

**Beijing Jiaotong University**

**机器学习线性模型实验报告**

**学 院：软件学院**

**姓 名：张鑫成**

**学 号：20271055**

**班 级：软件2005**

**专 业：软件工程**

**指导老师：鲍鹏**

2022年09月18日

1. **实验目的及要求**

实验一：

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。

数据集名称：Boston House Price Dataset

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

实验二：

提供文件介绍：housing-data.csv（数据）

数据集名称：Wisconsin Breast Cancer Dataset

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

两个实验分别进行手写代码以及调用API两种方法书写。

手写代码的基本实验步骤如下：第一步，导入实验所需包；第二步，定义模型，分别进行参数初始化、定义预测函数与损失函数、模型训练、模型保存与加载、模型预测与评估；第三步：数据处理，有获取数据集、缺省值处理、特征标准化、分割数据集的步骤。

调用API的基本实验步骤如下：第一步，导入实验所需包；第二部定义模型（调用API的方式）；第三步，数据处理；第四步，主函数。

同时，本实验需要同时从全部特征、部分特征的角度入手。

总的来说，本实验需要在是否采用全部或者部分特征，是否手写代码或者调用API，是否是线性回归或者逻辑回归，总共分为八个小实验，需要提交八个.py文件或者.ipynb。

1. **仪器用具及实验用具**

1. 系统：Window 11

2. 软件：Anaconda与JupyterNotebook的集成开发环境。

3. 依赖库：

1.numpy

2.matplotlib

3.pandas

4.scikit-learn

1. **实验原理**

线性回归（Linear Regression）是一种通过属性的线性组合来进行预测的线性模型，其目的是找到一条直线或者一个平面或者更高维的超平面，使得预测值与真实值之间的误差最小化。

利用线性回归定义，设定所求值与投入特征存在线性关系，并且利用损失函数，梯度下降或者最小二乘法等方法综合求解参数。最后完善模型使得模型具有不错的预测能力。

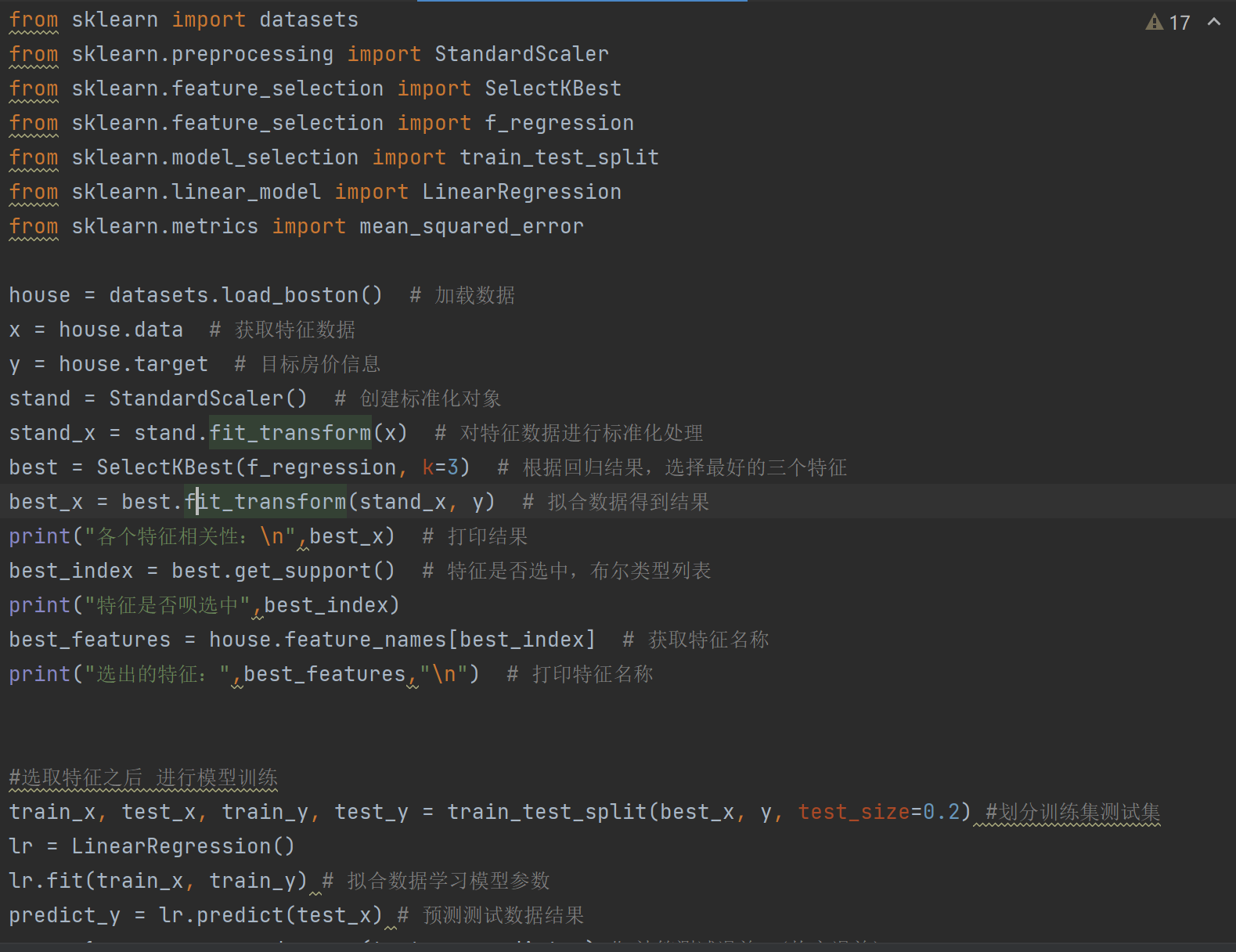
同样，在线性回归的基础上，利用sigmoid等函数将回归内容区间化，结合数据本身的特点使得其达到能够进行分类的目的。因为其分类问题的性，综合选择不一样的损失函数以及优化方法。

