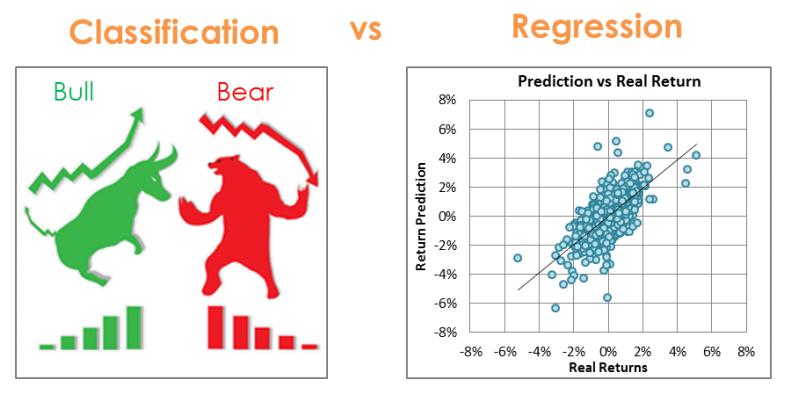


第12课：朴素贝叶斯分类器——从贝叶斯定理到分类模型

前面有关线性回归的课程中，我们讲了一个回归模型，我们现在来讲一个分类模型。

### 分类 vs 回归



****分类模型 VS 回归模型****，最根本的不同：前者是预测一个标签（类型、类别）；后者则是预测一个量。

换一个角度来看，分类模型输出的预测值是离散值；而回归模型输出的预测值则是连续值。

也就是说输入一个样本给模型，回归模型给出的预测结果是在某个值域（一般是实数域或其子集）上的任意值；而分类模型则是给出特定的某几个离散值之一。

上篇讲的线性回归模型，是用来做回归的。这次我们来讲一个做分类的模型：朴素贝叶斯分类器。

### 贝叶斯定理

在讲模型之前，我们先来看看概率统计中一个非常重要的定理：****贝叶斯定理****。

#### 贝叶斯公式

贝叶斯公式本身一目了然：

P(A|B)=P(B|A)P(A)P(B)P(A|B)=P(B|A)P(A)P(B)

用语言解释就是：在 B 出现的前提下 A 出现的概率，等于 A 和 B 都出现的概率除以 B 出现的概率。

换句话说就是后验概率和先验概率的关系。

#### 举例说明

##### ****一个简单的例子****

****例子1****：

我们假设：目前的全集是一个小学的小学一年级学生。

这个小学一年级一共100人，其中有男生30人。

穿白袜子的人数一共有20个，这20个人里面，有5个是男生。

那么请问，男生里面穿白袜子的人的出现概率为多少？

这不是废话嘛，一共30个男生，5个穿白袜子，出现概率是5/30=1/6啊。用得着贝叶斯公式吗？

如果我已经把人数都告诉你了，当然没必要算什么先后验概率。

但是我先不告诉你人数，我只告诉你：

(下面用 A 指代“穿白袜子”，B 指代“是男生”)

