1、数据是dat格式，我不清楚你用什么软件打开。如果打不开，可以把文件后缀改为TXT即可。

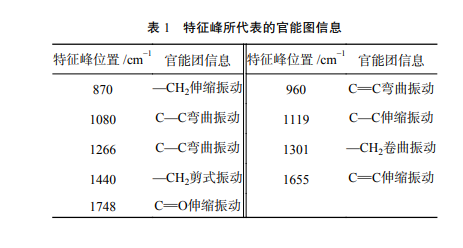
2、文件名称为“大豆油浓度-玉米油浓度-茶油浓度” 。你只用做三元混合的回归模型，但我把另外两种二元混合数据也发给你了，你可以放在一起试试，反正我偏最小二乘回归做出来效果肯定没那么好。

3、实验数据是原始数据，即照射的波长范围是从100多到1800cm-1范围。并且谱图只做了扣除背景一个预处理操作，后面阶段我还会给你求了一阶导之类的其他谱图预处理的数据给你，不过不急可以慢慢来。

现阶段你大致用不同的方法试试就行，因为我不清楚你Python建模不同方法需要花多少时间。如果一种方法需要花好几个小时那种，你就每种方法单独对茶油浓度回归分析就行，目前我们只是看看不同方法的效果。

4、建模时需要提取有代表信息的波段再建模，即我们之前说的含量有变化的那几个波峰。这一步最好你自己用主成分分析降维提取，因为比如9个特征波峰，并不是9个都用来建模效果才最好，可能这个模型用前六个波峰效果可能更好。

下图即为通过化学知识得到的9个特征波峰，我用PCA降维提取出来也是这个几个波峰。另外由于每台仪器不同，所以位置会有很小的变化，比如我用PCA提取的波峰是843、865、966、1077、1263、1301、1438、1651、1656这几个。但同一台仪器测不同次是一样的，其实中966、1263、1651这几个比较重要。



5、根据你看的文献你也可以发现，评判一个模型好坏的标准是看R²和均方根误差RMSE，只要R²大致能在0.95以上，在RMSE能小于1点多就行。

6、有问题及时沟通发语音都行，我基本上随时微信都在，因为可能你会有化学上的东西不太明白，然后没啥情况，我们争取7月13号前后在腾讯会议上交流一波。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 模型 | 数据集 | test\_size | learning\_rate | max\_depth | n\_estimators | RMSE | R² |
| XGBoost | 双掺 | 默认参数 | 默认参数 | 默认参数 | 默认参数 | 11.47 | 0.17 |
| 双掺 | 0.2 | 0.1 | 3 | 1000 | 11.4 | 0.12 |
| 双掺+玉米 | 默认参数 | 默认参数 | 默认参数 | 默认参数 | 10.86 | 0.22 |
| 双掺+玉米 | 0.2 | 0.1 | 5 | 50 | 9.44 | 0.43 |
| 双掺+大豆 | 默认参数 | 默认参数 | 默认参数 | 默认参数 | 10.85 | 0.71 |
| 双掺+大豆 | 0.2 | 0.01 | 3 | 5000 | 10.38 | 0.72 |
| 双掺+玉米+大豆 | 默认参数 | 默认参数 | 默认参数 | 默认参数 | 18.23 | 0.09 |
| 双掺+玉米+大豆 | 0.2 | 0.01 | 3 | 5000 | 18.04 | 0.12 |