1917—2017
100th Anniversary
Shanghai University of Finance and Economics
上海财经大学 100周年校庆

基于遗传规划的因子挖掘

算法分享

发言人：吴展鹏
日　期：2023.12.13

目录

基于遗传规划的因子挖掘

因子挖掘

(1)因子种类

- **演绎法：先有逻辑、后有公式**

如估值、成长、波动率等，可认为是投资者经验的演绎

- **归纳法：先有公式、后有逻辑**

1. 靠遗传规划等技术手段生成
2. 检验因子有效性
3. 试图解释有效因子内涵

基于遗传规划的因子挖掘

因子挖掘

(2)因子检验方法

- **回归法**

因子与目标值进行回归，系数显著不等于0，则因子有效

- **IC(Information Coefficient)值分析法**
  - 计算因子和目标值的相关性
  - 计算公式：
    - Pearson 相关系数：采用原始因子值，受极端值影响较大
    - Spearman 秩相关系数（Rank IC）：基于变量排名计算相关性，更稳健

- **分层回测法**

根据因子值将股票进行分组，若Top组和Bottom组的收益长期稳定区别于Middle组，则该因子对收益预测存在稳定的非线性规律

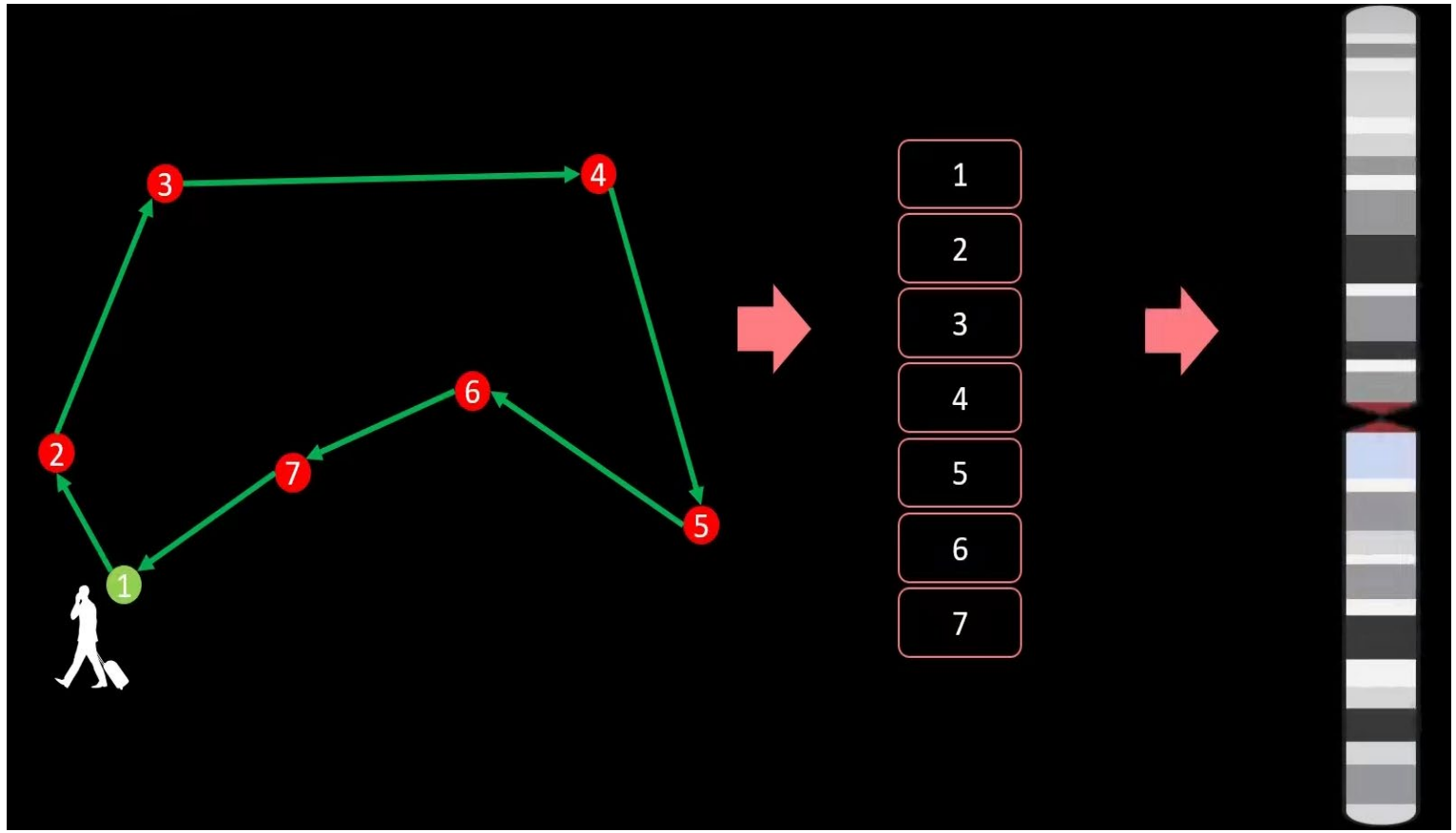
基于遗传规划的因子挖掘

基于遗传规划的因子挖掘

遗传算法

- **TSP商旅路径问题**

商人从城市1出发，前往其他多个城市出差，每个城市仅能途经一次，最后回到初始城市。则N座城市可能的路径为$\frac{(N-1)!}{2}$,目标是找到最短的路径



- **染色体**

将不同城市按一定顺序排列形成一个序列，将这样一个路径序列视为为一条染色体

