第4-1节 贝叶斯网络理论及方法

—— 参数学习(1)

4.1 贝叶斯网络定义

- 一个贝叶斯网络定义包括：

  ✓一个有向无环图 (DAG)，其网络结构为S。其中每一个节点表示一个随机变量x_i，$i = 1, 2, \ldots, n$

  ✓一个条件概率表集合，与每一个变量相联系

$$p(x_1, x_2, \ldots\ldots, x_n \mid S) = \prod_{i=1}^{n} p(x_i \mid Parents(x_i), S)$$

4.1.1贝叶斯网络学习类型：

- **Bayesian** 网络学习指的是通过分析数据儿获得**Bayesian** 网的过程，它包括以下两种情况
  - 参数学习：已知网络结构，确定网络参数
  - 结构学习：既要确定网络结构，又要确定网络参数

$P(x_i \mid Parents(x_i))$

4.1.2贝叶斯网络参数估计的两种方法：

1. 贝叶斯估计：

利用先验概率 $p(\theta|S)$，寻求给定拓扑结构 S 和训练样本集 D 时具有最大后验概率的θ参数取值。

2. 极大似然估计：

根据概率最大的事件最可能发生的原理，寻求使样本集 D 的似然函数取最大值的θ参数取值。

4.2 极大似然参数估计方法

设样本 $x_1, x_2, \ldots, x_n$ 组成的贝叶斯网络 S，其中的节点 x_i 共有 r_i 个取值1, 2, …, r_i，概率 $p(x_i \mid \text{Parents}(x_i), S)$ 中父节点取值共有 q_i 个组合，则独立参数的个数是：

$$\sum_{i=1}^{n} q_i (r_i - 1)$$

在此基础上，构造极大似然函数，求得极大似然估计即可。

4.2.1 极大似然法的提出

有一个箱子,装有形状相同的黑色球和白色球100个，其中一种颜色90个，另一种颜色球10个.现从箱中任取一球，结果所取得的球是黑色球。问：箱中黑球和白球的个数？

答：极有可能黑色球90个

极大似然估计法
(Method of Maximum Likelihood Estimation -- MLE)

极大似然估计法的依据就是——

概率最大的事件最可能发生，而一次试验就出现的事件(应该)有较大的概率。

4.2.2 极大似然原理及数学表述

若一个试验有 n 个可能的结果 $\theta_1, \theta_2, \ldots, \theta_m$。现做一次试验，若事件$\theta_i$ 发生了，则认为事件 θ_i 在这 m 个可能结果中出现的概率最大。

数学表达为：设一次采样的观测值为 $x_1, x_2, \ldots, x_n$，则利用这组数据推测参数θ的取值。在所有的 $\hat{\theta}(x_1, x_2, \ldots, x_n)$ 估计值中，选取使得样本$x_1, x_2, \ldots, x_n$出现的概率最大的θ估计值。

4.2.3 极大似然估计步骤

1. 若样本 x 为离散型

假设 x 分布为：$p\{x_i=\theta_j\}=p(x_i\mid\theta)\quad i=1,\ldots,n;\ j=1,\ldots,m$

其似然函数表示为：

$$L(\theta\mid x)=L(x_1,x_2,\quad,x_n;\ \theta)=\prod_{i=1}^{n}p(x_i\mid\theta)$$

若

$$L(x_1,x_2,\ldots,x_n;\hat{\theta})=\max_{\theta\in\Omega}L(x_1,x_2,\ldots,x_n;\theta)$$

则称 $\hat{\theta}(x_1,x_2,\ldots,x_n)$ 为参数 θ 的极大似然估计值。

1) 单参数的最大似然估计

设数据 D 由样本 $D = \{ x_1, x_2, \ldots, x_n \}$ 组成。在给定单参数 θ 下，数据 D 的条件概率 $p(D \mid \theta)$称为似然度，记为:

$$L(\theta \mid D)$$

上式称为单参数 θ 的似然函数，使其达到最大值的 $\hat{\theta}$ ，则称为参数 θ 的最大似然估计。

例4.1

如图4.1 所示为一个单随机变量X 构成的简单贝叶斯网。X 代表投掷图钉的结果：h 代表头朝上，t 代表尾朝上。网络只有一个待估计参数 θ，表示图钉头朝上的概率：$p(\theta) = p(X = h)$ 。为估计 θ，投掷图钉6 次，结果依次是：t 、h、t、t、h、t。根据上述数据对θ 进行极大似然估计。

