机器学习工程师毕业项目

基于回归分析的房价预测

康飞

2016 年 12 月 26 日

1 定义

1.1 项目概述

预测是一个常见的问题，例如预测学习毕业情况，预测股票的涨跌等等。本项目是预测房屋价格。问一个购房者来描述他们梦想的房子,他们可能不会关心从房屋地下室到天花板的高度或靠近一个铁路。但这些数据证明影响价格谈判，比如比卧室的数量或着白色尖篱笆的数量。

这个项目将关注算法方法,准确的机器学习算法广泛应用于信息科学和计算机科学。有许多分类算法如资讯、朴素贝叶斯、决策树等,所有这些都有其优点和缺点。

回归分析是研究自变量和因变量之间关系的一种预测模型技术。这些技术应用于预测，时间序列模型和找到变量之间关系。例如可以通过回归去研究超速与交通事故发生次数的关系。

本项目是采用的 kaggle 上面的竞赛数据，来自波士顿郊区的房屋信息，根据不同的特征类型预测房屋价格，根据提供房屋的地理位置，房屋面积等特征预测房屋价格。

数据集分成 4 个部分：

1、data_description 描述数据的一个文本

2、train.csv 训练数据集

3、test.csv 测试数据集

4、sample_submission.csv 一个最终提交的预测房屋对应价格的文档

文件描述：

train.csv——训练集

test.csv——测试集

data_description.txt -每一列的完整描述,最初由院长 De Cock 但轻轻编辑这里使用的列名称相匹配

1.2 问题陈述

波士顿郊区的房屋数据集的预测是一个有监督的回归问题,一个训练集和一个测试集，训练集有多个特征对应的房屋价格，我们的目的是提取适当的特征，构建一个有效的模型，预测每个房屋的正确销售价格。

利用马萨诸塞州波士顿郊区的房屋信息数据训练和测试一个模型，并对模型的性能和预测能力进行测试。通过该数据训练后的好的模型可以被用来对房屋做特定预测。

我们利用机器学习的回归方法来分析房价的价格趋势，通过竞赛 kaggle 提供的 dataset，根据训练数据中房屋的多个特征对房屋建模，使得最终收敛到一个真实的价格，由于在建模过程中，难免出现拟合不准的情况，我们在最终的建模过程中使用不同的策略来对比，使用误差最小的模型作为最终模型。

在开始探索数据集训练部分,然后提取有用的特征，分析自变量特征和因变量预测值的相关性，给予这些特征我们构造回归模型，比较这些回归算法在测试数据的效率和选择一个作为最终的模型。最后我将验证在不同的数据集分类器的性能和参数。

这里有一些使用回归分析的好处：它指示出自变量与因变量之间的显著关系；它指示出多个自变量对因变量的影响。回归分析允许我们比较不同尺度的变量，例如：价格改变的影响和宣传活动的次数。这些好处可以帮助市场研究者 / 数据分析师去除和评价用于建立预测模型里面的变量。

1.3 评价指标

在这个项目中，我使用均方误差(RMSE)作为衡量指标，这也是 kaggle 平台上要求的衡量指标。

准确性:提交评估使用均方误差(RMSE)预报值的对数与对数之间的销售价格。(意味着错误预测昂贵的房子,便宜的房子同样会影响结果。

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n} (x_{1,t} - x_{2,t})^2}{n}}.$$

