第八次作业

王原龙 05/30/2022

13.15 After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease and that the test is 99% accurate (i.e., the probability of testing positive when you do have the disease is 0.99, as is the probability of testing negative when you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people of your age. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

设得病事件为A，检测阳性事件为B，则有

$$P(A)=\frac{1}{10000},P(B|A)0.99,P(B|\bar{A})=0.01$$

于是由全概率公式可以得到先验概率

$$P(B)=P(A)P(B|A)+P(\bar{A})P(B|\bar{A})=\frac{1}{102}$$

由贝叶斯公式得到所求概率

$$P(A|B)=\frac{P(A)P(B|A)}{P(B)}=0.0098$$

13.18 Suppose you are given a bag containing n unbiased coins. You are told that $n-1$ of these coins are normal, with heads on one side and tails on the other, whereas one coin is a fake, with heads on both sides.

a. Suppose you reach into the bag, pick out a coin at random, flip it, and get a head. What is the (conditional) probability that the coin you chose is the fake coin?

b. Suppose you continue flipping the coin for a total of k times after picking it and see k heads. Now what is the conditional probability that you picked the fake coin?

c. Suppose you wanted to decide whether the chosen coin was fake by flipping it k times. The decision procedure returns *fake* if all k flips come up heads; otherwise it returns *normal*. What is the (unconditional) probability that this procedure makes an error?

同样是贝叶斯定理的应用，与前一题没有什么区别，这里就给出答案。

$$(a)\ \frac{2}{n+1}\quad (b)\ \frac{2^k}{n-1+2^k}\quad (c)\ \frac{n-1}{n2^k}$$

13.22 Text categorization is the task of assigning a given document to one of a fixed set of categories on the basis of the text it contains. Naive Bayes models are often used for this task. In these models, the query variable is the document category, and the "effect" variables are the presence or absence of each word in the language; the assumption is that words occur independently in documents, with frequencies determined by the document category.

a. Explain precisely how such a model can be constructed, given as "training data" a set of documents that have been assigned to categories.

b. Explain precisely how to categorize a new document.

c. Is the conditional independence assumption reasonable? Discuss.

易疏忽的点：

(a) 描述"使用频率估计概率"的方法，只说计算先验后验概率是不完整的。（补充：零概率处理）

(b) 写出朴素贝叶斯进行分类的计算公式

(c) 混淆独立与条件独立的关系

- 同学们的答案几乎可以如此概括：不合理，因为词之间不是独立的，比如xx和yy倾向于一起出现为xxyy的形式（此时$P(AB)?P(A)P(B)$）。
- xxyy包括但不限于"人工智能"、"机器学习"、"编译原理"等我觉得拆开也完全没有问题的词。