

上机实验报告

|  |  |
| --- | --- |
| 题目： | 使用岭回归对数据进行多项式拟合 |

|  |  |
| --- | --- |
| 姓 名： | **李梓焓** |
| 学 号： | **1700011735** |
| 组 别： | **无** |
| 实验日期： | **2019.09.19** |

**摘要** 本实验运用岭回归对数据进行多项式拟合，并分析不同多项式次数下最佳的训练结果及相应的测试结果，以及与的关系，从而得出结论：（1）本实验最佳模型为三次多项式模型；（2）改变值对欠拟合和较好拟合的模型的优化并不明显，而对过拟合模型改善显著。

**关键词** 岭回归，多项式拟合，正则化系数，均方根误差

1. **引言**

线性回归模型是回归分析中最常用的模型之一，然而实际运用中，样本数据往往不满足线性关系，因此要引入高次特征项进行拟合，这就是多项式回归。

一般情况下，使用多项式回归可以使模型更加复杂，从而提高模型准确度，但如果模型过于复杂，过于贴近训练数据，有可能导致对真实数据的预测准确度并不高，即欠拟合现象。为了解决模型过拟合的问题，人们引入了正则化的概念，通过在损失函数中添加惩罚项，从而对损失函数的某些参数进行限制，以提高模型的泛化能力。其中最常用的正则化方法便是岭回归，其损失函数可表示为（按Scikit-learn的定义）：

是正则化系数，它控制正则化项的占比，对模型泛化而言非常重要。过小，正则化的效果弱，得到的模型仍有严重的过拟合现象；过大，正则化的效果强，容易导致模型的精确度变差。在实际运用中，作为超参数，需要结合训练和测试的结果进行多次尝试，寻找最合适的结果，也可以采用交叉验证的方式，获取最合适的值。

本次上机实验采用Scikit-learn自带的Ridge和RidgeCV，结合PolynomialFeatures，对训练样本进行多项式拟合，然后用mean\_squared\_error和Numpy的sqrt()计算RMS，最后用Matplotlib对数据进行可视化处理。

**2 实验部分**

**2.1 仪器**

2.1.1硬件

Surface Pro（第5代，处理器参数：Intel® Core™ i5-7300U CPU @ 2.60GHz，2.71 GHz，2个内核，4个逻辑处理器；内存容量：8.00 GB）

2.2.2软件

操作系统：Windows 10家庭版，版本1903

开发环境：Visual Studio 2019 Community（已预装64位Miniconda 3，含64位Python 3.7.4、Conda 4.7.12、NumPy 1.16.5、SciPy 1.3.1、Scikit-learn 0.21.2、Matplotlib 3.1.1）

2.2.3训练和测试数据

训练数据：dB1-2.dat，含10组(x, y)数据。

测试数据：dB1-2-t1.dat，含10组(x, y)数据；dB1-2-t2.dat，含100组(x, y)数据。

**2.2 实验过程**

2.2.1训练和测试数据的读入

从dB1-2.dat读入训练数据，用于训练多项式岭回归模型；从dB1-2-t1.dat和dB1-2-t2.dat读入测试数据，以考察不同多项式次数下岭回归模型的最佳训练结果和相应的测试结果。为保证所作的折线图的准确性，我们用自己实现的快速排序，按从小到大的顺序，对训练数据和测试数据进行排序。

2.2.2计算并绘制不同多项式次数下最佳训练结果和相应的测试结果

使用Scikit-learn库中的RidgeCV（带交叉验证的岭回归），结合PolynomialFeatures，对训练数据进行多项式拟合，得到不同的多项式次数下最优的值，计算该值下模型预测值与训练数据的RMS，并将此时的预测值与训练数据绘制在同一张图上。

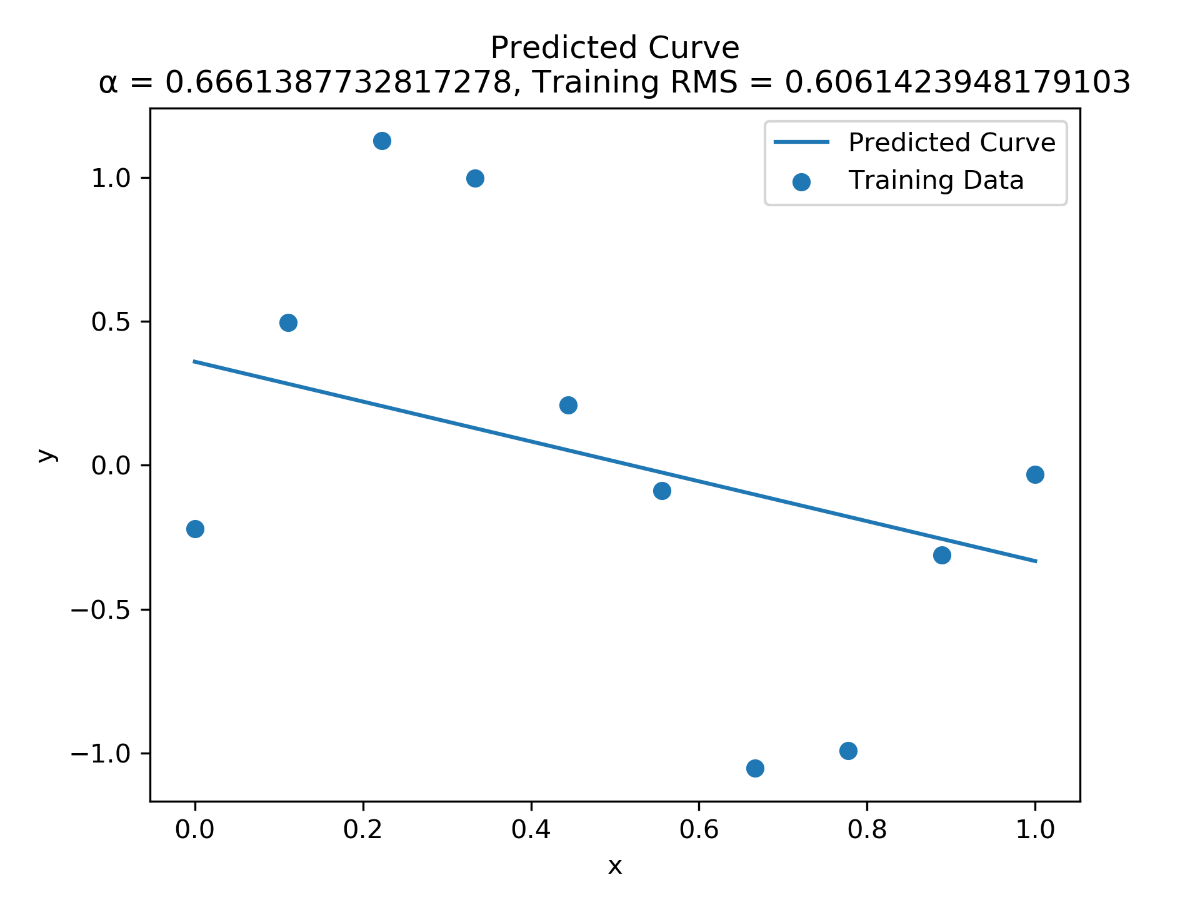
2.2.3计算并绘制和的关系图

使用Scikit-learn库中的Ridge（岭回归），结合PolynomialFeatures，在不同值和多项式次数下对训练数据进行多项式拟合，然后分别计算预测值与训练数据、测试数据1、测试数据2的RMS，并输出至dLR1a.dat和dLR1b.dat，绘制和的关系图。

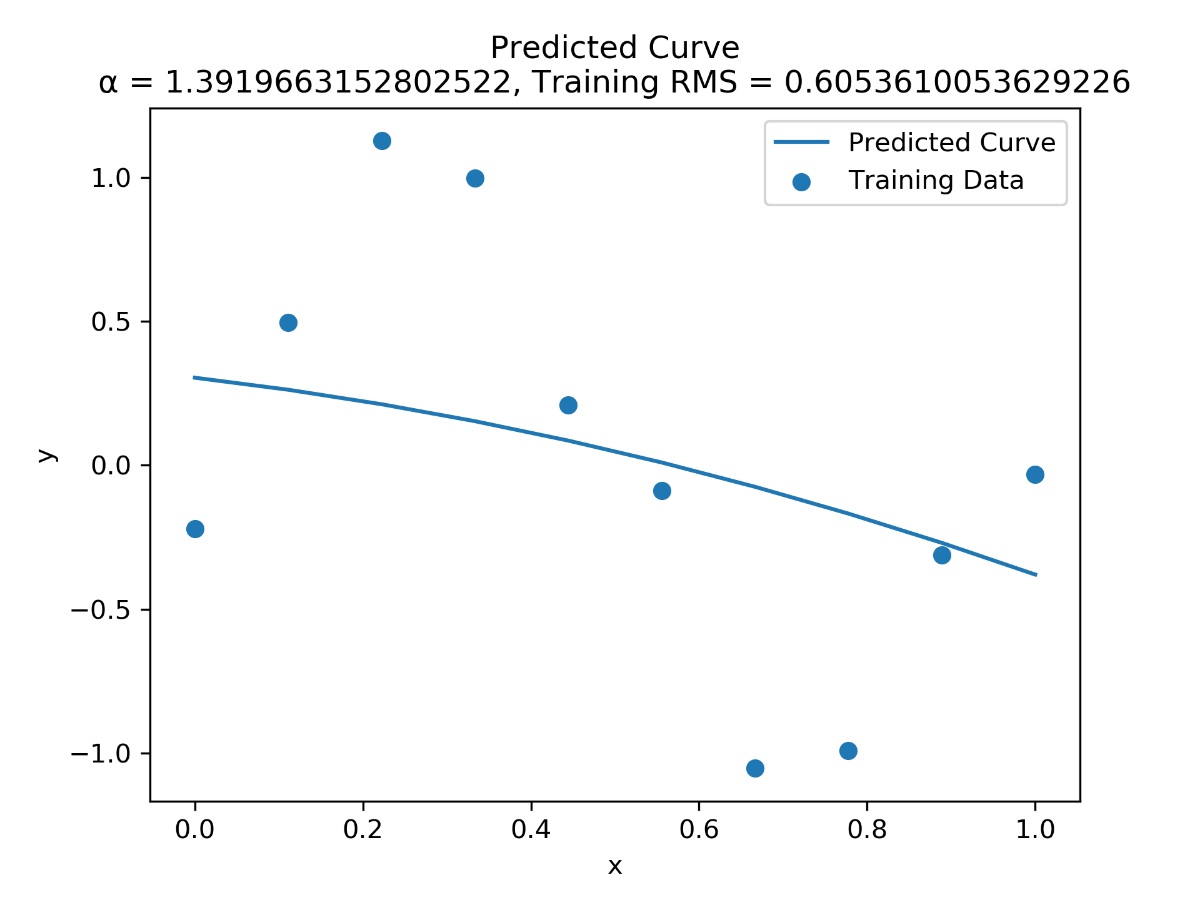
**3 数据与结果**

**3.1不同多项式次数下最佳训练结果和相应的测试结果**

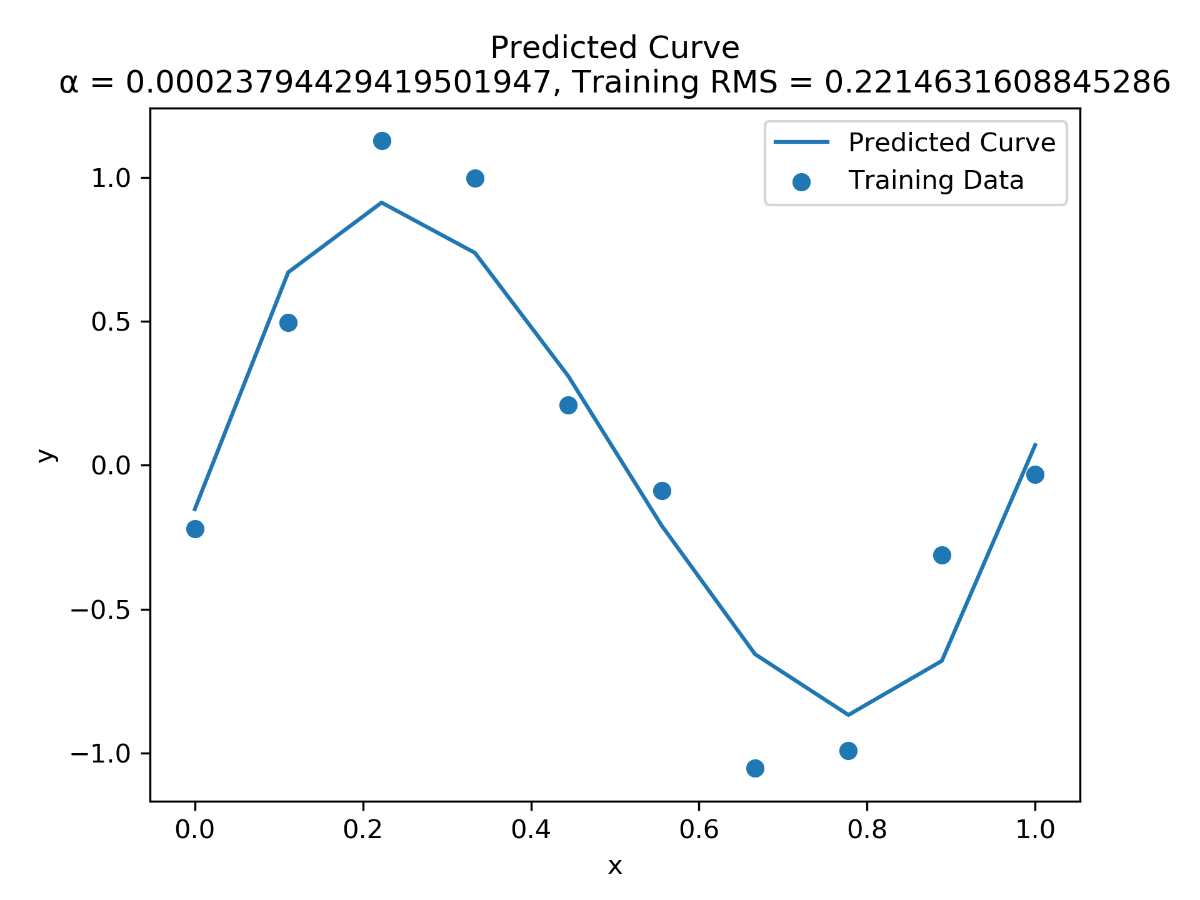
读入数据后，按2.2.2进行实验，其中多项式次数从1取到10，值取值集合为范围从到，项数为5000个的等比数列。最佳训练结果如图1-图10所示：



