22讲朴素贝叶斯：如何让计算机学会自动分类

你好，我是黄申。今天我们来聊聊朴素贝叶斯。

在开始正式的内容之前，我想问你一个问题，你是如何区分苹果、甜橙和西瓜的？你可能要说了，这个问题还用得着讲吗？是不是你们博士都喜欢将简单的问题复杂化？还真不是，如果你将计算机想象成一个两三岁的孩子，你会怎么教一个孩子区分这些水果呢？

比如我曾经就和一个小朋友有过这样一段对话：

小朋友：黄叔叔，你和我讲讲，什么样的水果才是苹果呀？

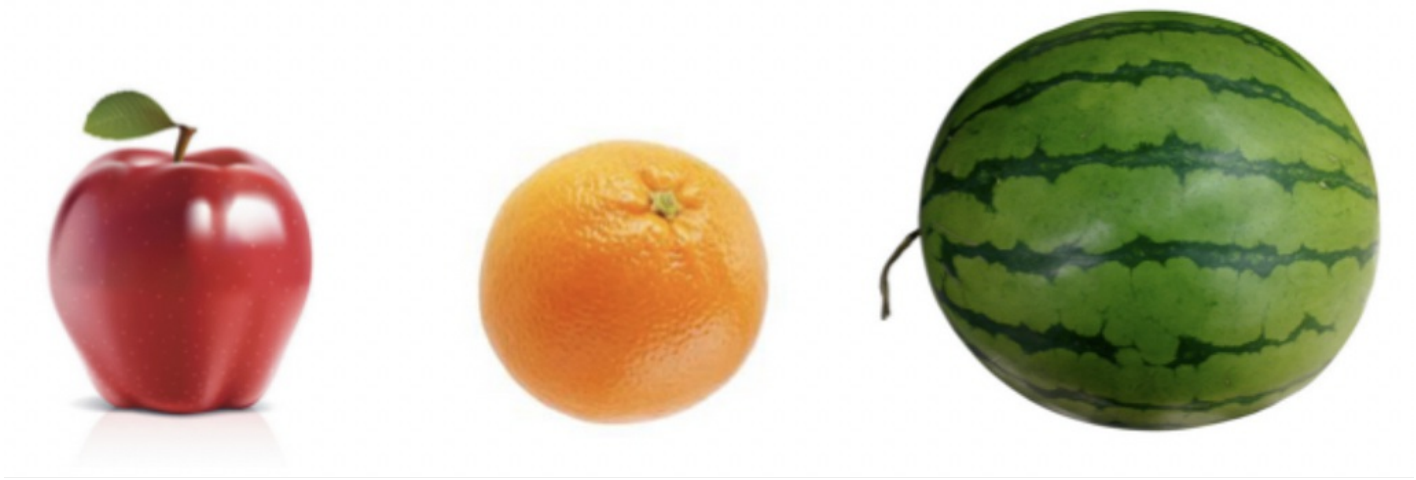
我：圆形的、绿色的水果。

小朋友：那西瓜也是圆形的、绿色的呀？

我：嗯……苹果也有可能是黄色或红色的，但西瓜不是。

小朋友：那甜橙也是圆形的、黄色的呀？

我：好吧，你看到的大部分情况下的甜橙都是黄色的，而苹果只有很少情况（少数品种）是黄色的。而且你还可以尝尝，它们的味道也是不同的。



哈哈，你是不是觉得想要描述清楚，并没有想象中的那么容易？但是，在这个对话中，有两点我觉得你需要关注一下：

- 我使用了“可能”“大部分情况”“很少情况”等等这种词语，这些词包含了**概率**的概念；
- 我使用了**多个条件**来判断一个水果属于哪个类别。

基于此，我接下来就要聊聊，我们是如何通过数学的思想和方法，系统性地解决这个问题的。这其中，**朴素贝叶斯**（Naive Bayesian）就提供了一个切实可行的方案。不过，在深入了解它之前，我们还需要做点准备工作。

如何将原始信息转化为计算机能看懂的数据?

