## 项目背景介绍：

一家店铺开通运营一年多，流量、订单数和销售额都显著增长。经过一段时间的观察，发现网店商品的销量和广告推广的力度息息相关。店家在微信公众号推广，也通过微博推广，还在一些其他网站上面投放广告。投入推广的资金越多，则商品总销售额越多。店家想：“能不能通过机器学习算法，根据**过去记录下来的广告投放金额和商品销售额**，来预测在未来的某个节点，一个特定的广告投放金额对应能实现的商品销售额？

请用线性回归解决以上问题。

本次作业包括3个任务，完成任务1，本次作业算及格，如果完成任务1-2，本次作业可判为优秀。提交作业的时候请提交该文档和对应的源码。

注：本作业仅限于使用python实现。数据读取可使用pandas包，该数据是csv格式的。可使用如下指令读取

Import pandas as pd

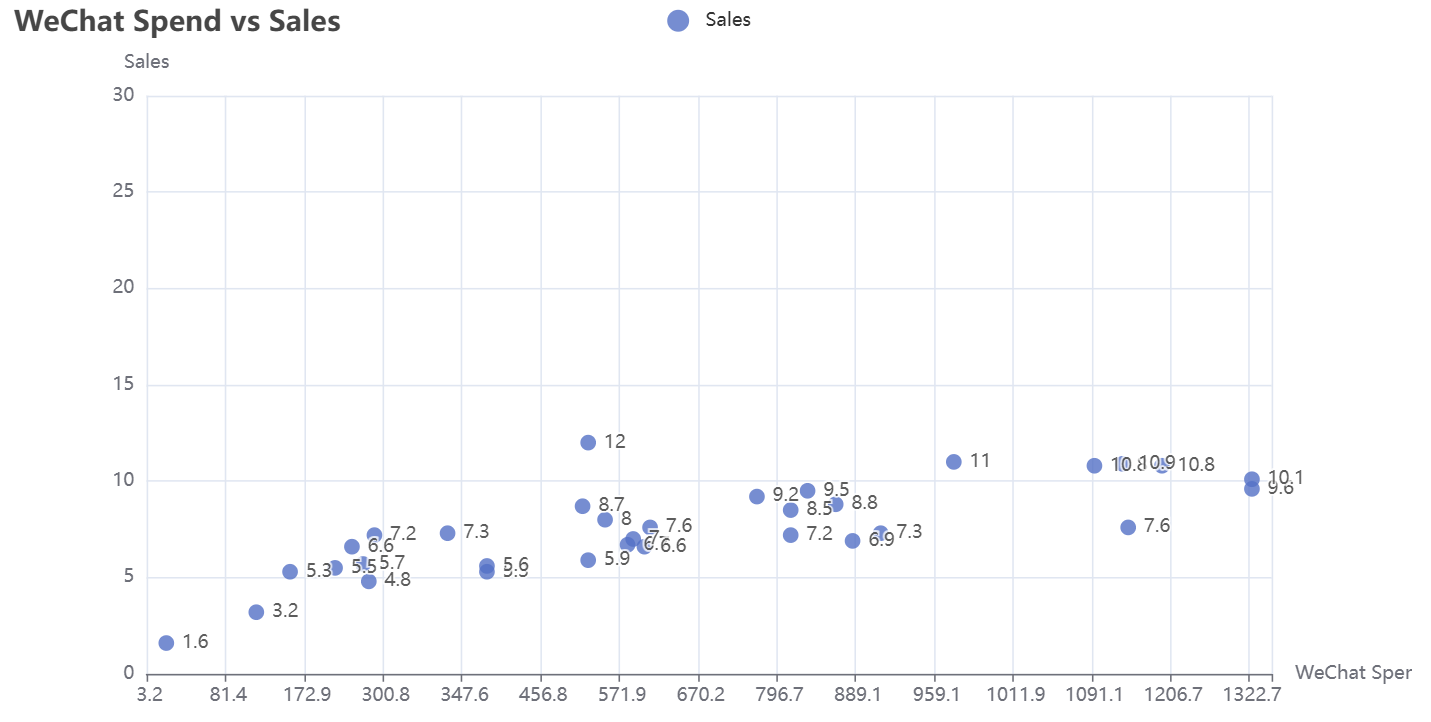
df=pd.read\_csv(“文件名字”)

