0、概述

朴素贝叶斯是一种多类分类算法。它是一种概率模型，并且是一种生成模型。朴素贝叶斯中的“朴素”一词来源于它假设了样本各维特征之间具有独立性（即条件独立性假设）。

概率模型与非概率模型的区别是看待样本数据的角度不同，或者说对数据建模的思路不同。

非概率模型认为对于样本数据，存在一个未知的模型（God模型），God模型对给定的不同输入x，就有不同的输出y。样本数据就是对God模型的输入和输出的多次采样的结果，但由于这些采样的结果跟God模型的真实的输入输出取值相比有误差，并且这些数据只是God模型所有可能的输入输出数据的一部分，因此我们不可能丝毫不差的求解出God模型，只能按照某些标准，求解“最优”的模型，该模型对给定的输入x，能够给出最接近God模型输出的输出。

概率模型则基于概率和统计的理论，将样本数据看成是随机向量X和随机变量Y的联合采样的观察值，模型根据观察值估计概率分布P(X,Y)或P(Y|X)。

非概率模型认为，在给定模型下，输入x和输出y具有因果关系，确定的输入有确定的输出；而概率模型认为对于给定的x，对应的y可以取不同的值，只不过y取不同值的可能性不一样。概率模型根据X和Y的概率关系给出对于给定的X的取值x，对应的Y取不同y值的可能性大小，并取可能性最大者为最优输出。

我认为，概率模型是更理性的看待问题的方式，因为现实中很多问题并不是非黑即白的，不能一刀切的看待。

朴素贝叶斯的基本思路是：

（1）在样本集上统计出样本取各类别的概率；

（2）在样本集上统计出样本在取各类别条件下，样本的输入的各维特征取不同值（标量）的概率，并根据条件独立性假设，计算出样本在取各类别的条件下，样本输入取不同值（向量）的概率；

（3）计算输入和输出的联合分布；

（4）根据贝叶斯定理，计算给定输入下，输出为各类别的概率。

注意，条件独立性假设与抽样理论中的样本独立性不同，前者是把样本的各维特征看成随机变量，并假设它们之间条件独立；后者是把训练集看成是多个独立同分布的随机变量（样本）的观察值，这些样本之间是互相独立的。

“条件独立性假设”是朴素贝叶斯如此简单的原因，也是它的一个缺点。由于“条件独立性假设”是一个非常强的假设，所以实际中模型的效果不一定很好。因为数据的各维特征不一定很好地满足条件独立性假设。我们在应用朴素贝叶斯模型时，最好提前对数据进行处理，使之尽量满足条件独立性假设。

最基本的朴素贝叶斯处理的是样本特征为离散的情况，对于特征是连续型的样本，我们使用高斯朴素贝叶斯作为一种扩展方法。

一、朴素贝叶斯模型的训练过程和预测过程

**1、训练过程：**

给定训练集，每个训练样本有n维特征，输出为1,2,…,K（K个类别）。训练过程就是在训练集上统计出概率P(Y)和P(X|Y)：

|  |  |
| --- | --- |
|  | (1) |

其中I是指示函数，的取值是：若，则取1，否则取0。

上式统计了Y取第k类的概率，它相当于所有样本中类别为k的样本的比例。

设样本的第j维特征可能取值的个数为，可能取值的集合为：，则有：

|  |  |
| --- | --- |
|  | (2) |

上式统计了在Y为第k类的条件下，X的第j维特征取值为的概率，它相当于在所有类别为第k类的样本中，第j维特征取值为的样本所占的比例。

**2、预测过程：**

给定样本的x，如何用训练好的朴素贝叶斯模型预测其呢？

根据条件独立性假设，有：

|  |  |
| --- | --- |
|  | (3) |

其中，是x的第j个分量。上式给出了当Y取k时，X取值为x的概率。

根据贝叶斯定理，有：

|  |  |
| --- | --- |
|  | (4) |

据此我们得出了朴素贝叶斯模型的预测过程：对任意给定的样本x，利用（4）式计算每个输出k的概率，然后看哪个概率大就把样本分到哪一类。这跟极大似然估计的思想是一样一样的。实际中，由于我们只需要比较各输出k的概率的大小，所以不必计算(4)式中的分母，只需计算其分子即可。

可以看出，朴素贝叶斯模型没有“明显”的参数。如果非要定义朴素贝叶斯模型的参数，那么可以这样理解：从训练集中统计出来的所有的和都是其模型参数。

以上就是朴素贝叶斯模型的训练过程和预测过程。根据上述流程，你就能在实际中应用这个模型解决分类问题了。但是你可能对整个流程有一些疑问：

**问题：**

（1）首先，既然我们的目标是求出P(Y|X)，那为什么不直接从训练集中统计P(Y|X)，而是要费劲拔力的统计P(X|Y)和P(Y)，得到P(X,Y)，再根据贝叶斯定理，拐弯抹角地计算P(Y|X)？