事实上，计算机并不像两三岁的小孩那样，可以看到水果的颜色、形状和纹理，或者能尝到水果的味道。**我们需要将水果的特征转化为计算机所能理解的数据。最常用的方式就是提取现实世界中的对象之属性，并将这些转化为数字。**

以水果为例，你会提取它们的哪些属性呢？我会考虑这些，比如：形状、外皮颜色、斑马纹理、重量、握感、口感。我手边刚好有一个苹果、一个甜橙和一个西瓜，我把它们的属性分别统计了一下，你可以看看。

水果名称	形状	外观颜色	斑马纹理	重量	握感	口感
苹果	不规则圆	红色	无	200.45克	较硬	酸甜
甜橙	圆形	橙色	无	150.92克	较软	甜
西瓜	椭圆形	绿色	条纹	6000.88克	较硬	甜

然后，我们需要这些属性转化为计算机能够理解的东西——数字，也就是说，我给每种属性都定义了具体的数值，用来代表它们的具体属性。

水果名称	形状	外观颜色	外观纹理	重量	握感	口感
	不规则圆：1 圆形：2 椭圆形：3	红色：1 橙色：2 绿色：3	无条纹：1 有条纹：2	小于200g：1 200g和500g间：2 大于500g：3	较硬：1 较软：2	酸甜：1 甜：2
苹果	1	1	1	2	1	1
甜橙	2	2	1	1	2	2
西瓜	3	3	2	3	1	2

比较细心的话，你可能已经发现了，我偷偷地把重量由连续值转化成了离散值，这是因为朴素贝叶斯处理的都是离散值。

好了，仅仅3个水果还不足以构成朴素贝叶斯分类所需的训练样本。为了保证训练的质量，我们可以继续扩展到10个水果。

水果的名称	形状 不规则圆：1 圆形：2 椭圆形：3	外观颜色 红色：1 橙色：2 绿色：3	外观纹理 无：1 条纹：2	重量 小于200g：1 200g和500g间：2 大于500g：3	握感 较硬：1 较软：2	口感 酸甜：1 甜：2
苹果a	1	1	1	2	1	1
苹果b	1	1	1	1	1	1
苹果c	2	3	1	1	2	1
甜橙a	2	2	1	1	2	2
甜橙b	2	2	1	2	2	2
甜橙c	1	2	1	2	1	1
西瓜a	3	3	2	3	1	2
西瓜b	3	3	2	3	1	1
西瓜c	3	3	2	3	1	2
西瓜d	1	3	2	3	2	2

朴素贝叶斯的核心思想

我们现在已经拿到了这10个水果的数据，那如果现在我手上有一个新的水果，它也有一定的形状、颜色、口感等等，你怎么判断它是哪种水果呢?

之前的文章我们讲过先验概率、后验概率、条件概率和贝叶斯法则，它们是朴素贝叶斯分类的核心组成部分。通过贝叶斯法则，我们可以根据先验概率和条件概率，推导出后验概率。首先让我们快速回想一下贝叶斯公式。

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

上一节，我已经详细解释了这个公式的推导和每一部分的含义，这里再强调一下贝叶斯定理的核心思想：**用先验概率和条件概率估计后验概率**。

那具体到这里的分类问题，我们该如何运用这个公式呢？为了便于理解，我们可以将上述公式改写成这样：

$$P(c|f)=\frac{P(f|c)\times P(c)}{P(f)}$$

其中，c表示一个**分类**（class），f表示属性对应的**数据字段**（field）。如此一来，等号左边的P(clf)就是待分类样本中，出现属性值f时，样本属于类别c的概率。而等号右边的P(flc)是根据训练数据统计，得到分类c中出现属性f的概率。P©是分类c在训练数据中出现的概率，P(f)是属性f在训练样本中出现的概率。

看到这里，你可能要问了，这里的贝叶斯公式只描述了单个属性值属于某个分类的概率，可是我们要分析的水果每个都有很多属性啊，这该怎么办呢?

别急，朴素贝叶斯在这里就要发挥作用了。这是基于一个简单假设建立的一种贝叶斯方法，并假定数据对象的不同属性对其归类影响时是相互独立的。此时若数据对象o中同时出现属性fi与fj，则对象o属于类别c的概率就是这样：

$$\begin{aligned}P(c|o)&=P(c|f_i,f_j)=P(c|f_i)\times P(c|f_j)\\&=\frac{P(f_i|c)\times P(c)}{P(f_i)}\times\frac{P(f_j|c)\times P(c)}{P(f_j)}\end{aligned}$$

现在，我们应该已经可以用10个水果的数据，来建立朴素贝叶斯模型了。

其中，苹果的分类中共包含3个数据实例，对于形状而言，出现2次不规则圆、1次圆形和0次椭圆形，因此各自的统计概率为0.67、0.33和0.00。我们将这些值称为，给定一个水果分类时，出现某个属性值的**条件概率**。以此类推，所有的统计结果就是下面这个表格中这样：

