黄冈师範学院
HUANGGANG NORMAL UNIVERSITY

人工智能与机器学习

Artificial Intelligence and Machine Learning

章节：实验5-线性回归模型之波士顿房价预测

教师：刘重

学院：计算机学院

厚德 博学 力行 致远

一、实验目的

- 1） 了解波士顿房价数据集
- 2） 掌握线性回归的原理
- 3） 利用多元线性回归模型建立一个预测房屋价值的模型，给出线性回归的指标，画出数据图

一、波士顿房价数据集介绍

二、实验步骤

1.数据分析

2.可视化处理特殊异常特征信息值（共14幅散点图）

3.导入线性回归模型进行训练

三、实验结果分析

- 所谓多元问题，就是输入有d个变量，如前述影响薪资水平的因素包括城市、学历、年龄和经验等。为方便矩阵化的最小二乘法的推导，可将参数w和b合并为向量表达形式$\hat{\boldsymbol{w}} = (\boldsymbol{w}; b)$。训练数据集D的输入部分可表示为一个$m \times (d+1)$维的矩阵$\boldsymbol{X}$，其中$d$为输入变量的个数。则矩阵$\boldsymbol{X}$和$y$可表示为：

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} x_1^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ x_m^{\mathrm{T}} & 1 \end{pmatrix} \qquad \boldsymbol{y} = (y_1; y_2; \ldots; y_m)$$

- 参数优化目标函数的矩阵化表达式为：

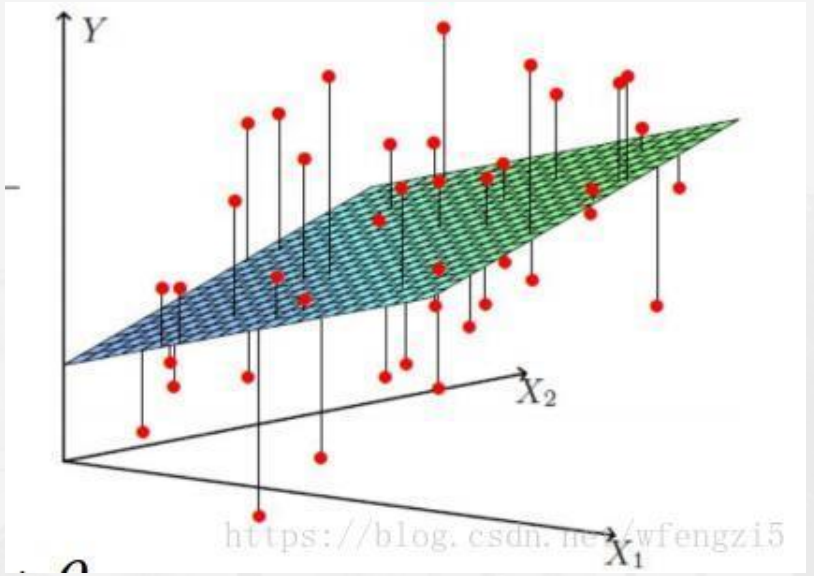
$$\hat{\boldsymbol{w}}^* = \arg\min (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})$$

- 令$L = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})$，对参数$\hat{\boldsymbol{w}}$求导，其推导过程如下：

$$L = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + \hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{w}}$$

- 求梯度

$$\frac{\partial L}{\partial \hat{\boldsymbol{w}}} = \frac{\partial \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}{\partial \hat{\boldsymbol{w}}} - \frac{\partial \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{w}}}{\partial \hat{\boldsymbol{w}}} - \frac{\partial \hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}}{\partial \hat{\boldsymbol{w}}} + \frac{\partial \hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{w}}}{\partial \hat{\boldsymbol{w}}}$$



- 最终学习得到的线性回归模型为：

$$\hat{\boldsymbol{w}}^* = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \qquad f(\hat{x}_i) = \hat{x}_i^T\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

四、波士顿房价数据集介绍

- 波士顿房价数据集统计的是20世纪70年代中期波士顿郊区房价的中位数，统计了城镇人均犯罪率、不动产税等共计13个指标，506条房价数据，通过统计出的房价，试图能找到那些指标与房价的关系。数据集中的每一行数据都是对波士顿周边或城镇房价的情况描述，下面对数据集变量进行说明，方便大家理解数据集变量代表的意义。