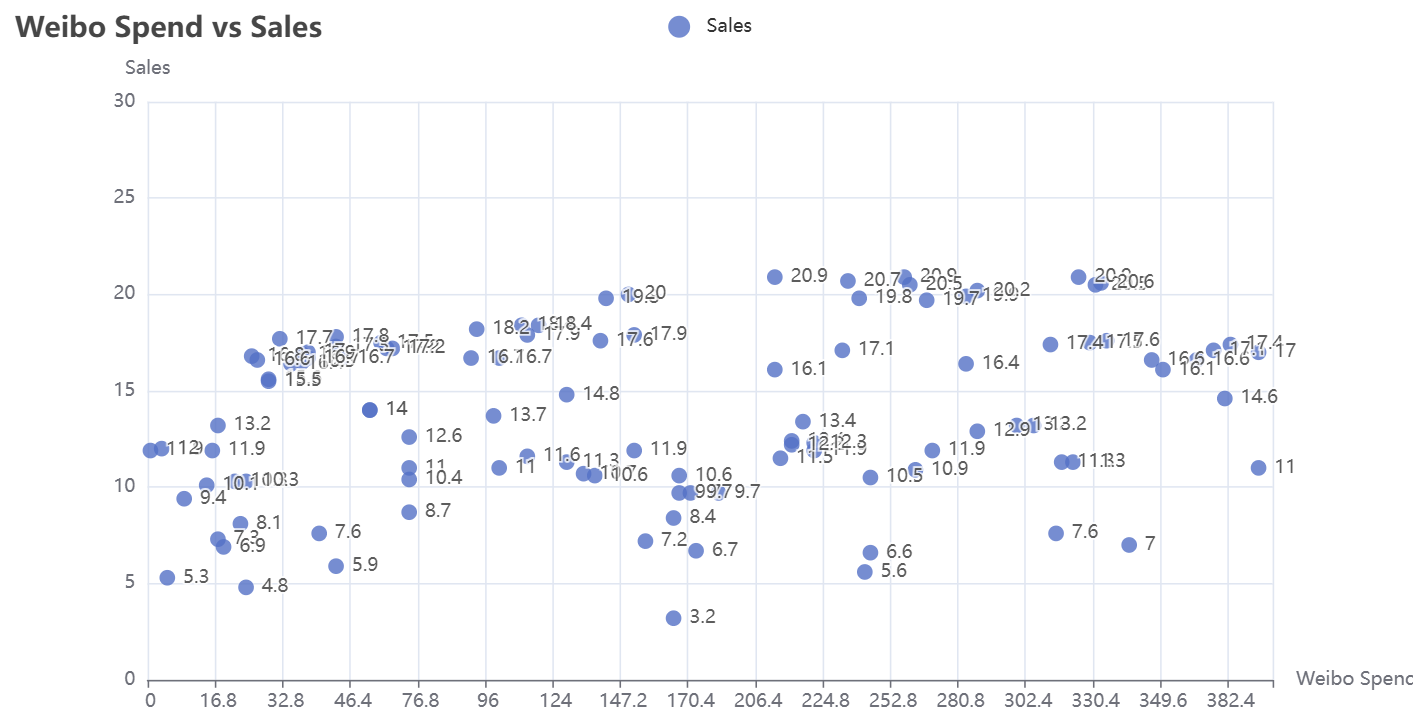
### 任务1(40’) 请使用机器学习包sklearn建立回归模型

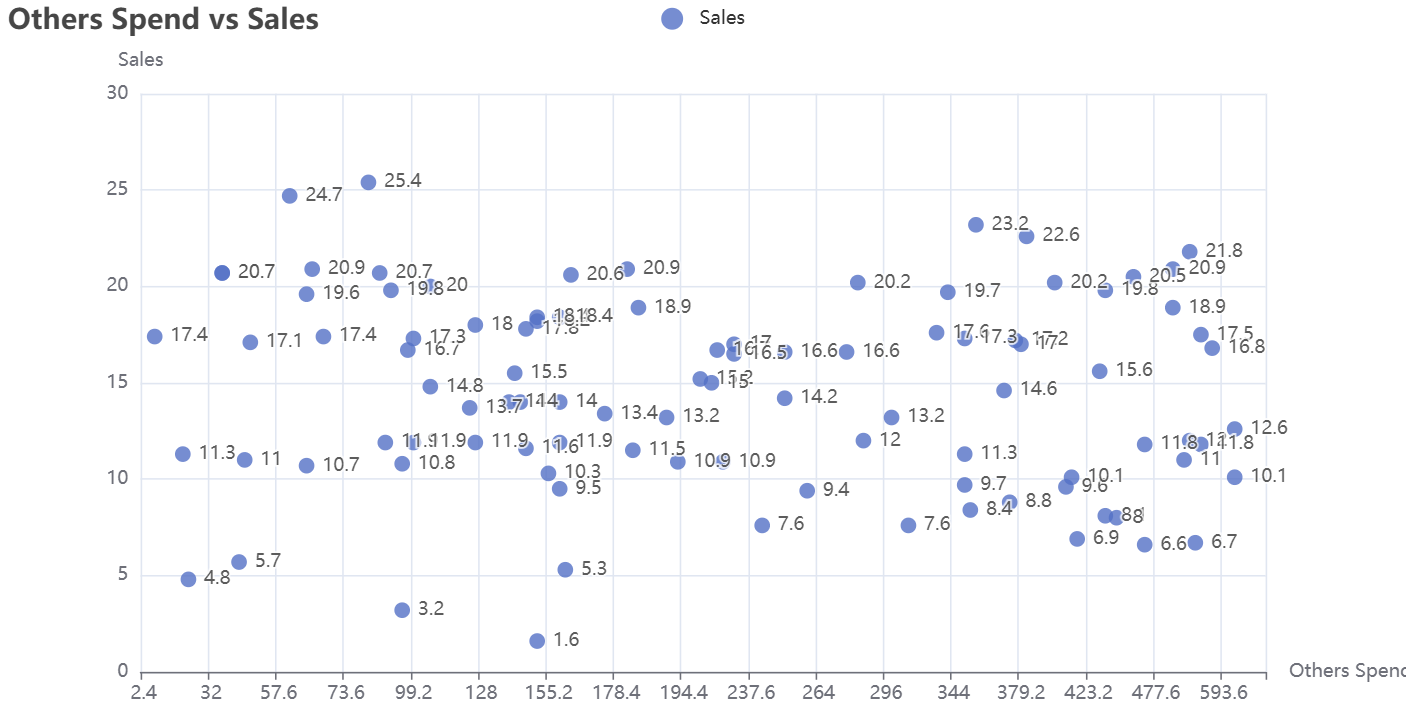
步骤：请先配置号自己的机器学习环境，然后访问<https://scikit-learn.org/stable/> 安装该机器学习包，并熟悉内置回归方法的使用方法。你需要完成一下任务