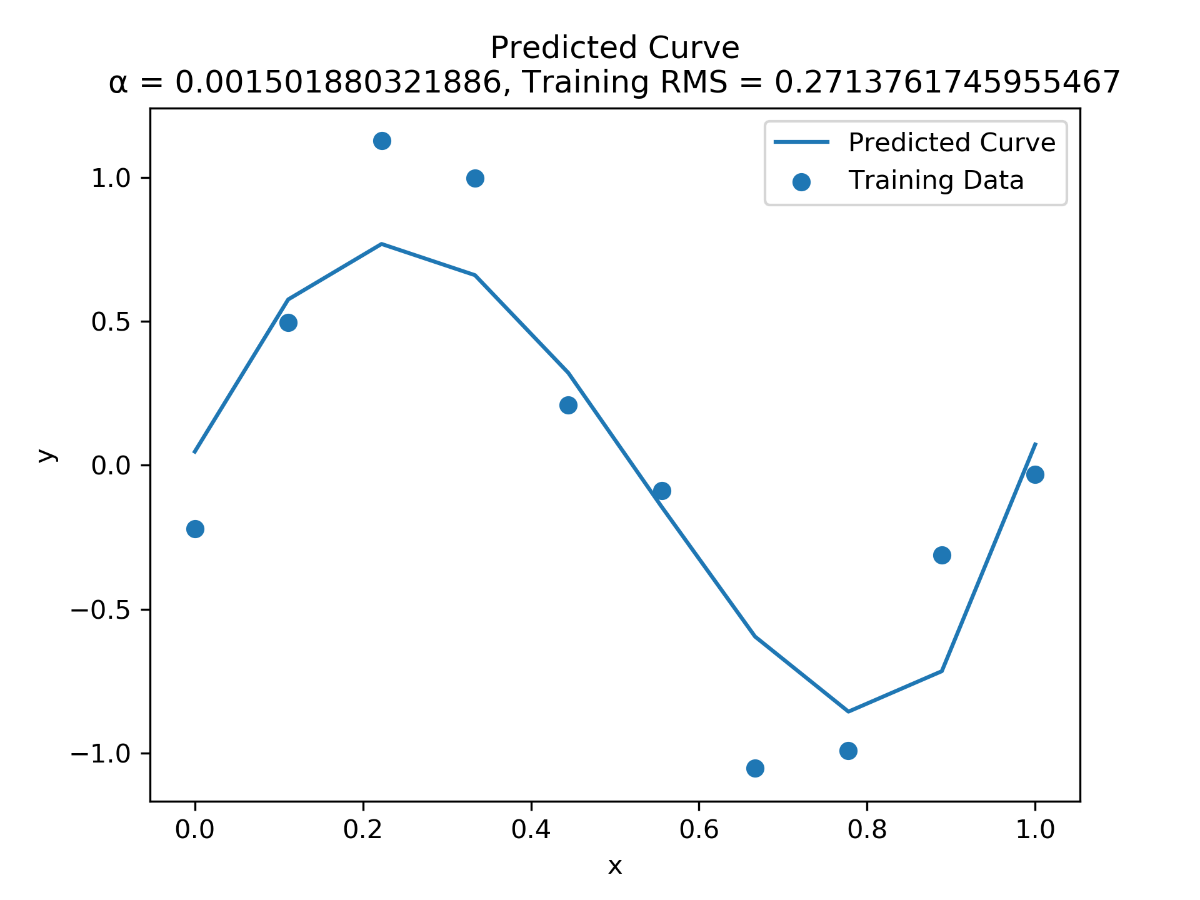
**图1** 多项式次数为1时最佳训练结果



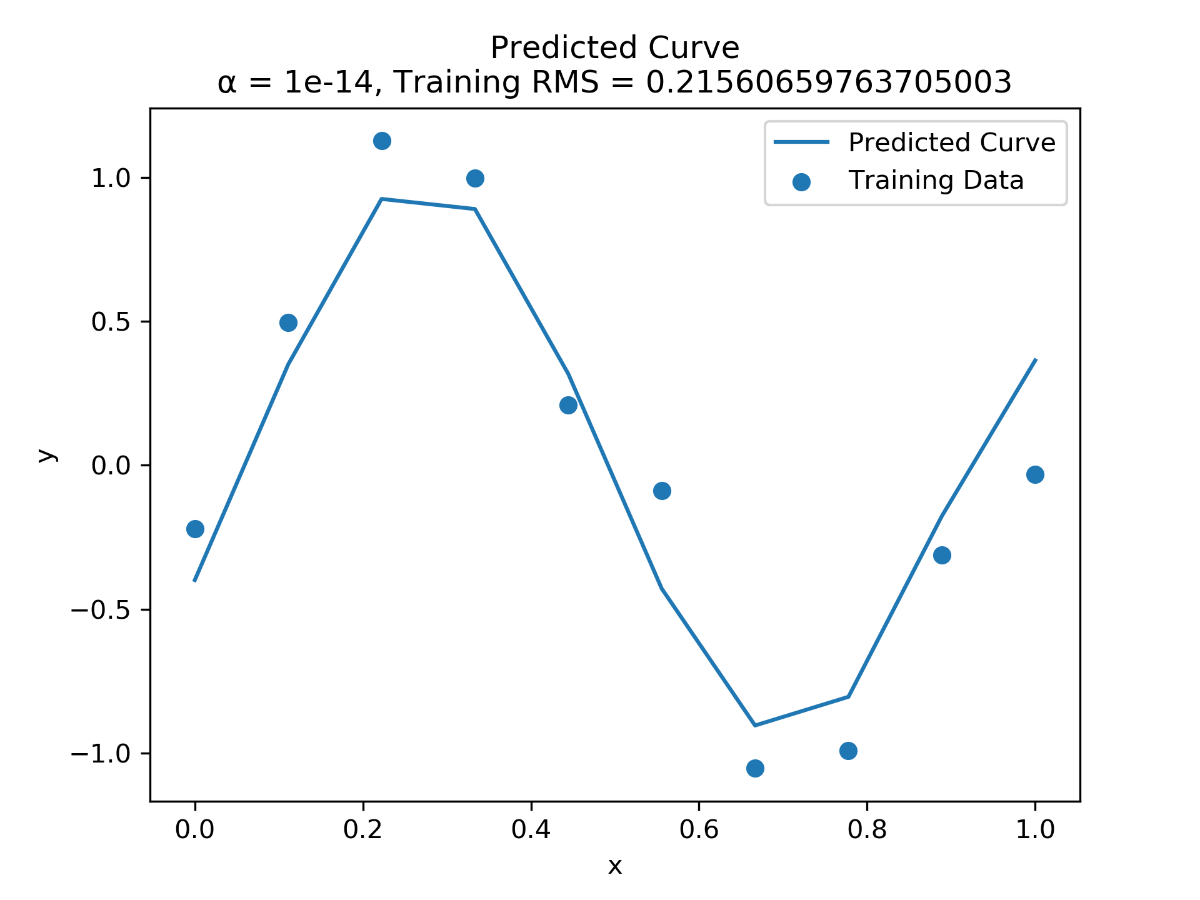
**图2** 多项式次数为2时最佳训练结果



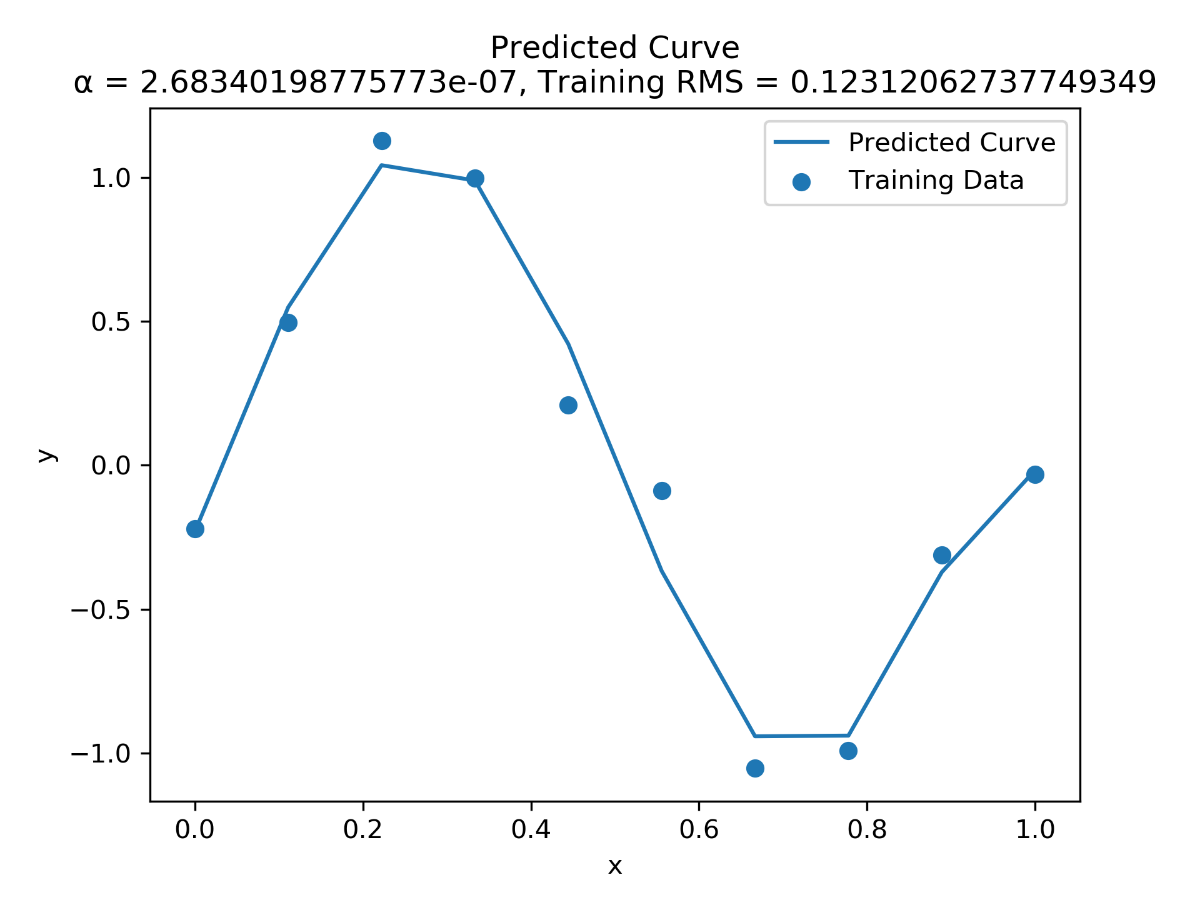
**图3** 多项式次数为3时最佳训练结果



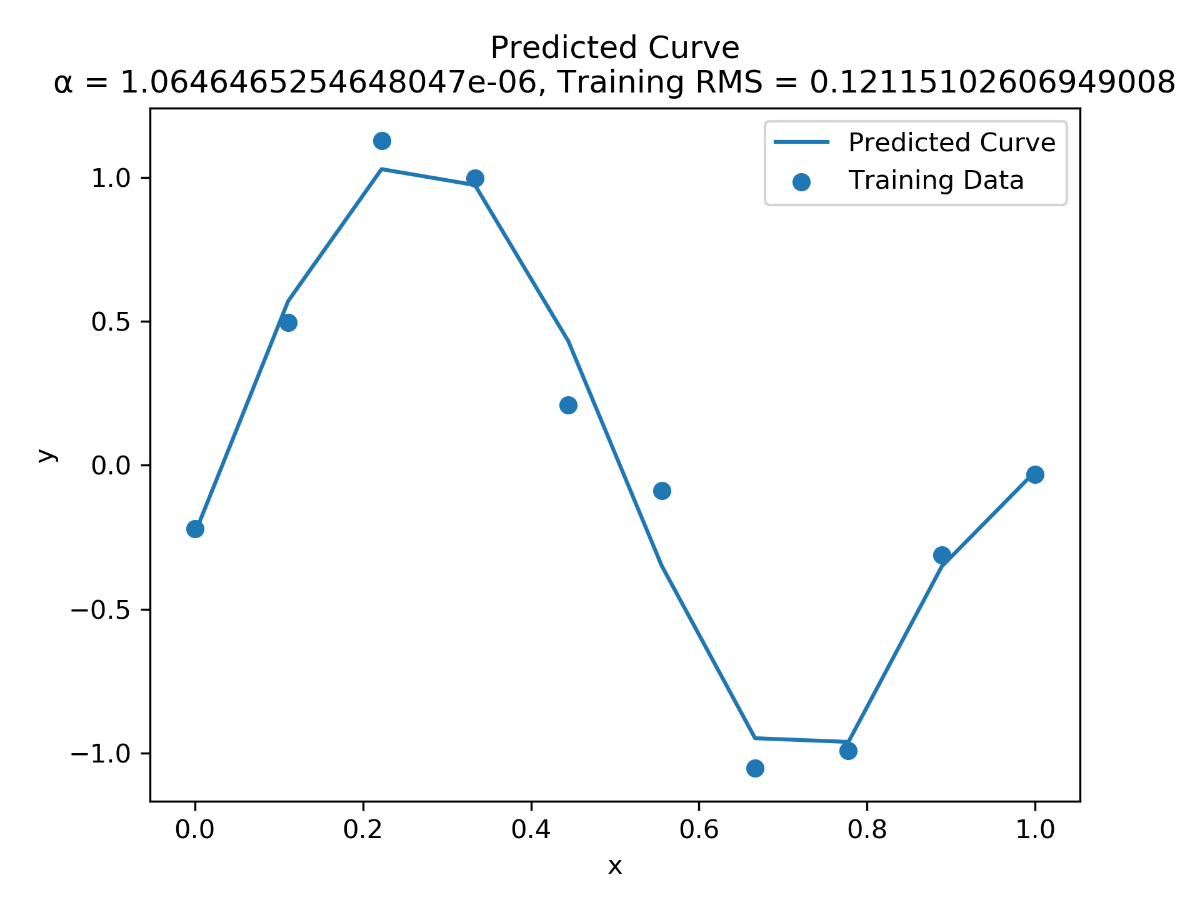
**图4** 多项式次数为4时最佳训练结果



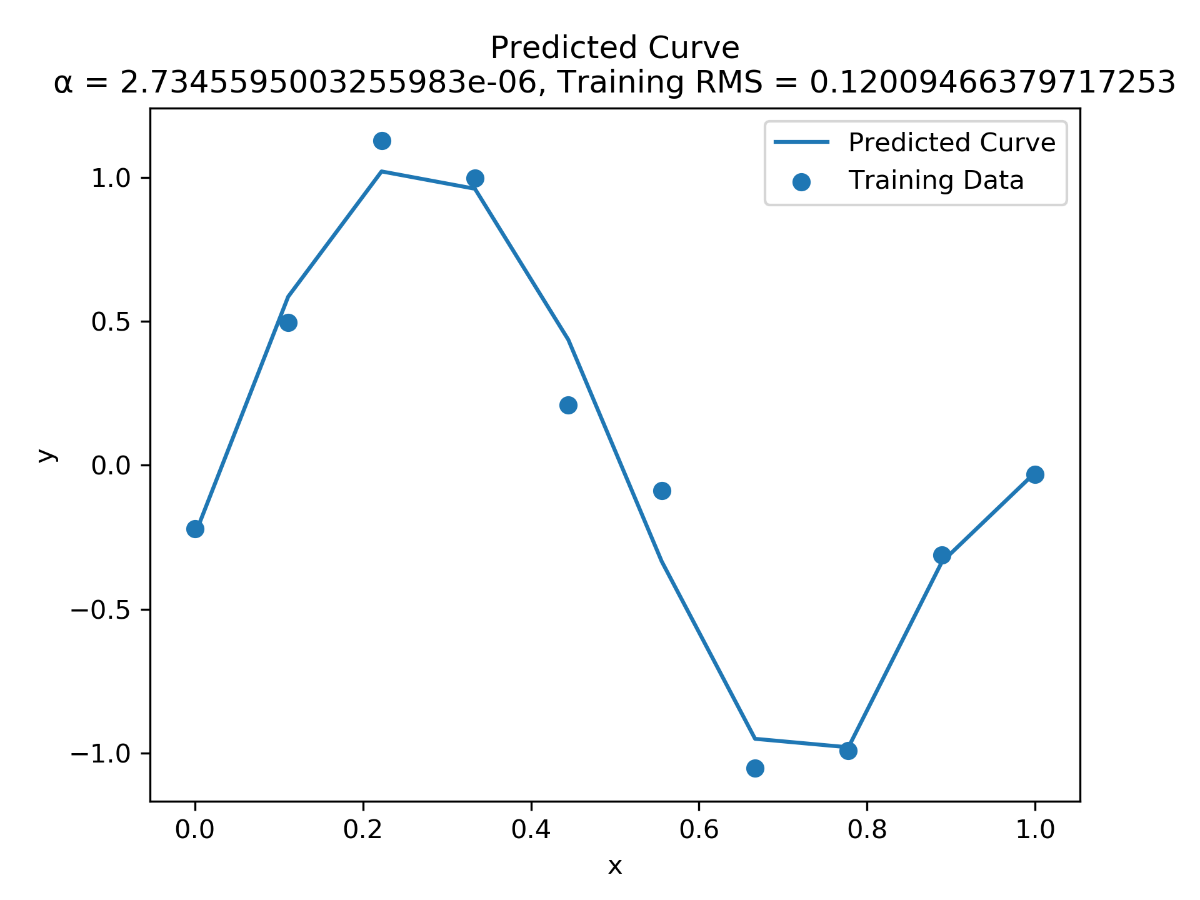
**图5** 多项式次数为5时最佳训练结果



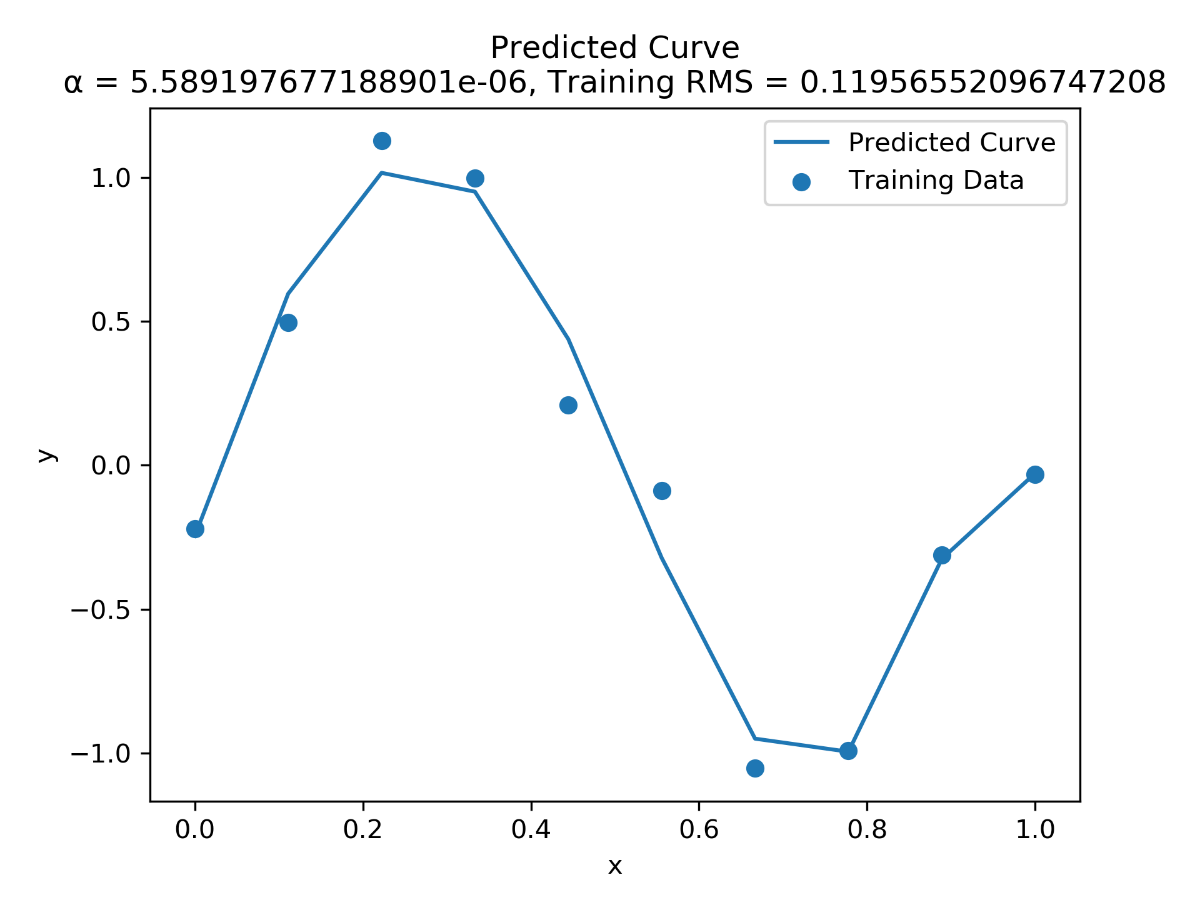
**图6** 多项式次数为6时最佳训练结果



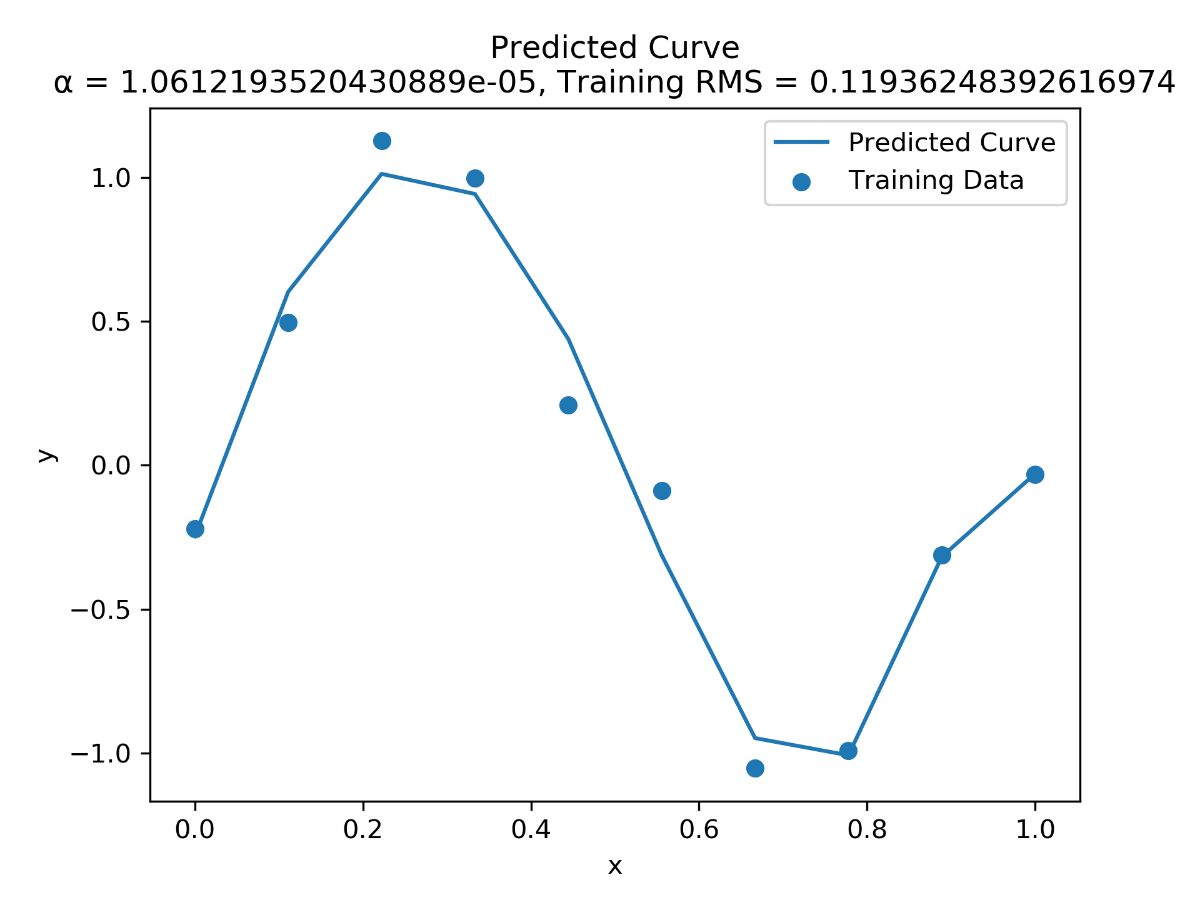
