智能传感与检测技术

Curve fitting: perspective from machine learning

2017010928 屈晨迪 自 71

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

热敏电阻在 0-100℃的阻值变化曲线如下图所示。

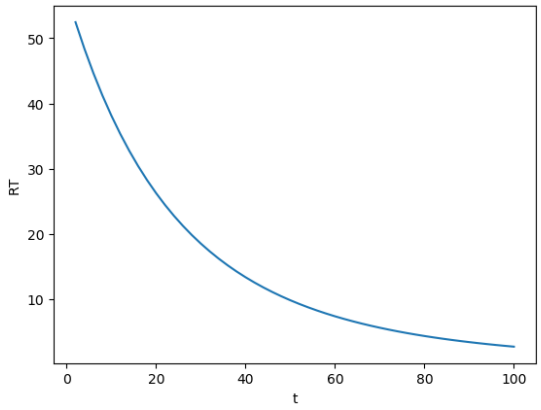


图 1 热敏电阻温度特性曲线

（2）在 1）中获得的 20℃~80℃ 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℃~80℃（训练集）上和温度范围 0℃~100℃ 刨除 20℃~80℃ 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；

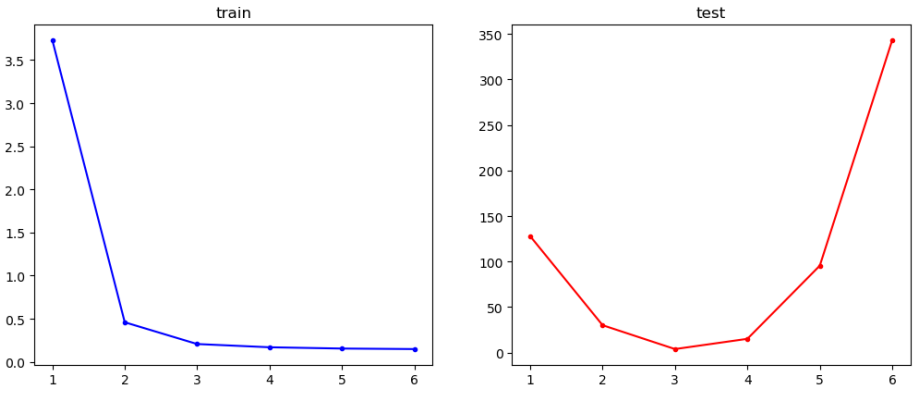


图 2 均方误差随模型阶次变化（左：训练集 右：测试集）

```
-0.3433 x + 28.88
          2
0.006323 x - 0.9756 x + 42.66
           3          2
-6.386e-05 x + 0.0159 x - 1.418 x + 48.81
           4           3          2
2.665e-06 x - 0.0005969 x + 0.0537 x - 2.532 x + 60.21
           5           4          3          2
-9.812e-08 x + 2.72e-05 x - 0.002946 x + 0.1607 x - 4.839 x + 78.93
           6           5          4          3          2
-5.951e-09 x + 1.687e-06 x - 0.0001883 x + 0.01039 x - 0.2839 x + 2.706 x + 28.17
```

