算法团队笔试题

1. 企业痛点：

对企业来说现有生产规模和管理制度有以下问题。首先机床不能及时满足下游的新产品需求，且存在设备老化问题；公司内部的维修程序耗费时间长；无关键零件的维修能力，过于依赖制造商会使生产不够灵活。以上四点都会降低企业产线的生产效率。

对管理人员来说，老张一个人管理大量数控机床，面对老化的设备使得管理人员工作量过大；同时生产产品不断更新，也增大了日常工作压力。由于公司管理制度使老张对机台故障的处理不够智能，厂内解决需要填写工单申请也会耗费时间精力。

2. 故障处理的优化方法及价值

采用工业物联网的模式使现有生产线实现智能化。基于大数据处理和机器学习、深度学习算法来预测生产设备故障出现的位置、类型、时间并预测整个生产线的剩余寿命，来最大化设备的生产效率。

在现有设备上安装传感器以采集操作环境信息，例如设备的振动信号、图像采集等。数据集成后通过计算可输出初步描述，工作和管理人员可以实时获取设备的工作状态。

使用协同模型来分层计算，可以重复使用相似算法来提高处理效率。假设可以通过收集历史数据和已知工作状态，使用有监督学习模型，例如线性回归、逻辑回归，KNN，SVM，决策树模型，贝叶斯算法等，不断调整算法参数来提高模型准确率。可以通过判断设备是否发生故障的混淆矩阵比较不同模型的优劣。选出适合的算法后，无论是管理人员还是维修人员都可以第一时间获取设备状态和故障类型。

通过设备智能化来实现设备的故障预测和健康管理，可有效减少管理流程中的冗余和故障维修所耗费的时间，同时提高了工厂的生产效率以及盈利能力，使工厂在同行业中更具有竞争力。

工厂人员排班问题

Q1（人员编号为 1-12）

	一	二	三	四	五	六	日
早班	1，2，7，12	11，10，8	11，7，9	3，2，1	3，4，5，8	5，6	5，4，6
中班	8，9，10，11	1，2，7	1，2，12	8，9	9，10，12	12，11	7，10
晚班	3，4，5	3，6	4，5	6，11，12	7，1，2	10	8，9

思路：

i. 3，4，5，6 每周两次晚班，补偿为每周工作 32 小时（除 5 外）且只上早班，其余人每周工作 40 小时。

ii. 其他八位人约为一次晚班，两次早班，两次中班。

Q2

至少需要 12 人。由于现阶段排班表需要 57 人次，需要 12 人每周上 4-5 天班。

Q3

i. 缩减人员开支可以通过提高绩效管理实现，例如改变“月薪相同”的管理办法。但需要结合工厂的生产情况，如果是钢厂、汽车生产线等不适合计件薪酬。

ii. 通过自动化、物联网减少值班人数，依靠技术升级和工人培训，更新换代机器设备，高效生产的同时也降低了人力成本。

Quiz Projects

1. 影响程度分析

i. 线性回归

表格 1 列出了由线性回归得出的和机器状态最相关的十个自变量，其中，Vat_Valve Step 4 STD 对机器状态的影响最大。

然而 OLS 回归模型调整后的 R_2 接近 0，说明线性回归并不适合逻辑变量。

表格 1 OLS

	feature
1.244231	Vat_Valve Step 4 STD
0.972775	TCP_Rfl_Pwr Step 4 Mean
0.739805	Cl2_Flow Step 5 Mean
0.714532	RF_Pwr Step 5 Max
0.707374	Cl2_Flow Step 5 STD
0.658274	Cl2_Flow Step 4 STD
0.540213	TCP_Rfl_Pwr Step 5 STD
0.413775	RF_Btm_Pwr Step 5 STD
0.360993	TCP_Rfl_Pwr Step 5 Mean
0.360502	Vat_Valve Step 4 Min

ii. 逻辑回归

逻辑回归估计中使用了 ridge 惩罚项，即 l2 正则，步长为 1，迭代次数为 100。

表格 2 列出了由逻辑回归得出的十个和机器状态最相关的自变量，其中 TCP_Load Step 5 Max 的系数最大，系数最大的前十个变量都与与机器的健康状态正相关。使用逻辑回归估计的模型正确率约为 91.4%。若按 70%和 30%的比例划分训练集和测试集，正确率分别为 92.1%和 84.6%，结果比较接近说明模型稳固可靠。

