第 5 章

贝叶斯网络：比赛结果预测

5.1 教 学 目 标

（1）掌握概率论在不确定性推理中的应用。

（2）能够建立贝叶斯网络模型，能够进行贝叶斯网络的精确求解。

（3）能够应用蒙特卡洛采样方法计算概率，包括拒绝采样方法、似然加权采样方法、Gibbs 采样方法等。

（4）能够分析研究不同的计算方案的特点。

5.2 实验内容与任务

三支足球队 A、B、C 两两之间各赛一场，总共需要赛三场，分别是 A 对 B、A 对 C、B 对 C。对一支球队来说，一场比赛的结果可能是胜、平、负之一。假设每场比赛的结果以某种概率取决于两队的实力，而球队实力为一个 0~3 的整数。现已知前两场比赛结果是 A 战胜了 B，A 和 C 战平，请预测最后一场比赛 B 对 C 的结果。

5.3 实验过程及要求

（1）实验环境要求：Windows/Linux 操作系统，Python 编译环境，numpy、random 等程序库。

（2）建立足球比赛的贝叶斯网络，设置贝叶斯网络的条件概率表。

（3）分别实现精确求解方法、拒绝采样方法、似然加权采样方法、Gibbs 采样方法，获得 B 对 C 比赛结果的后验分布。

（4）调整采样次数，观测几个近似方法相对于精确解的差距。

（5）撰写实验报告。

5.4　相关知识及背景

不确定性推理利用概率论知识来处理状态和采取的行为。通过完全联合概率分布可以计算多个变量的任何分布问题，但是当变量过多时，计算量是巨大的，最后可能多到不可操作。实际问题中，如果变量之间存在独立关系或者条件独立关系，则计算概率分布时的计算量要小很多。应用贝叶斯网络模型来表示变量之间的依赖关系，是进行不确定性推理的重要方法。

应用贝叶斯网络模型，进行概率分布的精确计算依然可能有较大的计算量，此时可以用采样的方法完成计算。当然采样计算是一种近似计算，但当采样规模足够大时，计算结果逼近精确结果。

5.5　实验教学与指导

5.5.1　贝叶斯网络

记三队的实力为 XA、XB、XC，其先验分别满足分布 PA(X)、PB(X)、PC(X)，其中 X 取值 0、1、2、3。一场比赛结果与队伍实力的关系的表现为条件分布，如 $P(s\mathrm{AB}|X\mathrm{A}, X\mathrm{B})$，其中 sAB 是 A 队对战 B 队时 A 队的结果，假设胜、平、负分别用 0、1、2 表示。根据实力和比赛结果的关系，构建贝叶斯网络（图 5.1）。实验任务为求 $P(s\mathrm{BC}|s\mathrm{AB}=0, s\mathrm{AC}=1)$。

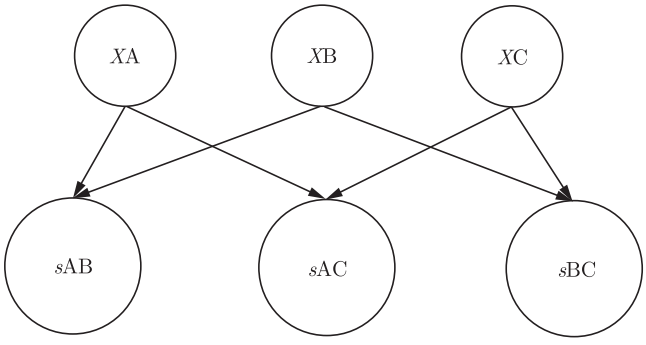

图 5.1　比赛问题的贝叶斯网络

假设在足球比赛问题中，XA、XB、XC 结点的条件概率表 PA、PB、PC 分别是：

```
PA=[0.3,0.3,0.2,0.2]
PB=[0.4,0.4,0.1,0.1]
PC=[0.2,0.2,0.3,0.3]
```

因为实验中比赛结果取决于实力，因此 $s\mathrm{AB}, s\mathrm{AC}, s\mathrm{BC}$ 共享一个条件概率表 PS，即

```
PS=\
[[[0.2,0.6,0.2],[0.1,0.3,0.6],[0.05,0.2,0.75],[0.01,0.1,0.89]],
[[0.6,0.3,0.1],[0.2,0.6,0.2],[0.1,0.3,0.6],[0.05,0.2,0.75]],
[[0.75,0.2,0.05],[0.6,0.3,0.1],[0.2,0.6,0.2],[0.1,0.3,0.6]],
[[0.89,0.1,0.01],[0.75,0.2,0.05],[0.6,0.3,0.1],[0.2,0.6,0.2]]]
```

PS 是一个 $4\times4\times3$ 的表，$PS[i][j]$ 是一个比赛结果的分布，表示参赛两队的实力为 i、j 时比赛结果的分布。

5.5.2 精确算法

贝叶斯网络的联合分布概率计算公式为

$$P(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}P(x_i|\mathrm{Parent}(x_i)) \tag{5.1}$$