这个小学一年级学生里面，男生的出现概率是 0.3 —— P(B)；

穿白袜子的人的出现概率是0.2 —— P(A)；

穿白袜子的人是男生这件事出现的概率是0.25 —— P(B|A)。

请问你，一个人是男生又穿白袜子的出现概率 —— P(A|B)是多少？

这个时候就该****贝叶斯公式****出场啦：

P(A|B)=P(B|A)P(A)P(B)P(A|B)=P(B|A)P(A)P(B) ==> P(A|B) = 0.25 \* 0.2 / 0.3 = 1/6

##### ****另一个简单的例子****

如果你问我，明明人数都知道了，为什么还要绕个弯算概率？那么再来看另一个例子。

****例子2****：

把场景从一个小学的一年级转换为某个大饭店的门口，我们根据以往数据，可以计算出：

所有来吃饭的客人中，会有10%的人喝酒 —— P(B)，

所有客人中，会有20%的人驾车前来—— P(A)，

开车来的客人中，会有5%喝酒 —— P(B|A)。

那么请问，在这个饭店喝过酒的人里，仍然会开车的比例—— P(A|B)是多少？

P(A|B)=P(B|A)P(A)P(B)P(A|B)=P(B|A)P(A)P(B) ==> P(A|B) = 5% \* 20% / 10% = 10%

#### 一般化的贝叶斯公式

更一般化的情况，假设事件 A 本身又包含多种可能性，即 A 是一个集合：A={A1,A2,…,An}A={A1,A2,…,An}，那么对于集合中任意的 AiAi，贝叶斯定理可用下式表示：

P(Ai|B)=P(B|Ai)P(Ai)∑jP(B|Aj)P(Aj)P(Ai|B)=P(B|Ai)P(Ai)∑jP(B|Aj)P(Aj)

这和之前的简化版公式：

P(A|B)=P(B|A)P(A)P(B)P(A|B)=P(B|A)P(A)P(B)

在使用上有什么区别呢？

我们还是来看一个例子，****例子3****：

某 AI 公司招聘工程师，来了8名应聘者，这8个人里，有5个人是985院校毕业的，另外3人不是。

面试官拿出一道算法题准备考察他们。根据以前的面试经验，面试官知道：985毕业生做对这道题的概率是80%，非985毕业生做对率只有30%。

现在，面试管从8个人里随手指了一个人——小甲，让 TA 出来做题。结果小甲做对了，那么请问，小甲是985院校毕业的概率是多大？

现在我们来看，这道题里面的小甲的毕业院校有两种可能，也就是 A={A1,A2}A={A1,A2},

A1A1 —— 被选中的人是985毕业的；

A2A2 —— 被选中的人不是985毕业的。

B —— 被选中的人做对了面试题

P(A1)P(A1) = 5/8

P(A2)P(A2) = 3/8

P(B|A1)P(B|A1) = 80% = 0.8（985毕业生做对该道面试题的先验概率）

P(B|A2)P(B|A2) = 30% = 0.3（非985毕业生做对该道面试题的先验概率）

因此：

P(A1|B)=P(B|A1)P(A1)∑jP(B|Aj)P(Aj)P(A1|B)=P(B|A1)P(A1)∑jP(B|Aj)P(Aj) = (0.8 \* 5/8) / (0.8 \* 5/8 + 0.3 \* 3/8)) = 0.8163

所以，小甲是985毕业的概率是81.6%

****注意****：上面的几个例子中，先验、后验事件的概率都是离散的。

事实上，贝叶斯定理一样可以应用于连续概率的情况，假设上面的具体事件的概率不是一个确定值，而是一个分布函数，也是一样的。只不过 sum 部分变为了对应函数的积分而已。

连续概率的贝叶斯定理的形式为（下面所说的 A 和 B 对应之前贝叶斯公式中的 A 与 B）：

f(x|y)=f(y|x)f(x)∫∞−∞f(y|x)f(x)dxf(x|y)=f(y|x)f(x)∫−∞∞f(y|x)f(x)dx

其中，f(x|y) 是给定 B=y 时，A 的后验分布；对应的 f(y|x) 是给定 A=x 时，B 的后验分布； f(x) 则是 A 的先验分布概率函数。

为了方便起见，这里的 f 在这些专有名词中代表不同的函数。

### 朴素贝叶斯分类器（Naïve Bayes Classifier）

“****朴素贝叶斯****”（****Naïve Bayes****）既可以是一种****算法****——朴素贝叶斯算法，也可以是一种****模型****——朴素贝叶斯分类模型（分类器）。

#### 朴素贝叶斯算法

首先我们来讲作为****算法****的 Naïve Bayes，朴素贝叶斯算法可以直接利用贝叶斯定理来实现。

先来看简洁版的贝叶斯定理：

P(A|B)=P(B|A)P(A)P(B)P(A|B)=P(B|A)P(A)P(B)