1. **实验方法与步骤**

**波士顿房价预测调用API全特征：**

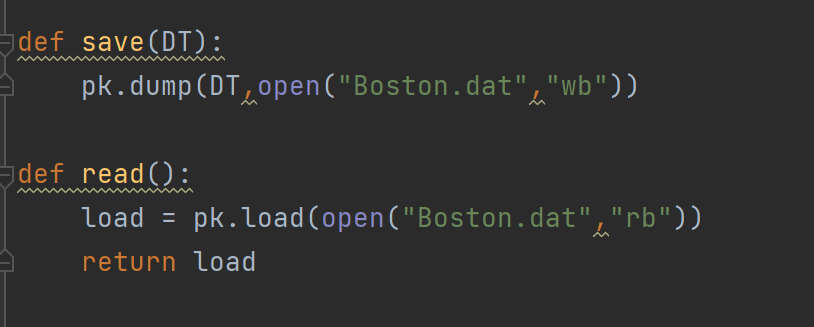
主要利用aklearn包，利用包中的数据装载、数据划分、创建模型、训练模型、预测比对、以及调用相关数学函数计算对应的均方误差。

**波士顿房价预测调用API部分特征：**



在全部特征的基础上，调用了 SelectKBest函数，从原先的特征里面选择了三个最好的特征，重新获取一个best\_x列表，代替之前的x然后进行上一类实验同样的步骤。

