

记录网站分析实践,分享Google Analytics应用与技巧

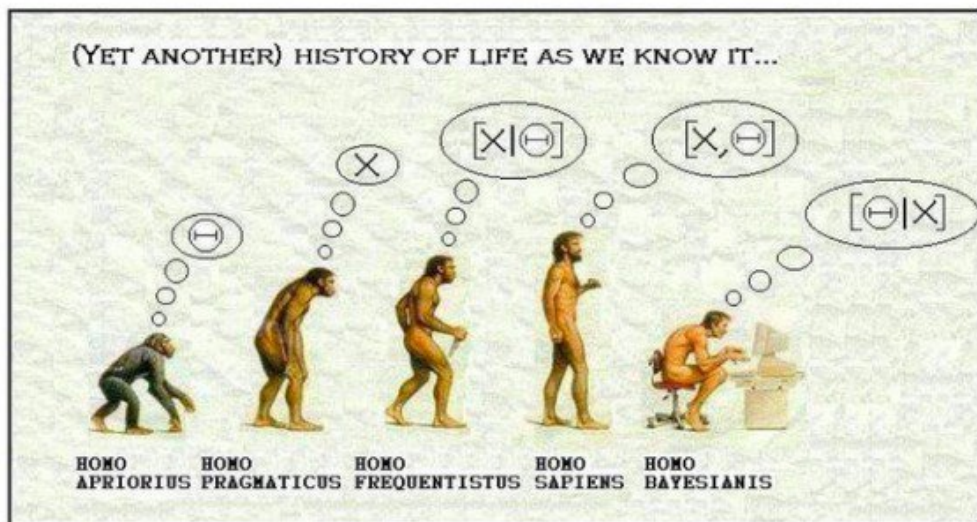
- [首页](#)
- [关于作者](#)
- [网站分析库](#)

•

## 朴素贝叶斯分类和预测算法的原理及实现

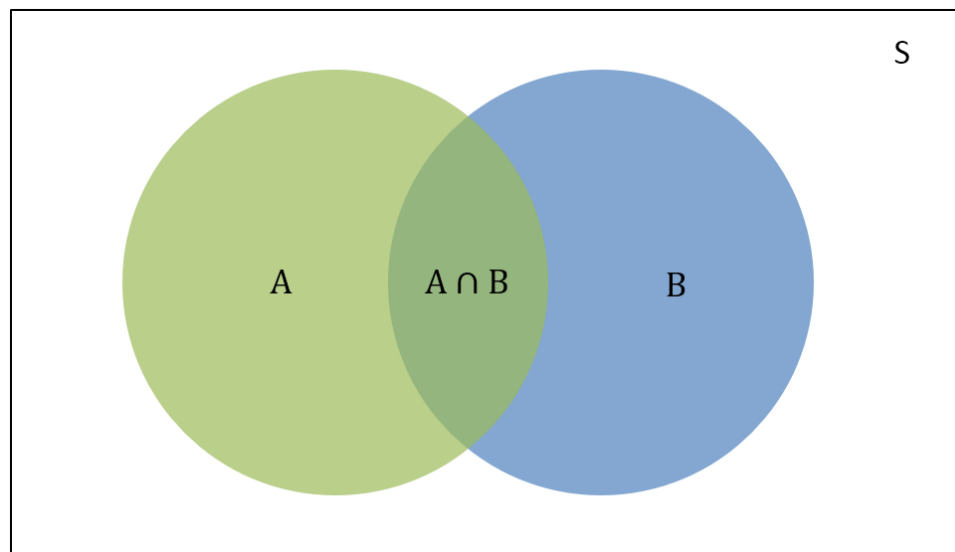
2016年4月1日 By [蓝鲸](#) [2 Comments](#)

