# 波士顿房价预测

# 数据集描述

本作品所用数据是一份源于美国某经济学杂志上，分析研究波士顿房价( Boston House Price)的数据集。数据集中的每一行数据都是对波士顿周边或城镇房价的描述：  
 CRIM: 城镇人均犯罪率  
 ZN: 住宅用地所占比例  
 INDUS: 城镇中非住宅用地所占比例  
 CHAS: CHAS 虚拟变量,用于回归分析  
 NOX: 环保指数  
 RM: 每栋住宅的房间数  
 AGE: 1940 年以前建成的自住单位的比例  
 DIS: 距离 5 个波士顿的就业中心的加权距离。  
 RAD: 距离高速公路的便利指数  
 TAX: 每一万美元的不动产税率  
 PRTATIO: 城镇中的教师学生比例  
 B: 城镇中的黑人比例  
 LSTAT: 地区中有多少房东属于低收入人群  
 MEDV: 自住房屋房价中位数（也就是均价）

· 总之，数据集包括506个样本，每个样本包括12个特征变量和该地区的平均房价。

# 库文件描述

所用到的库有tensorflow,matplotlib,numpy,pandas,sklearn。

TensorFlow是一个基于[数据流编程](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%B5%81%E7%BC%96%E7%A8%8B/22735640" \t "https://baike.baidu.com/item/TensorFlow/_blank)（dataflow programming）的符号数学系统，其前身是[谷歌](https://baike.baidu.com/item/%E8%B0%B7%E6%AD%8C/117920" \t "https://baike.baidu.com/item/TensorFlow/_blank)的神经网络算法库DistBelief。

Matplotlib主要用于绘图。

Numpy主要用于数组处理。

Pandas是一款开源的、基于BSD协议的Python库，能够提供高性能、易用的数据结构和数据分析工具。他能够从CSV文件、文本文件、MS Excel、SQL数据库，甚至是用于科学用途的HDF5格式中读取文件。对CSV文件加载能够自动识别列头，支持列的直接寻址。数据结构自动转换为Numpy的多维数组。

Sklearn是机器学习中常用的第三方模块，对常用的机器学习方法进行了封装，包括回归(Regression)、降维(Dimensionality Reduction)、分类(Classfication)、聚类(Clustering)等方法。他是简单高效的数据挖掘和数据分析工具，让每个人能够在复杂环境中重复使用，建立NumPy、Scipy、MatPlotLib之上。

# 逻辑结构描述

由于房价和多个特征变量相关，所以使用多元线性回归建模。其公式如下：

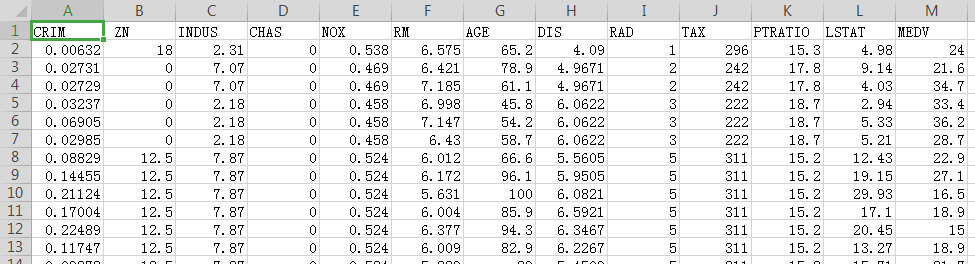


房价预测结果由不同特征的输入值和对应的权重相乘求和，加上偏置项计算求解。

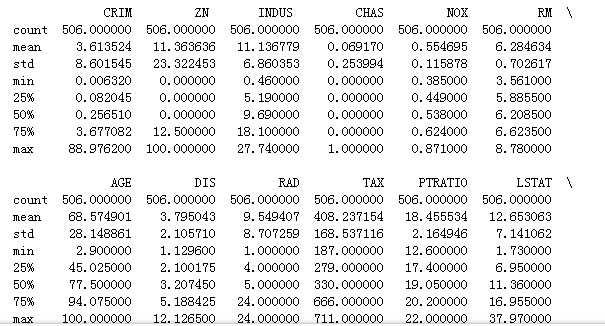
具体步骤：

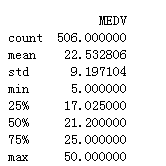
1. 读取数据

原数据如下：



读取出来的数据描述如下：





1. 数据预处理：归一化，打乱数据顺序

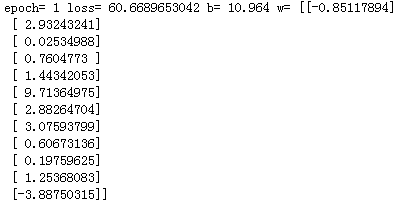
考虑到不同特征值取值范围大小不同，会有影响，所以要对数据进行归一化。记特征值为x,最小特征值为，最大特征值为，归一化后的结果为，则归一化公式为：



打乱数据顺序是为了防止过拟合。

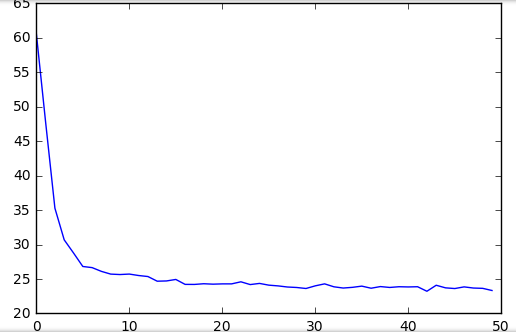
1. 定义模型：包括定义输入输出，模型结构与命名空间等。
2. 模型训练：这里设置训练轮次为50，学习率为0.01，定义均方差损失函数，选择梯度下降优化器。

部分训练过程中的训练数据如下：



1. 模型应用：可视化损失函数，对房价进行预测。

训练过程中损失函数图像如下所示：

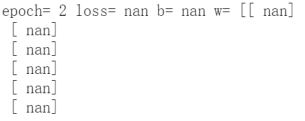


随机测试一个数据，其结果如下：



# 主要问题及解决过程

在第一次输出结果时，出现了如下的结果：



网上查阅资料后才知道，这是因为不同的特征值之间大小差别很大。所以需要进行归一化处理，处理后就解决了这个问题。

# 总结

通过这次作业，我完整的完成了一个小的房价预测模型的构建，主要用到了多元线性回归模型，知道了训练结果出现异常的一个原因。但是单纯的线性模型进行预测，在精度要求较高的情况下，就有误差较大的缺点了，所以可以尝试加上非线性模型进行改进。