# 程序解释

整个程序分为两个部分，全量模型和增量模型，分别为API.py和delta\_Y9\_2.py，两个除了思路类似其他的差别较大，训练数据，验证数据有很大不同。由此也得到两种pearson的验证代码pearson\_Y9.py, pearson\_Y2\_Y7.

## 功能

根据现有其他公司数据训练模型，进行交叉验证，在华为上验证以及预测华为未来数据。

## 输入输出

### 输入：

* 训练起始时间year\_start
* 训练结束时间 year\_end
* 数据开始时间 data\_start
* 自变量列号，分别在otherFeatureList和baseFeatureList,分别代表不随时间变化的量和随时间变化的量，表示在训练数据中数据开始时间的列号
* 因变量名和列号，col\_name\_dict表示因变量在训练数据中数据开始时的列号,var\_name表示当前变量名
* 训练数据路径
* 华为验证数据路径
* 华为预测数据路径

### 输出

交叉验证的结果txt，华为验证结果txt，预测未来几年结果（显示在命令行中）

### 原理：

1. 读入训练数据
2. 根据训练起始与结束时间算出持续时间year\_last和训练起始时间与数据开始时间之间的间隔year\_gap
3. 利用year\_gap生成实际数据起始列号，year\_last抽取相应列，形成中间变量（包括训练集和华为验证集）
4. 根据滑动窗口大小（这里设为1）抽取标签，构建训练数据集和标签集
5. 带入模型训练，交叉验证，华为验证，预测

## 流程

在预测烧钱率的所有程序中，都是遵照如下的流程：

输入训练起止时间，如1997-2016，注意由于使用到时间序列的滑窗，所以起始时间为最开始的时间，结束时间为最后截止时间加上滑动窗口表示的时间。比如预测1999-2001，滑动窗口大小为1，所以输入时时间为1999,2002.

输入数据起始时间，比如数据是从1997年开始，则将数据起始时间设为1997.

输入自变量因变量的列号，这里统一为数据起始时间的列号，比如某变量从1997年开始，则将1997年盖变量的列号作为输入，在运算时根据训练开始的时间进行推算得到准确的变量列号。

训练数据为除去华为剩下的1百多家公司，训练是从这里面随机抽出一部分作为交叉验证的验证集，其余作为训练集，得到交叉验证的误差。再次用华为作为验证时是用所有公司的在相应年份的变量数据作为训练，最后输入华为相应年份的数据得到预测值，比如1999-2001年，将所有公司数据输入作为训练后输入华为数据，得到华为2000-2002年的预测数据，再将它与华为的实际数据进行比对算出最大最小均值中值误差。对于2016-2020的数据同样如此，利用全部公司训练的模型，输入华为2016-2020的自变量数据，得到2017-2021的预测数据。

## 注意事项

预测数据集并未按照训练和验证那样按列抽取，而是自己构建的，预测时读取相应文件就行了