预测数值型数据：回归

1. 前言

分类一般是对标称型数据进行划分，例如性别、肿瘤等等。回归将对连续型的数据做出预测。

1. 线性回归
2. 首先要保证数据线性相关，否则对非线性的数据拟合不太好。
3. 目的是找到回归系数。
4. 利用分析模型的效果

模型平方和，预测值-样本均值的平方和：

残差平方和，真实值-预测值的平方和：

总平方和，真实值-样本均值的平方和：

而且，对于线性回归模型而言：

一般将决定系数定义为：

越接近于1对应更紧密的关联，接近0对应更弱的关联。

1. 这里利用正规方程的原理。

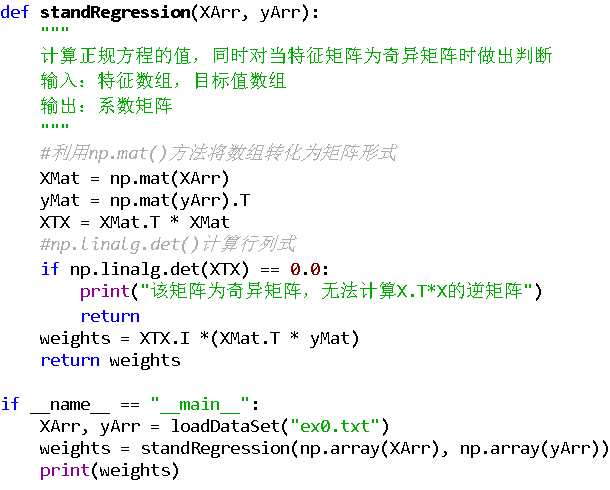
假设已知m个样本数据，特征数为n。其中上标表示第i个样本，下标表示该样本的第j个特征。

可以知道X是m\*(n+1)维，而w是(n+1)\*1维。想要利用线性代数的方法求显然不可取，因为只有方阵才可逆。所以在两边同时左乘，是(n+1)\*m维，此时如果可逆，则：

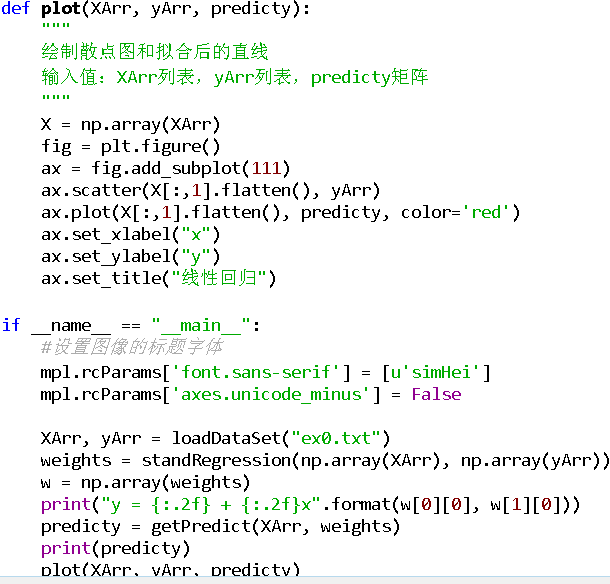
当然，也不一定可逆，所以在代码中要对此作出判断。

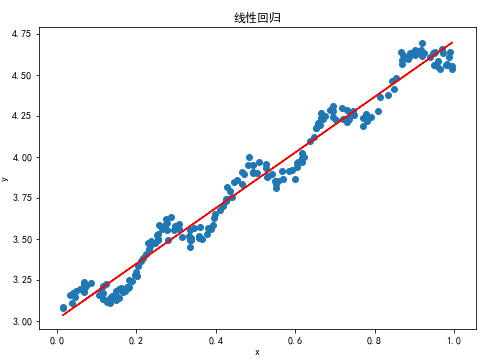
1. 最小二乘法（ordinary least squares）
2. Numpy提供了一个线性代数库linalg，可以直接调用linalg.det()计算行列式。还可以使用linalg.solve()来求解未知矩阵。





