朴素贝叶斯算法

朴素贝叶斯算法（Naive Bayesian algorithm）是在贝叶斯算法的基础上，假设特征变量相互独立的一种分类方法，是贝叶斯算法的简化，常用于文档分类和垃圾邮件过滤。当“特征变量相互独立”的假设条件能够被有效满足时，朴素贝叶斯算法具有算法比较简单、分类效率稳定、所需估计参数少、对缺失数据不敏感等种种优势。而在实务中“特征变量相互独立”的假设条件往往不能得到满足，在一定程度上降低了贝叶斯分类算法的分类效果，但这并不意味着朴素贝叶斯算法在实务中难以推广，反而是堪与经典的“决策树算法”比肩的应用最为广泛的分类算法之一。

目录

CONTENTS

PART 01

朴素贝叶斯算法的基本原理

贝叶斯方法的基本原理

朴素贝叶斯算法来自贝叶斯方法，贝叶斯方法与传统经典的统计方法有所区别，体现在对随机分布参数进行参数估计时，传统经典统计方法认为参数（如均值、方差）是固定的，或者说待估计参数是未知的常数，但数据是随机的，或者说用于估计参数的数据仅仅是总体的一个随机抽取样本，所以在估计参数时也会对既有样本进行“毫无偏见”的处理，而且因为总体不可观测，所以依据样本得到的参数估计量大概率会存在估计误差，传统经典统计理论中用置信区间表示这些误差的大小。在对概率的理解上，经典统计认为概率就是频率的稳定值。一旦离开了重复试验，就谈不上去理解概率。而贝叶斯方法则恰好相反，认为数据是固定的，但是待估计参数是随机的而不是常数，存在概率分布，所以概率也是一种人们的主观概率，会随着更多样本示例的实际响应情况不断进行更新。

贝叶斯方法的基本原理

举一个授信客户是否违约的例子，假设一家商业银行授信客户的历史违约概率为2%。如果该银行新增了100个授信客户，到期后都按时结清，没有违约，那么按照传统经典统计理论，100次未违约和0次违约的结果，就会得出该银行新增客户的违约概率为0%，未违约的概率为100%；而按照贝叶斯理论，授信客户有先验违约概率2%，随着每个客户还款结果的逐渐明朗，新增客户的违约概率也在不断更新，从一开始的2%逐渐下降，随着越来越多未违约客户的出现，银行人员对于新增客户的还款信心也会不断增加，违约概率会不断下降，会不断接近于0，但永远不会明确的等于0。所以，贝叶斯理论考虑先验概率，对概率的理解是，人们对事件的信任程度，或者说是对事物不确定性的一种主观判断，与个人因素等有关，这也是前述称为主观概率的原因之所在。

贝叶斯方法的基本原理

由此可以看出，传统经典统计理论基于样本估计参数，往往需要有大量的数据样本作为支持，统计抽样时所要求的样本独立同分布的条件也很难满足，或者一言以蔽之，需要使得样本能够充分代表总体，但在实务中这一要求往往难以满足，比如前面提到的授信客户的例子，按照新增客户0违约的频率表现，会得出授信客户违约概率为0的判断，这一判断显然会存在较大偏差，不符合商业银行的经营实践。

贝叶斯方法的基本原理

贝叶斯理论的优势在于能够充分利用现有信息，将统计推断建立在后验分布的基础上。相对于传统经典统计理论，贝叶斯除了利用样本信息之外，还充分运用了先验概率（上例中的历史违约概率）信息，利用了参数的历史资料或先验知识，而模型中的参数估计值是建立在后验分布基础上，后验分布由于共同综合了先验概率和样本信息的知识，即避免了只使用先验概率的主观偏见，也避免了单独使用样本信息的过拟合现象，可以对参数作出较先验分布更合理的估计。所以，如果样本示例全集容量比较小，或者说样本不足以充分代表总体，而是需要充分利用除样本之外的信息时，贝叶斯估计具有经典传统理论无可比拟的优势，不但可以减少因样本量小而带来的统计误差，而且在没有数据样本的情况下也可以进行推断，在对研究除观测数据外还具备较多信息的情况特别有效。