- **基因**

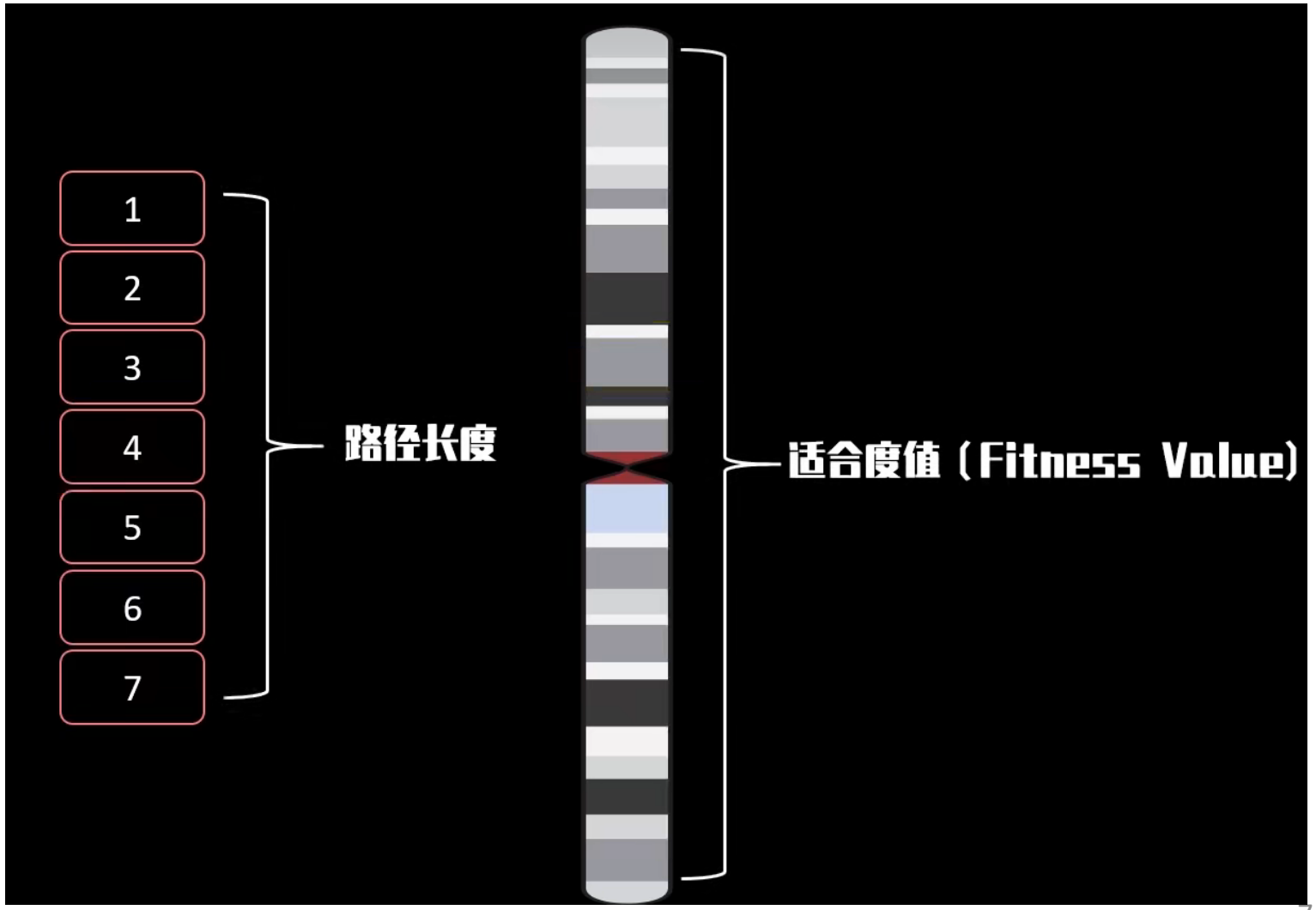
每个城市视为一个基因单位

基于遗传规划的因子挖掘

遗传算法

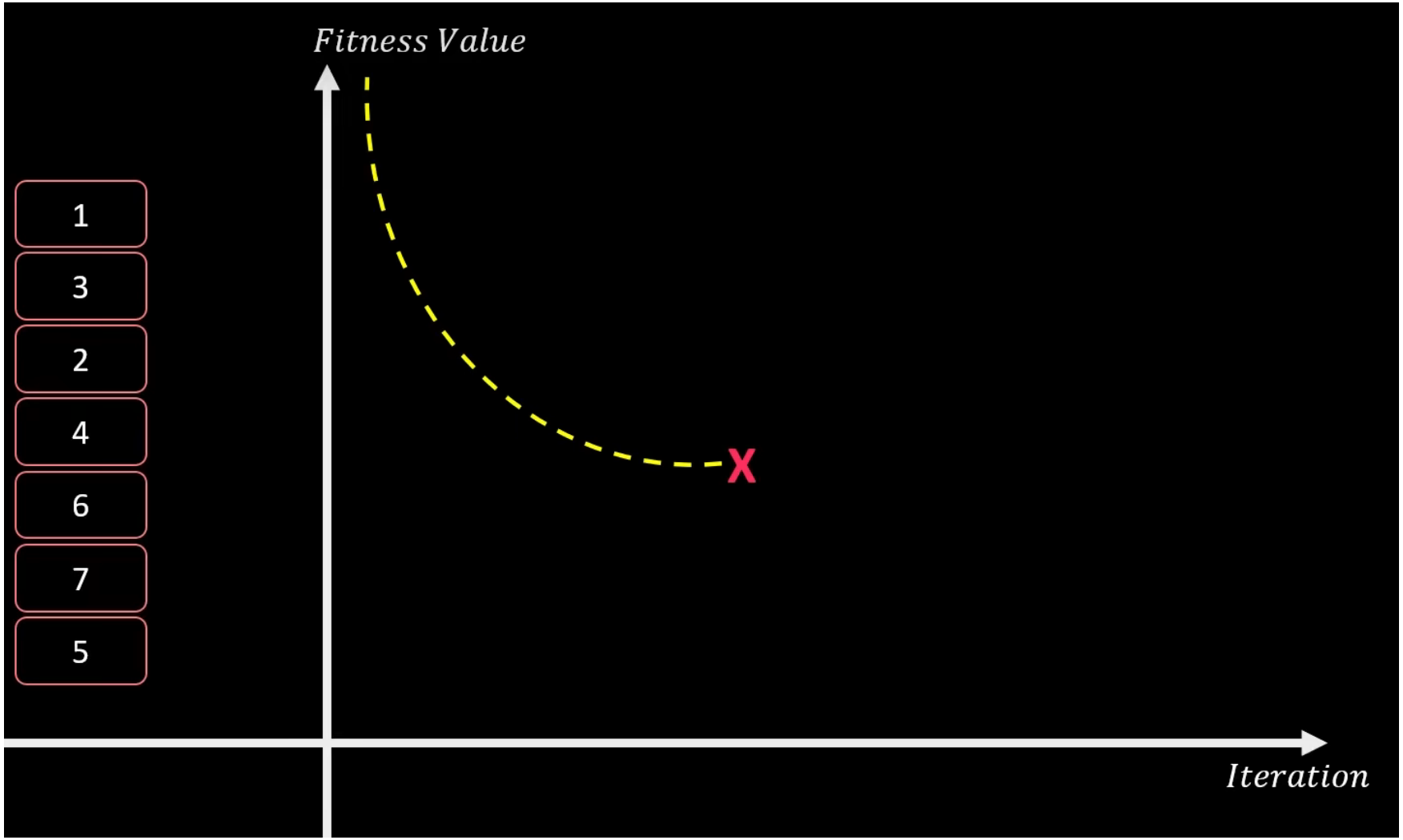
- **适应度值(Fitness value)**

类似目标函数，用来评判染色体的优劣



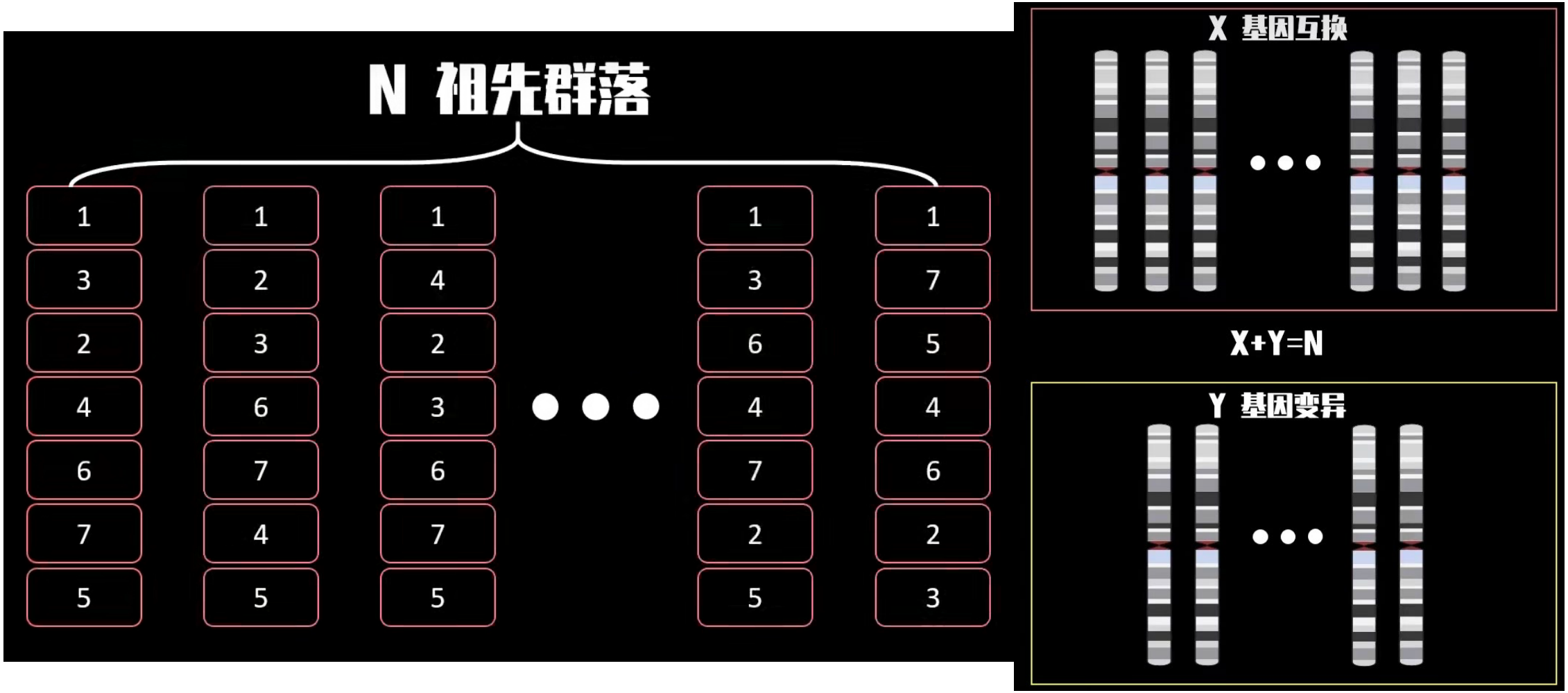
基于遗传规划的因子挖掘

遗传算法