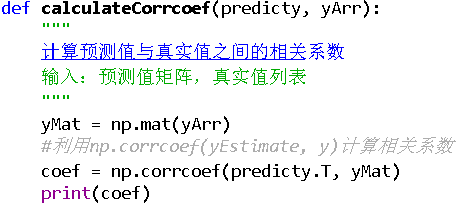
1. 之后绘制图像



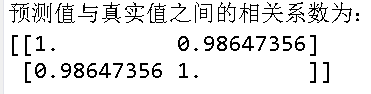


1. 计算相关系数查看拟合效果

Numpy提供了计算相关系数的方法，通过命令corrcoef(yEstimate,yActual)来计算预测值和真实值的相关性。







可以看到对角线上的元素为1，因为自己和自己的相关系数为1。预测值与真实值间的相关系数为98.6%，效果还不错。

1. 局部加权线性回归（Locally Weighted Linear Regression,LWLR）

线性回顾的一个问题是有可能出现欠拟合现象，因为求的是具有最小均方误差的无偏估计。可以在估计中引入一些偏差，从而降低预测的均方误差。

1. 给待预测点附近的每个点都赋予一定的权重，在这个子集上基于最小均方差来进行普通的回归。与knn一样，这种算法每次预测均要事先选取出对应的数据子集。该算法解出回归系数w如下：

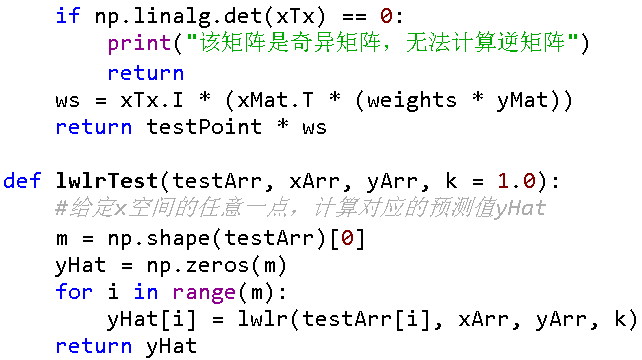
其中w是一个矩阵，用来给名数据点赋予权重。

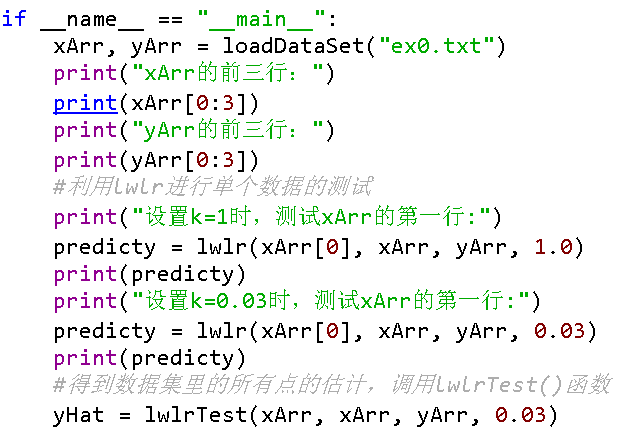
1. LWLR使用高斯核来对附近的点赋予更高的权重：

径向基函数：空间中任一点到某一中心x之间的欧式距离，其作用往往是局部的，当离x越远时，其值越小。x为核中心，k为函数的宽度参数，控制了函数的径向作用范围。用户需要指定参数k，决定了对附近的点赋予多大的权重，这也是唯一需要考虑的参数。