任务1.1(可视化) （10’）请分别画出以微信、微博和其他平台为x轴，销售额和y轴的数据可视化（给出代码，并展示3张图）

**代码见 Code 文件夹 data\_visualization.py**







任务1.2(10’) 请将数据集划分为70%训练，30%测试的数据(可调用函数train\_test\_split

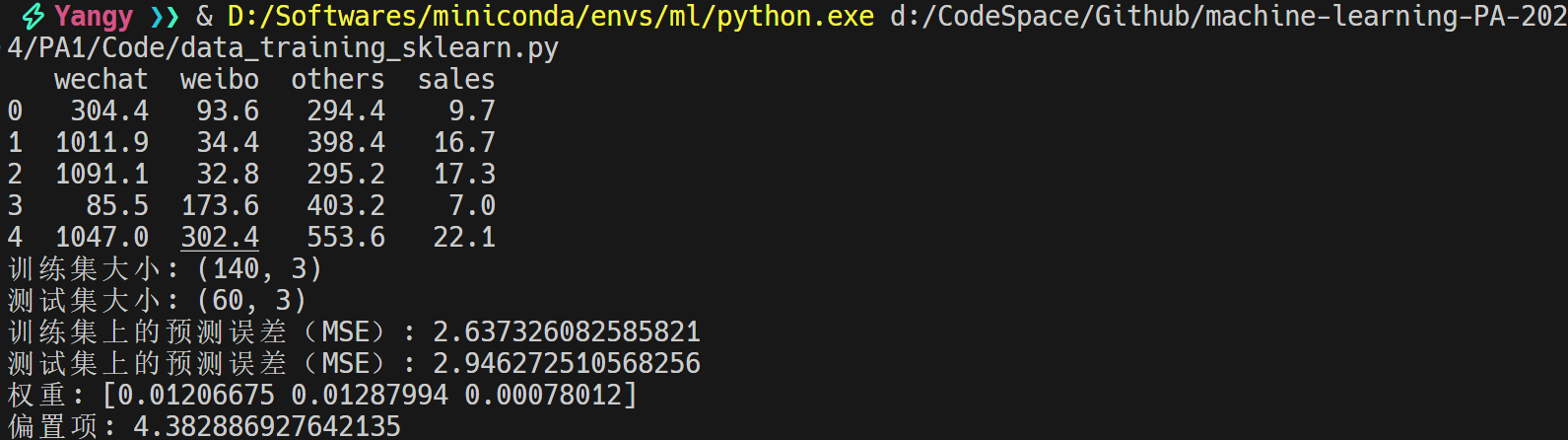
实现)

