机器学习第四次作业
——朴素贝叶斯分类器

2052902 韩意

一、 问题描述

试用 Python 编程实现拉普拉斯修正的朴素贝叶斯分类器，并以西瓜数据集 3.0 为训练集，对下面“测 n”样本进行类别判定。

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测n	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.361	0.371	?

二、 算法实现

本题使用 Python 实现，用到的库包括 NumPy、Matplotlib、pandas。

朴素贝叶斯分类器通过学习数据集的先验与似然来进行预测，而其中似然是学习的关键对象。

考虑到先验是对应标签不同类别的出现频率，用一个数组存储即可，而似然的存储相对较为复杂。根据朴素贝叶斯分类器的训练过程，需要针对每个类别的每个属性进行学习，而属性又要分为离散属性与连续属性进行学习。

其中，离散属性要学习每个取值的概率，并且要注意使用拉普拉斯修正，公式如下：

$$\hat{P}(x_i|c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$$

连续属性要学习均值$\mu_{c,i}$和方差$\sigma_{c,i}^2$，从而有

$$p(x_i|c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$$

基于以上分析，可知最适合用于存储似然数据的数据结构是字典。字典第一

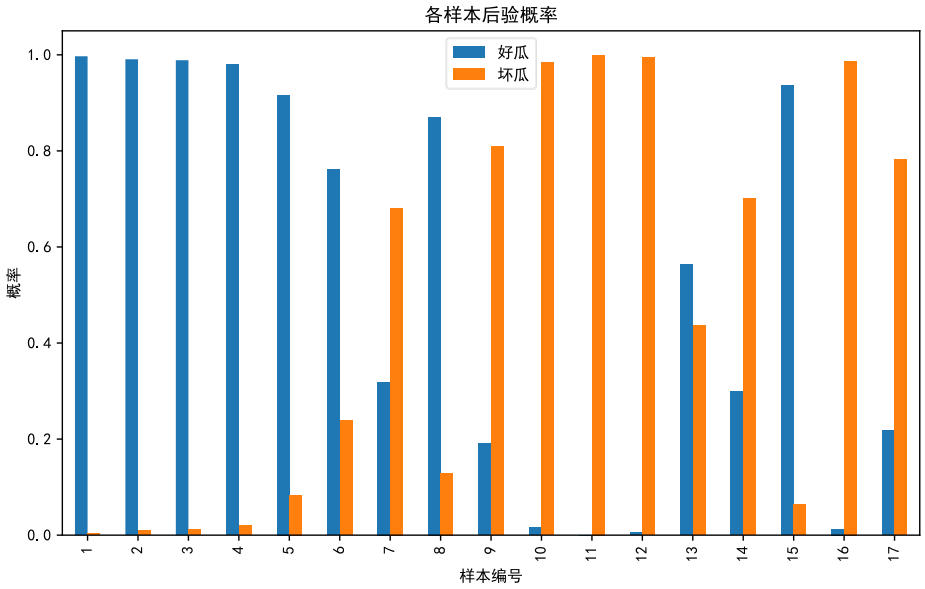
层对应不同类别，第二层对应不同属性，第三层对应相应属性下的参数，参数类型视属性为连续型还是离散型而定，连续型用嵌套字典保存均值和方差，离散型用 pandas Series 存对应属性的条件概率。似然的一部分训练结果打印如下：

```
{'好瓜': {'密度': {'mean': 0.57375, 'std': 0.12921051483086482}, '含糖率': {'mean': 0.27875, 'std': 0.10092394590553255}, '
色泽': 青绿    0.363636
乌黑    0.454545
浅白    0.181818
Name: 色泽, dtype: float64, '根蒂': 蜷缩    0.545455
稍蜷    0.363636
硬挺    0.090909
Name: 根蒂, dtype: float64, '敲声': 浊响    0.636364
沉闷    0.272727
清脆    0.090909
Name: 敲声, dtype: float64, '纹理': 清晰    0.727273
稍糊    0.181818
模糊    0.090909
Name: 纹理, dtype: float64, '脐部': 凹陷    0.545455
稍凹    0.363636
平坦    0.090909
Name: 脐部, dtype: float64, '触感': 硬滑    0.7
软粘    0.3
Name: 触感, dtype: float64}, '坏瓜': {'密度': {'mean': 0.49611111111111117, 'std': 0.19471867170641627}, '含糖率': {'mean':
0.1542222222222222, 'std': 0.10779468653159321}, '色泽': 青绿    0.333333
乌黑    0.250000
浅白    0.416667
```

预测阶段就是计算分类器的后验，选取后验最大的一个类别作为预测类别。为了避免连乘操作造成下溢，代码中使用对数求和。

基于以上原则，训练朴素贝叶斯分类器，计算并存储相应的先验与似然。训练集准确率为 0.82，“测 n”样本判定得到**好瓜**。

计算训练集的后验并进行概率归一化，得到下表：



前 8 个样本实际标签为好瓜，剩下的为坏瓜，由图可以清晰地看出 7、13、15 号三个样本判断错误，对应前文提到的训练集准确率 0.82。

注意，使用拉普拉斯修正后，求得的后验概率不可能为 0，这一点从图上可以得到验证。