

1. 贝叶斯理论

1. 贝叶斯定理

1.1 贝叶斯公式

“如果一个袋子中共有 10 个球，分别是黑球和白球，但是我们不知道它们之间的比例是怎么样的，现在，**仅通过摸出的球的颜色，是否能判断出袋子里面黑白球的比例？**”

上述问题可能与我们高中时期所接受的的概率有所冲突，因为你所接触的概率问题可能是这样的：“一个袋子里面有 10 个球，其中 4 个黑球，6 个白球，如果你随机抓取一个球，那么是黑球的概率是多少？”毫无疑问，答案是 0.4。这个问题非常简单，因为我们事先知道了袋子里面黑球和白球的比例，所以很容易算出摸一个球的概率，**但是在某些复杂情况下，我们无法得知“比例”，此时就引出了贝叶斯提出的问题。**

在统计学中有两个较大的分支：一个是“频率”，另一个便是“贝叶斯”，它们都有各自庞大的知识体系，而“贝叶斯”主要利用了**“相关性”**一词。

下面以通俗易懂的方式描述一下“贝叶斯定理”：通常，

事件 A 在事件 B 发生的条件下发生的概率： $P(A|B)$

事件 B 在事件 A 发生的条件下发生的概率： $P(B|A)$

它们两者的概率并不相同， $P(A|B)$ **不等于** $P(B|A)$

但是它们两者之间存在一定的相关性，并具有以下公式（称之为“贝叶斯公式”）：

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

首先我们要了解上述公式中符号的意义：

- $P(A)$ 这是概率中最基本的符号，表示 A 出现的概率
- $P(B|A)$ 是**条件概率**的符号，表示事件 A 发生的条件下，事件 B 发生的概率，条件概率是“贝叶斯公式”的关键所在，它也被称为**“似然度”**。
- $P(A|B)$ 是**条件概率**的符号，表示事件 B 发生的条件下，事件 A 发生的概率，这个计算结果也被称为“后验概率”。

有上述描述可知，贝叶斯公式可以预测事件发生的概率，**两个本来相互独立的事件，发生了某种“相关性”**，此时就可以通过“贝叶斯公式”实现预测。

1.2 先验概率

在贝叶斯看来，世界并非静止不动的，而是动态和相对的，他希望利用已知经验来进行判断，那么如何用经验进行判断呢？这里就必须提到“先验”和“后验”这两个词语。

我们先讲解“先验”，其实“先验”就相当于“未卜先知”，在事情即将发生之前，做一个概率预判。比如从远处驶来了一辆车，是轿车的概率是 45%，是货车的概率是 35%，是大客车的概率是 20%，在你没有看清之前基本**靠猜**，此时，我们把这个概率就叫做“先验概率”。

1.3 后验概率

在理解了“先验概率”的基础上，我们来研究一下什么是“后验概率？”

我们知道每一个事物都有自己的特征，比如前面所说的轿车、货车、客车，它们都有着各自不同的特征，距离过远的时候，我们无法用肉眼分辨，而当距离达到一定范围内就可以根据**各自的特征再次做出概率预判**，这就是后验概率。

比如轿车的速度相比于另外两者更快可以记做 $P(\text{轿车}|\text{速度快}) = 55\%$ ，而客车体型可能更大，可以记做 $P(\text{客车}|\text{体型大}) = 35\%$ 。

如果用**条件概率**来表述 $P(\text{体型大}|\text{客车})=35\%$ ，这种通过“车辆类别”推算出“类别特征”发生的概率的方法叫作“似然度”。这里的似然就是“可能性”的意思。

1.5 朴素+贝叶斯

了解完上述概念，你可能对贝叶斯定理有了一个基本的认识，**实际上贝叶斯定理就是求解后验概率的过程，而核心方法是通过似然度预测后验概率，通过不断提高似然度，自然也就达到了提高后验概率的目的。**

