机器学习大作业研究报告

Kaggle账号：Cihao

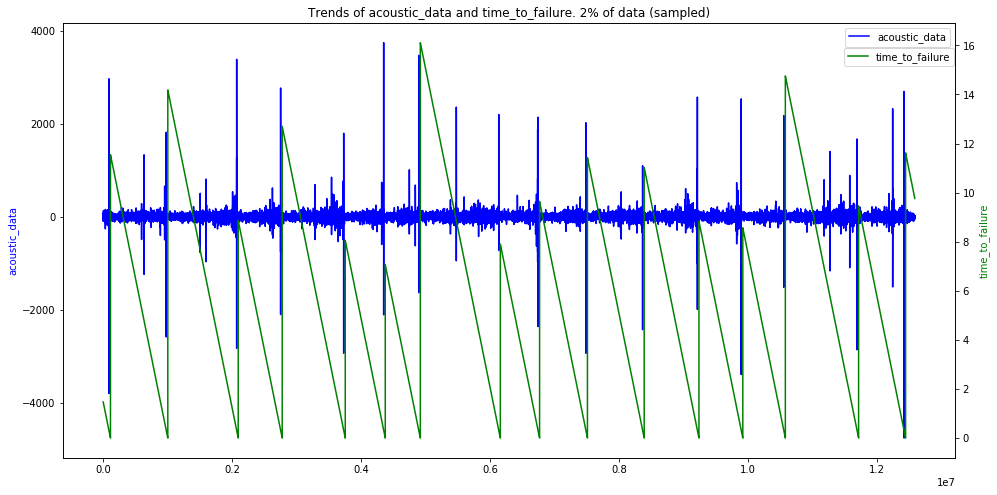
黄梓浩，郑昊思，黎炳华，贾浩

1. 问题提出

准确地预测地震能够避免较大的损失，地震预测问题一般包括时间、地点和规模，在该问题中集中关注地震预测的时间。该问题的数据来自Los Alamos国家实验室，需要我们根据实验中的声学信号数据来预测下次地震发生时间。

1. 现状与分析

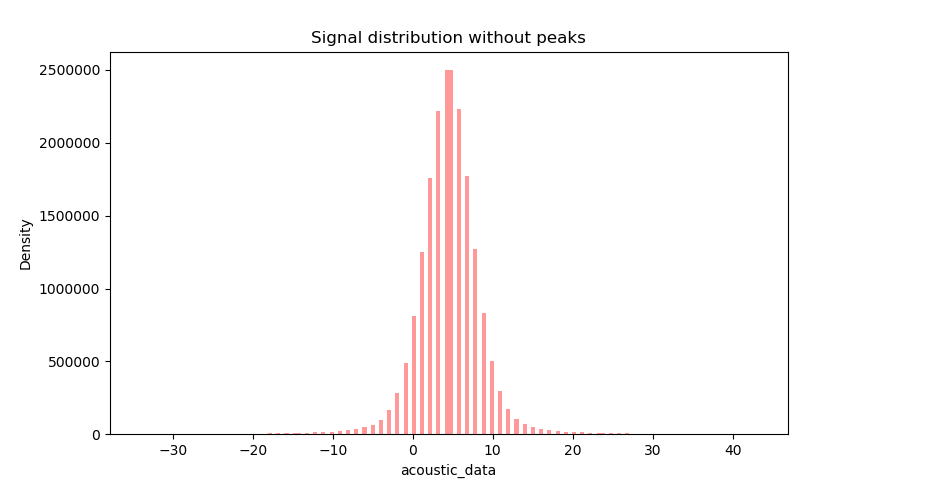
train.csv文件包含训练数据与标签，共六亿多行，第一列“acoustic\_data”是训练数据，含义是一段非周期性的声学信号；第二列“time\_to\_failure”是数据标签，含义是距离下次地震发生剩余的时间。测试集总共包含2624个csv文件，文件名与sample\_submission.csv中的“seg\_id”列对应，代表2624段声学信号。该问题简而言之就是一个回归任务，根据tran.csv中的数据训练回归模型预测2624段声学信号最后时刻距离下次地震发生的时间。



**图 1** 2%采样的训练数据

通过2%的采样绘制训练集数据，观察发现：

* 共16个点的time\_to\_failure为0，即训练集数据中一共发生16次地震
* 临近地震时信号波动较剧烈，有时在两次地震之间的某段时间也有较剧烈波动，故在特征工程中应考虑提取频域特征
* 临近地震发生时的信号一般幅度较大，远离地震发生的信号一般幅度较小，故在特征工程中应考虑提取时域特征