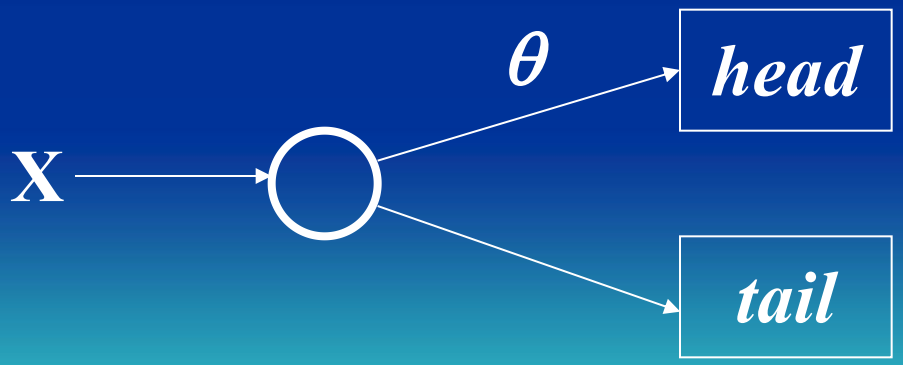


图4.1 单变量构成的简单贝叶斯网络

解： 图钉投掷的试验服从二项分布，根据头朝上的概率 $p(\theta)$ 得到 X 的似然函数表达为：

$$L(\theta \mid D) = p(D \mid \theta) = \prod_{i=1}^{n_h} p(D_i \mid \theta) = \theta^{n_h} \cdot (1-\theta)^{n_t}$$

其中：$n_h = 2$ 和 $n_t = 4$ 分别是 6 次采样中图钉头朝上和尾朝上的次数。θ 的估计值 $\hat{\theta}$ 应使 $L(\theta \mid D)$ 的值最大，二项分布的 θ 估计值如下：

$$\hat{\theta} = \frac{n_h}{n_h + n_t} = \frac{n_h}{n}$$

其中：$n = n_h + n_t$ 是总样本量，即 $\hat{\theta}$ 为头朝上出现的频率。

2) 多参数的最大似然估计

设样本 $x_1, x_2, \ldots, x_n$ 组成的贝叶斯网络 S，节点 x_i 共有 r_i 个取值1, 2, …, r_i，其父节点 $\pi(x_i)$ 的取值共有 q_i 个组合。若 x_i 无父节点，则 q_i =1。该网络的参数为：

$$\theta_{ijk} = p(x_i = k \mid \pi(x_i) = j)$$

其中：$i = 1, 2, \ldots, n$ 表示 n 个节点；$j = 1, 2, \ldots, q_i$ 表示父节点的 q_i 个组合；$k = 1, 2, \ldots, r_i$表示节点 x_i 的r_i 个取值。

对数似然函数为：

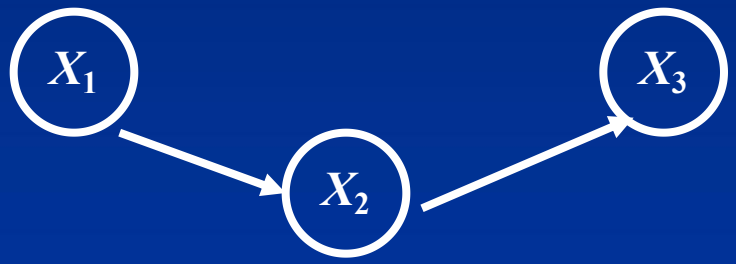
$$l(\theta)=\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}m_{ijk}\log\theta_{ijk}$$

其中，m_{ijk} 是数据中满足 $x_i=k$ 和 $\pi(x_i)=j$ 的样本数量。多项分布的 θ_{ijk} 估计值如下：

$$\hat{\theta}_{ijk}=\frac{\text{数据}D\text{中满足}x_i=k\text{和}\pi(x_i)=j\text{的样本数量}}{\text{数据}D\text{中满足}\pi(x_i)=j\text{的样本数量}}$$

例4.2

如图4.2 所示的多变量贝叶斯网络，其中所有变量均取 2 值，1或2。图 (b) 是一组共4个训练数据，对多参数 θ 进行极大似然估计。



(*a*) 网络结构

	X_1	X_2	X_3
D_1	1	1	1
D_2	2	2	2
D_3	1	1	2
D_4	2	2	2

(*b*) 样本数据

图4.2 多变量贝叶斯网络及训练样本集合

根据多参数最大似然估计公式，可得：

$p(X_1)$

X_1	1	2
	2/4	2/4

$p(X_2|X_1)$

X_1 \ X_2	1	2
1	2/2	0
2	0	2/2

$p(X_3|X_2)$

X_2 \ X_3	1	2
1	1/2	1/2
2	0	2/2

3) 不完整数据(有缺损)的最大似然估计

当样本数据存在随即缺失下，采用期望最大化(Expectation Maximization, EM)方法进行参数估计。

EM 算法先对数据进行修补，然后再进行极大似然估计。算法分为 E 步和 M 步：

- E步从随机产生的初始值开始迭代，重复地估计参数。每次重复时，根据给定学习变量的已知值和当前参数的估计，找到未学习变量的分布；
- M 步对极大似然度重新估计参数，不断重复直至达到局部最大值。