朴素贝叶斯是一种简单的贝叶斯算法，因为贝叶斯定理涉及到了概率学、统计学，其应用相对复杂，因此我们**只能以简单的方式使用它**，比如天真的认为，所有事物之间的特征都是相互独立的，彼此互不影响。

2. 朴素贝叶斯算法

2.1 多特征分类问题

下面我们使统计学的相关知识解决上述分类问题，分类问题的样本数据大致如下所示：

C++

- 1 [特征 X1 的值,特征 X2 的值,特征 X3 的值,.....,类别 A1]
- 2 [特征 X1 的值,特征 X2 的值,特征 X3 的值,.....,类别 A2]

解决思路：这里我们先简单的采用 1 和 0 代表特征值的有无，比如当 X1 的特征值等于 1 时，则该样本属于 A1 的类别概率；特征值 X2 值为 1 时，该样本属于类别 A1 的类别的概率。

依次类推，然后最终算出该样本对于各个类别的概率值，哪个**概率值最大**就可能是哪个类。

上述思路就是贝叶斯定理的典型应用，如果使用条件概率表达，如下所示：

Apache

- 1 $P(\text{类别A1} | \text{特征X1}, \text{特征X2}, \text{特征X3}, \dots)$

上述式子表达的意思是：**在特征 X1、X2、X3 等共同发生的条件下，类别 A1 发生的概率**，也就是**后验概率**。

2.2朴素贝叶斯算法

上一节我们已经了解了贝叶斯公式，下面使用**贝叶斯公式将多特征分类问题**表达出来，如下所示：

$$P(c | \mathbf{x}) = \frac{P(c)P(\mathbf{x} | c)}{P(\mathbf{x})}$$

来估计**后验概率** $P(\mathbf{x} | c)$ 的主要困难在于：类条件概率 $P(\mathbf{x} | c)$ 是**所有属性上的联合概率**，难以从有限的训练样本直接估计而得。

为避开这个障碍，**朴素贝叶斯分类器(naive Bayes classifier)**采用了**"属性条件独立性假设"(attribute conditional independence assumption)**：对已知类别，假设所有属性相互独立。

$$P(c|\mathbf{x}) = \frac{P(c)P(\mathbf{x}|c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^d P(x_i|c)$$

换言之，假设每个属性独立地对分类结果发生影响，其表达式为：**(目标方程)**

$$h_{nb}(\mathbf{x}) = \arg \max_{c \in \mathcal{U}} P(c) \prod_{i=1}^d P(x_i | c)$$

2.3朴素贝叶斯分类器的训练

朴素贝叶斯的训练就是基于数据集 D ，来估计“类先验概率 $P(x)$ ”，并为每个属性估计条件概率。
令 D_c 表示训练集 D 中第 c 类样本组成的集合，若有充足的独立同分布样本，则很容易估计出“类先验概率”

$$P(c) = \frac{|D_c|}{|D|}$$

- 对离散属性而言

令 D_{c,x_i} 表示 D_c 中在第 i 个属性上取值为 x_i 的样本组成的集

则条件概率 $P(x_i | c)$ 可估计为

$$P(x_i | c) = \frac{|D_{c,x_i}|}{|D_c|}$$

- 对连续属性

可考虑概率密度函数

假定

$$p(x_i | c) \sim \mathcal{N}(\mu_{c,i}, \sigma_{c,i}^2)$$

其中 $\mu_{c,i}$ 和 $\sigma_{c,i}^2$ 分别是第 c 类样本在第 i 个属性上取值的“均值”和“方差”

则有

$$p(x_i | c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$$

2.4手算朴素贝叶斯实例

训练集：

表 4.3 西瓜数据集 3.0									
编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

测试样例为：

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	？

1、计算“类先验概率”