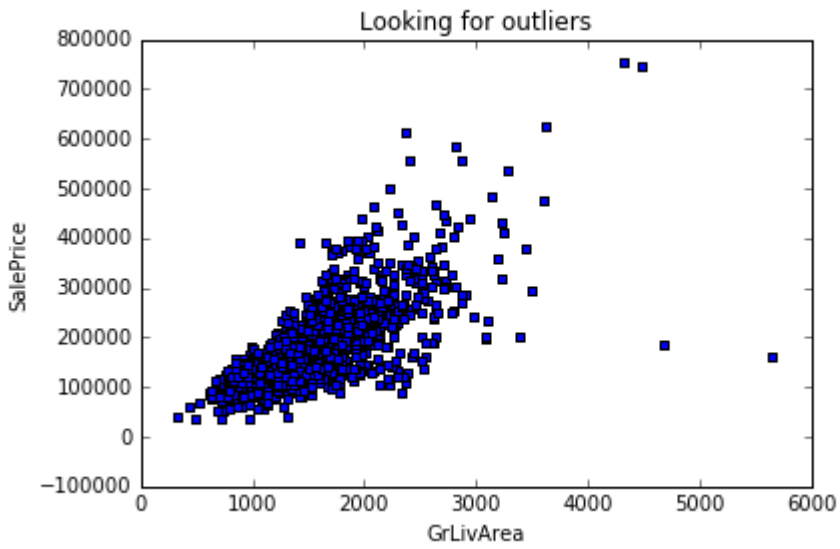
均方根误差，它是观测值与真值偏差的平方和观测次数 n 比值的平方根，在实际测量中，观测次数 n 总是有限的，真值只能用最可信赖（最佳）值来代替.方根误差对一组测量中的特大或特小误差反映非常敏感，所以，均方根误差能够很好地反映出测量的精密度。均方根误差，当对某一量进行甚多次的测量时，取这一测量列真误差的均方根差(真误差平方的算术平均值再开方)，称为标准偏差，以 σ 表示。σ 反映了测量数据偏离真实值的程度，σ 越小，表示测量精度越高，因此可用 σ 作为评定这一测量过程精度的标准。

2 分析

2.1 数据可视化

通过数据探索，发现数据集中有分类变量，缺失值和异常值，对分类变量移除了，缺失值补充为平均值，异常值移除了。

数据集使用了 numpy、pandas、matplotlib、scikit-llearn 工具进行分析，一共有 80 个特征。我们对房屋的价格和房屋的面积进行分析，判断价格和面积的相关度。



似乎有 2 极端异常值在右下角,很大的房子,卖的很便宜。所以删除任何超过 4000 平方英尺的房屋数据集。我们同时发现房屋的价格和面积有很高的相关度。

观察数据我发现有数字型的，有文本型的，还有缺失值，空置等等数据。我们需要进一步对数据进行探索分析。

由于数据中很有多缺失值，我们对数据进行了处理。

1、对数据的缺失值进行处理，将空值替换成为最有可能的值。

2、对现有的特征简化

3、组合存在的特征，比如吧一层的面积和二层的面积合并。

4、同时我们观察和销售价格和这些特征线性关系，找出最相关的特征。

Find most important features relative to target

SalePrice 1.000

OverallQual 0.819

AllSF 0.817

AllFlrsSF 0.729

GrLivArea 0.719

SimplOverallQual 0.708

ExterQual 0.681

GarageCars 0.680

TotalBath 0.673

KitchenQual 0.667

GarageScore 0.657

GarageArea 0.655

TotalBsmtSF 0.642

SimplExterQual 0.636

SimplGarageScore 0.631

BsmtQual 0.615

1stFlrSF 0.614

SimplKitchenQual 0.610

OverallGrade 0.604

SimplBsmtQual 0.594

FullBath 0.591

YearBuilt 0.589

ExterGrade 0.587

YearRemodAdd 0.569

FireplaceQu 0.547

GarageYrBlt 0.544

TotRmsAbvGrd 0.533

SimplOverallGrade 0.527

SimplKitchenScore 0.523

FireplaceScore 0.518

...

