**Boston房价回归课程报告**

学号：20307110315

姓名：周训哲

# 1.任务描述

主要采用SVR支持向量回归算法对boston多因子变量进行回归。其中，对数据的处理有两种处理，一种是简单的正则化处理，另一种是正则化处理后还加上pca降维处理。运用的核函数有线性核、多项式核、高斯核。除此而外，还利用图表将boston房价相关因子进行可视化分析。

# 2.数据集描述

**采用何种数据集开展实验？**

采用波士顿房价数据集。波士顿房价数据集来源于1978年美国某经济学杂志上。该数据集包含若干波士顿房屋的价格及其各项数据，每个数据项包含14个数据，分别是犯罪率、是否在河边和平均房间数等相关信息，其中最后一个数据是房屋中间价。

**数据集有什么特点？**

该数据集为多因子统计数据，总共包含14种房子相关的因素以及评分：

1.CRIM: per capita crime rate by town 每个城镇人均犯罪率

2.ZN: proportion of residential land zoned for lots over 25,000 sq.ft. 超过25000平方英尺用地划为居住用地的百分比

3.INDUS: proportion of non-retail business acres per town 非零售商用地百分比

4.CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) 是否靠近查尔斯河

5.NOX: nitric oxides concentration (parts per 10 million) 氮氧化物浓度

6.RM: average number of rooms per dwelling 住宅平均房间数目

7.AGE: proportion of owner-occupied units built prior to 1940 1940年前建成自用单位比例

8.DIS: weighted distances to five Boston employment centres 到5个波士顿就业服务中心的加权距离

9.RAD: index of accessibility to radial highways 无障碍径向高速公路指数

10.TAX: full-value property-tax rate per $10,000 每万元物业税率

11.PTRATIO: pupil-teacher ratio by town 小学师生比例

12.B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town 黑人比例指数

13.LSTAT: % lower status of the population 下层经济阶层比例

14.MEDV: Median value of owner-occupied homes in $1000's 业主自住房屋中值

**数据集的各项统计量是多少？**

总共包含了506组房价统计的信息，其中每组信息包含14项数据。

# 3.方法介绍

## 3.1数据预处理（如有）

**从原始数据到输入模型的数据之间，经过了哪些处理？这些处理有何作用？**

对于数据的获取有两种方式，一种是下载为本地csv，通过read\_csv（）函数读取本地文件，并且将数据进行列的命名。另一种是用load\_boston（）函数从网上读取数据，用.data和.target得到自变量和因变量。

在实验过程中，由于没有单独分开的训练集和测试集，人为随机将数据集利用train\_test\_split（）函数拆分为80%的训练集和20%的测试集，根据训练集拟合的函数对测试集数据进行回测，再用预测数据与测试数据进行比对和计算损失函数，得到拟合结果的优劣。

然后对数据进行正则化处理，

scaled\_x = [(i - mean) / std for i in x]

其中，mean表示平均值，std表示标准差，通过正则化处理，即可增强最相关的因子影响力，削弱不相关因子的影响，是模型准确率提高。

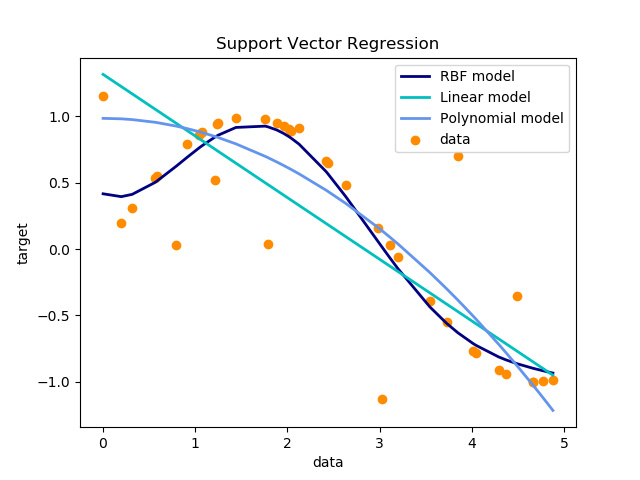
除此而外，还引入了pca降维处理，将数据经过降维后更加易于处理和辨认相关因子。

## 3.2算法描述

**算法原理**：

SVR的原理是寻找一个回归函数y=wx+b，以及其间隔带，在间隔带外面的数据才计算损失，而在间隔带里面的数据不计算损失。其中，间隔带的宽度类似SVM算法，为2/||w||，而与SVM不同的地方在于，SVR需要使得间隔带尽量的小，从而保证回归的准确性。也即寻找最大的||w||，在约束条件|y-(wx+b)|< ε下，通过拉格朗日乘子法计算lamda和ε的值，然后利用对偶问题进行转换，求导得到KKT条件，在KKT条件下，使用SMO算法即可求解w和b。代入即可。

