Python
机器学习实战

贝叶斯分类器

贝叶斯分类器

贝爷其人

贝叶斯 Thomas Bayes，英国数学家，1702年出生于伦敦，做过神甫．1742年成为英国皇家学会会员. 1763年4月7日逝世. 贝叶斯在数学方面主要研究概率论．他对统计推理的主要贡献是使用了“逆概率”这个概念, 在1763年提出了著名的贝叶斯公式.

贝叶斯分类器

概率的基本性质

- 事件的概率在0~1之间，即0≤P(A)≤1.
- 必然事件的概率为1
- 不可能事件的概率为0

- 概率的加法公式： P（A∪B)=P(A)+P(B) （当事件A与B互斥时）
- 事件B与事件A互为对立事件，P（A∪B）=1。由加法公式得到P(A)=1-P(B)

贝叶斯分类器

概率的基本知识

先验概率

- P(A)

后验概率（条件概率）

- P(A|B)

概率乘法公式

- P(AB)= P(A) P(B | A)

贝叶斯分类器

概率基本知识

两个人掷骰子，比点数大小，在比赛还未开始之前，问两个人各自获胜的概率以及打平手的概率是多少?

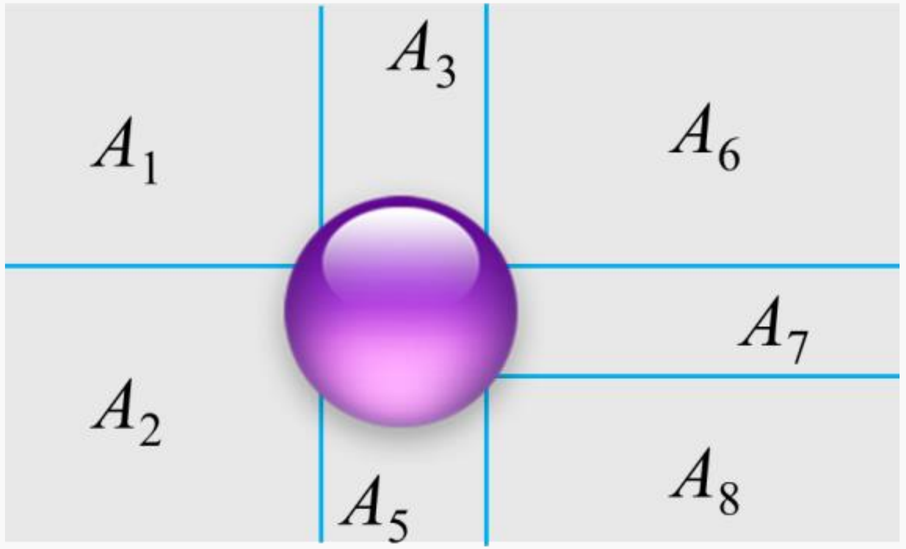
比赛开始，甲丢了个5点出来，问乙获胜的概率有多大。

贝叶斯分类器

全概率公式

$$P(B) = P(A_1B + A_2B + \cdots + A_nB) = \sum_{i=1}^{n} P(A_iB)$$

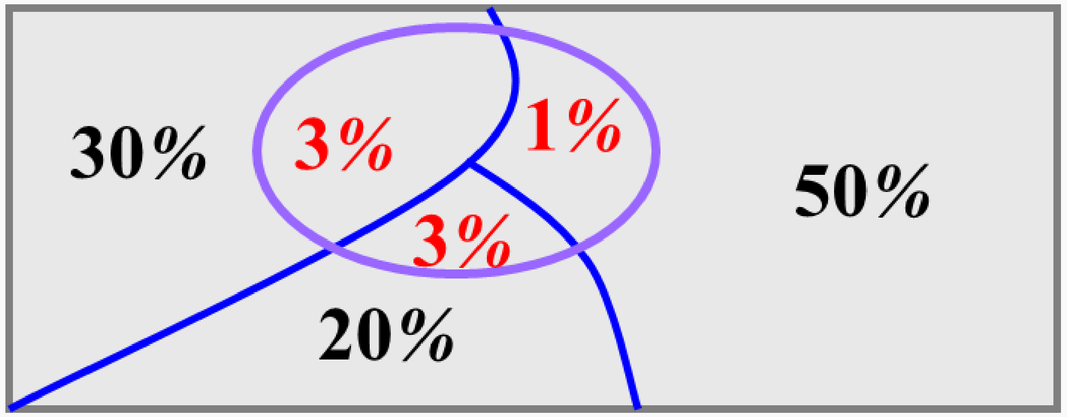
$$P(B) = \sum_{i=1}^{n} P(A_i)\,P(B \mid A_i)$$



全概率公式—例子

市场上有甲、乙、丙三家工厂生产的同一品牌产品,已知三家工厂的市场占有率分别为30%、20%、 50%,且三家工厂的次品率分别为 3%、3%、1%

试求市场上该品牌产品的次品率



贝叶斯定理

贝叶斯分类器

贝叶斯定理

已知三家工厂的市场占有率分别为30%、20%、50%, 次品率分别为3%、3%、1%.如果买了一件商品，发现是次品，问它是甲、乙、丙厂生产的概率分别为多少?