当然，如果样本示例全集的容量足够大，或者说样本能够充分代表总体，那么贝叶斯方法与经典传统统计方法的估计都是可靠的，估计出的参数也会是一致的。

贝叶斯定理

贝叶斯方法依据贝叶斯定理。关于贝叶斯定理解释如下：首先我们设定在事件 B 条件下，发生事件 A 的条件概率，即 P(A|B)，从数学公式上，此条件概率等于事件 A 与事件 B 同时发生的概率除以事件 B 发生的概率。

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

上述公式可以进行变换，得到事件 A 与事件 B 同时发生的概率，这一概率既等于“事件 B 发生的概率”乘以“事件 B 条件下，发生事件 A 的条件概率”，也等于“事件 A 发生的概率”乘以“事件 A 条件下，发生事件 B 的条件概率”，或者说，A 与 B 的角色可以互换。

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

也就是说：

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

这一公式即为贝叶斯定理。单纯从数学推导上看，相对并不复杂，或者说只是把常识用数学公式表达了出来。下面我们结合上一节中提到的先验概率、后验概率等概念，赋予公式的各个组成部分以具体含义：

P(A)：先验概率　　P(B)：证据

P(B|A)：条件概率　　P(A|B)：后验概率

即有：

$$\text{后验概率} = \frac{\text{条件概率} \times \text{先验概率}}{\text{证据}}$$

贝叶斯定理

下面以一个员工异常行为管理的案例说明其神奇。假设一家商业银行基于历史数据统计（案件、监管处罚、内外部审计、诚信举报、离职核查等各种渠道）发现其员工异常行为发生率为0.005，其搭建的“非现场监测模型系统+人工复核”员工行为管理体系的检查准确率为0.99。

P(A)：先验概率，员工异常行为发生率为 0.005；

1-P(A)：员工异常行为未发生率等于 0.995；

P(B|A)：条件概率，员工存在异常行为且被检查发现的概率为 0.99；

P(B)：证据，通过全概率公式计算得到

$$P(B) = P(A) \times P(B|A) + [1 - P(A)] \times P[B|(1 - P(A)]$$

$$= 0.005 * 0.99 + 0.995 * 0.01 = 0.00495 + 0.0095 = 0.0149$$

后验概率：$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} = \frac{0.99 \times 0.005}{0.0149} = 0.332215$

也就是说，虽然该银行员工行为管理体系的检查准确率高达 0.99，但令人遗憾的事实却是，如果某员工被该体系判定存在员工异常行为，但是其确实存在异常行为的概率只有不到三分之一（0.332215），被误判的可能性超过了三分之二。

贝叶斯定理

但这并不意味着员工异常行为管理体系的彻底失效，如果让该员工再次接受体系检查，那么上次的后验概率就成为了新的检查的先验概率，即用 0.332215 代替了 0.005，如果员工仍然被该体系判定存在员工异常行为，那么后验概率将变成：

$$P(A|B)=\frac{P(B|A)\times P(A)}{P(B)}=\frac{0.99\times 0.332215}{0.332215*0.99+0.667785*0.01}=0.980014942$$

也就是说，该员工被该体系前后两次判定存在员工异常行为，并且其确实存在异常行为的概率达到了 98%以上，被误判的可能性已经很小了。按照同样的逻辑，如果该员工被该体系前后三次或更多判定为存在员工异常行为，那么其被误判的可能性会继续下降，逐渐接近于 0。

这一原理也提示我们，在进行员工异常行为排查时，一是在界定员工异常行为时，为最大程度保护奋斗者干事创业的热情，不应该以一次发现而下定论，因为被“误判”的可能性较大，即使相应的监测模型已经非常成熟和完善（例子中达到了 99%以上）；二是应该高度重视前后多次排查存在异常行为的员工，这部分员工被“误判”的可能性较低，应该及时采取果断措施，防止引发案件风险。

朴素贝叶斯算法的基本原理

朴素贝叶斯方法是在贝叶斯算法的基础上进行了相应的简化，即假定给定目标值时特征变量之间相互条件独立。前面我们讲解了贝叶斯定理，我们把 A 作为响应变量，把 B 作为特征变量，即有