我们发现有很多特征相关，认为是一个不是简单的线性回归，可能是一个多元线性回归，所以人工在构造一些特征。

5、处理剩余的缺失值的数值特性使用中位数作为替代。

2.2 算法和技术

2.2.1 线性回归

线性回归用最适直线(回归线)去建立因变量 Y 和一个或多个自变量 X 之间的关系。可以用公式来表示：

Y=a+b*X+e

a 为截距，b 为回归线的斜率，e 是误差项。

重点：

1.自变量与因变量之间必须要有线性关系。

2.多重共线性、自相关和异方差对多元线性回归的影响很大。

3.线性回归对异常值非常敏感，其能严重影响回归线，最终影响预测值。

4.在多元的自变量中，我们可以通过前进法，后退法和逐步法去选择最显著的自变量。

2.2.2 岭回归

当碰到数据有多重共线性时，我们就会用到岭回归。所谓多重共线性，简单的说就是自变量之间有高度相关关系。在多重共线性中，即使是最

小二乘法是无偏的，它们的方差也会很大。通过在回归中加入一些偏差，岭回归酒会减少标准误差。

岭回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。'

岭回归是通过岭参数 λ 去解决多重共线性的问题。看下面的公式：

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

其中 loss 为损失函数，penalty 为惩罚项。

重点：

1.岭回归的假设与最小二乘法回归的假设相同除了假设正态性。

2.它把系数的值收缩了，但是不会为 0。

3.正则化方法是使用了 l2 正则。

2.2.3 LASSO 回归

和岭回归类似，Lasso(least Absolute Shrinkage and Selection Operator)也是通过惩罚其回归系数的绝对值。看下面的公式：

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

Lasso 回归和岭回归不同的是，Lasso 回归在惩罚方程中用的是绝对值，而不是平方。这就使得惩罚后的值可能会变成 0.

重点：

1.其假设与最小二乘回归相同除了正态性。

2.其能把系数收缩到 0，使得其能帮助特征选择。

3.这个正则化方法为 l1 正则化。

4.如果一组变量是高度相关的，lasso 会选择其中的一个，然后把其他的都变为 0.

LASSO 代表至少绝对收缩和选择算子。这是另一种正则化方法,我们简单地取代的平方权重加权的绝对值的总和。L1 与 L2 正规化,正则化收益率向量稀疏的特点:大部分能权重为零。稀疏可能是有用的在实践中如果我们有一个高维数据集与许多无关的功能。 这里应该比岭更有效率。

2.3 基准标准

我们使用 RMSE 作为测试基准。以上分别对不同的模型做了测试，我们最终以输出的 RMSE 分值作为衡量标准。分值越小，预测的准确率越高。

3 具体方法

3.1 数据预处理

我们这里有几个非数值的列需要做一定的转换！它们中很多是简单的 yes/no，比如 internet。这些可以合理地转化为 1/0（二元值，binary）值。其他的列，如 Mjob 和 Fjob，有两个以上的值，被称为分类变量（categorical variables）。处理这样的列的推荐方法是创建和可能值一样多的列（如：Fjob_teacher，Fjob_other，Fjob_services 等），然后将其中一个的值设为 1 另外的设为 0。这些创建的列有时候叫做 虚拟变量（dummy variables），我们将用pandas.get_dummies()函数来完成这个转换。

3.2 实现

我用线性回归、岭回归和 LASSO 回归络训练数据的方法,指标准确性和时间。

3.2.1 线性回归

首先导入 LinearRegression（）算法，然后 fit()训练数据进行训练，然后看预测训练和验证集，然后根据预测的结果画出散点图,对比预测值和真实值之间的残差。(残差是指实际观察值与估计值（拟合值）之间的差。)通过残差所提供的信息，分析出数据的可靠性、周期性或其它干扰。rems 分值为 0.40405824585252131。

RMSE 由于某种原因在训练集出现奇怪的错误。错误似乎随机分布和随机分散在中心线。这意味着我们的模型能够捕捉到的大部分解释信息。但是还是有很大误差。

3.2.2 岭回归

首先导入 RidgeCV（）算法，设置初始的 alphas = [0.01, 0.03, 0.06, 0.1, 0.3, 0.6, 1, 3, 6, 10, 30, 60]，然后训练模型，找出最优的 alphas 。然后使用最优的 alphas 调整参数，找出最优的 alphas，然后预测数据。对结果分析时候分别画出了预测值的散点图和重要系数的散点图。训练和测试结果表明,我们排除了大部分的过度拟合。视觉上,图似乎证实了这一想法。岭回归几乎使用了大部分存在的特征。rems 分值降低到 0.11642721377799541。

