第七次作业(13.15, 13.18, 13.21, 13.22，14.12，14.13)

许悦娇

13.15

在一年一度的体检之后，医生告诉你一个好消息和一个坏消息。坏消息是你在一种严重疾病的测试中结果呈阳性，而这个测试的精度为 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）。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？

已知: $P(test\ |disease)\ = 0.99, P(\neg test|\neg disease) = 0.99, P(disease) = 0.0001.$
求$P(disease|test)$
解答:

$$P(disease|test) = \frac{P(disease, test)}{P(test)} = \frac{P(test|disease)P(disease)}{P(test|disease)P(disease) + P(test|\neg disease)P(\neg disease)}$$

设$P(disease) = p$，则:

$$P(disease|test) = \frac{0.99p}{0.99p + 0.01(1-p)} = \frac{0.99}{0.98} - \frac{0.0099}{0.98(0.98p + 0.01)}$$

$P(disease|test)$与p正相关，所以“这种病很罕见”是一个好消息。

$p = 0.0001$代入得，$P(disease|test) = 0.009804$

13.18

假设给你一只装有 n 个无偏差硬币的袋子，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，并发现硬币落地后正面朝上。那么你拿到伪币的（条件）概率是多少？

b. 假设你不停地抛这枚硬币，拿到它之后一共抛了 k 次而且看到 k 次正面向上。那么现在你拿到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掷 k 次的方法来确定它是不是伪造的。如果 k 次抛掷后都是正面朝上，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. 有n个硬币，每个硬币抛掷有两种结果，总共$2n$种结果，其中$n+1$种是正面，$n-1$种是反面。而在正面朝上得$n+1$种中有两种是伪造的情况，所以拿到伪币得概率是$\frac{2}{n+1}$

b. 每个币掷k次有2^k种结果，有n个币，即有$n*2^k$种结果。其中伪币的2^k种结果都是全正面，$n-1$个真币每个币的2^k种结果中，只有一种是k次正面向上。即$n*2^k$种结果中有2^k

c. +n−1种是k次全正面向上，但是只有2^k种是拿到了伪币。所以$P(fake|k-heads)=\frac{2^k}{2^k+n-1}$

d. “这个过程发生错误”是指拿到了正常的硬币却判定为fake，而拿到伪造的硬币判定为fake并不是错误的。挑选一枚正常硬币的概率是$(n-1)/n$，而一个正常硬币抛掷k次均为正面的概率是$1/2^k$，故最后的结果应该是$\frac{n-1}{n*2^k}$

13.21

假设你在雅典的夜晚目击了一场交通肇事逃逸事故，雅典的出租车是蓝色和绿色的，你对天发誓那辆出租车是蓝色的。延伸实验结果表明，在暗淡的灯光下，搞混蓝绿色的概率为25%

a. 出租车最有可能的颜色？（提示：出租车是蓝色和出租车看起来是蓝色有区别）

b. 假设你知道雅典绿色出租车占总数九成，出租车最有可能的颜色？

解答：已知$P(appear = blue|blue) = P(appear = green|green) = 0.75$

a. $$P(blue|appear = blue) = \frac{P(appear = blue|blue)*P(blue)}{P(appear=blue)}$$

$$= 0.75\ \frac{P(blue)}{P(appear=blue)}$$

$$P(green|appear = blue) = \frac{P(appear = blue|green)*P(green)}{P(appear=blue)}$$

$$= 0.25\ \frac{1-P(blue)}{P(appear=blue)}$$

b. $P(blue) = 0.1,$

$$P(appear = blue) = P(blue, appear = blue) + P(green, appear = blue) = 0.1$$

13.22

文本分类是在文档所包含的文本基础上，把给定的文档分配到固定类别集合中某一个类别的任务。这个任务中常常用到朴素贝叶斯模型。在这些模型中，查询变量是文档类别，“结果”变量则是语言中每个词是否出现。我们假设文档中的词的出现都是独立的，其出现频率由文档类别确定。

a. 准确地解释当给定一组类别已经确定的文档作为“训练数据”时，这样的模型是如何构造的。

b. 准确地解释如何对新文档进行分类。

c. 这里独立性假设合理吗？请讨论。

a. 模型由先验概率$P(category)$和条件概率$P(word_i|category)$组成。对于每一类来说，利用所有文档中的属于类c的那一部分文档，来近似估计$P(category = c)$.类似的，用属于类 c 的那一部分文档中包含单词 i 的文档来近似估计$P(word_i|category = c)$

b. 当得到一个新文档时，判断文档是否包含某个词$word_i$，最后来计算条件概率 $P(category = c|\dots, w_i \dots, w_j, \dots)$

c. 不合理，每个词出现概率不独立，因为实际文档中上下文（context）的单词间存在关联性：显然$P(artificial\ intelligence) \neq P(artificial) * P(intelligence)$

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?

