

简述朴素贝叶斯算法原理

- 工作原理
 - 假设现在有样本 x 待分类
 - 假设样本有 m 个特征 ($a_1, a_2, a_3, \dots, a_m$) (特征独立, 强假设条件, 因此称之为朴素)
 - 再假设现在有分类目标 $Y = \{y_1, y_2, y_3, \dots, y_n\}$
 - 那么就 $\max(P(y_1 | x), P(y_2 | x), P(y_3 | x), \dots, P(y_n | x))$ 是最终的分类类别
 - 而 $P(y_i | x) = P(x | y_i) * P(y_i) / P(x)$, 因为 x 对于每个分类目标来说都一样, 所以就是求 $\max(P(y_i) * P(x | y_i))$
 - $P(x | y_i) * P(y_i) = P(y_i) * \prod (P(a_j | y_i))$, 而具体的 $P(a_j | y_i)$ 和 $P(y_i)$ 都是能从训练样本中统计出来, 而具体的 $P(a_j | y_i)$ 和 $P(y_i)$ 都是能从训练样本中通过极大似然估计计算出来 (这里)
 - $P(a_j | y_i)$ 表示该类别下该特征 a_j 出现的概率 $P(y_i)$ 表示全部类别中这个这个类别出现的概率, 这样就能找到应该属于的类别了

条件概率、先验概率、后验概率、联合概率、贝叶斯公式的概念

- 条件概率
 - $P(X | Y)$: 表示 Y 发生条件下 X 发生的概率
- 先验概率
 - 表示事件发生前的预判概率。这个可以使基于历史数据统计, 也可以由背景常识得出, 也可以是主观观点得出。一般都是单独事件发生的概率, 如 $P(X)$
- 后验概率
 - 基于先验概率求得的反向条件概率, 形式上与条件概率相同(若 $P(X | Y)$ 为正向, 则 $P(Y | X)$ 为反向)
- 联合概率
 - 事件 X 与事件 Y 同时发生的概率
- 贝叶斯公式

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- - $P(Y)$ 叫做**先验概率**: 事件 X 发生之前, 我们根据以往经验和分析对事件 Y 发生的一个概率的判断
 - $P(Y|X)$ 叫做**后验概率**: 事件 X 发生之后, 我们对事件 Y 发生的一个概率的重新评估
 - $P(X, Y)$ 叫做**联合概率**: 事件 X 与事件 Y 同时发生的概率
 - 先验概率和后验概率是相对的. 如果以后还有新的信息引入, 更新了现在所谓的后验概率, 得到了新的概率值, 那么这个新的概率值被称为后验概率。

为什么朴素贝叶斯如此"朴素"

因为它假定所有的特征在数据集中的作用是同样重要和独立的。正如我们所知, 这个假设在现实世界中是很不真实的, 因此, 说朴素贝叶斯真的很"朴素"。用贝叶斯公式表达如下:

$$P(Y|X_1, X_2) = \frac{P(X_1|Y)P(X_2|Y)}{P(X_1)P(X_2)}$$

而在很多情况下，所有变量几乎不可能满足两两之间的条件。

朴素贝叶斯模型(Naive Bayesian Model)的朴素(Naive)的含义是“很简单很天真”地假设样本特征彼此独立.这个假设现实中基本上不存在，但特征相关性很小的实际情况还是很多的，所以这个模型仍然能够工作得很好。

什么是贝叶斯决策理论

贝叶斯决策理论是主观贝叶斯归纳理论的重要组成部分。贝叶斯决策就是在不完全情报下，对部分未知的状态用主观概率估计，然后用贝叶斯公式对发生概率进行修正，最后再利用期望值和修正概率做出最优决策(选择概率最大的类别)。

贝叶斯决策理论方法是统计模型决策中的一个基本方法，其基本思想是：

- 已知类条件概率密度参数表达式和先验概率
- 利用贝叶斯公式转换成后验概率
- 根据后验概率大小进行决策分类

朴素贝叶斯算法的前提假设是什么

- 特征之间相互独立
- 每个特征同等重要

什么是朴素贝叶斯中的零概率问题? 如何解决

零概率问题: 在计算实例的概率时，如果某个量 x ，在观察样本库(训练集)中没有出现过，会导致整个实例的概率结果是0。

解决办法: 若 $P(x)$ 为零则无法计算。为了解决零概率的问题，法国数学家拉普拉斯最早提出用加1的方法估计没有出现过的现象的概率，所以加法平滑也叫做**拉普拉斯平滑**。

举例: 假设在文本分类中，有3个类，、 C_1 、 C_2 、 C_3 ，在指定的训练样本中，某个词语 K_1 ，在各个类中观测计数分别为0, 990, 10, K_1 的概率为0, 0.99, 0.01，对这三个量使用拉普拉斯平滑的计算方法如下：

$$\begin{aligned}1/1003 &= 0.001, \\ 991/1003 &= 0.988, \\ 11/1003 &= 0.011\end{aligned}$$