共有17个样本，其中8个好瓜，9个坏瓜

$$P(\text{好瓜} = \text{是}) = \frac{8}{17} \approx 0.471$$

$$P(\text{好瓜} = \text{否}) = \frac{9}{17} \approx 0.529$$

2、为每个属性估计条件概率

(1) 色泽 = 青绿

好瓜里“色泽=青绿”的有3个，好瓜共8个

坏瓜里“色泽=青绿”的有3个，坏瓜共9个

$$P_{\text{青绿}|\text{是}} = P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{是}) = \frac{3}{8} = 0.375$$

$$P_{\text{青绿}|\text{否}} = P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333$$

(2) 纹理 = 蜷缩

好瓜里“纹理=蜷缩”的有5个，好瓜共8个

坏瓜里“纹理=蜷缩”的有3个，坏瓜共9个

$$P_{\text{蜷缩}|\text{是}} = P(\text{纹理} = \text{蜷缩} | \text{好瓜} = \text{是}) = \frac{5}{8} = 0.625$$

$$P_{\text{蜷缩}|\text{否}} = P(\text{纹理} = \text{蜷缩} | \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333$$

(3) 敲声 = 浊响

(4) 纹理 = 清晰

(5) 脐部 = 凹陷

(6) 触感 = 硬滑

$$P_{\text{浊响}|\text{是}} = P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{浊响}|\text{否}} = P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{否}) = \frac{4}{9} \approx 0.444$$

$$P_{\text{清晰}|\text{是}} = P(\text{纹理} = \text{清晰} | \text{好瓜} = \text{是}) = \frac{7}{8} = 0.875$$

$$P_{\text{清晰}|\text{否}} = P(\text{纹理} = \text{清晰} | \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222$$

$$P_{\text{凹陷}|\text{是}} = P(\text{脐部} = \text{凹陷} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{凹陷}|\text{否}} = P(\text{脐部} = \text{凹陷} | \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222$$

$$P_{\text{硬滑}|\text{是}} = P(\text{触感} = \text{硬滑} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{硬滑}|\text{否}} = P(\text{触感} = \text{硬滑} | \text{好瓜} = \text{否}) = \frac{6}{9} \approx 0.667$$

(7) 密度 = 0.697

μ 和 σ^2 分别是“正样本”在密度属性上取值的“均值”和“方差”

$$\mu = 0.574, \sigma^2 = 0.129^2$$

$$P_{\text{密度: 0.697}|\text{是}} = p(\text{密度} = 0.697 | \text{好瓜} = \text{是})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.129} \exp\left(-\frac{(0.697 - 0.574)^2}{2 \cdot 0.129^2}\right) \approx 1.959$$

μ 和 σ^2 分别是“负样本”在密度属性上取值的“均值”和“方差”

$$P_{\text{密度: 0.697}|\text{否}} = p(\text{密度} = 0.697 | \text{好瓜} = \text{否})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.195} \exp\left(-\frac{(0.697 - 0.496)^2}{2 \cdot 0.195^2}\right) \approx 1.203$$

(8) 含糖率 = 0.460

$$P_{\text{含糖: 0.460}|\text{是}} = p(\text{含糖率} = 0.460 | \text{好瓜} = \text{是})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.101} \exp\left(-\frac{(0.460 - 0.279)^2}{2 \cdot 0.101^2}\right) \approx 0.788$$

$P_{\text{含糖: 0.460}|\text{否}} = p(\text{含糖率} = 0.460 | \text{好瓜} = \text{否})$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.108} \exp\left(-\frac{(0.460 - 0.154)^2}{2 \cdot 0.108^2}\right) \approx 0.066$$

3、连乘，得出最终结果

公式：

$$h_{nb}(x) = \arg \max_{c \in Y} P(c) \prod_{i=1}^d P(x_i | c)$$

好瓜：

$$P(\text{好瓜} = \text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}} \\ \times P_{\text{硬滑}|\text{是}} \times p_{\text{密度: 0.697}|\text{是}} \times p_{\text{含糖: 0.460}|\text{是}} \approx 0.038,$$

