c_1/c, c_2/c, c_3/c, c_4/c, c_5/c) = (0.1, 0, 0.1, 0.5, 0.3)$，即用频度近似概率值

求概率质量函数例子—两变量

$p(t,r) = p(t)p(r|t)$



给定的网络结构

问题描述

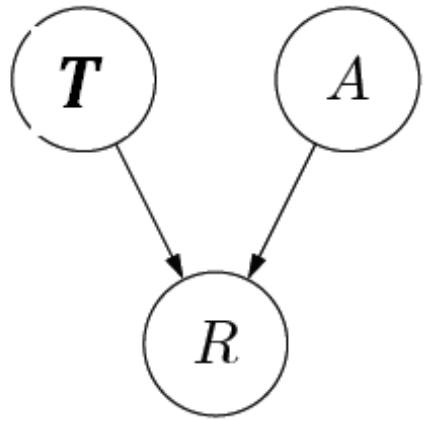
- 数据集：$D = \{(c,4),(c,4),(c,5),(s,1),(s,5)\}$
- 随机向量 $X = \{T, R\}$，随机分量 $R \in \{1,2,3,4,5\}$ 表示对一个餐馆的评分，随机分量 $T \in \{c,s\}$ 表示餐馆类型，假设类型是决定餐馆评分的唯一标准（简化）

求解过程

- 统计频度当边缘概率 $p(T) = (0.6, 0.4)$
- 令 $T = c$ 或 s 时，分别统计各个打分的频度，得到条件概率 $p(R|T) = ((0,0,0,0.67,0.33),(0.5,0,0,0,0.5))$

求概率质量函数粒子—三变量

$$p(t,r,a) = p(a)p(t)p(r|t,a)$$



给定的网络结构

问题描述

- 数据集：$D = \{(c,0,3),(c,1,5),(s,0,1),(s,0,5),(s,1,4)\}$
- 随机向量：$X = \{T, A, R\}$，A 表示是否获得认证，其他变量含义如前所述

求解过程

- 统计频度当边缘概率 $p(T) = (0.4, 0.6), p(A) = (0.6, 0.4)$
- 对 (T, A) 的任意给定的组合取值，统计 R 的出现频度当成条件概率
  $p(R|T, A) = ((0,0,1,0,0)_{c0}, (0,0,0,0,1)_{c1}, (\ldots)_{s0}, (\ldots)_{s1})$
- 条件概率质量函数中出现了大量的 0，这是非常不合理/不准确地对概率值的估计，有没有改良的办法？

拉普拉斯平滑

问题描述

- 条件概率表中存在很多不合理的估计值 0，原因是样本数量太少。如何处理？

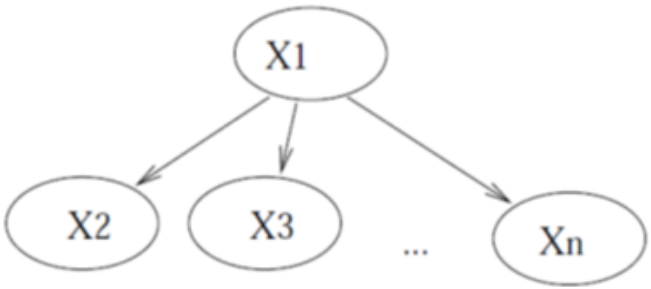
解决办法：拉普拉斯平滑

- 去掉这些 0 的方法称为平滑技术
- 若在统计每个值出现次数时，计数器的初始值设置为非零，这就是所谓的“拉普拉斯平滑”
- 通常设置计数器初始值为 1
- 如右图的例子，数据集为 $D = \{(d, 4), (d, 5), (c, 5)\}$

x_1	x_2	$p_2(x_2 \mid x_1)$
d	1	1/7
d	2	1/7
d	3	1/7
d	4	2/7
d	5	2/7
c	1	1/6
c	2	1/6
c	3	1/6
c	4	1/6
c	5	2/6