任务1.3(20’) 利用训练集训练模型，并通过测试集评估其性能。性能评估采用MSE，其计算公式为

预测误差 = （真实值-预测值）\*（真实值-预测值） （1）

利用公式(1)计算在训练集和测试集上的预测误差。并讨论如何减少在训练集和测试集上的预测误差。

代码执行结果



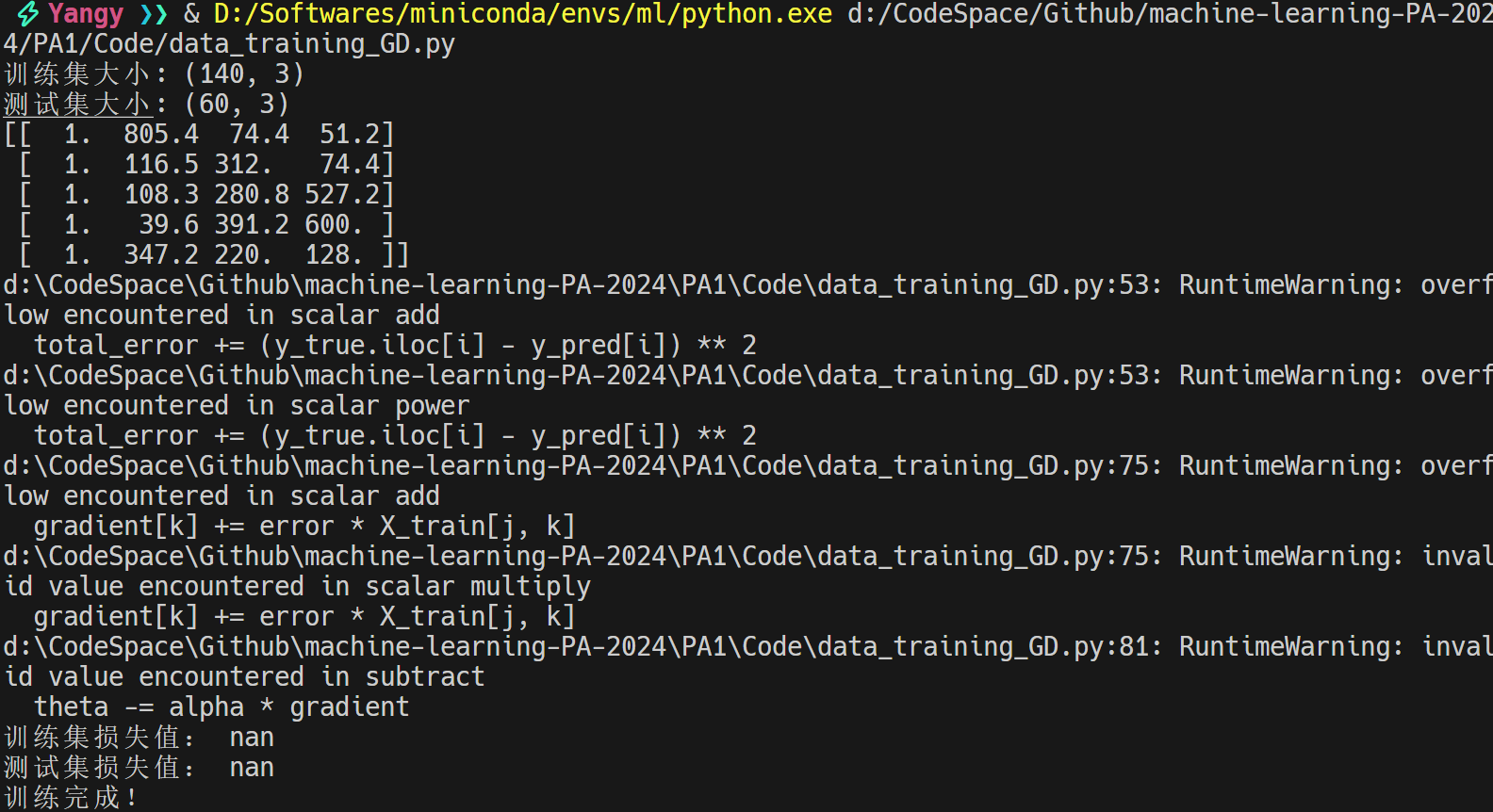