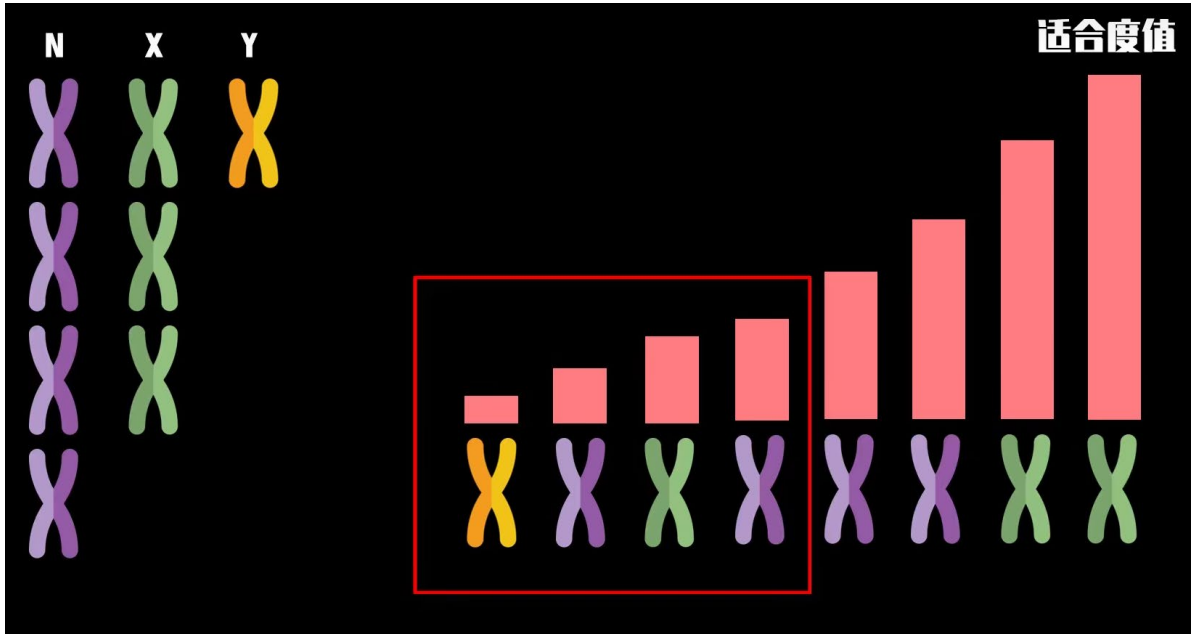
基于遗传规划的因子挖掘

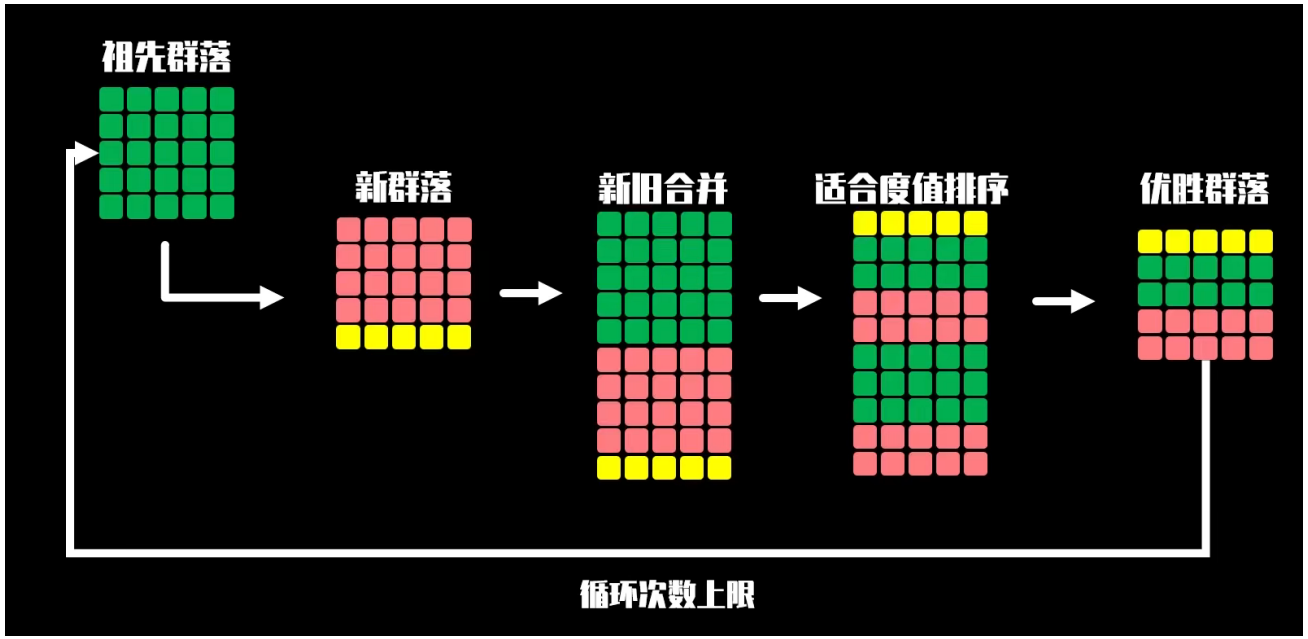
遗传算法



基于遗传规划的因子挖掘

遗传算法

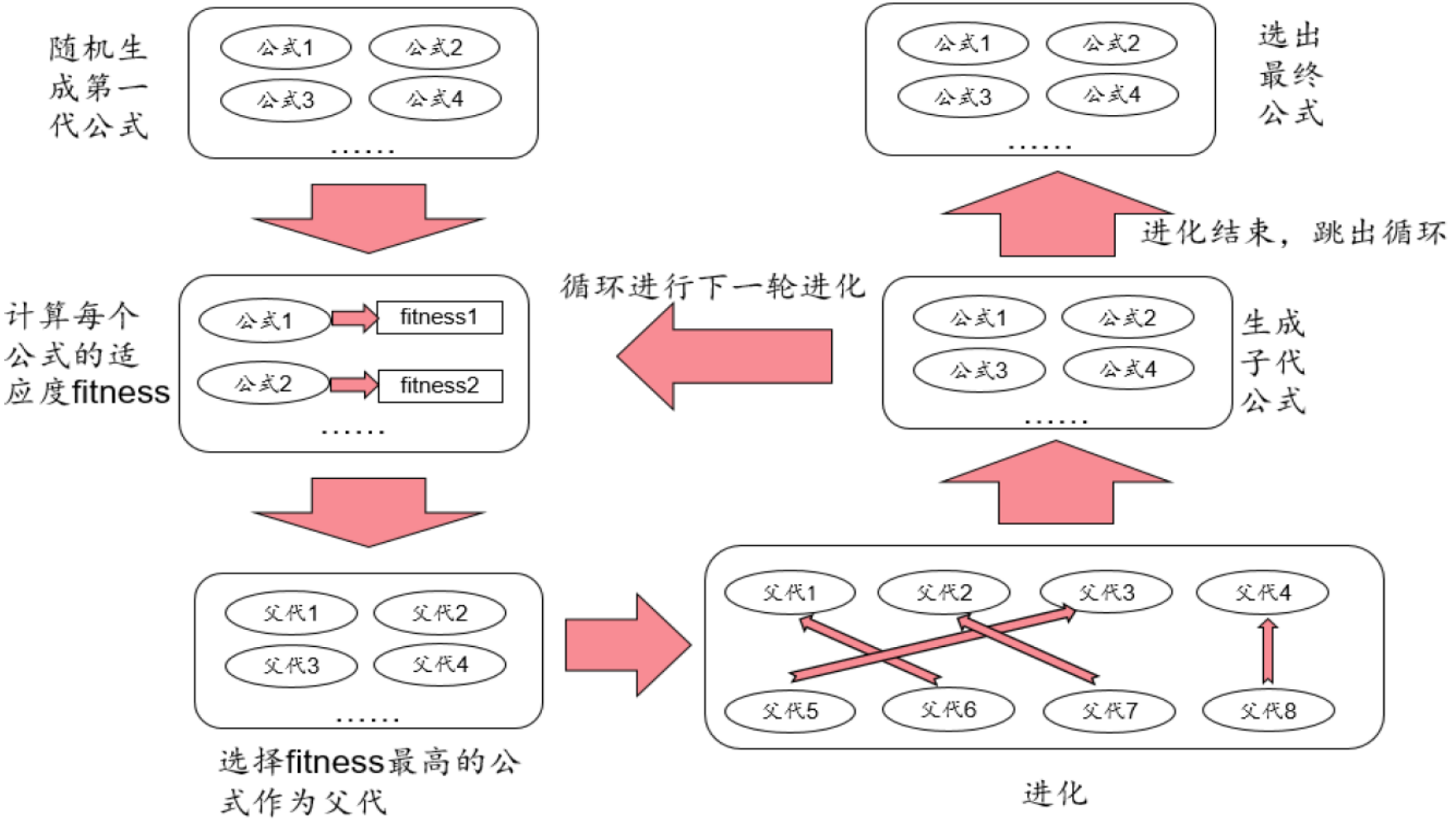




基于遗传规划的因子挖掘

遗传规划

遗传规划总体流程：



基于遗传规划的因子挖掘

遗传规划

公式/树的表示方式：

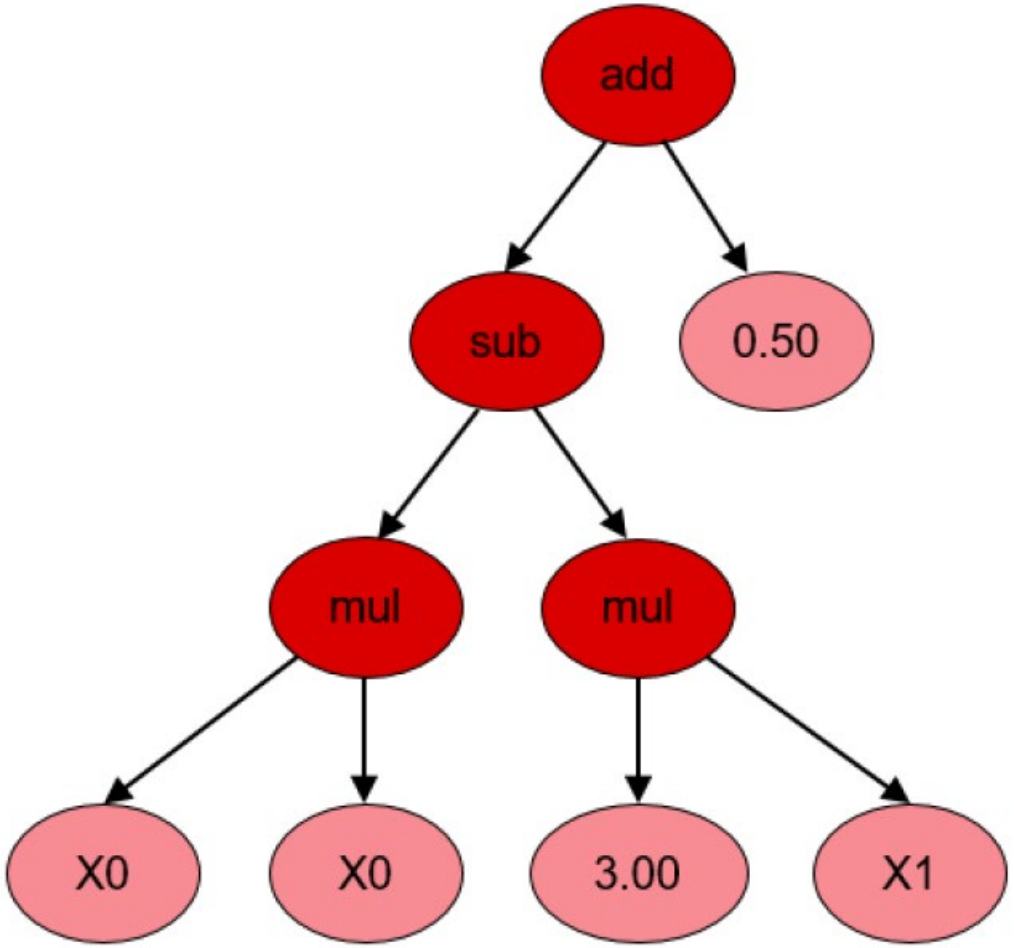
- 遗传规划中公式被表示为二叉树的形式，假若有特征X0和X1，预测目标y。则一个可能的公式为：

$$y = X_0^2 - 3 * X_1 + 0.5$$

- 在遗传规划中将改为S-表达式（S-expression）：

$$y = (+(-(\times X_0 X_0)(*3X_1))0.5)$$

- 二叉树：
  - **深色节点：运算符号**
  - **浅色节点（叶子）：变量OR常数**
- 属性：
  - **depth：根节点到叶节点的最长路径的长度**
  - **length：树中包含的总节点数**