贝叶斯分类器

贝叶斯定理

贝叶斯(Bayes)于1763年提出. 它是在观察到事件B已发生的条件下，寻找导致B发生的每个原因A_k的概率.

$$P(A_k \mid B) = \frac{P(A_k B)}{P(B)}$$

乘法定理

全概率公式

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)} \qquad (k = 1,2,\cdots,n)$$

贝叶斯分类器

已知三家工厂的市场占有率分别为30%、20%、50%, 次品率分别为3%、3%、1%.如果买了一件商品，发现是次品，问它是甲、乙、丙厂生产的概率分别为多少?

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(B)} = \frac{0.3 \times 0.03}{0.02} = 0.45,$$

$$P(A_2 \mid B) = \frac{0.2 \times 0.03}{0.02} = 0.3, \quad P(A_3 \mid B) = \frac{0.5 \times 0.01}{0.02} = 0.25.$$

所以这件商品最有可能是甲厂生产的.

$P(A_i)$: 0.3, 0.2, 0.5

$P(A_i \mid B)$: 0.45, 0.3, 0.25

贝叶斯分类器

应用场景

垃圾邮件过滤

新闻分类

金融风险识别

…… ……

贝叶斯分类器

事件 $A_1, A_2, \cdots, A_n$ 看作是导致事件 B 发生的“原因”，在不知事件 B 是否发生的情况下，它们的概率为 $P(A_1), P(A_2), \cdots, P(A_n)$，通常称为**先验概率**.

现在有了新的信息已知（B 发生），我们对 $A_1, A_2, \cdots, A_n$ 发生的可能性大小 $P(A_1 \mid B), P(A_2 \mid B), \cdots, P(A_n \mid B)$ 有了新的估价，称为“**后验概率**”.

全概率公式可看成“由原因推结果”，而贝叶斯公式的作用在于“由结果推原因”：现在一个“结果” A 已经发生了，在众多可能的“原因”中，到底是哪一个导致了这一结果？

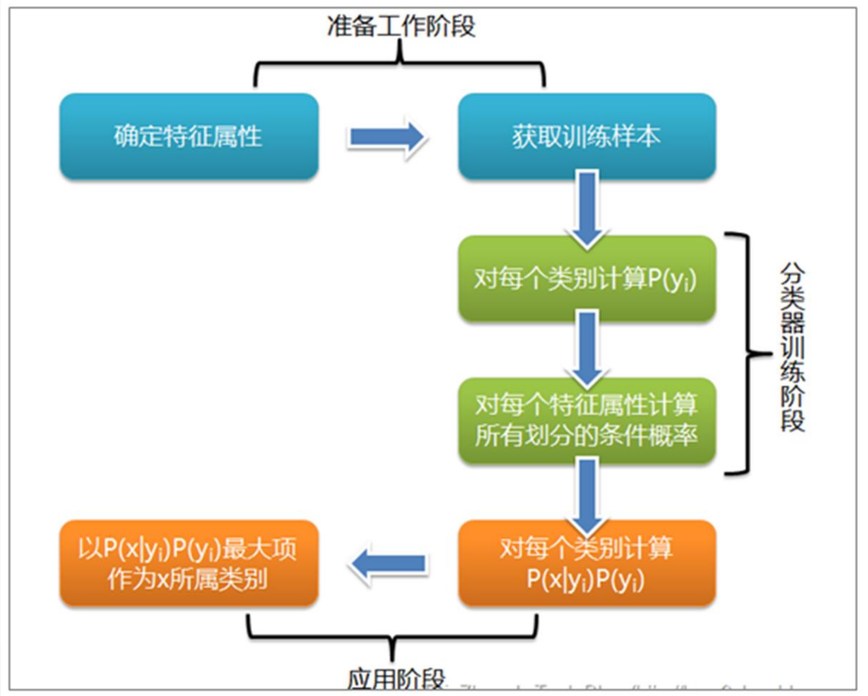
故贝叶斯公式也称为“**逆概公式**”.

实战

贝叶斯分类器

智能手环

```
both,sedentary,moderate,yes,i100
both,sedentary,moderate,no,i100
health,sedentary,moderate,yes,i500
appearance,active,moderate,yes,i500
appearance,moderate,aggressive,yes,i500
appearance,moderate,aggressive,no,i100
health,moderate,aggressive,no,i500
both,active,moderate,yes,i100
both,moderate,aggressive,yes,i500
appearance,active,aggressive,yes,i500
both,active,aggressive,no,i500
health,active,moderate,no,i500
health,sedentary,aggressive,yes,i500
appearance,active,moderate,no,i100
health,sedentary,moderate,no,i100
```



贝叶斯分类器

使用中的注意事项

1、加法平滑

2、对数加法

训练数据集

吸烟	饮酒	体重	运动	健康状况
不吸烟（10）	滴酒不沾（30）	偏瘦（20）	不运动（50）	亚健康(80)
偶尔（40）	偶尔（30）	适中（30）	经常运动(20)	
烟枪（30）	酒鬼（20）	肥胖（30）	运动机器(10)	
不吸烟	滴酒不沾	偏瘦	不运动	健康
偶尔	偶尔	适中	经常运动	
烟枪	酒鬼	肥胖	运动机器	
不吸烟	滴酒不沾	偏瘦	不运动	不健康
偶尔	偶尔	适中	经常运动	
烟枪	酒鬼	肥胖	运动机器	