在之前的几个例子中，为了便于理解，当 B 作为 A 的条件出现时，我们假定它总共只有一个特征。

但在实际应用中，很少有一件事只受一个特征影响的情况，往往影响一件事的因素有多个。假设，影响 B 的因素有 n 个，分别是 b1,b2,…,bnb1,b2,…,bn。

则 P(A|B)P(A|B) 可以写为：

P(A|b1,b2,…,bn)=P(A)P(b1,b2,…,bn|A)P(b1,b2,…,bn)P(A|b1,b2,…,bn)=P(A)P(b1,b2,…,bn|A)P(b1,b2,…,bn)

A 的先验概率 P(A) 和多个因素的联合概率 P(b1,b2,…,bn)P(b1,b2,…,bn) 都是可以单独计算的，与 AA 和 bibi 之间的关系无关，因此这两项都可以被看作常数。

对于求解 P(A|b1,b2,…,bn)P(A|b1,b2,…,bn)，最关键的是 P(b1,b2,…,bn|A)P(b1,b2,…,bn|A)。根据链式法则，可得：

P(b1,b2,…,bn|A)=P(b1|A)P(b2|A,b1)P(b3|A,b1,b2)...P(bn|A,b1,b2,...,bn−1)P(b1,b2,…,bn|A)=P(b1|A)P(b2|A,b1)P(b3|A,b1,b2)...P(bn|A,b1,b2,...,bn−1)

上面的求解过程，看起来好复杂，但是，如果从 b1b1 到 bnbn 这些特征之间，在概率分布上是条件独立的，也就是说每个特征 bibi 与其他特征都不相关。

那么，当 i≠≠j 时，有 P(bi|A,bj)=P(bi|A)P(bi|A,bj)=P(bi|A) —— 无关条件被排除到条件概率之外。因此，当 b1，b2，…,bnb1，b2，…,bn 中每个特征与其他 n-1 个特征都不相关时，就有：

P(A|b1,b2,…,bn)=1ZP(A)∏ni=1P(bi|A)P(A|b1,b2,…,bn)=1ZP(A)∏i=1nP(bi|A)

注意：此处的 Z 对应 P(b1,b2,…,bn)P(b1,b2,…,bn)。

#### 一款极简单的朴素贝叶斯分类器

上式中的 b1b1 到 bnbn 是特征（Feature），而 A 则是最终的类别（Class），所以，我们换一个写法：

P(C|F1,F2,…,Fn)=1ZP(C)∏ni=1P(Fi|C)P(C|F1,F2,…,Fn)=1ZP(C)∏i=1nP(Fi|C)

这个公式也就是我们的****朴素贝叶斯分类器的模型函数****！

它用来做预测时是这样的：

有一个朴素贝叶斯分类模型（器），它能够区分出 k 个类 （c1,c2,…,ck）（c1,c2,…,ck）, 用来分类的特征有 n 个：(F1,F2,…,Fn)(F1,F2,…,Fn)。

现在有个样本 s，我们要用 NB 分类器对它做预测，则需要先提取出这个样本的所有特征值 F1F1 到 FnFn，将其带入到下式中进行 k 次运算:

P(C=cj)∏ni=1P(Fi=fi|C=cj)P(C=cj)∏i=1nP(Fi=fi|C=cj)

然后比较这 k 次的结果，选出使得运算结果达到最大值的那个 cj（j=1,2,…,k）cj（j=1,2,…,k）—— 这个 cjcj 对应的类别就是预测值。

假设我们当前有一个模型，总共只有两个类别：c1c1 和 c2c2；有三个 Feature：F1、F2和F3F1、F2和F3。F1F1 有两种可能性取值：f11f11 和 f12f12；F2F2 有三种可能性取值：f21、f22、f23f21、f22、f23；F3F3 也有两种可能性取值：f31、f32f31、f32。

