**第六次实验报告**

闫嘉依 2021214417

**作业标题：python工具包使用2**

**作业说明**

**（一）文件中存储了美国小孩的姓名， 出生年份, 性别，出现次数。请使用pandas处理数据，文件以txt形式存储。完成以下任务：**

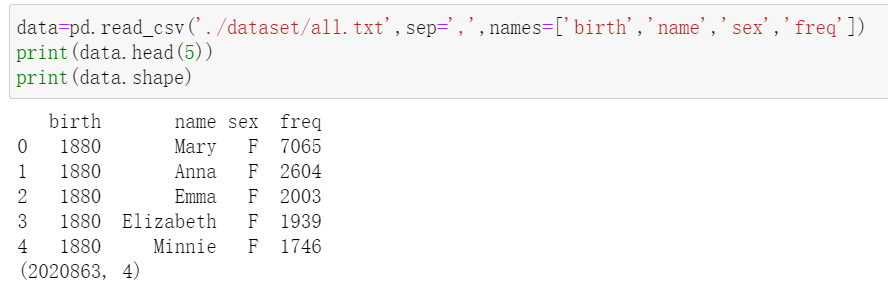
合并数据



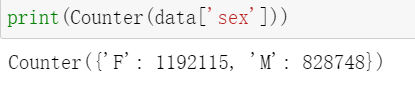
记录all.txt写入进度



用dataframe形式读txt，增加表头，输出df的前5行和维度大小

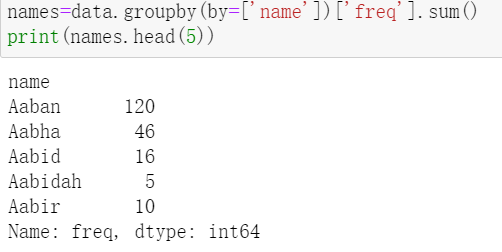


**（1）统计数据集中男孩名字和女孩名字各是多少。**

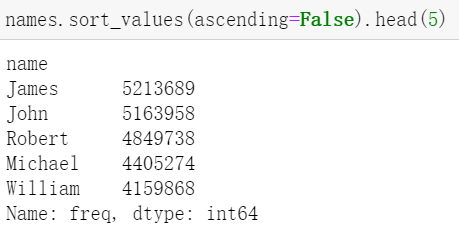


用counter统计sex列各值频率，男孩名字用F表示，女孩名字用M表示

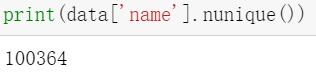
**（2）按照Name字段将数据集进行分组并求和赋值给变量names，最后输出前五行。**



**（3）按照每个名字被使用的次数（Count）对第二步中结果进行降序排序，得出最受欢迎的的五个名字。**



**（4）在数据集中，共出现了多少个名字？(不包含重复项)**



统计name列不重复元素个数

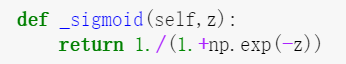
**（二）手动实现逻辑回归（除数据处理工具外不允许调库）来进行癌症分类预测，并使用工具库也实现一个逻辑回归，比较两者的结果。数据集有11个列，第1个列为id号，第2-10列为特征，11列为标签（2为良性、4为恶性）**

