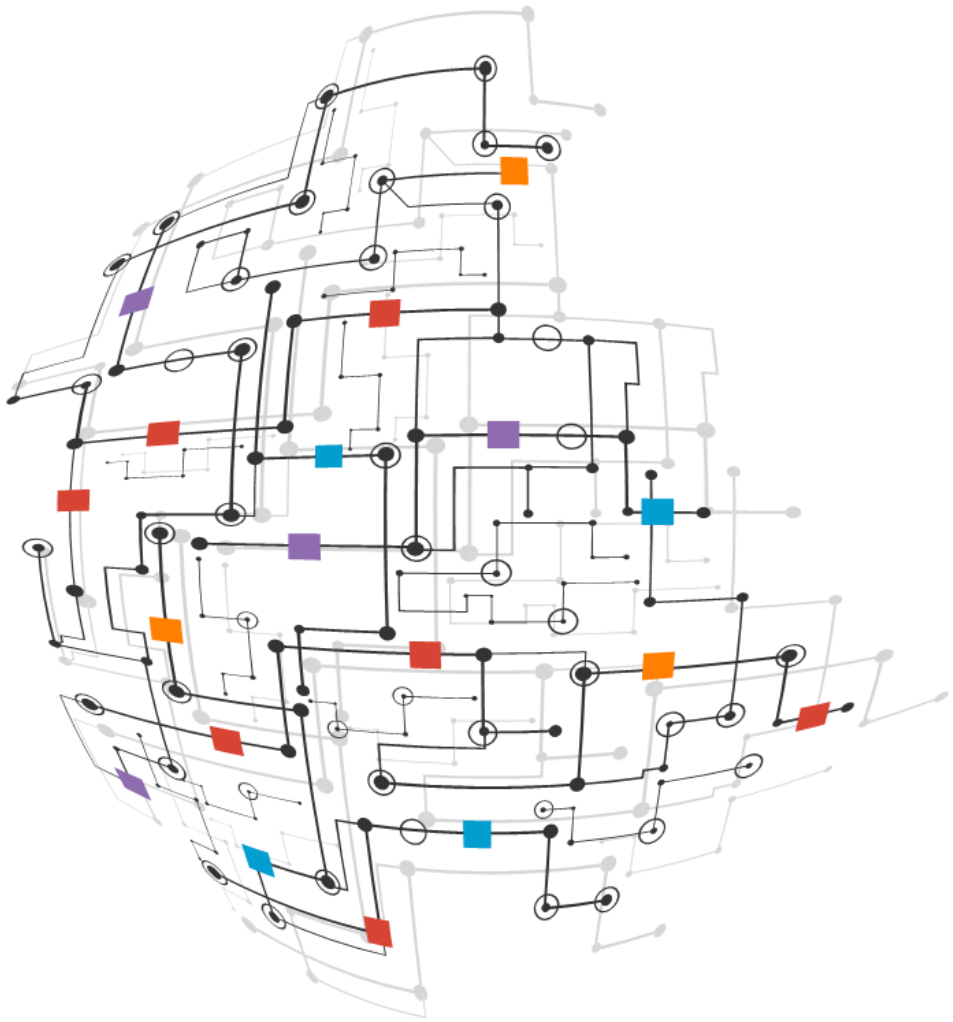
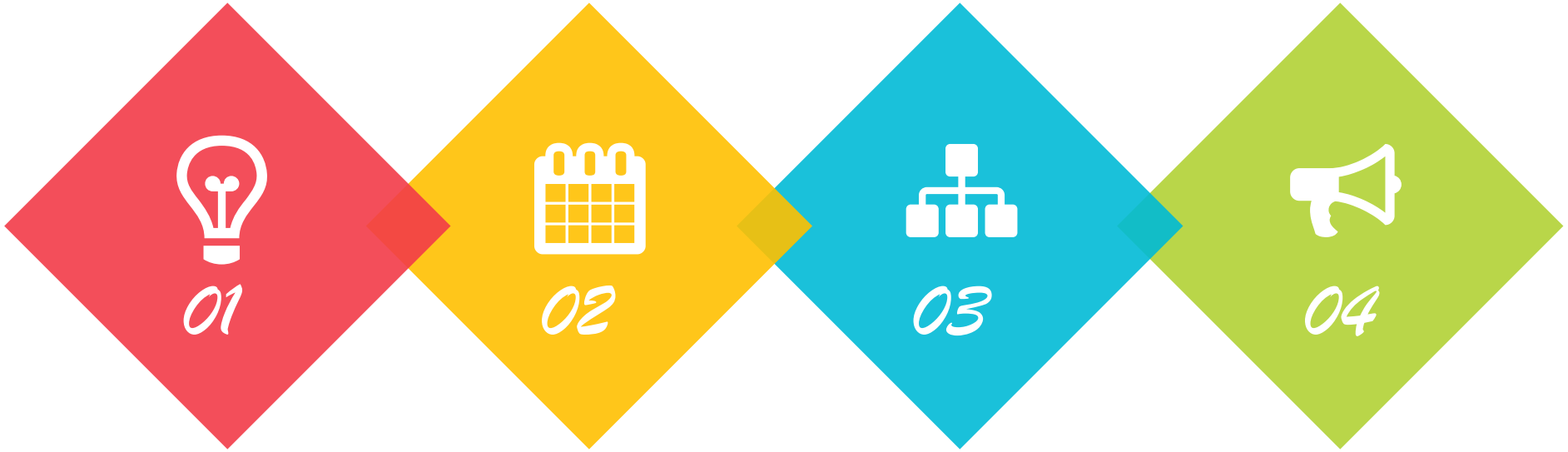


