朴素贝叶斯分类

万永权

目录

CONTENTS

托马斯·贝叶斯

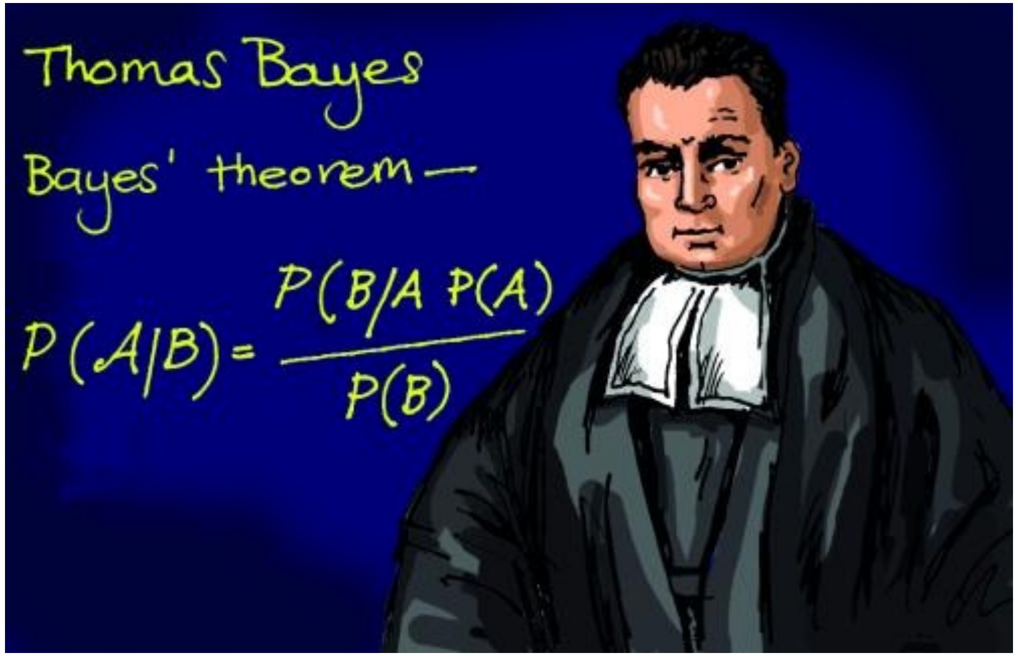
- 18世纪英国长老会牧师、业余数学家贝叶斯提出过一种看上去似乎显而易见的观点：
- 用客观的新信息，更新我们最初关于某个事物的信念后，我们就会得到一个新的、改进了的信念。

初始信念加上新证据等于新的改进信念

托马斯·贝叶斯
(Thomas Bayes，1702～1761)