为了便于使用，将手动逻辑回归算法封装成和sklearn调用一样的方式

实现原理：

模型形式：

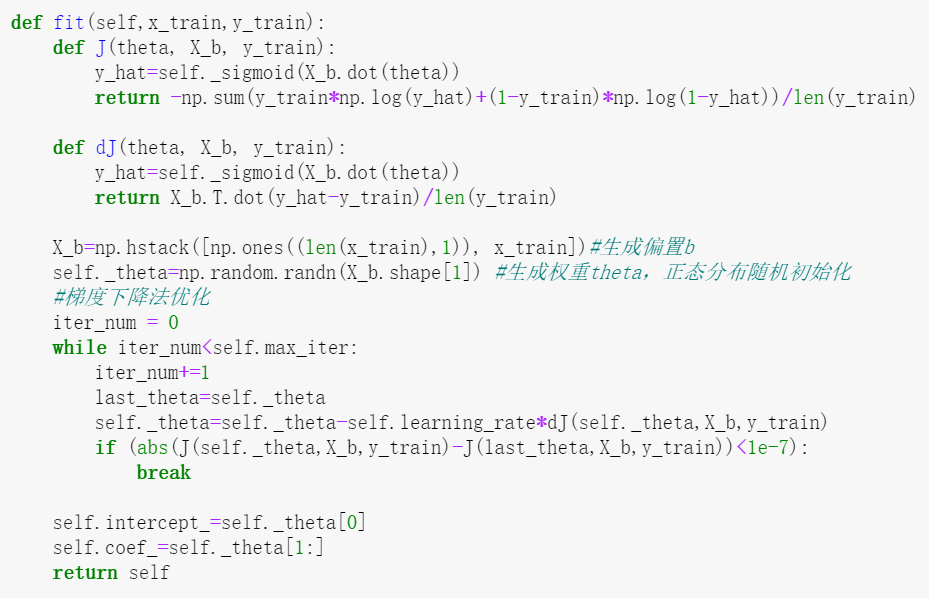
1.定义激活函数sigmoid



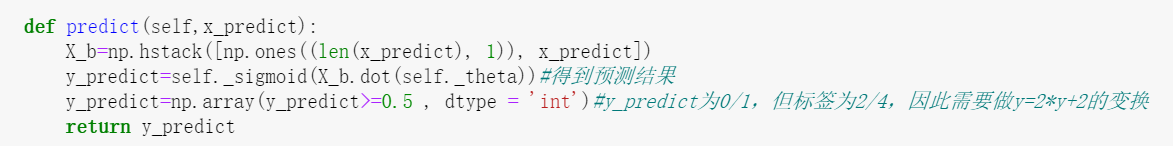
2.定义拟合方法，使用梯度下降法优化参数