遗传算法

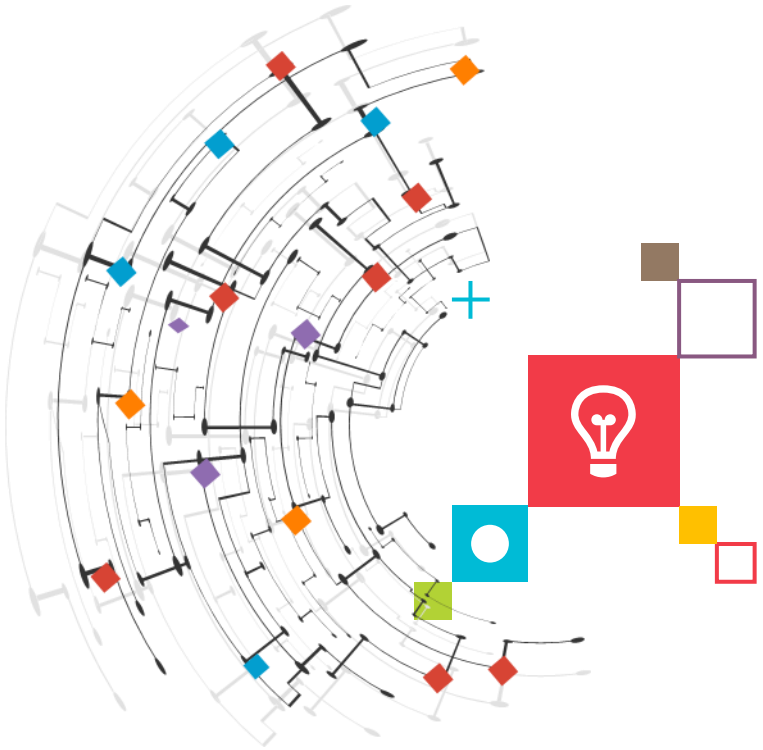
——TSP及符号回归的python实现

学号：31801341　　姓名：童峻涛

目录
CONTENT

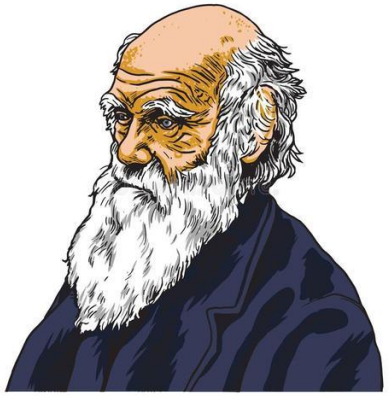


遗传算法简介　所选问题概述　具体实现分析　参考资料展示



01. 遗传算法简介

01. 算法简介



1859年，**达尔文**在《物种起源》中提出生物进化论学说，主要论证阐述了以下两个观点：

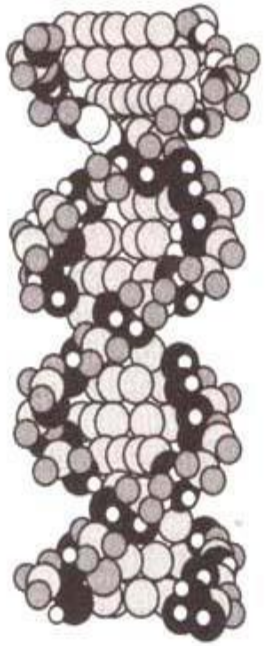
1.物种是可变的，生物是进化的。

2.自然选择是生物进化的动力。

一个多世纪后，20世纪70年代美国**霍兰德**教授认为既然“物竞天择”存在于有机体中，那么也可能存在计算机程序中。

于是在漫长研究后提出遗传算法，霍兰德也被誉为“遗传算法之父”。

01. 算法简介



在遗传算法中，染色体对应的是数据或数组
每个个体拥有属于自己的基因（individual）
一定数量的个体就组成了群体（population）
群体中的个体数目称为群体大小（population size）
各个个体对环境的适应程度叫适应度（fitness）

（根据我的认知）
遗传算法大致分位下列**6个步骤**:
（1）编码
（2）初始群体的生成
（3）适应度评估
（4）选择
（5）交叉
（6）变异

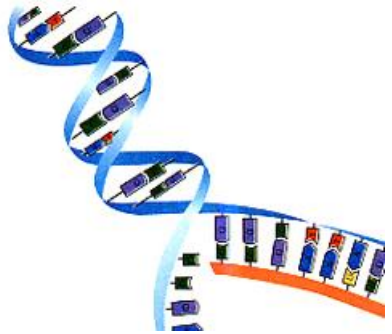
01. 算法简介

（1）编码

编码是应用遗传算法时要解决的首要问题，也是设计遗传算法时的一个关键步骤。

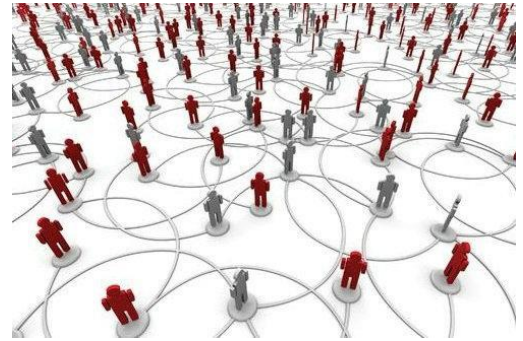
编码方法影响到交叉算子、变异算子等遗传算子的运算方法，大很大程度上决定了遗传进化的效率。