**图7** 多项式次数为7时最佳训练结果



**图8** 多项式次数为8时最佳训练结果

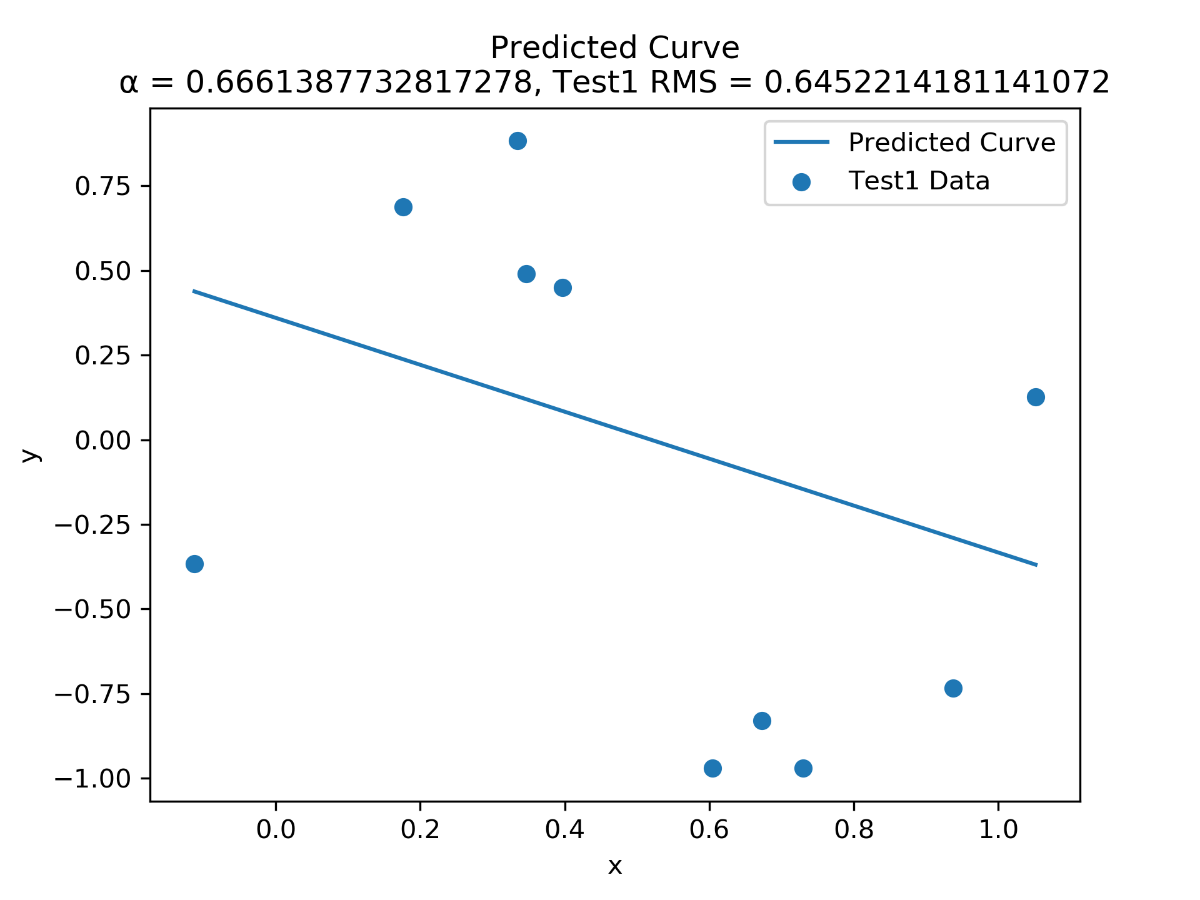


**图9** 多项式次数为9时最佳训练结果

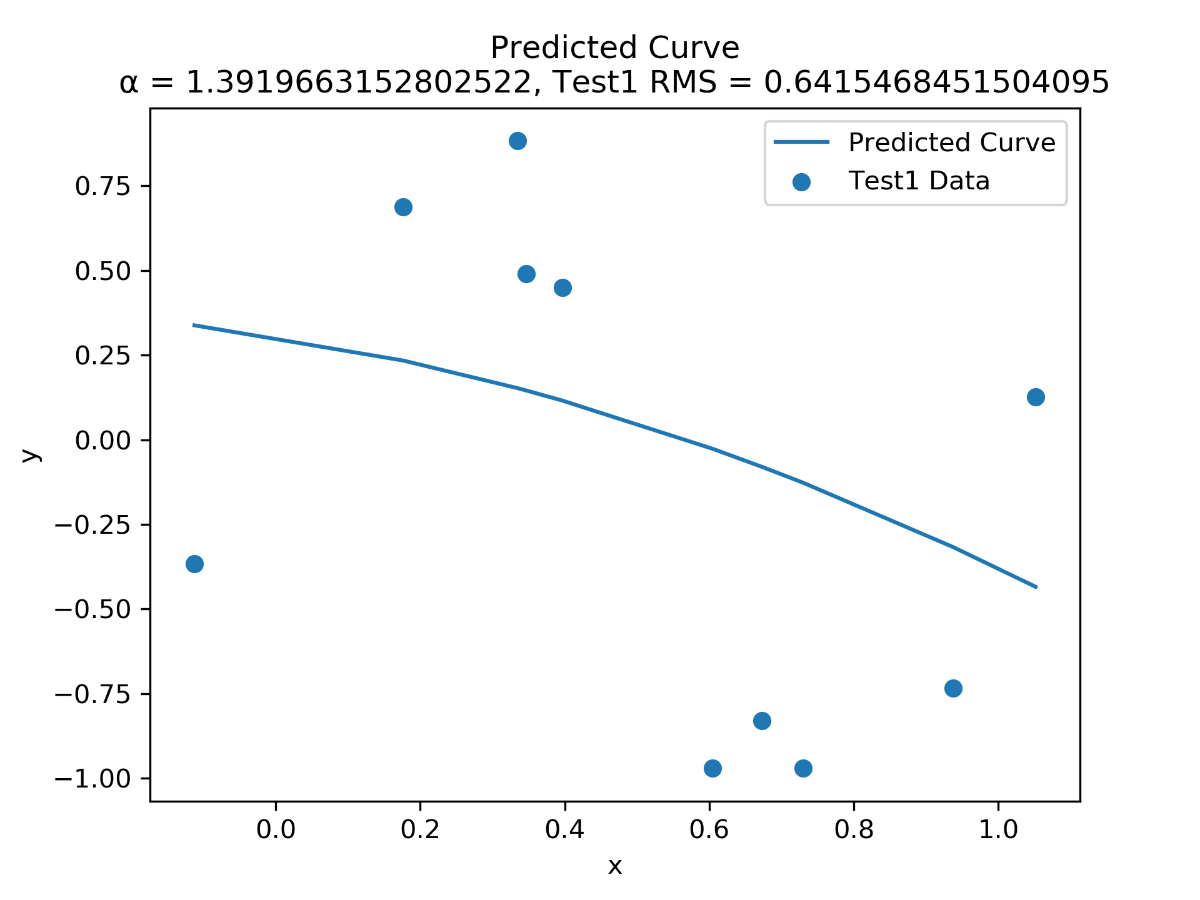


**图10** 多项式次数为10时最佳训练结果

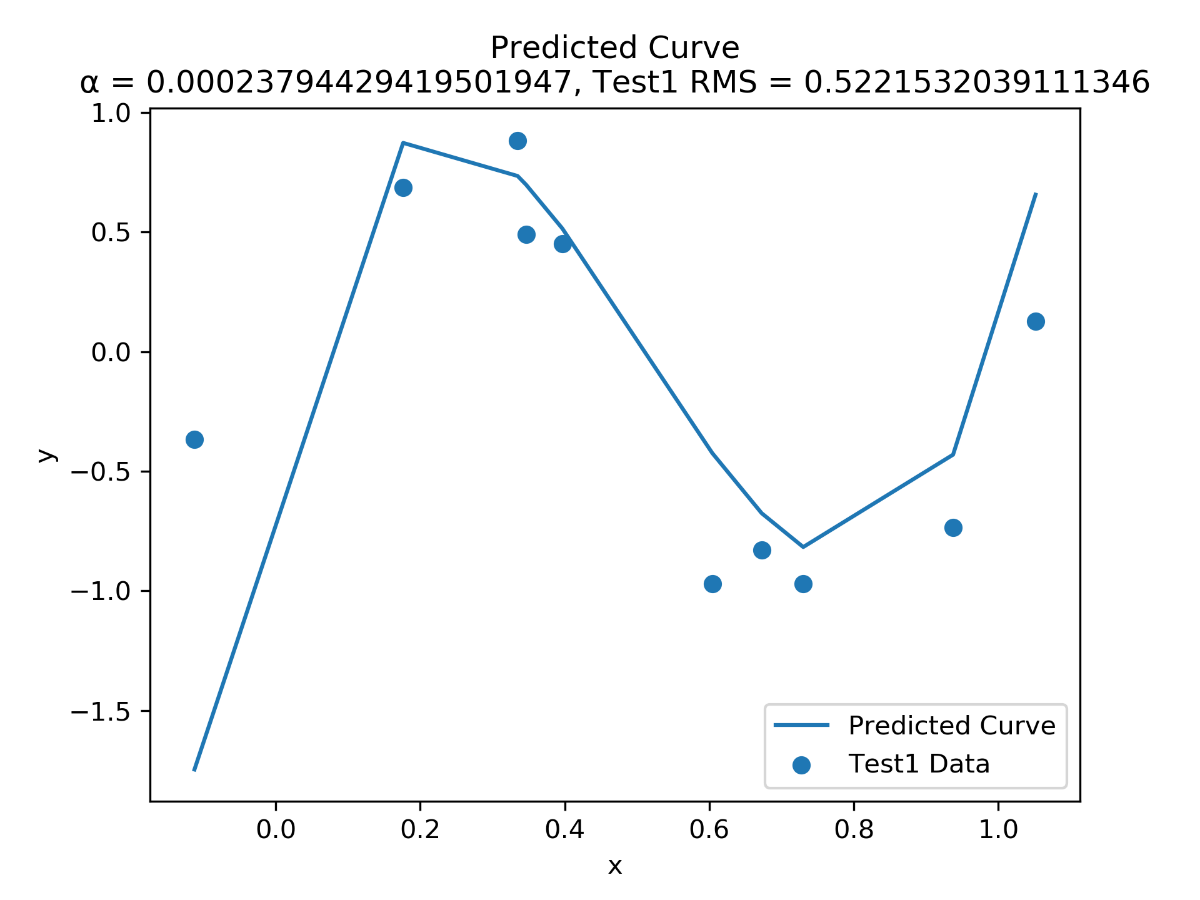
在最佳训练结果对应的值下，相应的测试结果1如图11-图20所示：



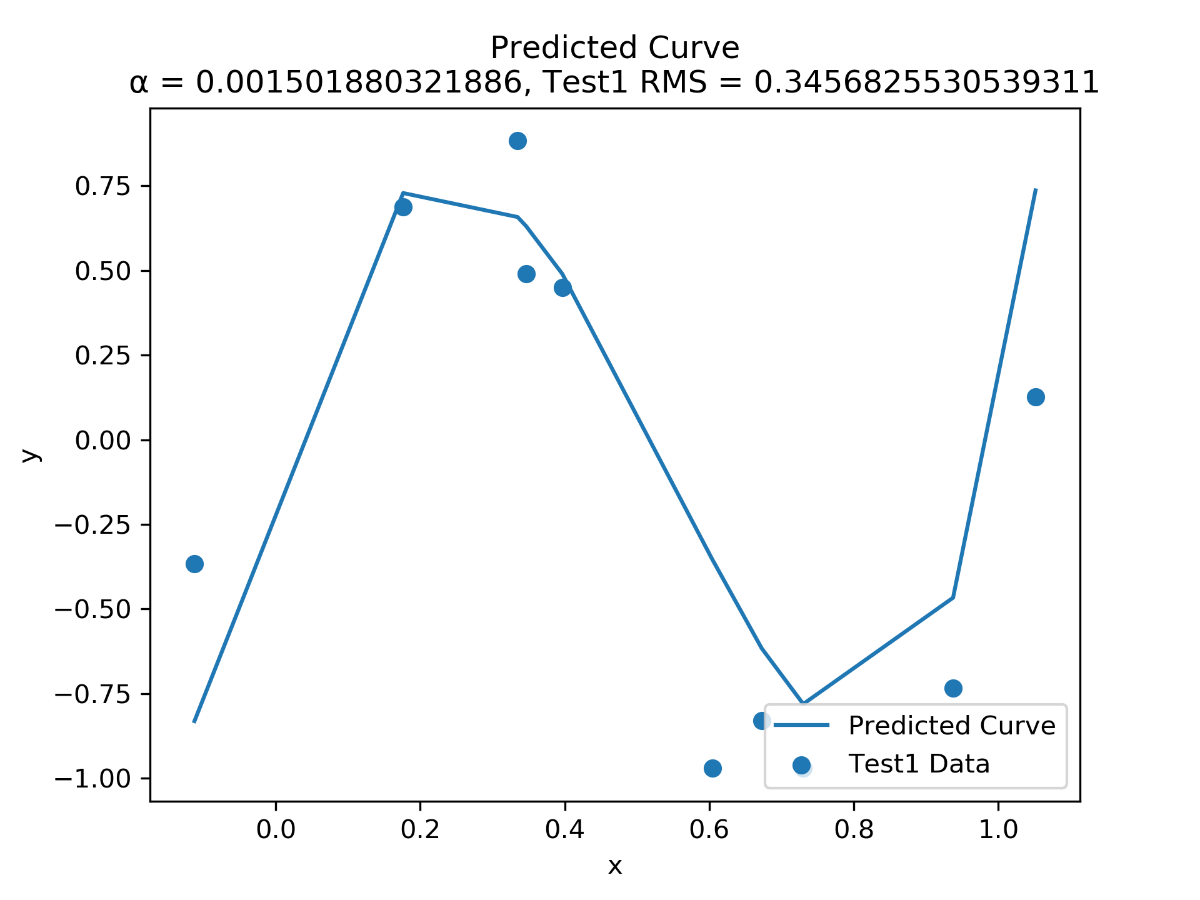
**图11** 多项式次数为1时测试结果1



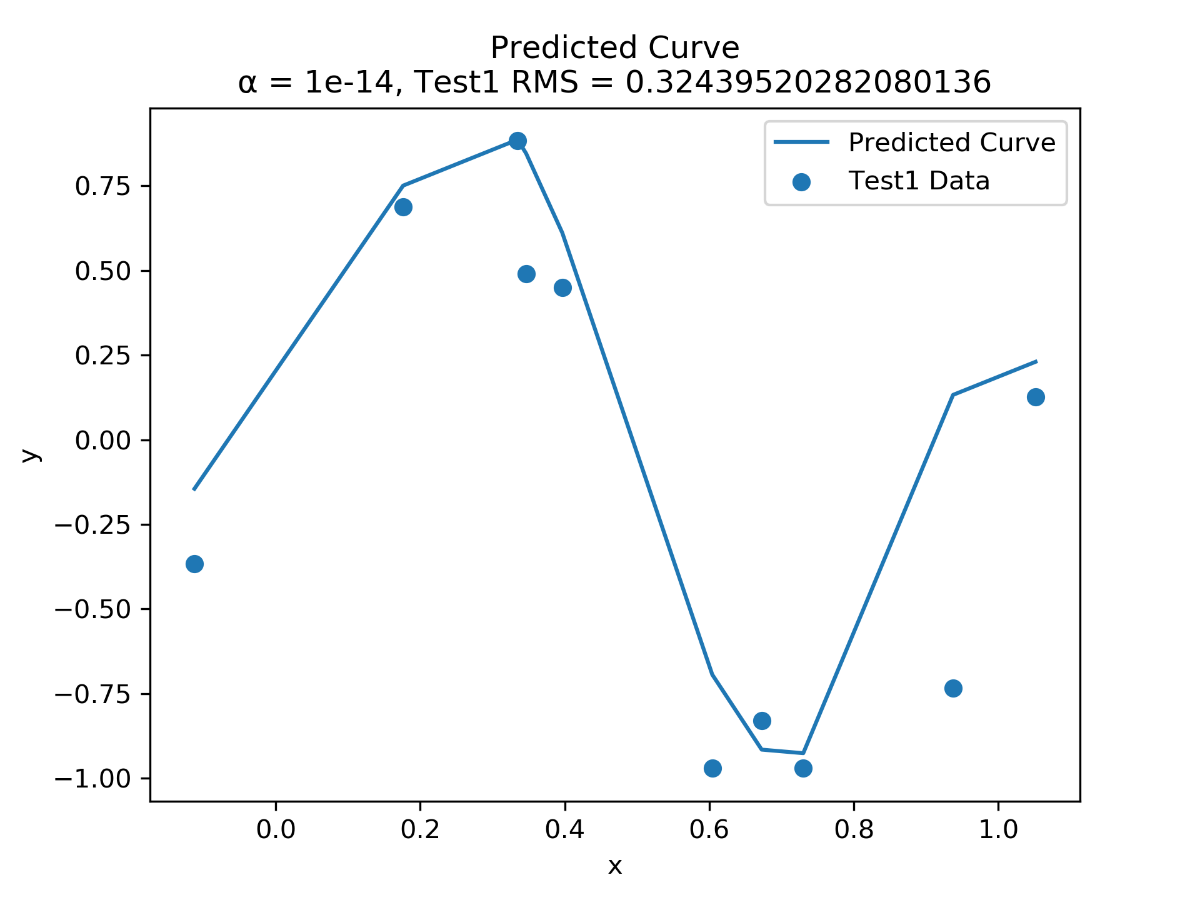
**图12** 多项式次数为2时测试结果1



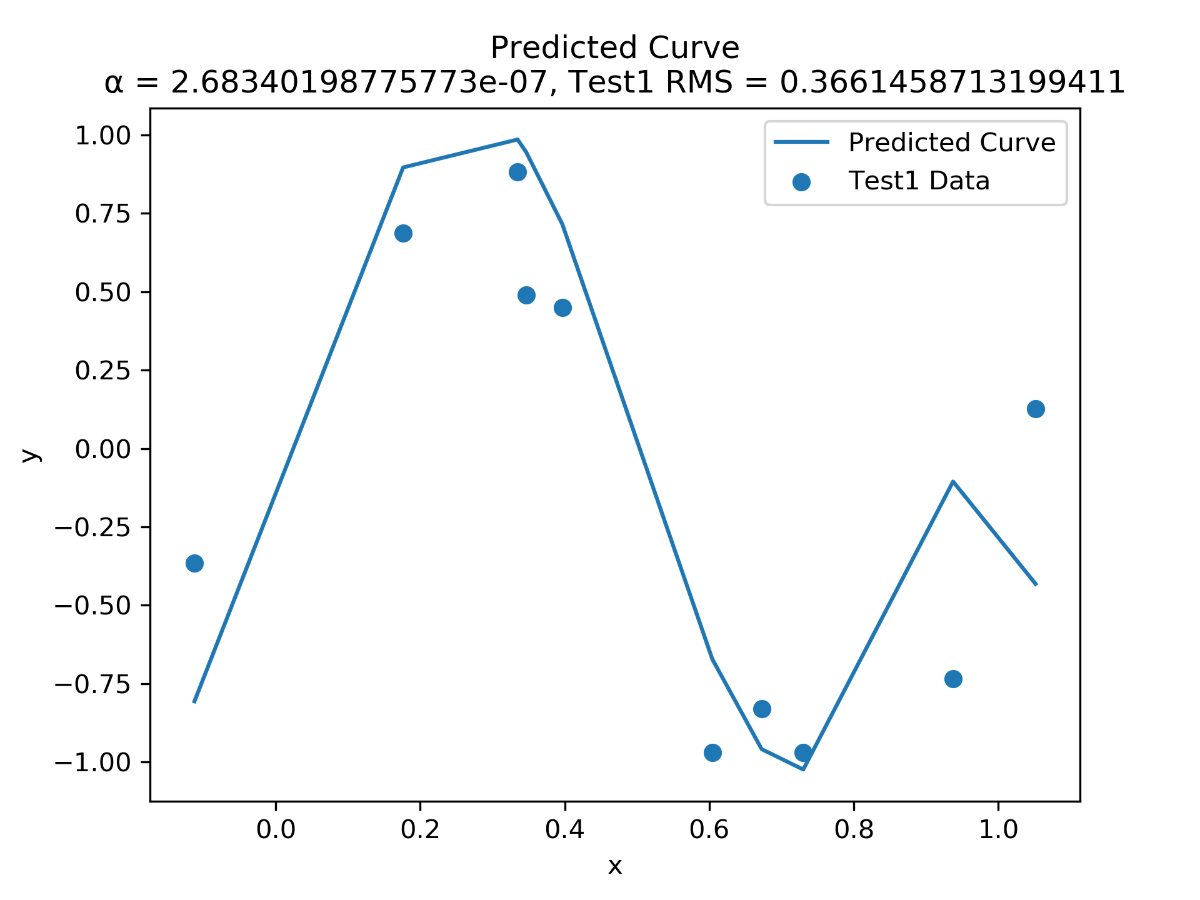
