作业 4　使用拉普拉斯修正的朴素贝叶斯分类器

1752931 胡斌

➢ 作业内容

本作业使用 PYTHON 语言完成，实现了根据给定数据集，训练一个朴素贝叶斯分类器模型，其中朴素贝叶斯分类器使用拉普拉斯修正以增强极端情况下模型的适用性。完整的程序由一系列函数和主函数组成，程序中包含的函数及各自的作用如下。

- `load_training_data(filename)`：加载训练集。
- `class_prior_probability_estimation(training_samples)`：计算各标签类别的类先验概率。
- `normal_distribution_parameter_estimation(data)`：对于一组数据，假设其服从某参数的正态分布，使用这组数据估计其服从的正态分布的参数（均值和标准差）。
- `num_of_sample_of_different_label(training_samples)`：根据训练集的数据，统计出训练样本中各标签类别的样本数量。
- `discrete_attribute_conditional_probability_estimation(training_samples, attributes)`：根据训练集的数据，估计各离散属性的条件概率。
- `continuous_attribute_distribution_parameter_estimation(training_samples)`：根据训练集的数据，估计各连续属性的正态分布参数。
- `load_testing_data(filename)`：加载测试集。
- `bayes_score(testing_sample, class_prior_probability, discrete_attribute_conditional_probability, parameters)`：对于某一个测试样本，计算预测其为各种可能的标签类别的（未用“证据”归一化）概率，作为预测该测试样本为各种可能的标签类别的得分。
- `classify(score)`：根据测试样本的得分情况，对测试样本进行判断，预测其为某一标签类别。

以上各函数的具体输入和输出的数据类型、使用方法、使用要求等内容，在代码相应位置均有详细注释，在此不赘述。

在主函数中，首先根据实际情况定义好训练集文件和测试集文件的文件名，然后使用 load_training_data 函数加载数据集，接着依次以上各函数，就能得到训练出来的朴素贝叶斯分类器模型，并对测试样本进行预测。

朴素贝叶斯分类器的训练思路是：首先估计类先验概率，然后再根据训练集数据估计各属性在不同类别标签下的条件概率。连续属性和离散属性的条件概率估计方法有所不同。离散属性的条件概率可直接根据训练集数据中的频率来估计，当训练集足够大且采样独立同分布时，能够较好有效地用频率估计概率。对于连续属性，本项目均假设其条件概率服从某参数的正态分布，使用训练集数据对其服从的正态分布参数进行无偏估计。在最后对测试样本进行预测时，对离散属性查找模型训练得到的条件概率估计值，对连续属性查找模型训练得到的对应正态分布参数并计算条件概率密度值。综合以上信息和类条件概率，对测试样本进行预测判断。

加载《机器学习》84 页表 4.3 给出的西瓜数据集 3.0，并以此作为训练集生成使用拉普拉斯修正的朴素贝叶斯分类器，然后使用 151 页中的“测 1”作为训练样本得到的结果如图 1 所示。

```
The score of Test Sample No.1 is: {'good': 0.025631024529740684, 'bad': 7.722360621178054e-05}
Test Sample No.1 is predicted to be a good watermelon.
```

图 1 “测 1”的预测结果

事实上，可以发现“测 1”就是西瓜数据集 3.0 的第一个训练样本，模型的预测结果是正确的，但是并不能表现模型的泛化性能。

➢ 作业说明

本作业的程序逻辑相对清晰，在代码中有充分注释，能够比较容易地看懂。下面列出比较关键的点或在完成过程中遇到的一些问题进行说明。

- 拉普拉斯修正

关于朴素贝叶斯分类器的拉普拉斯修正的相关论述，可见教材《机器学习》第 153 页至 154 页。在本项目中，拉普拉斯修正需要在 class_prior_probability_estimation 函数和 discrete_attribute_conditional_probability_estimation 函数中使用。在对连续属性进行条件概率正态分布参数估计的时候，不存在拉普拉斯修正的问题。

- 无偏估计

在训练朴素贝叶斯分类器时，针对连续属性的条件概率估计问题，本项目假设所有对应的条件概率函数均为正态分布的概率密度函数，然后根据训练集的数据估计对应正态分布的参数。本项目使用参数的无偏估计，即：正态分布的均值为样本均值，正态分布的方差为样本方差。需要注意的是，样本方差的计算公式与教材 150 页的公式略有区别。经验证，教材中 152 页中的相关计算结果均使用无偏估计。

- 深拷贝（deepcopy）

在第 2 次作业——决策树中，就提到过深拷贝的问题。本次作业中，又有一处需要用到深拷贝。由于程序设计的原因，在 bayes_score 函数中，第 303 行和第 308 行需要使用深拷贝，否则会因对拷贝对象误操作而出错。

- **效果分析**

如前所述，本作业的要求并不能体现出训练得到的朴素贝叶斯分类器的泛化性能，因为测试样本是训练集中的一个样本。为了探究朴素贝叶斯分类器的性能，额外进行了如下实验。

将西瓜数据集 3.0 中的第 16 个样本作为测试样本，其它样本组成训练集。训练好朴素贝叶斯分类器后，对测试样本进行预测，得到的结果如图 2 所示。

```
The score of Test Sample No.16 is: {'good': 0.00014065427973609937, 'bad': 0.0033637151176224515}
Test Sample No.16 is predicted to be a bad watermelon.
```

图 2 样本 16 的预测结果

将西瓜数据集 3.0 中的第 17 个样本作为测试样本，其它样本组成训练集。训练好朴素贝叶斯分类器后，对测试样本进行预测，得到的结果如图 3 所示。

```
The score of Test Sample No.17 is: {'good': 0.0017829933409240978, 'bad': 0.0016271233744638383}
Test Sample No.17 is predicted to be a good watermelon.
```

图 3 样本 17 的预测结果

进一步地，将西瓜数据集中的每一个样本分别作为测试样本，其余样本作为训练样本，使用交叉验证法进行模型的训练和预测，得到的结果记录于表 1。从表 1 中可以看到，17 个测试样本中，有 11 个样本预测正确，有 6 个样本预测错误，平均正确率约为 0.647。根据大数定律，需要足够多的数据才能较准确地估计各属性的条件概率，而本作业使用的训练样本较少，因此效果有待提高。

表 1 使用交叉验证法进行模型训练和预测

测试样本编号	实际类别	预测类别	是否预测正确
1	好瓜	好瓜	是
2	好瓜	好瓜	是
3	好瓜	好瓜	是
4	好瓜	好瓜	是
5	好瓜	好瓜	是
6	好瓜	坏瓜	否
7	好瓜	坏瓜	否
8	好瓜	好瓜	是
9	坏瓜	坏瓜	是
10	坏瓜	坏瓜	是
11	坏瓜	坏瓜	是
12	坏瓜	坏瓜	是
13	坏瓜	好瓜	否
14	坏瓜	好瓜	否
15	坏瓜	好瓜	否
16	坏瓜	坏瓜	是
17	坏瓜	好瓜	否

- **改进方向**

- 增加训练集数据量，使训练出的模型性能更好。
- 考虑半朴素贝叶斯分类器。
- 使用取对数的方法，将计算过程中的“连乘”变成“连加”，避免在计算 bayes_score 函数中出现数值下溢。
- ……