Curve fitting: perspective from machine learning

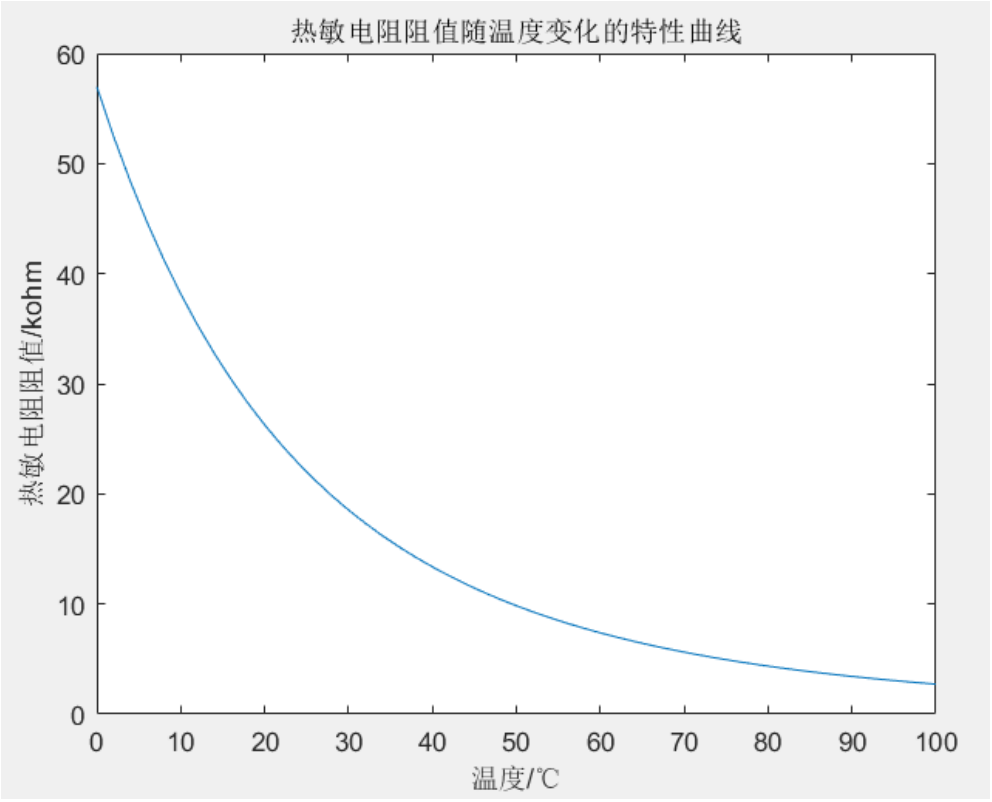
描述热敏电阻阻值与温度关系的模型可以表示为：

$$R_T = R_{T_0} e^{\beta\left(\frac{1}{T}-\frac{1}{T_0}\right)} \quad (1)$$

其中，T 为温度（单位为K），R_T 为温度为T 时热敏电阻阻值（单位为 kΩ），R_{T_0} 为温度为T_0时热敏电阻阻值（单位为 kΩ）。已知某种热敏电阻在 25°C 时的阻值为 22kΩ， $\beta = 3100$（K），试完成如下研究工作：

1） 以 2°C 作为间隔（步长），画出该种热敏电阻在温度范围为 0°C~100°C 间阻值随温度变化的特性曲线；

根据课件，T_0一般为$(25 + 273.15)K$，即对应的温度为 25°C，也就是说$R_{T_0} = 22\text{k}\Omega$，由公式（1）得到特性曲线如下：