图 3 预测 1-6 阶多项式模型

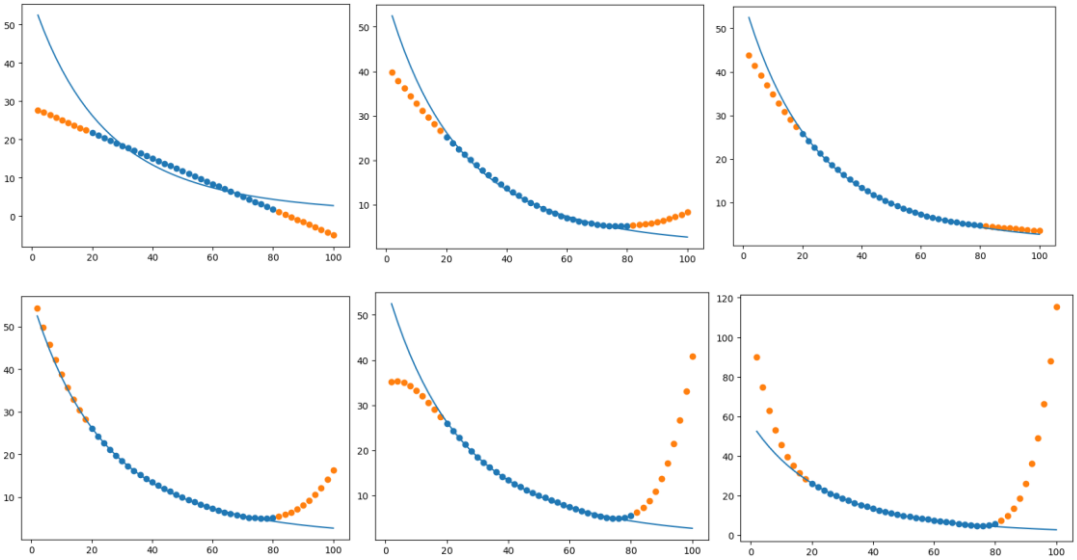


图 4 各阶次下预测值和真实值

从图 2 均方误差的变化曲线可以看到，随着模型阶次的升高，训练集的错误率逐渐降低，在阶次升到三之后下降放缓；而测试集上的误差先变低后升高，在次数等于三处达到最低（多次测试发现有时会在四阶次达到最低）。

预测误差包含了方差、偏差和噪声，其中噪声为定值。当阶次较低时，模型预测的方差较小，偏差较大，表现为欠拟合，在训练集和测试集上的误差均较大；而当阶次较高时，预测的偏差减小，方差变大，表现为过拟合，在训练集上误差很小，但测试集上误差大，模型泛化性能不佳。因此要找到复杂度合适的模型，其偏差和方差之和最小，在本例中多项式阶次为三时最佳。

（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；

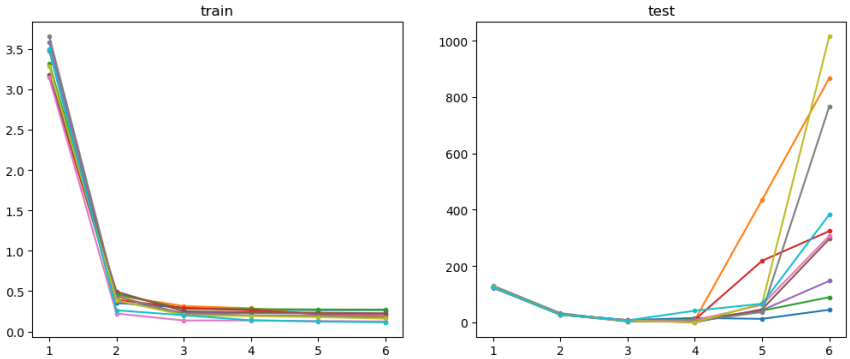


图 10 不同噪声拟合 10 次误差变化曲线

每次改变添加的噪声，重复 10 次，观察到均方误差随阶次变化的曲线基本不变，在训练集上均是逐渐下降，在阶次 3 之后基本不变，而测试集上的误差有时在阶次 3 最低，有时在阶次 4 处最低，随训练数据集的不同而不同。实际实验时可用 3、4 次都测试一下，对比选取最优的模型。同时观察到，在多次实验中，低阶模型参数比较稳定，但高阶如 6 阶多项式系数变化较大，这是因为每次添加的噪声不同，样本点不同，6 阶曲线尽可能地贴合每个

样本点，因此形状会出现较明显的变化。

（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；

在不同标准差下，测试集上均方误差随阶次变化如下表：

阶次 / 标准差	1	2	3	4	5	6
0.5	127.3	28.3	4.82	9.09	32.2	63.4
1	125.7	25.6	1.03	21.4	68.1	1828.1
2.5	129.8	39.3	40.4	43	1327	13175
5	119.5	18.6	30.7	90.9	2824.9	7071.9