对于非线性回归问题，普通的线性回归显然不可求解，则需要非线性回归方法。核函数则提供非线性回归的模型，对非线性问题进行回归求解。



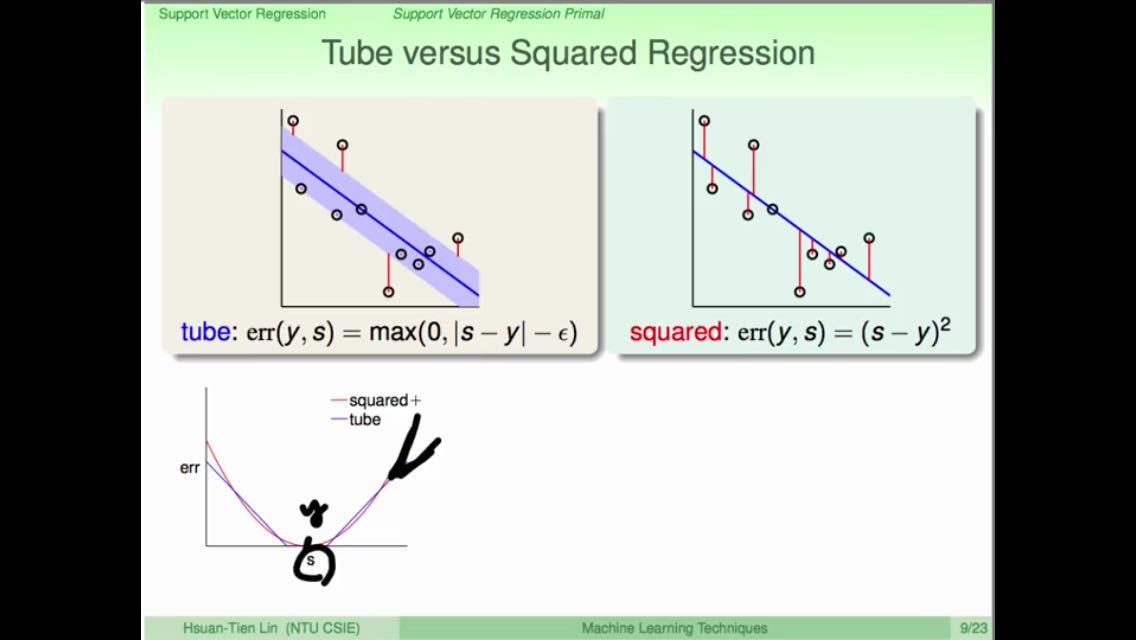
**整体流程：**

首先导入数据集，对数据集进行正则化或pca降维处理。然后将数据集进行人为拆分为训练集和测试集，调用库函数，建立一个回归模型，调用模型中的fit（）函数对训练集数据进行拟合。然后用模型中的predict（）函数对测试集进行预测，得到预测值。

**函数描述及优化：**

映射函数：对于变量data与价格price而言，则是一个多项式回归，f（x）=w1\*x1^t1+…wn\*xn^tn…。

损失函数：为了拟合的程度尽量的高，则需要寻求合适的系数w。引入损失函数loss\_function，其计算方式为，对于在间隔带外面的数据，计算数据与间隔带的误差的绝对值，而对于间隔带以内的数据，则不需要计算。



**训练及预测过程：**

训练过程即利用训练集进行拟合（fit（）），寻求损失函数最小的系数值，即完成训练过程。预测过程即将训练过程得到的系数对测试样例进行映射（predict（）），得到预测值，将预测值与实际值进行比对，运用打分函数（求r^2以及均方根误差）来判断拟合程度的高低。

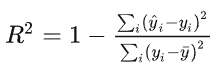
**缺点：**

同SVM，由于SVR在进行SMO计算时，需要对所有数据进行多次循环迭代以更新lamda，所以当数据集较大时，则可能耗费时间较长。

# 4.实验结果分析

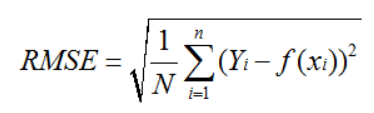
## 4.1评价指标

评价指标之一为拟合系数（R2）



用来判断拟合程度的高低，当R2越接近1，说明y^越接近y，即预测值越接近真实值，说明拟合程度很好。

评价指标之二为均方根误差



当均方根误差越大时，说明预测值与实际值的误差越大，拟合程度越低。

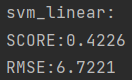
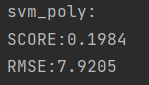
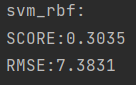
## 4.2定量评价结果