**图13** 多项式次数为3时测试结果1



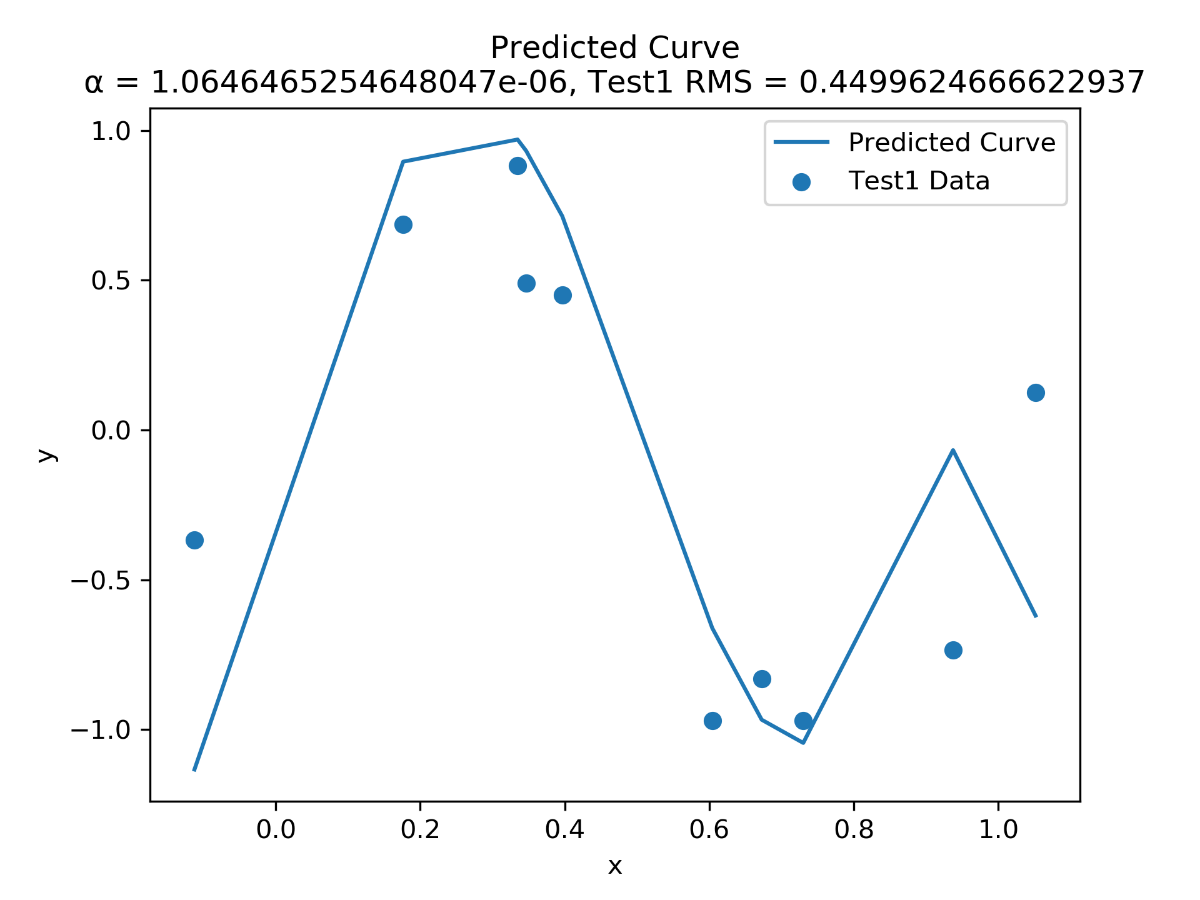
**图14** 多项式次数为4时测试结果1



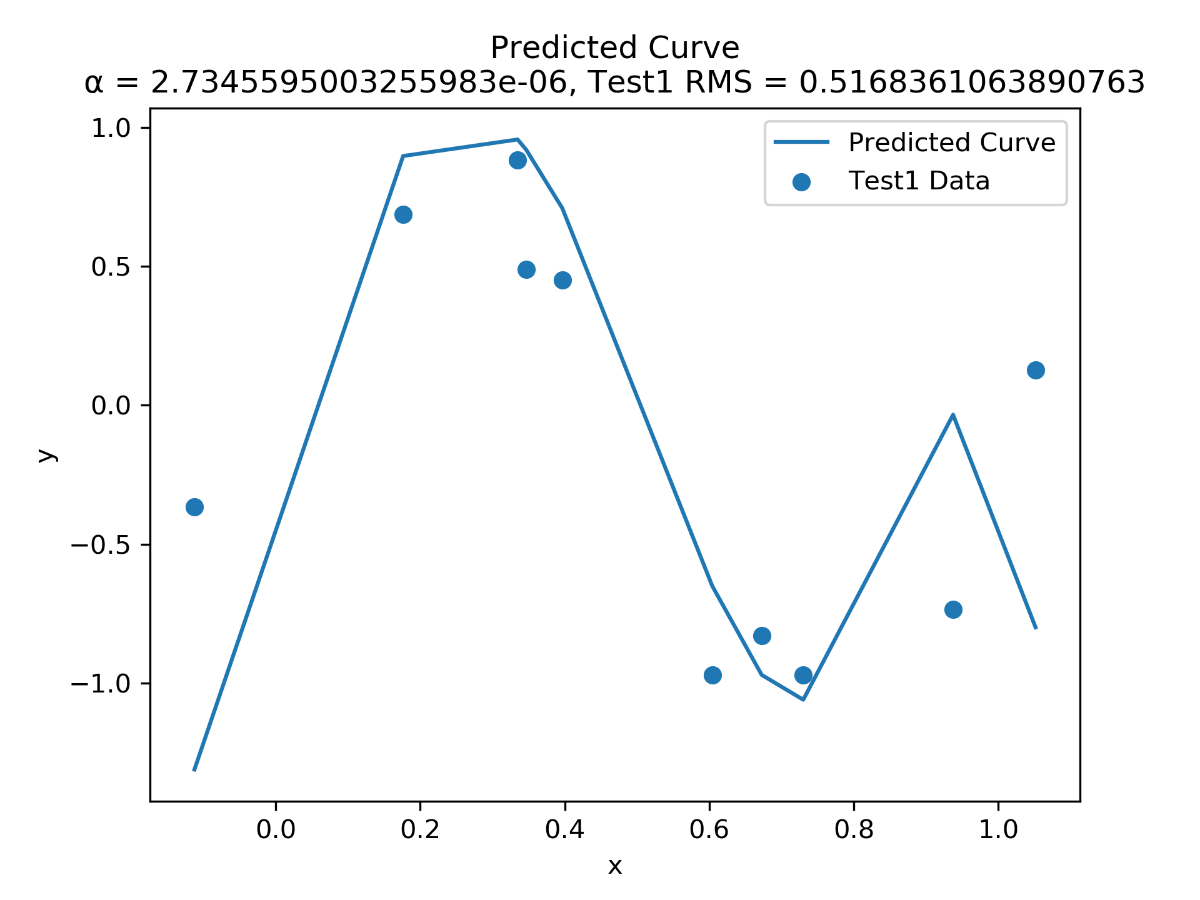
**图15** 多项式次数为5时测试结果1



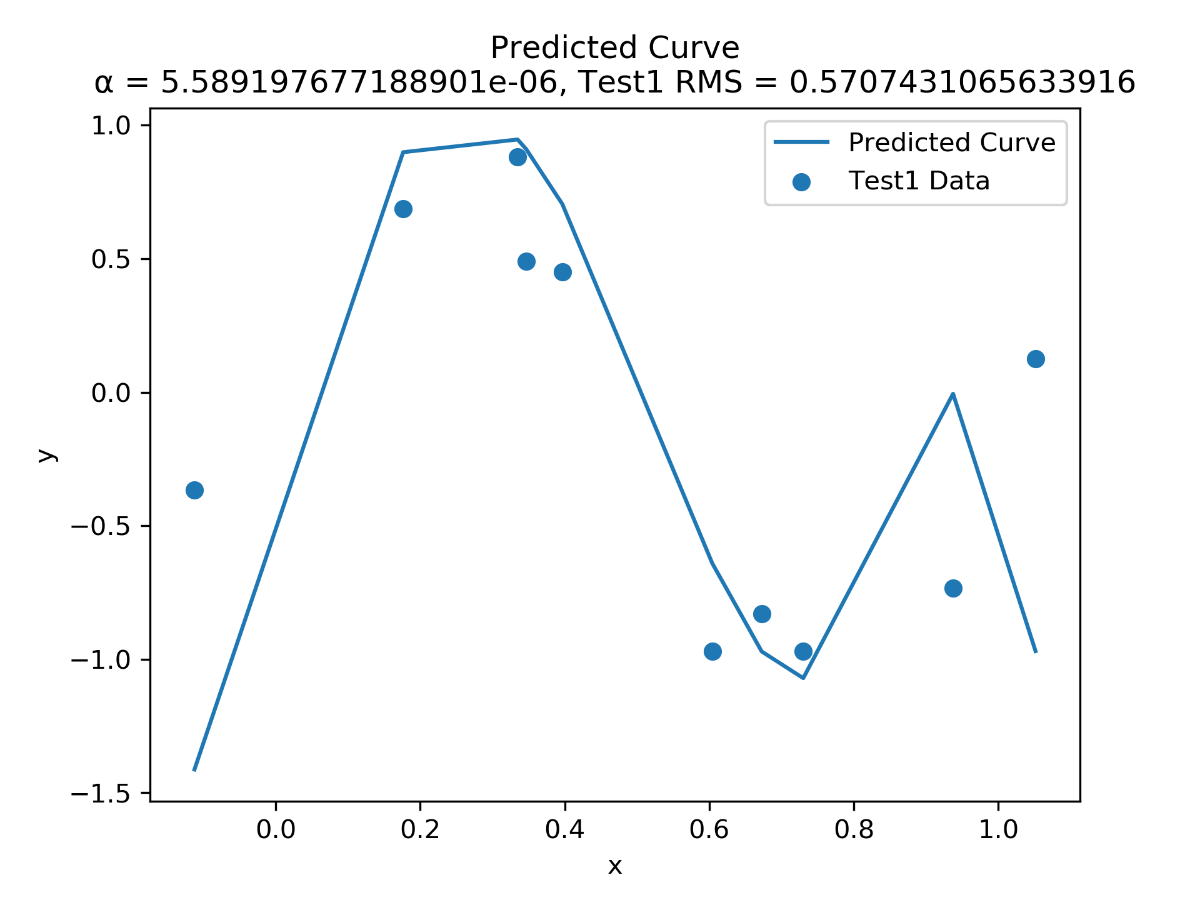
**图16** 多项式次数为6时测试结果1



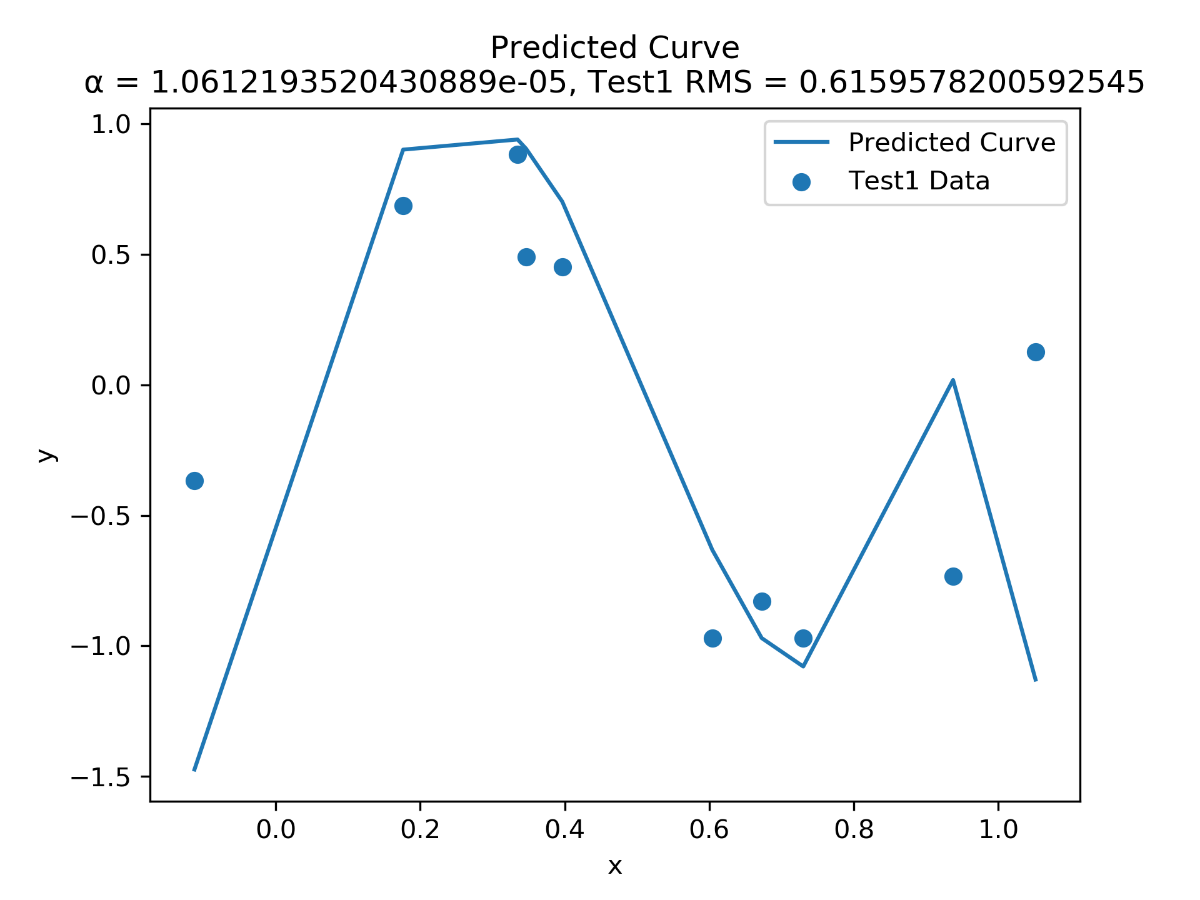
**图17** 多项式次数为7时测试结果1



**图18** 多项式次数为8时测试结果1

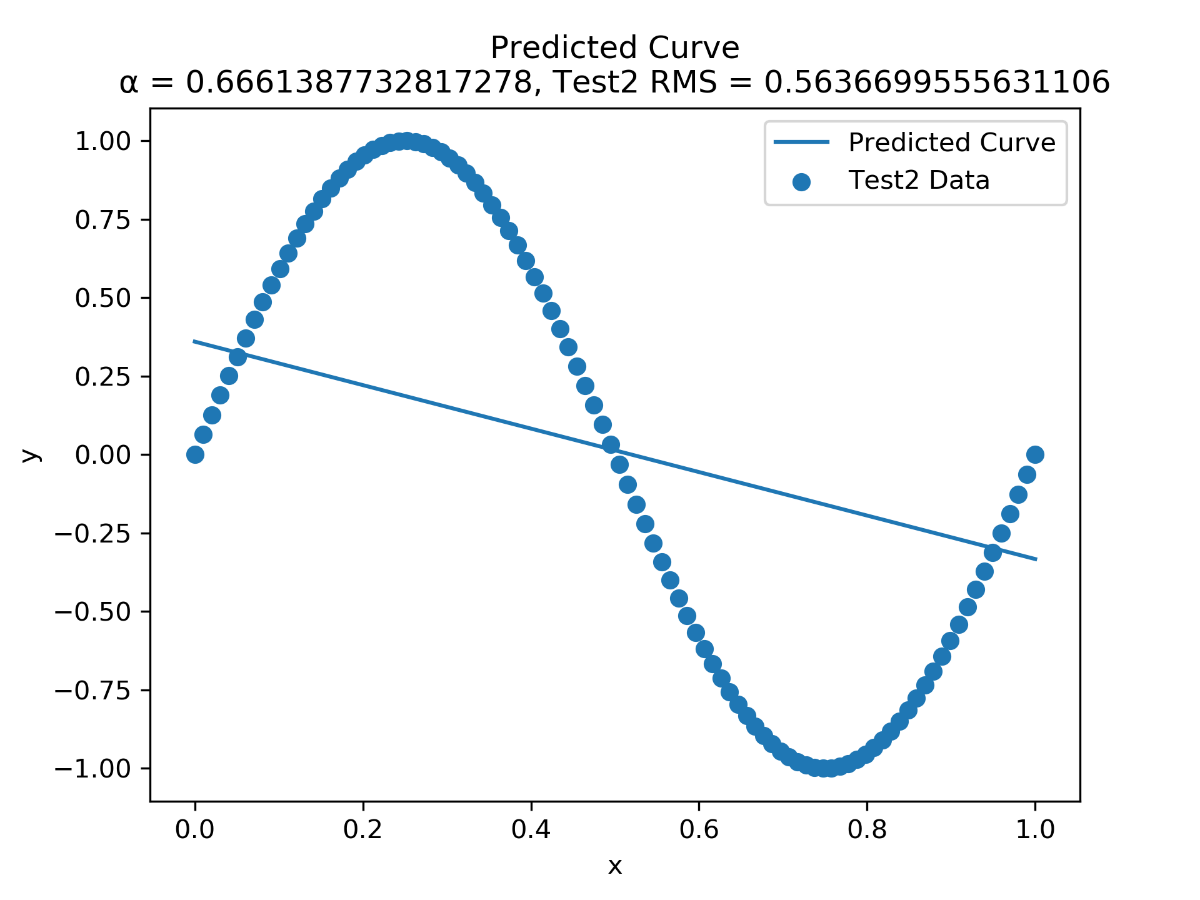


**图19** 多项式次数为9时测试结果1

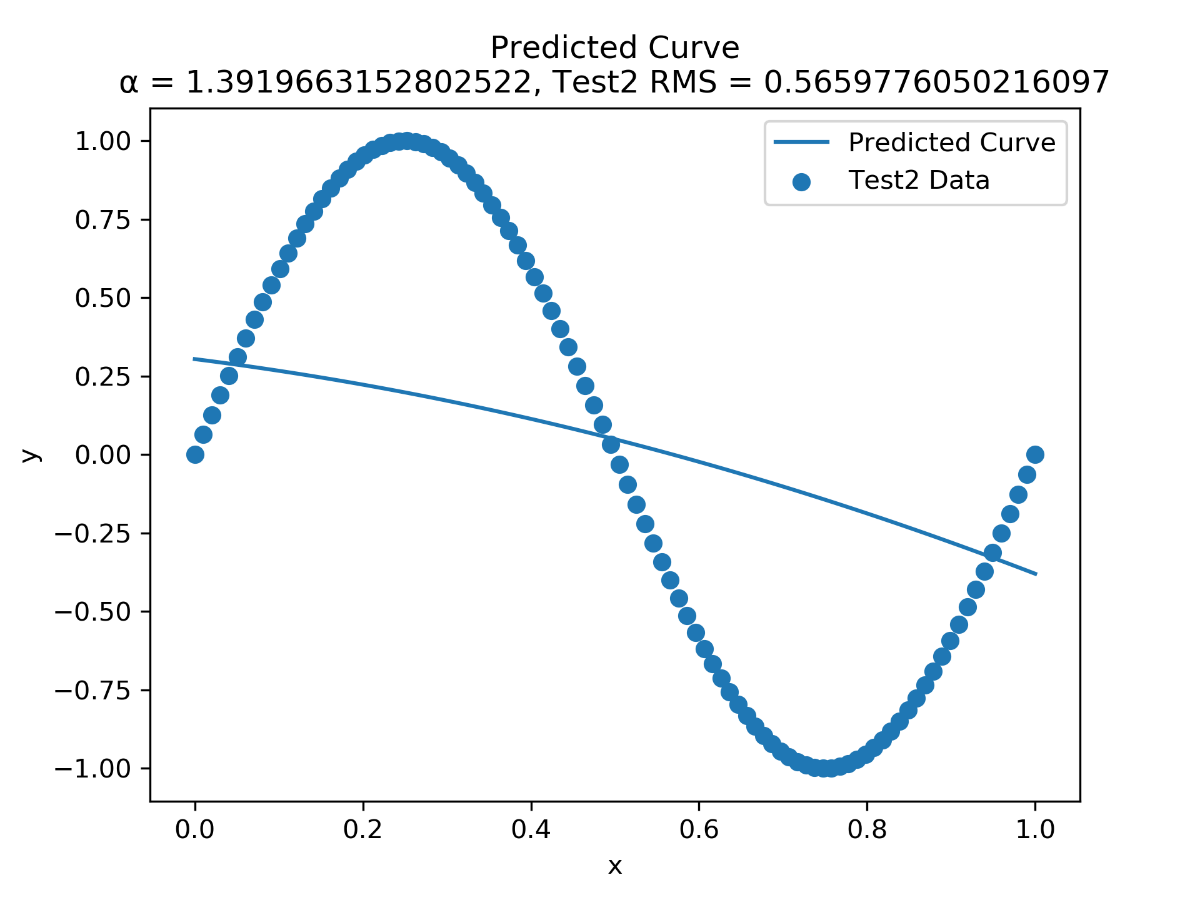