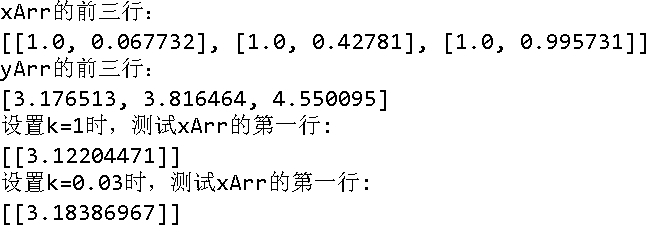
1. 利用代码解释原理



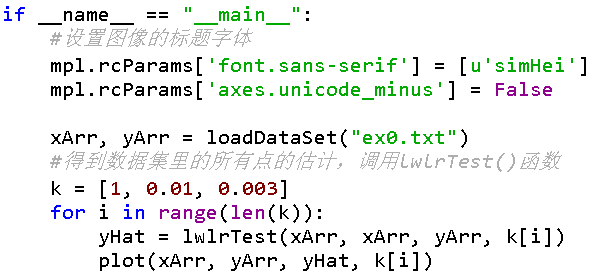
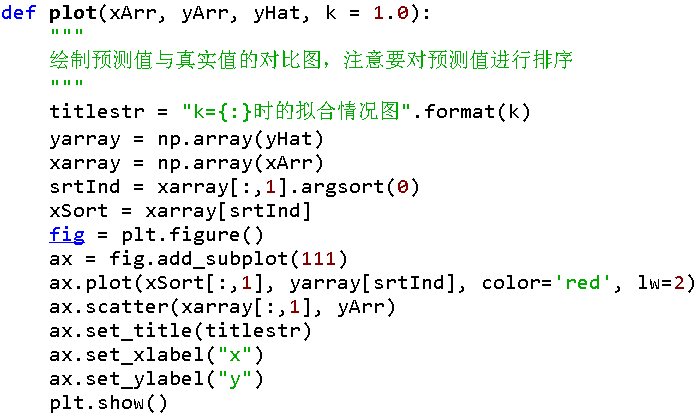