**图 2** acoustic\_data分布图

绘制前两千万行的acoustic\_data数据的分布图发现：

* 信号数据的分布类似高斯分布，均值在4左右，且取值较为离散。

生成训练样本：我们将训练序列分为16个段，每个段落分别代表一次记录完整的地震数据，从每个段落中随机选取300个点，这300个点往后记录149999个点组成300个训练样本，共4800个训练样本。

1. 特征工程
   1. 频域特征

首先选择4阶巴特沃斯低通滤波器滤除18000Hz以上的噪声。然后对数据进行傅里叶变换，获得频域幅值序列，并分为虚数和实数序列。

* + 1. 根据计算样本的频域幅值，然后每隔2500hz将样本频域分段，计算各段的1分位数、10分位数、90分位数和99分位数，以及各段的均值、标准差、最大值。
    2. 根据计算样本的相位幅值，然后每隔2500hz将样本频域分段，计算各段的均值和标准差。
    3. 分别取虚数序列和实数序列的均值、标准差、最大值和最小值。
    4. 分别取实数序列前6000个点和18000个点的均值、标准差、最大值和最小值。
  1. 时间域
     1. 记录样本中尖峰出现的次数，尖峰定义为周围10/20/50/100个点的值均小于该点时认为该点是一个尖峰。
     2. 计算自相关系数，公式如下



其中n为序列长度，l为滞后长度，为序列均值，为序列方差。我们设置l的值为5、10、50、100、500、1000、5000、10000计算序列的自相关系数。

* + 1. 计算时间序列的均值、标准差、最大值和最小值。
    2. 计算时间序列的汉明均值，公式如下：



* + 1. 计算时间序列的集合均值，公式如下：



* + 1. 计算时间序列的nth k-statistic，这里n的取值范围为1-4，公式为：



其中为样本均值，为样本的方差，为样本的第i个中心矩。

* + 1. 计算时间序列的1-4阶中心距，公式如下：



* + 1. 计算时间序列的1阶和2阶统计量方差的无偏估计量，公式如下：



其中和为2th k-statistic和4th k-statistic。

* + 1. 计算时间序列最大值跟最小值的比例的绝对值
    2. 计算时间序列最大值跟最小值绝对值的差
    3. 记录时间序列中绝对值大于500的尖峰的数量
    4. 计算时间序列各点的总和。
    5. 计算时间序列各点的微分的均值。
    6. 计算时间序列各点变化率的均值。
    7. 计算时间序列绝对值的最大值和最小值。
    8. 计算时间序列前10000、50000个点和后50000、10000个点的均值、标准差、最大值和最小值。
    9. 计算时间序列和时间序列的绝对值的1, 5, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, 95, 99分位数。
    10. 建立一个由10、50、100、500、1000、5000个点组成的窗口，在样本上步长为1滑动取每个窗口的平均值和标准差。在获得的新样本上，计算标准差序列的均值、标准差、最大值和最小值，以及1, 5, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, 95, 99分位数，计算标准差的微分均值和变化率。计算均值序列的均值、标准差、最大值和最小值，1, 5, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, 95, 99分位数，计算均值的微分均值和变化率。

**最后每个样本获得441个特征。**

1. 模型描述
   1. LightGBM

XGBoost是一类GBDT模型，其基于预排序的分裂点搜索方法让XGBoost模型能更精确地搜索到最优的分裂点，达到较好的分类效果。然而，精确的分类点极容易导致模型过拟合，且预排序算法及模型中决策树节点的层生长模式使得模型的计算代价和存储代价巨大。

LightGBM是一类更为轻便的GBDT模型，其基于直方图优化对XGBoost进行了改善，在搜索分裂点时不再基于预排序算法，而按特征将数据分块装桶，相当于每个特征获得一个直方图表达。基于这种方式的分裂点搜索既大大降低了计算代价和存储代价，又防止了精确分裂点带来的过拟合。

* 1. 模型对比与实验设计

为了更好地对比LightGBM和XGBoost之间的优劣并取得更好的学习效果，我们采用了三个模型对声学数据进行回归。

1. XGBoost
2. LightGBM
3. XGBoost + LightGBM

每个模型均采用上述讨论得到的特征进行训练，损失函数采用平均绝对值误差(MAE)，最终损失计算采用五折交叉验证得出。关于五折交叉验证，我们采用的不是基于时间点的交叉，而是基于地震次数的交叉。即，我们按地震次数将数据分作16段，每三段地震数据作为一折，最后一折取4段地震数据。每次取其中四折训练，剩下一折作为测试集计算损失值。模型的最终损失为五个损失值的均值。