决策树和朴素贝叶斯是最常用的两种分类算法，本篇文章介绍朴素贝叶斯算法。贝叶斯定理是以英国数学家贝叶斯命名，用来解决两个条件概率之间的关系问题。简单的说就是在已知 $P(A|B)$ 时如何获得 $P(B|A)$ 的概率。朴素贝叶斯（Naive Bayes）假设特征 $P(A)$ 在特定结果 $P(B)$ 下是独立的。



# 1. 概率基础：

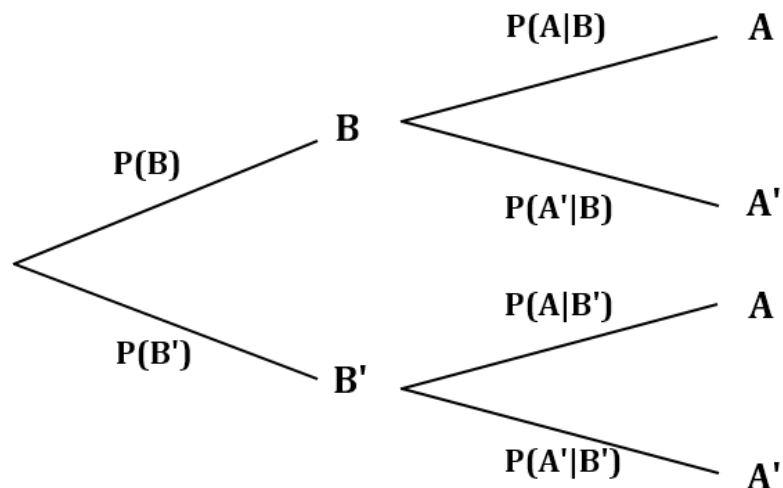
在开始介绍贝叶斯之前，先简单介绍下概率的基础知识。概率是某一结果出现的可能性。例如，抛一枚匀质硬币，正面向上的可能性多大？概率值是一个0-1之间的数字，用来衡量一个事件发生可能性的大小。概率值越接近1，事件发生的可能性越大，概率值越接近0，事件越不可能发生。我们日常生活中听到最多的是天气预报中的降水概率。概率的表示方法叫维恩图。下面我们通过维恩图来说明贝叶斯公式中常见的几个概率。



在维恩图中：

- $S$ ： $S$ 是样本空间，是所有可能事件的总和。
- $P(A)$ ：是样本空间 $S$ 中 $A$ 事件发生的概率，维恩图中绿色的部分。
- $P(B)$ ：是样本空间 $S$ 中 $B$ 事件发生的概率，维恩图中蓝色的部分。
- $P(A \cap B)$ ：是样本空间 $S$ 中 $A$ 事件和 $B$ 事件同时发生的概率，也就是 $A$ 和 $B$ 相交的区域。
- $P(A|B)$ ：是条件概率，是 $B$ 事件已经发生时 $A$ 事件发生的概率。

对于条件概率，还有一种更清晰的表示方式叫概率树。下面的概率树表示了条件概率 $P(A|B)$ 。与维恩图中的 $P(A \cap B)$ 相比，可以发现两者明显的区别。 $P(A \cap B)$ 是事件 $A$ 和事件 $B$ 同时发现的情况，因此是两者相交区域的概率。而事件概率 $P(A|B)$ 是事件 $B$ 发生时事件 $A$ 发生的概率。这里有一个先决条件就是 $P(B)$ 要首先发生。



因为条件概率 $P(A|B)$ 是在事件B已经发生的情况下，事件A发生的概率，因此 $P(A|B)$ 可以表示为事件A与B的交集与事件B的比率。

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

该公式还可以转换为以下形式，以便我们下面进行贝叶斯公式计算时使用。

$$P(A \cap B) = P(A | B) \times P(B)$$

## 2. 贝叶斯公式：

贝叶斯算法通过已知的 $P(A|B)$ ， $P(A)$ 和 $P(B)$ 三个概率计算 $P(B|A)$ 发生的概率。假设我们现在已知 $P(A|B)$ ， $P(A)$ 和 $P(B)$ 三个概率，如何计算 $P(B|A)$ 呢？通过前面的概率树及 $P(A|B)$ 的概率可知， $P(B|A)$ 的概率是在事件A发生的前提下事件B发生的概率，因此 $P(B|A)$ 可以表示为事件B与事件A的交集与事件A的比率。

