**形状

中度可信度描述已自动生成**

**智能制造质量预测**

组长： 张庆伟 3210103414

组员： 葛涛 3210103407

组员： 刘祥盛 321010XXXX

2023年11月08日

目录

[**一、** **问题介绍** 3](#_Toc150461872)

[**1.1.** **问题背景** 3](#_Toc150461873)

[**1.2.** **问题描述** 4](#_Toc150461874)

[**1.3.** **评估指标** 4](#_Toc150461875)

[**二、** **数据处理** 5](#_Toc150461876)

[**三、** **线性回归** 9](#_Toc150461877)

[**3.1** **立项阶段** 9](#_Toc150461878)

[**3.2** **外部对接1** 9](#_Toc150461879)

[**3.3** **外部对接2** 9](#_Toc150461880)

[**3.4** **筹备阶段** 9](#_Toc150461881)

[**3.5** **活动阶段1** 9](#_Toc150461882)

[**3.6** **活动阶段2** 9](#_Toc150461883)

[**3.7** **总结阶段** 9](#_Toc150461884)

[**四、** **Lasso回归** 10](#_Toc150461885)

[**五、** **Ridge回归** 11](#_Toc150461886)

[**5.1 Ridge回归介绍** 11](#_Toc150461887)

[**5.2 代码分析** 11](#_Toc150461888)

[**5.3 运行结果** 12](#_Toc150461889)

[**5.4 分析总结** 13](#_Toc150461890)

[**六、** **小组分工** 15](#_Toc150461891)

1. **问题介绍**
   1. **问题背景**

半导体产业是一个信息化程度高的产业。高度的信息化给数据分析创造了可能性。基于数据的分析可以帮助半导体产业更好的利用生产信息，提高产品质量。

现有的解决方案是，生产机器生产完成后，对产品质量做非全面的抽测，进行产品质量检核。这往往会出现以下状况，一是不能即时的知道质量的好坏，当发现质量不佳的产品时，要修正通常都为时以晚，二是在没有办法全面抽测的状况下，存在很大漏检的风险。

在机器学习，人工智能快速发展的今天，我们希望着由机器生产参数去预测产品的质量，来达到生产结果即时性以及全面性。更进一步的，可基于预先知道的结果，去做对应的决策及应变，对客户负责，也对制造生产更加敏感。

**痛点与挑战：**

（1）TFT-LCD（薄膜晶体管液晶显示器）的生产过程较为复杂，包含几百道以上的工序。每道工序都有可能会对产品的品质产生影响，故算法模型需要考虑的过程变量较多。

（2）另外，这些变量的取值可能会存在异常（如测点仪表的波动导致、设备工况漂移等现象），模型需要足够稳定性和鲁棒性。

（3）产线每天加工的玻璃基板数以万计，模型需要在满足较高的精准度前提下尽可能实时得到预测结果，这样才能给在实际生产中进行使用。

**价值：**

（1）如果能够建立算法模型准确预测出特性值，便可以实现生产过程的实时监控和预警，提前发现当前工序的问题、避免问题流入到后道工序，减少生产资源浪费的同时也优化了产品良率。

（2）基于预测模型得到的关键参数，工艺人员能够快速地针对那些电性表现不佳的产品进行问题溯源分析，重点分析和调整那些关键的影响因子，加快不良问题的处理、提高整体工艺水平。

（3）该预测模型在部署后也可以用于减少特性检测相关的工序，能够节约检测资源并且对提升产线整体的产能有正面作用。

**问题/数据集来源：**

天池智能制造质量预测数据集：<https://tianchi.aliyun.com/dataset/140667>

具体处理见第二部分“数据处理”。

* 1. **问题描述**

本题目提供的数据集提供了生产线上的抽样数据，反应机台的温度，气体，液体流量，功率，制成时间等因子。通过这些因子，需要研究人员设计出模型，准确的预测与之相对应的特性数值。这是一个典型的回归预测问题。