变量名	变量描述
CRIM	城镇人均犯罪率
ZN	住宅地超过25000平方英尺的比例
INDUS	城镇非零售商用土地的比例
CHAS	查理斯河空变量（如果边界是河流，则为1，否则为0）
NOX	一氧化碳浓度
RM	住宅平均房间数
AGE	1940年之前建成的自用房屋比例
DIS	到波士顿五个中心区区域的加权距离
RAD	辐射性公路的接近指数
TAX	每10000美元的全值财产产税率
PTRATIO	城镇师生比例
B	城镇中黑人的比例
LSTAT	人口中地位低下者的比例
target	自住房的平均房价，以千美元计

- 1.数据分析
- 首先导入数据集，对数据进行分析：

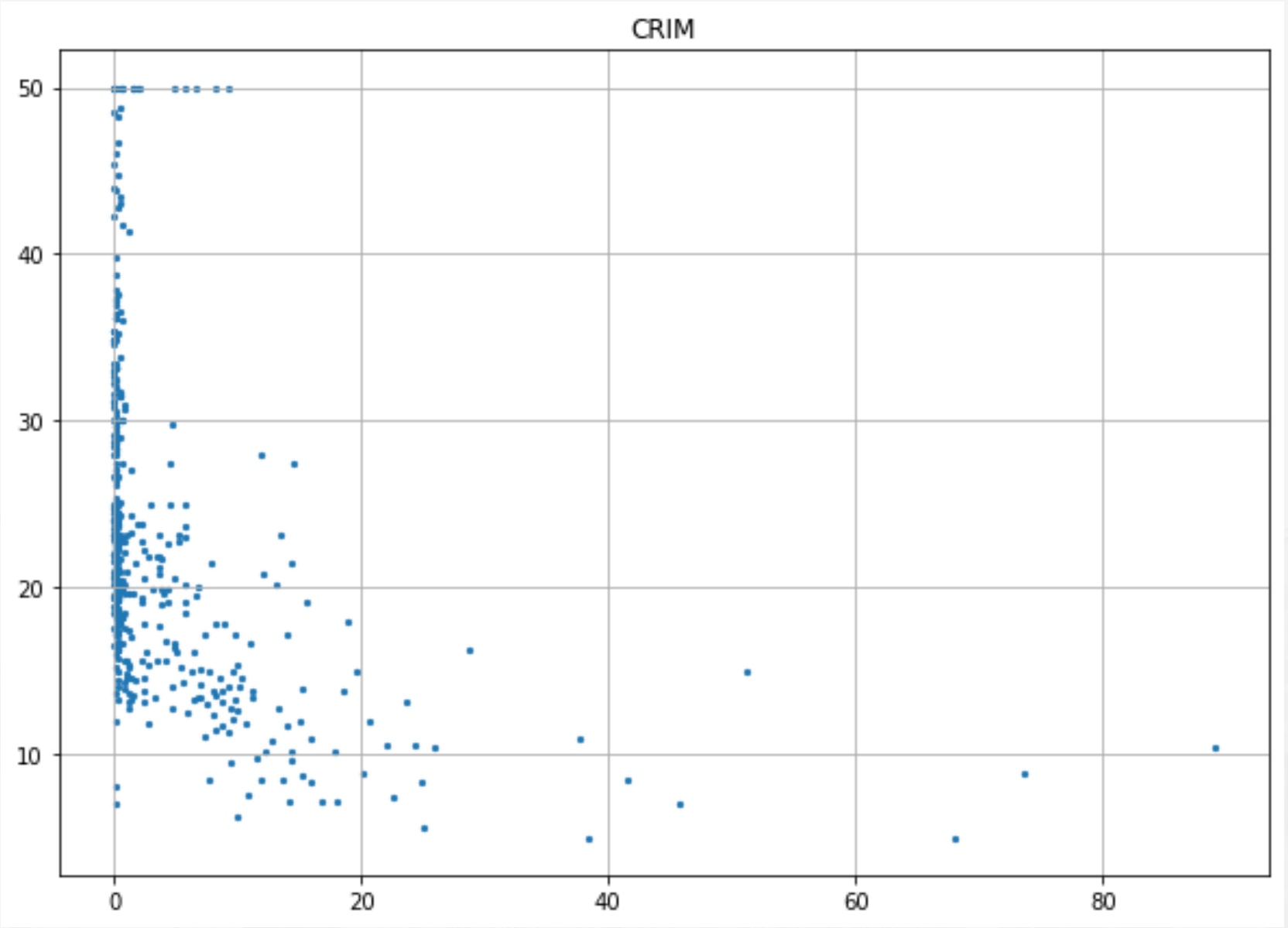
```
 8  import pandas as pd
 9  import numpy as np
10  from sklearn.datasets import load_boston  # 导入数据集
11  import matplotlib.pyplot as plt
12  from matplotlib.pyplot import MultipleLocator
13
14  boston = load_boston()
15  print(boston.feature_names)  # 查看boston数据集特征变量
16  print(boston.data.shape)  # 分析数据集样本总数，特征变量总数
17  v_bos = pd.DataFrame(boston.data)  # 查看波士顿数据集前5条数据，查看这13个变量数据情况
18  print(v_bos.head(5))
19
```

```
In [1]: runfile('C:/Users/liuzh/Desktop/波士顿房价预测.py', wdir='C:/Users/
['CRIM' 'ZN' 'INDUS' 'CHAS' 'NOX' 'RM' 'AGE' 'DIS' 'RAD' 'TAX' 'PTRATIO'
 'B' 'LSTAT']
(506, 13)
        0     1     2    3      4   ...   8      9     10      11    12
0  0.00632  18.0  2.31  0.0  0.538  ...  1.0  296.0  15.3  396.90  4.98
1  0.02731   0.0  7.07  0.0  0.469  ...  2.0  242.0  17.8  396.90  9.14
2  0.02729   0.0  7.07  0.0  0.469  ...  2.0  242.0  17.8  392.83  4.03
3  0.03237   0.0  2.18  0.0  0.458  ...  3.0  222.0  18.7  394.63  2.94
4  0.06905   0.0  2.18  0.0  0.458  ...  3.0  222.0  18.7  396.90  5.33

[5 rows x 13 columns]
```