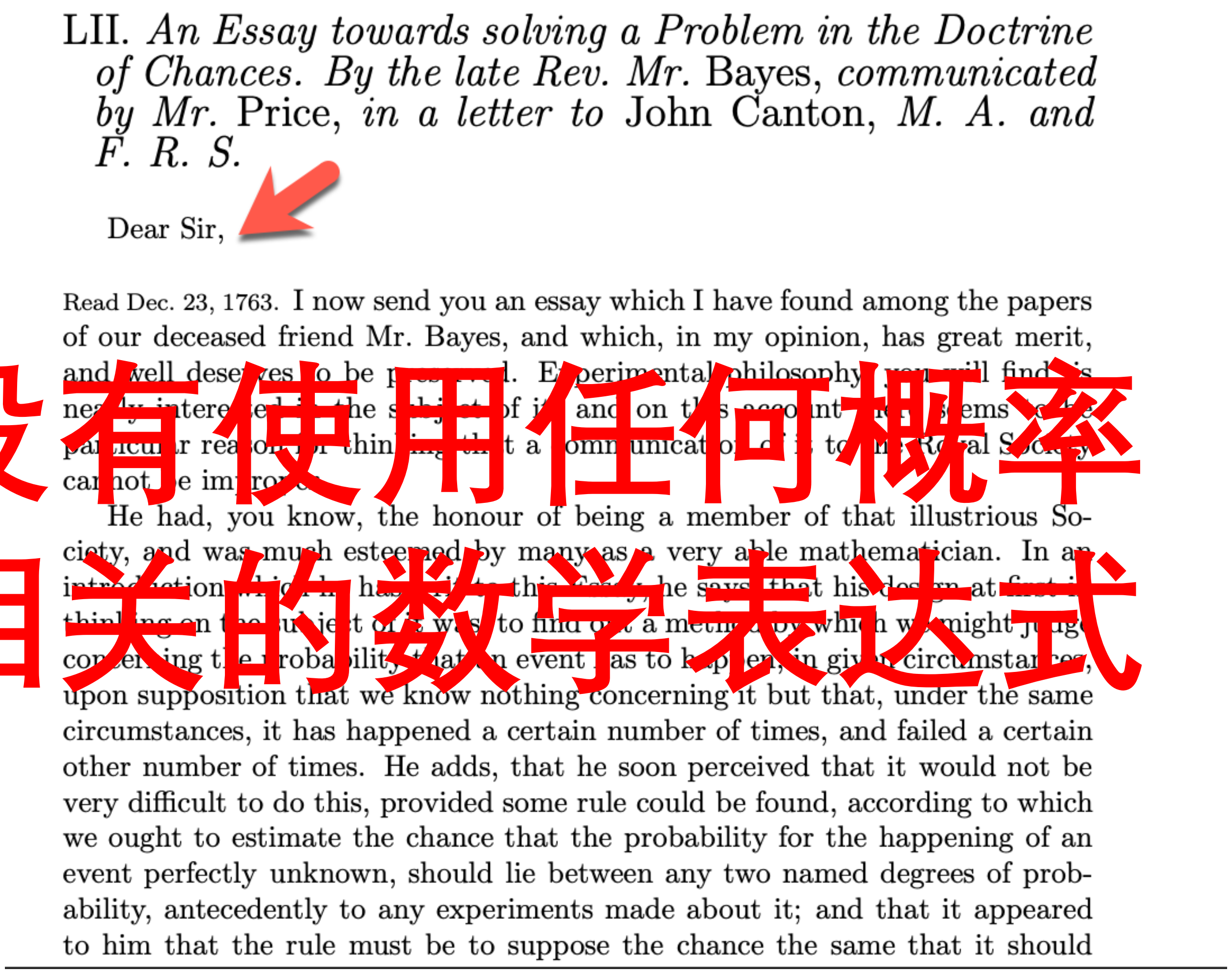
托马斯·贝叶斯的“神作”

一篇解决概率学说问题的文章



Bayes T. Essay towards solving a problem in the doctrine of chances[J]. Biometrika, ~~1763~~, 45: 293-315.

由他的一位朋友理查德·普莱斯（Richard Price）在他去世后发表的



LII. *An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *communicated by Mr.* Price, *in a letter to* John Canton, *M. A. and F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved. Experimental philosophy, you will find, is nearly interested in the subject of it; and on this account there seems to be particular reason for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illustrious Society, and was much esteemed by many as a very able mathematician. In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times. He adds, that he soon perceived that it would not be very difficult to do this, provided some rule could be found, according to which we ought to estimate the chance that the probability for the happening of an event perfectly unknown, should lie between any two named degrees of probability, antecedently to any experiments made about it; and that it appeared to him that the rule must be to suppose the chance the same that it should

没有使用任何概率相关的数学表达式

你看到的公式并不是贝叶斯给出的

概率，只不过是把常识用数学公式表达了出来。

——拉普拉斯

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

拉普拉斯

概率论

- 概率论是研究随机现象中的数量规律的一门学科，反应了事物的不确定性。
- 随机现象：
  - 想同的条件下重复进行某种试验时，试验结果不一定完全相同且不可预知的现象。
  - 抛硬币、掷色子、买彩票