那么对于这个模型，我们要做的就是通过训练过程，获得下面这些值：

P(C=c1)P(C=c1) P(C=c2)P(C=c2)

P(F1=f11|C=c1)P(F1=f11|C=c1)

P(F1=f12|C=c1)P(F1=f12|C=c1)

P(F2=f21|C=c1)P(F2=f21|C=c1)

P(F2=f22|C=c1)P(F2=f22|C=c1)

P(F2=f23|C=c1)P(F2=f23|C=c1)

P(F3=f31|C=c1)P(F3=f31|C=c1)

P(F3=f32|C=c1)P(F3=f32|C=c1)

P(F1=f11|C=c2)P(F1=f11|C=c2)

P(F1=f12|C=c2)P(F1=f12|C=c2)

P(F2=f21|C=c2)P(F2=f21|C=c2)

P(F2=f22|C=c2)P(F2=f22|C=c2)

P(F2=f23|C=c2)P(F2=f23|C=c2)

P(F3=f31|C=c2)P(F3=f31|C=c2)

P(F3=f32|C=c2)P(F3=f32|C=c2)

把这些概率值都算出来以后，就可以用来做预测了。

比如我们有一个需要预测的样本 X，它的特征值分别是 f11、f22、f31f11、f22、f31，那么：

样本X被分为c1的概率是：P(C=c1|x)=P(C=c1|F1=f11,F2=f22,F3=f31)∝P(C=c1)P(F1=f11|C=c1)P(F2=f22|C=c1)P(F3=f31|C=c1);样本X被分为c1的概率是：P(C=c1|x)=P(C=c1|F1=f11,F2=f22,F3=f31)∝P(C=c1)P(F1=f11|C=c1)P(F2=f22|C=c1)P(F3=f31|C=c1);

样本X被分为c2的概率是：P(C=c2|x)=P(C=c2|F1=f11,F2=f22,F3=f31)∝P(C=c2)P(F1=f11|C=c2)P(F2=f22|C=c2)P(F3=f31|C=c2)样本X被分为c2的概率是：P(C=c2|x)=P(C=c2|F1=f11,F2=f22,F3=f31)∝P(C=c2)P(F1=f11|C=c2)P(F2=f22|C=c2)P(F3=f31|C=c2)

两者都算出来以后，只需要对比 P(C=c1|x)和P(C=c2|x)P(C=c1|x)和P(C=c2|x) 谁更大，那么这个样本的预测值就是对应类别。假设 P(C=c2|x)>P(C=c1|x)P(C=c2|x)>P(C=c1|x)，则 x 的预测值为 c2c2，这个样本被分类器分为了 c2c2。

上面那些先验概率和条件概率如何得到呢？通过在训练样本中间做统计，就可以直接获得了！

现在让我们把上面那个抽象的例子具象化，来看一个新的例子，****例子4****：

假设有一家小公司招收机器学习工程师，为了在更广泛的范围内筛选人才，他们写了一些爬虫，去各个招聘平台、职场社交平台爬取简历，然后又写了一个简单的分类器，来筛选他们感兴趣的候选人。