**图20** 多项式次数为10时测试结果1

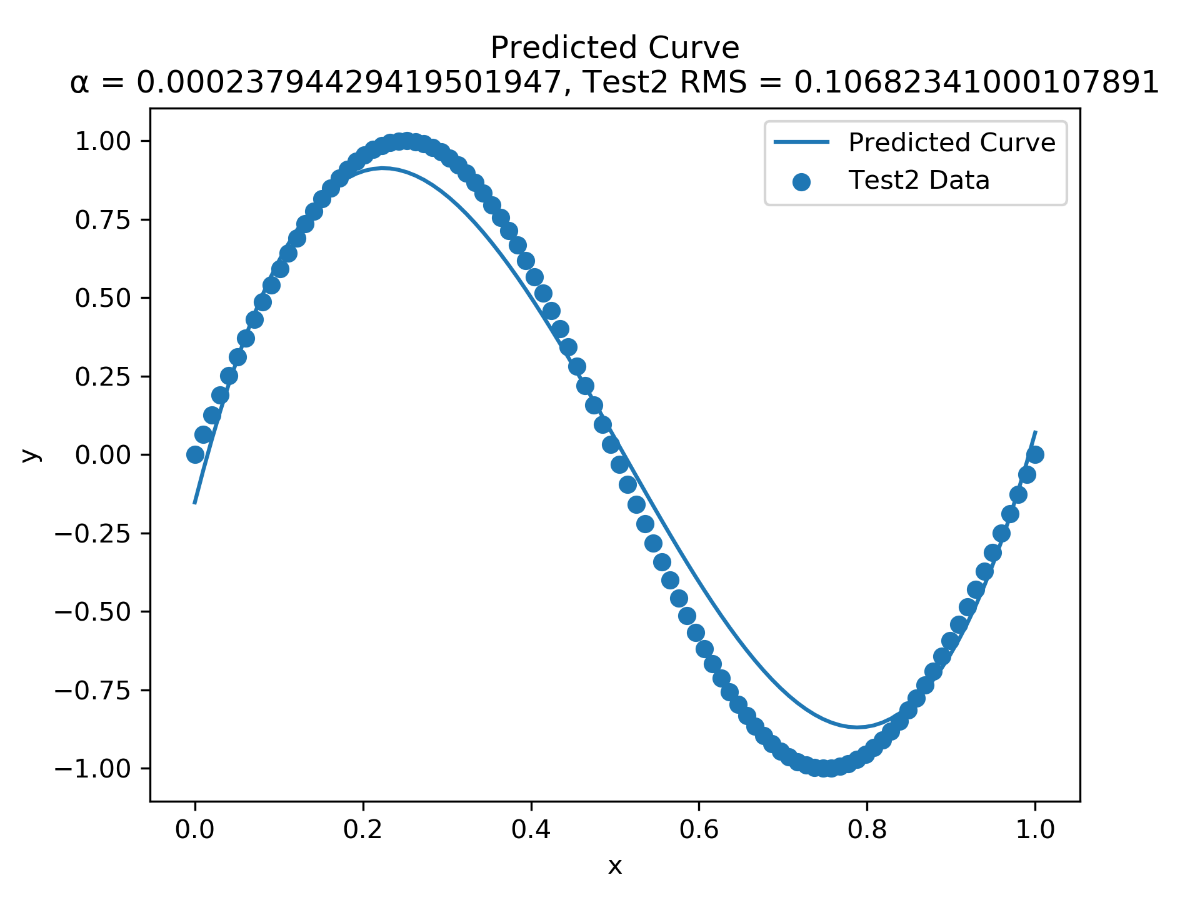
在最佳训练结果对应的值下，相应的测试结果2如图21-图30所示：



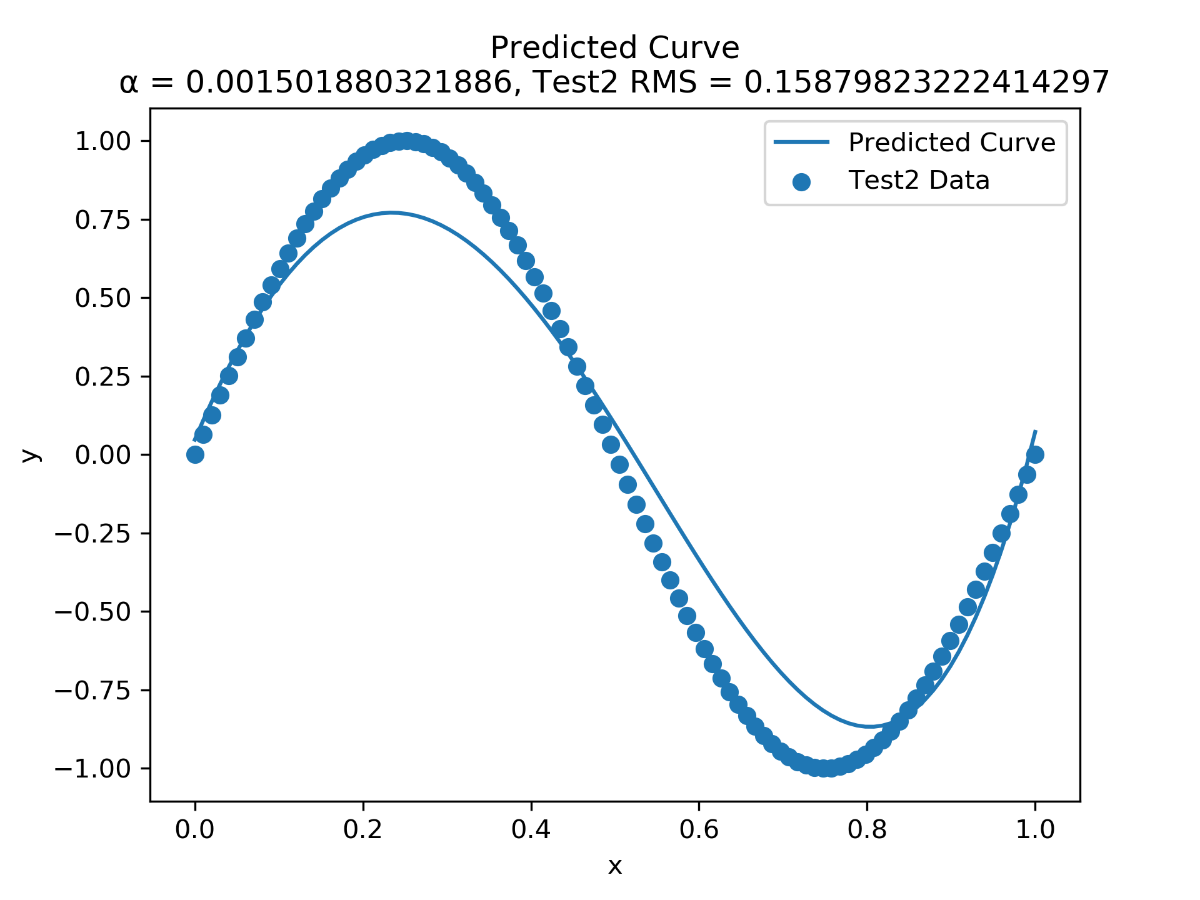
**图21** 多项式次数为1时测试结果2



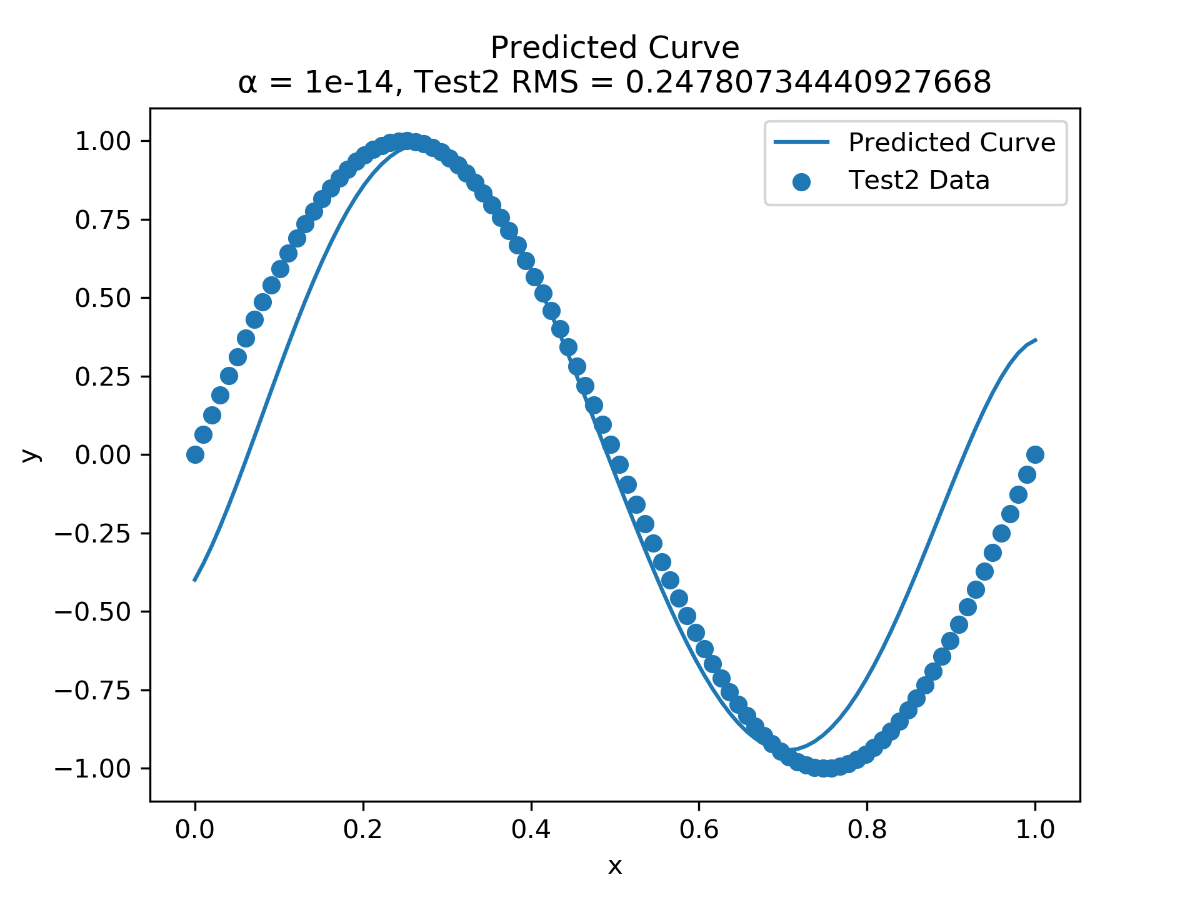
**图22** 多项式次数为2时测试结果2



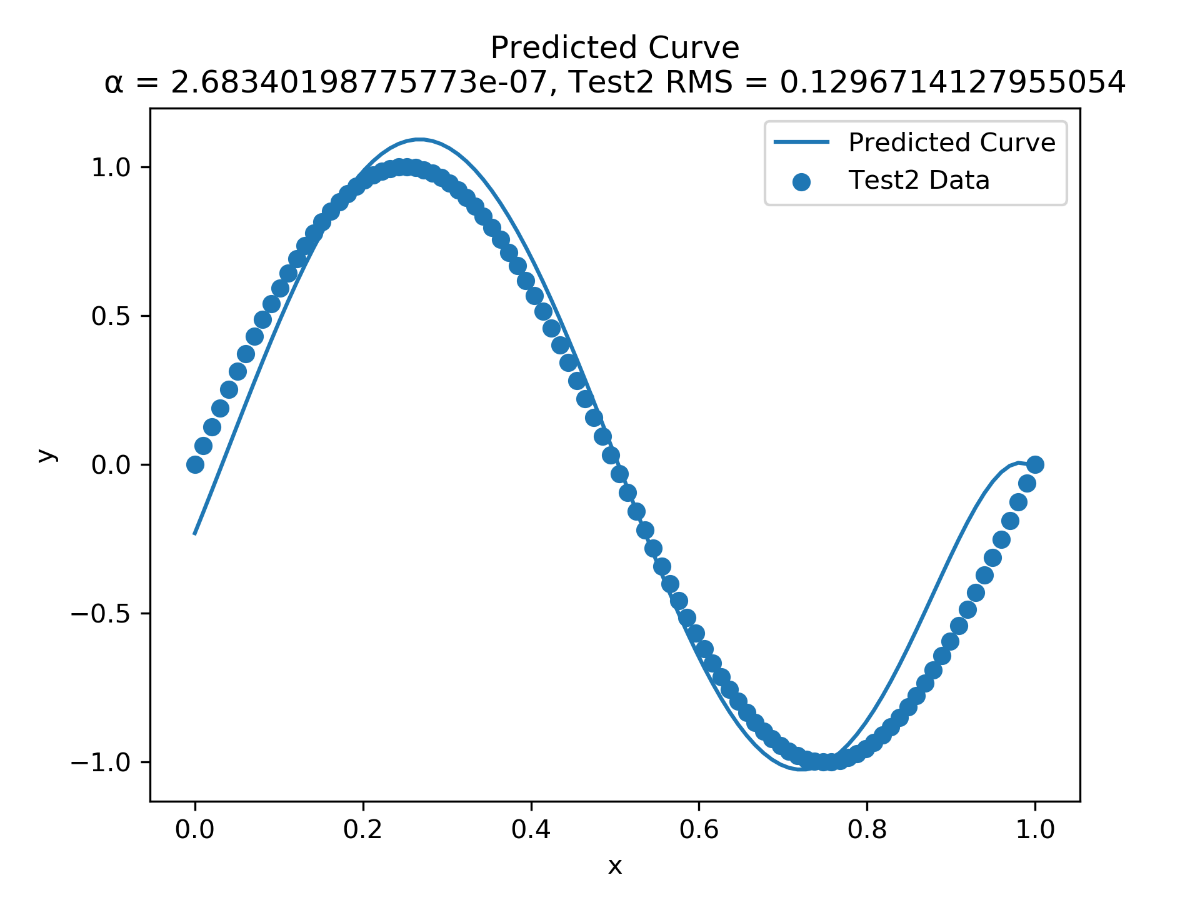
**图23** 多项式次数为3时测试结果2



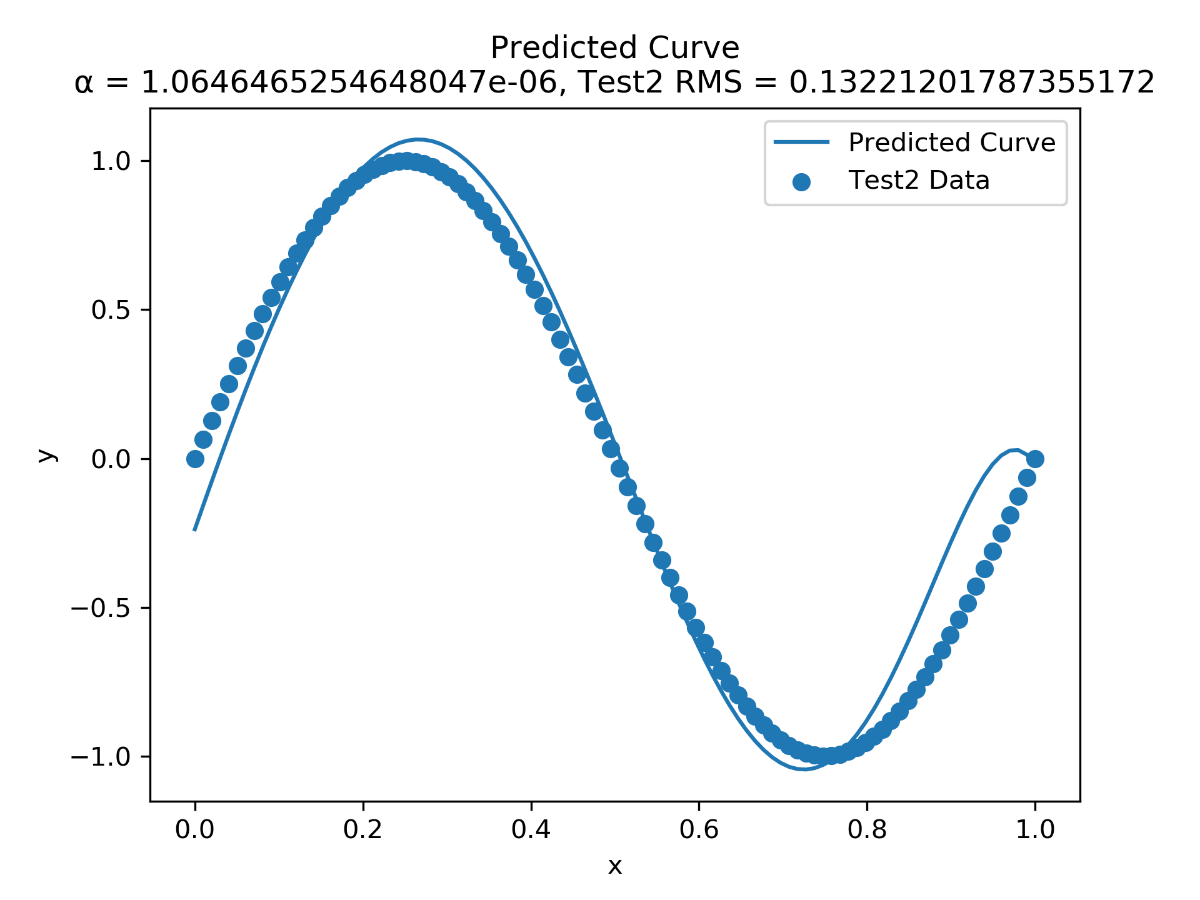
**图24** 多项式次数为4时测试结果2



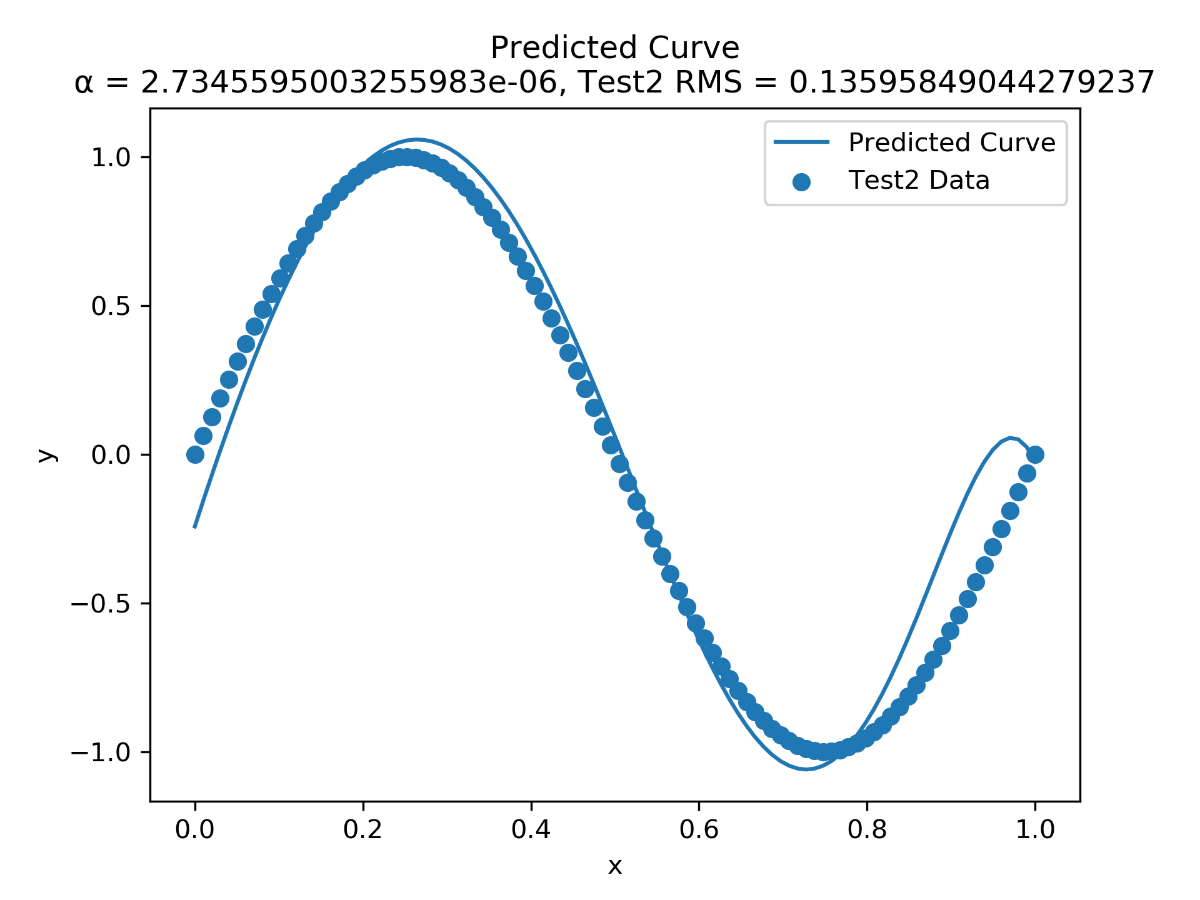