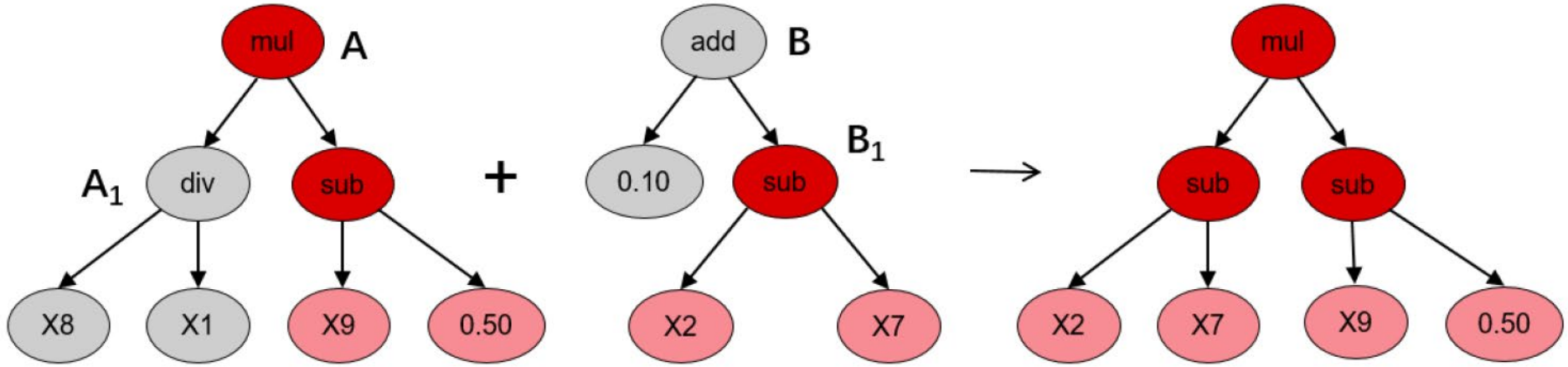


基于遗传规划的因子挖掘

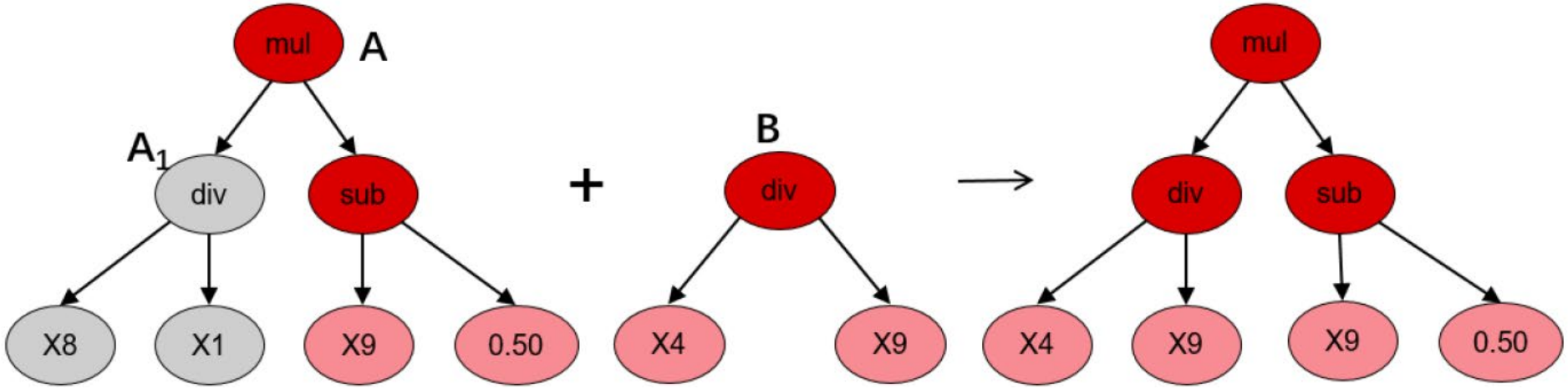
遗传规划

进化方法

- **交叉：**



- **子树变异：**

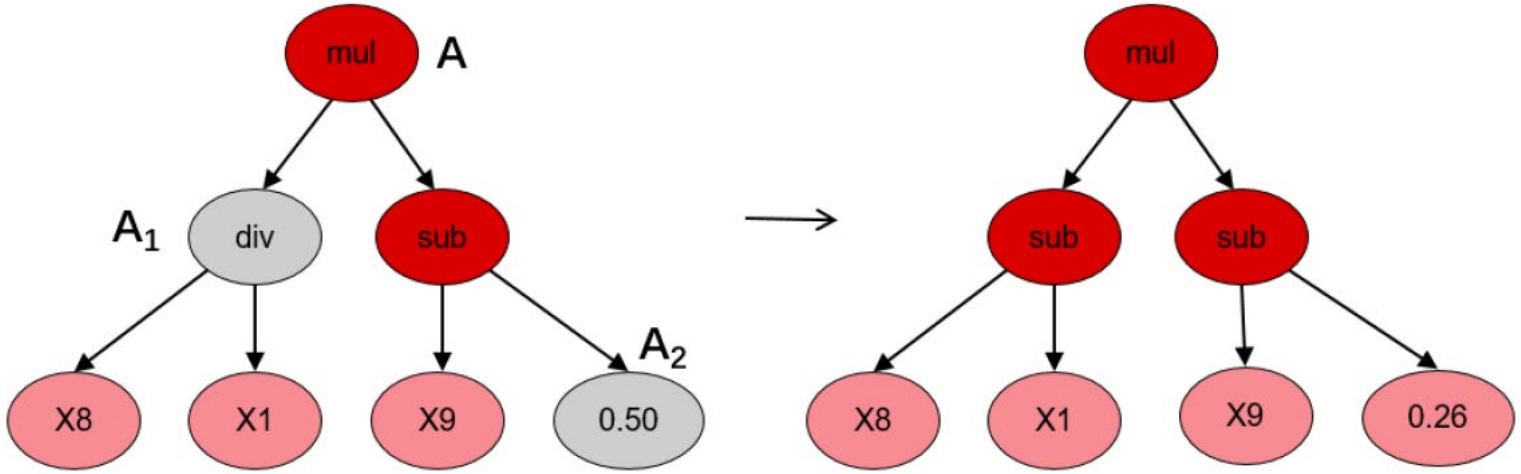


基于遗传规划的因子挖掘

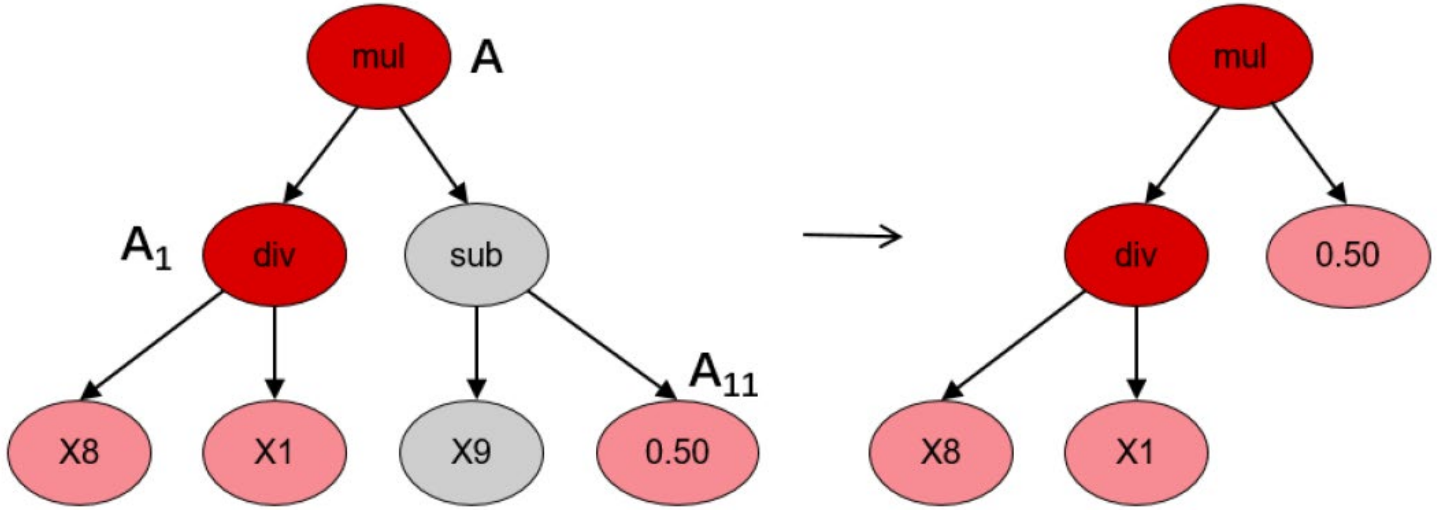
遗传规划

进化方法

- **点变异：**



- **Hoist变异：**



基于遗传规划的因子挖掘

gplearn

模型分类：

- **Symbolic Regressor**
  - 专注于回归，用于拟合数据和生成符号式模型
  - Fitness适应度：
    - MAE(default)
    - MSE
    - RMSE
    - Pearson
    - Spearman
- **Symbolic Transformer**
  - 可一次性得到多个因子，用于挖掘因子
  - Fitness适应度：
    - Pearson(default)
    - Spearman
- **Symbolic Classifier**
  - 用于分类任务，构建符号式分类模型，以进行分类预测。

基于遗传规划的因子挖掘

gplearn

主要参数：

generations	公式进化的世代数量。
population_size	每一代公式群体中的公式数量。
n_components	最终筛选出的最优公式数量。
hall_of_fame	选定最后的 n_components 个公式前，提前筛选出的备选公式的数量，n_components < hall_of_fame < population_size。
function_set	用于构建和进化公式时使用的函数集。
parsimony_coefficient	节俭系数，用于惩罚过于复杂的公式。
tournament_size	每一代的所有公式中，tournament_size 个公式会被随机选中，其中适应度最高的公式能进行变异或繁殖生成下一代公式。
random_state	随机数种子。