这个筛选分类器是朴素贝叶斯分类器，训练数据是现在公司里的机器学习工程师和之前来面试过这一职位，没有被录取的人员的简历。

全部数据集如下图所示：



对应符号表达含义：

c1：被录取；c2：不被录取；c1：被录取；c2：不被录取；

f11：是985毕业；f12：不是985毕业；f11：是985毕业；f12：不是985毕业；

f21：本科；f22：硕士；f23：博士；f21：本科；f22：硕士；f23：博士；

f31：技能为C++；f32：技能为Javaf31：技能为C++；f32：技能为Java

P(C=c1)=0.6P(C=c1)=0.6

P(C=c2)=0.4P(C=c2)=0.4

P(F1=f11|C=c1)=0.67P(F1=f11|C=c1)=0.67

P(F1=f12|C=c1)=0.33P(F1=f12|C=c1)=0.33

P(F2=f21|C=c1)=0.33P(F2=f21|C=c1)=0.33

P(F2=f22|C=c1)=0.33P(F2=f22|C=c1)=0.33

P(F2=f23|C=c1)=0.33P(F2=f23|C=c1)=0.33

P(F3=f31|C=c1)=0.17P(F3=f31|C=c1)=0.17

P(F3=f32|C=c1)=0.83P(F3=f32|C=c1)=0.83

P(F1=f11|C=c2)=0.25P(F1=f11|C=c2)=0.25

P(F1=f12|C=c2)=0.75P(F1=f12|C=c2)=0.75

P(F2=f21|C=c2)=0.5P(F2=f21|C=c2)=0.5

P(F2=f22|C=c2)=0.5P(F2=f22|C=c2)=0.5

P(F2=f23|C=c2)=0P(F2=f23|C=c2)=0

P(F3=f31|C=c2)=0.75P(F3=f31|C=c2)=0.75

P(F3=f32|C=c2)=0.25P(F3=f32|C=c2)=0.25

假设这时候一个样本x的特征值为f11，f22，f31（985毕业，硕士，掌握C++），那么:假设这时候一个样本x的特征值为f11，f22，f31（985毕业，硕士，掌握C++），那么:

P(C=c1|x)∝P(C=c1)P(F1=f11|C=c1)P(F2=f22|C=c1)P(F3=f31|C=c1)=0.6∗0.67∗0.33∗0.17=0.022;P(C=c1|x)∝P(C=c1)P(F1=f11|C=c1)P(F2=f22|C=c1)P(F3=f31|C=c1)=0.6∗0.67∗0.33∗0.17=0.022;

P(C=c2|x)∝P(C=c2)P(F1=f11|C=c2)P(F2=f22|C=c2)P(F3=f31|C=c2)=0.4∗0.25∗0.5∗0.75=0.038P(C=c2|x)∝P(C=c2)P(F1=f11|C=c2)P(F2=f22|C=c2)P(F3=f31|C=c2)=0.4∗0.25∗0.5∗0.75=0.038

P(C=c2|x)>P(C=c1|x)，所以预测值为c2：不能被录取。P(C=c2|x)>P(C=c1|x)，所以预测值为c2：不能被录取。

体现的****思路****是：在训练样本的基础上做一系列概率运算，然后用这些算出来的概率按朴素贝叶斯公式“拼装”成分类模型——这就成了朴素贝叶斯分类器。

#### 频率 VS 概率

****这也太简单了吧****。

朴素贝叶斯分类器这个模型的训练过程都不需要先从模型函数推导目标函数，再优化目标函数求 Cost 最小的解吗？朴素贝叶斯公式就是朴素贝叶斯分类器的训练算法啦？

上述例子之所以这样简单，是因为****我们简单地将频率当成了概率****。

但在现实应用中，这种方法往往不可行，因为这种方法实际上默认了“未被观测到”的就是“出现概率为0”的。这样做显然是不合理的。

比如：上面例子中，由于样本量太小，“博士”候选人只有两位，而且全部被录取，因此对于“未被录用”的情况而言，学历是博士的条件概率就变成了0。这种情况使得学历是否是博士成了唯一决定因素，显然不合理。

虽然我们可以靠做一些简单的变换——比如加一平滑法（就是把概率计算为：对应类别样本中该特征值出现次数 + 1 /对应类别样本总数）——来规避除以0的情况，但是这样做的“准确性”仍然非常不可靠。

我们有没有更加精准的方法呢？当然有！

[邀请好友一起学，获得 25% 返现奖励](https://gitbook.cn/m/mazi/columns/5bc6ac7442d7d32f50f19a98/topics/5bf281c5fd72950cafdccd8b)

[IMG_259](https://gitbook.cn/m/mazi/comp/column?columnId=5bc6ac7442d7d32f50f19a98%26tag=2#catalog)