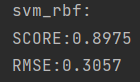
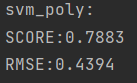
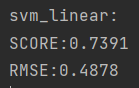
同时设定测试集与训练集分配的比例与随机分配的随机种子相同来进行测试，以控制数据集相同，以防非算法本身的误差所导致结果不同。



若对数据不进行处理，计算r2以及均方根误差的结果并输出：

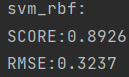
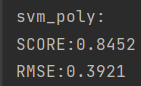
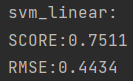
  

进行正则化处理，计算r2以及均方根误差的结果并输出：



可见，对于boston房价数据集，如果采用线性核模型，其回归效果明显不如高斯核的结果。所以boston房价数据集可能是一个非线性模型。

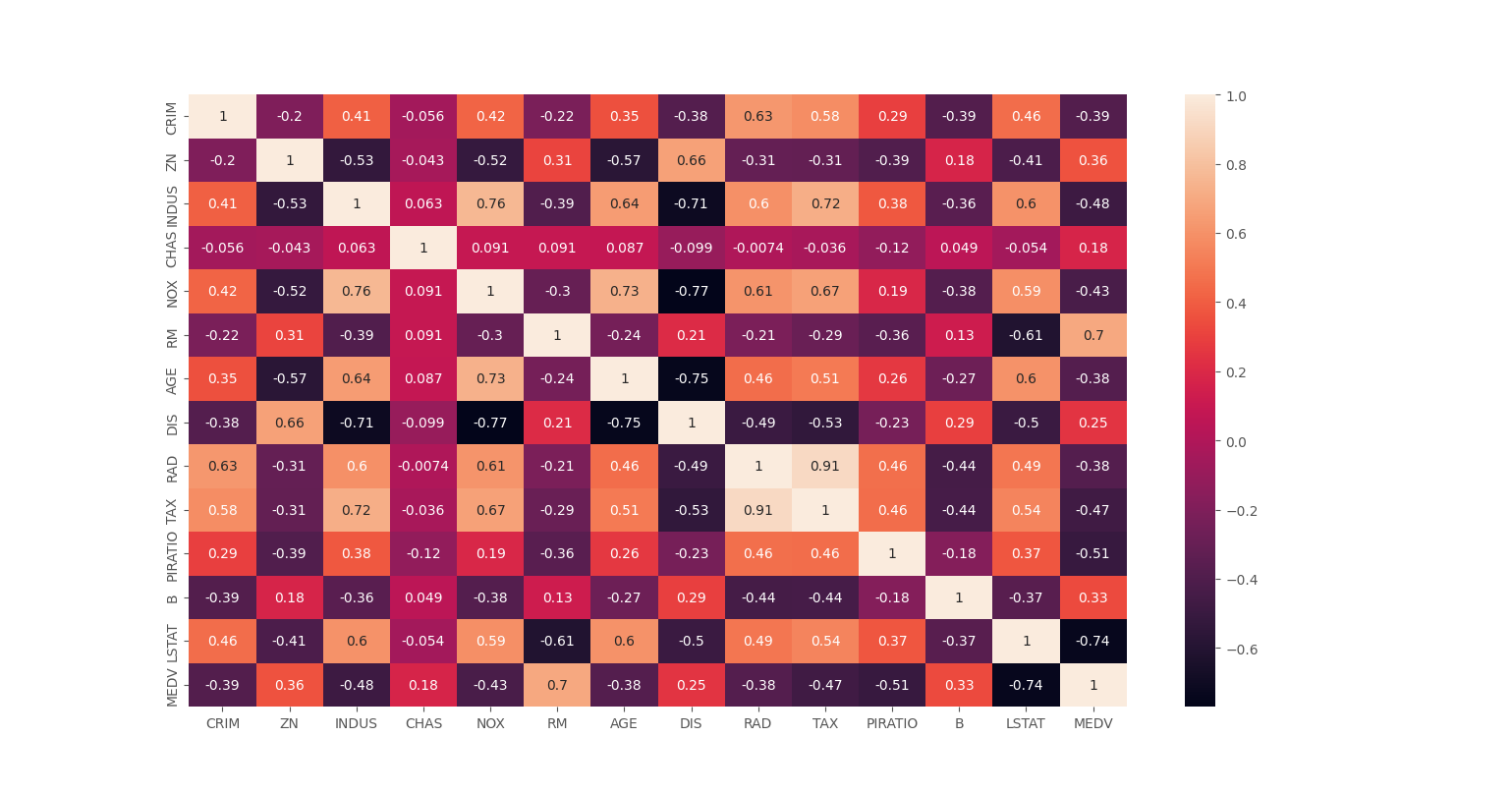
然后进行pca降维分析，检查回归效果：



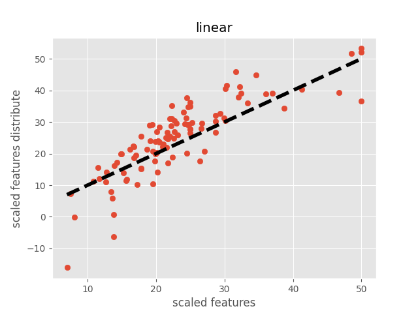
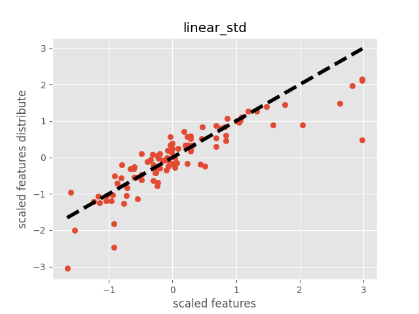
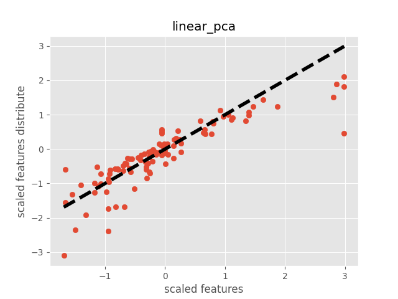
可以发现，回归结果都要好于只进行正则化处理的结果，说明pca降维处理是有效的；同时正则化处理的结果又优于不做处理的结果，说明使用正则化处理也是有效的。

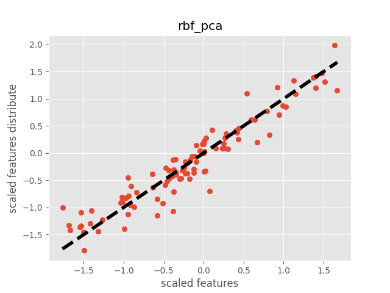
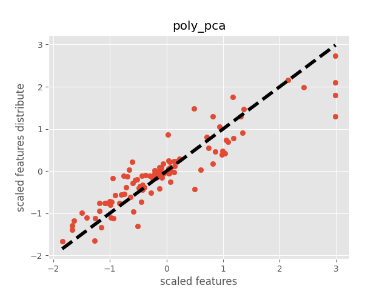
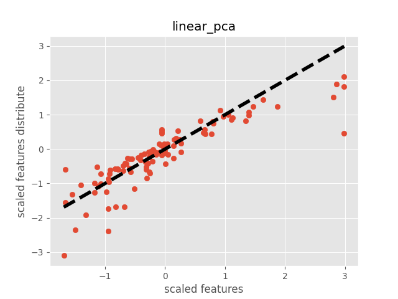
仔细分析原因，是因为通过正则化和pca处理，使得相关性好的因素更为凸显，抛去无关因子后，其回归的准确率会更好。

