山东大学

信息科学与工程学院

2022－2023 学年第一学期

实 验 报 告

课程名称： 信息基础II

专 业 班 级 崇新学堂

学 生 学 号

学 生 姓 名

课 程 报 告 homework6

1. Homework6.1

1.1 Gibbs 采样

确定初始值 $X^{(1)}$；假设已经得到样本 $X^{(i)}$，记下一个样本为 $X^{(i+1)}=(x_1^{(i+1)},x_2^{(i+1)}\cdots x_n^{(i+1)})$，于是可将其看作一个向量，对其中某一分量 $x_j^{(i+1)}$,可通过在其他分量已知的条件下该分量的概率分布来抽取该分量，对于此条件概率，我们使用样本 $X^{(i+1)}$ 中已经得到的分量 $x_1^{(i+1)}$ 到 $x_{j-1}^{(i+1)}$ 以及上一样本 $X^{(i)}$ 总的分量 $x_{j+1}^{(i)}$ 到 $x_n^{(i)}$，即条件概率 $P(x_j^{(i+1)}\mid x_1^{(i+1)},\cdots x_{j-1}^{(i+1)},x_{j+1}^{(i)},\cdots x_n^{(i)})$，重复上述过程 k 次，如果仅仅考虑其中部分变量，则可以得到这些变量的边缘分布。此外，我们还可以对所有样本求某一变量的平均值来估计该变量的期望值。

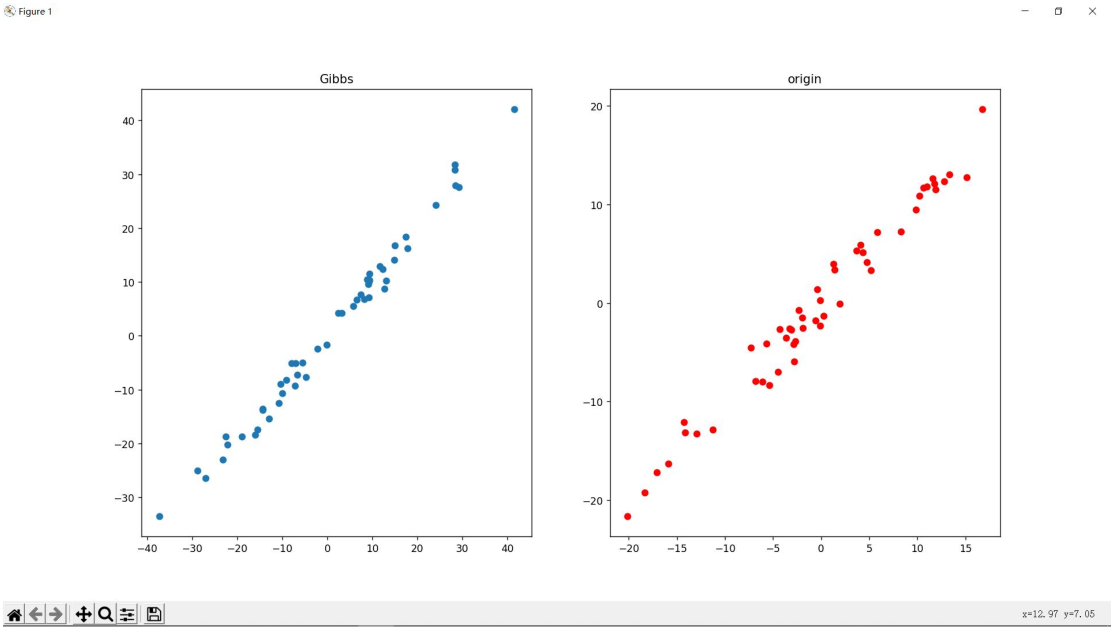
1.2 代码实现

由于待采样的数据为二维，且确定某一特征之后另一特征的条件分布已知，因此只需要利用 np.random.normal（）函数重复迭代即可，确定 X1，之后保持 X1 不变，利用条件分布求取 X2，再确定 X2 不变，对 X1 进行更新，每个点迭代更新 100 次，总共产生 50 个点，最终进行绘图和展示。值得一提的是，使用 Gibbs 采样的同时，也使用 np.random.multivariate_noramal()函数，生成了标准的二维高斯分布点，同时绘图，用来检验 Gibbs 采样的可行性。