* 1. **评估指标**

本任务采用常用的MSE指标作为评估指标，具体来讲：是我们建立回归模型预测的值，是真实的值。的值越小，代表预测结果和真实值越接近，效果越好。其中是：

1. **数据处理**

我们选用天池智能制造质量预测数据集， 本次大作业所用的所有数据集均在DataSet文件夹中，这里做一点说明：

训练集.xlsx：共5954列，801行；

测试集A.xlsx：共5953列，301行；

测试集B.xlsx：本次大作业中未使用；

测试集A\_答案.csv：官方提供的测试集A的答案；

train：经过数据处理后得到的训练集，用于后续的模型训练；

testa：经过数据处理后得到的测试集a，用于后续的模型训练；

训练集中共有800条样本数据，每条数据包含5954列字段。第一个字段为ID号码，最后一列为要预测的值Y。其余的5952列数据为用于预测Y的变量X。这些变量一共由多道工序组成，字段的名字可以区分不同的工序，例如 210X1, 210X2。300X1,300X2（即标题行）。字段中的TOOL\_ID或者Tool为每道工序使用的机台。

测试集A中共有300条样本数据，每条数据包含的字段数与训练集相同，但是缺少最后一列的Y值。

下面是对训练集数据处理的介绍（数据处理的所有源代码均在中）：

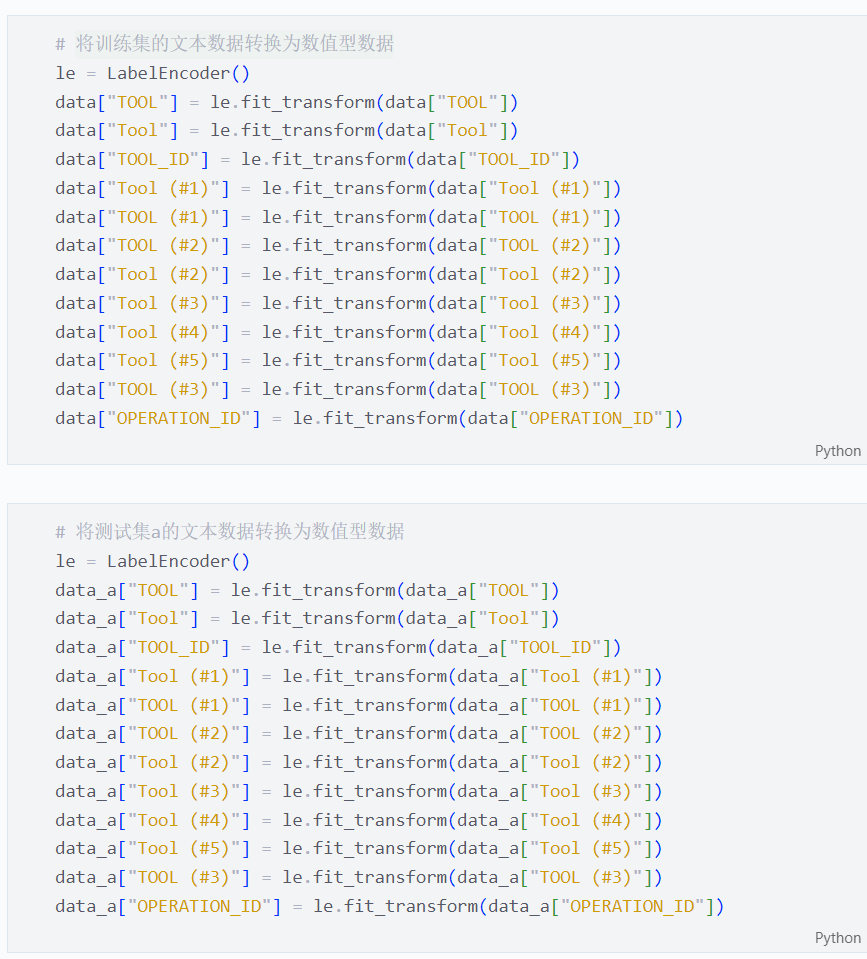
1. **预处理**

导入外部库，导入训练集.xlsx为train(data)，导入测试集A为testa(data\_a)；

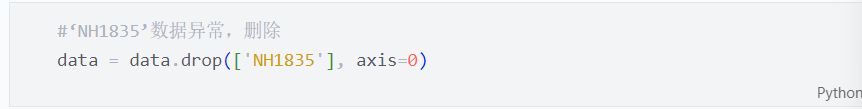


1. **类型转换**

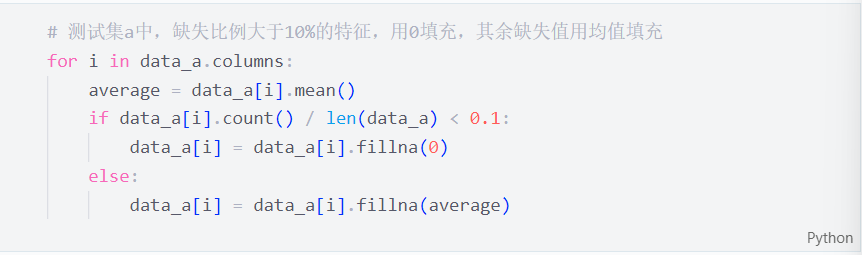