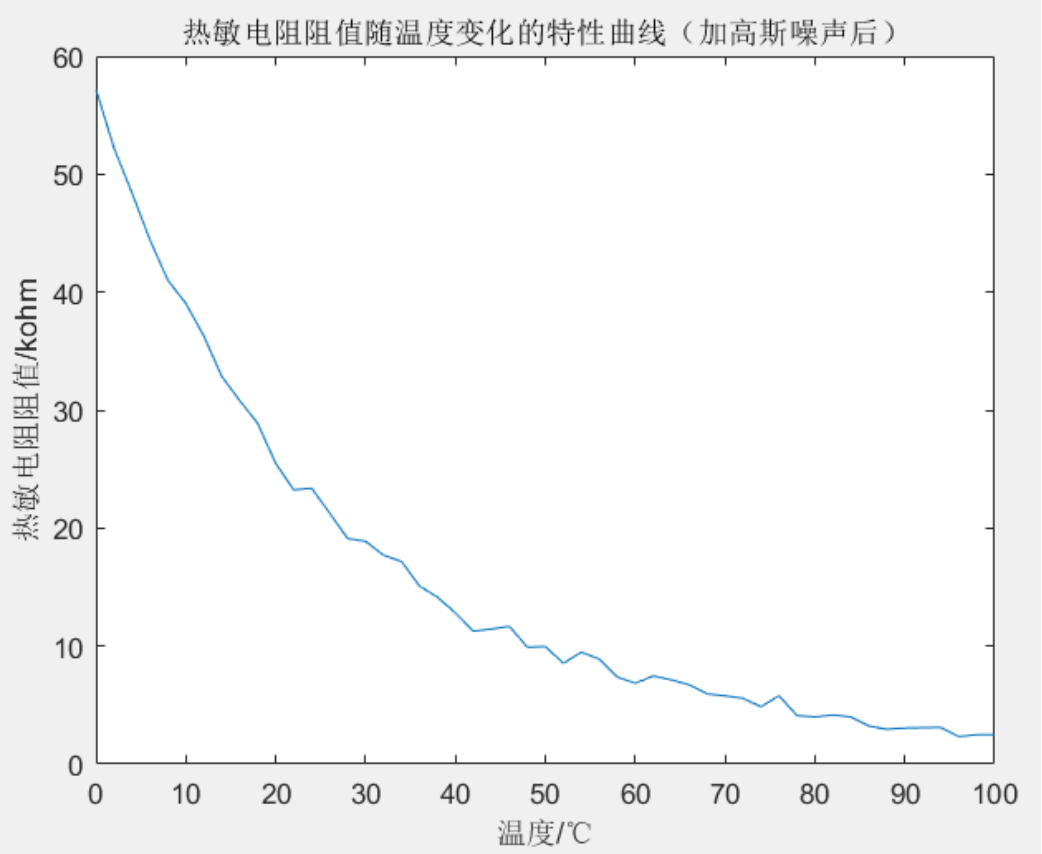
2）假设我们事先并不知道（1）式所描述的热敏电阻阻值—温度模型，现通过测量热敏电阻在不同温度下的阻值的实验方法对其特性加以研究，实验温度范围为 20°C~80°C。现采用如下多项式模型描述热敏电阻阻值与温度关系

$$R_t = a_n t^n + a_{n-1} t^{n-1} + \cdots + a_1 t + a_0 \quad (2)$$

其中，t 为温度（单位为℃），R_t 为温度为 t ℃ 时热敏电阻阻值（单位为 kΩ），n 为模型阶次，a_n 为不同阶次项系数。

在 1）中获得的 20°C~80°C 范围的数据上添加适当噪声（以零均值、标准偏差取 500Ω的高斯噪声为例），用添加噪声后的数据模拟实验数据（添加噪声模拟实际测量过程）。针对（2）式描述的多项式模型，用模拟的实验数据作为训练数据集，采用曲线拟合最小二乘法**分别获得模型阶次 $n=1,2,3,4,5,6$ 时传感器特性曲线对应的多项式模型；分别计算不同阶次模型在温度范围 20°C~80°C（训练集）上和温度范围 0°C~100°C 刨除 20°C~80°C 温度范围后（测试集）上的误差（均方误差, mean squared error）**，观察训练集和测试集上误差随模型阶次的变化规律并加以讨论；（注：可能用到的 matlab 函数：polyfit；randn）