现在我们已经添加正则化发现我们得到一个更好的 RMSE 结果。训练和测试结果表明,我们排除了大部分的过度拟合。视觉上,图似乎证实了这一想法。岭回归几乎使用了大部分存在的特征。

3.2.3 LASSO 正则化

更进一步的分析，我们使用了 LASSO 正则化来分析数据，导入了 LassoCV（）包，同样找出最优的 alphas 值，然后去训练模型，对于训练集和测试集都有了更好的提升。rems 分值降低到 0.11583213221750714。

RMSE 结果对于训练集和测试集都有了更好的提升。

3.3 改进

在这个项目中,我没有考虑分类变量对销售价格的影响的相关系数，LotConfig、LandSlope、Condition 等等。,如果我将它们合并到一起或者使用 PCA 降维的方式，,我可以减少冗余的功能和提高效率的模型。

后续可以使用 ElasticNet 回归，ElasticNet 回归是 Lasso 回归和岭回归的组合。它会事先训练 L1 和 L2 作为惩罚项。当许多变量是相关的时候，Elastic-net 是有用的。Lasso 一般会随机选择其中一个，而 Elastic-net 则会选在两个。与 Lasso 和岭回归的利弊比较，一个实用的优点就是 Elastic-Net 会继承一些岭回归的稳定性。