主要由：二进制编码、浮点数编码、符号编码等。



（2）初始群体的生成

随机产生数据个体，由个体组成群体。



（3）适应度评估

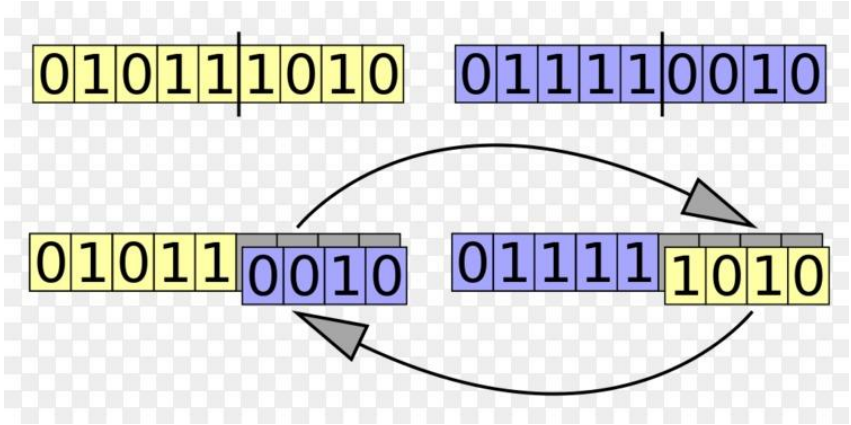
适应度表明个体或解的优劣性。

对于不同问题适应度函数的定义也不尽相同。

01. 算法简介

（4）选择