我们借助中的将训练集和测试集A的文本数据（类型，即数据中的工具列“TOOL”等），转换为数值型数据



1. **异常处理**
2. 删除重复样本，共六条；
3. 删除Y值异常的样本，共一条；

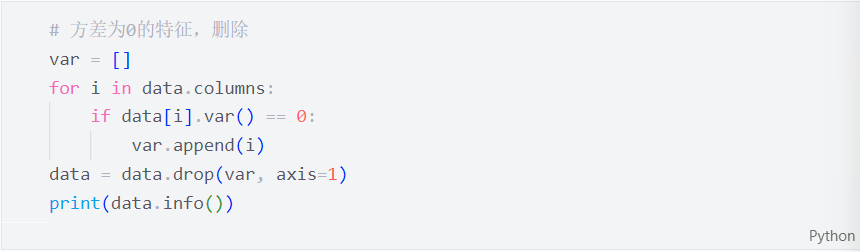


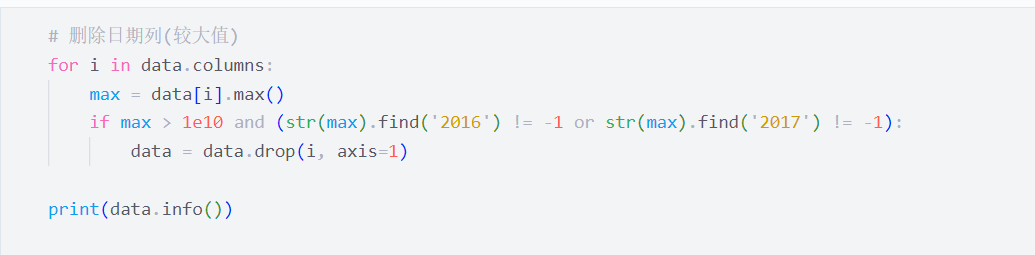
1. 测试集A和训练集中，所有特征均存在缺失值，对于缺失比例大于10%的特征，用0填充；其余缺失值用均值填充；





1. 删除方差为0的特征（即列中数据完全一样）；





1. 删除特征之间相关性（线性化）程度过高的特征，只保留一条以显示这几个特征所传递的信息；

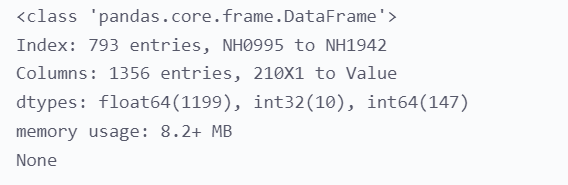


1. 后期处理中，我们注意到训练集和测试集A中的某些特征差距很大，极大地影响了回归模型的稳定性，因此我们删除训练集中与测试集A差距很大的特征。



1. 数据输出

通过数据处理，我们将训练集从原来的5954×801压缩到了1356×793。



1. **线性回归**
   1. **立项阶段**

。

* 1. **外部对接1**

。

* 1. **外部对接2**

动发生意外。

* 1. **筹备阶段**

全。

* 1. **活动阶段1**

。

* 1. **活动阶段2**

：

* 1. **总结阶段**

。

1. **Lasso回归**
2. **Ridge回归**

**5.1 Ridge回归介绍**

回归是一种用于处理多重共线性问题的回归方法，它通过在最小二乘法的损失函数中加入一个正则化项，来减小回归系数的大小，从而提高模型的稳定性和泛化能力，这非常适合本题在线性回归模型表现不佳的情况。