样本空间和随机事件

- 试验中每一个可能出现的结果称为试验的一个样本点，由全部样本点构成的集合称为样本空间。
  - 抛硬币： 2个
  - 色子： 6个
  - 彩票： ？ 中彩票的概率是多少?
- 随机事件:
  - 随机试验中可能发生也可能不发生的事件
  - 随机事件和样本空间的子集有一一对应关系。
  - 某次试验中，若事件包含的某一个样本点出现，则称事件发生。

统计概率

- 同一组条件下所做的大量重复试验中，事件A出现的频率P(A) 总是在[0,1]上的一个确定常数附近，且稳定，称为事件A的概率或先验概率。
- P(!A) = 1- P(A)

条件概率

- 假设A与B是某个随机试验的两个事件，如果在事件B发生的条件下考虑A发生的概率，就称它为事件A的条件概率P(A|B)。
- P(A|B) = P(A∩B) / P(B)

例： 1-7这7个数字中，取一个数字

样本空间 S = （1，2，3，4，5，6，7）

A: 取3的倍数　　P(A) = 2/7

B: 取偶数　　P(B) = 3/7

C: 既是3的倍数，又是偶数：　　P(C)=P(A∩B) = 1/7

D: B发生的条件下，A发生的概率：P(A|B) = 1/3

贝叶斯公式

- 先验概率：由以往的数据分析得到的概率
- 后验概率：得到“结果”的信息后重新修正的概率

- P(A) 是 A 的先验概率（边缘概率）
- P(B) 是 B 的先验概率（边缘概率）

- P(A|B) 是已知 B 发生后 A 的条件概率，由于得自 B的取值而被称为 A的后验概率
- P(B|A) 是已知 A 发生后 B 的条件概率，由于得自 A 的取值而被称为 B的后验概率

贝叶斯公式

条件概率

- P(A|B) 表示事件B已经发生的前提下，事件A发生的概率，叫做事件B发生下事件A的条件概率。
- 求解公式：

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

贝叶斯公式

$$P(B \mid A) = \frac{P(AB)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A)}$$

贝叶斯公式

$$P(A \mid B) = \frac{P(B, A)}{P(B)} \qquad P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

- 解决了两个事件条件概率的转换问题
- 贝叶斯定理是基于假设的先验概率、给定假设下观察到不同数据的概率，提供了一种计算后验概率的方法。
- 在人工智能领域，贝叶斯方法是一种非常具有代表性的不确定性知识表示和推理方法。

全概率公式

如果事件组B_1，B_2，... B_n 满足

1. B_1，B_2....两两互斥，即 $B_i \cap B_j = \emptyset$ ，$i \neq j$ ， $i,j=1，2，....$，且 $P(B_i)>0,i=1,2,....$;
2. $B_1 \cup B_2 \cup ....=\Omega$ ，则称事件组 $B_1,B_2,...$是样本空间Ω的一个划分。

A为任一事件，则:

$$P(A)=\sum_{i=1}^{n} P(B_i)P(A \mid B_i)$$

事件A分解成几个小事件，通过求小事件的概率，然后相加从而求得事件A的概率

即: $P(A)=P(AB_1)+P(AB_2)+....+P(AB_n)$

$=P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+...+P(A|B_n)P(PB_n)$

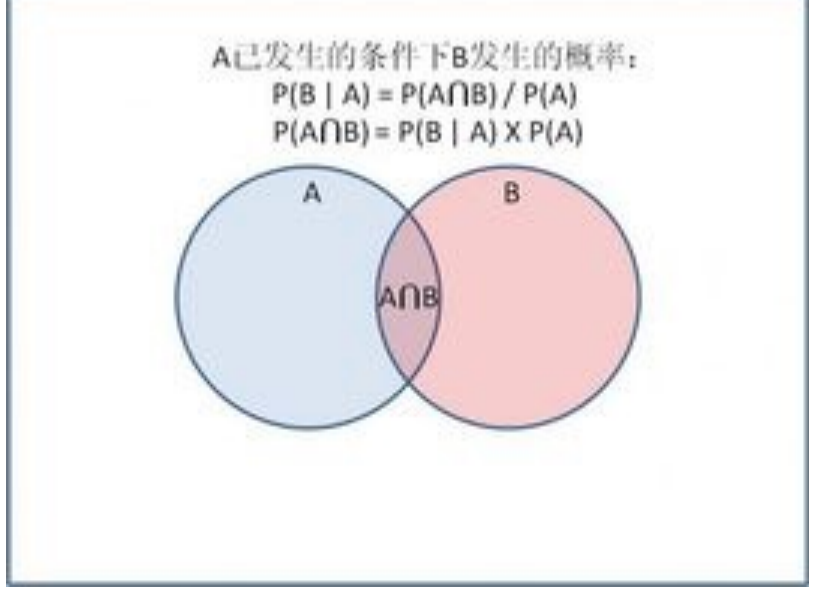
贝叶斯（Bayes）公式

- 若B_1，B_2，... B_n满足全概率公式中的条件，则对任意事件A，下式成立，该式称为Bayes公式：

$$P(B_i|A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_j P(B_j)P(A|B_j)}$$

- A发生的情况下，事件 B_i 的概率是多少？

贝叶斯公式推导



乘法定理

$$\mathrm{P}(AB) = P(A|B) \cdot P(B) \Rightarrow P(A|B) = \frac{\mathrm{P}(AB)}{P(B)}$$

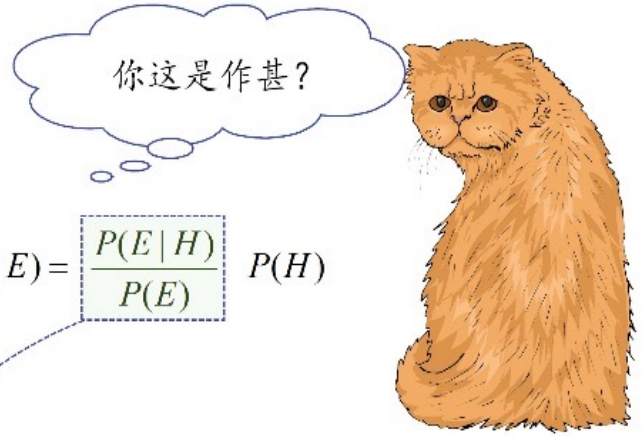
全概率

$$\mathrm{P}(B) = P(AB) + P(\bar{A}B) \Rightarrow P(A|B) = \frac{\mathrm{P}(AB)}{P(AB) + P(\bar{A}B)}$$