x_1	$p_1(x_1)$
d	3/5
c	2/5

求概率质量函数例子—朴素 Bayes 网络



给定的 Bayes 网络结构

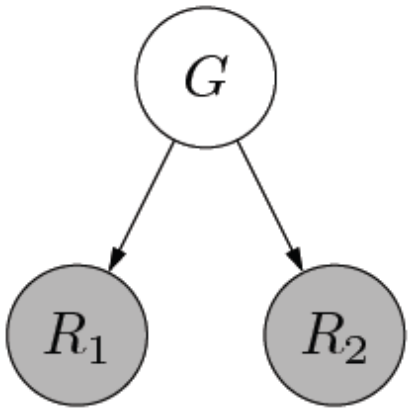
问题描述

- 给定数据集 D 和 Bayes 网络结构如图，求边缘/条件概率质量函数/表。

求解过程

- 首先在数据集中统计 X_1 各种取值出现的频度，并以之替代概率，得到 $p(X_1)$；
- 依据 X_1 的不同取值，将数据集 D 分为若干组，同组 X_1 的值相同；
- 统计每组内 $X_i, i > 1$ 的各种取值出现的频度，得到条件概率质量函数 $p(X_i|X_1)$

求概率质量函数例子—隐变量



给定的 Bayes 网络结构
给定数据集为
$D = \{(?, 4, 5), (?, 4, 4), (?, 5, 3), (?, 1, 2), (?, 5, 4)\}$

问题描述

- 给定数据集 D 和 Bayes 网络结构如图，求边缘/条件概率质量函数/表。

问题分析

- 数据集中有隐变量存在，即没有任何的观测到的分量数据，图中 $H = \{G\}, E = \{R_1, R_2\}$
- 典型的代表问题：隐马尔可夫模型，聚类问题
- 思想：将条件概率质量函数视为参数/变量，在这些参数/变量的取值中找一组最好的参数，使得出现观测到输出变量（已知数据集）的概率最大。

隐变量学习：EM 算法

算法思想

- 隐变量 H 和条件概率质量函数都是未知变量，在算法依次/轮流更新；随机设置条件概率质量函数的初值，不妨设为 cpt_0；
- E-step：
  - 对所有 $H = h$，计算 $p(E = e|h, cpt_{i-1})$，给定第 $i-1$ 次迭代获得的各个" 小表"cpt_{i-1}，我们可以计算该条件概率值；
  - 比较不同 h 的概率值 $p(E = e|h, cpt_i)$，取最大概率值对应的 $h*$ 来构建随机向量 (H, E) 的取值 $(h*, e)$，这是概率中的 "极大似然估计"，也就是说隐藏的嫌疑犯 H 就是最有可能造成观测到的证据 $E = e$ 出现的人；
- M-step：
  - 获得了样本数据集 (H, E)，用它来计算/更新 Bayes 网络的各种边缘/条件概率质量函数，即 cpt_i
- E-step 和 M-step 迭代循环，i 为循环控制变量；
- 算法停止条件：收敛到极值。

EM 算法

评述

- 理解 EM 算法思想：假设我们想估计知道 A 和 B 两个参数，在开始状态下二者都是未知的，但如果知道了 A 的信息就可以得到 B 的信息，反过来知道了 B 也就得到了 A。可以考虑首先赋予 A 某种初值，以此得到 B 的估计值，然后从 B 的当前值出发，重新估计 A 的取值，这个过程一直持续到收敛为止。
- 聚类算法中 k-means 算法就是 EM 的特例；Bayes 网络及推理可以用来解聚类问题。

编码实现概率推理

问题描述

- 给定完整的联合概率质量函数（表格），用程序代码（C，java，SQL或伪代码）描述某个特定概率值的计算。

典型的两类概率值计算问题

- 边缘化：计算某些随机分量的边缘概率质量函数；
- 条件概率：计算某些分量 A 的取值/取值范围已知时，其他某些分量 B 取某个/某些值的条件概率。
- 请总结方法。