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

进一步的，我们假设 A 可以分为 m 类，B 拓展成为一个特征变量集，共有 d 个特征变量，即有

$$A = \{a_1, a_2, \ldots\ldots a_m\}$$

$$B = \{b_1, b_2, \ldots\ldots b_d\}$$

朴素贝叶斯方法假定给定目标值时特征变量之间相互条件独立

$P(B|A) = \prod_{i=1}^{d} P(b_i|A)$，即有后验概率为：

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} = \frac{\prod_{i=1}^{d} P(b_i|A) \times P(A)}{P(B)}$$

基于上式，某样本示例属于类别a_i的朴素贝叶斯计算公式为：

$$P(a_i|b_1, b_2, \ldots\ldots b_d) = \frac{\prod_{i=1}^{d} P(b_i|a_i) \times P(a_i)}{\prod_{j=1}^{d} P(b_j)}$$

朴素贝叶斯算法的基本原理

当“特征变量相互独立”的假设条件能够被有效满足时，朴素贝叶斯算法具有算法比较简单、分类效率稳定、所需估计参数少、对缺失数据不敏感等种种优势。而在实务中“特征变量相互独立”的假设条件往往不能得到满足，在一定程度上降低了贝叶斯分类算法的分类效果。但这并不意味着朴素贝叶斯算法在实务中难以推广，反而是堪与经典的“决策树算法”比肩的应用最为广泛的分类算法之一，这是因为：

（1）实务中机器学习中使用的特征变量可能非常多，比如在商业银行授信业务客户违约分类中可能需要用到上百个特征变量，或者说数据集的维数会非常多，这些特征变量之间的协方差矩阵难以估计，或者说一般的贝叶斯方法是难以落地实施的。

（2）虽然多个特征变量之间不太可能做到完全独立，但是在很多场景下，特征变量之间的相关性可能会起到抵消的效果，从而虽然未满足独立条件，但仍然能够得到较为准确的预测。

（3）前面在第三章中我们提到，机器学习领域关注的更多的是模型的预测效果，也就是说拟合值与实际值的差距，而对于假设条件是否能够得到满足、具体特征之间的相关性等并未过多考虑。或者说，“预测效果好坏”才是判定模型优劣的根本标准。

拉普拉斯修正

不难发现，朴素贝叶斯计算公式中的分子为条件概率和先验概率的乘积，如果其中一项取值为 0，那么整个公式的值也将取值为 0。事实上，我们在进行机器学习时，如果把样本示例全集随机划分为训练样本和测试样本，那么在训练样本中很可能会出现某一类别没有样本进入的情形，但这种情形只是因为相应类型的样本没有在训练样本中被观测到，而不是其取值概率确实为 0，从而造成分类结果的失真。

拉普拉斯修正，又叫拉普拉斯平滑，是为了解决上述零概率问题而引入的处理方法。如果样本示例全集为 D，训练集中的样本类别数为 N，N_i 表示训练集样本第 i 个属性上的取值个数，拉普拉斯修正修正原理为：

原来的先验概率 $P(A)=\frac{|D_y|}{|D|}$ 拉普拉斯修正为：$P(A)=\frac{|D_y|+1}{|D|+N}$

原来的条件概率 $P(B|A)=\frac{|D_{y,x}|}{|D_c|}$ 拉普拉斯修正为：$P(B|A)=\frac{|D_{y,x}|+1}{|D_c|+N_i}$

也就是说，通过拉普拉斯修正，即使响应变量的某类别没有出现，先验概率或条件概率也不会等于 0，从而解决了前述分类结果失真问题。当然，拉普拉斯修正中分子“+1”也可以根据实际情况设置成“+c”，c 为一个比较小的整数了，也就是：

$$P(A)=\frac{|D_y|+c}{|D|+N\times c}$$

$$P(B|A)=\frac{|D_{y,x}|+c}{|D_c|+N_i\times c}$$

朴素贝叶斯算法分类及适用条件

与其他机器学习算法相比，朴素贝叶斯算法所需要的样本量比较少，当然样本量肯定是多多益善，如果样本量少于特征变量数目时，估计效果也会被削弱。对比支持向量机、随机森林等算法，朴素贝叶斯算法往往估计效果偏弱，但胜在运行速度更快。Python的sklearn模块有四种朴素贝叶斯算法，包括高斯朴素贝叶斯、多项式朴素贝叶斯、补集朴素贝叶斯、二项式朴素贝叶斯。