表格 2 logistics

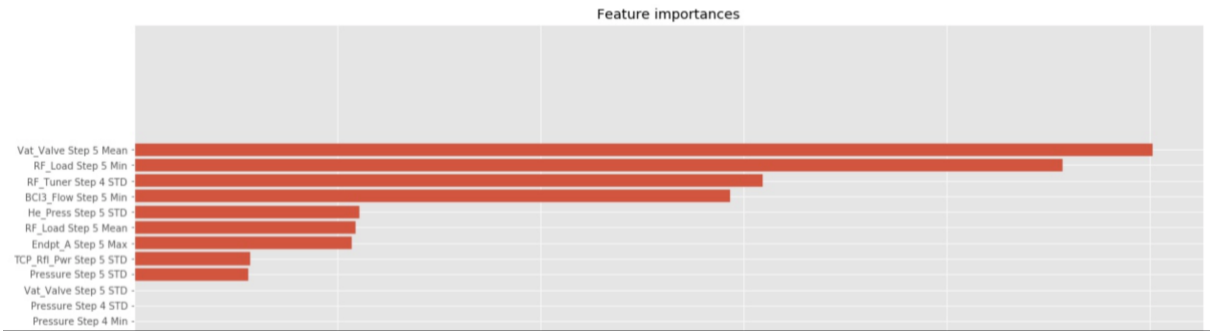
	values
TCP_Load Step 5 Max	0.003906
TCP_Load Step 5 Min	0.003627
RF_Phase_Err Step 5 Min	0.002587
RF_Phase_Err Step 4 Min	0.002568
RF_Impedance Step 5 Min	0.002285
TCP_Impedance Step 5 Max	0.002003
RF_Impedance Step 5 Max	0.001994
RF_Phase_Err Step 4 Mean	0.001764
TCP_Impedance Step 4 Mean	0.001752
RF_Tuner Step 5 Max	0.001641

iii. 决策树回归

使用决策树模型计算变量重要性。重要性排序是根据自上而下结点测试的顺序。图 3 的横坐标是重要性水平，即 Gini 系数的下降值，也是特征选择的标准。

由 图表 1 可读出前九个与机器状态相关的变量，分别为：Vat_Valve Step 5 Mean, RF_Load Step 5 Min, RF_Tuner Step 4 STD, BCI3_Flow Step 5 Min, He_Press Step 5 STD, RF_Load Step 5 Mean, Endpt_A Step 5 Max, TCP_Rfl_Pwr Step 5 STD, Pressure Step 5 STD，其中 Vat_Valve Step 5 Mean 对机器状态影响最大。

决策树估计的正确率为 1，但若按相同比例划分训练测试集，测试集模型准确性降低约 13%，变化较大所以可能存在过拟合。



图表 1 变量重要性

iv. 结论

通过比较三种模型的结果，逻辑回归的估计结果更稳定，正确率较高，因此以下变量与机器状态最相关：

表格 3 相关变量

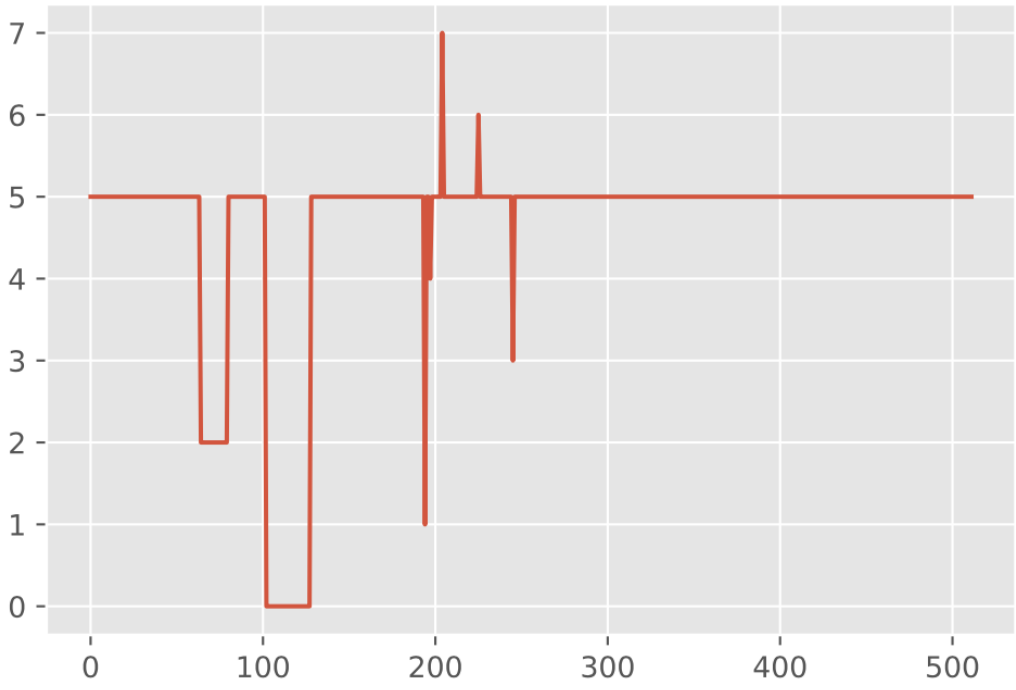
TCP_Load Step 5 Max
TCP_Load Step 5 Min
RF_Phase_Err Step 5 Min
RF_Phase_Err Step 4 Min
RF_Impedance Step 5 Min
TCP_Impedance Step 5 Max
RF_Impedance Step 5 Max
RF_Phase_Err Step 4 Mean
TCP_Impedance Step 4 Mean
RF_Tuner Step 5 Max