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

该公式同样可以转化为以下形式：

$$P(B \cap A) = P(B | A) \times P(A)$$

到这一步，我们只需要证明 $P(A \cap B) = P(B \cap A)$ 就可以证明在已知 $P(A|B)$ 的情况下可以通过计算获得 $P(B|A)$ 的概率。我们将概率树转化为下面的概率表，分别列出 $P(A|B)$ ,  $P(B|A)$ ,  $P(A)$ , 和  $P(B)$  的概率。

Relative size	Case B	Case $\bar{B}$	Total
Condition A	w	x	w+x
Condition $\bar{A}$	y	z	y+z
Total	w+y	x+z	w+x+y+z

$$\begin{array}{c}
 \begin{array}{|c|} \hline \text{shaded} \\ \hline \end{array} \\
 P(A|B) \times P(B) = \frac{w}{w+y} \times \frac{w+y}{w+x+y+z} = \frac{w}{w+x+y+z}
 \end{array}$$

$$\begin{array}{c}
 \begin{array}{|c|c|} \hline \text{shaded} & \\ \hline \end{array} \\
 P(B|A) \times P(A) = \frac{w}{w+x} \times \frac{w+x}{w+x+y+z} = \frac{w}{w+x+y+z}
 \end{array}$$

通过计算可以证明 $P(A|B) \times P(B)$ 和 $P(B|A) \times P(A)$ 最后求得的结果是概率表中的同一个区域的值，因此：

$$P(A \cap B) = P(B \cap A)$$

我们通过 $P(A \cap B) = P(B \cap A)$ 证明了在已知 $P(A|B)$ ， $P(A)$ 和 $P(B)$ 三个概率的情况下可以计算出 $P(B|A)$ 发生的概率。整个推导和计算过程可以说得通。但从统计学的角度来看， $P(A|B)$ 和 $P(B|A)$ 两个条件概率之间存在怎样的关系呢？我们从贝叶斯推断里可以找到答案。

### 3. 贝叶斯推断：

贝叶斯推断可以说明贝叶斯定理中两个条件概率之间的关系。换句话说就是我们为什么可以通过 $P(A|B)$ ， $P(A)$ 和 $P(B)$ 三个概率计算出 $P(B|A)$ 发生的概率。

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

在贝叶斯推断中，每一种概率都有一个特定的名字：

- $P(B)$ 是“先验概率” (Prior probability)。
- $P(A)$ 是“先验概率” (Prior probability)，也作标准化常量(normalized constant)。
- $P(A|B)$ 是已知B发生后A的条件概率，叫做似然函数(likelihood)。
- $P(B|A)$ 是已知A发生后B的条件概率，是我们要求的值，叫做后验概率。
- $P(A|B)/P(A)$ 是调整因子，也被称作标准似然度 ( standardised likelihood ) 。

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

贝叶斯推断中有几个关键的概念需要说明下：

- 第一个是先验概率，先验概率是指我们主观通过事件发生次数对概率的判断。
- 第二个是似然函数，似然函数是对某件事发生可能性的判断，与条件概率正好相反。通过事件已经发生的概率推算事件可能性的概率。

**维基百科中对似然函数与概率的解释：**

**概率：**是给定某一参数值，求某一结果的可能性。

例如，抛一枚匀质硬币，抛10次，6次正面向上的可能性多大？

**似然函数：**给定某一结果，求某一参数值的可能性。

例如，抛一枚硬币，抛10次，结果是6次正面向上，其是匀质的可能性多大？

- 第三个是调整因子：调整因子是似然函数与先验概率的比值，这个比值相当于一个权重，用来调整后验概率的值，使后验概率更接近真实概率。调整因子有三种情况，大于1，等于1和小于1。
  1. 调整因子 $P(A|B)/P(A) > 1$ ：说明事件可能发生的概率要大于事件已经发生次数的概率。
  2. 调整因子 $P(A|B)/P(A) = 1$ ：说明事件可能发生的概率与事件已经发生次数的概率相等。
  3. 调整因子 $P(A|B)/P(A) < 1$ ：说明事件可能发生的概率与事件小于已经发生次数的概率。