ElasticNet 主要特点：

1.在选择变量的数量上没有限制

2.双重收缩对其有影响

3.除了这 7 个常用的回归技术，你也可以看看贝叶斯回归、生态学回归和鲁棒回归。

4 结果

4.1 模型评价与验证

根据我之前的研究中,我写的函数来计算性能的几个模型与不同大小的训练数据。

每个模型的学习和测试精度被绘制。

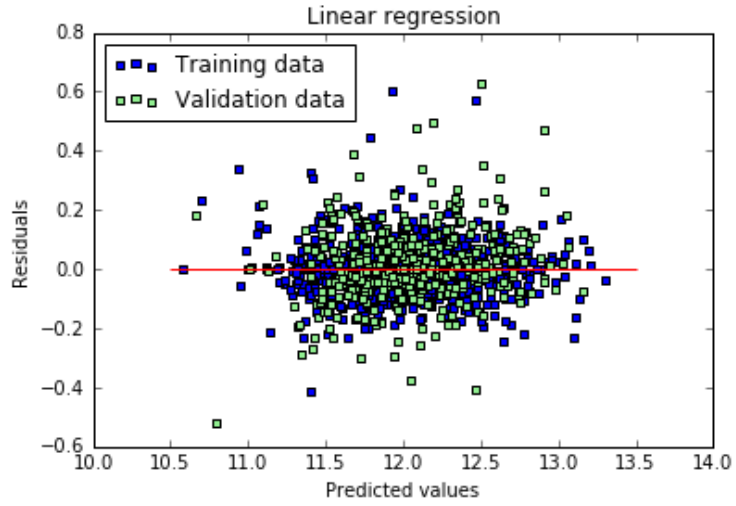
绘制不同模型的图，线性回归和岭回归

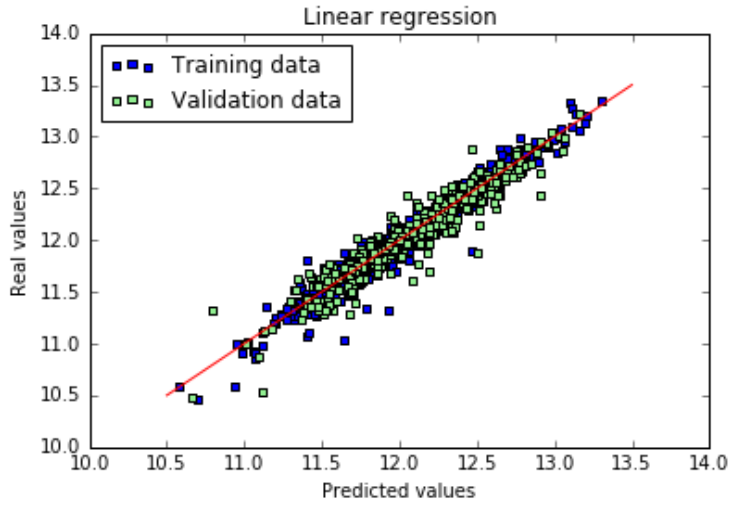
4.2 结果分析

4.2.1 线性回归

线性回归处理的结果如下图所示：

```
('RMSE on Training set :', 0.38851553131013078)
('RMSE on Test set :', 0.38716808616012294)
```