$$P(A|B) = \frac{\mathrm{P}(A)\mathrm{P}(B|\mathrm{A})}{P(AB) + P(\bar{A}B)}$$

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

（a）数学课本上的公式



$$P(H \mid E) = \frac{P(E \mid H)}{P(E)} P(H)$$

（b）具有物理意义的公式

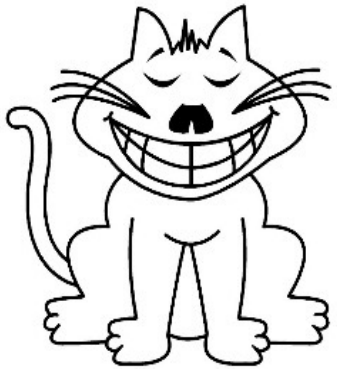
（c）贝叶斯原理解读

看到证据E后的更新概率

没有新证据的概率

$$P(H \mid E) = \frac{P(E \mid H)}{P(E)} P(H)$$

后验概率 = 调整因子 x 先验概率



使用贝叶斯汤姆断案

$$P(A|B) = \mathrm{P}(A)\frac{\mathrm{P}(B|\mathrm{A})}{P(AB) + P(\bar{A}B)}$$

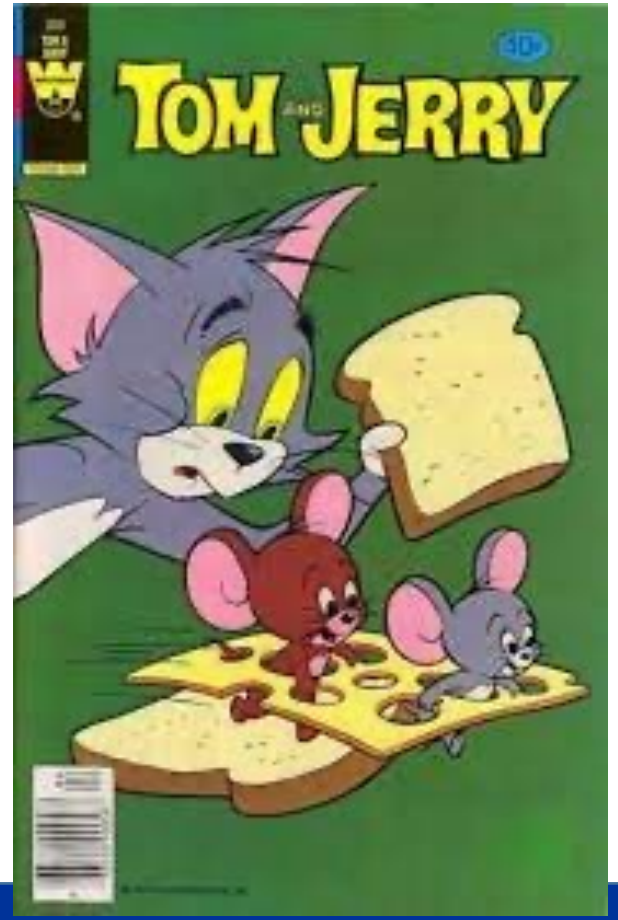
事件A：杰瑞偷了汤姆面包

事件B：杰瑞身上有奶油的味道

目标：计算条件概率P(A|B)

P(A):杰瑞偷面包的概率。假设这个汤姆很相信杰瑞，那么这个值就很低，比如说，$P(A)$ =1%

$P(B)$：表示杰瑞身上有奶油味道的概率。



Cont.

P(A)=1%， P(B|A)= 60%

$$P(A|B) = \mathrm{P}(A)\frac{\mathrm{P}(B|\mathrm{A})}{P(B|A) * P(A) + P(B|\bar{A}) * P(\bar{A})}$$

分母部分：P(B|A)= 60%:杰瑞偷了蛋糕的前提下，身上留下奶油味道的概率。

$P(B|\bar{A})$：表示杰瑞没有偷汤姆面包的前提下依然身上带有奶油味道的概率=10%

$$\mathrm{P(A|B)}=0.01*\frac{0.6}{0.6*0.01+0.1*0.99}=0.057$$

结论：（1）偷蛋糕的概率很低，不足6%，（2）但怀疑的概率提升了，从1%提升到6%

练习

- 一个村子，一共有3个小偷。A1小张，A2小英，A3小郑。警局已经对他们的偷窃能力有备案：小张去偷东西成功的概率为0，小英去偷东西成功的概率是1/2，小郑去偷东西成功的概率是1。某一天，村子一个人大喊：失窃啦！！！试问：这三人中，与这次失窃案件有关的概率是多少?

由题目我们知道，三个人去偷东西的概率是都是1/3，所以我们有：

$P(B)=\frac{1}{2}$ $P(A_1)=P(A_2)=P(A_3)=\frac{1}{3}$ $P(B|A_1)=0, P(B|A_2)=\frac{1}{2}, P(B|A_3)=1$

我们需要求的是后验概率：$P(A_1|B), P(A_2|B), P(A_3|B)$

$$P(A_1|B)=\frac{P(A_1B)}{P(B)}=\frac{P(A_1)P(B|A_1)}{\sum_{i=1}^{3}P(A_i)P(B|Ai)}=\frac{\frac{1}{3}*0}{\frac{1}{2}}=0$$

$$P(A_2|B)=\frac{P(A_2B)}{P(B)}=\frac{P(A_2)P(B|A_2)}{\sum_{i=1}^{3}P(A_i)P(B|Ai)}=\frac{\frac{1}{3}*\frac{1}{2}}{\frac{1}{2}}=\frac{1}{3}$$

$$P(A_3|B)=\frac{P(A_3B)}{P(B)}=\frac{P(A_3)P(B|A_3)}{\sum_{i=1}^{3}P(A_i)P(B|Ai)}=\frac{\frac{1}{3}*1}{\frac{1}{2}}=\frac{2}{3}$$