岭回归的损失函数可以表示为：

其中，​是第个观测值，是第个样本的特征向量,是回归系数向量，是样本数量，是特征数量，是正则化参数，也称为岭参数，它控制了正则化项对损失函数的影响程度。当时，岭回归就退化为普通的最小二乘法；当时，岭回归就会对回归系数进行收缩，使得它们更接近于零，从而降低模型的复杂度和方差，增加模型的偏差，达到一种偏差-方差权衡的效果。

岭回归的回归系数可以通过以下公式求解：

其中，是样本的特征矩阵，是样本的观测值向量，是单位矩阵。可以看出，岭回归通过在的对角线上加上，避免了的奇异性，从而保证了逆矩阵的存在，使得回归系数可以估计出来。

岭回归的一个重要问题是如何选择合适的值，一般来说,值越大，回归系数的收缩程度越高，模型的方差越小，但是模型的偏差也越大，可能导致欠拟合的问题；反之，值越小，回归系数的收缩程度越低，模型的方差越大，但是模型的偏差也越小，可能导致过拟合的问题。因此，需要正确选择的值。

岭回归是一种简单而有效的正则化方法，它可以有效地缓解多重共线性问题，提高模型的稳定性和泛化能力，但是它也有一些局限性，比如它不能实现特征选择，因为它不会将回归系数完全压缩到零，而是保留了所有的特征。