减少预测误差的可能方法有，增加迭代次数，调整学习率，进行特征工程等。

### 任务2(60’) 请使用机器学习包sklearn建立回归模型

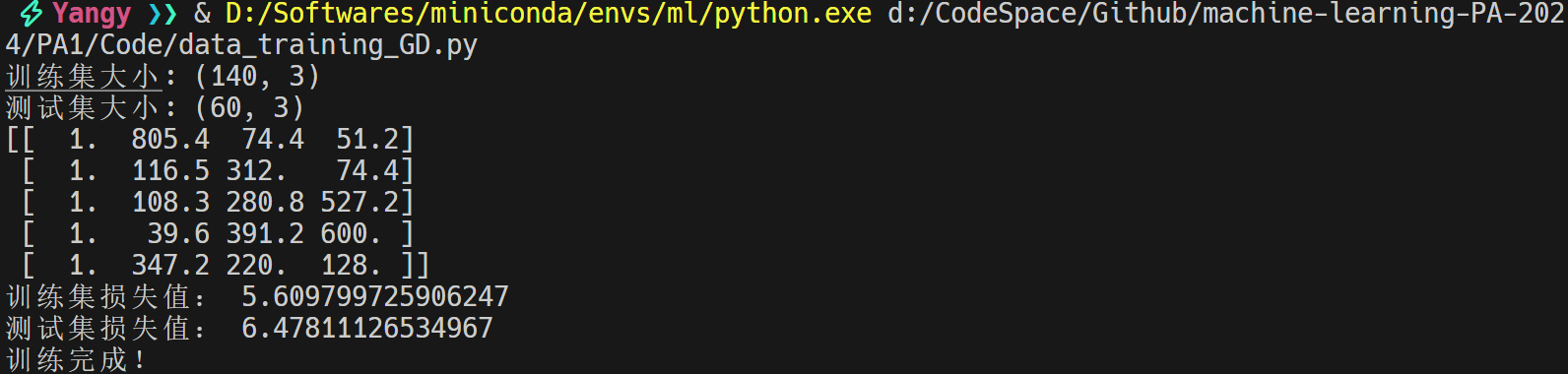
任务2.1 请编程实现用梯度下降法求解你建立的模型，并完成以下内容：

1. 梯度下降法中学习速率可能的选取范围，并给出依据。

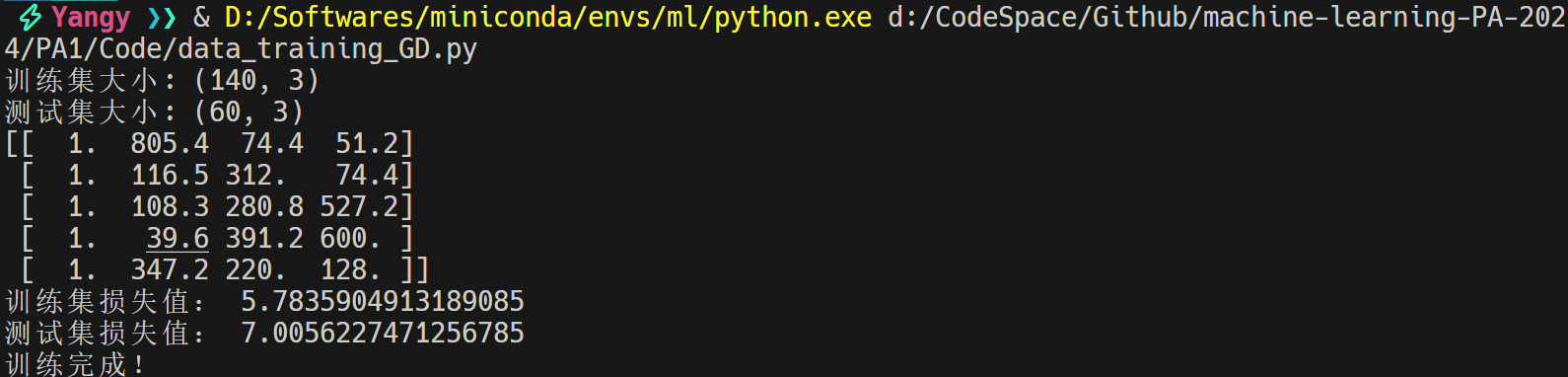