在实际的使用中也经常使用加 λ ($\lambda \geq 0$) 来代替简单加1。如果对 N 个计数都加上 λ ，这时分母也要记得加上 $N * \lambda$ 。

将朴素贝叶斯中的所有概率计算应用拉普拉斯平滑即可以解决零概率问题。

朴素贝叶斯中概率计算的下溢问题如何解决

下溢问题: 在朴素贝叶斯的计算过程中，需要对特定分类中各个特征出现的概率进行连乘，小数相乘，越乘越小，这样就造成了下溢出。

为了解决这个问题，对乘积结果取自然对数。通过求对数可以避免下溢出或者浮点数舍入导致的错误。

$$\prod_{i=1}^n p(x_i|y_j)$$

解决办法: 对其取对数:

$$\log \prod_{i=1}^n p(x_i|y_j) = \sum_{i=1}^n \log p(x_i|y_j)$$

将小数的乘法操作转化为取对数后的加法操作，规避了变为零的风险同时并不影响分类结果。

当数据的属性是连续型变量时，朴素贝叶斯算法如何处理？

当朴素贝叶斯算法数据的属性为连续型变量时，有两种方法可以计算属性的类条件概率。

- 第一种方法：把一个连续的属性离散化，然后用相应的离散区间替换连续属性值。但这种方法不好控制离散区间划分的粒度。如果粒度太细，就会因为每个区间内训练记录太少而不能对 $P(X|Y)$ 做出可靠的估计，如果粒度太粗，那么有些区间就会有来自不同类的记录，因此失去了正确的决策边界。
- 第二种方法：假设连续变量服从某种概率分布，然后使用训练数据估计分布的参数，例如可以使用高斯分布来表示连续属性的类条件概率分布。
 - 高斯分布有两个参数，均值 μ 和方差 σ^2 ，对于每个类 y_i ，属性 X_i 的类条件概率等于：

 image-20210616215248145


 image-20210616215310629

通过高斯分布估计出类条件概率。

朴素贝叶斯有哪几种常用的分类模型

三种常用模型: 高斯, 多项式, 伯努利. 三个模型都是为了解决连续性变量引起的问题的.

- 多项式模型
 - 思想：将一个连续值视为特征的一个类别，这样会引起很多概率为0的问题。所以用拉普拉斯平滑解决。
 - 其中 α 为拉普拉斯平滑，加和的是属性出现的总次数，比如文本分类问题里面，不光看词语是否在文本中出现，也得看出现的次数。如果总词数为 n ，出现词数为 m 的话，说起来有点像掷骰子 n 次出现 m 次这个词的场景。

 image-20210616215447457

- 高斯模型
 - 思想：是要得到一个概率值，但是直接把一个值当做一个类别会引起零概率问题，于是假设并模拟一个分布，求其参数，进而求其概率。
 - 处理包含连续型变量的数据，使用高斯分布概率密度来计算类的条件概率密度

- 伯努利模型

- 伯努利模型特征的取值为布尔型，即出现为true没有出现为false，在文本分类中，就是一个单词有没有在一个文档中出现。
- 伯努利模型适用于离散特征情况，它将重复的词语都视为只出现一次。



我们看到，“发票”出现了两次，但是我们只将其算作一次。我们看到，“发票”出现了两次，但是我们只将其算作一次。

为什么说朴素贝叶斯是高偏差低方差

在统计学习框架下，大家刻画模型复杂度的时候，有这么个观点，认为 $\text{Error} = \text{Bias} + \text{Variance}$

- Error反映的是整个模型的准确度
- Bias反应的是模型在样本上的输出与真实值之间的误差，即模型本身的精准度
- Variance反应的是模型每一次输出结果与模型输出期望(平均值)之间的误差，即模型的稳定性，数据是否集中
- 对于复杂模型，充分拟合了部分数据，使得他们的偏差较小，而由于对部分数据的过度拟合，对于部分数据预测效果不好，整体来看可能引起方差较大。
- 对于朴素贝叶斯了。它简单的假设了各个数据之间是无关的，是一个被严重简化了的模型，简单模型与复杂模型相反，大部分场合偏差部分大于方差部分，也就是说高偏差而低方差。

朴素贝叶斯为什么适合增量计算

因为朴素贝叶斯在训练过程中实际只需要计算出各个类别的概率和各个特征的类条件概率，这些概率值可以快速的根据增量数据进行更新，无需重新全量训练，所以其十分适合增量计算，该特性可以使用在超出内存的大量数据计算和按小时级等获取的数据计算中。