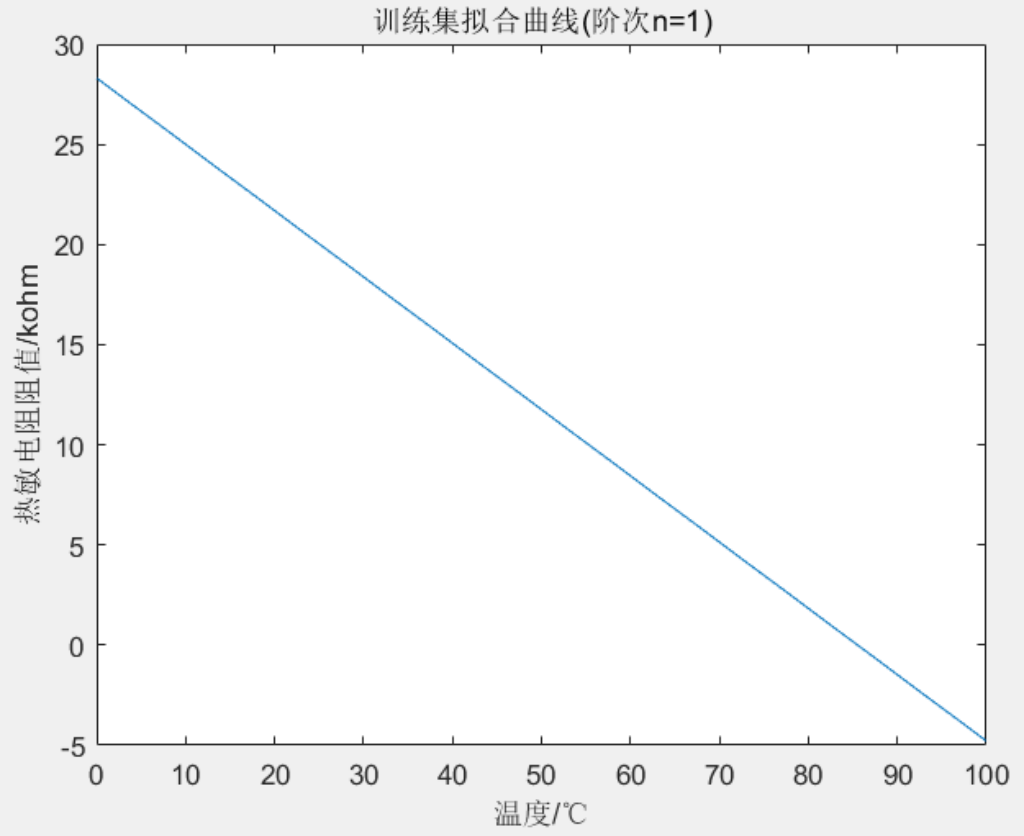
首先得到添加了零均值、标准偏差 500Ω的高斯噪声后的特性曲线如下：



然后利用 polyfit 进行曲线拟合，得到各阶次对应的曲线方程和图像如下：

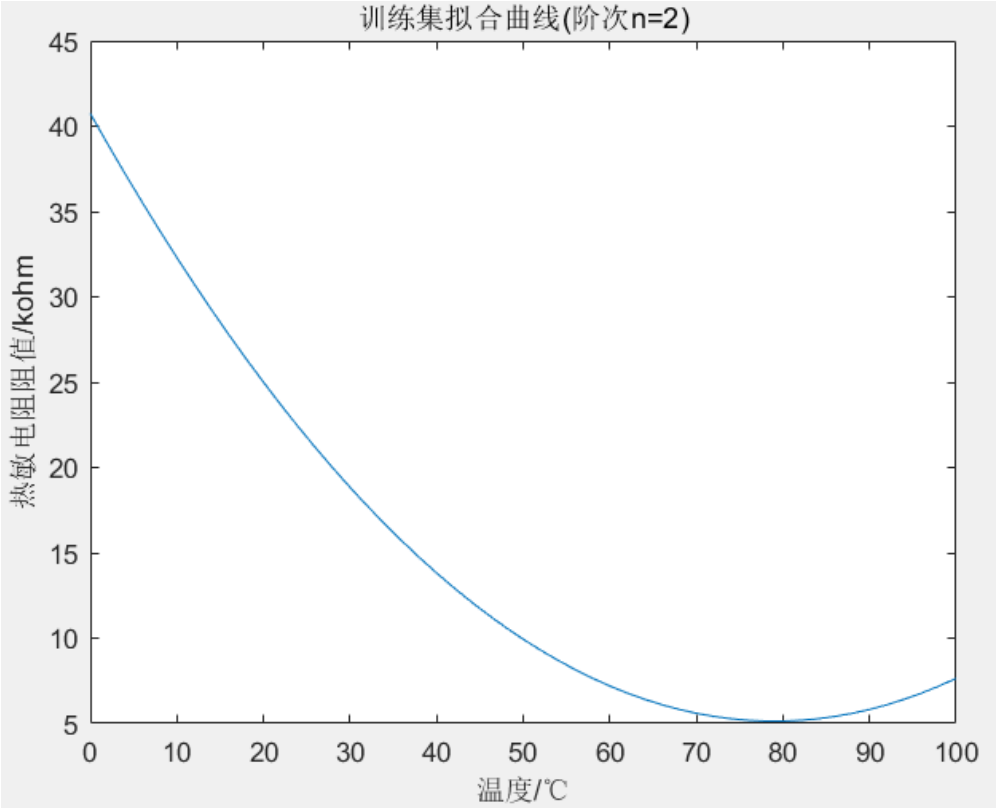
① $n=1$

$$R_t = -0.331t + 28.3084$$



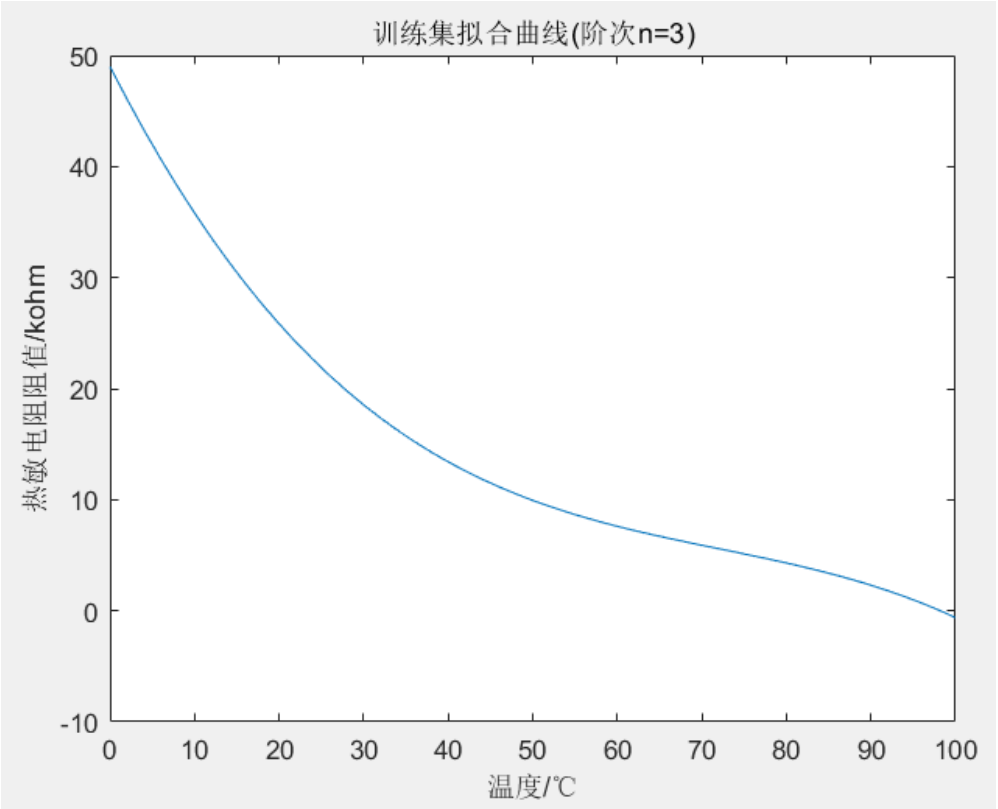
② $n = 2$

$$R_t = 0.0057t^2 - 0.9012t + 40.7394$$



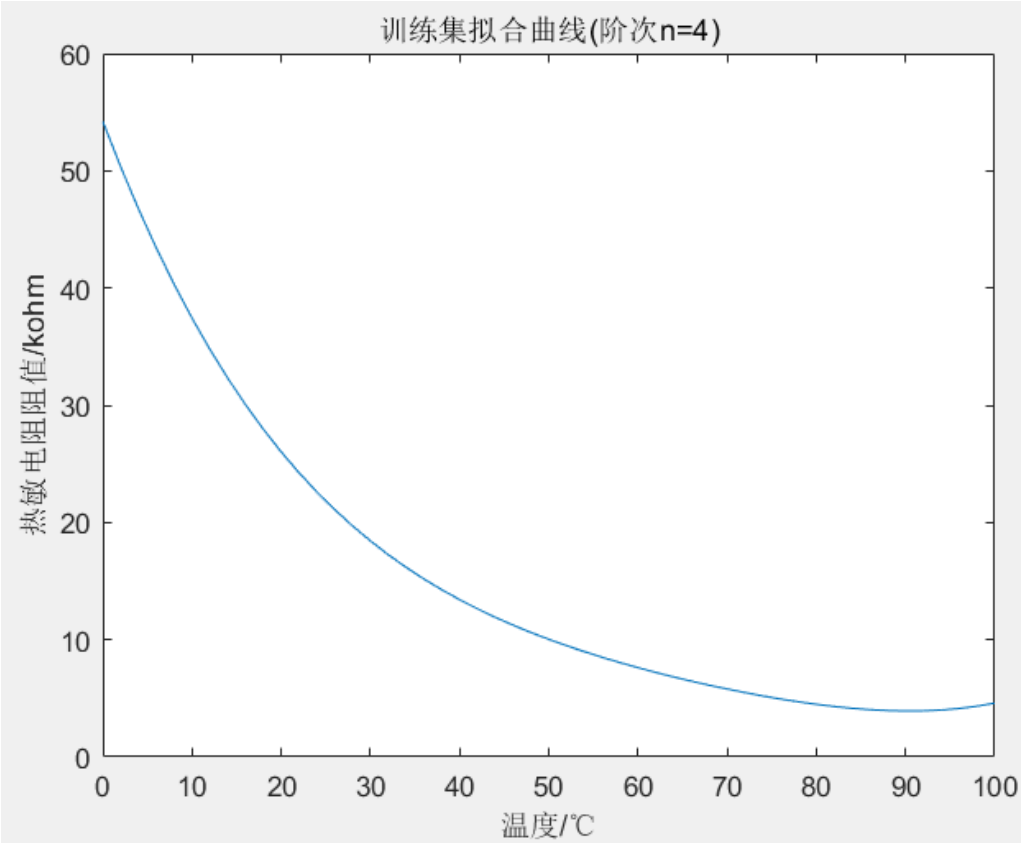
