贝叶斯分类任务

概述

- 利用贝叶斯分类算法对 wine 数据集中的测试集进行分类。

数据说明

- wine 葡萄酒数据集是 UCI 上的公开数据集。数据集包含由三种不同葡萄酿造的葡萄酒，通过化学分析确定了葡萄酒中含有的 13 种成分的含量。数据集的相关信息如表 1 所示：

表 1 wine 数据集相关信息

样例数量	特征维度	特征类型	类别数量
178	13	数值	3

- 数据集已被划分为训练集和测试集，分别存储于 data 文件夹中的 train_data.csv,和 test_data.csv。其中，训练集包含 120 个样例，测试集包含 58 个样例，每个样例包含各个维度的特征值及样例标签（标签为 1、2 或 3），假定各维度的特征属性之间条件独立。

任务说明

- 基于贝叶斯分类原理，实现一个**贝叶斯分类器**。在**训练集**中进行训练，尽可能提高模型准确率，并在**测试集**上进行测试。在朴素贝叶斯分类模型中，当属性是连续型时，有两种方法可以计算属性的类条件概率：第一种方法是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，之后用频率去表示类条件概率，但这种方法不好控制离散区间划分的粒度；第二种方法是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，通过高斯分布估计出类条件概率。
- 本实验规定采用**高斯分布**估计类条件概率。其中，均值和方差分别用训练集的**样本均值**和**样本方差**估计。
- 实验报告要求对**贝叶斯分类模型的过程**进行推导，并计算各个属性各个类别的**类条件密度**（高斯分布），同时，给出测试集的**预测准确率。**
- 测试集预测结果文件需要包含每个测试样例的**预测类别**及**分属于三个类别的概率值。**

作业提交格式要求

- 需提供完整的**代码文件**、**测试集分类结果文件**和**实验报告**，将以上内容打包压缩，**压缩文件命名格式：学号-姓名-贝叶斯分类任务实验**；

- 提交测试集分类结果文件时，请将文件命名为**test_prediction.csv**，文件格式参照sample_submission.csv；
- 尽量以相对路径的形式索引数据集，便于我们对代码进行复现；
- **代码若有雷同，一律按 0 分处理。**

Tips

- 推荐语言：Matlab、Python（可采用 Numpy，Pandas，Matplotlib 等基础代码集成库）、C++；
- 不得使用集成度较高，函数调用式的代码库（如 Python 环境下的 sklearn，PyTorch，Tensorflow 等）；
- 代码可加适当注释，提高阅读性；
- 实验报告尽量语言简洁、逻辑清晰、计算和推导过程详尽。