根据得到的个体适应度，从当前群体中选出优良个体，使他们有机会作为父代为下一代繁衍子孙。

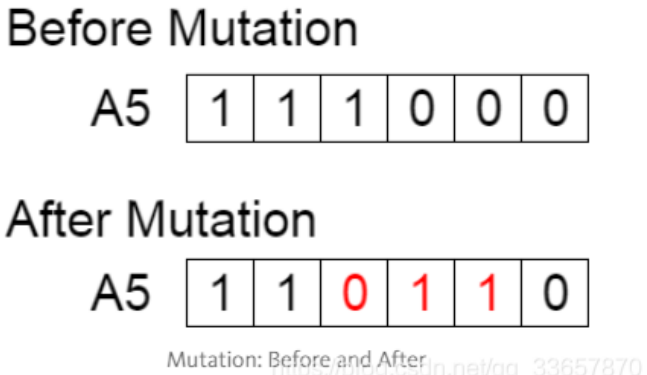


（5）交叉

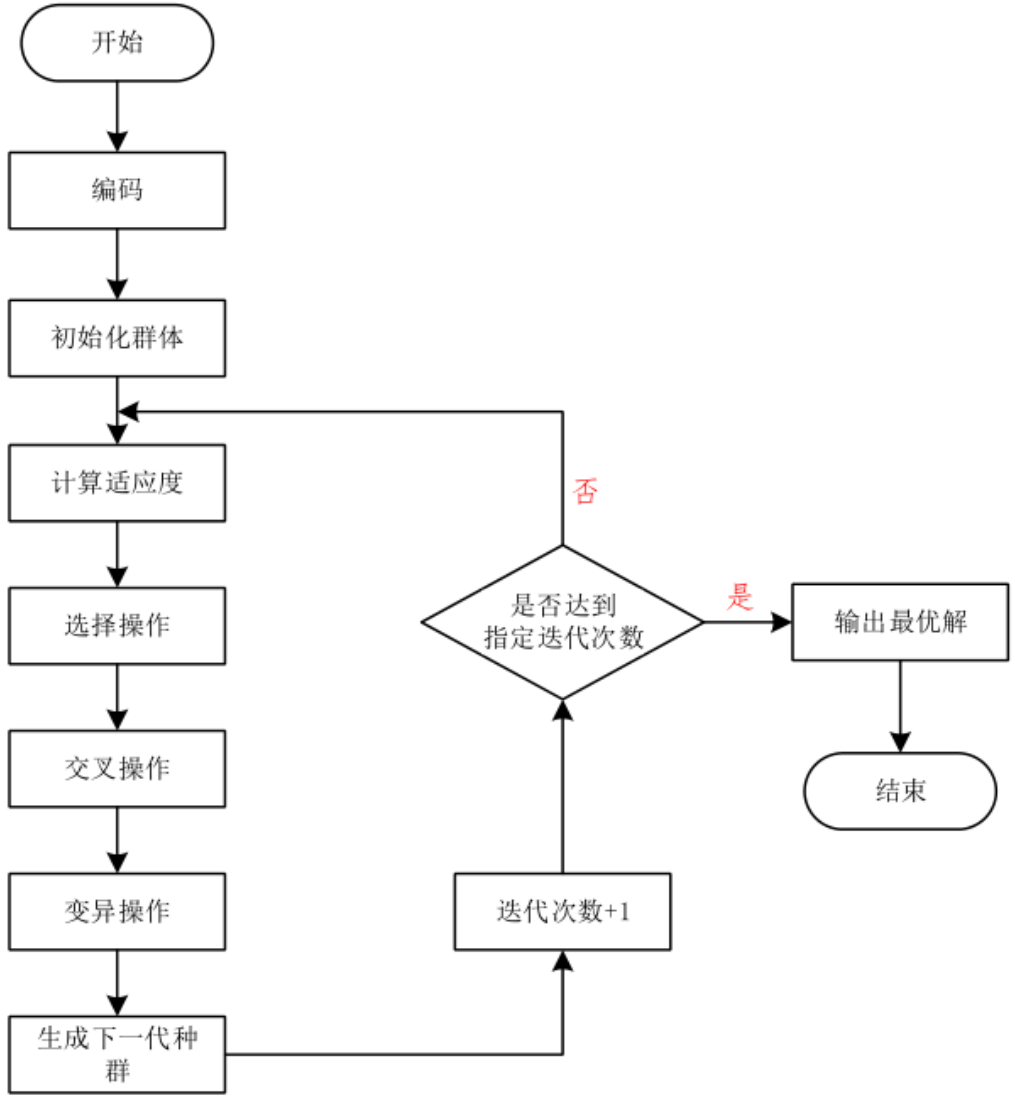
交叉操作是遗传算法中最主要的遗传操作。通过交叉操作可以得到新一代个体，新个体组合了其父辈个体的特性，体现了信息交换的思想。