**图25** 多项式次数为5时测试结果2



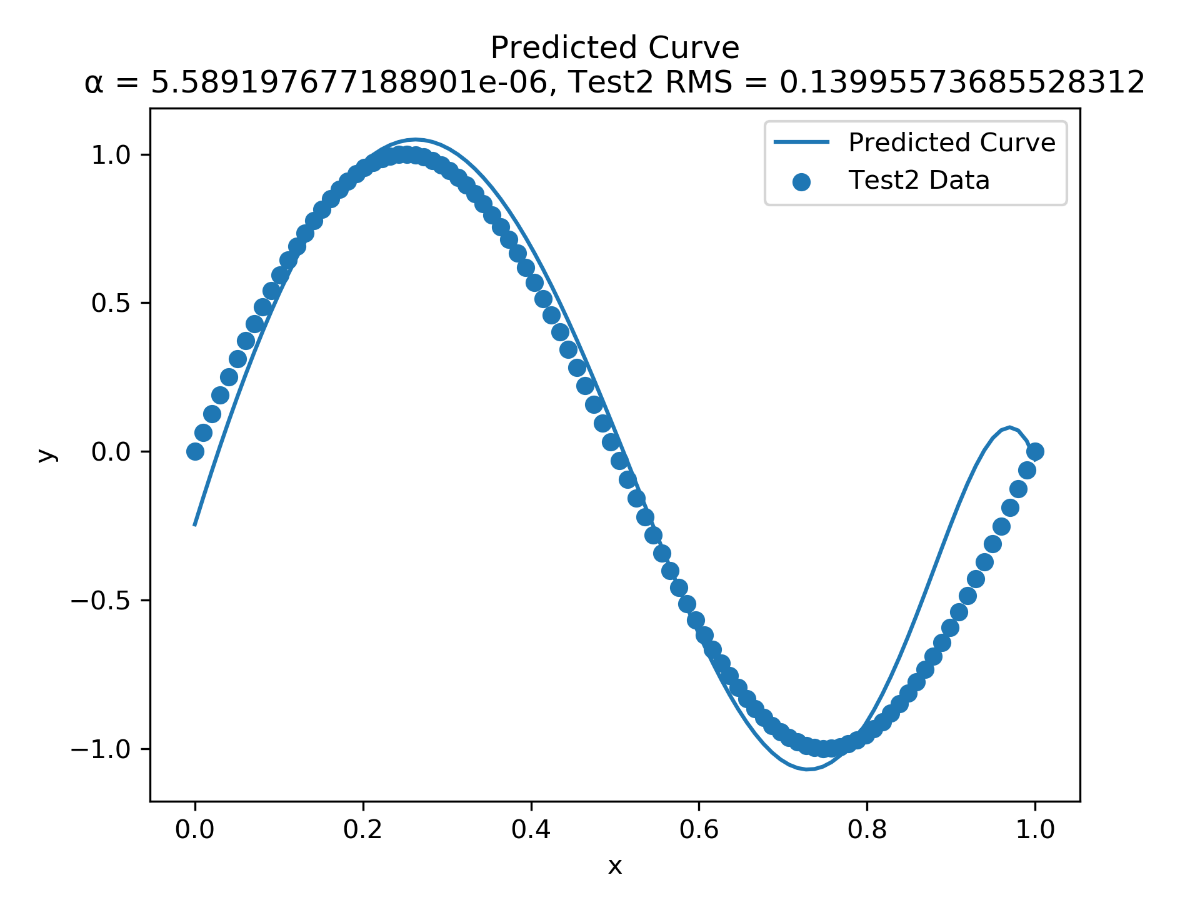
**图26** 多项式次数为6时测试结果2



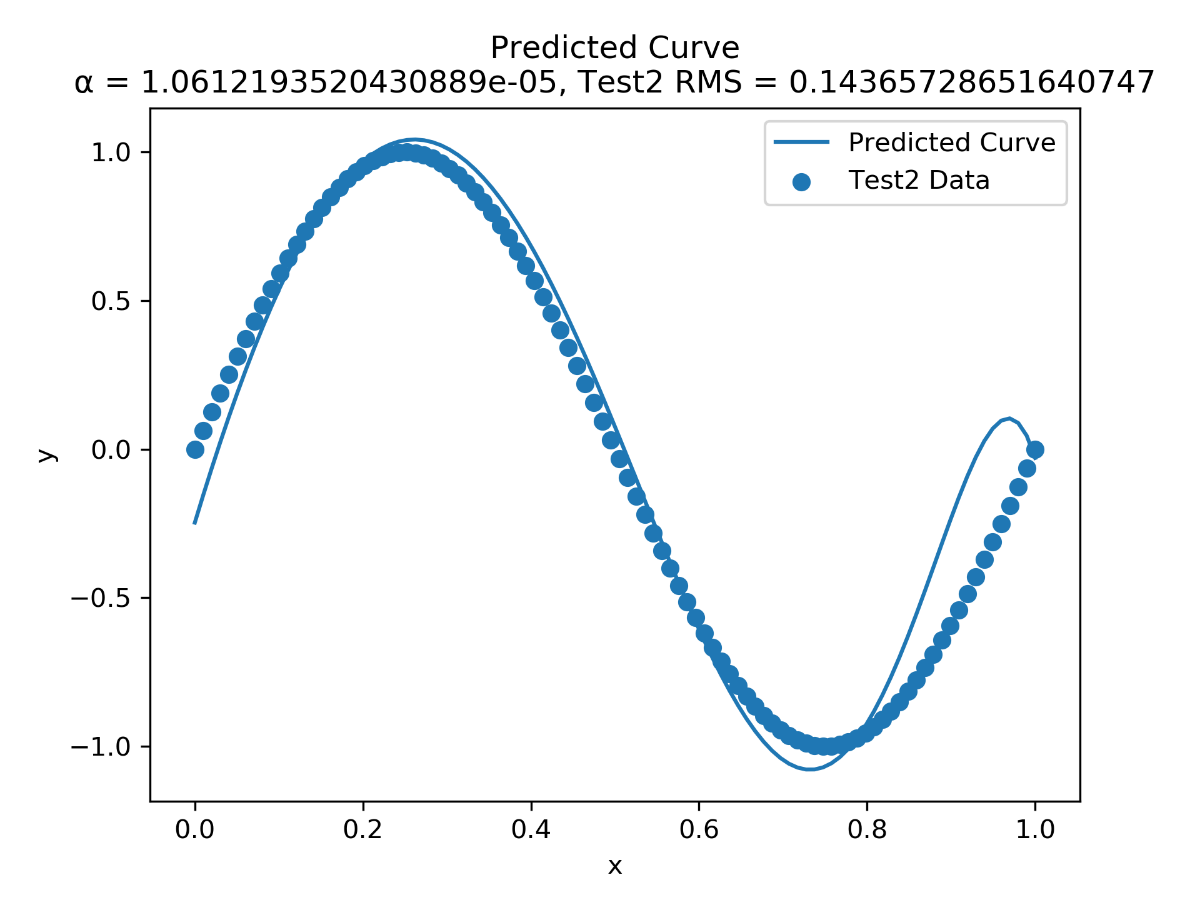
**图27** 多项式次数为7时测试结果2



**图28** 多项式次数为8时测试结果2



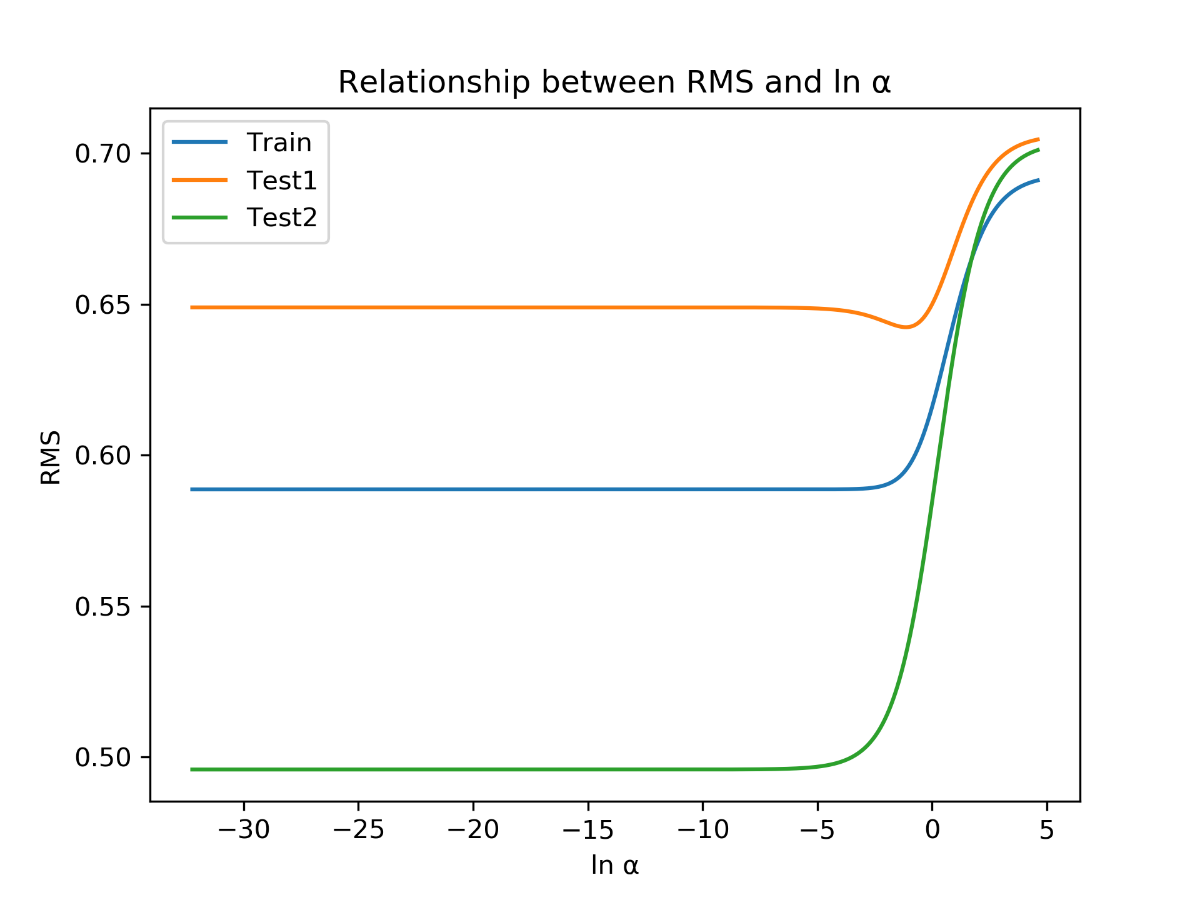
**图29** 多项式次数为9时测试结果2



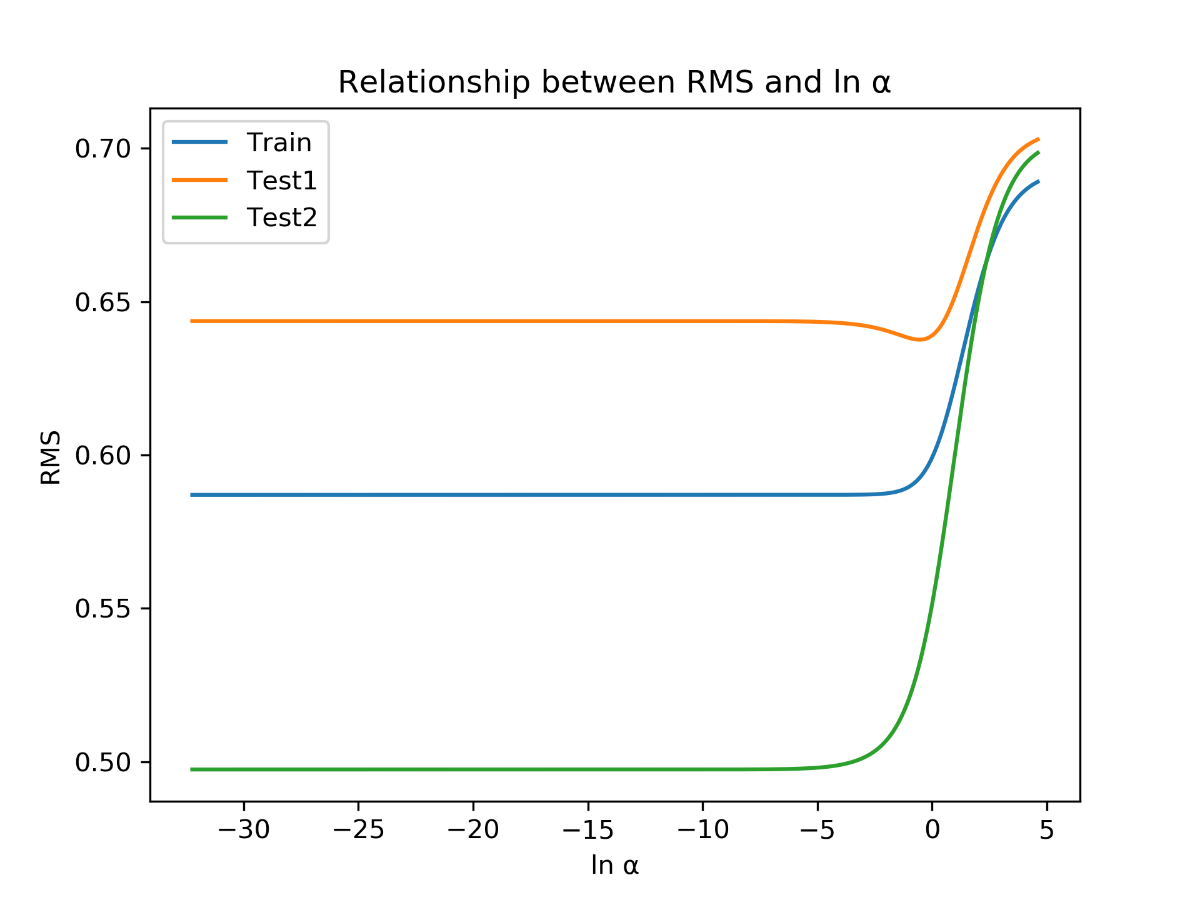
**图30** 多项式次数为10时测试结果2

**3.2 和的关系图**

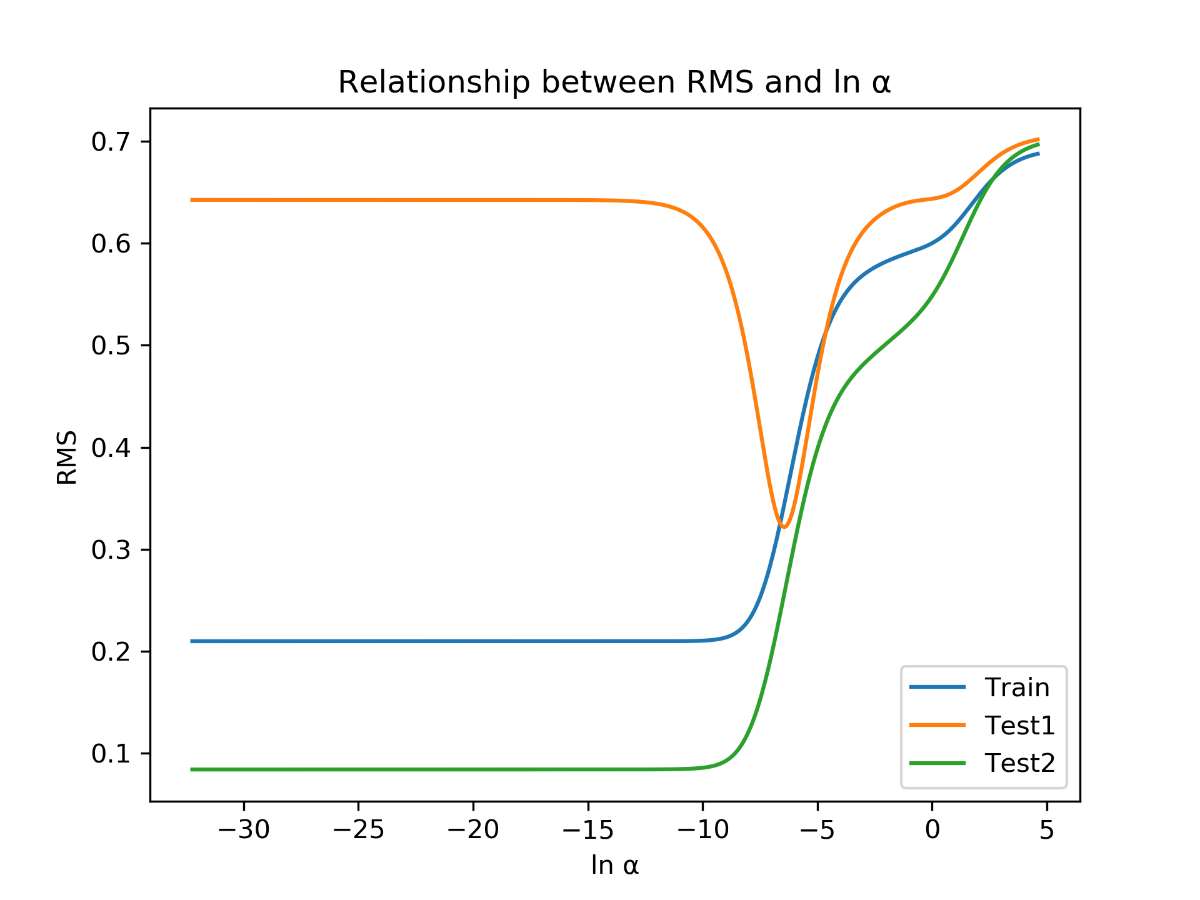
按2.2.3进行实验，其中多项式次数从1取到10，值取值集合为范围从到，项数为5000个的等比数列，由此得和的关系图，如图31-图40所示：