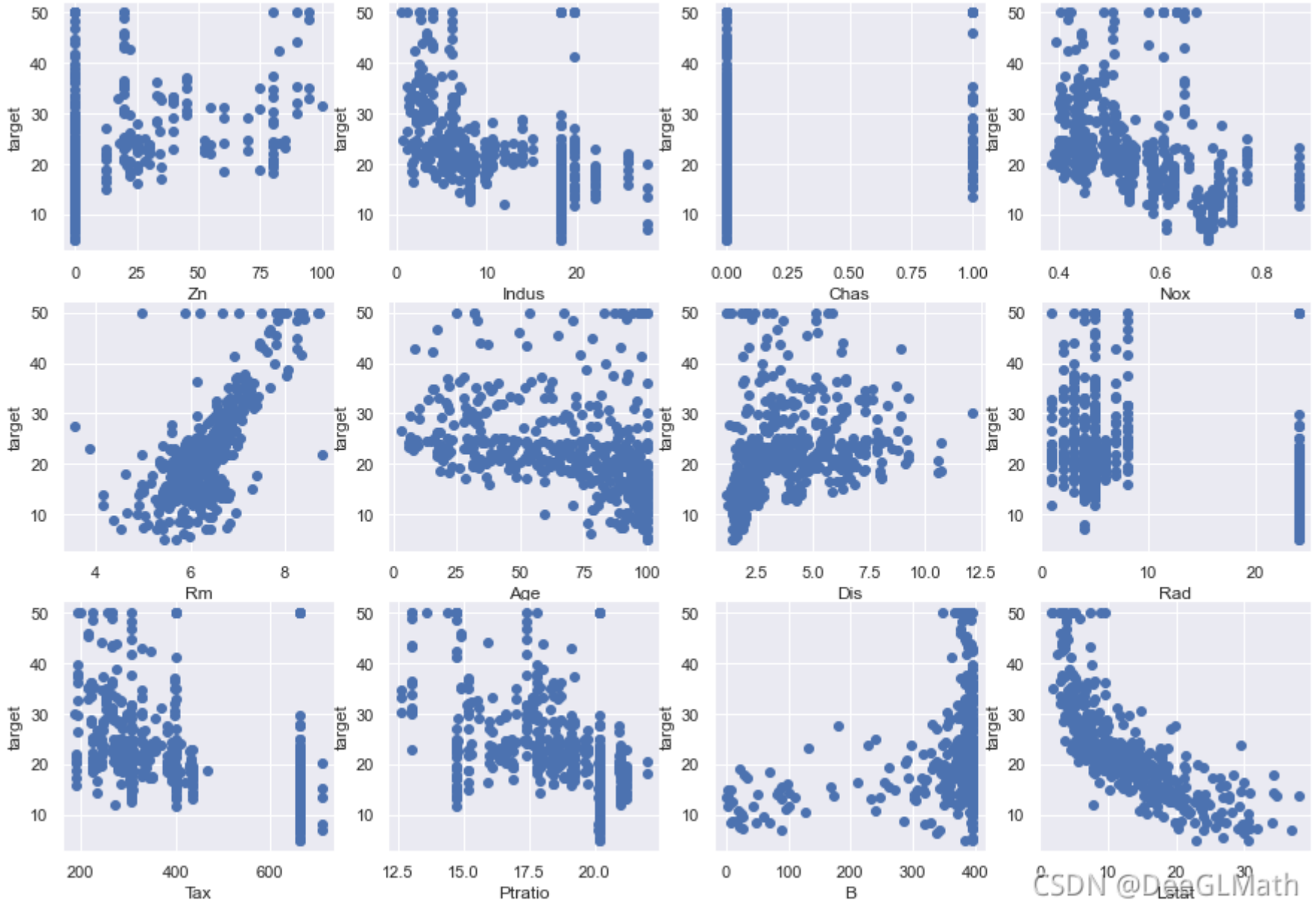
根据程序输出结果，查看数据集数据样本总数，与特征变量个数；以及通过数据集前5条数据，查看13个特征变量数据情况。