（6）变异

变异是在群体中随机选择一个个体，对于选中的个体以一定的概率随机地改变其编码中某个值。发生变异的概率通常很低所以取值一般很小。



01. 算法简介



算法流程图

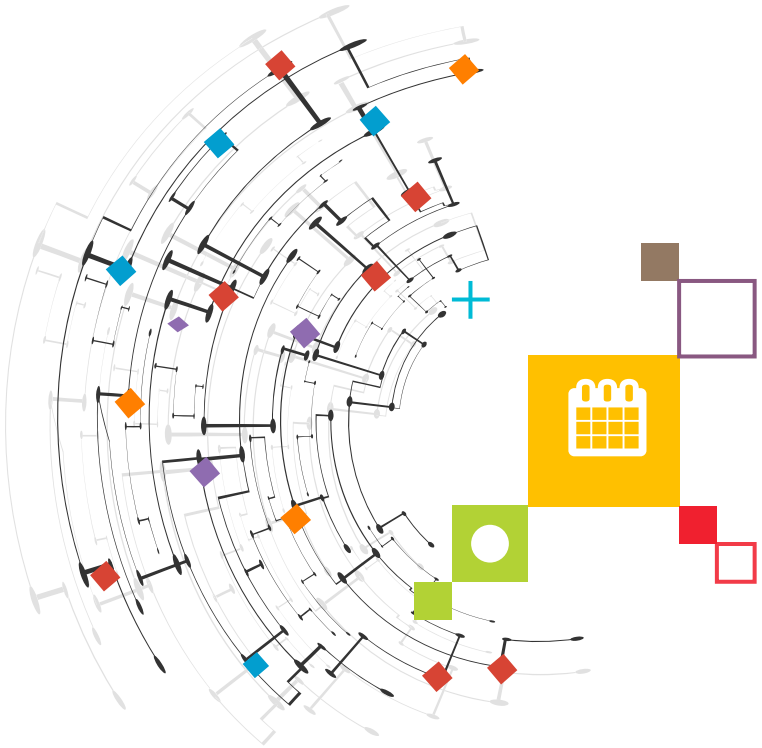
01. 算法简介

遗传算法的特点

（1）覆盖面大，减少了误入局部最优解的风险，利于全局择优。

（2）使用适应度函数值来评价个体，不受连续可微的约束，并且可以任意设定

（3）具有自组织、自适应和自学习性。



02. 所选问题概述

02. 问题概述

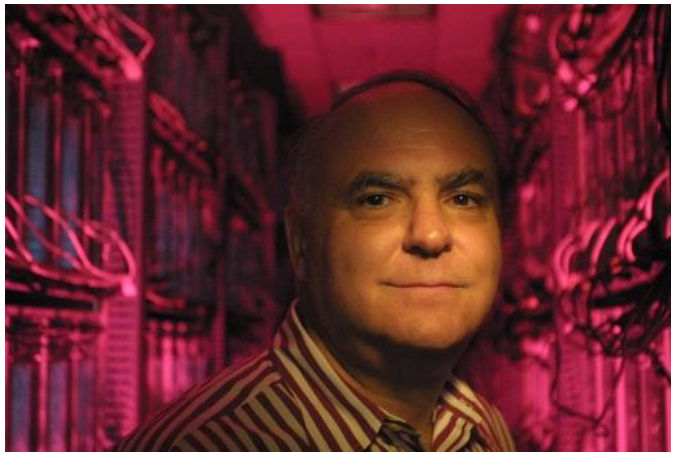
1. 旅行商问题概述

旅行商问题(Traveling Saleman Problem，TSP)是指一个旅行商需要到n个城市去推销商品，要求从某一城市出发，经过n-1个城市，旅行商在途径的n-1个城市能且仅能经过一次，使得旅行商的路程最短。

与传统TSP不太一样，我在这里进行了简化，从权值最小的Hamilton回路变成了随机点的最短路径，这里从随机的一点后出发一点结束，同时保证每个城市都经过。

02. 问题概述

2. 基于遗传算法的符号回归问题概述



在这个问题中论文选自1994年在《Statistics and Computing》中John Koza所著的《Genetic programming as a means for programming computers by natural selection》一文，其主要思想是通过基于遗传算法的符号回归来由计算机得到一个适合的计算公式，来对相关数据进行预测。

在解决过程中将会用到gplearn库，是目前Python内最成熟的符号回归算法实现。

Link：https://github.com/trevorstephens/gplearn

Genetic Programming in Python,
with a scikit-learn inspired API:
gplearn

02. 问题概述

2. 基于遗传算法的符号回归问题概述

根据该领域的先验知识，对于金额特征、日期特征进行比值操作生成的特征通常能够提升验证集和测试集合的分数，所以论文中对1959年至1988年的美国经济就行了回归预测。