EM算法两个步骤：

① 基于θ^{t-1}对数据进行修补，使之完整。

设D_l为某一缺值样本，X_l为D_l中所有缺值变量的集合，对X_l的一个取值x_l，将$X_l = x_l$加入D_l，就得到一个完整样本(D_l，$X_l = x_l$)，这个过程称为数据修补。由于X_l有多个可能取值，D_l有多种修补方式。EM算法考虑所有可能的修补结果，并给每一结果附加一个权重ω_{xl}，得到加权样本(D_l，$X_l = x_l$) [ω_{xl}]，然后通过迭代可确定这个权重。由此得到的加权样本又称为碎权样本。

② 基于修补后的完整数据计算θ的极大似然估计，得θ^t

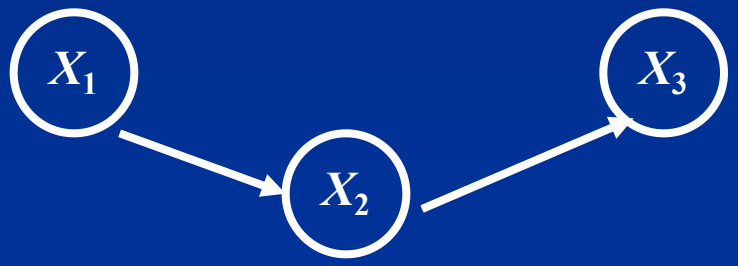
数据经过修补后，每个缺值样本都被一系列完整的碎权样本所代替，所以修补后的样本都是完整的，而且每个样本都带有一个权重。EM 算法用下式来计算基于修补后数据的极大似然估计θ^t：

$$\hat{\theta}_{ijk}^{t}=\begin{cases}\dfrac{m_{ijk}^{t-1}}{\sum_{k=1}^{r_i} m_{ijk}^{t-1}} & \text{若} \sum_{k=1}^{r_i} m_{ijk}^{t-1}>0 \\ \dfrac{1}{r_i} & \text{若否}\end{cases}$$

其中：$m_{ijk}{}^{t\text{-}1}$ 是基于$\theta^{t\text{-}1}$ 修补后数据中所有满足 $X_i=k$ 且 $\pi(X_i)=j$的样本的权重之和。

例4.3

如图4.3 所示的多变量贝叶斯网络，其中所有变量均取 2 值，1或2。图 (b) 是由两个完整样本D_1、D_2和两个缺损样本D_3、D_4组成的训练数据。用EM方法近似对参数 θ 进行极大似然估计。