loss——交叉熵损失函数，其中，y是真值，是预测值

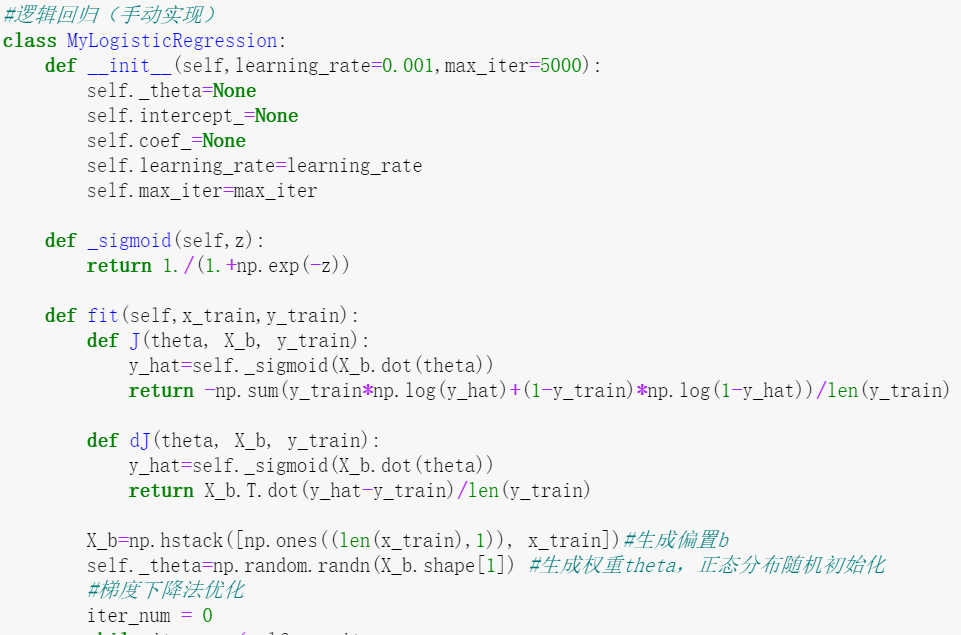
梯度下降优化公式：

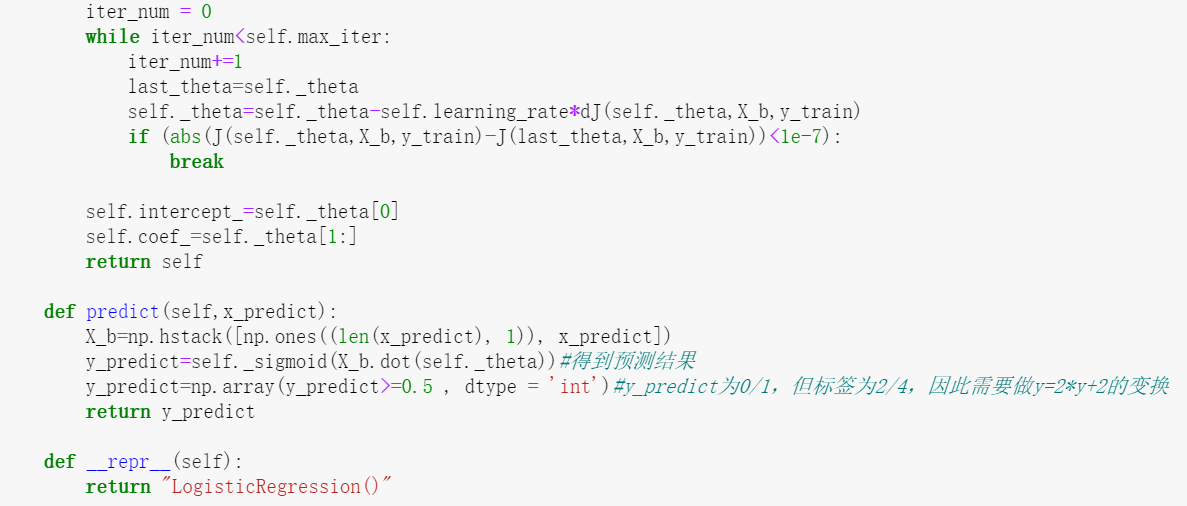


定义预测方法



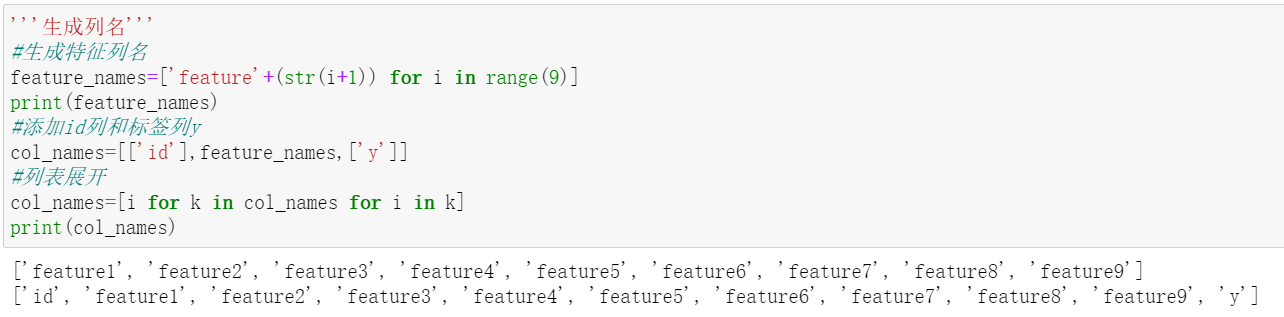