③ $n = 3$

$$R_t = -0.000085844t^3 + 0.0186t^2 - 1.4956t + 49.001$$



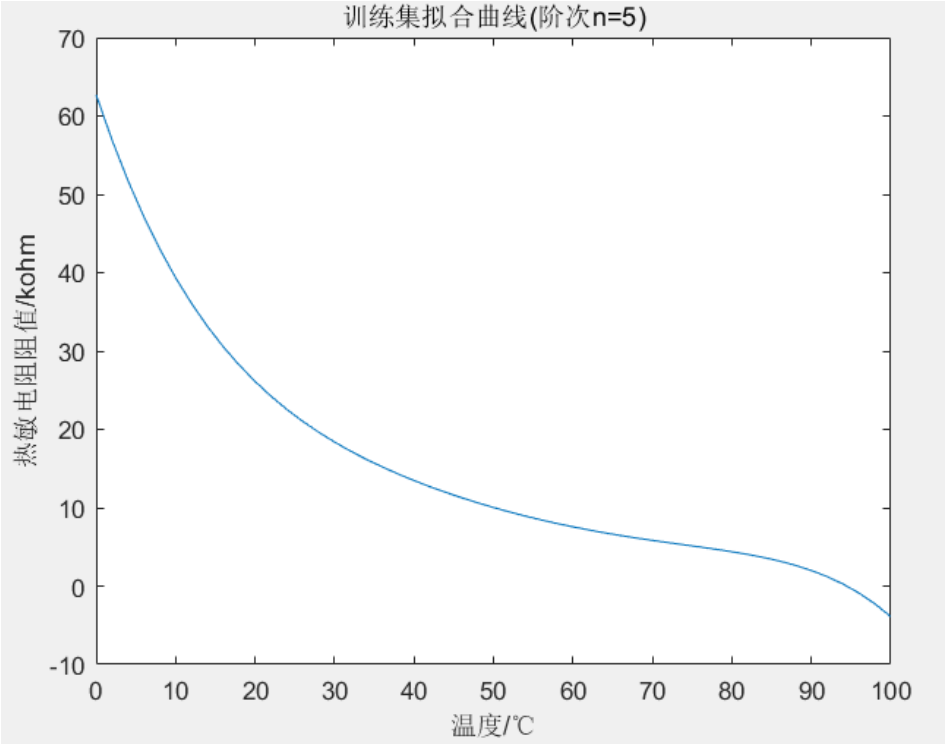
④ $n = 4$

$$R_t = -0.000001218t^4 + 0.0003294t^3 + 0.0358t^2 - 2.0048t + 54.2118$$



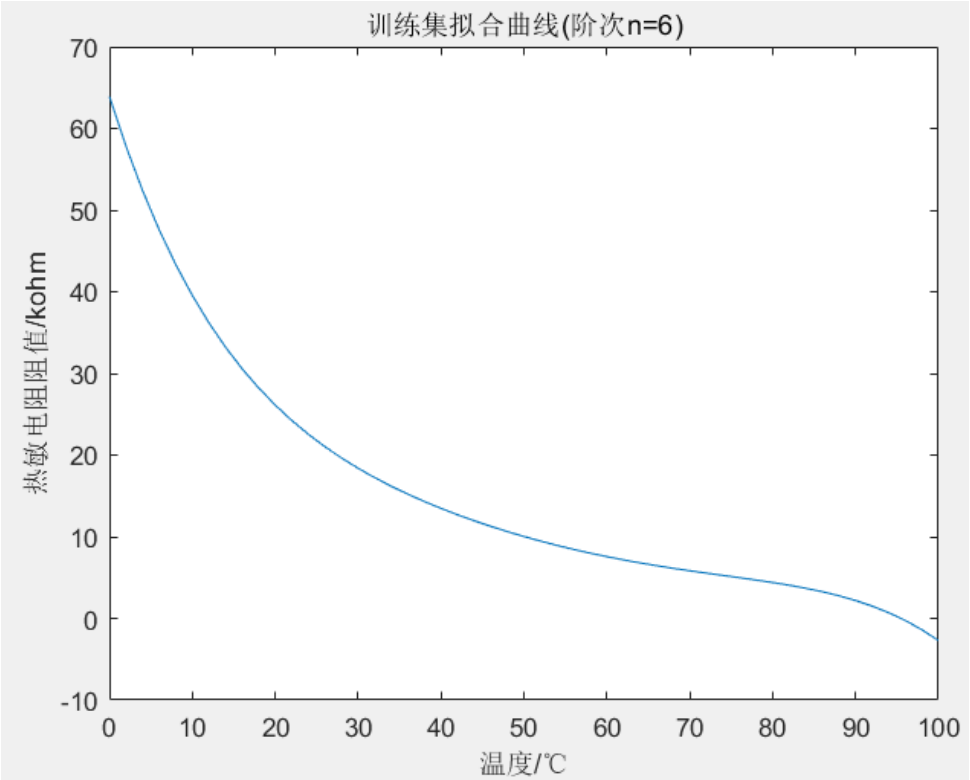
⑤ $n = 5$

$$R_t = -4.4545 \times 10^{-8}t^5 + 1.2354 \times 10^{-5}t^4 - 0.0014t^3 + 0.0844t^2 - 3.0523t + 62.7119$$



⑥ $n = 6$

$$R_t = -1.3967 \times 10^{-10}t^6 - 8.645 \times 10^{-8}t^5 + 1.7411 \times 10^{-5}t^4 - 0.0017t^3 + 0.0949t^2 - 3.2294t + 63.9023$$



对比各阶次拟合曲线的直观感受是$n=4$时的图像最接近原始数据对应的特性曲线，计算训练集和测试集上的均方误差有：

n	训练集	测试集
1	3.009158108288936	162.313622538625
2	0.353790210316391	45.6340535194501
3	0.206470845787103	45.6340535194501
4	0.199353390495343	1.87681105502791
5	0.199353390495343	8.40990687992179
6	0.197097994956285	8.47434166528563

首先观察训练集数据，可以看到阶次提高后均方误差明显变小了，理论上，多项式可以无限逼近一组确定的数据，但多项式阶次过高也会导致过拟合的问题，这一点在测试集的均方误差中有所体现，可以看到 5、6 阶多项式对应的误差比 4 阶时更大了。此外，选择不合适的低阶次模型时训练集对应的误差可能不是很显著，但测试集可能会明显地体现出误差，即该模型对整体数据的不适用性。综合图像和表格数据来看，$n=4$是最合适的模型。

