## 实验2：基于回归分析的大学综合得分预测实验报告

陈泓宇 2022/7/4

### 实验任务

本次实验的任务是：根据所给大学综合情况及得分数据，使用线性回归模型，进行得分预测。

数据分析

在本次实验给出的数据中，有共9879组数据，每个数据对应有一个标签（blueWins）和38个特征。其中红队与蓝队特征各19项。

**数据集划分**

将原始数据集以8：2划分为训练集和测试集。

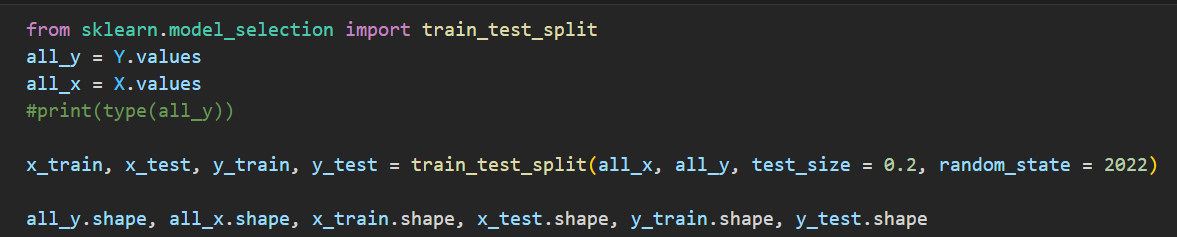
**结果分析**

对线性回归模型的系数进行分析

### 实验过程

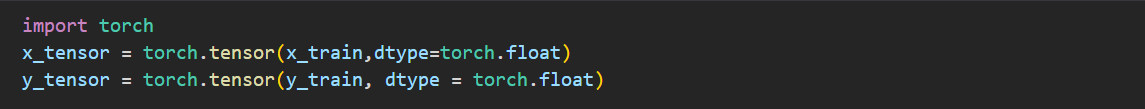
#### 数据集划分

同上次实验一样，使用sklearn.model\_selection中的train\_test\_split将数据按8：2分成训练集与测试集。

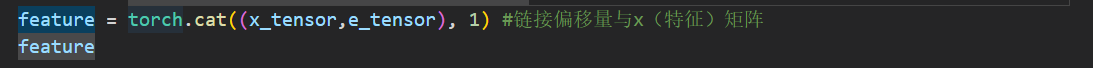


#### 数据预处理

划分好的数据集是numpy.ndarry格式的数据，为了方便之后的矩阵运算，需要将训练集的数据转为torch.tensor



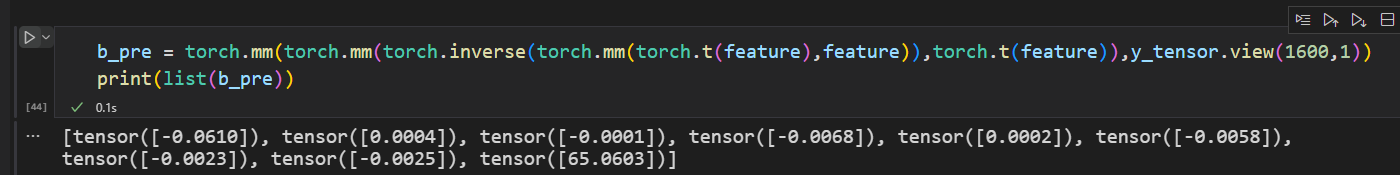
在需要求解的线性模型中，除了具有每个特征的权重系数，还有一个常数像项，在计算之前可以在原数据的特征矩阵X后加上一列单位矩阵，将它作为feature矩阵进行系数求解。