时间复杂度

- 如何降低扫描整个表格的时间代价？表格有指数量级的行，在复杂问题下，扫描一遍都不可行。
- 算法总是和数据结构相关的。

概率推理的例子

$$\mathbb{P}(S,R)=$$

s	r	$\mathbb{P}(S=s,R=r)$
0	0	0.2
0	1	0.08
1	0	0.7
1	1	0.02

$$\mathbb{P}(S)=$$

s	$\mathbb{P}(S=s)$
0	0.28
1	0.72

$$\mathbb{P}(S \mid R=1)=$$

s	$\mathbb{P}(S=s \mid R=1)$
0	0.8
1	0.2

Figure: 两类典型的概率推理题

解释说明

- 左图字母含义如下：S—sunny，R—rain
- 左图中三个表格分别是：联合概率质量函数，sunny 的边缘概率质量函数，rain $(R=1)$ 时 sunny 的条件概率质量函数；

更详细的概率推理题

b	e	a	$\mathbb{P}(B=b, E=e, A=a)$
0	0	0	$(1-\epsilon)^2$
0	0	1	0
0	1	0	0
0	1	1	$(1-\epsilon)\epsilon$
1	0	0	0
1	0	1	$\epsilon(1-\epsilon)$
1	1	0	0
1	1	1	ϵ^2

解释说明

- $p(B) =?$，即发生/不发生入室盗窃的概率各是多少？
- $p(B|A=1) =?$，当警铃响了的时候，发生/不发生入室盗窃的概率各是多少？
- $p(B|A=1, E=1)$，当警铃响了，发生地震了，发生/不发生入室盗窃的概率各是多少？

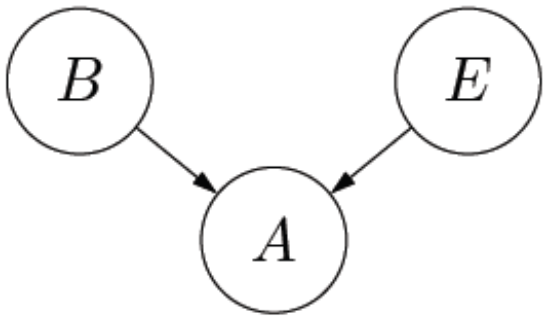
结果

- $p(B=1) = \epsilon,\ p(B=0) = 1-\epsilon$
- $p(B=1|A=1) = \frac{1}{2-\epsilon},\ p(B=0|A=1) = \frac{1-\epsilon}{2-\epsilon}$
- $p(B=1|A=1, E=1) = \epsilon,\ p(B=0|A=1, E=1) = 1-\epsilon$

题外话

- 你能解释 ϵ 很小，如 0.00001 时，上面三个概率吗？

有了 Bayes 网络，再次考虑概率推理



b	$p(b)$
1	ϵ
0	$1-\epsilon$

e	$p(e)$
1	ϵ
0	$1-\epsilon$

b	e	a	$p(a \mid b,e)$
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

$p(B) =?$，即发生/不发生入室盗窃的概率各是多少？

$p(B|A=1) =?$，当警铃响了的时候，发生/不发生入室盗窃的概率各是多少？

$p(B|A=1,E=1)$，当警铃响了，发生地震了，发生/不发生入室盗窃的概率各是多少？

计算/推理过程

- $p(B)$，直接查询边缘概率质量函数/表格
- $p(B|A=1)$，连接第一、二、三张表中 $A=1$ 的那些行（连接条件：$1.b=3.b, 2.e=3.e$），添加新列 $p(b)p(e)p(a|b,e)$，并归一化得到表的新概率列，依据新表 A,B 列取值不同，进行分组，组内新概率求和。
- $p(B|A=1,E=1)$，类似上一个例子处理。
- 上述过程可以优化：合理安排 SQL 中“投影、“连接”和“分组”等操作的次序，可以减少操作的数据量。

