HW7

13.15

13.15 在一年一度的体检之后，医生告诉你一些坏消息和一些好消息。坏消息是你在一种严重疾病的测试中结果呈阳性，而这个测试的准确度为99%(即当你确实患这种病时，测试结果为阳性的概率为0.99；而当你未患这种疾病时测试结果为阴性的概率也是0.99)。好消息是，这是一种罕见的病，在你这个年龄段大约10000人中才有1例。为什么“这种病很罕见”对于你而言是一个好消息？你确实患有这种病的概率是多少？

因为病很罕见的情况下，即使测试结果为阳性，由于发病率太低，所以实际患病概率也不大

A表示实际患病，B表示测试结果为阳性

则：在为阳性的情况下，我实际患病的概率为：

$$
\begin{aligned}
P(A|B) &= \frac{P(B|A) \times P(A)}{P(B)} \\
&= \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|\neg A) \times P(\neg A)} \\
&= \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} \\
&= 0.9804\%
\end{aligned}
$$

13.18

13.18 假设给你一只袋子，装有 n 个无偏差的硬币，并且告诉你其中 $n-1$ 个硬币是正常的，一面是正面而另一面是反面。不过剩余1枚硬币是伪造的，它的两面都是正面。

a. 假设你把手伸进口袋均匀随机地取出一枚硬币，把它抛出去，硬币落地后正面朝上。那么你取出伪币的（条件）概率是多少？

b. 假设你不停地抛这枚硬币，一共抛了 k 次，而且看到 k 次正面向上。那么你取出伪币的条件概率是多少？

c. 假设你希望通过把取出的硬币抛 k 次的方法来确定它是不是伪造的。如果抛 k 次后都是正面朝上，那么决策过程返回 $fake$ （伪造），否则返回 $normal$ （正常）。这个过程发生错误的（(无条件）概率是多少？

a

$$
\begin{aligned}
P(是伪币|正面) &= \frac{P(伪币 \wedge 正面)}{P(正面)} \\
&= \frac{\frac{1}{n} \times 1}{\frac{1}{n} \times 1 + \frac{n-1}{n} \times \frac{1}{2}} \\
&= \frac{2}{n+1}
\end{aligned}
$$

b

$$
\begin{aligned}
P(\text{是伪币}|k\text{ 次正面}) &= \frac{P(\text{伪币} \wedge k\text{ 次正面})}{P(k\text{ 次正面})} \\
&= \frac{\frac{1}{n} \times 1^k}{\frac{1}{n} \times 1^k + \frac{n-1}{n} \times (\frac{1}{2})^k} \\
&= \frac{2^k}{n + 2^k - 1}
\end{aligned}
$$

c

$$
\begin{aligned}
P(\text{不是伪币} \wedge k\text{ 次正面}) &= P(k\text{ 次正面}|\text{不是伪币})P(\text{不是伪币}) \\
&= \left(\frac{1}{2}\right)^k \times \frac{n-1}{n} \\
&= \frac{n-1}{2^k n}
\end{aligned}
$$

13.21

13.21 （改编自Pearl (1988)的著述。）假设你是雅典一次夜间出租车肇事逃逸的交通事的目击者。雅典所有的出租车都是蓝色或者绿色的。而你发誓所看见的肇事出租车是蓝色的。大量测试表明，在昏暗的灯光条件下，区分蓝色和绿色的可靠度为75%。

a. 有可能据此计算出肇事出租车最可能是什么颜色吗？（提示:请仔细区分命题“肇事车是蓝色的”和命题“肇事车看起来是蓝色的”。）

b. 如果你知道雅典的出租车10辆中有9辆是绿色的呢？

a

B表示出租车是蓝色的

LB表示出租车看起来是蓝色的

则$P(LB \mid B) = 0.75$，$P(\neg LB \mid \neg B) = 0.75$

则看起来是蓝色时，实际确实为蓝色的概率：（设p为蓝色出租车概率）

$0.75^* p/(0.75^* p + 0.25^*(1-p))$

由于不知道蓝色出租车概率，所以无法计算

b

告知了p=1/10，所以带入上面的式子就可知：

看起来是蓝色，实际确实为蓝色的概率为0.25，实际是绿色的概率为0.75

13.22

13.22 文本分类是基于文本内容将给定的一个文档分类成固定的几个类中的一类。朴素贝叶斯模型经常用于这个问题。在朴素贝叶斯模型中，查询（query）变量是这个文档的类别，而结果（effect）变量是语言中每个单词的存在与否；假设文档中单词的出现是独立的，单词的出现频率由文档类别决定。

a. 给定一组已经被分类的文档，准确解释如何构造这样的模型。

b. 准确解释如何分类一个新文档。

c. 题目中的条件独立性假设合理吗?请讨论。

a

模型由先验概率P(category=A)和条件概率P(word_i|caategory=A)