#### 模型求解

根据课堂上给出的公式（最小二乘法）：

求出各项系数



在得出的系数中，系数为正数的，说明大学的综合得分与该项正相关，反之则负相关

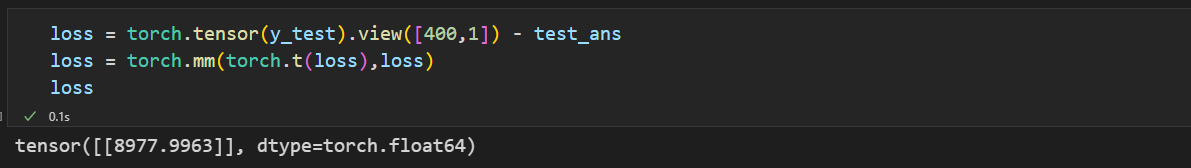
#### 结果分析与改进的尝试

将所得系数与测试集数据相乘，可以得到测试集的预测数据

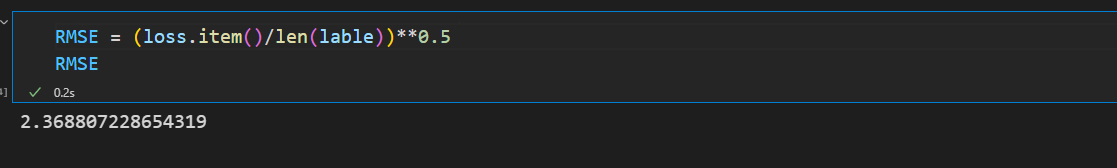


带入loss公式：

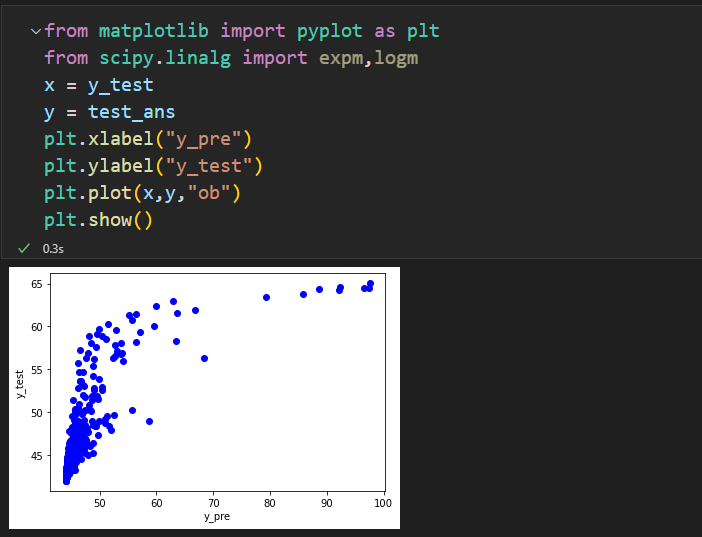
得到loss的值为8977



计算均方根误差RMSE

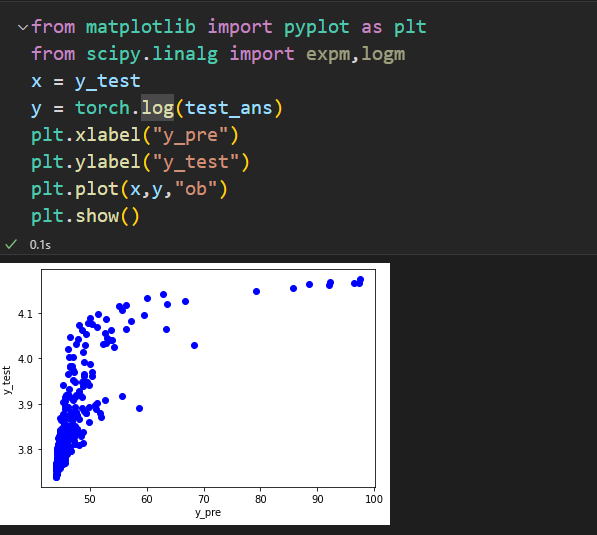


由于没有对比的loss及RMSE作为参考，我使用matplotlib中的pyplot将预测值与参考值画了下来



可以看到这个图像并不是很符合“线性”。

我尝试使用对数使数据看起来更线性但是失败了



由于时间问题我尝试将训练集的标签数据取指数后重新计算系数，使用以下模型

但是失败了