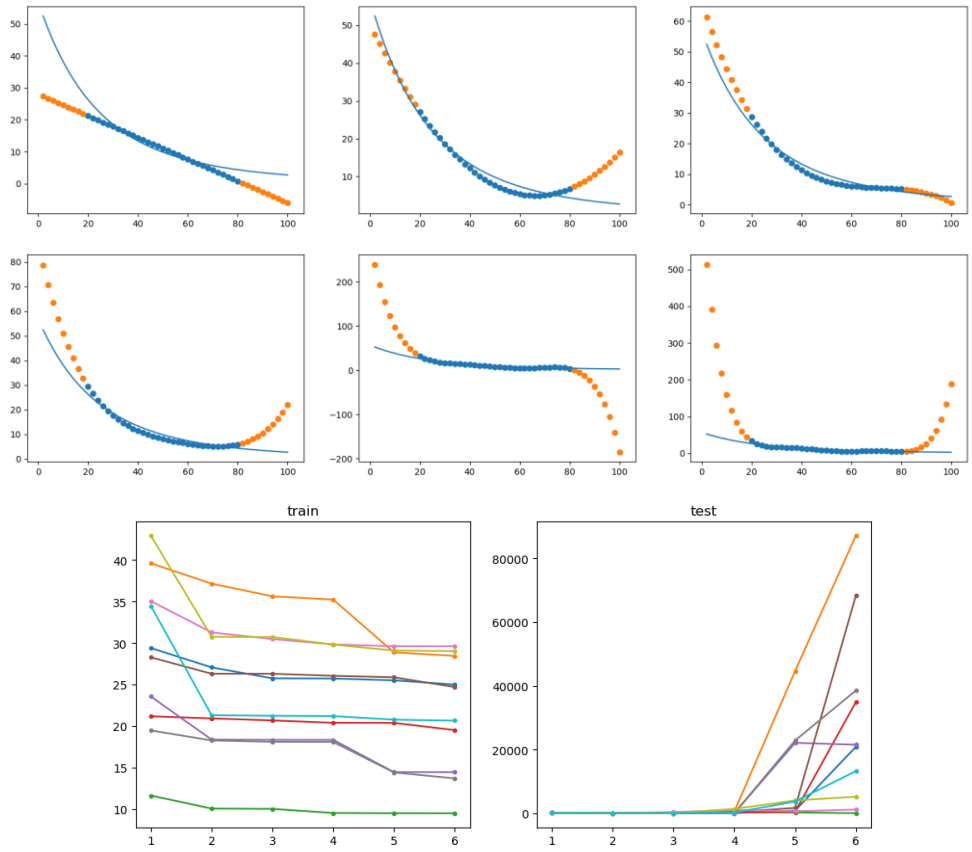


图 5 噪声标准差=5 时的误差曲线（10 次）

观察上表和图 5 曲线，可以看到在噪声标准差小于一定范围内，多项式模型能较好地对阻值进行预测，随着模型复杂度提高，均是在训练集上误差逐渐降低，测试集上先降低后升高，只是最优模型阶次可能有所不同；但当噪声标准差较大，如等于 5 时，训练数据中温度和阻值相关关系太弱，难以拟合出合适的模型，很容易出现过拟合的现象，即训练集上误差不断下降，但测试集上误差升高，尤其在 6 阶模型上误差急剧攀升。因此在实验过程中应尽量控制噪声强度，避免数据失去原有的相关性。