当学习率大于等于 0.00001 时，训练结果如下图所示。



当学习率为 0.000001 时，训练结果如下图所示。



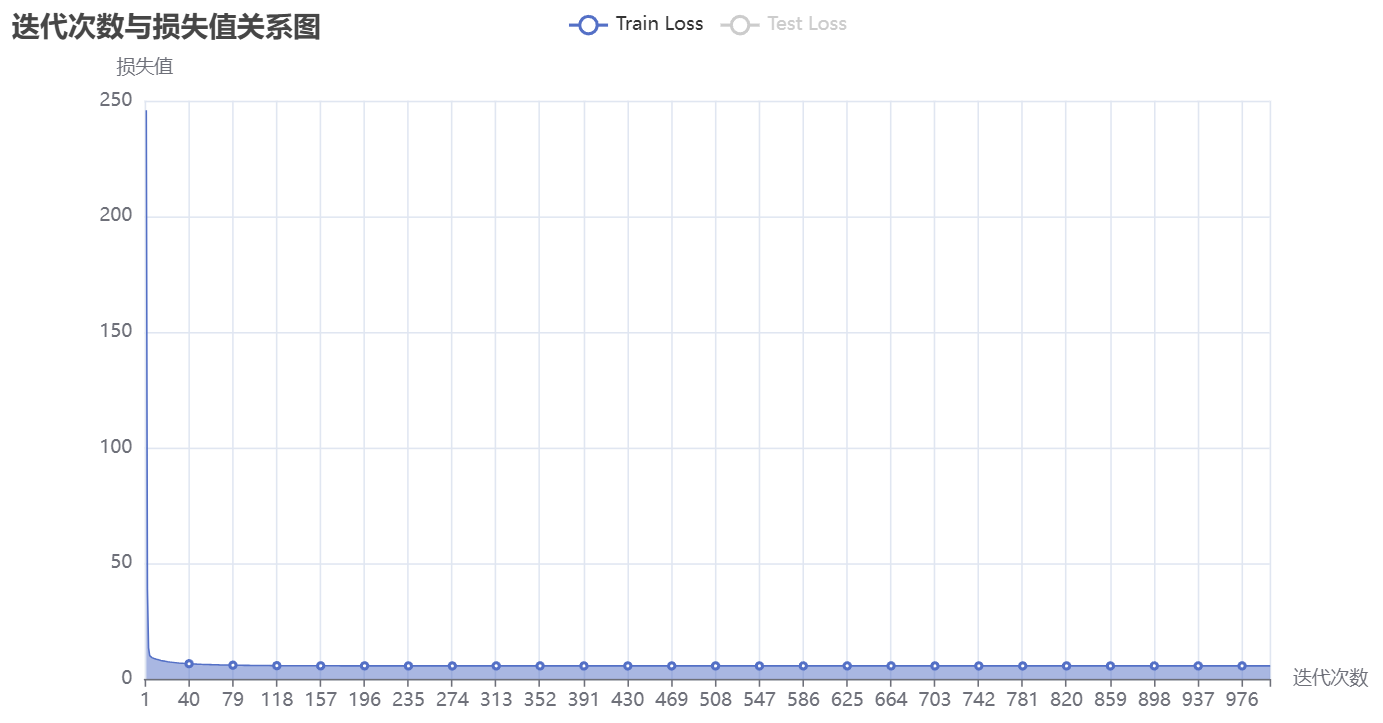
当学习率为 0.0000001 时，训练结果如下图所示。



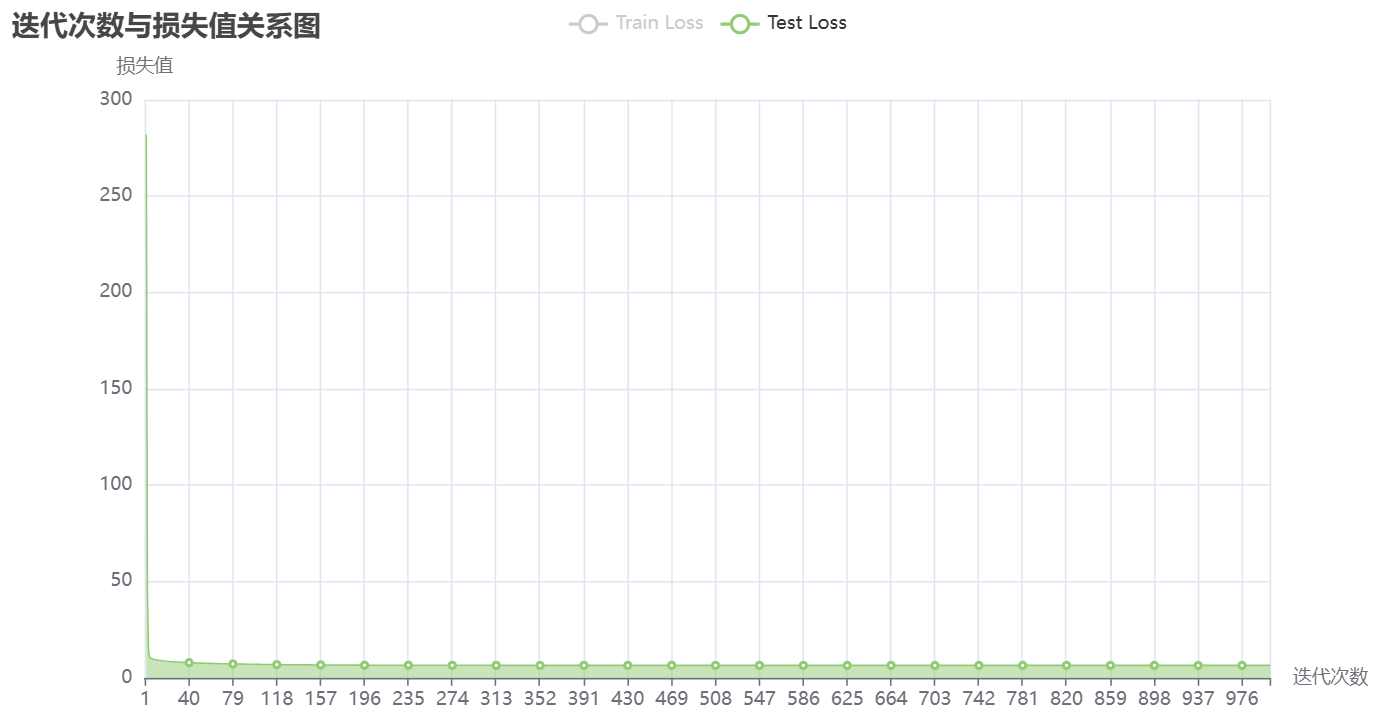