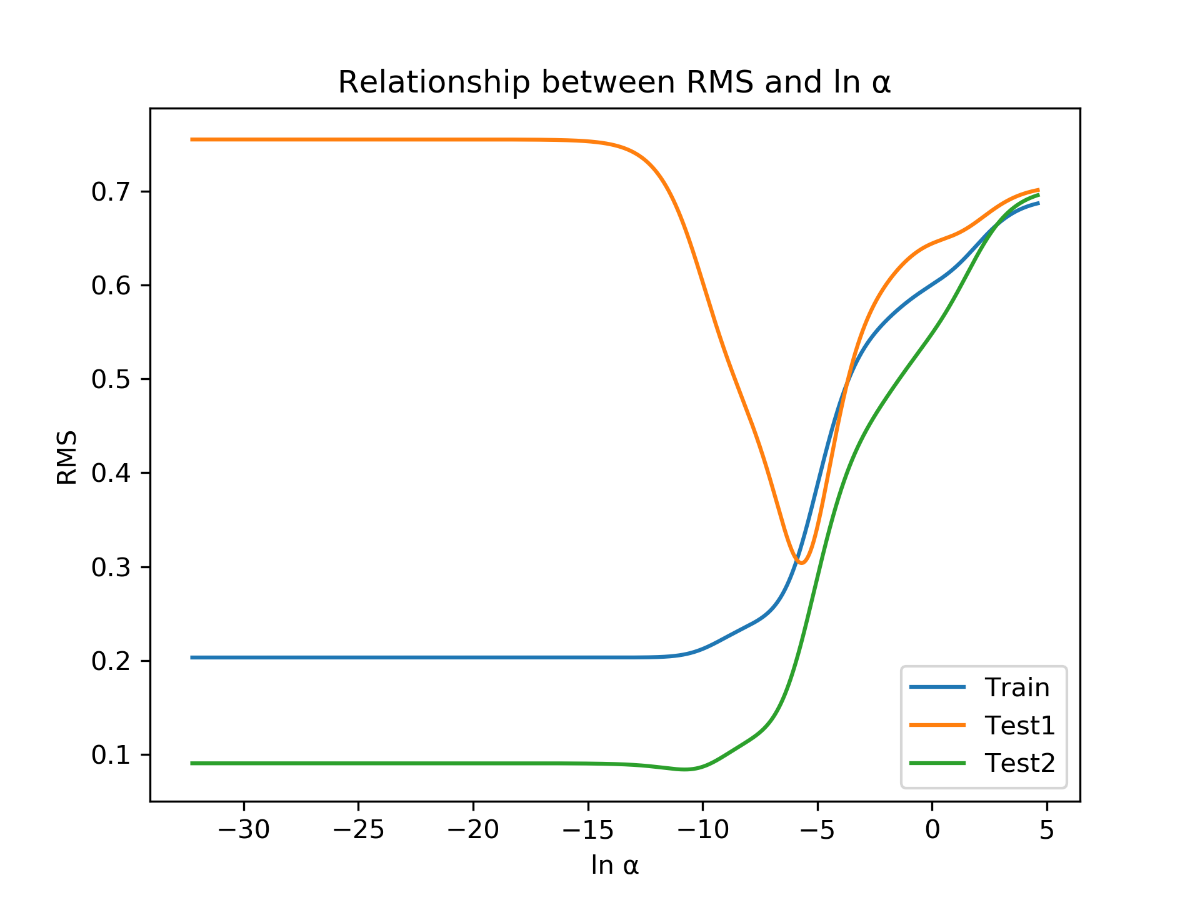
**图31** 多项式次数为1时和的关系图



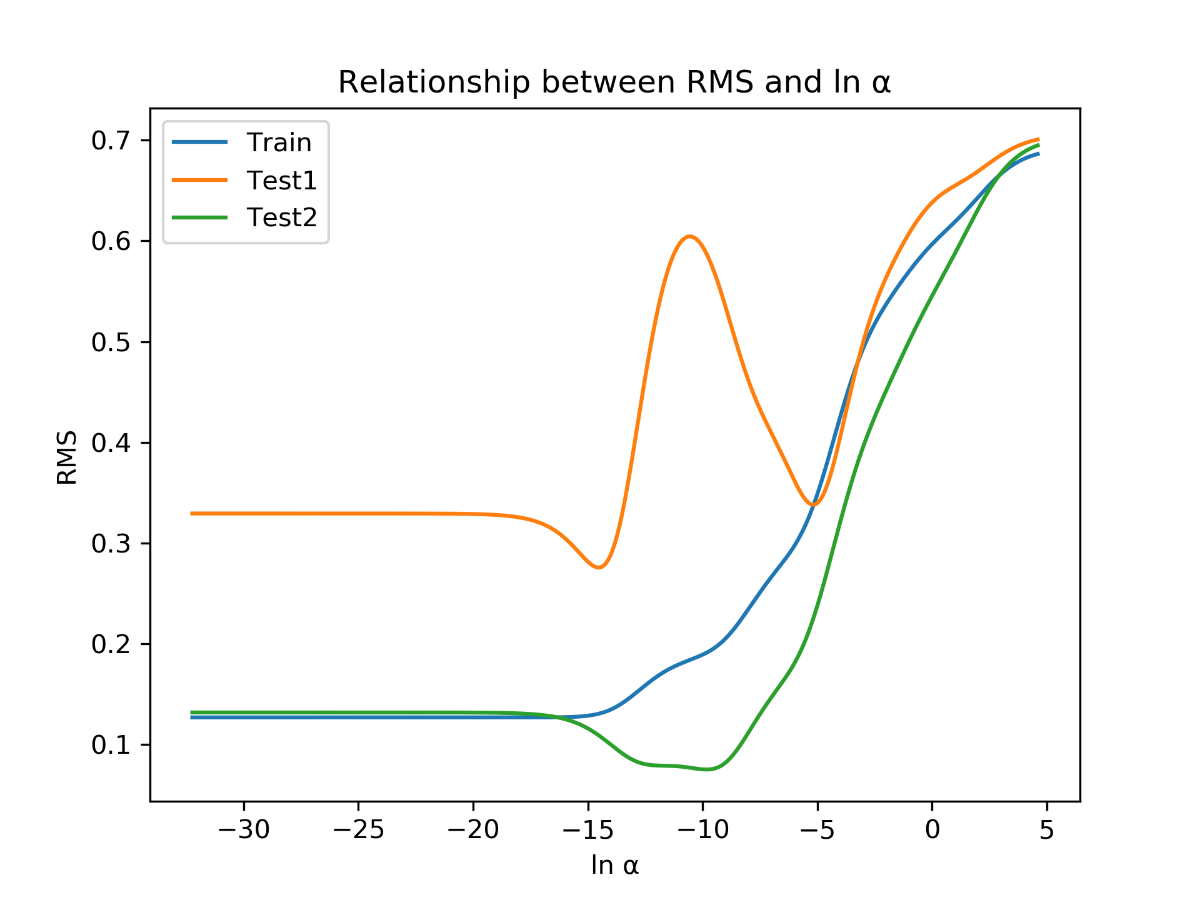
**图32** 多项式次数为2时和的关系图



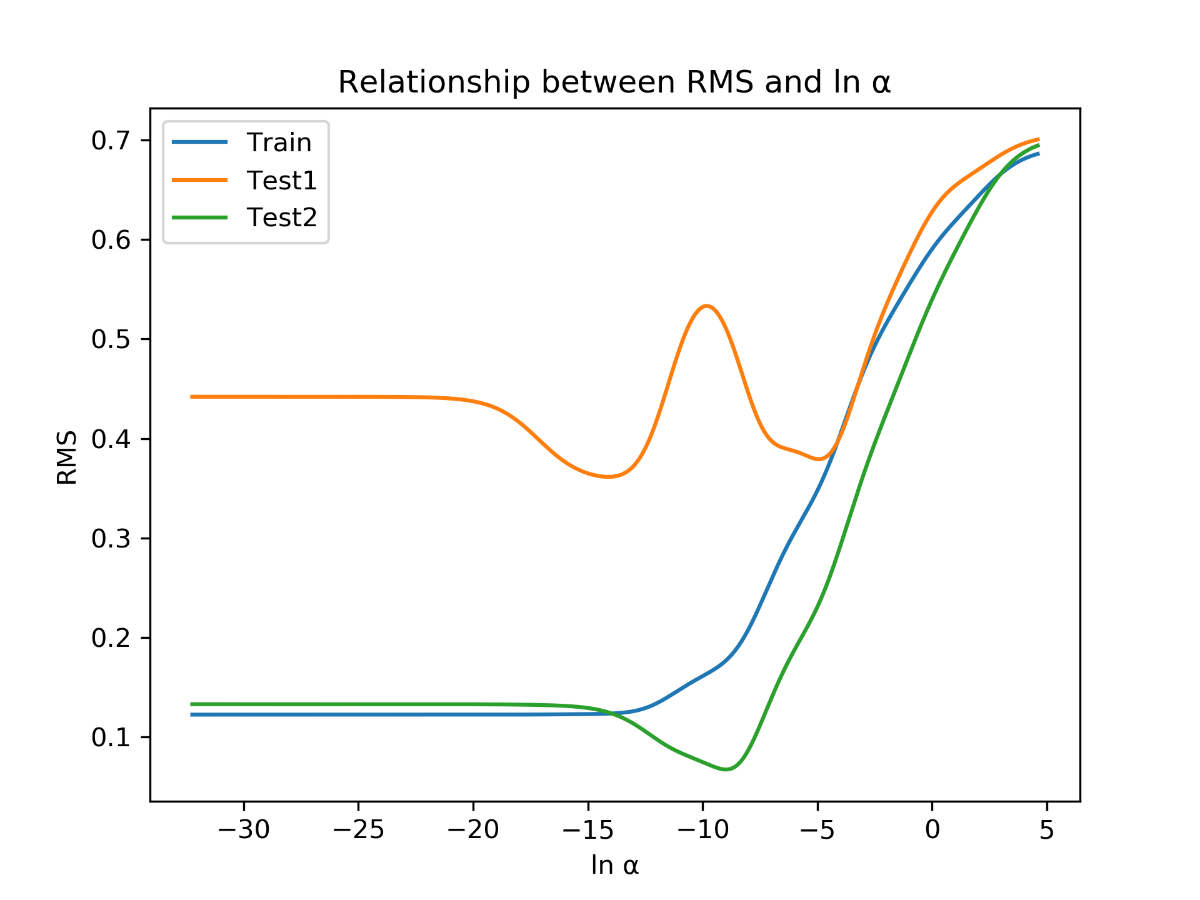
**图33** 多项式次数为3时和的关系图



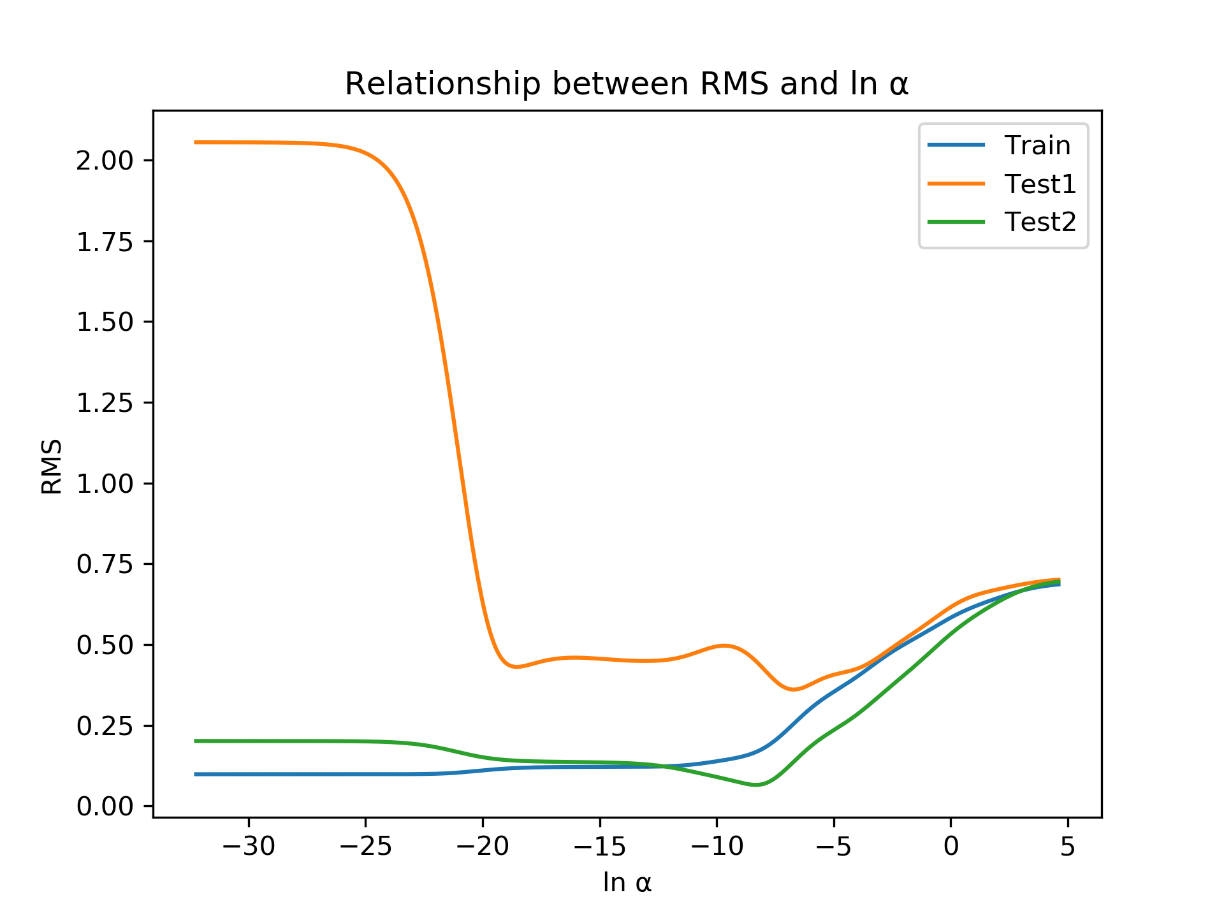
**图34** 多项式次数为4时和的关系图



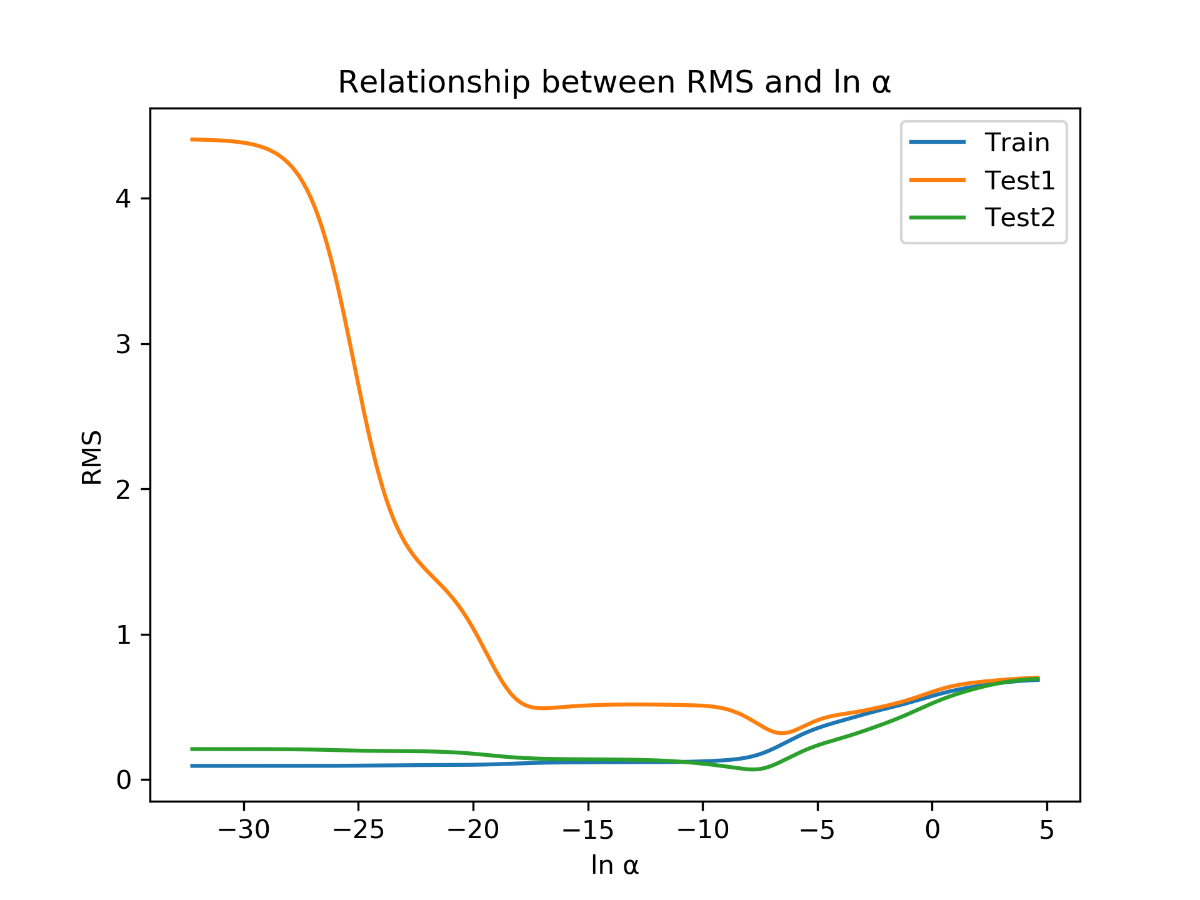
**图35** 多项式次数为5时和的关系图



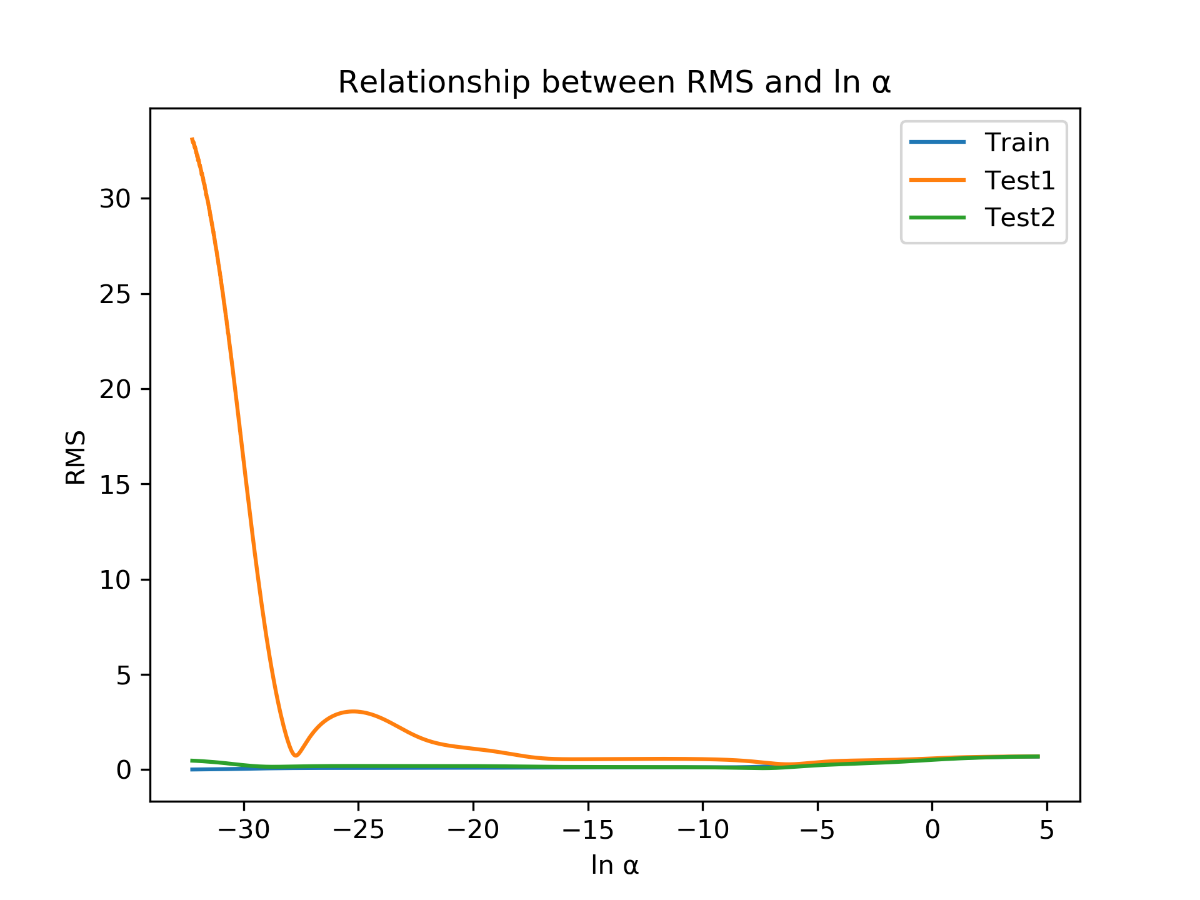
**图36** 多项式次数为6时和的关系图



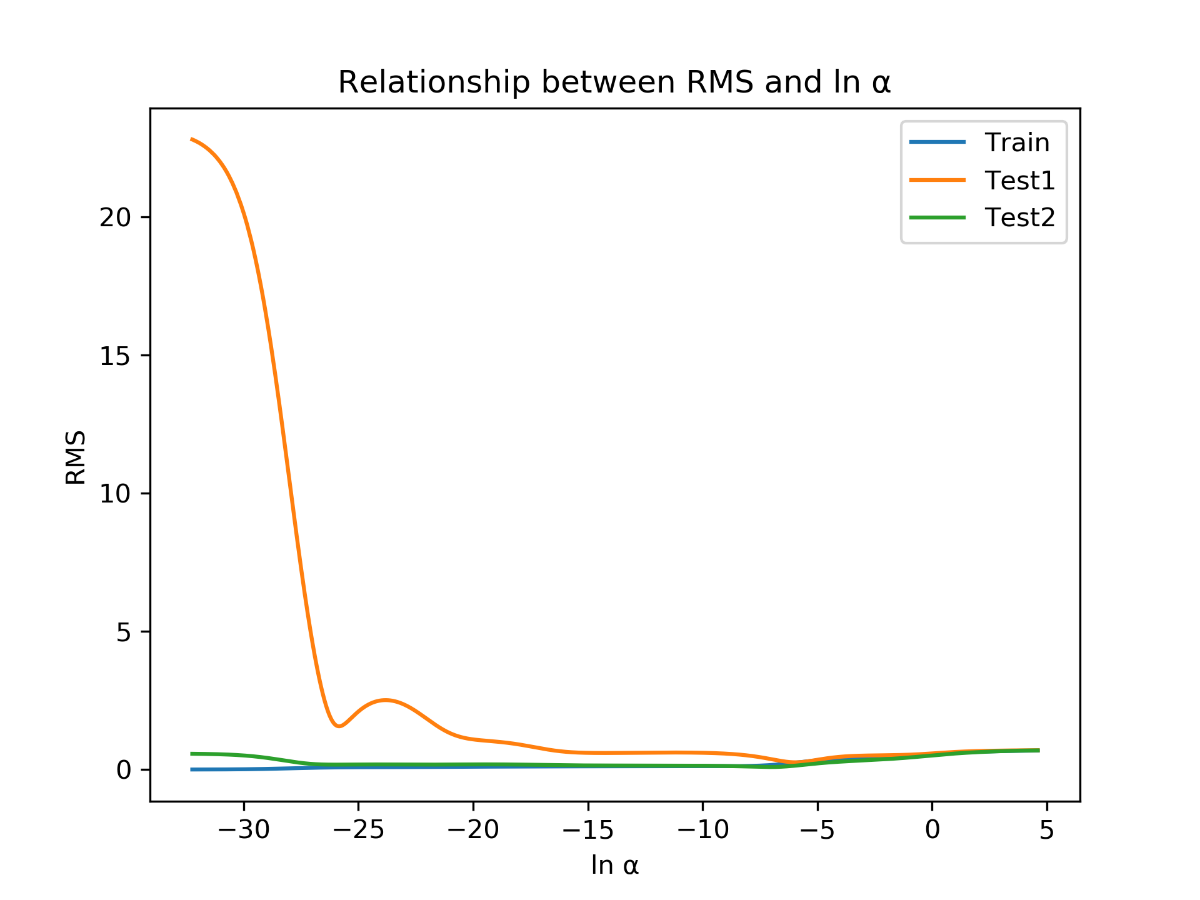
**图37** 多项式次数为7时和的关系图



**图38** 多项式次数为8时和的关系图



**图39** 多项式次数为9时和的关系图



**图40** 多项式次数为10时和的关系图

**4 结果与讨论**

**4.1 最佳训练结果及相应的测试结果与多项式次数的关系**

由3.1可知，在训练时，若模型的多项式次数少（例如只有一次项、二次项），那么即使采用岭回归，也难以与训练数据贴合，这就是欠拟合现象；而当模型的多项式次数足够多的时候（在本例中，多项式次数至少要不小于3），模型与训练数据的拟合程度较高，误差较小，可以进行进一步的测试。

而在测试时，对样本数少的测试数据1，我们发现，对于多项式次数少的模型（如线性函数模型或二次函数模型），其预测的数值与实际值偏差非常大，说明这些模型无法准确预测；而当模型的多项式次数足够多的时候，RMS先是从三次函数模型的0.5222降至五次函数模型的0.3244，随后又上升至十次函数模型的0.6160，且从六次函数模型开始，曲线的形状由侧“S”形变为“M”形，与训练数据体现出的可能的曲线特征不符，也与训练模型的侧“S”形不符，表明即使引入了二次惩罚项，过高的多项式次数仍然会导致模型的过拟合，从而在测试时无法得出准确的预测结果。

如果用样本数多的测试数据2进行测试，那么欠拟合和过拟合现象将表现得更加明显。显而易见，测试数据2大致分布在一条类似于正弦函数的曲线附近。对于次数较少的线性函数模型或二次函数模型，其预测结果是一条从中间穿过的平缓斜线，与测试数据2明显不相符。而对于三次函数模型，其预测结果是侧“S”形的曲线，与测试数据2反映的趋势基本吻合。继续增加多项式次数到四次、五次，则预测曲线发生了较为严重的偏离。再增加到六次及以上，虽然偏离有所减少，且对于峰区的拟合优于三次函数模型，但是对于谷区的拟合，六次及以上函数模型并没有比三次函数模型好，且或多或少出现了上凸的预测峰，与测试数据2反映的趋势不相符。

由此，我们可得出结论：（1）对于本次上机实验采用的训练数据和测试数据，三次函数模型的拟合效果最好；（2）当多项式次数小于三次时，由于模型的参数过少，会发生欠拟合现象，导致无法较好地描述训练数据及测试数据的趋势；（3）当多项式次数大于三次时，由于模型的参数过多，过于贴近训练数据，会发生过拟合现象，导致模型预测结果与训练数据出现极大偏差。

**4.2** **和及多项式次数的关系**

由3.2可知，对于不同的多项式次数，和的关系是不同的。就训练数据而言，越大，则模型与训练数据的拟合程度越小。而就样本量不大的测试数据1而言，若从误差曲线的形状出发，我们会发现，当多项式次数不大于4时，误差曲线呈“V”形，即随增大，先减小后增大，且存在一个（不一定与训练模型的最佳相同），使最小；当多项式次数为5~6时，误差曲线呈“W”形，即随增大，先减小至第一个谷值，随后上升至第一个峰值，之后再次减小至第二个峰值，最后上升至一个“平台”；当多项式次数为7~8时，误差曲线与logistic曲线的水平镜像相似，即随增大，先是迅速下降，随后在一个稳定值附近波动，或是缓慢上升；当多项式次数为9~10时，误差曲线的形状较为奇怪，随增大，先是迅速下降，然后略微上升至一个较小的峰值，最后下降至一个“平台”。