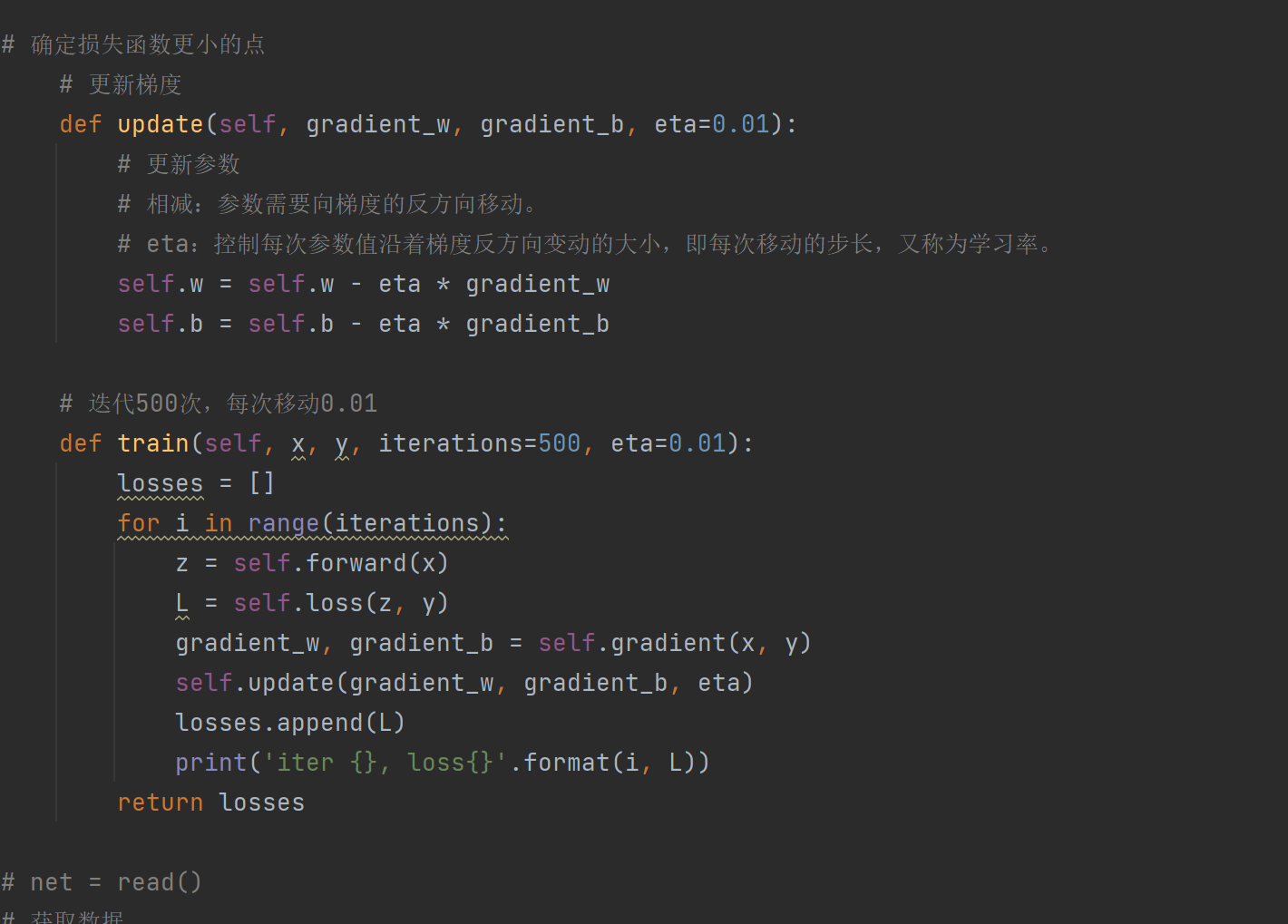
**波士顿房价预测手写代码全特征：**

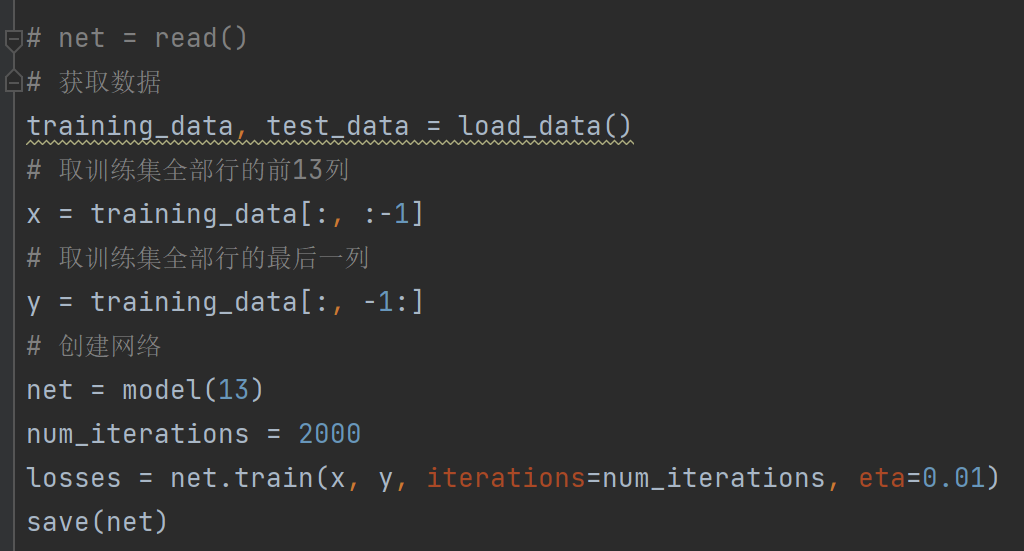


定义并使用了模型的装载于加载函数



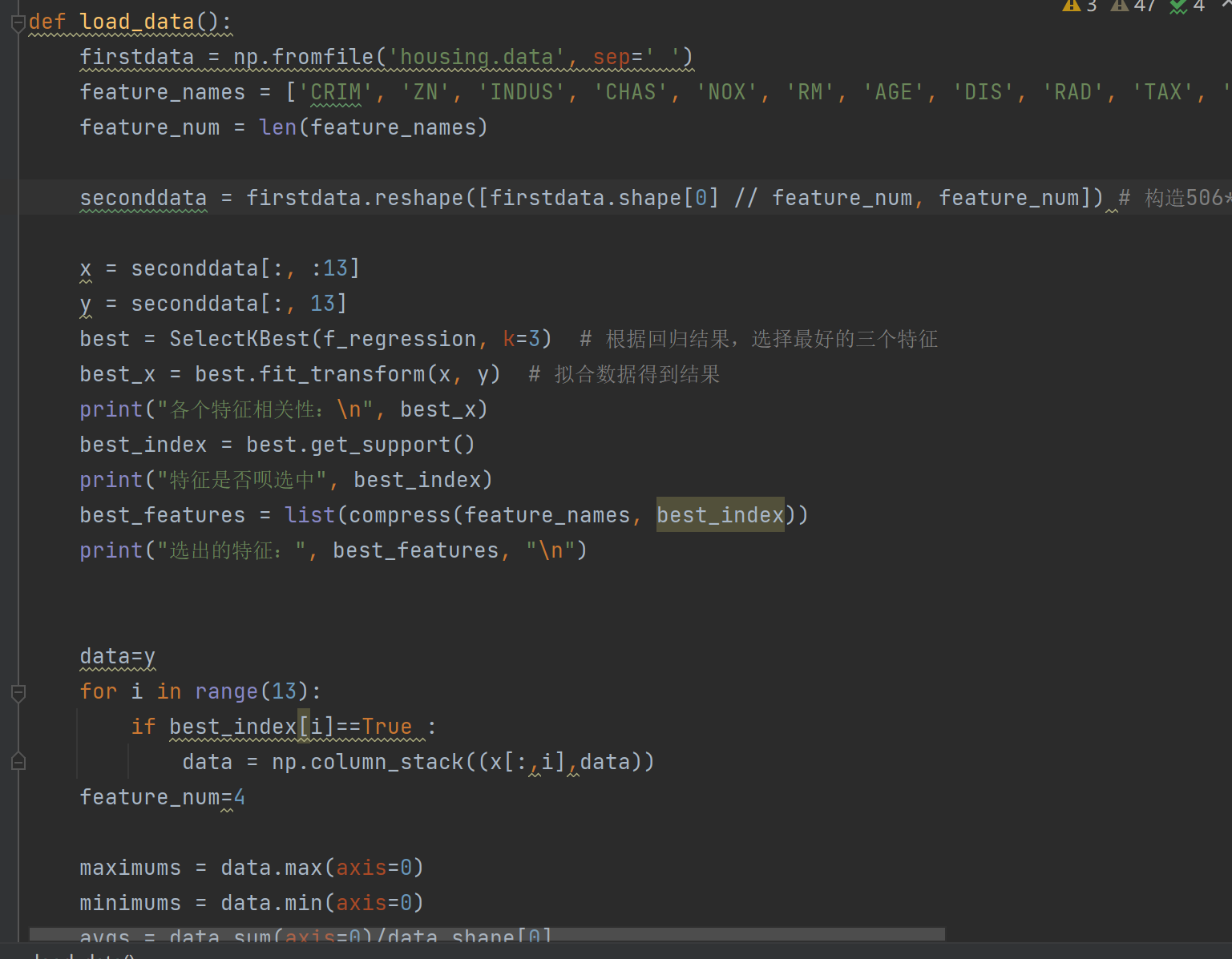
利用np从文件中读取文件，标志好需要的特征及其姓名，然后对每一列特征都进行归一化处理值域为[-0.5,0.5]，划分其中80%作为训练集、20%作为数据集。

定义模型，其中包括模型的初始化，forward函数计算y的值，loss函数（根据均方误差的定义返回对应误差），gradient函数计算w，b的梯度变换的计算，以及update函数更新梯度，以及训练函数（模型迭代500次