- 2.可视化处理特殊异常特征信息值（共14幅散点图）
- 然后对自变量进行特征分析，并画出散点图，分析特征变量与房价之间的相关性，把不相关的数据进行剔除。

```
x = boston['data']      # 导入特征变量
y = boston['target']    # 导入目标变量房价

name = boston['feature_names']
for i in range(13):
    plt.figure(figsize=(10, 7))
    plt.grid()
    plt.scatter(x[:, i], y, s=5)  # 横纵坐标和点的大小
    plt.title(name[i])
    print(name[i], np.corrcoef(x[:i]), y)
    plt.show()
```


CRIM
50
40
30
20
10
0
20
40
60
80

五、实验步骤



- 经过分析“房价特征信息图”，将房价大于或者等于50的数据视为异常数据，在划分训练集和测试集之前我们需要先把这些数据从数据集中进行剔除。

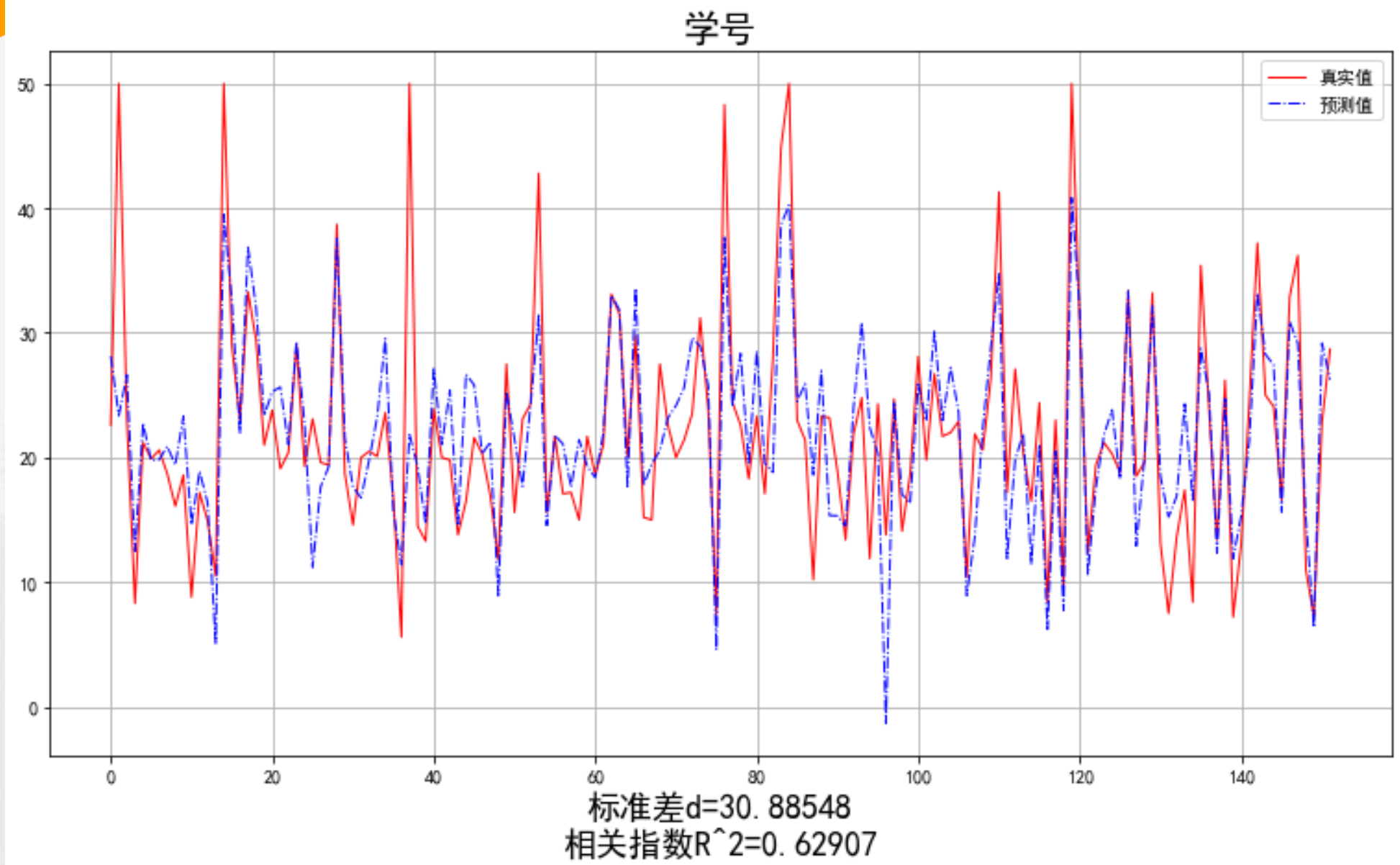
- 3.导入线性回归模型进行训练
- 接着通过上述散点图分析，对异常数据进行处理，完成数据的预处理。最后通过导入线性回归模型搭建波士顿房价预测模型

```
88    # 将数据进行拆分，一份用于训练，一份用于测试和验证
89    # 测试集大小为30%,防止过拟合
90    # 这里的random_state就是为了保证程序每次运行都分割一样的训练集和测试集。
91    # 否则，同样的算法模型在不同的训练集和测试集上的效果不一样。
92    x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=0)
93
94    # 线性回归模型
95    lf = LinearRegression()
96    lf.fit(x_train, y_train)  # 训练数据,学习模型参数
97    y_predict = lf.predict(x_test)  # 预测
98
99    # 与验证值作比较
100   error = mean_squared_error(y_test, y_predict).round(5)  # 平方差
101   score = r2_score(y_test, y_predict).round(5)  # 相关系数
```