若从的波动范围这一角度看，我们有新的发现：当多项式次数为1~2时，的取值范围大致在0.64~0.70左右，说明二次惩罚项对模型的校正能力不强；当多项式次数为3~6时，上界大致在0.64~0.75，而下界均大致在0.30左右，说明通过调节二次惩罚项的系数，可使模型既能较好贴合训练数据，又能具有一定的泛化能力，即对于不同数据均能给出误差较小的预测结果；当多项式次数为7~10时，若，则上界将高得非常离谱，分别在2.10、4.50、34.00、23.00左右，说明在高次多项式的拟合中，二次惩罚项的作用已经不能忽视，若不加上二次惩罚项，所得的模型将出现严重的过拟合现象，而若，则的波动范围迅速与3~6次多项式模型一致，在0.30~0.75左右，这说明当二次惩罚项的系数足够大时，模型的泛化能力得到了迅速提升，过拟合现象得到了缓解。

就样本量大的测试数据2而言，若从误差曲线的形状出发，我们会发现，当多项式次数为1~2时，误差曲线的形状与logistic曲线相似，即随增大，先是大致保持不变，随后迅速上升，最后保持在一个稳定值；当多项式次数为3~4时，误差曲线的形状与logistic曲线略有不同，随增大，先是大致保持不变，随后迅速上升，但上升一段之后速度突然下降，最后在稳定值附近波动（也有可能是平缓上升）；当多项式次数为5~8时，误差曲线的形状大致呈左低右高的“V”形，即随增大，先是大致保持不变，随后下降至一个小谷，然后迅速上升，且存在一个（不一定与训练模型的最佳相同），使最小；当多项式次数为9~10时，由于测试数据1的过大，无法看出误差曲线的具体形状，只能大致推定为“V”形。

若从的波动范围这一角度看，我们则发现，当多项式次数为1~2时，的取值范围大致在0.50~0.70左右，说明在样本量大的条件下，二次惩罚项对模型的校正能力仍然不强；当多项式次数为3~4时，若，则仅略小于0.10，而若，则将逐渐升高至0.70左右，说明在样本量大的条件下，3~4次多项式模型不需要使用过大的即可获得较好的预测结果，使用过大的反而适得其反，使模型的预测准确度下降；当多项式次数为5~8时，若，则只有使用较大的才能使下降至0.10以下，否则明显比3~4次多项式模型的高（一般为0.13~0.25左右），而若，则同样升高至0.70左右，说明在样本量大的条件下，5~8次多项式模型必须要引入较大的，否则模型容易出现过拟合现象，但也不宜过高，以免导致模型对训练数据的贴合度不高，造成预测误差；当多项式次数为9~10时，无法从图象得出准确的最优值，但可以看出必定存在一个使最小。

由此，我们可得出结论：（1）对于欠拟合的情形（如多项式拟合中多项式次数不高），调节的大小并不能改善预测结果，相反，当逐渐增大时，模型的预测结果与实际值的偏差反而增大；（2）对于拟合结果较好的情形，并不需要太大，微调即可获得最佳结果，一味增加值同样会造成模型与训练数据的偏离，导致模型的预测能力下降；（3）对于过拟合的情形（如多项式拟合中多项式次数较高），的大小对拟合结果起到至关重要的作用。过小，二次惩罚项对拟合结果的改善有限，模型中仍然出现严重的过拟合现象；过大，二次惩罚项对损失函数的影响太大，容易掩盖训练数据蕴含的信息，导致模型与训练数据的贴合程度不高。

**4.3 实验结论**

（1）对于本次上机实验的数据，选三次多项式模型能得到最好的训练和预测结果；

（2）值在欠拟合和较好拟合的情形中效果并不明显，相反容易导致模型与训练数据的偏离，使模型的预测能力下降；而在过拟合的情形中效果显著，可有效改善模型的过拟合现象，但要注意值不宜过高，否则容易导致模型与训练数据的贴合程度不高。