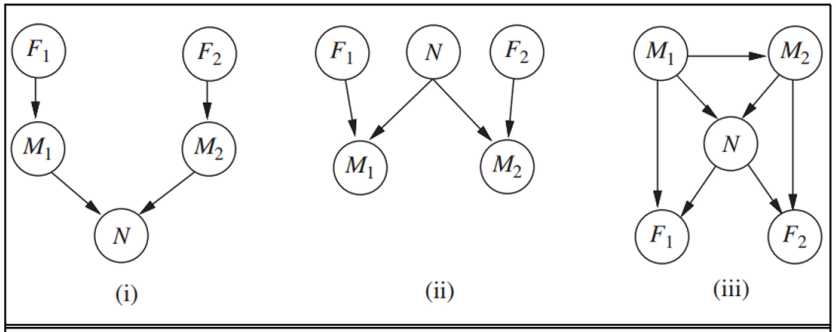

Figure 14.22 Three possible networks for the telescope problem.

- a. （i）不是。因为它表示在给定$M1$和$M2$时，N与$F1$, $F2$ 无关，这是不对的。（ii）是。因为它正确的表达了恒星观测数量$M1$,$M2$受到实际数量N以及望远镜对焦情况$F1$, $F2$的影响。（iii）是。因为它同样正确表达了五者之间的相互影响关系，虽然更为复杂。
- b. （ii）最好，因为它需要的参数最少，网络结构更简单
- c.

当 $N \in \{1,2,3\}$, $M_1 \in \{0,1,2,3,4\}$ 时， 写出 $P(M_1 \mid N)$ 的条件概率表（用参数 e 和 f 表示）

容易得到

$$P(M_1 \mid N) = P(M_1 \mid N, F_1)P(F_1 \mid N) + P(M_1 \mid N, \neg F_1)P(\neg F_1 \mid N)$$
$$= P(M_1 \mid N, F_1)P(F_1) + P(M_1 \mid N, \neg F_1)P(\neg F_1)$$

- f: 望远镜失焦概率（失焦时 $M_1 = 0$）
- e: 计数误差概率（正常时）：
  - e 概率少数1颗
  - e 概率多数1颗
  - $1-2e$ 概率准确计数

$M_1 \setminus N$	$N=1$	$N=2$	$N=3$
0	$f+e(1-f)$	f	f
1	$(1-2e)(1-f)$	$e(1-f)$	0
2	$e(1-f)$	$(1-2e)(1-f)$	$e(1-f)$
3	0	$e(1-f)$	$(1-2e)(1-f)$
4	0	0	$e(1-f)$

- d.对于$M_1 = 1$，考虑到测量误差和失焦误差，可能的N的取值为$\{2\} \cup \{n|n \geq 4, n \in N\}$;
- 对于$M_2 = 3$，考虑到测量误差和失焦误差，可能的N的取值为$\{2, 4\} \cup \{n|n \geq 6, n \in N\}$;
- 则可能的恒星数量N的取值为上述两者的交集，即$\{2, 4\} \cup \{n|n \geq 6, n \in N\}$。

14.12

- 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。

- 由于先验分布P(N)未知，因此无法计算最可能的恒星数目。若假设对于先验分布P(N=2),P(N=4),P(N≥6)三者相差不大，则可估计后验概率如下:

- $P(N=2|M_1=1,M_2=3)=\alpha\cdot e^2(1-f)^2\cdot P(N=2)$

- $P(N=4|M_1=1,M_2=3)=\alpha\cdot ef\cdot P(N=4)$

- $P(N\geq 6|M_1=1,M_2=3)=\alpha\cdot f^2\cdot P(N\geq 6)$

- 其中$\alpha=P(M_1=1,M_2=3)$，考虑到$f\ll e$，则有$P(N=2|M_1=1,M_2=3)$最大，因此在该假设下，最可能的恒星数目为2。

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$\mathbf{P}(N|M_1 = 2, M_2 = 2)$。

$$P\left(N \mid M_1 = 2, M_2 = 2\right) = \alpha \sum_{f_1, f_2} P\left(f_1, f_2, N, M_1 = 2, M_2 = 2\right)$$

$$= \alpha \sum_{f_1, f_2} P\left(f_1\right) P\left(f_2\right) P(N) P\left(M_1 = 2 \mid f_1, N\right) P\left(M_2 = 2 \mid f_2, N\right)$$

展开，化简后得到：

$$P\left(N \mid M_1 = 2, M_2 = 2\right)$$

$$= \frac{e^2}{6e^2 - 4e + 1}[P(N = 1) + P(N = 3)] + \frac{(1 - 2e)^2}{6e^2 - 4e + 1} P(N = 2)$$