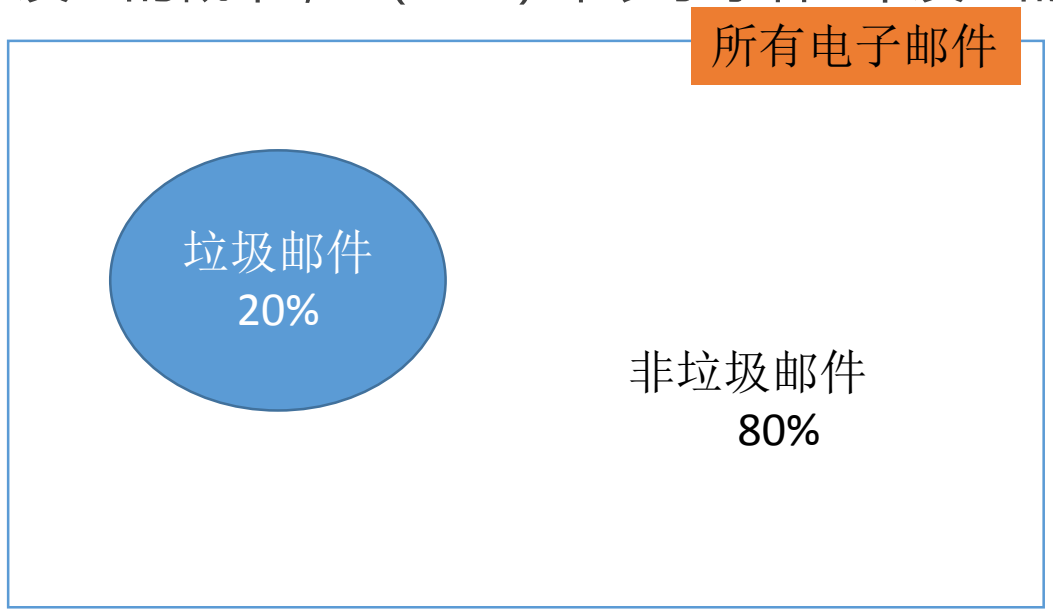
天善智能
TIANSHAN SOFT
专注于商业智能BI和大数据的垂直社区平台

朴素贝叶斯分类

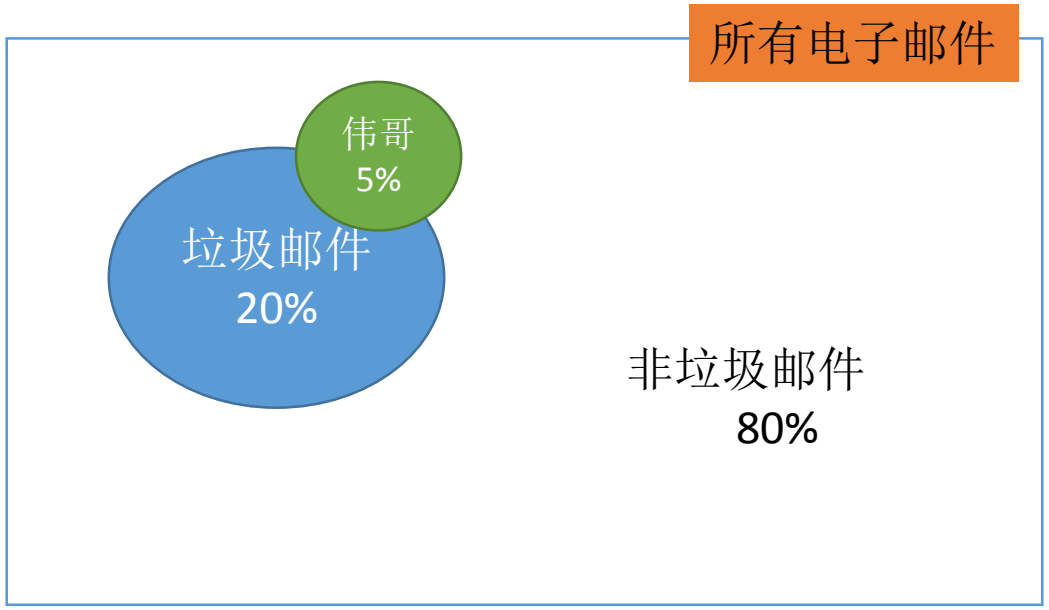
谢佳标（Daniel.xie）

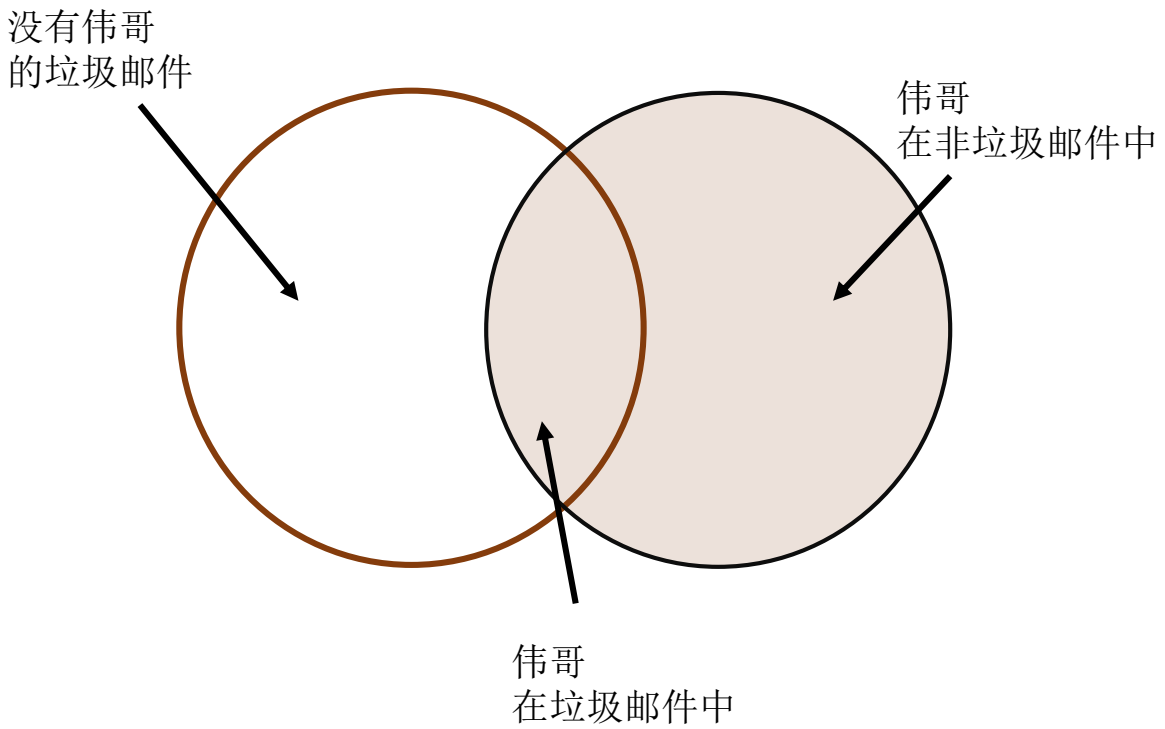
贝叶斯方法的基本概念

- 贝叶斯概率理论植根于这样一个思想，即一个事件的似然估计应建立在已有证据的基础上。
- 事件(event)就是可能的结果
- 试验(trail)就是事件发生一次的机会
- 一个事件发生的概率可以通过观测到的数据来估计，即用该事件发生的试验的次数除以试验的总次数
- 一个试验的所有可能结果的概率之和一定为100%
- 用符号P(A)来表示事件A发生的概率，P（¬A）来表示事件A不发生的概率



联合概率