朴素贝叶斯算法分类及适用条件

1、高斯朴素贝叶斯（Gaussian naive Bayes）：该算法假设每个特征变量的数据都服从服从高斯分布（也就是正态分布），来估计每个特征下每个类别上的条件概率。高斯朴素贝叶斯的决策边界是曲线，可以是环形也可以是弧线。高斯朴素贝叶斯擅长处理连续型特征变量。相对于前面介绍的logistic回归，我们算法的目的是为了获得对概率的预测，并且希望越准确越好，那么应该首选logistic算法。而如果数据十分复杂，或者满足稀疏矩阵的条件（在矩阵中，若数值为0的元素数目远远多于非0元素的数目，并且非0元素分布没有规律时，则称该矩阵为稀疏矩阵），那么朴素贝叶斯算法就更占优势。

朴素贝叶斯算法分类及适用条件

2、多项式朴素贝叶斯（multinomial naive Bayes），通常被用于文本分类，该算法假设所有特征变量是离散型特征变量，所有特征变量都符合多项式分布。多项式分布来源于统计学中的多项式实验，多项式实验的概念是，在n次重复试验中每项试验都有不同的可能结果，但在任何给定的试验中，特定结果发生的概率是不变的。多项式朴素贝叶斯算法擅长处理分类型特征变量，但受到样本不均衡问题（分类任务中不同类别的训练样例数目差别很大的情况，一般地,样本类别比例(Imbalance Ratio)(多数类vs少数类)明显大于1:1(如5:1)就可以归为样本不均衡的问题）影响较为严重。

朴素贝叶斯算法分类及适用条件

3、补集朴素贝叶斯（Complement Naive Bayes），该算法是前述多项式朴素贝叶斯算法的改进，不仅能够解决样本不均衡问题，还在一定程度上放松了“所有特征变量之间条件独立的朴素假设”。补集朴素贝叶斯在召回率方面表现较为出色，如果算法的目的是为了找到少数类（存在异常行为的员工、存在洗钱行为等），则补集朴素贝叶斯算法是一种不错的选择。

朴素贝叶斯算法分类及适用条件

4、二项式朴素贝叶斯（Bernoulli Naive Bayes），也称伯努利朴素贝叶斯，该算法假设所有特征变量是离散型特征变量，所有特征变量都符合伯努利分布（二项分布，取值为两个，注意并不必然为0、1取值，也可为1、2取值等）。二项式朴素贝叶斯算法要求将特征变量取值转换为二分类特征向量，如果某特征变量本身不是二分类的，那么就可以使用类中专门用来二值化的参数binarize来转换数据，使其符合算法。

PART 02

数据准备

数据准备

本节我们以“数据8.1”和“数据8.2”为例进行讲解。“数据8.1”与上一章中的“数据7.1”类似，同样记录的是某商业银行在山东地区的部分支行经营数据（虚拟数据，不涉及商业秘密），变量包括这些商业银行全部支行的转型情况（V1）、存款规模（V2）、EVA（V3）、中间业务收入（V4）、员工人数（V5）。但与“数据7.1”不同的是，转型情况（V1）分为两个类别：“0”表示“未转型网点”；“1”表示“已转型网点”。

针对“数据8.1”的朴素贝叶斯模型，我们以转型情况（V1）为响应变量，以存款规模（V2）、EVA（V3）、中间业务收入（V4）、员工人数（V5）为特征变量。不难发现，各个特征变量均为连续型变量，所以我们使用高斯朴素贝叶斯算法进行拟合。