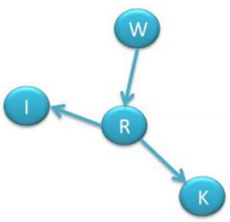
```
# 利用np中的函数生成50个高斯随机点，用于后续验证结果
x = np.random.multivariate_normal(np.array([0, 0]), np.array([[100, 99], [99, 100]]), 50)
# 利用Gibbs Sampling产生50个高斯随机点
for j in range(50):
    # 由于迭代过程只需要保持一个特征不变，来确定另一个参数，也就是说初始点只用随机选取一个特征即可
    x1 = 0
    # x2为初始点随机选取的值
    x2 = np.random.randint(-20, 20)
    # 每个点迭代100次
    for i in range(100):
        # 保证x2不变，确定x1
        x1 = np.random.normal(0.99*x2, 100-99**2/100)
        # 保证x1不变，确定x2
        x2 = np.random.normal(0.99*x1, 100-99**2/100)
    #将迭代完成的点添加到绘图列表中
    x01.append(x1)
    x02.append(x2)
# 将利用高斯函数产生的点添加到另一个绘图列表中
for i in range(len(x)):
    x11.append(x[i][0])
    x12.append(x[i][1])
# 绘图，进行对比。
plt.subplot(121)
plt.scatter(x01, x02)
plt.title("Gibbs")
plt.subplot(122)
plt.scatter(x11, x12, color="red")
plt.title("origin")
plt.show()
```

1.3 结果分析

实验结果如下图，左侧为 Gibbs 采样的结果，右侧为标准的二维高斯分布，对比两者的结果，可以发现两者的分布几乎相同，因此可以证明 Gibbs 采用在实际使用中是完全可行的。



2. Homework6.2



W=w	W=s
0.9	0.1

	R=w	R=s
W=w	0.8	0.2
W=s	0.3	0.7

	I=w	I=s
R=w	0.8	0.2
R=s	0.3	0.7

	K=w	K=s
R=w	0.8	0.2
R=s	0.3	0.7

给定一个简单的BayesNet，条件概率入上图所示

给定K=w, 计算W=w的条件概率 P(W=w|K=w)

***Answer*：**

贝叶斯如上所示，因此可以获得该贝叶斯网络的联合概率分布为：

$$P(W,R,I,K)=P(W)P(R\,|\,W)P(I\,|\,R)P(K\,|\,R) \tag{1}$$

条件概率可以表示为：

$$P(W=w\,|\,K=w)=\frac{P(W=w,K=w)}{P(K=w)} \tag{2}$$

之后我们求取分母表达式：

$$P(K)=\sum_{W}\sum_{R}\sum_{I}P(W,R,K,I) \tag{3}$$

同时也可以获得分子表达式：

$$P(K,W)=\sum_I\sum_R P(W,R,K,I) \tag{4}$$

代入以上公式可以获得如下结果

$$\begin{aligned}P(K)&=\sum_W\sum_R\sum_I P(W,R,K,I)\\&=0.9\times0.8\times0.8+0.9\times0.2\times0.3\\&\quad+0.1\times0.3\times0.8+0.1\times0.7\times0.3\\&=0.675\end{aligned}$$

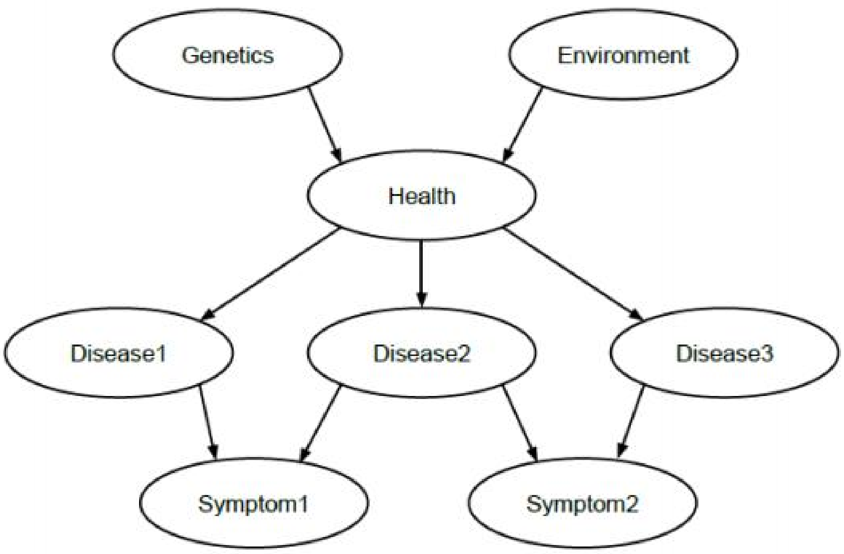
$$\begin{aligned}P(K,W)&=\sum_I\sum_R P(W,R,K,I)\\&=0.9\times0.8\times0.8+0.9\times0.2\times0.3\\&=0.63\end{aligned}$$

