数据科学作业2--预测房价

1.模型建立与公式推导

通过构建损失函数，来求解损失函数最小时的参数w和b：

$$\hat{y} = \omega x + b$$

y^为预测值，自变量x和因变量y是已知的，而我们想实现的是预测新增一个x，其对应的y是多少。因此，为了构建这个函数关系，目标是通过已知数据点，求解线性模型中w和b两个参数。

求解最佳参数，需要一个标准来对结果进行衡量，为此我们需要定量化一个目标函数式，使得计算机可以在求解过程中不断地优化。

针对任何模型求解问题，都是最终都是可以得到一组预测值y^ ，对比已有的真实值 y ，数据行数为 n ，可以将损失函数定义如下：

$$L = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$$

即预测值与真实值之间的平均的平方距离，统计中一般称其为MAE(mean square error)均方误差。把之前的函数式代入损失函数，并且将需要求解的参数w和b看做是函数L的自变量，可得：

$$L(\omega, b) = \frac{1}{n}\sum_{i=1}^{n}(\omega x_i + b - y_i)^2$$

梯度下降核心内容是对自变量进行不断的更新（针对w和b求偏导），使得目标函数不断逼近最小值的过程：

$$\omega \leftarrow \omega - \alpha\frac{\partial L}{\partial \omega}$$

$$b \leftarrow b - \alpha\frac{\partial L}{\partial b}$$

$$(\omega^*, b^*) = argmin(\omega, b)\sum_{i=1}^{n}(\omega x_i + b - y_i)^2$$

y表示我们要求的销售价格，x表示特征值。需要调用sklearn库来进行训练。

2.数据预处理

对数据data2.csv进行合理的划分，训练数据（kc_train.csv）主要包括10000条记录，14个字段，测试数据（kc_test.csv）主要包括3000条记录，13个字段，跟训练数据的不同是测试数据并不包括房屋销售价格，需要通过由训练数据所建立的模型以及所给的测试数据，得出测试数据相应的房屋销售价格预测值。在训练过程中是不需要销售价格的，把第二列删除掉，新建一个csv文件存放销售价格这一列，作为验证数据进行后面的结果对比。

3.所用到的库

```
# 导入相关python库
import os
import numpy as np
import pandas as pd

#设定随机数种子
np.random.seed(36)

#使用matplotlib库画图
import matplotlib
import seaborn
import matplotlib.pyplot as plot

from sklearn import datasets
```

4.数据处理

之后对数据进行读取，首先先读取数据，查看数据是否存在缺失值，然后进行特征缩放统一数据维度。

```
#读取数据
housing = pd.read_csv('kc_train.csv')
target=pd.read_csv('kc_train2.csv')  #销售价格
t=pd.read_csv('kc_test.csv')         #测试数据

#数据预处理
housing.info()    #查看是否有缺失值

#特征缩放
from sklearn.preprocessing import MinMaxScaler
minmax_scaler=MinMaxScaler()
minmax_scaler.fit(housing)   #进行内部拟合，内部参数会发生变化
scaler_housing=minmax_scaler.transform(housing)
scaler_housing=pd.DataFrame(scaler_housing,columns=housing.columns)
```

5.模型训练

使用sklearn库的线性回归函数进行调用训练。梯度下降法获得误差最小值。最后使用均方误差法来评价模型的好坏程度，并画图进行比较。

```
#选择基于梯度下降的线性回归模型
from sklearn.linear_model import LinearRegression
LR_reg=LinearRegression()
#进行拟合
LR_reg.fit(scaler_housing,target)


#使用均方误差用于评价模型好坏
from sklearn.metrics import mean_squared_error
preds=LR_reg.predict(scaler_housing)   #输入数据进行预测得到结果
mse=mean_squared_error(preds,target)   #使用均方误差来评价模型好坏，可以输出mse进行查看评价值

#绘图进行比较
plot.figure(figsize=(10,7))        #画布大小
num=100
x=np.arange(1,num+1)               #取100个点进行比较
plot.plot(x,target[:num],label='target')      #目标取值
plot.plot(x,preds[:num],label='preds')        #预测取值
plot.legend(loc='upper right')  #线条显示位置
plot.show()
```

6.可视化结果

对测试数据进行输出,以进行验证。

```
#输出测试数据
result=LR_reg.predict(scaler_t)
df_result=pd.DataFrame(result)
df_result.to_csv("result.csv")
```

使用测试数据进行目标函数预测输出，观察结果是否符合预期。通过画出对比函数进行结果线条对比。最后输出的图是这样的： 从这张结果对比图中就可以看出模型是比较好地得到精确的目标函数，能够比较精确地精确预测房价。

```
#绘图进行比较
plot.figure(figsize=(10,7))        #画布大小
num=100
x=np.arange(1,num+1)               #取100个点进行比较
plot.plot(x,target[:num],label='target')       #目标取值
plot.plot(x,preds[:num],label='preds')         #预测取值
plot.legend(loc='upper right')  #
```