Python 程序设计实验报告

波士顿房价预测

院系：人工智能学院

姓名：张运吉

学号：211300063

班级：21 级人工智能学院 AI2 班

邮箱：211300063@smail.nju.edu.cn

时间：2022 年 5 月 25 日

目录

1 问题类型：

波士顿房价预测问题是一个回归问题。因为是预测房价，需要给出一个具体数值，而不是像分类问题给出属于某个类别。

2 初始模型：

首先使用sklearn库中的线性回归模型。

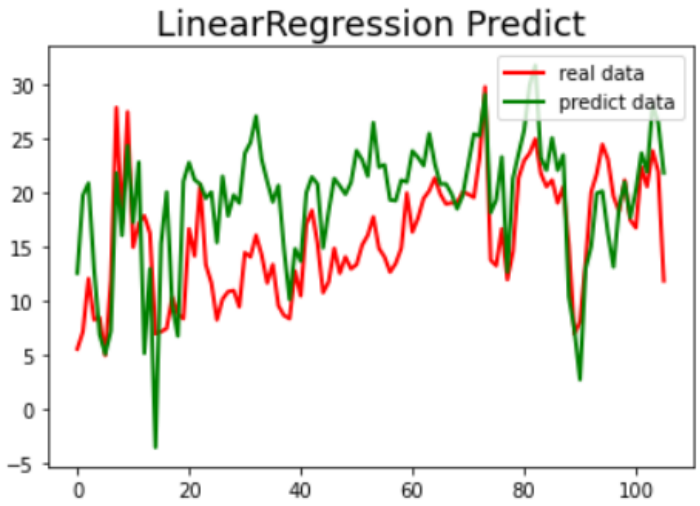
得到的结果如下图所示：

```
In [9]: error=np.sqrt(mean_squared_error(y_test, pred))
        print(" Linear Regression Error:",error)

        Linear Regression Error: 6.1557922804137615
```

```
In [10]: t = range(len(y_test))
         plt.plot(t, y_test, 'r-', linewidth=2, label='real data')
         plt.plot(t, pred, 'g-', linewidth=2, label='predict data')
         plt.legend(loc='upper right')
         plt.title('LinearRegression Predict', fontsize=18)

Out[10]: Text(0.5, 1.0, 'LinearRegression Predict')
```



由图可以看出，预测值和真实值相差较大，预测值的曲线和实际值的曲线变化趋势大约一致，但是预测值平均比实际值大 6 左右。

这里使用的模型评估标准是计算均方误差：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3 数据降维、归一化：

通过对数据的分析，发现只有三种特征对于价格的相关性大于0.5(“rm”,”ptratio”, “lstat”)

```
corr = data.corr()['medv']
corr
```

```
crim      -0.388305
zn         0.360445
indus     -0.483725
chas       0.175260
nox       -0.427321
rm         0.695360
age       -0.376955
dis        0.249929
rad       -0.381626
tax       -0.468536
ptratio   -0.507787
black      0.333461
lstat     -0.737663
medv       1.000000
Name: medv, dtype: float64
```

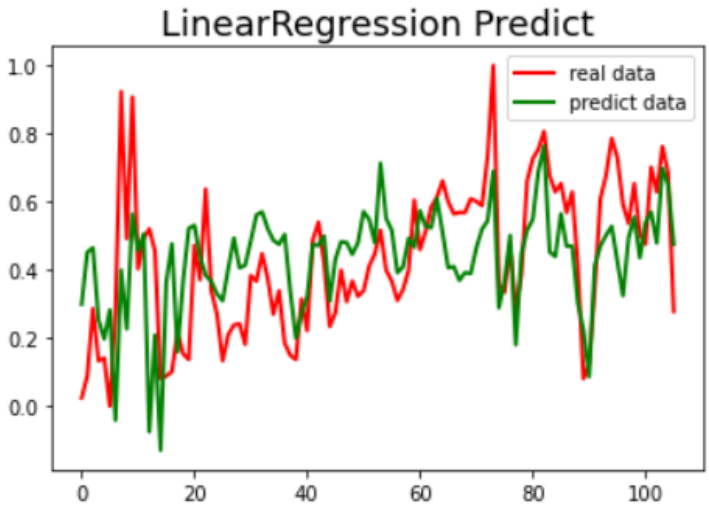
所以可以忽略其他特征以达到数据降维的目的。

由于各维属性的取值范围差别很大，这里就要用到一个常见的操作：归一化。归一化的目标是把各位属性的取值范围放缩到差不多的区间，例如[-0.5,0.5]。这里使用一种很常见的操作方法：减掉均值，然后除以原取值范围。

In [12]:
```
# 数据归一化
from sklearn import preprocessing
min_max_scaler = preprocessing.MinMaxScaler()
x1_train = min_max_scaler.fit_transform(x1_train)
y1_train = min_max_scaler.fit_transform(y1_train.to_numpy().reshape(-1,1))
x1_test = min_max_scaler.fit_transform(x1_test)
y1_test = min_max_scaler.fit_transform(y1_test.to_numpy().reshape(-1,1))
```

4. 最后效果：

```
Out[14]: Text(0.5, 1.0, 'LinearRegression Predict')
```




```
In [15]: error1=np.sqrt(mean_squared_error(y1_test, pred1))
         print(" Linear Regression1 Error:",error1)

          Linear Regression1 Error: 0.18647687469839822

In [16]: r2_score(pred1, y1_test)

Out[16]: -0.5520922147746643
```

效果比降维前好了很多，而且这个模型的决定系数为 0.5521，说明线性关系可以解释房价的 55.21%

5. 能否使用 KNN：

我觉得这个问题可以使用knn，对于一个待预测的样本，求出k个邻居后，可以以这k个邻居的房价的均值作为待预测样本的预测值。