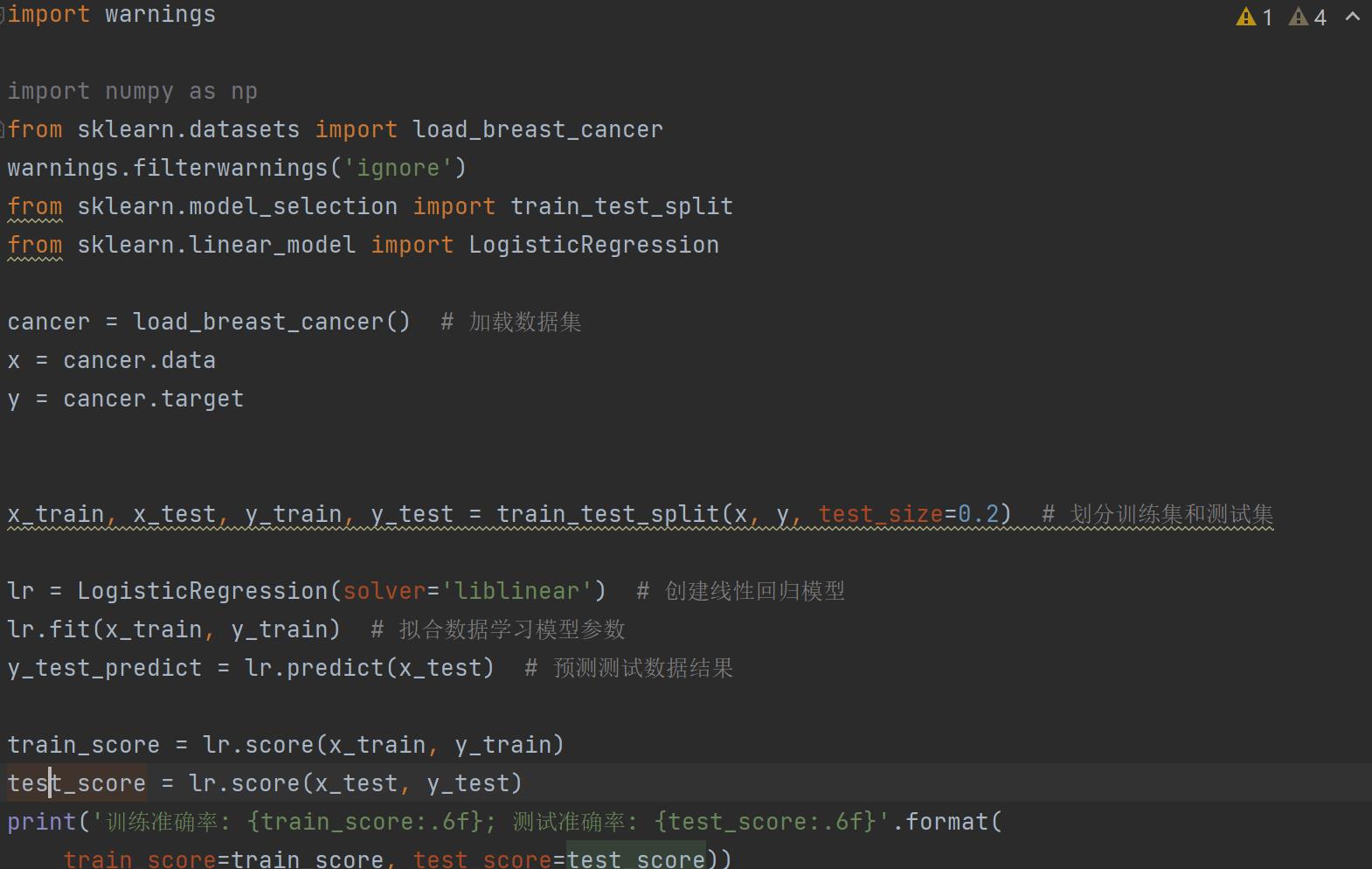
每次移动0.01，每一次都进行梯度的下降，并且输出对应的损失函数）。

主函数，进行相关的调用。

**波士顿房价预测手写代码部分特征：**

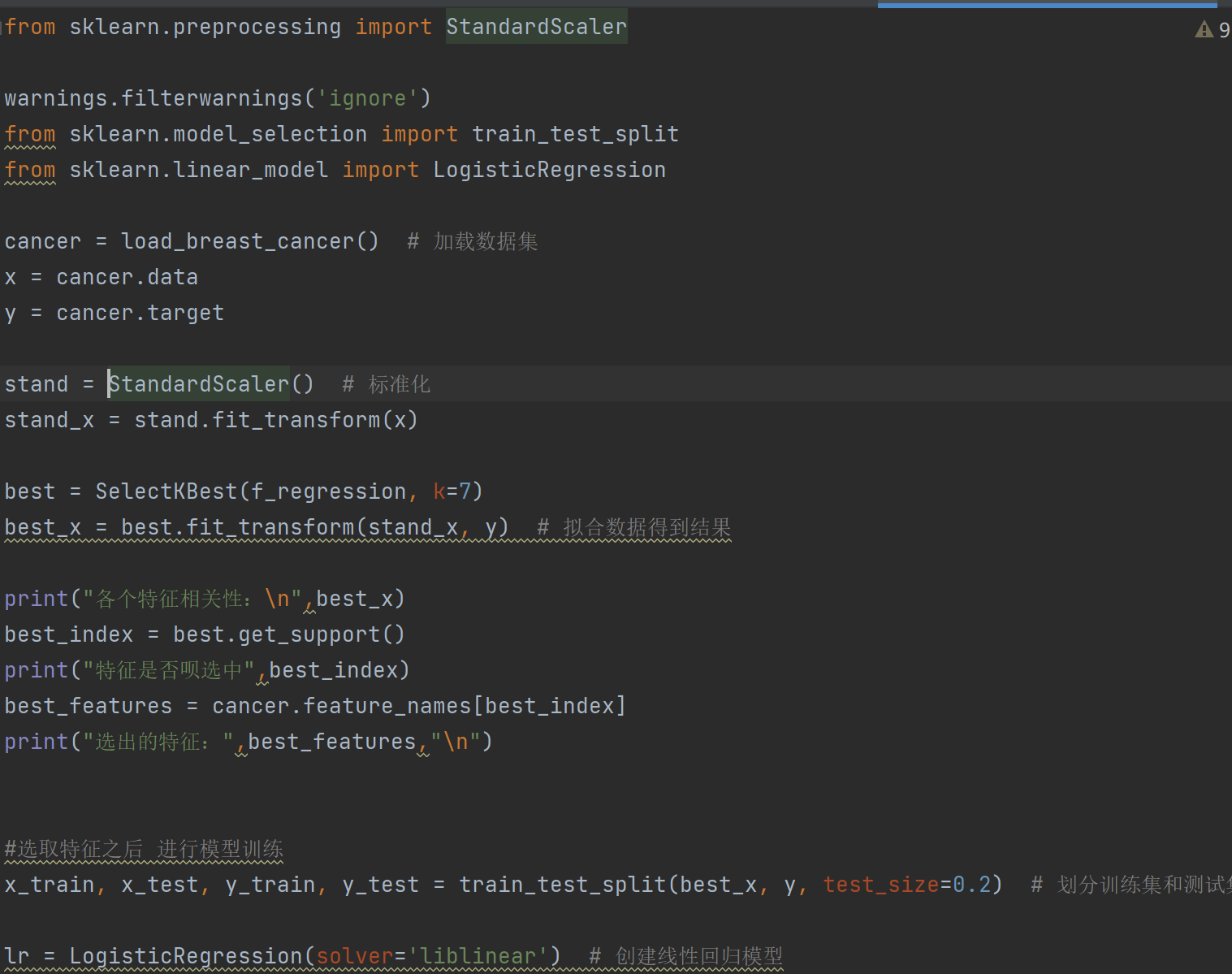
在上一个实验的基础之上，利用SelectKBest函数，选择最好的三个特征，并且改变x列表内容，用best\_x代替然后进行与上个实验一样的步骤。