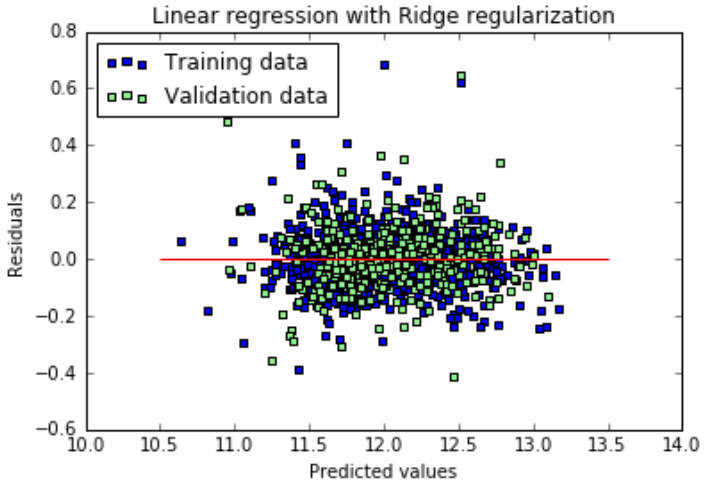


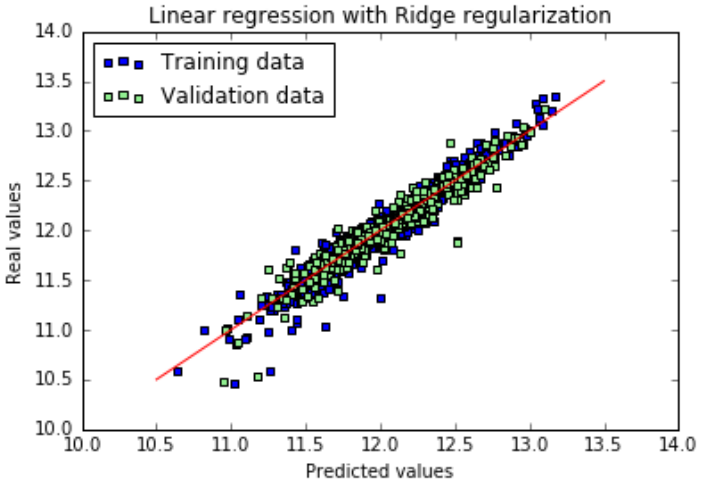


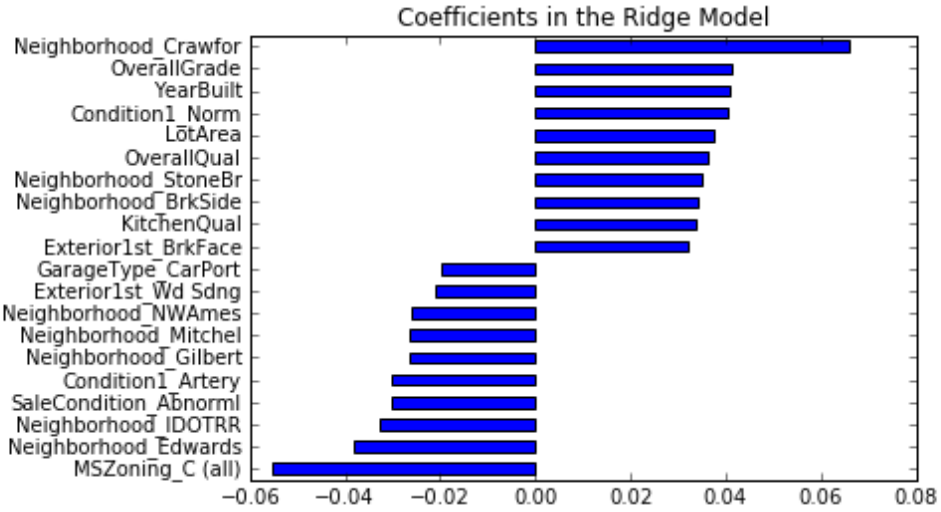
4.2.2 岭回归

岭回归处理的结果如下图所示：

```
('Best alpha :', 30.0)
Try again for more precision with alphas centered around 30.0
('Best alpha :', 24.0)
('Ridge RMSE on Training set :', 0.11540572328450793)
('Ridge RMSE on Test set :', 0.11642721377799554)
```