水果 \ 特征	形状 不规则圆：1 圆形：2 椭圆形：3	外观颜色 红色：1 橙色：2 绿色：3	外观纹理 无：1 条纹：2	重量 小于200g：1 200g和500g间：2 大于500g：3	握感 较硬：1 较软：2	口感 酸甜：1 甜：2
苹果	1：0.67 2：0.33 3：0.00	1：0.67 2：0.00 3：0.33	1：1.00 2：0.00	1：0.67 2：0.33 3：0.00	1：0.67 2：0.33	1：1.00 2：0.00
甜橙	1：0.33 2：0.67 3：0.00	1：0.00 2：1.00 3：0.00	1：1.00 2：0.00	1：0.33 2：0.67 3：0.00	1：0.33 2：0.67	1：0.33 2：0.67
西瓜	1：0.25 2：0.00 3：0.75	1：0.00 2：0.00 3：1.00	1：0.00 2：1.00	1：0.00 2：0.00 3：1.00	1：0.75 2：0.25	1：0.25 2：0.75
总共	1：0.40 2：0.30 3：0.30	1：0.20 2：0.30 3：0.50	1：0.60 2：0.40	1：0.30 2：0.30 3：0.40	1：0.60 2：0.40	1：0.50 2：0.50

对于上表中出现的0.00概率，在做贝叶斯公式中的乘积计算时，会出现结果为0的情况，因此我们通常取一个比这个数据集里最小统计概率还要小的极小值，来代替“零概率”。比如，我们这里取0.01。在填充训练数据中从来没有出现过的属性值的时候，我们就会使用这种技巧，我们给这种技巧起个名字就叫作**平滑**（Smoothing）。

有了这些条件概率，以及各类水果和各个属性出现的先验概率，我们已经建立起了朴素贝叶斯模型。现在，我们就可以用它进行朴素贝叶斯分类了。

假设我们有一个新的水果，它的形状是圆形，口感是甜的，那么根据朴素贝叶斯，它属于苹果、甜橙和西瓜的概率分别是多少

呢？我们先来计算一下，它属于苹果的概率有多大。

$$
\begin{aligned}
P(apple \mid o) &= P(apple \mid shape-2, taste-2) \\
&= P(apple \mid shape-2) \times P(apple \mid taste-2) \\
&= \frac{P(shape-2 \mid apple) \times P(apple)}{P(shape-2)} \times \frac{P(taste-2 \mid apple) \times P(apple)}{P(taste-2)} \\
&= \frac{0.33 \times 0.30}{0.30} \times \frac{0.01 \times 0.30}{0.50} \\
&= 0.33 \times 0.006 \\
&= 0.00198
\end{aligned}
$$

其中，apple表示分类为苹果，shape-2表示形状属性的值为2（也就是圆形），taste-2表示口感属性的值为2。以此类推，我们还可计算该水果属于甜橙和西瓜的概率。

$$
\begin{aligned}
P(orange \mid o) &= \frac{P(shape-2 \mid orange) \times P(orange)}{P(shape-2)} \times \frac{P(taste-2 \mid orange) \times P(orange)}{P(taste-2)} \\
&= \frac{0.67 \times 0.30}{0.30} \times \frac{0.67 \times 0.30}{0.50} \\
&= 0.67 \times 0.402 \\
&= 0.26934
\end{aligned}
$$

$$
\begin{aligned}
P(watermelon \mid o) &= \frac{P(shape-2 \mid watermelon) \times P(watermelon)}{P(shape-2)} \\
&\quad \times \frac{P(taste-2 \mid watermelon) \times P(watermelon)}{P(taste-2)} \\
&= \frac{0.01 \times 0.40}{0.30} \times \frac{0.75 \times 0.40}{0.50} \\
&= 0.0133 \times 0.6 \\
&= 0.00798
\end{aligned}
$$

比较这三个数值，0.00198<0.00798<0.26934，所以计算机可以得出的结论，该水果属于甜橙的可能性是最大的，或者说，这个水果最有可能是甜橙。

你可能已经注意到了，这几个公式里的概率乘积通常都非常小，在物品的属性非常多的时候，这个乘积可能就小到计算机无法处理的地步。因此，在实际运用中，我们还会采用一些数学手法进行转换（比如取log将小数转换为绝对值大于1的负数），原理都是一样的。