答：理论上，我们当然可以给出统计P(Y|X)的方法：

|  |  |
| --- | --- |
|  | (5) |

使用这个式子似乎可以直接统计出给定样本的x时，其类别为第k类的概率。但实际中利用它来计算概率P()是不现实的，因为很多情况下，训练样本没有那么多，如果某个取值x在训练集中出现的次数很少，甚至都没有出现，那么这样统计出的概率误差非常大，这就没意义了。

（2）好吧，我知道不能从训练集中直接统计P(Y|X)了，但是为什么不能直接统计联合分布P(X,Y)，而是要费劲拔力的统计两个概率P(X|Y)和P(Y)，再把它们乘起来？

答：我们仍然可以给出在训练集上直接统计P(X,Y)的公式：

|  |  |
| --- | --- |
|  | (6) |

但是还是出于同样的原因，某个样本的x可能在训练集上出现的次数很少，尤其是当x没有出现在训练集中时，统计出来的概率就是0。这是由于数据集的不充分导致的。当然，即使先统计两个概率P(X|Y)和P(Y)，再把它们乘起来，仍然可能会出现概率=0的情况，但是针对后面的计算方法，我们可以通过拉普拉斯平滑来解决概率=0的问题。（实际上，平滑是针对各维特征的概率的计算分别进行的，具体的平滑方法后面会介绍）。

（3）好吧，我已经明白了，必须要先统计P(X|Y)和P(Y)，然后作平滑，然后间接地计算P(X,Y)，但是在统计P(X|Y)时，为什么不直接统计它，而是费劲拔力的先统计P(Xj|Y)，然后再根据条件独立性假设，拐弯抹角地计算P(X|Y)？

答：其实这三个问题的答案都是一样的，我们这么费劲拔力、拐弯抹角地计算，就是为了对付因数据稀疏性导致的可能出现概率为0的情况。对于问题（3），实际中经常会出现这样的情况：一个样本的x没有出现在训练集中，但是其实它只是其中一个或几个特征的取值在训练集中没有出现，而大部分特征的取值其实在训练集中都出现了。对此，我们自然地想到将样本的各维特征分别考虑，对于没有出现的特征，对其概率的计算进行平滑，然后根据条件独立性假设，将各维特征的概率乘起来，得到该样本的概率P(X|Y)。

二、朴素贝叶斯模型的特殊性

朴素贝叶斯模型与常规模型有一定的区别。由于其参数的特殊性，其不具有中规中矩的目标函数，也不使用梯度下降法等优化方法。

具体来说，朴素贝叶斯模型与常规模型相比有哪些特殊性呢？我们从其训练过程和预测过程两方面进行理解：

（1）训练过程：

在朴素贝叶斯模型的训练过程中，它在训练集上学习，并学出了参数。我们知道，常规模型的训练过程实际就是求解目标函数的最优化问题的过程，一般来讲，它是个“反演”过程；而朴素贝叶斯模型求解参数，是一个“正演”过程——它直接就把参数计算出来了。

（2）预测过程：

根据朴素贝叶斯模型的预测过程，我们的目的是对给定的x，求使P(Y=k|X=x)最大的k，即：

|  |  |
| --- | --- |
|  | (7) |

我们可以看出：

第一、我们求的是k，但k并不是模型的参数，模型的参数是和，它们在模型的训练阶段就求出来了；而k的求取是在模型做预测的阶段。因此式中的函数不能理解成是目标函数。

第二、我们求最优的k，使用的是最大似然估计的思想，这是发生在模型预测阶段，而对常规模型来说，最大似然估计是用来导出模型的目标函数的，是属于理论推导阶段，甚至在模型训练阶段之前。

通过上面论述可以看出，朴素贝叶斯模型与常规机器学习模型相比，确实是一个比较特殊的模型。

三、朴素贝叶斯模型的改进

**1、拉普拉斯平滑**

在上面我们已经提到，朴素贝叶斯在计算概率和时可能出现概率=0的情况，因此我们对它们进行拉普拉斯平滑：

|  |  |
| --- | --- |
|  | (8) |

以及：

|  |  |
| --- | --- |
|  | (9) |

平滑过后的概率就不会成为0值了。注意式中分母上加的数并不是随便给的，它是为了保证概率的归一化，即保证所有可能的事情发生的概率加起来等于1。

**2、高斯朴素贝叶斯**

高斯朴素贝叶斯方法是为了处理样本特征为连续型特征的情况。其思路很简单：基本的朴素贝叶斯方法通过在训练集上直接统计的方式计算概率。而当特征为连续型特征时，改为这样计算：假设特征服从高斯分布，通过训练集样本计算出该分布的均值和方差，这就相当于拟合出了特征的分布的具体形式，然后根据这具体的形式就可以计算特征取任意值的概率。