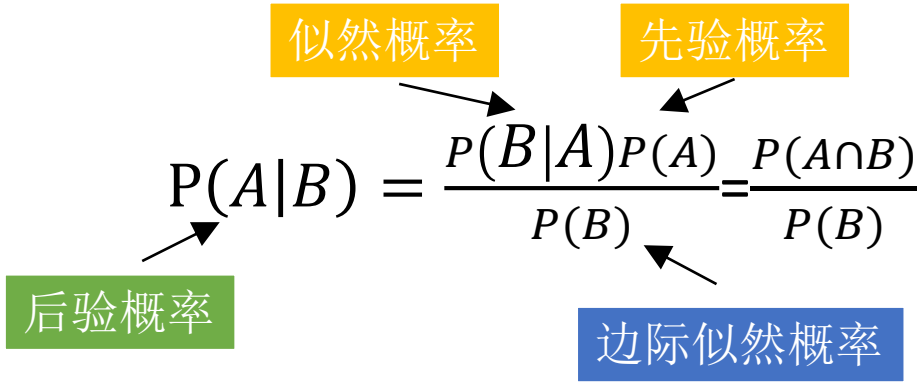




- 我们希望估计P(垃圾邮件)和P(伟哥)同时发生的概率，记为P(垃圾邮件∩伟哥)；
- 概率P(垃圾邮件∩伟哥)的计算取决于这两个事件的联合概率，即如何将一个事件发生的概率和另一个事件发生的概率连续在一起。如果这两个事件是完全不相关，我们成为独立事件。
- 如果P(垃圾邮件)和P(伟哥)是相互独立的，即P(垃圾邮件∩伟哥)= P(垃圾邮件)*P(伟哥)=0.2*0.05=0.01。

基于贝叶斯定理的条件概率

- 相关事件之间的关系可以用贝叶斯定理来描述，如下面的公式所示。符号P(A|B)表示在事件B已经发生的条件下，事件A发生的概率。这就是条件概率，因为事件A发生的概率依赖于事件B的发生。



$$P(\text{垃圾邮件}|\text{伟哥}) = \frac{P(\text{伟哥}|\text{垃圾邮件})P(\text{垃圾邮件})}{P(\text{伟哥})}$$