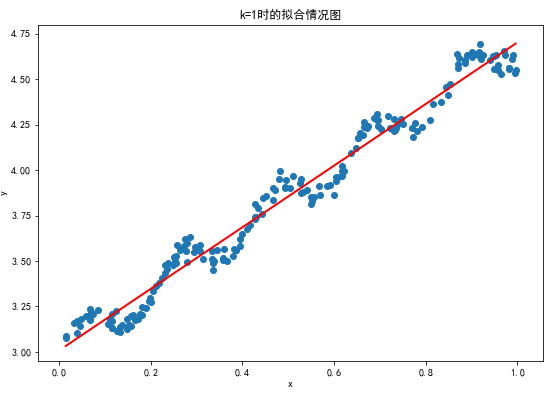
可以看到结果：

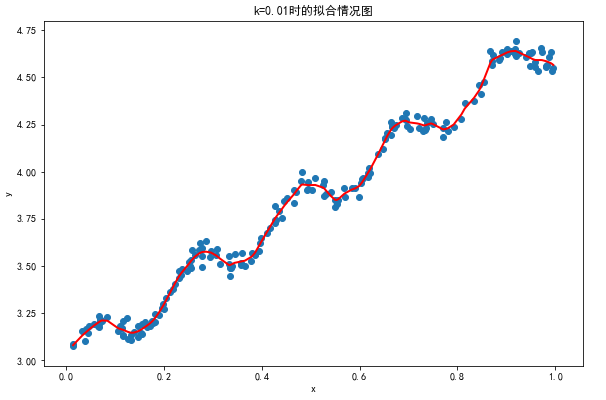


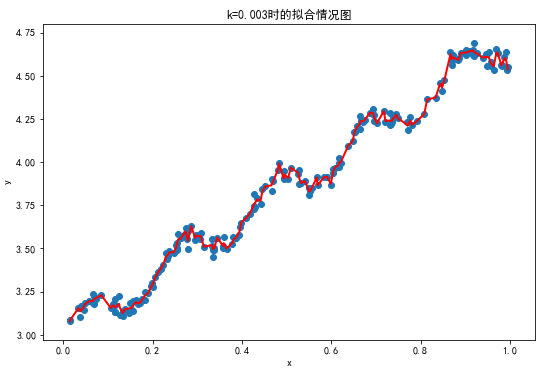
1. 可以设置不同的k值，查看程序运行的结果



实验所得的图像如下：







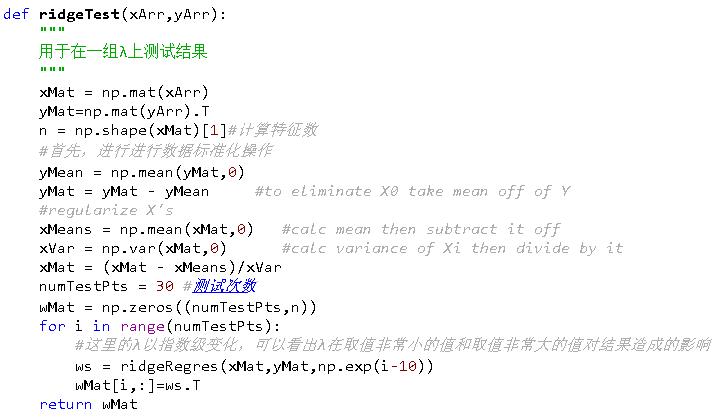
可以看到k=1时模型与最小二乘法差不多，k=0.01时该模型可以挖出数据的潜在规律，k=0.003时考虑到了太多的噪声，有些过拟合。

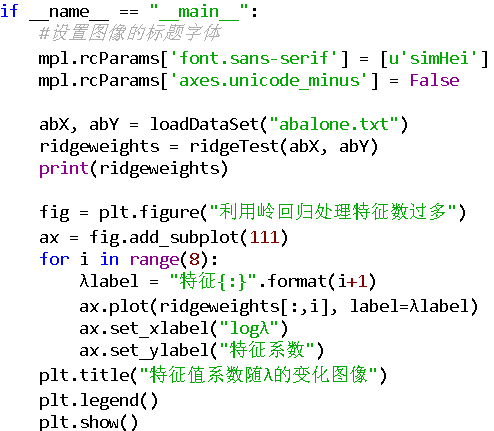
1. 预测鲍鱼的年龄

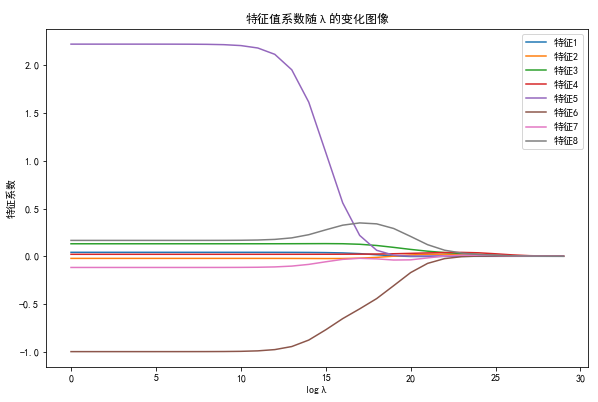
数据集过于庞大，可以只用前100行数据来作为测试：



1. 岭回归(ridge regression)
2. 当样本的特征数n比样本数m还多时，怎样利用线性回归方法？当然不能直接用，因为此时计算的逆矩阵可能会出错。
3. 岭回归就是在矩阵上加一个从而使得矩阵非奇异，进而可以对求逆。其中I是一个（n+1）（n+1）阶的单位矩阵。回归系数计算公式：
4. 岭回归最先用来处理特征数多于样本数的情况，现在也用于在估计中加入偏差，从而得到更好的估计。通过引入λ来限制所有的w之和，通过引入该惩罚项，能够减少不必要的参数。这个技术也叫缩减。
5. 缩减方法可以去掉不重要的参数，因此能更好地理解数据。因此，与简单的线性回归相比，缩减法能取得更好的预测效果。通过选取不同的λ来重复上述测试过程，最终得到一个使测试误差最小的λ。