**肿瘤预测调用API全特征：**

****

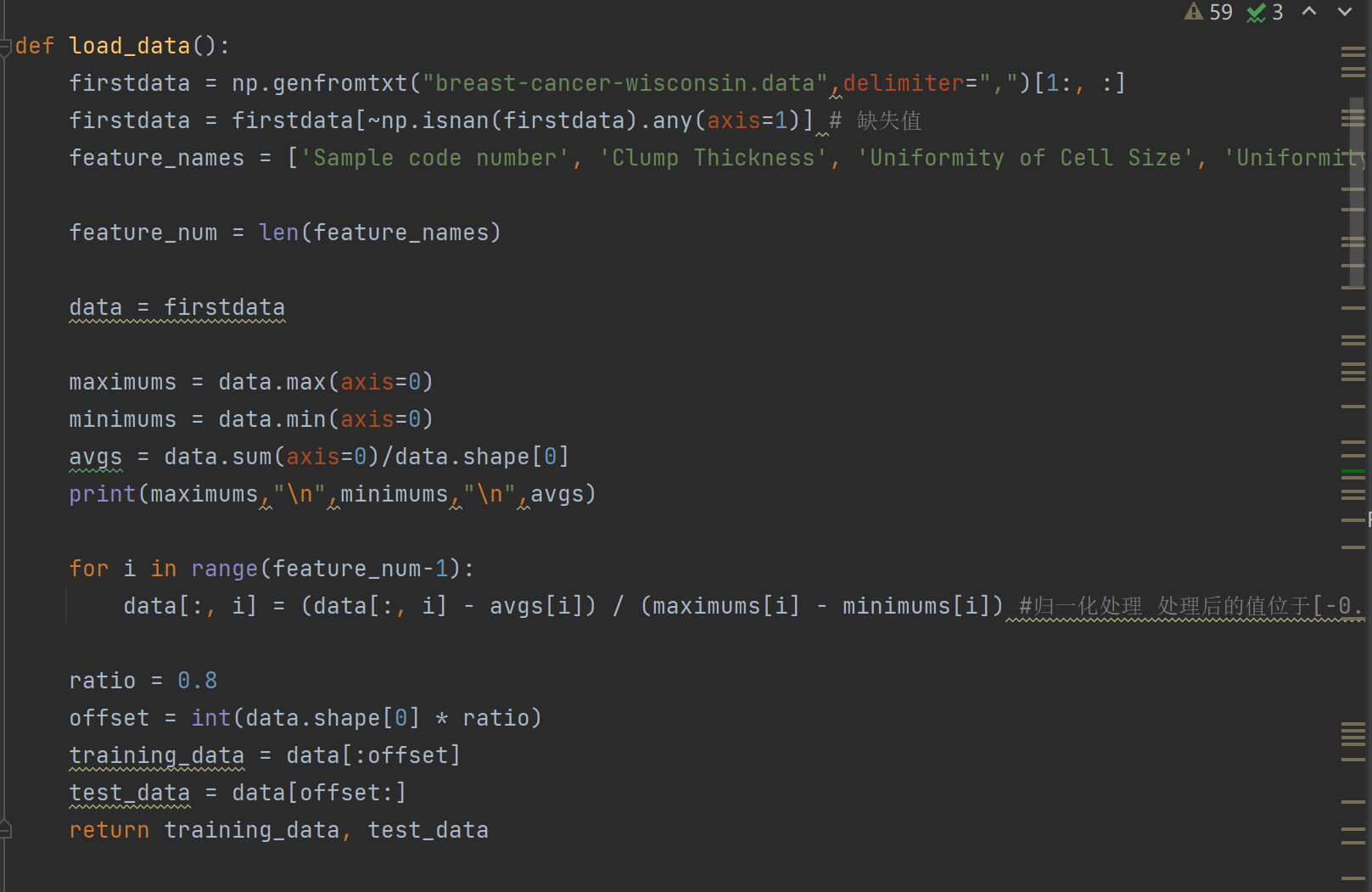
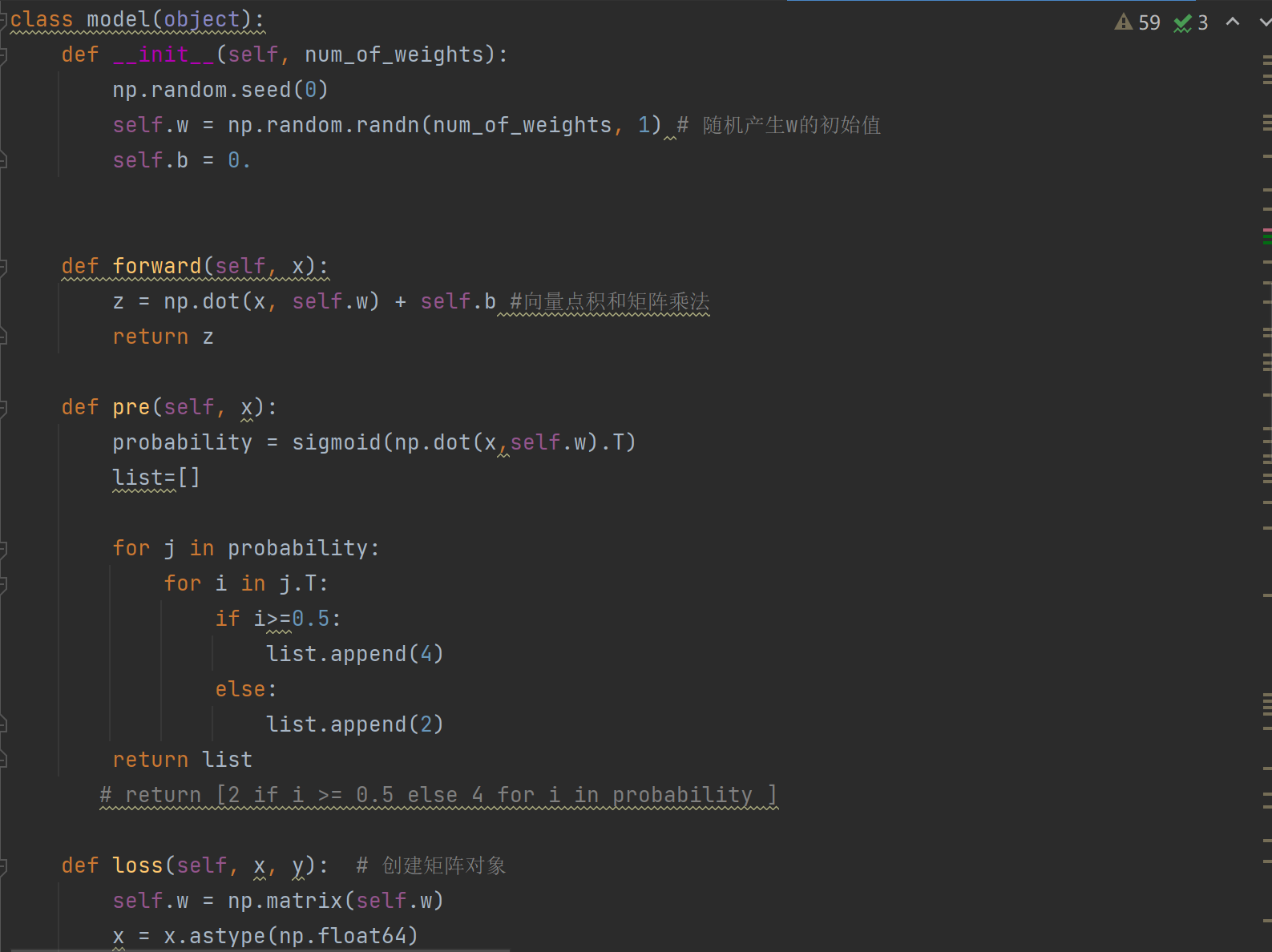
主要利用aklearn包，利用包中的数据装载、数据划分、创建模型、训练模型、预测比对、以及计算准确率。