朴素贝叶斯分类

朴素贝叶斯分类需要用数学描述，如Class是某个类别集合，可以表示为Class={类别1，类别2，类别3，，，类别n}，待分类某个样本特征集合={特征1，特征2，特征3，，，特征n}。

$$P(B \mid A) = \frac{P(A \mid B)p(B)}{P(A)}$$

$$P(\text{类别} \mid \text{特征}) = \frac{P(\text{特征} \mid \text{类别})P(\text{类别})}{P(\text{特征})}$$

- 已知样本具有某些特征，求它属于什么类别，可以将其转换为：
  - 求已知类别概率*P*(类别)，这个可以根据已知样本统计得到；
  - *P*(特征|类别)，也可以根据已知样本统计出已知类别中某个特征的条件概率。
  - *P*(特征)对于分类来讲这个特征是一样的，不受影响。
  - 朴素贝叶斯算法成立的前提是各属性之间互相独立。

朴素贝叶斯分类算法

朴素贝叶斯分类算法是基于贝叶斯定理的，它的工作过程如下：

（1）每个数据样本用一个n维特征向量$X=\{x_1,x_2,\ldots,x_n\}$表示，分别描述对n个属性$A_1,A_2,\ldots,A_n$样本的n个度量。

（2）假定有m个类$C_1,C_2,\ldots,C_m$。给定一个未知的数据样本 X (即没有类标号)，分类法将预测 X 属于具有最高后验概率(条件 X 下)的类，即，朴素贝叶斯分类将未知样本分配给类 C_{i}，当且仅当

$$P(C_i|X)>P\left(C_j\middle|X\right),1\leq j\leq m,j\neq i$$

$P(C_{\mathrm{i}}|X)$最大的类C_{i}称为最大后验假定。根据贝叶斯定理

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$

朴素贝叶斯

根据 天气 (X) 决定 是否打网球 (Y)

No.	天气	气温	湿度	风	类别	No.	天气	气温	湿度	风	类别
1	晴	热	高	无	N	8	晴	适中	高	无	N
2	晴	热	高	有	N	9	晴	冷	正常	无	P
3	多云	热	高	无	P	10	雨	适中	正常	无	P
4	雨	适中	高	无	P	11	晴	适中	正常	有	P
5	雨	冷	正常	无	P	12	多云	适中	高	有	P
6	雨	冷	正常	有	N	13	多云	热	正常	无	P
7	多云	冷	正常	有	P	14	雨	适中	高	有	N

天气		温度		湿度		有风		打网球	
P	N	P	N	P	N	P	N	P	N
晴 2/9	3/5	热 2/9	2/5	高 3/9	4/5	否 6/9	2/5	9/14	5/14
云 4/9	0/5	暖 4/9	2/5	正常 6/9	1/5	是 3/9	3/5		
雨 3/9	2/5	凉 3/9	1/5						

天气	温度	湿度	有风	打网球
晴	凉	高	是	?

$$P(是|x_1,x_2,x_3,x_4)=\frac{P(x_1,x_2,x_3,x_4|是)*P(是)}{P(x1,x2,x3,x4)}$$

$$=\frac{P(x_1|是)P(x_2|是)P(x_3|是)P(x_4|是)\ *\ P(是)}{P(x_1,x_2,x_3,x_4)}$$

$$=\frac{(\frac{2}{9}*\frac{3}{9}*\frac{3}{9}*\frac{3}{9})\ *\ \frac{9}{14}}{P(x_1,x_2,x_3,x_4)}=\frac{0.00529}{P(x_1,x_2,x_3,x_4)}$$

天气		温度		湿度		有风		打网球	
P	N	P	N	P	N	P	N	P	N
晴 2/9	3/5	热 2/9	2/5	高 3/9	4/5	否 6/9	2/5	9/14	5/14
云 4/9	0/5	暖 4/9	2/5	正常 6/9	1/5	是 3/9	3/5		
雨 3/9	2/5	凉 3/9	1/5						

