遗传规划项目使用说明

1. 项目概览

* 项目文件夹中一共包括4个配置文件夹和5个Python程序，配置文件夹中又包括其他相关Python程序和数据文件
* 配置文件夹database用来保存程序中调用的数据文件（复权后）
* factor\_result文件夹中保存了存储因子表达式有关信息的csv，以及运行因子复现函数后得到的MATLAB数据格式因子
* markbaka为一个Python包，其中包含所有可以调用的函数
* structure\_code为遗传规划配置文件，其中包含了遗传规划的构建程序
* factor\_realize为因子复现的程序
* genetic\_programming\_settings为设置遗传规划参数的文件
* load\_data为导入遗传规划数据的函数
* load\_data\_realize为导入因子复现数据的函数
* main为运行遗传规划的主程序

1. 部分操作说明
2. 导入数据

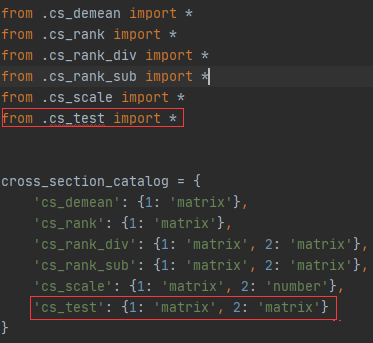
* 数据应当保存为numpy格式（.npy）文件，注意应当保存复权后的数据文件
* 数据统一放入dadabase文件夹中
* 如果添加了新的文件，则需要更改三处文件：

1. load\_data文件中增加对该数据的导入
2. load\_data\_realize文件中增加对该数据的导入，命名需要与load\_data相同
3. Genetic\_programming\_settings文件中的matrix变量增加对应名称的字符串， 注意字符串的内容应当与load\_data和load\_data\_realize中的对应变量名称保持 一致
4. 添加函数

* 添加函数时，需要将函数放入markbaka的Python包中对应的子包
* 需要在该子包的\_\_init\_\_文件下将该函数导入，且更新对应的函数对应参数字 典
* 例子：如添加了横截面类型函数cs\_test(x, y)，x与y均要求为矩阵类型，则除

了在cross\_section下添加对应的函数之外，还需要在cross\_section下的\_\_init\_\_

文件中添加下面红色的部分：

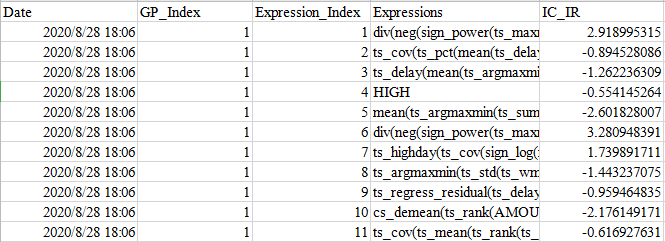
* genetic\_programming\_settings中的operator为使用函数集合的变量，应当为列 表类型，可以结合希望使用的函数自行调整

1. 参数调整

* 遗传规划的所有参数均设置在genetic\_progamming\_settings文件中，如下：

|  |  |
| --- | --- |
| operator | 引用函数的名称汇总，列表 |
| matrix | 引用矩阵类型数据名称的汇总，列表 |
| number | 引用数值类型数据的汇总，列表 |
| operator\_max | 每个表达式包含的算子的个数上限 |
| operator\_min | 每个表达式包含的算子的个数下限 |
| survive\_rate | 每代的存活率 |
| variation\_rate | 每代变异环节变异的概率 |
| generations | 遗传代数 |
| expression\_number | 每代初始生成表达式的个数 |
| step | 持仓周期 |
| date\_range | 日期范围，为一个数组，表示取哪些行 |

1. 结果生成

* 每次运行后，会在factor\_result中Factor\_Expressions.csv存入如下形式的信息：

分别对应记录日期，遗传规划运行编号，表达式编号，表达式内容，表达式对应因子IC\_IR值；每个遗传规划运行编号和表达式编号唯一对应一个表达式

* 在运行factor\_realize，输入对应的GP\_Index与Expression\_Index后，可以得到存放于factor\_result中的MATLAB文件

1. 备注

* 本版本使用年化IC\_IR作为评价因子好坏的指标，如果希望使用其他计算指标，则需要更改structure\_code中的fitness函数以及genetic\_programming中的调用fitness函数的部分；若希望在Factor\_Expressions.csv中记录对应的新指标值，则需要对genetic\_programming做对应记录部分的修改
* Structure\_code中的genetic\_programming中的fitness\_collection\_all字典存放有本次遗传规划生成的所有表达式以及对应的适应度值，但是没有输出此信息，如果需要获取生成过程中所有的表达式信息，则可以查看此变量
* 当step不为1时，可能在时间序列上的滚动计算方面有一定的隐患，暂时还没有好的处理方法，后续需要继续改进
* 当所有截面都运行函数时，效率比较低，暂时的处理方法为仅仅对最后的730个交易日进行计算