由此可以获得:

$$P(W=w\mid K=w)=\frac{P(W=w,K=w)}{P(K=w)}=\frac{14}{15}$$

3. Homework6.3

Consider the Bayesian network shown in Figure 1, which illustrates the incense of genetics and environment on an individual's overall health. Specially, this model encodes a set of independence relationships among the following variables: Genetics (G), Environment (E), Health (H), Disease1 (D_1), Disease2 (D_2), Disease3 (D_3), Symptom1 (S_1), Symptom2 (S_2).



1. **Write down an expression for P(S_1) in terms of the un-factorized joint distribution.**

   对于未分解的联合概率分布，P(S_1)的表达式如下:

$$P(S_1)=\sum_G\sum_E\sum_H\sum_{D_1}\sum_{D_2}\sum_{D_3}\sum_{S_2}P(E)P(G)P(H\mid E,G)P(D_1\mid H)P(D_2\mid H)P(D_3\mid H)P(S_1\mid D_1,D_2)P(S_2\mid D_2,D_3)$$

2. **Starting with the expression you gave in the last question, write down an expression for P(S_1) in terms of the factorized joint distribution and simplify it by pushing summations into the product of terms whenever possible.**

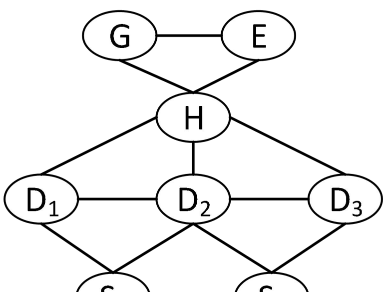
$$
\begin{aligned}
P(S_1) &= \sum_G \sum_E \sum_H \sum_{D_1} \sum_{D_2} \sum_{D_3} \sum_{S_2} P(E)P(G)P(H \mid E,G)P(D_1 \mid H)P(D_2 \mid H)P(D_3 \mid H)P(S_1 \mid D_1,D_2)P(S_2 \mid D_2,D_3) \\
&= \sum_G \sum_E \sum_H \sum_{D_1} \sum_{D_2} \sum_{D_3} P(E)P(G)P(H \mid E,G)P(D_1 \mid H)P(D_2 \mid H)P(D_3 \mid H)P(S_1 \mid D_1,D_2)\sum_{S_2} P(S_2 \mid D_2,D_3) \\
&= \sum_G \sum_E \sum_H \sum_{D_1} \sum_{D_2} \sum_{D_3} P(E)P(G)P(H \mid E,G)P(D_1 \mid H)P(D_2 \mid H)P(D_3 \mid H)P(S_1 \mid D_1,D_2)m_{S_2}(D_2,D_3) \\
&= \sum_G \sum_E \sum_H \sum_{D_1} \sum_{D_2} P(E)P(G)P(H \mid E,G)P(D_1 \mid H)P(D_2 \mid H)P(S_1 \mid D_1,D_2)\sum_{D_3} P(D_3 \mid H)m_{S_2}(D_2,D_3) \\
&= \sum_G \sum_E \sum_H \sum_{D_1} \sum_{D_2} P(E)\ P(G)P(H \mid E,G)\ P(D_1 \mid H)P(D_2 \mid H)P(S_1 \mid D_1,D_2)m_{D_3}(H,D_2) \\
&= \sum_E \sum_H \sum_{D_1} \sum_{D_2}\ P(E)m_G(H,E)\ P(D_1 \mid H)P(D_2 \mid H)P(S_1 \mid D_1,D_2)m_{D_3}(H,D_2) \\
&= \sum_H \sum_{D_1} \sum_{D_2}\ m_E(H)P(D_1 \mid H)P(D_2 \mid H)m_{D_3}(H,D_2)\ P(S_1 \mid D_1,D_2) \\
&= \sum_{D_1} \sum_{D_2} m_H(D1,D2)P(S_1 \mid D_1,D_2) \\
&= \sum_{D_2} m_{D_1}(S_1,D_2)
\end{aligned}
$$