## 4.3可视化结果



首先可以对原始数据进行分析，其中1表示相关性最高，0表示没有相关性，可以看出哪些数据相关性大，哪些数据相关性小或负相关。主要考虑MEDV项，可以发现RM与其相关性最高，为0.7，LSTAT与其负相关性最高，为-0.74。



该图片横坐标为房价实际值，纵坐标为房价预测值，黑色虚线为y=x，当点阵集中在黑色虚线附近时说明预测值与真实值相接近，拟合程度较好。

经过横向和纵向对比，可以发现非线性核拟合效果更好，pca+正则化数据处理效果更好。

# 总结

通过回归分析，经过横向和纵向对比，可以发现非线性核拟合效果更好，pca+正则化数据处理效果更好。这给我以后实现机器学习、深度学习提供一个思路，即在实现算法之前可以进行数据处理，找出相关性最高的因子，然后再进行算法代入。还有应该使用多种算法比对其效率，进行归因分析，才能得到效果最好的模型。

通过本次作业，我了解到对于数据回归中的SVR回归方法，了解了其原理以及python代码实现，为之后的学习又添上一个基础知识。

参考文献：

[1]<https://blog.csdn.net/weixin_45508265/article/details/113255859>

[2] [Linear Regression and PCA - Boston Housing | Kaggle](https://www.kaggle.com/code/agnihotri/linear-regression-and-pca-boston-housing#Exploratory-Data-Analysis)

[3]<https://baike.baidu.com/item/Lasso%E7%AE%97%E6%B3%95/22685468>

[4]<https://baike.baidu.com/item/%E5%B2%AD%E5%9B%9E%E5%BD%92?fromModule=lemma_search-box>