因此，贝叶斯推断可以理解为通过先验概率和调整因子来获得后验概率。其中调整因子是根据事件已经发生的概率推断事件可能发生的概率（通过硬币正面出现的次数来推断硬币均匀的可能性），并与已经发生的先验概率（硬币正面出现的概率）的比值。通过这个比值调整先验概率来获得后验概率。

$$\text{后验概率} = \text{先验概率} \times \text{调整因子}$$

## 4. 实例1：垃圾邮件分类

贝叶斯分类器比较有名的实验场景是对垃圾邮件进行分类和过滤。这里我们简单介绍下通过贝叶斯算法过滤垃圾邮件的过程。贝叶斯分类器需要依赖历史数据进行学习，假定包含关键词“中奖”的就算作垃圾邮件。我们先经过人工筛选找出10封邮件，并对包含关键词“中奖”的邮件标注为垃圾邮件（Spam）。

关键词 “中奖”	邮件类别
No	Email
Yes	Spam
Yes	Spam
No	Email
No	Spam
Yes	Spam
No	Email
Yes	Email
No	Email
No	Email

我们将普通邮件和垃圾邮件中出现“中奖”关键词的频率进行汇总，分别记录普通邮件中出现和未出现该关键词的次数和垃圾邮件中出现和未出现该关键词的次数，并分别进行汇总。

	Email	Spam	
Yes	1	3	4
No	5	1	6
	6	4	10

根据频率表计算出贝叶斯算法中所需的关键概率值，这里我们已知普通邮件的概率 $P(\text{Email})$ ，垃圾邮件的概率 $P(\text{Spam})$ ，出现关键词的概率 $P(\text{Yes})$ ，未出现关键词的概率 $P(\text{No})$ ，以及垃圾邮件出现关键词的概率 $P(\text{Yes}|\text{Spam})$ 。

		$P(\text{Yes} \text{Spam})$		
	Email	Spam		
Yes	0.17	0.75	0.40	$P(\text{Yes})$
No	0.83	0.25	0.60	$P(\text{No})$
	0.60	0.40		
	$P(\text{Email})$	$P(\text{Spam})$		

按照贝叶斯公式，已知 $P(B|A)$ ， $P(A)$ 和 $P(B)$ 的概率。求 $P(A|B)$ 的概率。

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

我们将贝叶斯公式套用到垃圾邮件分类中，已知垃圾邮件中出现“中奖”关键词的概率，和垃圾邮件及“中奖”关键词的概率，求出现“中奖”关键词是垃圾邮件的概率。

- $P(A)=P(\text{垃圾邮件})=0.40$
- $P(B)=P(\text{出现关键词})=0.40$
- $P(B|A)=P(\text{出现关键词}|\text{垃圾邮件})=0.75$
- $P(A|B)=P(\text{垃圾邮件}|\text{出现关键词})$
- 

$$P(\text{垃圾邮件} | \text{出现关键词}) = \frac{P(\text{出现关键词} | \text{垃圾邮件}) \times P(\text{垃圾邮件})}{P(\text{出现关键词})}$$

$$P(\text{垃圾邮件} | \text{出现关键词}) = \frac{0.75 \times 0.40}{0.40} = 0.75$$

## 5. 实例2：病情预测

除了垃圾邮件分类，再来看一个病情预测的实例。通过历史数据已知几类疾病的病症及患病职业。那么如果新来的一位打喷嚏的建筑工人，如何通过贝叶斯算法通过历史数据来预测这位打喷嚏的建筑工人患感冒的概率呢？以下是6位历史病例的数据。