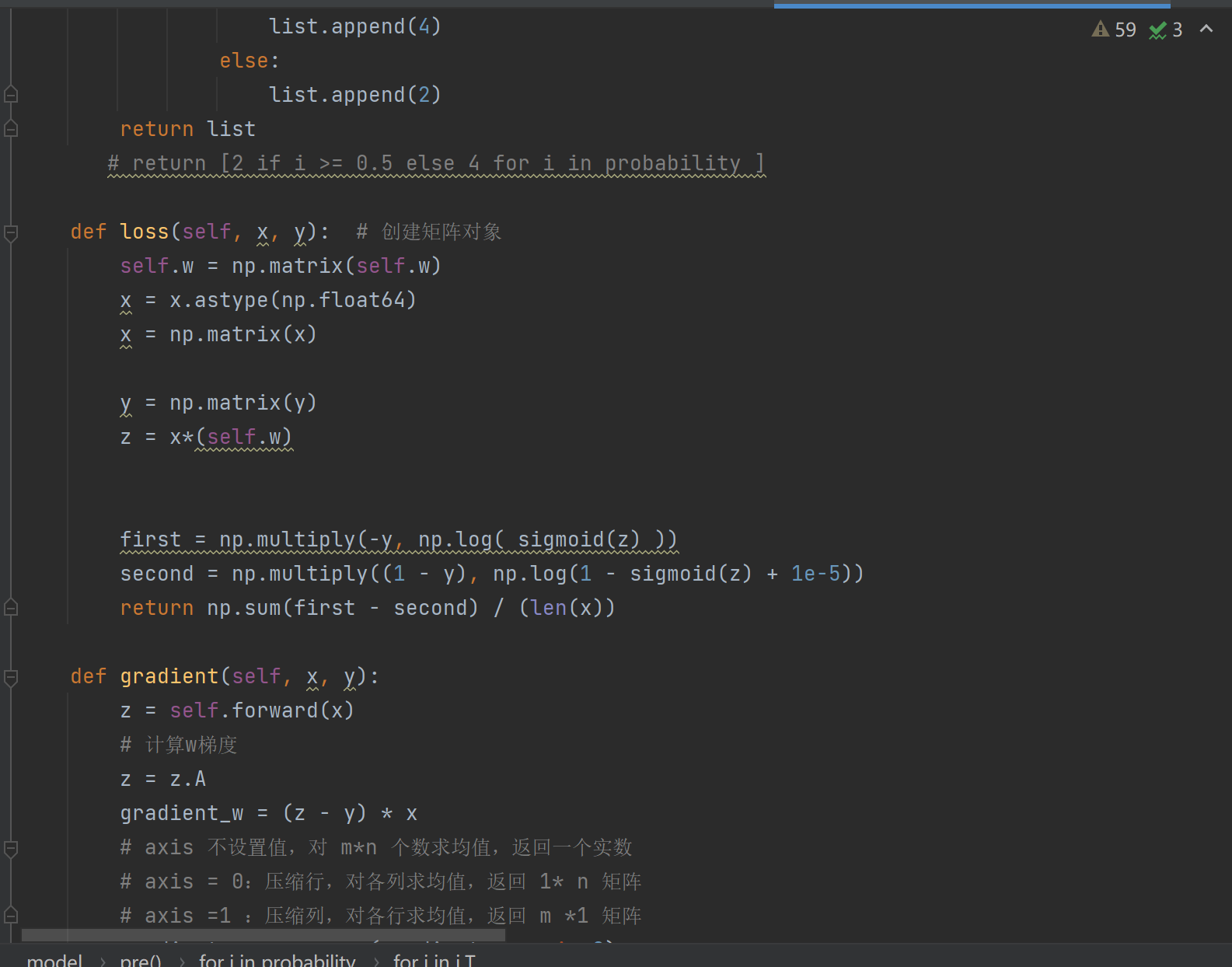
**肿瘤预测调用API部分特征：**

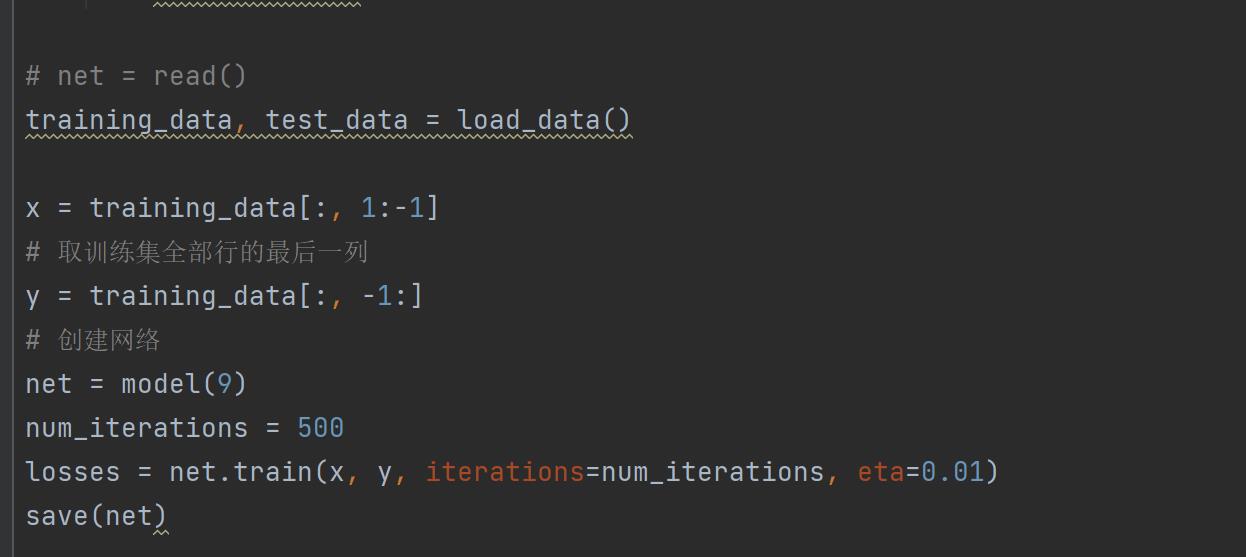


在上一个实验的基础上，调用SelectKBest函数，选取其中较好的七个特征。然后划分出新的特征集best\_x代替x完成接下来的预测。

**肿瘤预测手写代码全特征：**







利用np从文件中读取文件，标志好需要的特征及其姓名，然后对每一列特征都进行归一化处理值域为[-0.5,0.5]，划分其中80%作为训练集、20%作为数据集。

定义模型，其中包括函数的初始化，随机w的产生；定义了预测函数，即经过了sigmoid函数之后，如果最后的值大于0.5则将其归类于4的类别，反之则为2；损失函数，按照交叉熵的定义通过数学函数返回损失值；以及gradient函数计算w，b的梯度变换的计算，以及update函数更新梯度；以及训练函数，125轮，学习率0.001.以及经过计算得出准确率。

**肿瘤预测手写代码部分特征：**

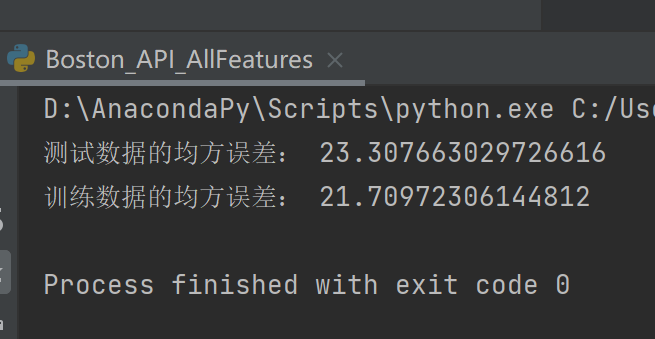
在上一个实验的基础上，在数据处理部分增加了一部分内容，利用SelectKBest函数选出了8个最合适的特征并且用best\_x列表代替x，然后继续接下来的模型训练。

