**带约束条件曲线拟合的简单尝试**

——wcy1122

1. **简介**

我们需要解决这样的一个问题，给定一个数列的前四项1991,1992,2004,2009，要求通过相关方法推测出这个数列的第五项。此外，该问题有一个额外的隐藏条件，数列的第五项>=2019。

~~拿到这个问题，我首先去OEIS上查询了一下这个数列，很遗憾，没有找到。~~

这是一个经典的预测回归类问题，解决此类问题的方法不外乎如下几种：（1）脑洞大开，找规律瞎猜；（2）插值；（3）曲线拟合。在我看来，插值法并不是一个适合于回归预测的方法，刻意经过数据点会带来严重的过拟合问题。关于曲线拟合的方法前面同学也已经展示了很多方法，比如使用多项式，sigmod，tanh函数，atan函数，三角函数等。

但我注意到，大部分的回归，第五项的值都是小于2019的，也就是说这些回归其实是错误的。因此，我尝试在这个方向上寻找突破口。

曲线拟合本质上是一个多元非线性最优化问题，我们首先设计一个损失函数，比如L2范数，然后通过最小化损失函数，让曲线接近数据点。注意到本题中的隐藏条件，数列的第五项>=2019。这是一个限制条件，转化为不等式也就是f(5)<=2019。带不等式约束的最优化问题其实是一个经典问题，我们可以使用众所周知的拉格朗日乘子法解决。因此我使用拉格朗日乘子法，配合上一些常见的拟合函数，尝试解决这个回归问题。

**2 拉格朗日乘子法**

拉格朗日乘子法是一类经典的解决带约束最优化问题的方法。对于每个不等式gi(x)<=0，我们引入一个非负参数λi，对于每个等式fi(x)=0，我们引入一个非负参数μi。我们可以根据原问题中的最优化函数f(x)构造拉格朗日函数。

我们将拉格朗日函数的最小化问题转换为对偶问题的最大化问题，找一种数值计算方法将其最优化即可。

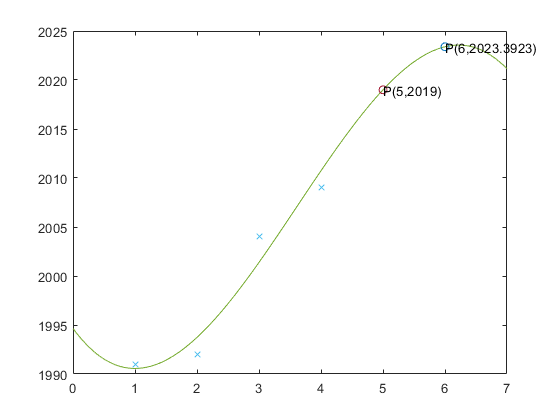
对于本问题，我们可以使用matlab下自带的带约束最优化函数fmincon，将原函数，初始值，约束条件传入函数，运行即可。

对于本问题，约束不等式和损失函数如下。

**3 实验结果**

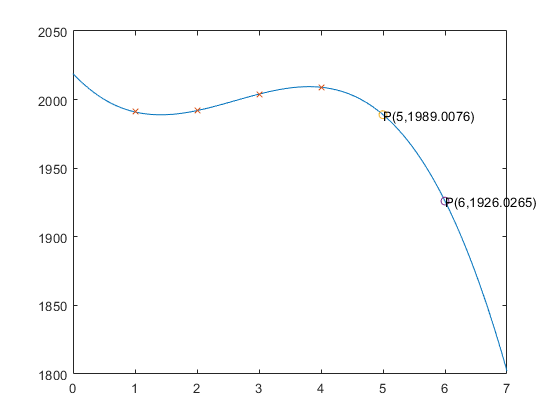
**3.1 多项式拟合**

将四个点进行插值会得到一个三次多项式，因此我首先使用三次多项式进行测试。定义f(x)=a\*x^3+b\*x^2+c\*x+d，结果如下。



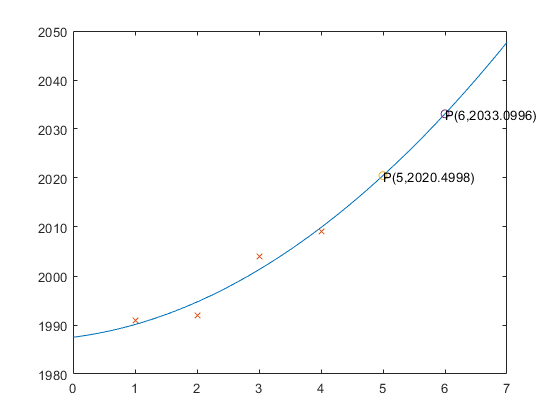
|  |
| --- |
| L2\_loss=13.0435  x=5,y=2019  x=6,y=2023.3922 |