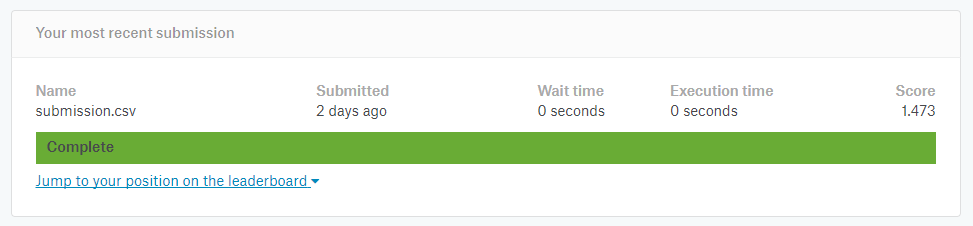
1. 实验结果与分析
   1. 实验结果

模型的最终提交分数如表1：

**表 1** 3种模型的Kaggle提交分数

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | XGBoost | LightGBM | XGBoost+LightGBM |
| 提交分数 | 1.489 | 1.473 | 1.507 |

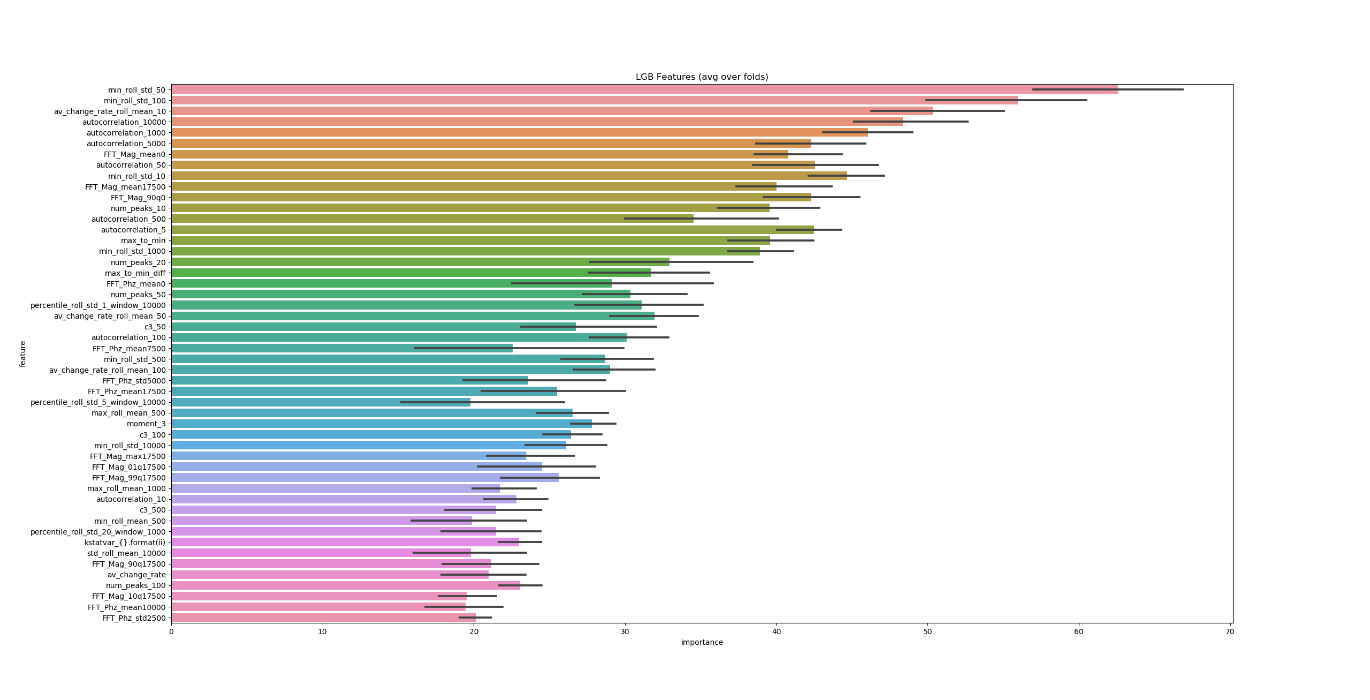
从提交结果来看，LightGBM模型在改组特征中表现较好，优于XGBoost模型与XGBoost和LightGBM的混合模型。最后将LightGBM的预测结果提交到Kaggle，具体如图3。



**图 3** LightGBM的Kaggle提交分数

* 1. 结果分析与工作展望

1．考虑到我们提取了较多特征，可能会导致特征冗余的情况。采用LightGBM评估的特征重要性如图4（由于特征太多，这里只是可视化了前50个特征）。



**图 4** 前50个特征的LightGBM特征重要性分数

后续考虑对重要性较低的特征进行筛选，降低特征维数。

2．考虑结合深度学习模型，如LSTM、CNN等模型，与LightGBM进行模型混合。

3．尝试改变每个样本的窗口大小，不再采用150000个连续时刻点为一个样本，而是用50000、100000个连续点为一个样本，用同样的特征进行学习。