应用条件概率、边缘概率及联合分布率计算公式为

$$\begin{aligned}
&P(s\mathrm{BC}|s\mathrm{AB}=0,s\mathrm{AC}=1)\\
&=\alpha P(s\mathrm{BC},s\mathrm{AB}=0,s\mathrm{AC}=1) &(5.2)\\
&=\alpha\sum_{X\mathrm{A}=0}^{3}\sum_{X\mathrm{B}=0}^{3}\sum_{X\mathrm{C}=0}^{3}P(s\mathrm{BC},s\mathrm{AB}=0,s\mathrm{AC}=1,X\mathrm{A},X\mathrm{B},X\mathrm{C}) &(5.3)\\
&=\alpha\sum_{X\mathrm{A}=0}^{3}\sum_{X\mathrm{B}=0}^{3}\sum_{X\mathrm{C}=0}^{3}P(s\mathrm{BC}|X\mathrm{B},X\mathrm{C})P(s\mathrm{AB}=0|X\mathrm{A},X\mathrm{B})\\
&\quad P(s\mathrm{AC}=1|X\mathrm{A},X\mathrm{C})P\mathrm{A}(X\mathrm{A})P\mathrm{B}(X\mathrm{B})P\mathrm{C}(X\mathrm{C}) &(5.4)
\end{aligned}$$

利用条件概率表，则精确计算方法为：

```
def direct_cal():
    res=[0,0,0]
    for XA in range (4):
        for XB in range (4):
            for XC in range (4):
                for sBC in range (3):
                    res[sBC] += PA[XA]*PB[XB]*PC[XC]\
                                *PS[XA][XB][0]\
                                *PS[XA][XC][1]\
```

```
                                    *PS[XB][XC][sBC]
    return normal(res)    #normal(X)=X/sum(X) 将计数变成概率
```

5.5.3　拒绝采样方法

5.5.2节给出的精确算法的计算式包括多层累加，当变量较多时，计算复杂性是指数级的。蒙特卡洛算法通过采样的方式，给出近似解，能降低算法的复杂性。拒绝采样方法按贝叶斯网络结点的顺序对所有变量进行采样，获得一个事件。经过 N 次采样后，对所有采样事件进行统计，获得查询结果。

```
def reject_sampling():
    n=5000
    res=[0,0,0]
    for i in range(n):
        XA=np.random.choice(4,p=PA)
        XB=np.random.choice(4,p=PB)
        XC=np.random.choice(4,p=PC)
        sAB=np.random.choice(3,p=PS[XA][XB])
        sAC=np.random.choice(3,p=PS[XA][XC])
        sBC=np.random.choice(3,p=PS[XB][XC])
        if sAB==0 and sAC==1:
            res[sBC]+=1
    return normal(res)
```

5.5.4　似然加权采样方法

拒绝采样方法最后统计的是出现证据 $s\mathrm{AB}=0$ 且 $s\mathrm{AC}=1$ 的样本点，其他的被拒绝，因此造成计算浪费。似然加权方法固定证据变量，只对非证据变量进行采样。然而每个事件与证据有不同的吻合程度，在计数时须被考虑，因此对证据变量计算权值，最后算到结果中。

```
def likehood_weighting():
    n=5000
    res=[0,10,0]
    for i in range(n):
        w=1
        XA=np.random.choice(4,p=PA)
        XB=np.random.choice(4,p=PB)
        XC=np.random.choice(4,p=PC)

        w=w*PS[XA][XB][0]  #sAB 加权
```

```
        w=w*PS[XA][XC][1]   #sAC 加权

        sBC=np.random.choice(3,p=PS[XB][XC])
        res[sBC]+=w
    return normal(res)
```

5.5.5 Gibbs 采样方法

Gibbs 采样从一个初始样本出发，每次更改一个非证据变量形成一系列的采样点，然后对查询变量进行统计。采样一个非证据变量时，以其马尔可夫覆盖为条件进行采样。

```
def Gibs():
    n=4999
    res=[0,0,0]
    XA,XB,XC,sAB,sAC,sBC=0,0,1,0,1,1
    for k in range(n):
        _PA=normal([PA[i]*PS[i][XB][sAB]*PS[i][XC][sAC] \
                    for i in range(4)])
        XA=np.random.choice(4,p=_PA)

        _PB=normal([PB[i]*PS[XA][i][sAB]*PS[i][XC][sBC] \
                    for i in range(4)])
        XB=np.random.choice(4,p=_PB)

        _PC=normal([PC[i]*PS[XA][i][sAC]*PS[XB][i][sBC] \
                    for i in range(4)])
        XC=np.random.choice(4,p=_PC)

        sBC=np.random.choice(3,p=PS[XB][XC])
        res[sBC]+= 1
    return normal(res)
```

5.6 实验报告要求

实验报告须包含实验任务、实验平台、实验原理、实验步骤、实验数据记录、实验结果分析和实验结论等部分，特别是以下重点内容。

（1）正确建立比赛问题的贝叶斯网络模型。

（2）实现精确求解方法、拒绝采样方法、似然加权采样方法、Gibbs 采样方法。

（3）分析各种算法的时间复杂性，分析近似方法相对于精确解的差距。

（4）对各种算法的优缺点进行分析。

5.7　考核要求与方法

实验总分 100 分，通过实验报告进行考核，标准有如下 3 点。

（1）报告的规范性 10 分。报告中的术语、格式、图表、数据、公式、标注及参考文献是否符合规范要求。

（2）报告的严谨性 40 分。结构是否严谨，论述的层次是否清晰，逻辑是否合理，语言是否准确。

（3）实验的充分性 50 分。实验是否包含“实验报告要求”部分的 4 个重点内容，数据是否合理，是否有创新性成果或独立见解。

5.8　案例特色或创新

本实验的特色在于：培养学生应用概率论的知识进行推理，建立了足球比赛问题的贝叶斯网络模型，要求学生使用精确求解方法、拒绝采样方法、似然加权采样方法、Gibbs 采样方法进行贝叶斯网络的近似计算，培养学生应用贝叶斯网络对复杂问题进行建模和分析计算的能力。