前者描述所有文档中类别A 的比例

后者表示类别为A的文档中包含word_i的比例

b

通过统计其中各单词出现频率，而由于单词出现的频率是由文档类别决定的，所以可以判别文档的类别

c

不合理，文档中的词呈现一定的相关性，复合词出现的概率并不一定等于各部分出现概率之积

14.12

14.12 两个来自世界上不同地方的宇航员同时用他们自己的望远镜观测了太空中某个小区域内恒星的数目 N 。他们的测量结果分别为 M_1 和 M_2 。通常，测量中会有不超过1颗恒星的误差，发生错误的概率 e 很小。每台望远镜可能出现(出现的概率 f 更小一些)对焦不准确的情况(分别记作 F_1 和 F_2)，在这种情况下科学家会少数三颗甚至更多的恒星(或者说，当 N 小于3时，连一颗恒星都观测不到)。考虑**图14.22**所示的三种贝叶斯网络结构。

a. 这三种网络结构哪些是对上述信息的正确(但不一定高效)表示?

b.哪一种网络结构是最好的？请解释。

c.当 $N \in 1,2,3$ ， $M_1 \in 0,1,2,3,4$ 时，请写出 $\mathbf{P}(M_1|N)$ 的条件概率表。概率分布表里的每个条目都应该表达为参数 e 和或 f 的一个函数。

d.假设 $M_1=1$ ， $M_2=3$ 。如果我们假设 N 取值上没有先验概率约束,可能的恒星数目是多少?

e.在这些观测结果下，最可能的恒星数目是多少？解释如何计算这个数目，或者，如果不可能计算，请解释还需要什么附加信息以及它将如何影响结果。

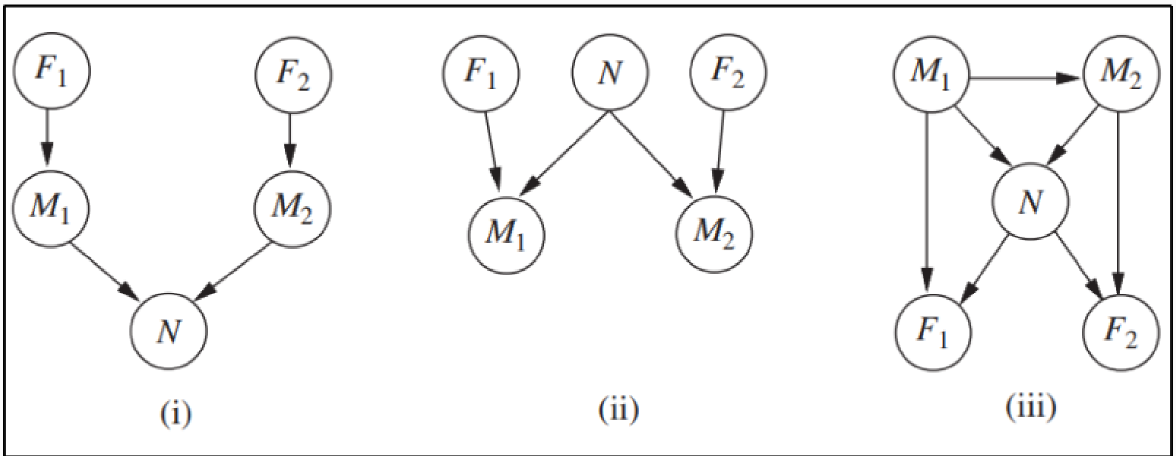

Figure 14.22 Three possible networks for the telescope problem.

a

（ii）和（iii）都是正确的表示

b

（ii）是最好的表示，需要参数更少，也正确

c

	N=1	**N=2**	**N=3**
M1=1	f+(1-f)e	f	f
M1=2	(1-2e)(1-f)	e(1-f)	0
M1=3	e(1-f)	(1-2e)(1-f)	e(1-f)
M1=4	0	e(1-f)	(1-2e)(1-f)
M1=5	0	0	e(1-f)

d

N可能是2或4或者大于等于6

e

$P(N|M_1,M_2)=\frac{P(M_1,M_2|N)\times P(N)}{P(M_1,M_2)}$

不知道P（N）的情况下无法求得，所以需要给定P（N），这样使得上式分子最大的N就是最有可能的

14.13

14.13 考虑 **图14.22(ii)** 的网络,假设两个望远镜完全相同。 $N \in 1,2,3$ ， $M_1, M_2 \in 0,1,2,3,4$ ， CPT表和**习题14.12**所描述的一样。使用枚举算法(**图14.9**)计算概率分布 $\mathbf{P}(N|M_1=2,M_2=2)$ 。

$$
\begin{aligned}
P(N|M1=2,M2=2) &= \frac{1}{P(M1=2,M2=2)}P(N,M1=2,M2=2) \\
&= \frac{1}{\mathrm{P}(M1=2,M2=2)}\sum_{F1,F2}P(N,M1=2,M2=2,F1,F2) \\
&= \frac{1}{\mathrm{P}(M1=2,M2=2)}\sum_{F1,F2}P(N)(1-f)^2P(M1=2|F1=f,N)P(M2=2|F2=f,N)
\end{aligned}
$$

所以：

设P=$\frac{1}{P(M1=2,M2=2)}$

$P(N|M1=2,M2=2)=P^{*}p_1(1-f)^2e^2, N=1, P(N=1)=p_1$
$P(N|M1=2,M2=2)=P^{*}p_2(1-f)^2(1-2e)^2$，$N=2, P(N=2)=p_2$
$P(N|M1=2,M2=2)=P^{*}p_3(1-f)^2e^2$，$N=3, P(N=3)=p_3$