将模型过程包装成类

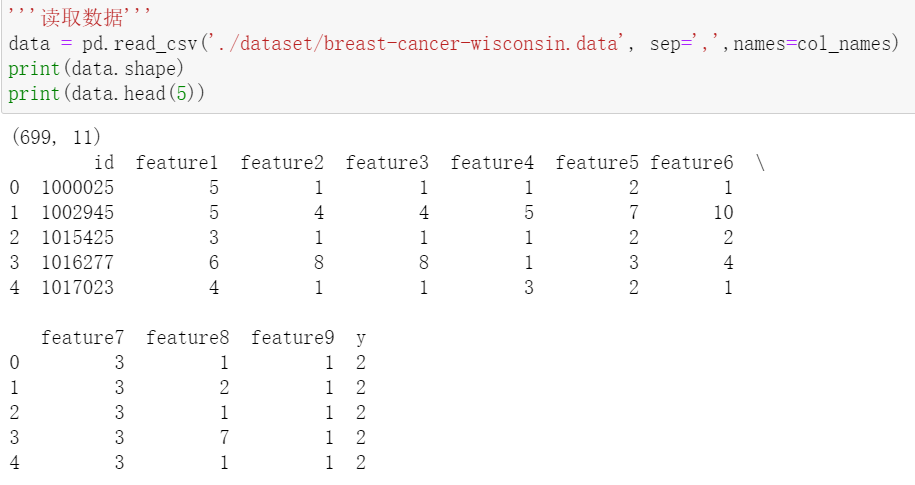




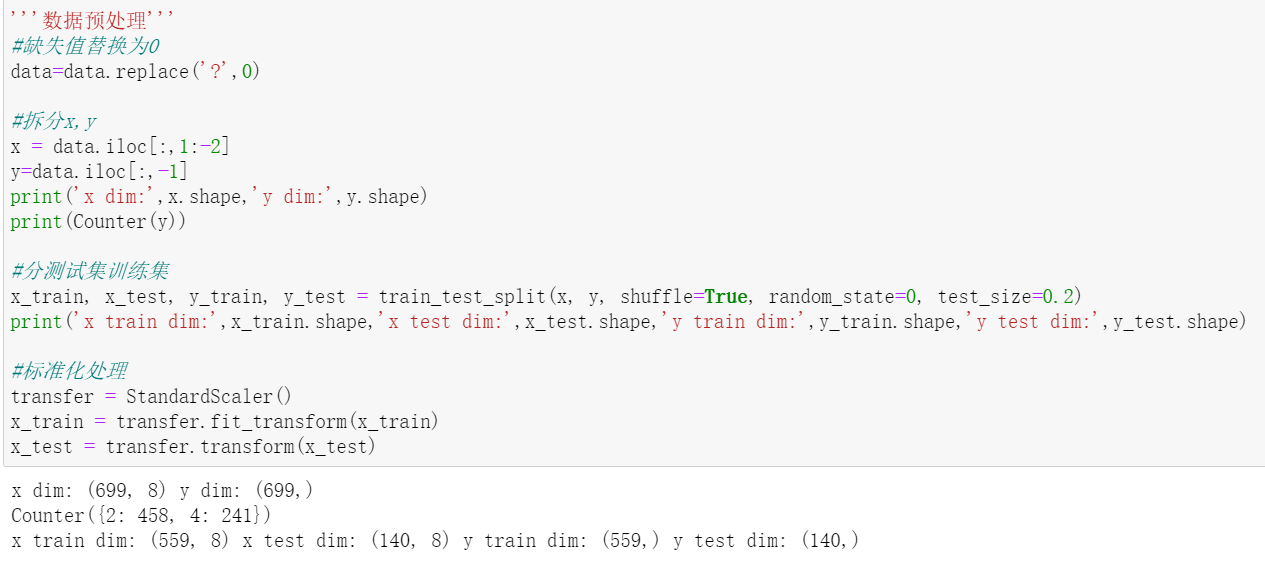
数据预测过程如下：

读取数据

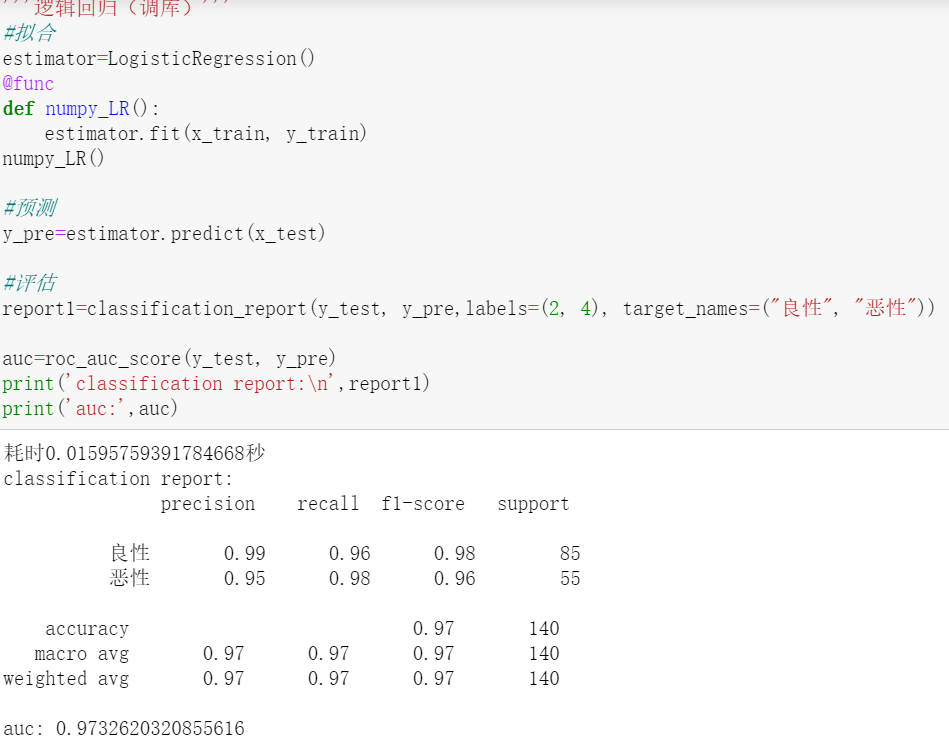