坏瓜：

$$P(\text{好瓜} = \text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}} \\ \times P_{\text{硬滑}|\text{否}} \times p_{\text{密度: 0.697}|\text{否}} \times p_{\text{含糖: 0.460}|\text{否}} \approx 6.80 \times 10^{-5}.$$

结果：

$$0.038 > 6.80 \times 10^{-5}$$

可以看出“好瓜”概率更高，因此最终判别结果为“好瓜”

2.5拉普拉斯修正

2.5.1引入

上面的过程存在一个隐患！若某个属性值在训练集的某个类“没有出现过”，则

$$P(x_i|c) = 0$$

例如 好瓜里没有“敲声 = 清脆”的例子

$$P_{\text{清脆}|\text{是}} = P(\text{敲声} = \text{清脆} | \text{好瓜} = \text{是}) = \frac{0}{8} = 0$$

而在连乘时，如果存在这一项，则整个结果永远为 0。即“无论其他属性怎么样，永远不选择‘瓜=好’这个结果（因为 结果=0）”，这显然是不合理的

2.5.2拉普拉斯修正

为了避免这种情况，在估计概率值时通常需要进行“平滑”

拉普拉斯修正：

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N},$$
$$\hat{P}(x_i | c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_c}.$$

N 表示训练集 D 中可能的类别数

N_i 表示第 i 个属性可能的取值数

理解：

其实就是加入了

$$\frac{1}{\text{类别数}}$$

例如：

1、共有17个样本，其中8个好瓜，9个坏瓜，类别数2

$$\hat{P}(\text{好瓜} = \text{是}) = \frac{8+1}{17+2} \approx 0.474, \quad \hat{P}(\text{好瓜} = \text{否}) = \frac{9+1}{17+2} \approx 0.526$$

2、色泽 = 青绿

好瓜里“色泽=青绿”的有3个，好瓜共8个

坏瓜里“色泽=青绿”的有3个，坏瓜共9个

色泽的可能取值数为3（青绿、乌黑、浅白）

$$\hat{P}_{\text{青绿}|\text{是}} = \hat{P}(\text{色泽} = \text{青绿} | \text{好瓜} = \text{是}) = \frac{3+1}{8+3} \approx 0.364$$

$$\hat{P}_{\text{青绿}|\text{否}} = \hat{P}(\text{色泽} = \text{青绿} | \text{好瓜} = \text{否}) = \frac{3+1}{9+3} \approx 0.333$$

效果

拉普拉斯修正避免了因“训练集样本不充分”而导致“概率估计值为零”的问题

并且当训练集变大时，修正过程引入的先验影响也会逐渐变得可以忽略

使得估计值趋向于实际概率值

3.sklearn应用朴素贝叶斯算法

在 sklearn 库中，基于贝叶斯定理的算法集中在 sklearn.naive_bayes 包中，根据对“似然度 $P(x_i|y)$ ”计算方法的不同，我们将朴素贝叶斯大致分为三种：多项式朴素贝叶斯（MultinomialNB）、伯努利分布朴素贝叶斯（BernoulliNB）、高斯分布朴素贝叶斯（GaussianNB）。

另外一点要牢记，**朴素贝叶斯算法的实现是基于假设而来，在朴素贝叶斯看来，特征之间是相互独立的，互不影响的。**

C++

- 1 高斯朴素贝叶斯适用于特征呈正态分布的，多项式贝叶斯适用于特征是多项式分布的，伯努利贝叶斯适用于二项分布。

3.1 算法使用流程

使用朴素贝叶斯算法，具体分为三步：

- 统计样本数，即统计先验概率 $P(y)$ 和 似然度 $P(x|y)$ 。
- 根据待测样本所包含的特征，对不同类分别进行后验概率计算。
- 比较 y_1, y_2, \dots, y_n 的后验概率，哪个的概率值最大就将其作为预测输出。

3.2朴素贝叶斯算法应用

下面通过鸢尾花数据集对朴素贝叶斯分类算法进行简单讲解。如下所示：

Python

[illegible]