实验一 线性模型

一、 实验目的

1、掌握线性回归和逻辑回归的实现过程与应用场景

2、知道线性回归与逻辑回归的不同之处

3、理解两种回归算法的评估标准

4、熟悉 scikit-learn、numpy、pandas 等库的使用

二、 实验内容

（一）线性回归

1. 定义问题：波士顿房价预测，用可用的工具进行统计分析，建立优化模型，基于该模型评估客户房产的最佳售价。
2. 数据集介绍：

- 名称：Boston House Price Dataset
- 属性：

属性名称	含义
CRIM	城镇人均犯罪率
ZN	住宅用地所占比例
INDUS	城镇中非住宅用地所占比例
CHAS	CHAS 虚拟变量,用于回归分析
NOX	环保指数
RM	每栋住宅的房间数
AGE	1940 年以前建成的自住单位的比例
DIS	距离 5 个波士顿的就业中心的加权距离
RAD	距离高速公路的便利指数
TAX	每一万美元的不动产税率
PRTATIO	城镇中的教师学生比例
B	城镇中的黑人比例
LSTAT	地区中有多少房东属于低收入人群

PRICE	房屋价格

- 数据描述：共 506 条样本，14 列数据，1~13 列为帮助预测的属性，最后一列为房屋价格。没有缺省值。
- 下载链接：

  https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data
- 提供文件介绍：housing-data.csv（数据）
- 选择 80%的数据作为训练集，20%的数据作为测试集

（二）逻辑回归

1. 定义问题：恶性乳腺癌肿瘤预测，用可用的工具进行统计分析，建立优化模型，基于该模型预测该肿瘤为良性还是恶性。

2. 数据集介绍：

- 名称：Wisconsin Breast Cancer Dataset
- 属性：

属性名称	含义
Sample code number	示例编号
Clump Thickness	团块厚度
Uniformity of Cell Size	细胞大小的均匀性
Uniformity of Cell Shape	细胞形状的均匀性
Marginal Adhesion	边际附着力
Single Epithelial Cell Size	单个上皮细胞大小
Bare Nuclei	裸核
Bland Chromatin	平淡的染色质
Normal Nucleoli	正常核仁
Mitoses	有丝分裂症
Class	所属类别

- 数据描述：共 699 条样本，11 列数据，第一列为检索的 id，后 9 列为与肿瘤相关的医学特征，最后一列表示肿瘤类型的数值。有缺省值。
- 下载链接：

https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

- 提供文件介绍：breast-cancer-wisconsin.data（数据）
  breast-cancer-wisconsin.names（数据集介绍）
- 选择 80%的数据作为训练集，20%的数据作为测试集

三、 实验环境

1. 系统：Window 10
2. 软件：Anaconda 与 JupyterNotebook 的集成开发环境。
3. 依赖库：
   - numpy
   - matplotlib
   - pandas
   - scikit-learn

四、 实验步骤

（一）手写代码

1. 导入实验所需包

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

2. 定义模型

- 参数初始化
- 定义预测函数与损失函数
  - 思考：如何选择预测函数？如何选择损失函数？
- 模型训练
  - 思考：采用何种优化方式？
- 模型保存与加载
- 模型预测与评估

3. 数据处理

- 获取数据集
- 缺省值处理
- 特征标准化
- 分割数据集
  - 思考：如何进行特征选择？如何处理数据？

4. 主函数

（二）调用 API

1. 导入实验所需包

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression #线性回归
from sklearn.linear_model import LogisticRegression #逻辑回归
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error
```

2. 定义模型：调用 API
3. 数据处理
4. 主函数

五、 实验要求

（一）提交要求：

- 代码源文件：
  1. 采用全部特征进行线性回归或逻辑回归
  2. 采用部分特征进行线性回归或逻辑回归
  3. 文件格式：.py 文件或者.ipynb
- 实验报告：描述实验方案和结果
- 将上述文件放在一个文件夹中，并命名为:学号_姓名，上传至课程平台。