（5）将实验数据温度 20℃~80℃ 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；

• 温度范围缩小为 30-70℃

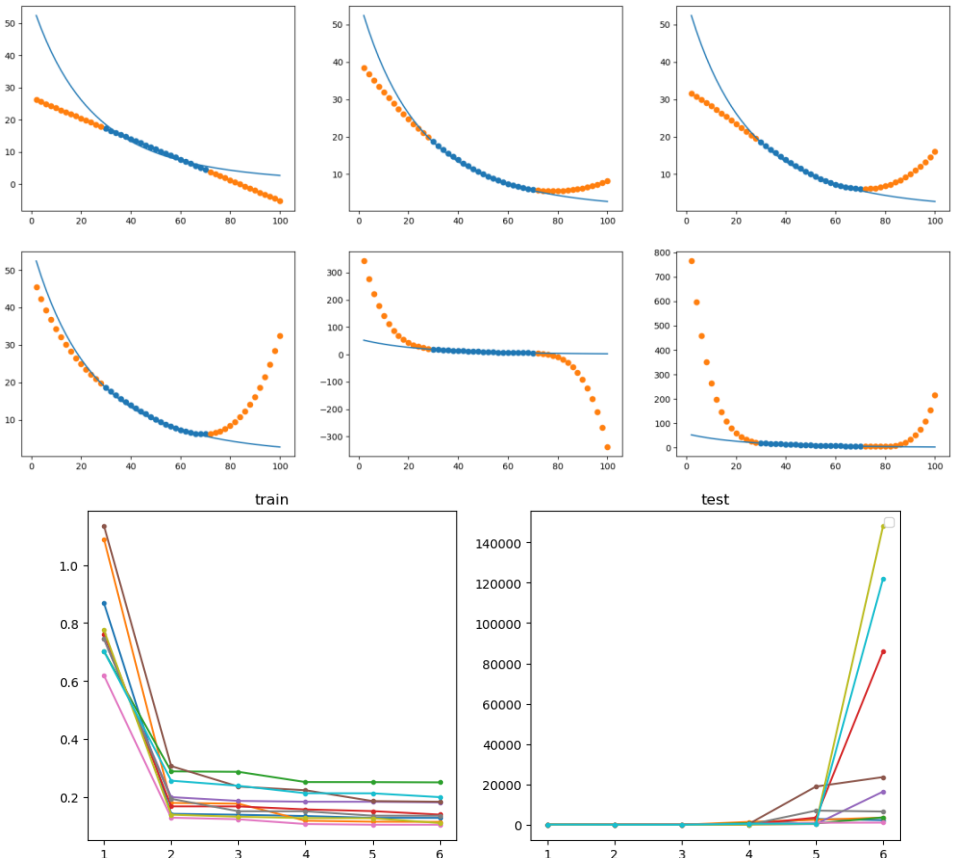
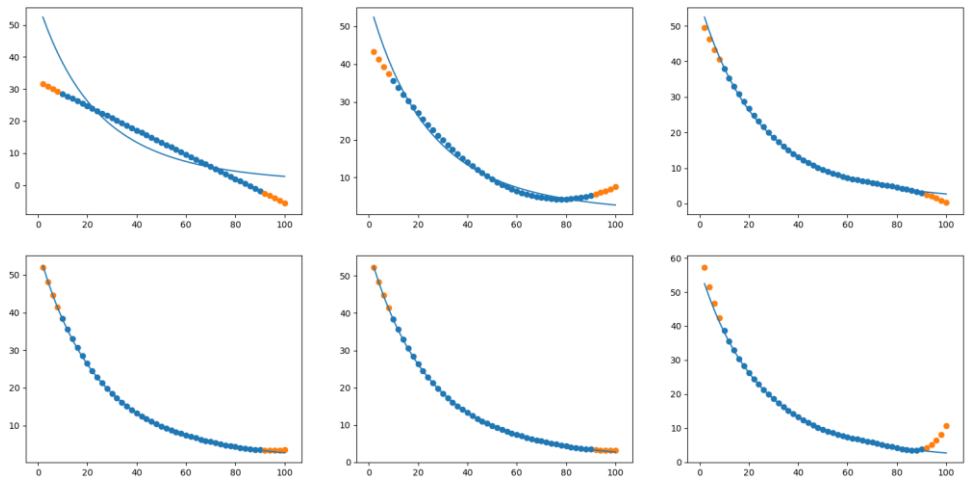