以上的迭代次数都是1000，分析可知学习率的选取范围是 学习率 <= 0.000001，当然也不能太小太小会造成训练速度过慢的问题。当学习率 > 0.000001，不能收敛，无法正常训练。

1. 请画出迭代次数和损失函数关系图.

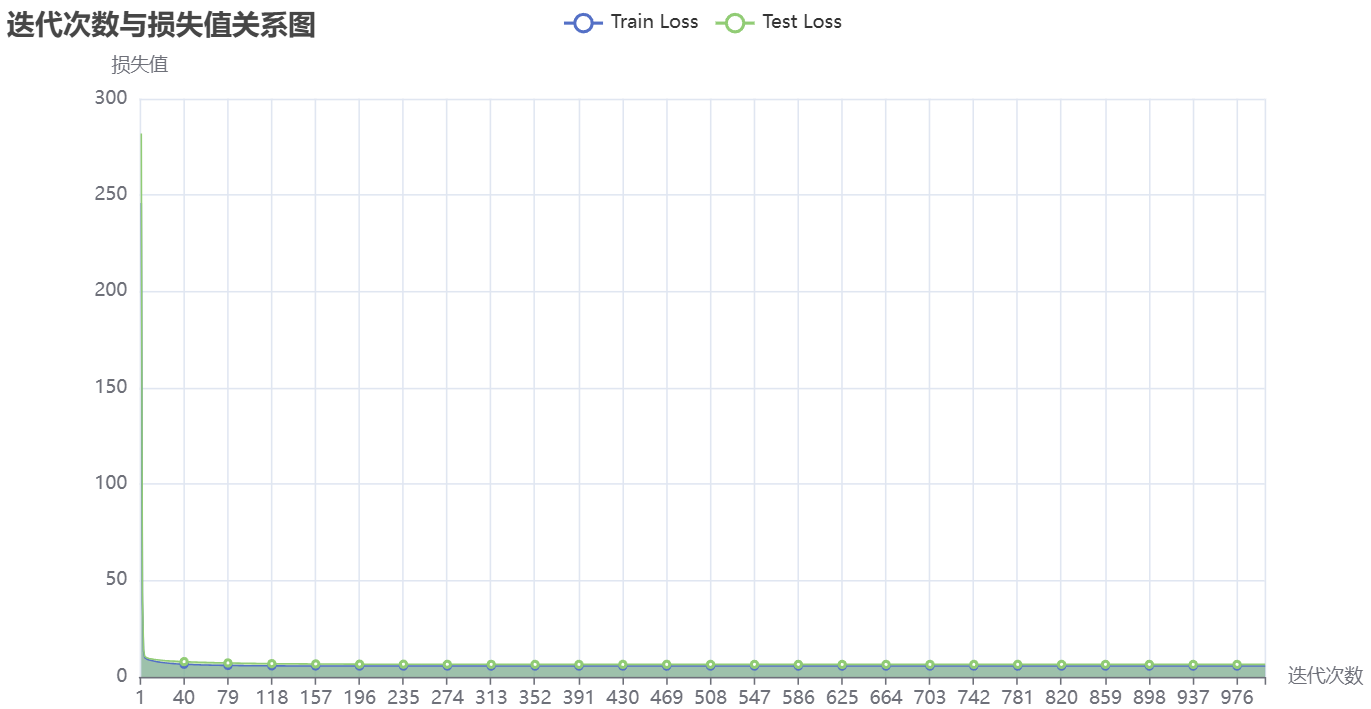
**画图代码见 Code 文件夹 data\_visualization\_GD.py**