内容比较多，我稍微总结一下。朴素贝叶斯分类主要包括这几个步骤：

- 准备数据：针对水果分类这个案例，我们收集了若干水果的实例，并从水果的常见属性入手，将其转化为计算机所能理解

的数据。这种数据也被称为**训练样本**。

- 建立模型：通过手头上水果的实例，我们让计算机统计每种水果、属性出现的先验概率，以及在某个水果分类下某种属性出现的条件概率。这个过程也被称为**基于样本的训练**。
- 分类新数据：对于一颗新水果的属性数据，计算机根据已经建立的模型进行推导计算，得到该水果属于每个分类的概率，实现了分类的目的。这个过程也被称为**预测**。

朴素贝叶斯分类VS其他分类算法

用朴素贝叶斯进行分类的内容差不多就是这样，你可能要问了，朴素贝叶斯是唯一的分类算法吗？现实中需要分类的场景那么多，朴素贝叶斯都适用吗？确实，我们有很多种分类算法，它们也都有各自的优劣。我这里就把朴素贝叶斯和常用的几种分类算法做个总结和比较。

- 和**KNN最近邻**相比，朴素贝叶斯需要更多的时间进行模型的训练，但是它在对新的数据进行分类预测的时候，通常效果更好、用时更短。
- 和**决策树**相比，朴素贝叶斯并不能提供一套易于人类理解的规则，但是它可以提供决策树通常无法支持的模糊分类（一个对象可以属于多个分类）。
- 和**SVM支持向量机**相比，朴素贝叶斯无法直接支持连续值的输入。所以，在前面的案例中，我将连续值转化成了离散值，便于朴素贝叶斯进行处理。

为了便于你理解记忆，我这里也做一下总结。

如果一个分类的应用场景中，待分类对象的属性值大部分都是离散的（或者很容易转化为离散的）、需要支持模糊分类，并且需要快速可靠的实时分类，那么这种场景通常就非常适合使用朴素贝叶斯方法。

总结

今天我从一个看似非常简单的判断水果的例子出发，介绍了如何通过物体的属性及其数值，让计算机理解现实世界中的事物，并通过朴素贝叶斯方法来对其进行分类。

在朴素贝叶斯方法的推导过程中，我给你讲了如何使用贝叶斯法则，将后验概率的估计转换为先验概率和条件概率。朴素贝叶斯训练过程包括基于样本数据的先验概率和条件概率统计，分类过程就包括了使用贝叶斯法则，结合新样本的属性数据以及训练好的模型数据，进行最终的预测。

最后，我将朴素贝叶斯和其他常见分类算法，比如KNN近邻、决策树、SVM向量机，做了对比。朴素贝叶斯适用离散属性值中，训练过程耗时长但是分类预测速度快，支持模糊分类。这一节的内容比较偏理论，下一节，我会着重来讲朴素贝叶斯的应用场景，告诉你哪些场合下更适合使用朴素贝叶斯法。

思考题

除了文本分类，你还知道什么地方可以使用朴素贝叶斯方法来处理分类问题？

欢迎留言和我分享，也欢迎你在留言区写下今天的学习笔记。你可以点击“请朋友读”，把今天的内容分享给你的好友，和他一起精进。

精选留言

temool

越看到后面越吃力，前面的也要再重新捋一遍

2019-02-11 10:04

作者回复

反复阅读和练习，就能加深印象，加油！

2019-02-12 01:27

Joe

朴素贝叶斯，朴素的原因是假设各个特征是相互独立的。

2019-02-11 08:51

时间被海绵吃了

老师好，已知形状是圆形，求是苹果的概率这个问题，请问一下用贝叶斯公式算和我直接数出样本数据里形状是圆形的个数，然后数圆形里面是苹果的个数，用这两个数相除算出来的结果好像是一样的，那它们有什么区别呢？谢谢！

2019-02-19 09:16

作者回复

对于单个特征确实如此，如果要考虑多个特征，比如形状、口味、纹理等等就会不同了

2019-02-20 01:46

胡

加油继续

2019-02-17 13:05

胡

真的是要反复理解，

2019-02-17 13:05

唯她命

朴素贝叶斯 必须各个特征相互独立的吗?

2019-02-16 15:58

作者回复

是的，朴素贝叶斯的“朴素”或者说naive就是指这个“天真”的假设。当然，很多时候并不成立，我们可以使用多阶马尔科夫模型来稍作修改，后面几节会有介绍。