图 6 训练数据集缩小

当缩小训练数据集时，多次重复实验，结果大致如图 6 所示，可以看到随着模型复杂度升高，在训练集上误差逐步降低，但在测试集上，低阶模型有较好的表现，高阶模型误差很大，整体相对于 20-80℃训练集误差变大。当自变量温度范围缩小时，阻值变化范围也相应缩小，二者关系更加近似于线性关系，不能很好地展现整段温度范围内的变化趋势，因此预测出来的模型误差会偏大，而且数据集规模小也容易造成过拟合的现象，实验时应在较大范围内取值，并尽量扩大数据集规模。

• 温度范围扩大至 10-90℃



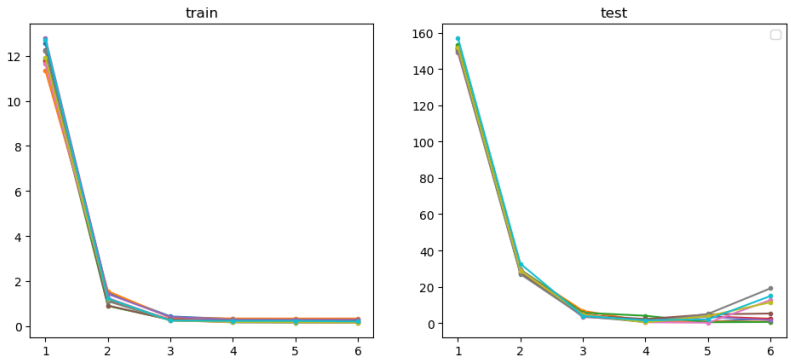


图 7 温度范围扩大

温度范围扩大，训练集变大，测试集缩小，与温度范围缩小相反，在测试集的表现上高阶模型较好，而低阶如 1 阶多项式有较大误差，但总体在训练集和测试集上的表现均有提升，这是因为训练集中包含了 80%的数据，基本已经可以反映出整体数据的变化趋势，因此模型在全范围的预测上误差较小。但如果实验中选取已经有过拟合趋势的 5 阶或 6 阶模型，在更大温度范围上的表现可能不佳，有可能效果不如 3 阶模型，应根据实验要求考虑权衡。

（6）选做：采用梯度下降算法，重复 2），3）内容，探讨模型参数初值、学习率对结果的影响。

• 学习率

采用梯度下降求解多元线性回归，拟合多项式模型。以一阶多项式为例，记录每次迭代的 cost，绘制曲线如图 8，控制其他参数不变，改变学习率，分别设为 0.01、0.001、0.0001，可以看到随着学习率的降低，达到损失函数最小值处的速度越慢，可见学习率代表了模型学习、更新参数的快慢，在迭代式中$\theta_{new} = \theta - \alpha\nabla J$，学习率$\alpha$越大，参数$\theta$变化更新的越快，在更复杂的模型训练中，可以动态改变学习率，在一开始较大，叫快速地更新参数，逐渐接近最优值时减小，使参数趋于稳定。有时学习率太大会无法收敛，一般凭借经验而定。

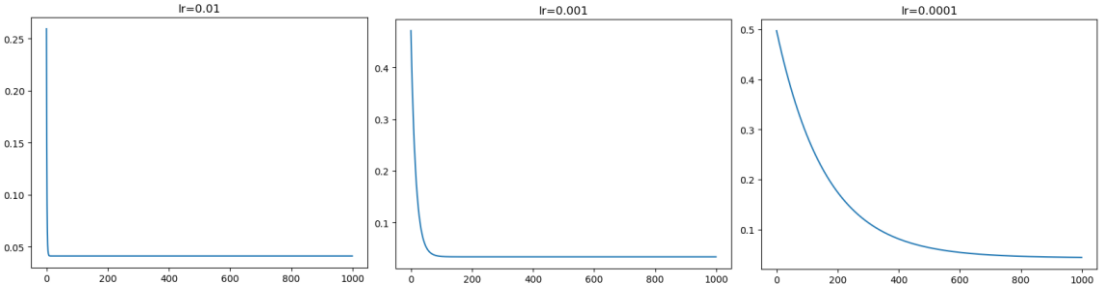


图 8 损失代价随迭代次数的变化曲线

• 初值

在本程序中，梯度下降的初值均设为 0，因为对输入样本点的 X 做了归一化，再进行多元线性回归，因此第一列常数项为全 0，如果初值设为其他值常数项会随之改变，导致结果有误。