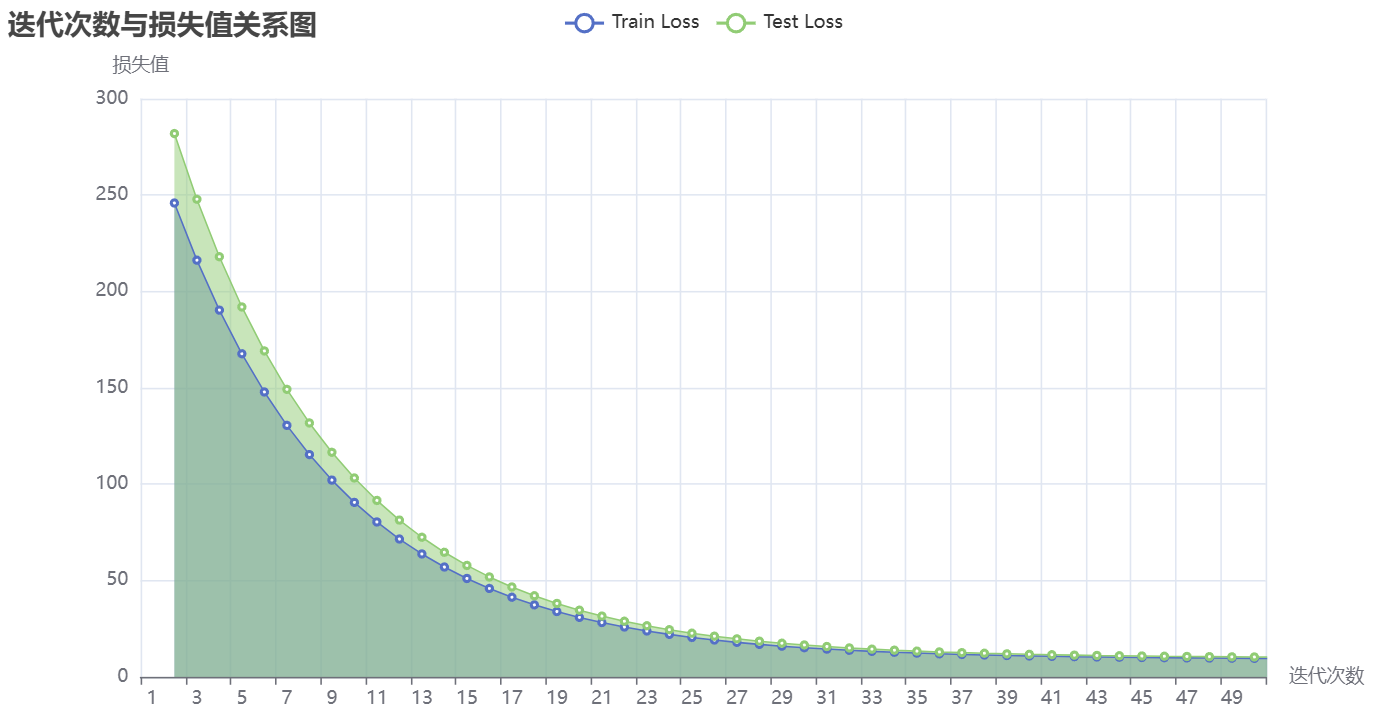
1. 请画出迭代次数和测试集预测结果之间的关系。（即每迭代一次，会得到一个模型参数，利用该模型参数计算在测试集上的预测误差，如此类推，会得到所有迭代次数对应的预测误差）



下图是迭代次数与训练集和测试集预测结果的对比，可以看出在测试集上面的表现（绿色）几乎和在训练集（蓝色）上面的表现一致，说明此模型在面对没有见过的数据的时候也可以保持较高的正确率。



附一张迭代次数为 50 次的关系图，可以更明显的看出下降趋势。



任务2.2 请使用矩阵形式的线性回归模型求解该问题。

**代码见 Code 文件夹 data\_training\_GD\_matrix.py**