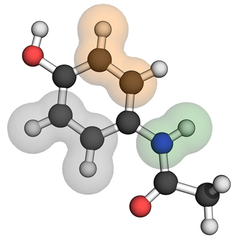
### 思路



任务描述:

我们需要使用一些原子的信息来预测原子间的scalar\_coupling\_constant( 就是原子与原子间的磁力相互作用)。

数据描述:

1. train.csv 里面有466万的数据，每一行有id（序号）, molecule\_name(分子名字), atom\_index\_0(第1个原子在对应分子里面的序号), atom\_index\_1(第2个原子在对应原子里面的序号), type(耦合键的类型), scalar\_coupling\_constant(耦合系数常量).
2. test.csv 里面有251万的数据，除了scalar\_coupling\_constant这个没有，因为这个是需要我们去预测的。
3. structure.csv 里有236万行，molecule\_name(分子名字), atom\_index(原子在分子里面的序号), atom(原子类型: H,C,N,O,F), x,y,z(最后3个为原子坐标)

也就是说，通过把train.csv, 和test.csv与structure.csv就可以知道atom\_index\_0，atom\_index\_1对应的原子类型，以及对应的x,y,z。还有一些其它的文件，用于生成QM9特征。

思路:

1．生成距离特征:先计算两个原子的中心，然后找到离这个中心最近的N个原子(x,y,z)，以及计算N个原子间的两两距离(d)。

2. 加入QM9特征(这个涉及比较多的专业知识，我直接参考了别人的)，并删除重复特征

1. 使用xgboost,lightgbm ,nn等模型进行数据拟合，分别对不同的原子耦合类型(如1JHC,1JHN,2JHC等等，总共8个)搭不同的模型。
2. nn模型:

4层神经网络，神经元个数分别为2048->1024->512->1

层与层之间使用BatchNormalization。

使用adam优化器，进行优化。

1. lightgbm模型

最大迭代次数4000(比xgboost收敛得快),200次效果没有提升后结束更新,subsample取0.9，eta(学习率)取0.1，最大深度11，mae作为损失函数。

1. xgboost模型

最大迭代次数6000,200次效果没有提升后结束更新,subsample取0.9，eta(学习率)取0.1，最大深度11，mae作为损失函数。

模型集成部分:

11．stack集成

将lightgbm,xgboost,nn作为基学习器，使用median集成，就是对lightgbm, xgboost, nn等生成的测试文件，取中值。试了其它办法，但是不好。

其它:比赛的评价指标是log(mae)，但是xgboost,lightgbm,nn里面用的是mae作为损失函数,catboost也做了，但是效果不好，跑得比较慢。