3） 重复 2）相应内容 10 次（每次重新添加噪声模拟不同批次实验数据），观察并讨论由于采用不同训练数据给拟合（学习）结果带来的影响；

取 10 次拟合对应均方误差的平均值：

n	训练集	测试集
1	3.78207411586178	153.464645824503
2	1.19938963572177	38.6101208188555
3	1.02719248923754	11.9447220254672
4	1.01955107767820	9.49031025791962
5	0.999139468660806	176.269496445182
6	0.856473880849460	8082.04292691442

多次训练后的结果进一步验证了前面的讨论：提高多项式阶次可以保证训练集上的误差无限减小，但同时会造成过拟合，使得测试集上的误差增大，由此表同样可以得到$n=4$最合适的结论。

4） 改变噪声强度（通过改变所加噪声的标准偏差实现），重复 2），3）内容，观察并讨论数据中不同噪声强度给拟合（学习）带来的影响；

（1）、（2）、（3）中使用的高斯噪声标准偏差为 500Ω，现在添加两组对比，分别为 100Ω和 800Ω，100Ω标准偏差噪声对应的均方误差（10 次重复的平均值）如下：

n	训练集	测试集
1	3.08588447408278	159.836046154237
2	0.152993250086732	38.8542121045535
3	0.0140077331022293	7.81198960672528
4	0.00798109068571335	1.46133747326329
5	0.00737498451819573	3.58005795764210
6	0.00719791503759217	15.1324734656026

可以看到数据变化的趋势与前面的情况相同，但减小噪声强度后训练集上均方误差的值有比较明显的减小，此时$n = 4$的模型比较合适。

800Ω标准偏差噪声对应的均方误差（10 次重复的平均值）如下：

n	训练集	测试集
1	3.82076784674907	160.316383025518
2	0.687156906776367	41.0728951762039
3	0.524293770777689	10.4494096322854
4	0.502430248418757	18.5491223686049
5	0.497188938309326	51.0883502011773
6	0.484905427078906	728.498551107517

误差值仍然保持着相同的变化趋势，更高的噪声强度使得训练集的吻合程度下降了，此时$n = 3$的模型比较合适。

5） 将实验数据温度 20°C~80°C 范围进行调整（扩大或缩小），重复 2），3）内容（需要对训练集及测试集范围进行对应调整），观察并讨论由于采用不同规模训练数据给拟合（学习）结果带来的影响；

下面添加两组对比，训练集分别为 30℃~70℃（缩小了范围）和 10℃~90℃（扩大了范围）。缩小训练集范围后的均方误差如下表（10 次重复的平均值）：

n	训练集	测试集
1	1.48461212646382	140.554958183774
2	1.05680413457904	43.9794369685876

3	0.998318623781837	225.165003385260
4	0.937026905615812	2238.25279495129
5	0.915981607449039	12657.2953595752
6	0.875071607013973	381830.118008264

由于训练数据的减少，训练集反映整体情况的能力下降了，这直接导致了测试集上的误差值明显变大，甚至得到一个不符合实际情况的“最适模型”。

扩大训练集范围后的均方误差如下表（10 次重复的平均值）：

n	训练集	测试集
1	14.9452833137929	205.435110820471
2	2.27631991619052	42.5027958543569
3	1.29091322597615	5.00435436931189
4	1.25158607402685	3.24649986197769
5	1.20853359604292	8.87874604936487
6	1.20441936959361	13.1470897688420

扩大范围后训练集反映整体情况的能力提高了，测试集上的误差也会因此减小，同时因为训练内容更复杂，训练集自身的误差可能会提高，上表显示$n = 4$是合适的模型。

6） 思考：假如实验前已事先了解热敏电阻测温机理并掌握其阻值与温度的关系符合（1）式所描述的模型，你将如何考虑从实验数据获得热敏电阻的阻值与温度关系模型？

可以对公式（1）等式两边取对数得到

$$\ln\left(R_T\right) = \ln\left(R_{T_0}\right) + \beta\left(\frac{1}{T} - \frac{1}{T_0}\right)$$

再令$x = \frac{1}{T}$，$\mathrm{y} = \ln\left(R_T\right)$得到

$$y = \beta x + \ln\left(R_{T_0}\right) - \frac{\beta}{T_0}$$

其中$\ln\left(R_{T_0}\right) - \frac{\beta}{T_0}$是一个常数，这样就得到了一个线性方程，然后就可以利用实验数据首先求出这个线性模型，再倒推得到最终的模型。