**练习1：**

在iris数据集中，选择恰当的属性或属性组合，训练逻辑回归模型，区分山鸢尾和维吉尼亚鸢尾，并测试模型性能，以恰当的形式展现训练过程和结果。

要求：

(1)编写代码实现上述功能；

(2)记录实验过程和结果：

尝试调试超参数，综合考虑准确率、交叉熵损失、和训练时间等，使模型在测试集达到最优的性能，并以恰当的方式记录和展示实验结果；

(3)分析和总结：

区分山鸢尾和维吉尼亚鸢尾，至少需要几种属性？说明选择某种属性或属性组合的依据；对比以上的两种方案，以表格或其他合适的图表形式展示。通过以上结果，可以得到什么结论，或对你有什么启发。

**练习2：**

在iris数据集中，分别选择2种、3种和4种鸢尾花的属性，训练逻辑回归模型，区分变色鸢尾和维吉尼亚鸢尾，并测试模型性能，以恰当的形式展现训练过程和结果。

要求：

(1)编写代码实现上述功能；

(2)记录实验过程和结果：

尝试调试超参数，综合考虑准确率、交叉熵损失、和训练时间等，使模型在测试集上达到最优的性能，并以恰当的方式记录和展示实验结果；

(3)分析和总结：

比较分别使用2种、3种和4种鸢尾花的属性训练模型时，学习率、迭代次数等参数、以及在训练集和测试集上的交叉熵损失、准确率和训练时间等结果，以表格或合适的图表形式展示。通过以上结果，可以得到什么结论，或对你有什么启发。

**选做题：**

乳腺癌肿瘤数据集，由威斯康辛大学麦迪逊医院的William博士提供，可在UCI数据库(加州大学欧文分校提出的用于机器学习的数据库)里下载。

下载地址：<https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data>

该数据集中有699个乳腺癌肿瘤样本，每个样本包含10个属性和1个肿瘤标签，其结构如图1所示。第1列为id号，第2-10列为肿瘤特征，第11列为肿瘤的标签。每个属性的属性值均为0-9之间的整数，标签值取2或4，2表示良性、4表示恶性。

图1 乳腺癌肿瘤数据集(前5行数据）

要求：

⑴ 下载数据集：

下载数据集，合理划分为训练集和测试集。

⑵ 数据预处理：

数据集中有16处缺失值，用“？”表示，在输入模型之前，需要对这些缺失值进行处理。首先将“？”替换为NaN，再进行丢弃，最后对已经丢弃缺失值的数组索引进行重置。

相关函数：

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 函数 | 函数功能 | 函数相关库 |
| (1) | 对象名.replace(to\_replace=’?’,value=np.nan) | 将问号替换为NaN | Python内置函数 |
| (2) | 对象名.dropna() | 丢弃缺失值 | Pandas库 |
| (3) | 对象名.reset\_index() | 索引重置 | Pandas库 |

(3)建立逻辑回归模型，使用属性“肿块密度”和“细胞大小的均匀性”训练模型，综合考虑准确率、交叉熵损失、和训练时间等，使模型在测试集达到最优的性能，并以合适的形式展现训练过程和结果。

(4)选择其他属性或属性组合训练模型:

尝试选择数据集中的其他属性或者属性组合，训练和测试逻辑回归模型，并展现训练过程和结果。

(5)分析和总结：

比较采用不同的属性或属性组合训练模型时，学习率、迭代次数，以及交叉熵损失、准确率和模型训练时间等，以表格或其他合适的图表形式展示。通过以上结果，可以得到什么结论，或对你有什么启发。