2. 基于振动信号的故障诊断

i. k-means 算法

随机初始化个聚类中心；对于每个数据点，寻找离它最近的聚类中心，将其归入该类；更新聚类中心的值为所有属于类的数据点的平均值，重复 2、3 步直到收敛或者达到最大迭代次数。

根据无监督学习结果，设置类别数为 8 后计算得健康状态为“5”，其余七种故障或故障组合中出现“0”故障的记录最多。



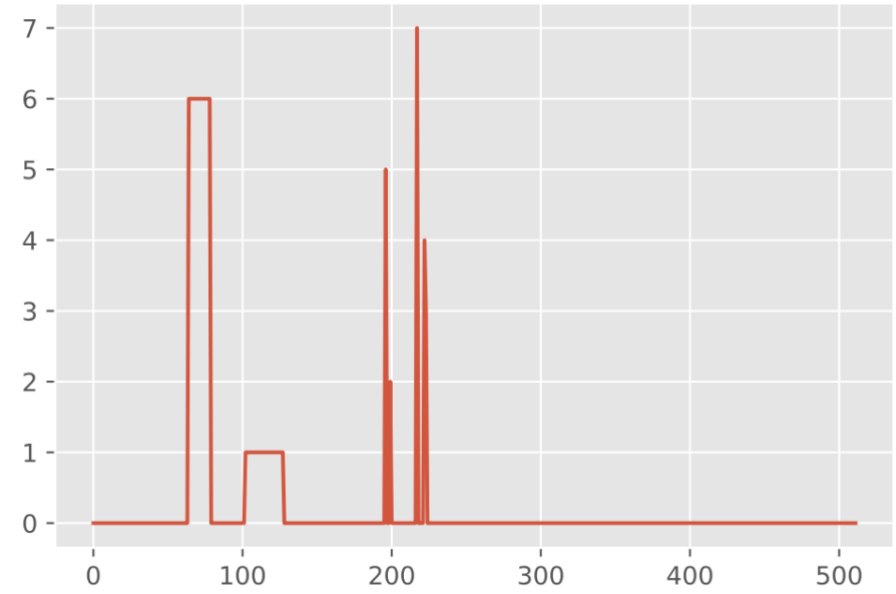
图表 2 原数据 kmeans

选择轮廓系数作为评价聚类效果的标准，使用原数据进行 k-means 学习的轮廓系数为 0.0433.

ii. PCA

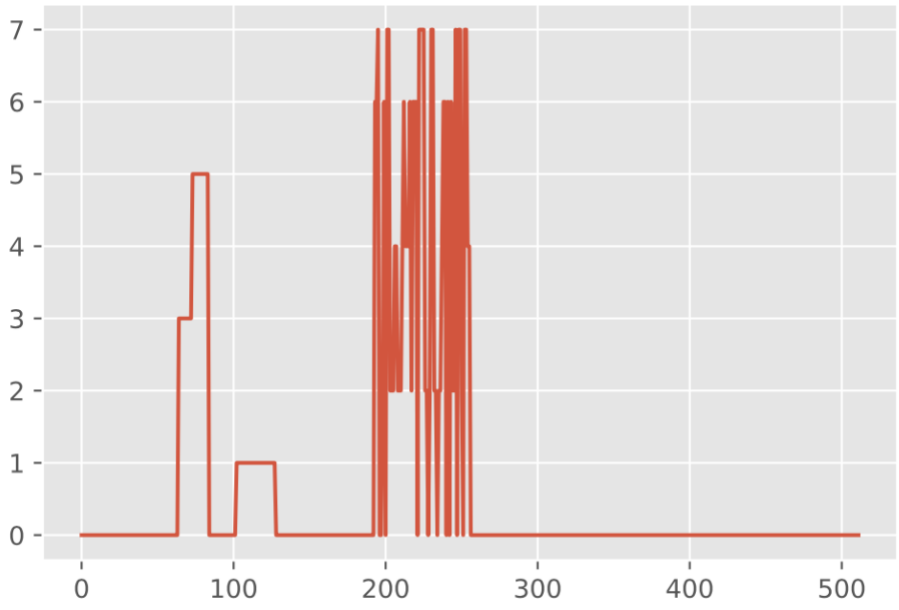
使用主成分分析帮助原数据降维来提高分类准确度。

首先根据贡献率，选出大于 95%的主成分个数，得到的新数据有 313 个特征列；再使用 kmeans 聚类，得到结果如 图表 3 所示，分类“0”为健康状态，类别“1”的故障状态数量最多。