形式化：概率推理问题

形式化描述

- 输入：
  - Bayes 网络（包括网络结构和相关的条件/边缘概率质量函数），也就是完整联合概率质量函数的约简描述形式 $p(X)$；
  - 观察到的证据 $E = e, E \subset X$，即 E 是随机向量 X 的一个分量子集；
  - 查询 $Q \subset X$ 是指定想要了解的对象（原因/现象）；
- 输出：
  - 对所有的 q，求条件概率 $p(Q = q|E = e)$

一般性方法

- 总结前面的例子所用的方法，得到一般性的推理计算过程。

推理算法描述

符号说明

- 随机向量 $X = E \cup Q \cup M$
- E 观察到的现象，Q 查询，M 不关心的因素

主要思想: 变量消元法

- 首先在各“小表”中选择 $E = e$ 的行留下（消去 E)；所有的“小表”通过“连接”操作，组合成一张“大表”，制作一个新的概率值列，即联合概率质量函数；
- 再将不关心的因素 M 去掉（消去 M)：有多行满足 $Q = q$，这些行的 M 值不一样，若这些行的概率值之和为 p ，则产生新的行 (q, p)；实际上我们并没有组合出完整的联合概率质量函数，因此，依次消去 M 中的某个分量时，通常只需要访问一部分“小表”即可，减少了时间代价。
- 最终得到新表 (Q, P) 即是所求。

Bayes 网络及其上推理算法的复杂度

空间复杂度

- 当网络中某个节点的（入）度为 k 时，会产生一个 $k+1$ 列的条件概率“小表”，表的行数是 k 的指数函数，因此 Bayes 网络描述联合概率质量函数并非完美解决了存储代价问题。

时间复杂度

- 遍历所有的“小表”是推理的基本操作之一，因此推理的时间复杂度和各个表的大小相关；
- 从网络结构的角度看，单连通网络（任何两个节点之间的路径只有一条）通常能线性时间解决；非单连通网络需要指数时间复杂度。

蒙特卡罗逼近

精确计算一个概率值常常不可能，能不能计算近似值？

- 已知条件：Bayes 网络。
- 目标：计算某个概率值。
- 应用范围：当计算概率值的时间代价是指数的复杂度时，采用计算概率近似值的方法。

蒙特卡罗逼近的基本概念

- 粒子：随机向量的一个值向量。
- 采样：利用 Bayes 网络，产生各个随机分量，最终获得一个完整的“粒子”。
- 重复采样过程，获得大量粒子构成的集合；所谓“采样方法”就是指产生一个粒子及多个粒子的过程。
- 统计/计数粒子集合，用各种比例代表各种对应的概率值。

如何写一个抽样程序？

假设我们有一个产生 (0,1) 之间均匀分布实数的函数，接下来.....

- 正态分布？box-muller 变换。$Y_1 = \sqrt{-2\ln X_1}cos(2\pi X_2)$, $Y_2 = \sqrt{-2\ln X_2}cos(2\pi X_1)$, X_1, X_2 是均匀 (0,1) 间分布的样本，Y_1, Y_2 服从标准正态分布。
- MCMC/马尔可夫蒙特卡罗方法，通用的抽样方法，从任意概率质量函数 $p(x)$ 中抽样。
  - 利用马尔可夫链，随机过程的概念，每个样本是随机过程的一个状态，大量的样本/状态的分布服从随机过程的稳态分布 $p(x)$，即我们要抽样的分布。
  - 从当前状态（粒子），以一定的概率转移到下一个状态（粒子），并不一定每次都转移成功，有一个转移接受率。
  - 具体实现 MCMC 的一个算法是 Metropolis-Hastings 算法。一维随机变量。
- 随机向量采样：Gibbs 抽样。当前粒子为 $(v_1, v_2, \ldots, v_n)$，下一个粒子只改变当前粒子的一个分量的值，改变的方法为对应的条件概率用一次随机抽样。
- 粒子滤波，通用抽样方法。