症状	职业	疾病
打喷嚏	护士	感冒
打喷嚏	农夫	过敏
头痛	建筑工人	脑震荡
头痛	建筑工人	感冒
打喷嚏	教师	感冒
头痛	教师	脑震荡

根据疾病的种类，我们分别对不同病症和不同职业患病的频率进行了统计。以下分别是不同症状与对应疾病发生的频率表，和不同职业与所对应疾病发生的频率表。

	感冒	过敏	脑震荡	
打喷嚏	2	1	0	3
头痛	1	0	2	3
	3	1	2	6

	感冒	过敏	脑震荡	
护士	1	0	0	1
农夫	0	1	0	1
建筑工人	1	0	1	2
教师	1	0	1	2
	3	1	2	6

根据两个频率表分布计算出贝叶斯算法中所需的概率值，这里我们已知每种疾病的概率，不同职业和不同症状的概率，以及患感冒后打喷嚏和职业为建筑工人的概率。

	P(打喷嚏   感冒)			
	感冒	过敏	脑震荡	
打喷嚏	0.67	1.00	0.00	P(打喷嚏)
头痛	0.33	0.00	1.00	P(头疼)
	0.50	0.17	0.33	
	P(感冒)	P(过敏)	P(脑震荡)	

	P(建筑工人   感冒)			
	感冒	过敏	脑震荡	
护士	0.33	0.00	0.00	P(护士)
农夫	0.00	1.00	0.00	P(农夫)
建筑工人	0.33	0.00	0.50	P(建筑工人)
教师	0.33	0.00	0.50	P(教师)
	0.50	0.17	0.33	
	P(感冒)	P(过敏)	P(脑震荡)	

按照贝叶斯公式，已知 $P(B \times C|A)$ ， $P(A)$ 和 $P(B \times C)$ 的概率。求 $P(A|B \times C)$ 的概率。

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

我们假设护士和打喷嚏这两个特征在感冒这个结果下是独立的，因此，上面的贝叶斯公式可以转化为朴素贝叶斯公式：

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

我们将贝叶斯公式套用到疾病预测中：

- $P(A)=P(\text{感冒})=0.5$
- $P(B)=P(\text{打喷嚏})=0.5$

- $P(C)=P(\text{建筑工人})=0.33$
- $P(B|A)=P(\text{打喷嚏}|\text{感冒})=0.67$
- $P(C|A)=P(\text{建筑工人}|\text{感冒})=0.33$


$$P(\text{感冒}|\text{打喷嚏} \times \text{建筑工人}) = \frac{P(\text{打喷嚏}|\text{感冒}) \times P(\text{建筑工人}|\text{感冒}) \times P(\text{感冒})}{P(\text{打喷嚏}) \times P(\text{建筑工人})}$$

$$P(\text{感冒}|\text{打喷嚏} \times \text{建筑工人}) = \frac{0.67 \times 0.33 \times 0.5}{0.5 \times 0.33} = 0.67$$

—【所有文章及图片版权归 蓝鲸（王彦平）所有。欢迎转载，但请注明转自“[蓝鲸网站分析博客](#)”。】—

Filed Under: [网站数据分析](#) Tagged With: [朴素贝叶斯](#)

## Comments

1.  [zengda](#) says:  
[2016年4月6日 at 上午11:11](#)

不错，不错，看看了！

[回复](#)

2.  [samSmith](#) says:  
[2017年3月6日 at 上午4:22](#)

简直就是精髓啊！！！！

[回复](#)

## Speak Your Mind

Name \*

Email \*

Website

发表评论

**出版物**



## 《人人都是网站分析师》



## [《网站分析实战》](#)



## [《流量的秘密》第二版](#)

[Return to top of page](#)

Copyright © 2018 [Genesis Framework](#) · [WordPress](#) · [Log in](#)

[站长统计](#)