和不带约束拟合，即等价的拉格朗日插值的进行对比。



|  |
| --- |
| L2\_loss=1.118e-06  x=5,y=1989.0076  x=6,y=1926.0265 |

我又尝试拟合了2次函数。定义f(x)=a\*x^2+b\*x+c，结果如下。



|  |
| --- |
| L2\_loss=16.2  x=5,y=2020.4998  x=6,y=2033.0996 |

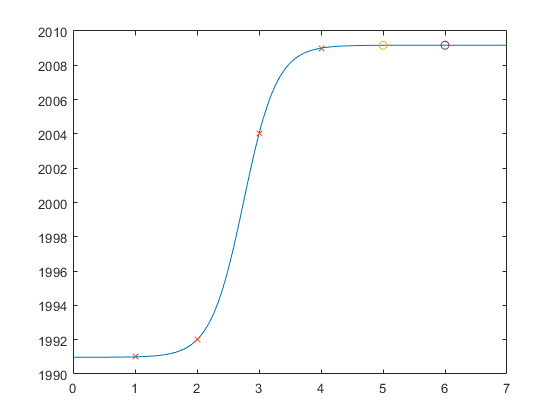
**3.2 tanh函数。**

我首先对不带约束的tanh函数进行拟合。Tanh的拟合方式如下。

非多项式函数需要调试初始值，否则一般情况下不会收敛。我设置的初始值：

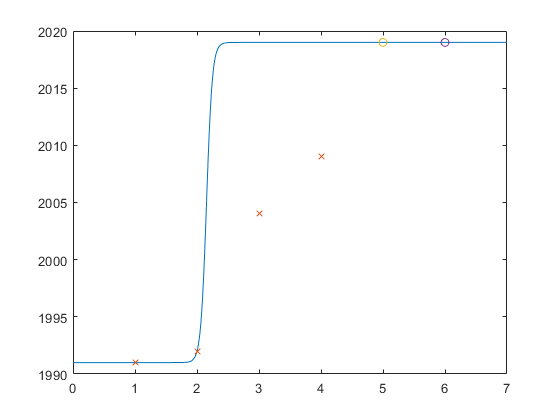
|  |
| --- |
| a=1,b=2,c=20,d=2000 |

结果如下。



|  |
| --- |
| L2\_loss=2.6552e-10  x=5,y=2009.1656  x=6,y=2009.1695 |

加入约束条件后，其实拟合效果并不是很好。为了满足约束，loss没有收敛到0。



|  |
| --- |
| L2\_loss=325  x=5,y=2019  x=6,y=2019 |

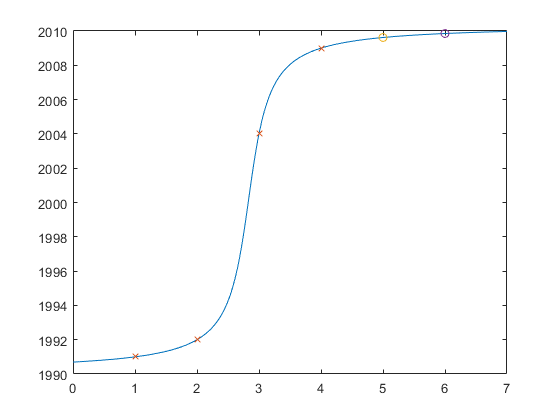
**3.3 atan函数。**

我首先对不带约束的tanh函数进行拟合。Tanh的拟合方式如下。

初始值：

|  |
| --- |
| a=0.5,b=1,c=2,d=2000 |

结果如下。



|  |
| --- |
| L2\_loss=1.7142e-09  x=5,y=2009.6205  x=6,y=2009.853 |

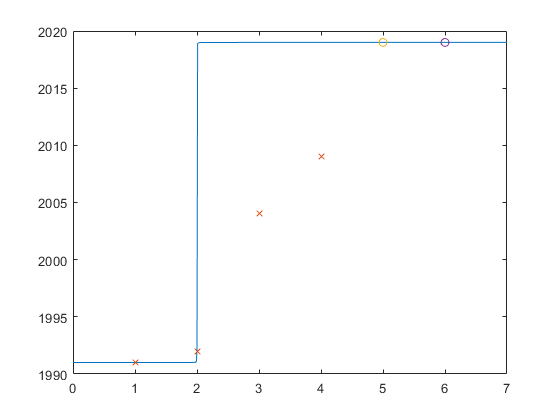