例子

	伟哥		
频数	Yes	No	总计
垃圾邮件	4	16	20
非垃圾邮件	1	79	80
总计	5	95	100

	伟哥		
似然	Yes	No	总计
垃圾邮件	4/20	16/20	20
非垃圾邮件	1/80	79/80	80
总计	5/100	95/100	100

$$\text{P}(\text{垃圾邮件} \cap \text{伟哥}) = P(\text{伟哥}|\text{垃圾邮件})P(\text{垃圾邮件}) = \left(\frac{4}{20}\right) * \left(\frac{20}{100}\right) = 0.04$$

$$\text{P}(\text{垃圾邮件}|\text{伟哥}) = \text{P}(\text{垃圾邮件} \cap \text{伟哥})*\text{P}(\text{伟哥}) = P(\text{伟哥}|\text{垃圾邮件})P(\text{垃圾邮件})*\text{P}(\text{伟哥})$$
$$= (4/20)*(20/100)*(5/100)=0.002$$

朴素贝叶斯算法

- 朴素贝叶斯(Baïve Bayes,NB)算法描述应用贝叶斯定理进行分类的一个简单应用。

优点	缺点
简单、快速、有效	依赖于一个常用的错误假设，即一样的重要性和独立性特征
能处理好噪声数据和缺失的数据	应用在含有大量数值特征的数据集时并不理想
需要用来训练的例子相对较少，但同样能处理好大量的例子	概率的估计值相对于预测的类而言更加不靠谱
很容易获得一个预测的估计概率值	

朴素贝叶斯分类-例子1

	W1		W2		W3		W4		
似然	Yes	No	Yes	No	Yes	No	Yes	No	总计
垃圾邮件	4/20	16/20	10/20	10/20	0/20	20/20	12/20	8/20	20
非垃圾邮件	1/80	79/80	14/80	66/80	8/80	71/80	23/80	57/80	80
总计	5/100	95/100	24/100	76/100	8/100	91/100	35/100	65/100	100

有一条消息包含单词W1和W4，不包含W2和W3，请判断此消息属于垃圾邮件还是非垃圾邮件？

垃圾邮件

$$P(\text{垃圾邮件}|W_1 \cap \neg W2 \cap \neg W3 \cap W4) = \frac{P(W_1 \cap \neg W2 \cap \neg W3 \cap W4|\text{垃圾邮件})P(\text{垃圾邮件})}{P(W_1 \cap \neg W2 \cap \neg W3 \cap W4)}$$

$$= \frac{P(\text{W1垃圾邮件})P(-W2\text{垃圾邮件})P(\neg W3\text{垃圾邮件})P(w4\text{垃圾邮件})P(\text{垃圾邮件})}{P(W1)P(\neg W2)P(\neg W3)P(W4)}$$ 0.012

$$= \frac{(4/20)(10/20)(20/20)(12/20)(20/100)}{(\frac{5}{100})(\frac{76}{100})(\frac{91}{100})(\frac{35}{100})}$$

0.012/(0.012+0.002)=0.857

$$P(\text{非垃圾邮件}|W_1 \cap \neg W2 \cap \neg W3 \cap W4) = \frac{P(W_1 \cap \neg W2 \cap \neg W3 \cap W4|\text{非垃圾邮件})P(\text{非垃圾邮件})}{P(W_1 \cap \neg W2 \cap \neg W3 \cap W4)}$$

$$= \frac{P(\text{W1非垃圾邮件})P(-W2\text{非垃圾邮件})P(\neg W3\text{非垃圾邮件})P(w4\text{非垃圾邮件})P(\text{非垃圾邮件})}{P(W1)P(\neg W2)P(\neg W3)P(W4)}$$ 0.002

朴素贝叶斯分类-例子2

性别	身高（英尺）	体重（磅）	脚掌（英寸）
男	6	180	12
男	5.92	190	11
男	5.58	170	12
男	5.92	165	10
女	5	100	6
女	5.5	150	8
女	5.42	130	7
女	5.75	150	9

这里的困难在于，由于身高、体重、脚掌都是连续变量，不能采用离散变量的方法计算概率。而且由于样本太少，所以也无法分成区间计算。怎么办？

这时，可以假设男性和女性的身高、体重、脚掌都是正态分布，通过样本计算出均值和方差，也就是得到正态分布的密度函数。有了密度函数，就可以把值代入，算出某一点的密度函数的值。

已知某人身高6英尺、体重130磅，脚掌8英寸，请问该人是男是女？

$$P(\text{男性}|\text{身高}=6 \cap \text{体重}=130 \cap \text{脚掌}=8) = 1.239414e{-}08$$

$$P(\text{女性}|\text{身高}=6 \cap \text{体重}=130 \cap \text{脚掌}=8) = 0.001075582$$

有99.99%的概率说明该人是女性