一般而言，参数对初值的选取较为鲁棒，即不同初值取值均能得到相近的结果，在多次实验观察到，初值选择不同时，可能导致收敛速度不同，如果初值选取不当，可能出现不收敛的情况，如果损失函数不是凸函数，也有可能陷入局部最优点。实际实验中应多次尝试，取较好的初值。

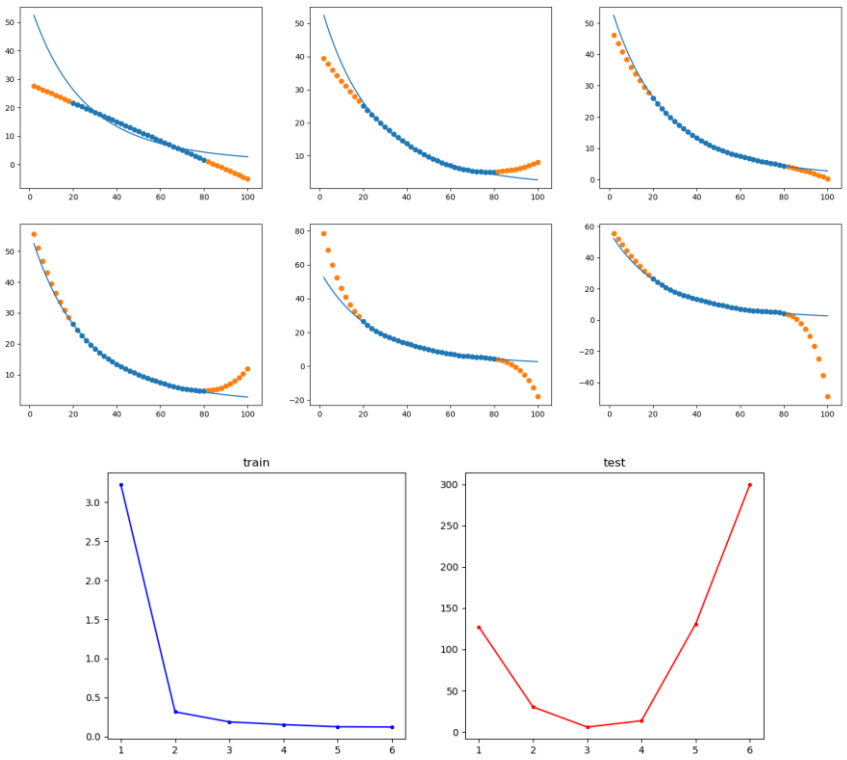


图 9 梯度下降法拟合结果

图 9 是用梯度下降法拟合出的模型结果，与最小二乘法基本一致，均方误差在训练集上随模型阶次增高而下降，在测试集上在 3-4 阶次左右最低。程序中梯度迭代函数用步长二范数小于$1e-4$作为停止判定，在实验中观察到，模型阶次越高时需要的迭代次数较多，如果迭代次数不够可能达不到最终稳态值。5、6 阶模型需要很长时间才能得到最终结果，运行时需要耐心等待，或增大步长停止条件，但得到的结果有一定偏差，一味提高学习率可能不收敛，后续可在加快迭代速度上做一些改进。

（7）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？

由题，如果已经知道热敏电阻阻值与温度的关系符合式子$R_T = R_{T_0} e^{\beta\left(\frac{1}{T}-\frac{1}{T_0}\right)}$，可以两边取对数，化为$\ln R_T = \ln R_{T0} \cdot \beta(\frac{1}{T} - \frac{1}{T_0})$，然后取一定范围温度的导数$1/T$为自变量，阻值的对数$\ln R_T$为因变量，两者符合$y = ax + b$的线性关系，此时用一阶多项式对二者进行拟合，可以得到相对应的参数。