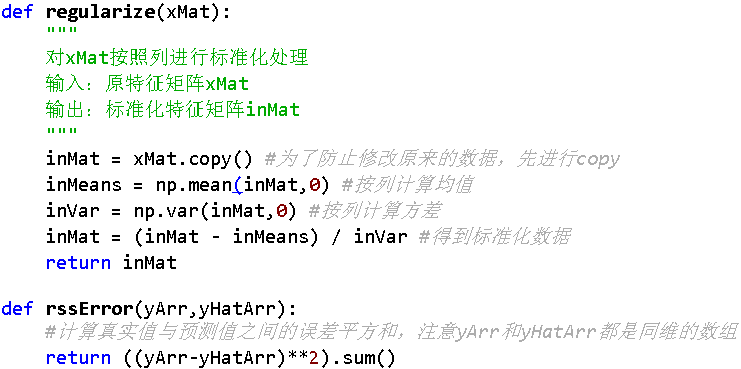


最左边λ最小时，可以看到所有系数的原始值（与线性回归一致）；而在最右边，系数全部缩减为0，在中间部分的某值可以取得最好的预测效果。

1. 前向逐步回归

属于一种贪心算法，每一步都尽可能的减少误差。一开始，所有权重都设置为1，然后每一步所做的决策就是对某个权重增加或者减少一个很小的值。

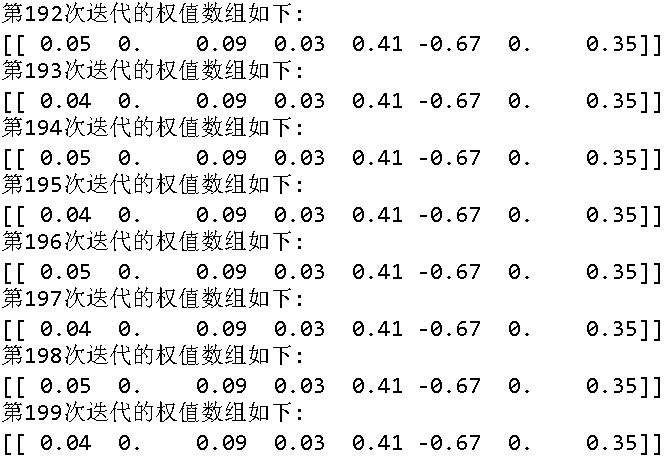
1. 程序测试：



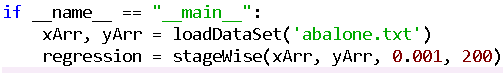


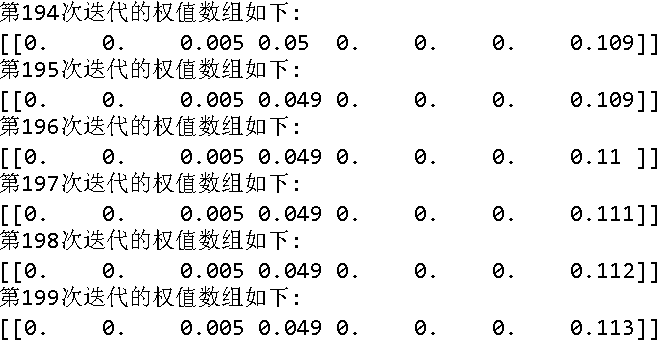


执行了200次迭代后的结果



可以看到w1和w6都是0，这表明它们不对目标值造成任何影响，这些特征可能是不需要的。步长设置为0.01，可能经过一段时间后系数就已经饱和，而且在特定值之间来回震荡，这是因为步长太长的原因。我们可以尝试缩小步长：





1. 主要优点在于可以帮助我们理解现有的模型并对之作出改进。及时停止对那些不重要特征的收集。
2. 方差是指模型之间的差异，偏差是指预测值和数据真实值之间的差异。