Homework 6

PB17000297 罗晏宸

April 6 2020

1 Exercise 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的疾病的测试结果呈阳性，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？

解 作为测试结果呈阳性的人，我关心的是在这个条件下的我患病的条件概率，即 $P(\text{患病}|\text{阳性})$，而已知 $P(\text{阳性}|\text{患病}) = 0.99$，$P(\text{阴性}|\text{未患病}) = 0.99$，$P(\text{患病}) = \frac{1}{10000}$，有

$$\begin{aligned}
P(\text{患病}|\text{阳性}) &= \frac{P(\text{患病} \wedge \text{阳性})}{P(\text{阳性})} \\
&= \frac{P(\text{阳性}|\text{患病})P(\text{患病})}{P(\text{阳性}|\text{患病})P(\text{患病}) + P(\text{阴性}|\text{未患病})P(\text{未患病})} \\
&= \frac{0.99 \times \frac{1}{10000}}{0.99 \times \frac{1}{10000} + (1 - 0.99) \times (1 - \frac{1}{10000})} \\
&= \frac{1}{102} \\
&\approx 0.009804
\end{aligned}$$

由式可见，我患病的条件概率随 $P(\text{患病})$ 减小而减小，因此“这种病很罕见”对我而言是一个好消息，事实上当测试的准确度比个体不患病的概率小

很多时，阳性的测试结果很大可能是误诊。从结果来看，我患这种病的概率仅为 $\frac{1}{102}$，不足 1%。

2 Exercise 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？

解

a 下面用 $P(\text{伪币}|\text{正面})$ 简记随机取出一枚硬币，抛出落地后正面朝上的条件下，取出的是伪币的条件概率。

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

b

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

c 过程发生错误即取出了一枚正常硬币但依然有抛 k 次后都是正面朝上的结果（取出伪币但是有反面朝上的结果是不存在的），返回 *fake*。

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

3 Exercise 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。

解

a 模型的概率分布由先验概率 $P(\text{类别} = A)$ 和条件概率 $P(\text{单词}_i|\text{类别} = A)$，前者描述所有文档中类别 A 的比例，后者表示类别为 A 的文档中包含单词 i 的比例。

b 对于一个给定的新文档，其内容是已知的，通过统计其中各单词的出现频率，由于单词的出现频率是由文档类别决定的，可以对于文档的类别给出断言。

c 不合理，单词不一定是原子的，复合词的出现概率并不一定等于其部件出现概率之积。