天气	温度	湿度	有风	打网球
晴	凉	高	是	?

$$P(否|x_1,x_2,x_3,x_4)=\frac{P(x_1,x_2,x_3,x_4|否)*P(否)}{P(x_1,x_2,x_3,x_4)}$$

$$=\frac{P(x_1|否)P(x_2|否)P(x_3|否)P(x_4|否)\ *\ P(否)}{P(x_1,x_2,x_3,x_4)}$$

$$=\frac{(\frac{3}{5}*\frac{1}{5}*\frac{4}{5}*\frac{3}{5})\ *\ \frac{5}{14}}{P(x_1,x_2,x_3,x_4)}=\frac{0.0206}{P(x_1,x_2,x_3,x_4)}$$

朴素贝叶斯

$$P(\text{是}|x_1, x_2, x_3, x_4) = \frac{0.00529}{P(x_1, x_2, x_3, x_4)}$$

$$P(\text{否}|x_1, x_2, x_3, x_4) = \frac{0.0206}{P(x_1, x_2, x_3, x_4)}$$

由于：$P(\text{是}|x_1, x_2, x_3, x_4) + P(\text{否}|x_1, x_2, x_3, x_4) = 1$

得到：$P(\text{是}|x_1, x_2, x_3, x_4) = 0.205$

$P(\text{否}|x_1, x_2, x_3, x_4) = 0.795$

朴素贝叶斯预测的结果是 不去打网球。

练习：贝叶斯分类

举个栗子：一对男女朋友，男生向女生求婚，男生的四个特点分别是不帅，性格不好，身高矮，不上进，请你帮助女生来决定是嫁还是不嫁？

帅？	性格好？	身高？	上进？	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
不帅	不好	矮	不上进	不嫁
帅	好	高	不上进	嫁
不帅	好	高	上进	嫁
帅	好	高	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

朴素贝叶斯算法经典应用

Scikit-learn提供了3种朴素贝叶斯模型，分别是GaussianNB、MultinomialNB和BernoulliNB。这3种模型适用的分类场景各不相同，其中：

- GaussianNB：先验为高斯分布的朴素贝叶斯，一般地，如果样本特征的分布大部分是连续值，使用GaussianNB会比较好。
- MultinomialNB：先验为多项式分布的朴素贝叶斯，如果样本特征的大部分是多元离散值，使用MultinomialNB比较合适。
- BernoulliNB：先验为伯努利分布的朴素贝叶斯，如果样本特征是二元离散值或者很稀疏的多元离散值，使用BernoulliNB比较合适。
- predict()：我们最常用的预测方法，直接输出测试集的预测类别。
- predict_proba()：它会给出测试集样本在各个类别上预测的概率。容易理解，predict_proba()预测出的各个类别概率里的最大值对应的类别，也就是predict()得到类别。
- predict_log_proba()：和predict_proba()类似，它会给出测试集样本在各个类别上预测的概率的一个对数转化。转化后predict_log_proba()预测出的各个类别对数概率里的最大值对应的类别，也就是predict()得到类别。

【例】 GaussianNB的实现。

```
from sklearn import datasets
iris=datasets.load_iris()
X=iris.data
y=iris.target

from sklearn.naive_bayes import GaussianNB
gnb=GaussianNB()
gnb.fit(X,y)
y_pred=gnb.fit(X,y).predict(iris.data)

print("Number of mislabeled points out of a total %d points:%d" % (iris.data.shape[0],  (iris.target!=y_pred).sum()))
```

运行程序，输出如下：

```
Number of mislabeled points out of a total 150 points:6
```

此外，GaussianNB的一个重要的功能是有 partial_fit()，这个函数一般用在训练集数据量非常大、一次不能全部载入内存的时候。这时我们可以把训练集分成若干等分，重复调用partial_fit()来一步步地学习训练集，非常方便。接下来介绍的MultinomialNB和BernoulliNB也有类似的功能。