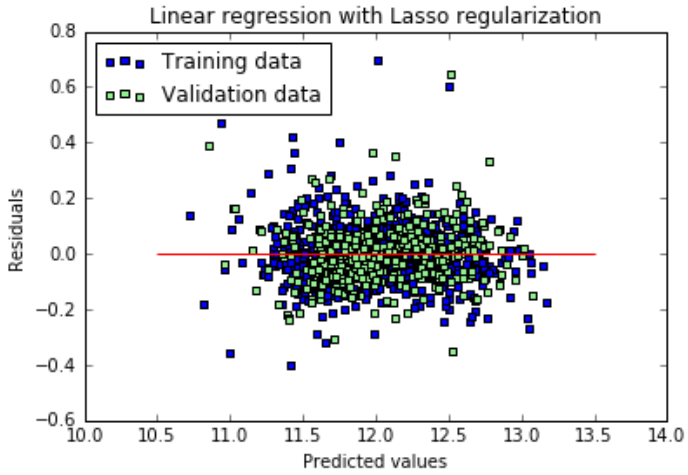


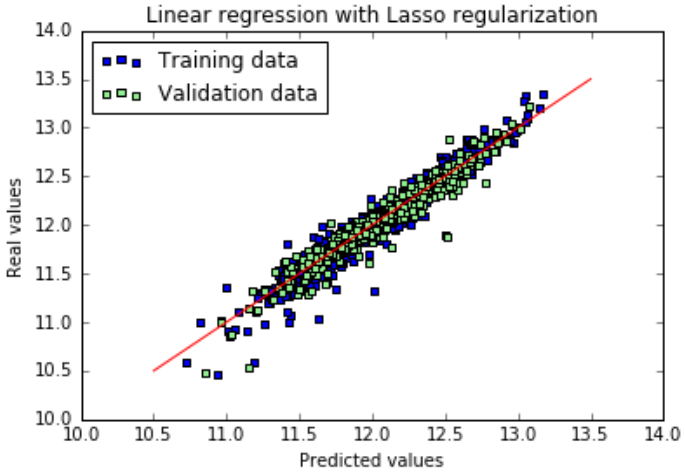


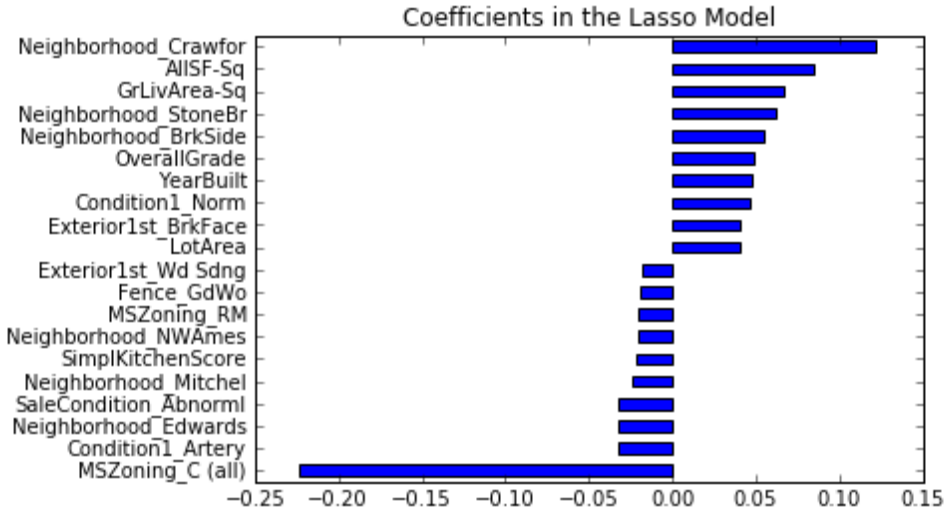
4.2.3 LASSO 正则化

Lasso 正则化处理的结果如下图所示：

```
('Best alpha :', 0.00059999999999999995)
Try again for more precision with alphas centered around 0.0006
('Best alpha :', 0.00059999999999999995)
('Lasso RMSE on Training set :', 0.11411150837458055)
('Lasso RMSE on Test set :', 0.11583213221750706)
```







5 结论

5.1 结论

这个项目是解决根据房屋的不同的特征预测房屋价格的问题，我分别使用了线性回归和岭回归算法来预测结果，正则化是一个非常有用的方法来处理共线性,过滤掉噪音数据,最终防止过度拟合。

正则化背后的概念是引入附加信息(偏见)惩罚极端参数权重。我们得到一个更好的 RMSE 结果现在我们已经添加正则化。很小的区别训练和测试结果表明,我们排除了大部分的过度拟合。视觉上,图似乎证实了这一想法。岭使用几乎所有的现有功能。

5.2 后续改进

在这个项目中,我没有考虑分类变量对销售价格的影响的相关系数，LotConfig、LandSlope、Condition 等等。,如果我将它们合并到一起或者使用 PCA 降维的方式，,我可以减少冗余的功能和提高效率的模型。

后续可以使用 ElasticNet 回归，ElasticNet 回归是 Lasso 回归和岭回归的组合。它会事先训练 L1 和 L2 作为惩罚项。当许多变量是相关的时候，Elastic-net 是有用的。Lasso 一般会随机选择其中一个，而 Elastic-net 则会选在两个。与 Lasso 和岭回归的利弊比较，一个实用的优点就是 Elastic-Net 会继承一些岭回归的稳定性。

ElasticNet 主要特点：

1.在选择变量的数量上没有限制

2.双重收缩对其有影响

3.除了这 7 个常用的回归技术，你也可以看看贝叶斯回归、生态学回归和鲁棒回归。

参考资料

https://docs.google.com/document/d/1B-vEOscvfqctGEMHTFDS9Nw7aqcE2iuwPRfp0jK8nf4/pub?embedded%3Dtrue

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

https://github.com/woshikangfei/student_intervention/blob/master/student_intervention.ipynb

https://en.wikipedia.org/wiki/Root-mean-square_deviation

http://baike.baidu.com/link?url=4_wLmmZ2aQmKgxXkApfcn1dmRhqIsQkn8Mv_sZQBPBbi5qMAyYVJ_IFjOBfwgPGYpuEr0dyw28rh-DfKpAqUlK

https://github.com/nd009/machine-learning/blob/zh-cn/projects/capstone/report-example-1.pdf