3. **What variable elimination ordering does the resulting expression correspond to?**
Ordering:

$$S_2, D_3, G, E, H, D_1, D_2$$

4. **Walk through an execution of the variable elimination algorithm for evaluating P(S_1) using the ordering you specified in the last question. Start by writing out the initial set of factors and drawing the moralized graph. Then go through each step of the elimination algorithm and (1) specify the intermediate factors that are generated by the product and sum operations, in that order, (2) write down the new set of factors that remain after the entire step has been executed, and (3) draw the new graph induced by eliminating the variable. For clarity, please use φ (·) to denote initial factors, ψ (·) to denote intermediate factors generated by a product operation, and τ (·) to denote intermediate factors generated by a sum operation**

Step1: moralized



initial set of factors:

$$P(S_1) = \sum_G \sum_E \sum_H \sum_{D_1} \sum_{D_2} \sum_{D_3} \sum_{S_2} P(E)P(G)P(H \mid E,G)P(D_1 \mid H)P(D_2 \mid H)P(D_3 \mid H)P(S_1 \mid D_1,D_2)P(S_2 \mid D_2,D_3)$$

Step2: eliminate S_2

(1) the intermediate factors:

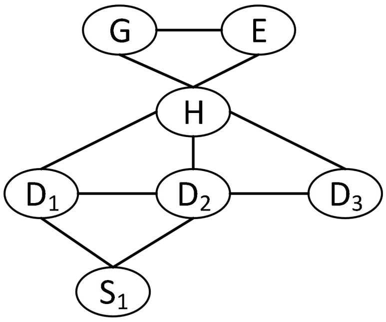
$$\psi_{S_2}(D_2, D_3) = P(S_2 \mid D_2, D_3)$$

$$\tau_{S_2}(D_2, D_3) = \sum_{S_2} \psi_{S_2}(D_2, D_3) = \sum_{S_2} P(S_2 \mid D_2, D_3)$$

(2) the new set of factors:

$$P(S_1) = \sum_G \sum_E \sum_H \sum_{D_1} \sum_{D_2} \sum_{D_3} P(E)P(G)P(H \mid E, G)P(D_1 \mid H)P(D_2 \mid H)P(D_3 \mid H)P(S_1 \mid D_1, D_2)\tau_{S_2}(D_2, D_3)$$

(3)



Step3: eliminate D_3

(1) the intermediate factors:

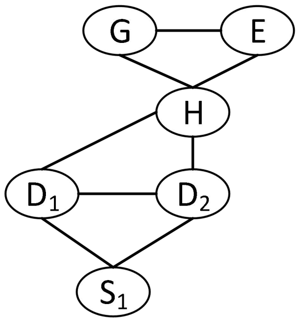
$$\psi_{D_3}(D_2, H) = P(D_3 \mid H)\tau_{S_2}(D_2, D_3)$$

$$\tau_{D_3}(D_2, H) = \sum_{D_3} \psi_{D_3}(D_2, H) = \sum_{D_3} P(D_3 \mid H)\tau_{S_2}(D_2, D_3)$$

(2) the new set of factors:

$$P(S_1) = \sum_G \sum_E \sum_H \sum_{D_1} \sum_{D_2} P(E)P(G)P(H \mid E, G)P(D_1 \mid H)P(D_2 \mid H)P(S_1 \mid D_1, D_2)\tau_{D_3}(D_2, H)$$

(3)



Step4: eliminate G

(1) the intermediate factors:

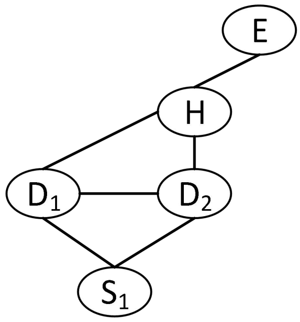
$$\psi_G(H, E) = P(G)P(H \mid E, G)$$

$$\tau_G(H, E) = \sum_G \psi_G(H, E) = \sum_G P(G)P(H \mid E, G)$$

(2) the new set of factors:

$$P(S_1) = \sum_E \sum_H \sum_{D_1} \sum_{D_2} P(E)\tau_G(H, E)P(D_1 \mid H)P(D_2 \mid H)P(S_1 \mid D_1, D_2)\tau_{D_3}(H, D_2)$$

(3)



Step5: eliminate E

(1) the intermediate factors:

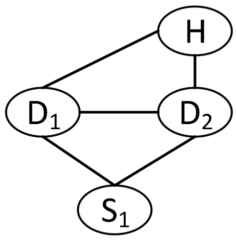
$$\psi_E(H) = P(E)\tau_G(H,E)$$

$$\tau_E(H) = \sum_E \psi_E(H) = \sum_E P(E)\tau_G(H,E)$$

(2) the new set of factors:

$$P(S_1) = \sum_H \sum_{D_1} \sum_{D_2} \tau_E(H)P(D_1 \mid H)P(D_2 \mid H)\tau_{D_3}(H,D_2)P(S_1 \mid D_1,D_2)$$

(3)



Step6: eliminate H

(1) the intermediate factors:

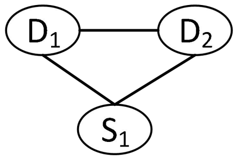
$$\psi_H(D1,D2) = \tau_E(H)P(D_1 \mid H)P(D_2 \mid H)\tau_{D_3}(H,D_2)$$

$$\tau_H(D_1,D_2) = \sum_H \psi_H(D_1.D_2) = \sum_H \tau_E(H)P(D_1 \mid H)P(D_2 \mid H)\tau_{D_3}(H,D_2)$$

(2) the new set of factors:

$$P(S_1) = \sum_{D_1} \sum_{D_2} \tau_H(D1,D2)P(S_1 \mid D_1,D_2)$$

(3)



Step7: eliminate D_1

(1) the intermediate factors:

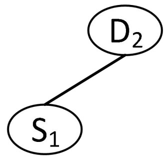
$$\psi_{D_1}(S_1,D2) = \tau_H(D1,D2)P(S_1 \mid D_1,D_2)$$

$$\tau_{D_1}(S_1,D_2) = \sum_{D_1} \psi_{D_1}(S_1.D_2) = \sum_{D_1} \tau_H(D1,D2)P(S_1 \mid D_1,D_2)$$

(2) the new set of factors:

$$P(S_1) = \sum_{D_2} \tau_{D_1}(S_1, D_2)$$

(3)



Step8: eliminate D_2

(1) the intermediate factors:

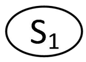
$$\psi_{D_2}(S_1) = \tau_{D_1}(S_1, D_2)$$

$$\tau_{D_2}(S_1) = \sum_{D_2} \psi_{D_2}(S_1) = \sum_{D_2} \tau_{D_1}(S_1, D_2)$$

(2) 已经求得 P（S_1）：

$$P(S_1) = \tau_{D_2}(S_1)$$

(3)



5. Assuming each variable can take m values, what is the size of the largest intermediate factor generated during this execution of the variable elimination algorithm, i.e. how many entries does this factor have?

The size of the largest intermediate factor is: m^3

6. What is the computational complexity (in terms of the number of possible assignments, m, and the number of variables in the network, n) of this execution of variable elimination?

The computational complexity of this execution of variable elimination is:

$$m^n$$

7. Can you do better than this? Specifically, is there another elimination ordering that yields a lower-complexity algorithm? If so, give the best such ordering and evaluate its computational complexity.

利用最大势和最小缺边搜索法，没有找到计算复杂度更低的顺序，最佳的顺序为：

$$S_2, D_3, G, E, H, D_1, D_2$$

计算复杂度为：（假设每个点有 m 种可能）

$$com = m^3 + m^3 + m^3 + m^2 + m^3 + m^3 + m^2$$