本次回归通过中的创建。

**5.2 代码分析**

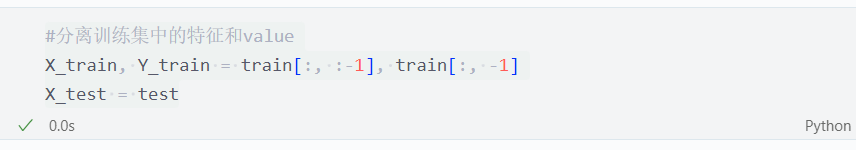
首先导入所需库：



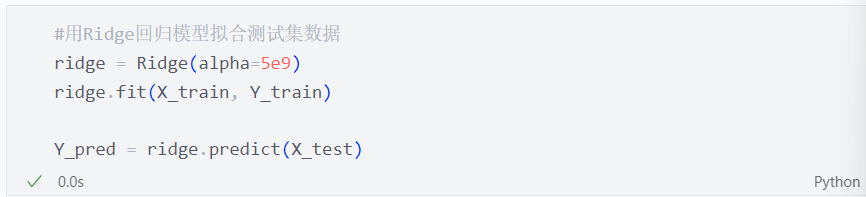
读取数据并删除测试集中训练集没有的特征



通过一系列数据转换，将导入的数据从𝑝𝑎𝑛𝑑𝑎𝑠的𝐷𝑎𝑡𝑎𝐹𝑟𝑎𝑚𝑒类型转换为𝑁𝑢𝑚𝑝𝑦的浮点类型𝑛𝑑𝑎𝑟𝑟𝑎𝑦，并将训练集中的特征数据与Y值分离。

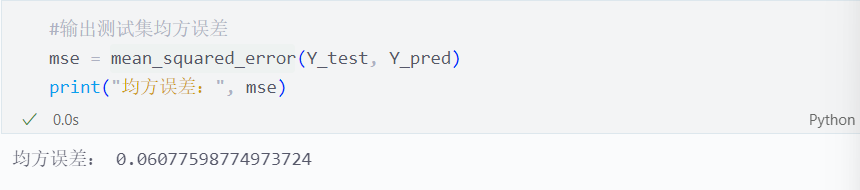


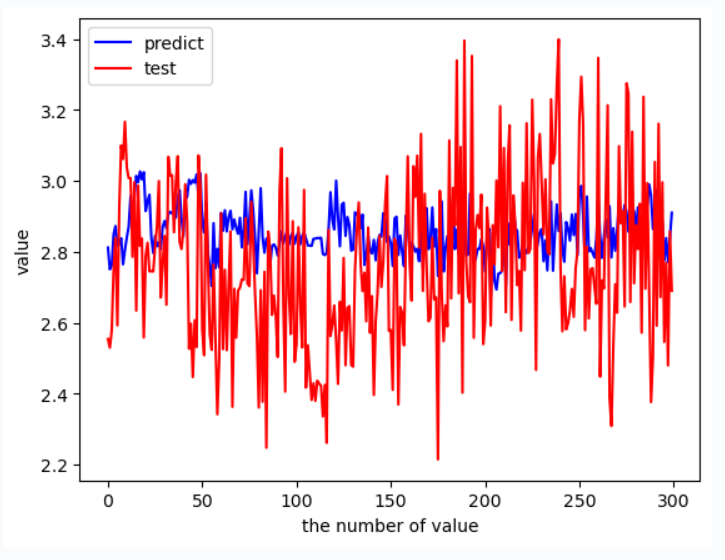
然后通过𝑠𝑘𝑙𝑒𝑎𝑟𝑛.𝑙𝑖𝑛𝑒𝑎𝑟\_通过中的𝑅𝑖𝑑𝑔𝑒创建模型，并以此模型对训练集进行训练，最后得到预测值。



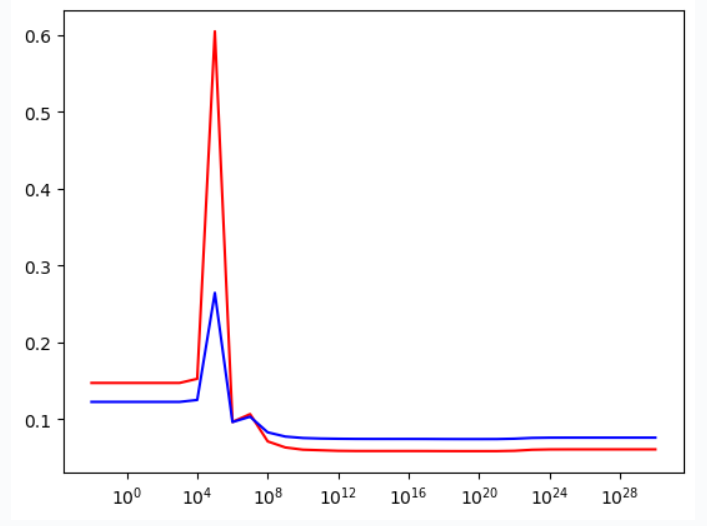
**5.3 运行结果**

得到预测值后，计算预测值与真值之间的均方误差，大约为0.0608：





在𝑅𝑖𝑑𝑔𝑒模型的𝑎𝑙𝑝ℎ𝑎(𝐿2)参数选择时，我遍历了从1𝑒−2至1𝑒30之间（十倍增长）的所有𝑎𝑙𝑝ℎ𝑎值，最终选择了5𝑒9作为我的𝑅𝑖𝑑𝑔𝑒模型的正则参数。



**5.4 分析总结**

数据显示，𝑅𝑖𝑑𝑔𝑒回归模型预测的结果与线性模型相比较好，其均方误差较小，为0.0608；但是将本模型的预测值与真实值放在一起对比，不难观察到除去第1-50组和250-300组的数据，其余预测数据与真实值相差略大，不能很好地跟踪其变化。

其好与坏的原因出自𝑅𝑖𝑑𝑔𝑒回归模型的特性。𝑅𝑖𝑑𝑔𝑒回归模型可以缓解多重共线性问题，提高模型的稳定性和泛化能力，并通过正则化参数控制模型的复杂度，防止过拟合的问题，这从数据和图像中可以看出，相对于线性回归模型，𝑅𝑖𝑑𝑔𝑒回归几乎没有过拟合的现象；且在本题提供如此高维数据的情况下（数据间的共线性性质强），𝑅𝑖𝑑𝑔𝑒回归模型仍能提供较好的预测值（𝑅𝑖𝑑𝑔𝑒回归可以处理特征数大于样本数的情况）。但是本题中我选择了较大的正则系数，这导致了预测数据的毛刺较大，没能更好地拟合真实值的大小。

1. **小组分工**

葛涛：选题，线性回归模型及其PPT、文字报告

刘祥盛：选题，Lasso回归模型及其PPT、文字报告

张庆伟：问题介绍、数据处理、Ridge回归模型，及其PPT、文字报告