加入约束条件后，结果如下。和tanh类似，效果也不是很好，没有收敛到0。

不同的初始值得到不同的解。

初始值：

|  |
| --- |
| a=0.5,b=1,c=2,d=2000 |

图像：

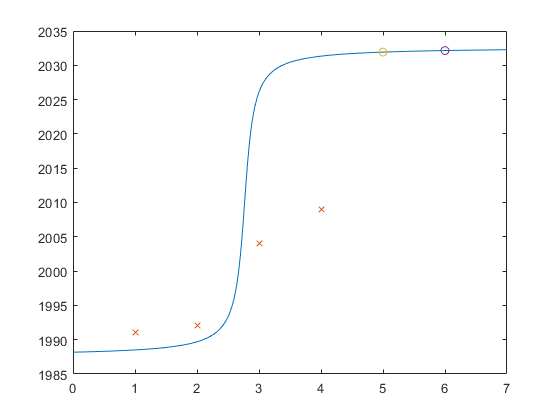


|  |
| --- |
| L2\_loss=325.0079  x=5,y=2019.0005  x=6,y=2019.0006 |

初始值：

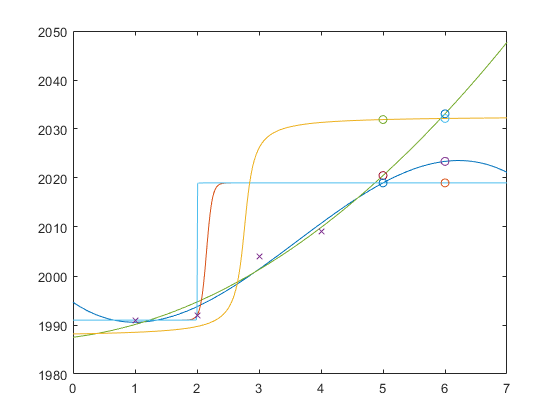
|  |
| --- |
| a=0.5,b=10,c=2,d=2000 |

图像：



|  |
| --- |
| L2\_loss=998.7521  x=5,y=2031.9241  x=6,y=2032.15 |

总体实验结果：



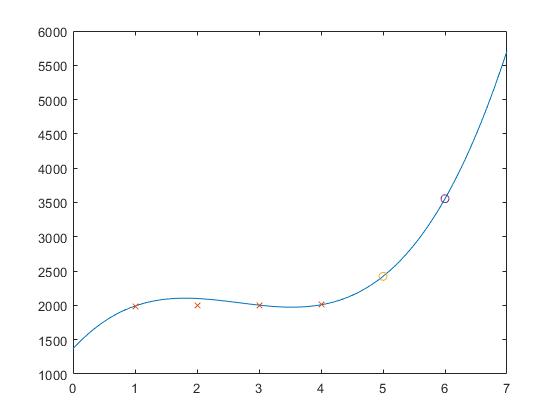
**4 一些新的尝试**

我尝试使用一个新的loss function进行测试。

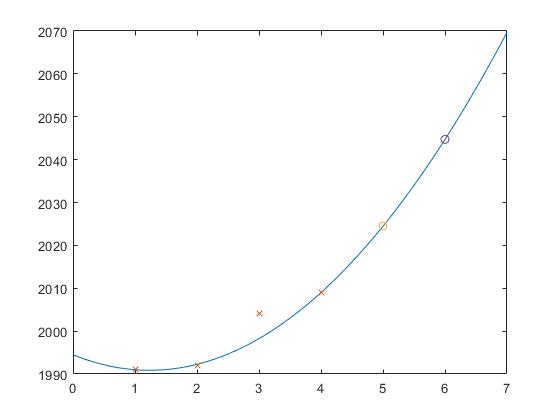
**4.1 L1误差**

以下是相关测试结果。

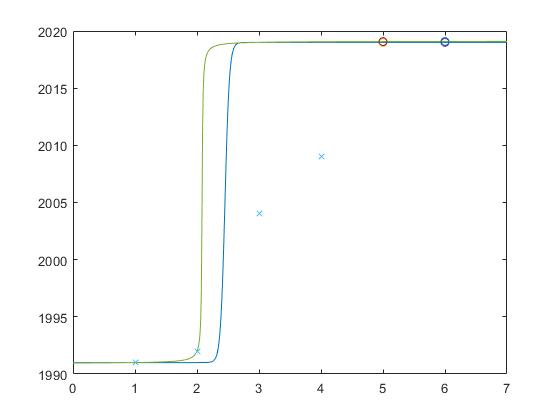
|  |  |  |  |
| --- | --- | --- | --- |
| **函数** | **Loss** | **f(5)** | **F(6)** |
| 三次函数 | 106.4972 | 2415.5256 | 3524.6907 |
| 二次函数 | 6 | 2024.6301 | 2045.0752 |
| Tanh函数 | 25.9999 | 2019 | 2019 |
| Arctan函数 | 25.0698 | 2019.0759 | 2019.0837 |