(*a*) 网络结构

	X_1	X_2	X_3
D_1	1	1	1
D_2	2	2	2
D_3	1	–	1
D_4	2	–	2

(*b*) 样本数据

图4.3 数据缺损的贝叶斯网络

对于D_3样本，假设取其中一组初始参数值θ^0，如下：

$p(X_1)$

X_1	1	2
	1/2	1/2

$p(X_2 | X_1)$

X_1 \ X_2	1	2
1	2/3	1/3
2	1/3	2/3

$p(X_3 | X_2)$

X_2 \ X_3	1	2
1	2/3	1/3
2	1/3	2/3

从θ^0出发，EM开始迭代。在第一次迭代中，先用θ^0修补数据。由于：

$p(X_2=1 \mid D_3, \theta^0) = 4/5$，$p(X_2=2 \mid D_3, \theta^0) = 1/5$

所以D_3被如下两个碎权样本所替换：

$$D_{3,1} = (1,\ 1,\ 1)\cdot\begin{bmatrix}\frac{4}{5}\end{bmatrix}$$

$$D_{3,2} = (1,\ 2,\ 1)\cdot\begin{bmatrix}\frac{1}{5}\end{bmatrix}$$

类似地，D_4也被两个碎权样本替换。

全部修补完成后得到的碎全完整数据如下：

	X_1	X_2	X_3	权重
D_1	1	1	1	1
D_2	2	2	2	1
$D_{3,1}$	1	1	1	4/5
$D_{3,2}$	1	2	1	1/5
$D_{4,1}$	2	1	2	1/5
$D_{4,2}$	2	2	2	4/5

第一次迭代完成后的估计θ^1如下：

$p(X_1)$

X_1	1	2
	1/2	1/2

$p(X_2|X_1)$

X_1 \ X_2	1	2
1	9/10	1/10
2	1/10	9/10

$p(X_3|X_2)$

X_2 \ X_3	1	2
1	9/10	1/10
2	1/10	9/10

EM接着进入第二次迭代。先用θ^1对数据进行修补。由于：

$p(X_2=1 \mid D_3, \theta^1) = 81/82$　$p(X_2=2 \mid D_3, \theta^1) = 1/82$

$p(X_2=1 \mid D_4, \theta^1) = 1/82$　$p(X_2=2 \mid D_4, \theta^1) = 81/82$

得到修补后的数据如下：

	X_1	X_2	X_3	权重
D_1	1	1	1	1
D_2	2	2	2	1
$D_{3,1}$	1	1	1	81/82
$D_{3,2}$	1	2	1	1/82
$D_{4,1}$	2	1	2	1/82
$D_{4,2}$	2	2	2	81/82

于是，在第二次迭代完成后的估计θ^2如下：

$p(X_1)$

X_1	1	2
	1/2	1/2

$p(X_2|X_1)$

X_1 \ X_2	1	2
1	163/164	1/164
2	1/164	163/164

$p(X_3|X_2)$

X_2 \ X_3	1	2
1	163/164	1/164
2	1/164	163/164

如此迭代下去，容易看到这个过程快速收敛于：

$p(X_1)$

X_1	1	2
	1/2	1/2

$p(X_2|X_1)$

X_1 \ X_2	1	2
1	1	0
2	0	1

$p(X_3|X_2)$

X_2 \ X_3	1	2
1	1	0
2	0	1

- 在实际中，由于 EM 算法不一定收敛于全局最优，也可能收敛于局部最优或鞍点。因此，它给出的估计可能与极大似然估计相差甚远。为提高估计质量，人们往往从不同的初始点出发多次运行 EM 算法，然后把所得结果进行比较，选择似然度最大的那个估计作为最后的结果。
- EM 算法的收敛速度开始比较快，后来逐渐放慢，而且其收敛速度与数据缺失的多少有关，通常数据缺失越多，收敛速度越慢。

2. 若样本 x 为连续型

假设 x 分布密度函数为：$f(x_i \mid \theta)$　$i = 1, \ldots, n$

其似然函数表示为：

$$L(\theta \mid x) = L(x_1, x_2, \quad , x_n; \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

似然函数取对数得：$\ln L(\theta \mid x)$

令　　$\dfrac{d \ln L(\theta \mid x)}{d\theta} = 0$

解似然方程得到参数 θ 的极大似然估计值 $\hat{\theta}$ 。

例4.4

设总体 X 服从参数为 $\lambda(\lambda>0)$ 的泊松分布。$x_1, x_2, \ldots, x_n$ 是来自于 X 的一组样本值，求 λ 的极大似然估计值。

解： 因为 X 的分布密度函数为：

$$p\{X=x\}=\frac{\lambda^x}{x!}\mathrm{e}^{-\lambda} \quad (x=0,1,2,\ldots,n)$$

所以 λ 的似然函数为：

$$L(\lambda \mid x)=\prod_{i=1}^{n}\left(\frac{\lambda^{x_i}}{x_i!}\mathrm{e}^{-\lambda}\right)=e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}(x_i!)}$$

λ 的似然函数取对数得：

$$\ln L(\lambda \mid x) = -n\lambda + \left(\sum_{i=1}^{n} x_i\right)\ln\lambda - \sum_{i=1}^{n}(x_i!)$$

令：
$$\frac{d}{d\lambda}\ln L(\lambda \mid x) = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0$$

解得 λ 的极大似然估计值为：

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

这一估计值与矩估计值是相同的。

例4.5

设总体 $X \sim N(\mu, \sigma^2)$ 的正态分布，μ, σ 为未知参数。$x_1, x_2, \ldots, x_n$ 是来自于 X 的一组样本值，求 μ, σ 的极大似然估计值。

解： X 的概率密度为：

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

似然函数为：

$$L(\mu, \sigma^2 \mid x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

取对数得：

$$\ln L(\mu,\sigma^2 \mid x) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

令：$\begin{cases} \dfrac{\partial}{\partial\mu}\ln L(\mu,\sigma^2 \mid x) = 0 \\ \dfrac{\partial}{\partial\sigma}\ln L(\mu,\sigma^2 \mid x) = 0 \end{cases}$ 得到：$\begin{cases} \dfrac{1}{\sigma^2}\left[\sum_{i=1}^{n} x_i - n\mu\right] = 0 \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \end{cases}$

解得 μ 和 σ 的极大似然估计值为：

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

这一估计值与矩估计值是相同的。