图表 3 PCA 95%

降维后的聚类轮廓系数为 0.0854，说明 PCA 方法有效，因此继续降维，根据震动信号的基本信息，选出 6 个主成分，得到结果如下：

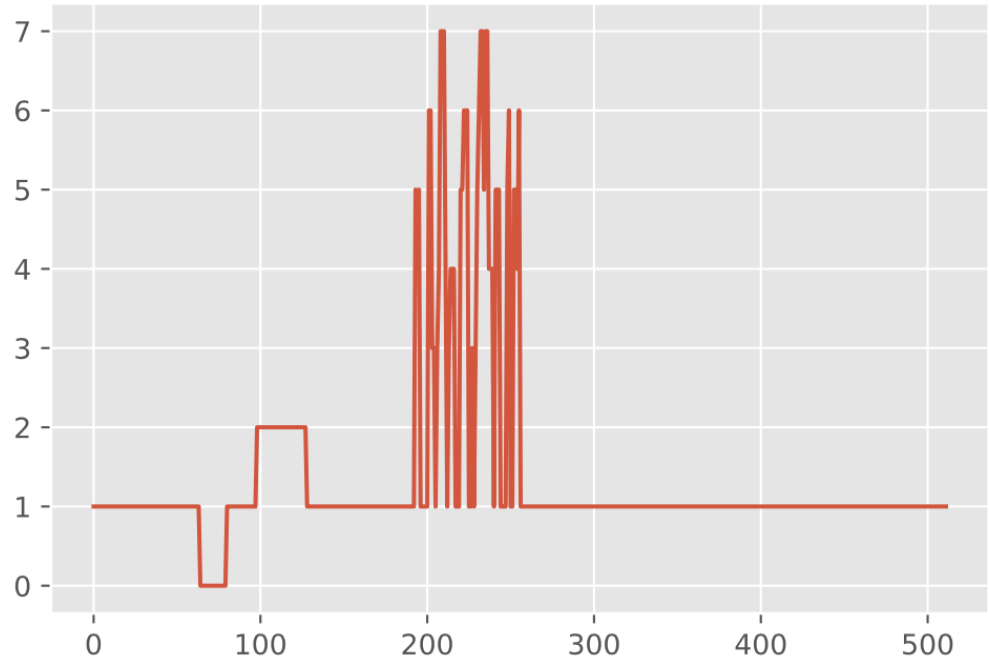


图表 4 PCA (6)

使用被显著降维的数据聚类，轮廓系数显著上升为 0.6383。同样为类别“0”为健康状态。

iii. 继承聚类

继承聚类的思路是从上至下学习，将每一个元素单独定为一类，每一轮都合并距离最小的类，直到所有的元素都归为同一类。其中，合并距离的参数选择为两类对象之间的最小距离，即‘ward’连接，同样设置类别总数为 8，得到结果图表 5 所示。继承聚类下，类别“1”为健康状态，聚类的轮廓系数为 0.6339.



图表 5 继承聚类

iv. 混合高斯模型（GMM）

混合高斯模型的聚类思路为，将所有分布看做多个高斯分布的综合，通过迭代计算八种分布的期望和方差来分类。输入降维后的数据得到的结果轮廓系数约为 0.0085，显著小于其他算法，且不能从分类图中区分健康和故障状态，因此 GMM 不适合本震动信号的故障判断。

v. 结论与待补充内容

以轮廓系数为标准可以看出，经过主成分分析降维至 6 维后使用 k-means 算法的聚类效果最好，保留学习模型，将收集到的新的振动信号输入至学习模型可自动判断故障类型。

通过划分训练集和测试集，模拟自动判断振动信号对应轴承的故障类型，得到测试集轮廓系数为 0.6084，与之前结果接近说明学习模型稳定。

在振动信号的特征分析中，主成分分析降维并不能与振动信号的基本信息联系起来是学习模型的不足之处；以及继承聚类中尝试使用 scipy 中的算法，但由于无监督学习没有列名称，不能提取出有价值的信息；最后，多次尝试同一 kmeans 聚类会得到截然不同的结果，说明学习模型不稳定。以上都是振动信号的故障诊断课题中待改进的内容。