- 4、实验结果分析

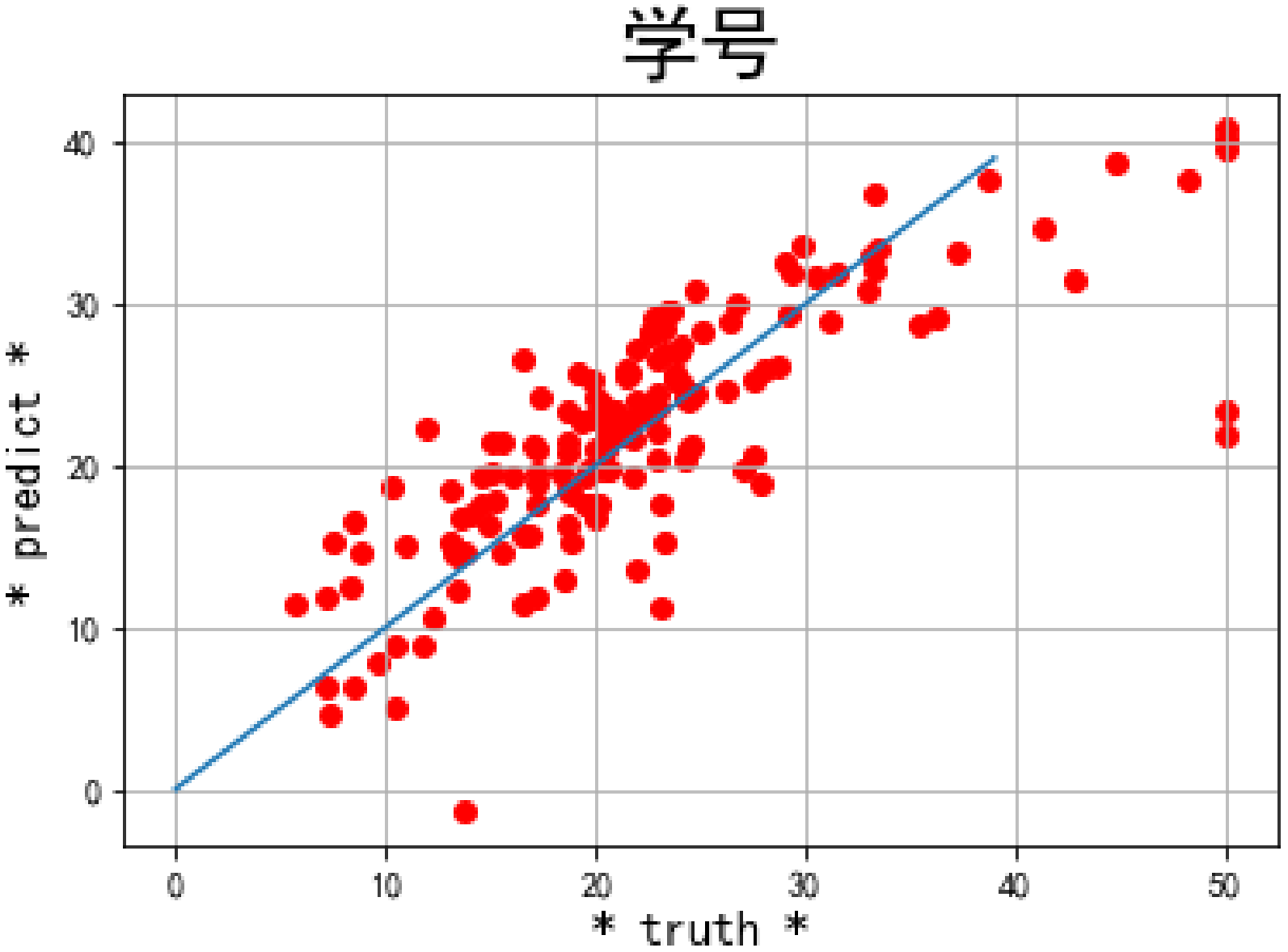
```
103    # 绘制真实值和预测值的对比图
104    fig = plt.figure(figsize=(13, 7))
105    plt.rcParams['font.family'] = "sans-serif"
106    plt.rcParams['font.sans-serif'] = "SimHei"
107    plt.rcParams['axes.unicode_minus'] = False  # 绘图
108    plt.plot(range(y_test.shape[0]), y_test, color='red', linewidth=1, linestyle='-')
109    plt.plot(range(y_test.shape[0]), y_predict, color='blue', linewidth=1, linestyle='dashdot')
110    plt.legend(['真实值', '预测值'])
111    plt.title("学号", fontsize=20)
112    error = "标准差d=" + str(error)+"\n"+"相关指数R^2="+str(score)
113    plt.xlabel(error, size=18, color="black")
114    plt.grid()
115    plt.show()
116
117    plt2.rcParams['font.family'] = "sans-serif"
118    plt2.rcParams['font.sans-serif'] = "SimHei"
119    plt2.title('学号', fontsize=24)
120    xx = np.arange(0, 40)
121    yy = xx
122    plt2.xlabel('* truth *', fontsize=14)
123    plt2.ylabel('* predict *', fontsize=14)
124    plt2.plot(xx, yy)
125    plt2.scatter(y_test, y_predict, color='red')
126    plt2.grid()
127    plt2.show()
```

六、运行截图



r2_score()函数可以表示特征模型对特征样本预测的好坏，即确定系数。根据预测值和真实值的对比图，如果其中线性回归模型的决定系数为0.60，说明线性关系可以解释房价的60%。

六、运行截图



其中，每个点的横坐标表示同一类房屋真实价格，纵坐标表示线性回归模型根据特征预测的结果，当二者值完全相等的时候就会落在红色实线上。所以模型预测得越准确，则点离红色实线越近。

六、实验报告要求

- 1、实验目的
- 2、实验内容
- 3、实验原理
- 4、实验代码
- 5、运行结果与分析
- 6、实验小结

- 说明：每个学生都要交电子版的实验报告，命名格式：
- 01/02-XXXX（学号）-XXX（姓名）

黄冈师範学院
HUANGGANG NORMAL UNIVERSITY

Q&A

厚德 博学 力行 致远