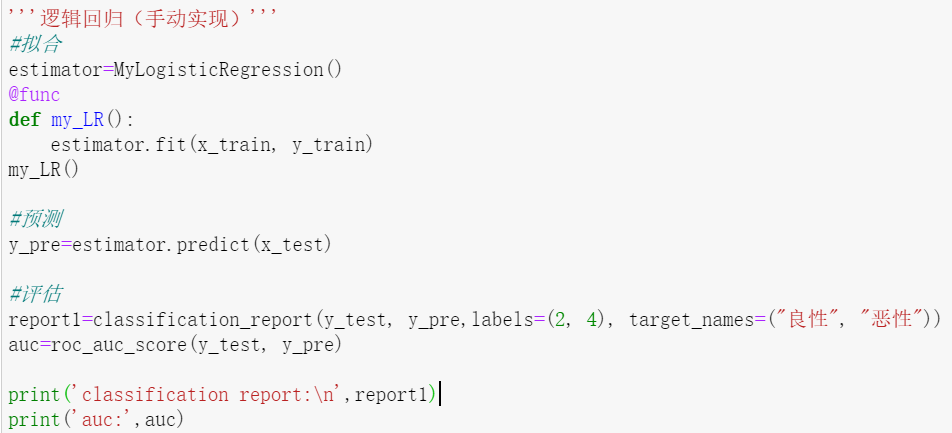
数据预处理

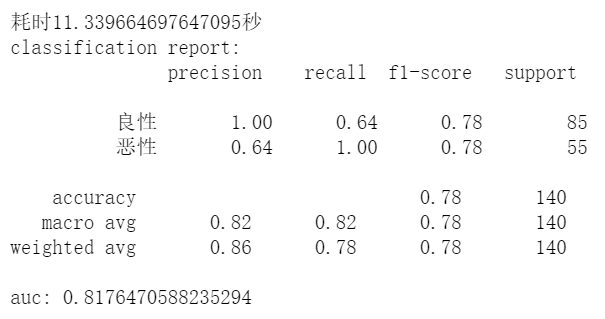


调库逻辑回归，并完成预测，输出运行时间和预测性能



手动逻辑回归





可以发现，手动实现的逻辑回归运行时间比基于库的长很多，为了平衡预测性能和运行时间，将迭代次数设为5000，学习率设为0.001