数据准备

“数据8.2”的案例数据是来自XX在线小额贷款金融公司（虚拟名，如有雷同纯属巧合）2417个的存量客户的信息数据，具体包括客户的“信用情况（V1）”、“年龄（V2）”、“贷款收入比（V3）”、“名下贷款笔数（V4）”、“教育水平（V5）”、“是否为他人提供担保（V6）”等。由于客户信息数据涉及客户隐私和消费者权益保护，也涉及商业机密，所以在本章介绍时进行了适当的脱密处理，对于其中的部分数据也进行了必要的调整。

数据准备

针对“数据8.2”的朴素贝叶斯模型，我们以“信用情况（V1）”为响应变量，其中分类为“0”表示“未违约客户”，分类为“1”表示“违约客户”；“年龄（V2）”、“贷款收入比（V3）”、“名下贷款笔数（V4）”、“教育水平（V5）”、“是否为他人提供担保（V6）”为特征变量，其中“年龄（V2）”为连续变量，其他变量为分类变量，“贷款收入比（V3）”取值为“1”“2”“3”分别表示“40%及以下”“40%~70%”“70%及以上”；“名下贷款笔数（V4）”取值为“1”“2”分别表示“3笔及以下”“4笔及以上”；“教育水平（V5）”取值为“1”“2”分别表示“大学专科及以下”“大学本科及以上”；“是否为他人提供担保（V6）”取值为“1”“2”分别表示“有对外担保”“无对外担保”。

数据准备

不难发现，数据集中大部分特征变量为分类变量，其中“贷款收入比（V3）”有3类取值，而“名下贷款笔数（V4）”、“教育水平（V5）”、“是否为他人提供担保（V6）”三个变量均为二项分布，所以我们使用多项式朴素贝叶斯算法、补集朴素贝叶斯算法、二项式朴素贝叶斯算法进行拟合。

载入分析所需要的模块和函数

在进行分析之前，我们首先载入分析所需要的模块和函数，读取数据集并进行观察。

示例

参阅教材内容

PART 03

高斯朴素贝叶斯算法示例

数据读取及观察

大家首先需要将本书提供的数据文件放入安装python的默认路径位置，并从相应位置进行读取。

示例

参阅教材内容

将样本示例全集分割为训练样本和测试样本

示例

参阅教材内容

高斯朴素贝叶斯算法拟合

示例

参阅教材内容

绘制ROC曲线

示例

参阅教材内容

运用两个特征变量绘制高斯朴素贝叶斯决策边界图

示例

参阅教材内容

PART 04

多项式、补集、二项式朴素贝叶斯算法示例

数据读取及观察

大家首先需要将本书提供的数据文件放入安装python的默认路径位置，并从相应位置进行读取。

示例

参阅教材内容

将样本示例全集分割为训练样本和测试样本

示例

参阅教材内容

多项式、补集、二项式朴素贝叶斯算法拟合

示例

参阅教材内容

寻求二项式朴素贝叶斯算法拟合的最优参数

示例

参阅教材内容

最优二项式朴素贝叶斯算法模型性能评价

示例

参阅教材内容

PART 05

习　题

习题

使用“数据5.1”数据文件，把响应变量设定为“V1征信违约记录”，将其他变量作为特征变量，具体包括“V2资产负债率”、“V3行业分类”、“V4实际控制人从业年限”、“V5企业经营年限”、“V6主营业务收入”、“V7利息保障倍数”、“V8银行负债”、“V9其他渠道负债”，构建朴素贝叶斯算法模型。

1、载入分析所需要的库和模块

2、数据读取及观察。

3、将样本示例全集分割为训练样本和测试样本

4、构建高斯朴素贝叶斯算法模型。

（1）高斯朴素贝叶斯算法拟合

（2）绘制ROC曲线

（3） 只运用“V7利息保障倍数”“V8银行负债”两个特征变量开展高斯朴素贝叶斯算法，并绘制高斯朴素贝叶斯决策边界图。

习题

5、构建多项式朴素贝叶斯算法模型。

6、构建补集朴素贝叶斯算法模型。

7、构建二项式朴素贝叶斯算法模型。

感谢聆听

THANKS