init_depth	公式树的初始化深度，init_depth 是一个二元组(min_depth, max_depth)，树的初始深度将处在[min_depth, max_depth]区间内。
metric	适应度指标。
const_range	公式中常数的取值范围，默认为(-1,1)，如果设置为 None，则公式中不会有常数。
p_crossover	交叉变异概率，即父代进行交叉变异进化的概率。
p_subtree_mutation	子树变异概率，即父代进行子树变异进化的概率。
p_hoist_mutation	Hoist 变异概率，即父代进行 Hoist 变异进化的概率。
p_point_mutation	点变异概率，即父代进行点变异进化的概率。
p_point_replace	点替代概率，即点变异中父代每个节点进行变异进化的概率。

基于遗传规划的因子挖掘

代码实证

- gplearn改进：扩充了gplearn 的函数集(function_set)，提供了更多特征计算方法，以提升其因子挖掘能力。用上了gplearn 提供的所有基础计算函数(加、减、乘、除、开方、取对数、绝对值等)，还自定义了一些计算函数用于处理时间序列数据，以捕捉因子在时间维度上的可能特征。

- 拟合预测：使用Symbolic Regressor对代码为3股票进行滚动窗口预测实验，并汇报其MSFE。

- 因子挖掘：将数据划分为训练集和测试集。先对训练集单独使用Ridge回归，并在测试集汇报其MSFE；接着使用SymbolicTransformer进行因子挖掘，将挖掘到的因子和原有初始因子合并，重新进行Ridge回归，汇报MSFE，与单独进行Ridge回归对比。结果MSFE得到改进。汇报最终得到的全部因子。

- 纯样本内应用：将全部数据进行Ridge回归拟合，汇报Adjusted R-squared；再对进行过因子挖掘的数据进行Ridge回归拟合，汇报Adjusted R-squared。发现Adjusted R-squared得到显著提升。

1917—2017
100th Anniversary
Shanghai University of Finance and Economics
上海财经大学 100周年校庆

谢 谢!

Thank You