1. **实验结果**

**波士顿房价预测调用API全特征：**

测试数据的均方误差： 23.307663029726616

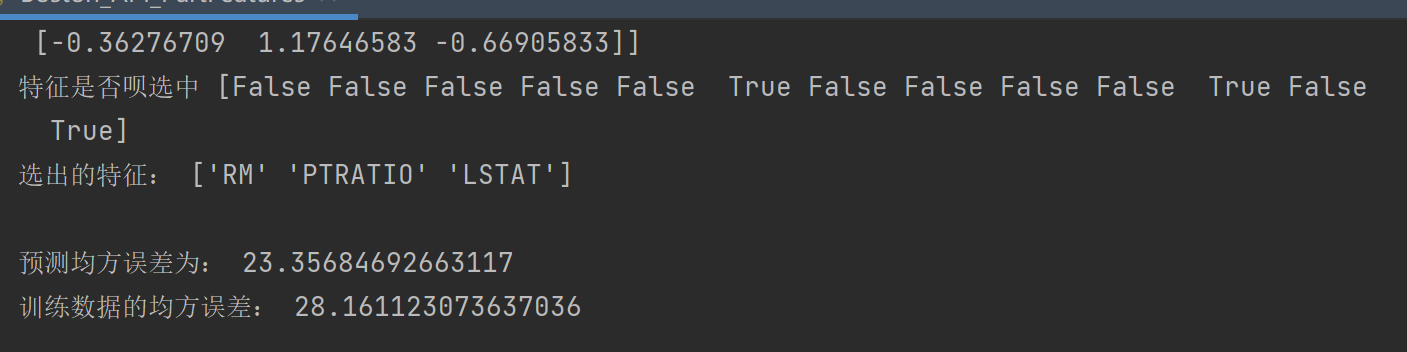
训练数据的均方误差： 21.70972306144812



**波士顿房价预测调用API部分特征：**

**预测均方误差为： 23.35684692663117**

**训练数据的均方误差： 28.161123073637036**



**波士顿房价预测手写代码全特征：**

**iter 1990, loss0.09726575790646104**

**iter 1991, loss0.09722052427317665**

**iter 1992, loss0.09717533182657757**

**iter 1993, loss0.09713018051567933**

**iter 1994, loss0.09708507028958013**

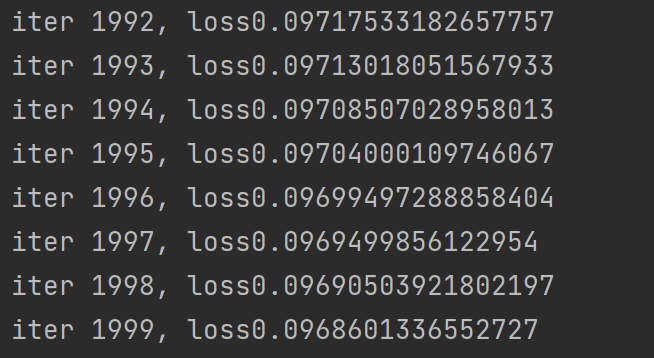
**iter 1995, loss0.09704000109746067**

**iter 1996, loss0.09699497288858404**

**iter 1997, loss0.0969499856122954**

**iter 1998, loss0.09690503921802197**

**iter 1999, loss0.0968601336552727**



**波士顿房价预测手写代码部分特征：**

**iter 1991, loss0.044948213210765746**

**iter 1992, loss0.04492762898319595**

**iter 1993, loss0.04490706564724133**

**iter 1994, loss0.044886523177790565**

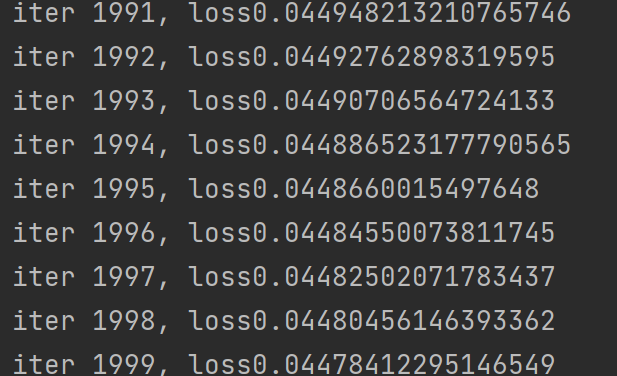
**iter 1995, loss0.0448660015497648**

**iter 1996, loss0.04484550073811745**

**iter 1997, loss0.04482502071783437**

**iter 1998, loss0.04480456146393362**

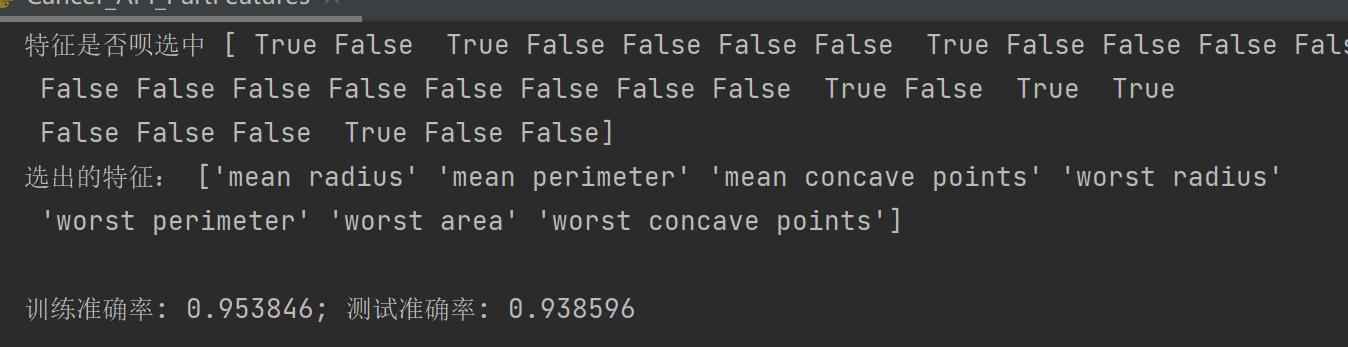
**iter 1999, loss0.04478412295146549**



**肿瘤预测调用API全特征：**

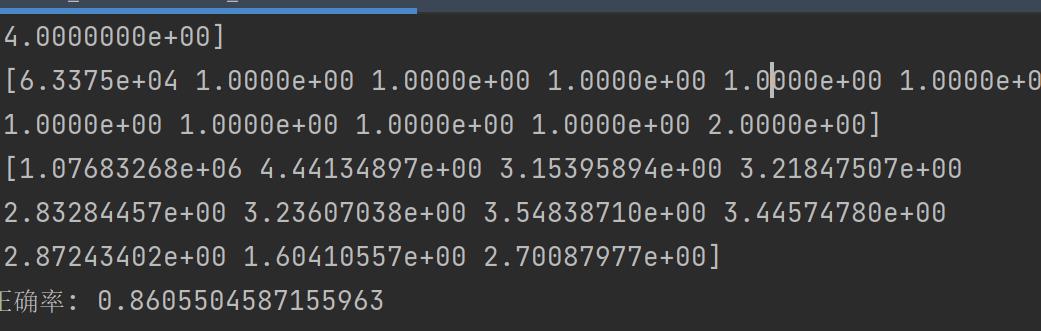
**训练准确率: 0.951648; 测试准确率: 0.947368**

**肿瘤预测调用API部分特征：**



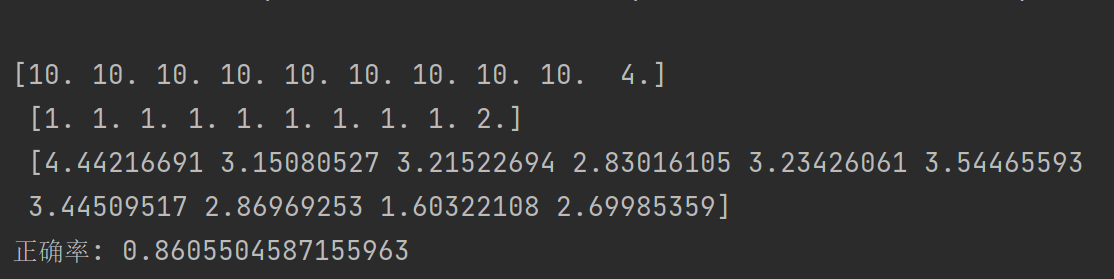
**肿瘤预测手写代码全特征：**

**正确率: 0.8605504587155963**

****

**肿瘤预测手写代码部分特征：**

**正确率: 0.8605504587155963**



1. **讨论与结论**

**经过八个python文件的实验之后，得出提下结论：**

1. **在大多数情况下，进行特征选取之后，模型的训练效果会更好，排除干扰项选取较好的特征能在一定层度上加快学习的效率，但同时特征过少也会导致模型欠拟合，难以学到内容。**
2. **API的调用非常方便，代码量，花费的时间都大大少于手写代码，并且API调用以及手写代码的差距很明显，其中手写代码线性回归更优良，API逻辑回归更优良。**
3. **逻辑回归线性回归两者看似是解决不同问题，实际上息息相关，核心思路都是非常接近，甚至在我感觉，只是最后一步的操作不同而已。**
4. **本次实验的线性回归以及逻辑回归的模型训练以及编写圆满完成，线性回归能够将均方差减少为0.04左右，逻辑回归正确率也有86%左右，感觉也算是达到机器学习的使用标准，而没有白费一场。**