用python3.6开发，代码要有最基本的注释

面对一个较大含有错误的分类数据集，在上面进行贝叶斯统计建模。

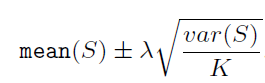
对于一个数据集，划分成训练集和测试集，考虑到脏数据的存在会影响最后的效果(精度、召回率等)，所以先进行清洗。在训练集中采样，对采集的样本进行清洗，进而得到一个较精确的贝叶斯统计模型。

具体方法如下：

一个有N个元组的数据集，记作P，从P中取一个大小为K(K为元组的个数)的子集S，并使每个元组被取到的概率相同。

1. 对于连续属性

用样本的均值和方差来估计整体的均值和方差：



其中95%的置信度等同于λ=1.96

因此，整体估计的均值f(P)可表示为f(P)=

(t)=t[a]，t[a]表示一个属性在某一元组中具体的值

主要处理的错误类型为值错误和重复错误，定义如下：

Correct(t)：对于一个元组t，返回其中正确的属性值

Numdup(t)：对于一个元组t，返回其重复的次数

(t)= t[a]表示清洗后的值

值错误：(t)=(Correct(t))

重复错误：令表示重复的次数（在P中），那么()=d× ，其中d=，=

举例说明：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 有房 | 婚姻状况 | 年收入 | 拖欠贷款 |
| 1 | 是 | 单身 | 125K(123K,2) | No |
| 2 | 否 | 已婚 | 100K(100K,1) | No |
| 3 | 否 | 单身 | 70K(70K,2) | No |
| 4 | 是 | 已婚 | 120K(120K,1) | No |
| 5 | 否 | 离婚 | 95K(104K,1) | Yes |
| 6 | 否 | 已婚 | 60K(60K,1) | No |
| 7 | 是 | 离婚 | 220K(220K,3) | No |
| 8 | 否 | 单身 | 85K(85K,1) | Yes |
| 9 | 否 | 已婚 | 75K(90K,2) | No |
| 10 | 否 | 单身 | 90K(90K,1) | Yes |

此表为采集出的一个样本，假设年收入这一列服从正态分布(一般都假设服从正太分布)，括号内第一个值为正确值，第二个值为重复的次数

d====1.28

(S)={,……,}

错误较多时，直接计算(S)的均值和方差

得到置信区间λ

错误较少时，通过(S)={125,100,……,90}和(S)来计算Q(S)

Q(S)={()-(),()-(),……,()-()}

计算Q(S)的均值和方差

因为f(P)=f()+

所以=

得到置信区间f(P)-λ

用这种区间的方法来估计条件概率，先验概率和边缘概率，再通过P(h|D)=公式，得到后验概率的区间(两个区间可通过左端点对左端点，右端点对右端点的加减乘除运算得到新的区间)

1. 对于离散属性

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 有房 | 婚姻状况 | 年收入 | 拖欠贷款 |
| 1 | 是 | 单身 | 125K | No |
| 2 | 否(是，2) | 已婚 | 100K | No |
| 3 | 否 | 单身 | 70K | No |
| 4 | 是 | 已婚 | 120K | No |
| 5 | 否 | 离婚 | 95K | Yes |
| 6 | 否 | 已婚 | 60K | No |
| 7 | 是 | 离婚 | 220K | No |
| 8 | 否 | 单身 | 85K | Yes |
| 9 | 否 | 已婚 | 75K | No |
| 10 | 否 | 单身 | 90K | Yes |

对概率P (有房=是|No) 进行估计

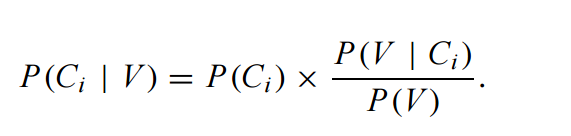
(t)=3(统计频率)

(t)=

类似于对连续属性的估计，最后在计算先验概率、条件概率、边缘概率的区间时除以样本总数即可

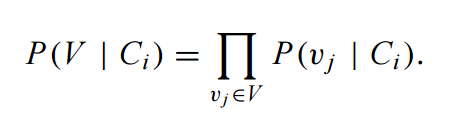
对于数据集中属性相关和不相关，在计算后验概率时的公式不同.

朴素贝叶斯的公式：

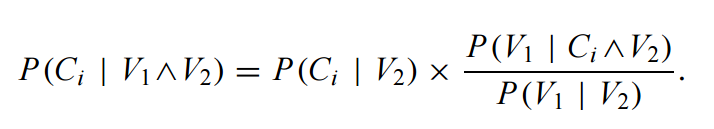


其中为分类属性，V为特征属性。

由于朴素贝叶斯假设属性之间是相互独立的，则对于条件概率有：



半朴素贝叶斯分类器是假设属性之间一部分是相关的，可以说是一种弱条件独立性假设，其公式为（以两个属性，为例）：



对于最后得到后验概率的决策，使用如下方法判定：

设a=[]和b=[]，且记L(a)=-，L(b)=-，则称



为a≥b的可能度。

输出：用这种区间方法得到的精度、召回率；不清洗数据集的精度、召回率；清洗整个数据集的精度、召回率。(过程中需要存储对先验概率，边缘概率，条件概率的统计值)