高度相关的特征对朴素贝叶斯有什么影响

假设有两个特征高度相关，相当于该特征在模型中发挥了两次作用(计算两次条件概率)，使得朴素贝叶斯获得的结果向该特征所希望的方向进行了偏移，影响了最终结果的准确性，所以朴素贝叶斯算法应先处理特征，把相关特征去掉。

朴素贝叶斯的应用场景有哪些

- **文本分类/垃圾文本过滤/情感判别：**
 - 这大概是朴素贝叶斯应用最多的地方了，即使在现在这种分类器层出不穷的年代，在文本分类场景中，朴素贝叶斯依旧坚挺地占据着一席之地。因为多分类很简单，同时在文本数据中，分布独立这个假设基本是成立的。而垃圾文本过滤(比如垃圾邮件识别)和情感分析(微博上的褒贬情绪)用朴素贝叶斯也通常能取得很好的效果。
- **多分类实时预测：**
 - 对于文本相关的多分类实时预测，它因为上面提到的优点，被广泛应用，简单又高效。
- **推荐系统：**
 - 朴素贝叶斯和协同过滤是一对好搭档，协同过滤是强相关性，但是泛化能力略弱，朴素贝叶斯和协同过滤一起，能增强推荐的覆盖度和效果。

朴素贝叶斯有什么缺点

- 优点
 - 对数据的训练快，分类也快
 - 对缺失数据不太敏感，算法也比较简单
 - 对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，尤其是数据量超出内存时，可以一批批的去增量训练
- 缺点：
 - 对输入数据的表达形式很敏感
 - 由于朴素贝叶斯的“朴素”特点，所以会带来一些准确率上的损失
 - 需要计算先验概率，分类决策存在错误率

朴素贝叶斯与LR区别

- **朴素贝叶斯是生成模型**，根据已有样本进行贝叶斯估计学习出先验概率 $P(Y)$ 和条件概率 $P(X|Y)$ ，进而求出联合分布概率 $P(X,Y)$ ，最后利用贝叶斯定理求解 $P(Y|X)$ ，而**LR是判别模型**，根据极大化对数似然函数直接求出条件概率
- 朴素贝叶斯是基于很强的**条件独立假设**(在已知分类Y的条件下，各个特征变量取值是相互独立的)，而 LR 则对此没有要求
- 朴素贝叶斯适用于数据集少的情景，而LR适用于大规模数据集。

贝叶斯优化算法(参数调优)

- 网格搜索和随机搜索：在测试一个新点时，会忽略前一个点的信息；
- 贝叶斯优化算法：充分利用了之前的信息。贝叶斯优化算法通过对目标函数形式进行学习，找到使目标函数向全局最优值提升的参数。
- 学习目标函数形式的方法：
 - 首先根据先验分布，假设一个搜集函数；
 - 每一次使用新的采样点来测试目标函数时，利用这个信息来更新目标函数的先验分布
 - 算法测试由后验分布给出的全局最值最可能出现的位置的点。

对于贝叶斯优化算法，有一个需要注意的地方，一旦找到了一个局部最优值，它会在该区域不断采样，所以很容易陷入局部最优值。为了弥补这个缺陷，贝叶斯优化算法会在探索和利用之间找到一个平衡点，“探索”就是在还未取样的区域获取采样点；而“利用”则是根据后验分布在最可能出现全局最值的区域进行采样。

朴素贝叶斯分类器对异常值敏感吗

朴素贝叶斯是一种**对异常值不敏感**的分类器，保留数据中的异常值，常常可以保持贝叶斯算法的整体精度，如果对原始数据进行降噪训练，分类器可能会因为失去部分异常值的信息而导致泛化能力下降。

朴素贝叶斯算法对缺失值敏感吗

朴素贝叶斯是一种**对缺失值不敏感**的分类器，朴素贝叶斯算法能够处理缺失的数据，在算法的建模时和预测时数据的属性都是单独处理的。因此**如果一个数据实例缺失了一个属性的数值，在建模时将被忽略**，不影响类条件概率的计算，在预测时，计算数据实例是否属于某类的概率时也将忽略缺失属性，不影响最终结果。

一句话总结贝叶斯算法

贝叶斯分类器直接用贝叶斯公式解决分类问题。假设样本的特征向量为 x ，类别标签为 y ，根据贝叶斯公式，样本属于每个类的条件概率（后验概率）为：

 image-20210616221450148

分母 $p(x)$ 对所有类都是相同的，**分类的规则是将样本归到后验概率最大的那个类**，不需要计算准确的概率值，只需要知道属于哪个类的概率最大即可，这样可以忽略掉分母。分类器的判别函数为：

 image-20210616221502727

在实现贝叶斯分类器时，**需要知道每个类的条件概率分布 $p(x|y)$ 即先验概率**。一般假设样本服从正态分布。训练时确定先验概率分布的参数，一般用最大似然估计，即最大化对数似然函数。

贝叶斯分类器是一种生成模型，可以处理多分类问题，是一种非线性模型。