****

三次



二次



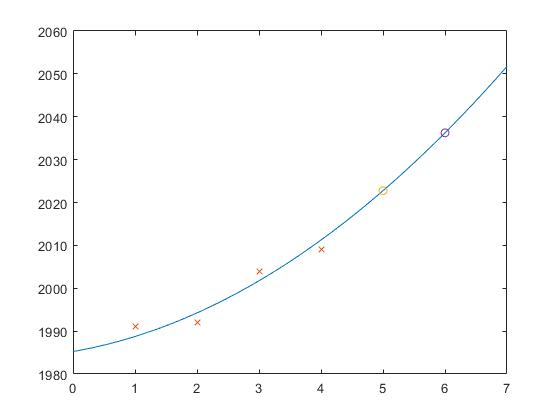
tanh和atan

**4.2 最大值误差**

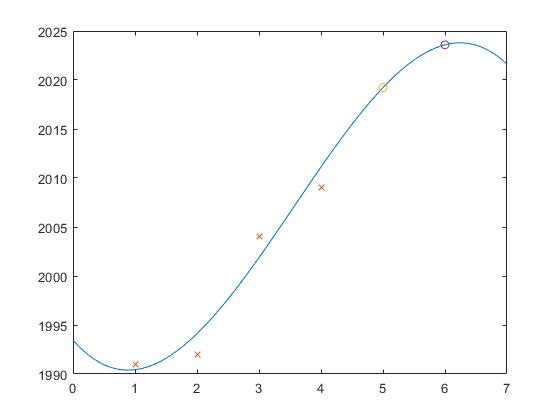
借鉴了SVM的损失函数，我使用以下函数作为损失函数。

以下是相关测试结果。

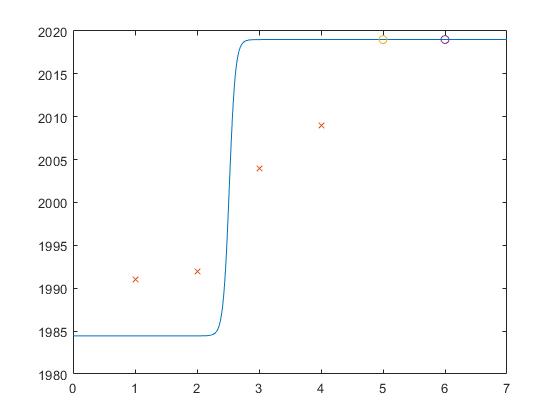
|  |  |  |  |
| --- | --- | --- | --- |
| **函数** | **Loss** | **f(5)** | **F(6)** |
| 三次函数 | 2.025 | 2019.3505 | 2025.1525 |
| 二次函数 | 2.2504 | 2022.7443 | 2036.2395 |
| Tanh函数 | 15 | 2019.0019 | 2019.0019 |
| Arctan函数 | 15 | 2023.8151 | 2024.4745 |

****

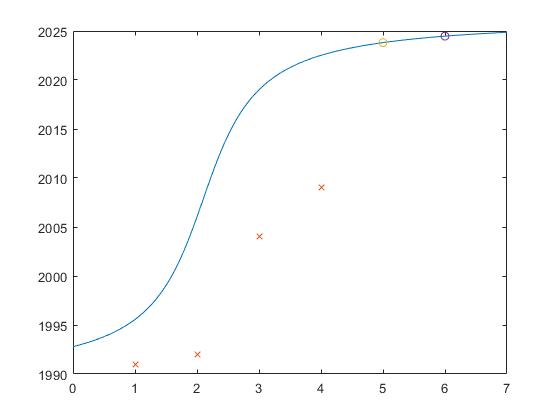
三次

****

二次

****

tanh

****

atan

**5 结论**

根据我们进行的预测，第五个数据点大概集中在2020年左右，第六个点出现在2030年左右的概率更高一些。以上只是一些带约束曲线拟合的简单尝试，可以看到拟合效果并不算很好，基本都没有收敛，当然我也没有特别仔细地调参。非线性约束条件在拉格朗日乘子法下的总体表现并不是很理想，对初始值的要求很高且很难收敛。或许可以考虑使用一些更优秀的最优化方法，比如牛顿法，或者搭一个深度神经网络。