复现则是采用1989年至2018年的美国经济作为数据集。

Link：https://fred.stlouisfed.org/

FRED® ECONOMIC DATA | ST. LOUIS FED

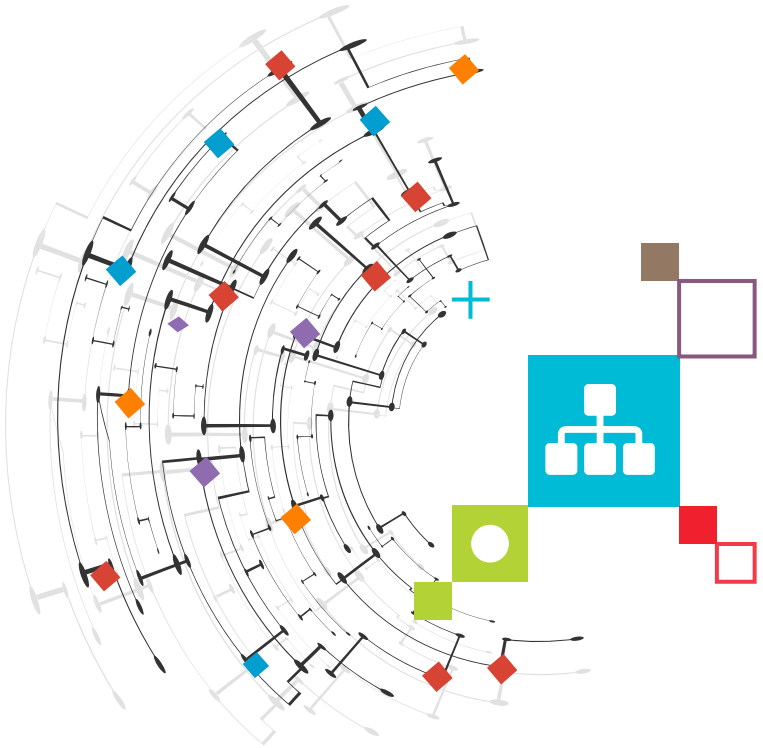
相关经济术语描述：

GNP：国民生产总值 = 国内生产总值 + 来自国外的净要素收入，反映了现代产业结构的变化。

GNPDEF：国民生产总值平减指数 = 报告期价格计算的国民生产总值 / 基准年的国民生产总值，反映价值指标增减过程中与物量变动同时存在的价格变动趋势和程度的价格指数。

M2：广义货币供应量 = 流通的现金（M0）+ 企业活期存款（M1）+ 准货币（定期存款、居民储蓄存款、其他存款），通常反映的是社会总需求变化和未来通胀的压力状态。

3MTB：三个月国库券，国库券的利率是市场利率变动情况的集中反映。



03. 具体实现分析

03. 实现分析

1. 旅行商问题实现

（1）DNA编码

每一个城市有一个ID,那经历的城市顺序就是按ID排序。
比如说商人要经过2个城市,我们就有以下两种方式:

a. 0-1
b. 1-0

（2）适应度函数的确定

根据路径的长度判断，总路径越短越优秀。

```
fitness = 1/total_distance
```

使用exp函数将数值扩大，以至于在轮盘赌策略时，尽可能多停留在优质基因区间。

```
fitness = np.exp(self.DNA_size * 2 / total_distance)
```

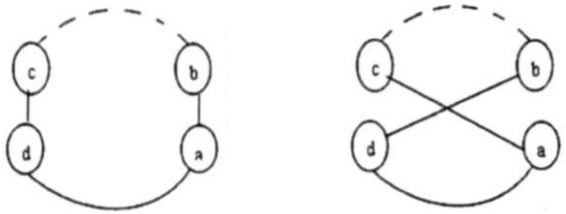
03. 实现分析

1. 旅行商问题实现

（3）交叉和变异

先选择来自对应位置父序列的城市序号，排列在子序列的前面位置，后将未访问的城市序号按照其在母序列中的顺序填入子序列中得到完整的子序列。这样就能避免存在未访问城市的问题。

```
1. p1=[0,1,2,3] #baba
2. cp=[ ,b, ,b] #from baba
3. c1=[1,3, , ] #to be continued
4. p2=[3,2,1,0] #mama
5. cp=[m, ,m, ] #from mama
6. c1=[1,3,2,0] #child
```



在变异阶段，采取的措施是随机将两个城市序号交换，以防止产生重复或者不存在的编号。

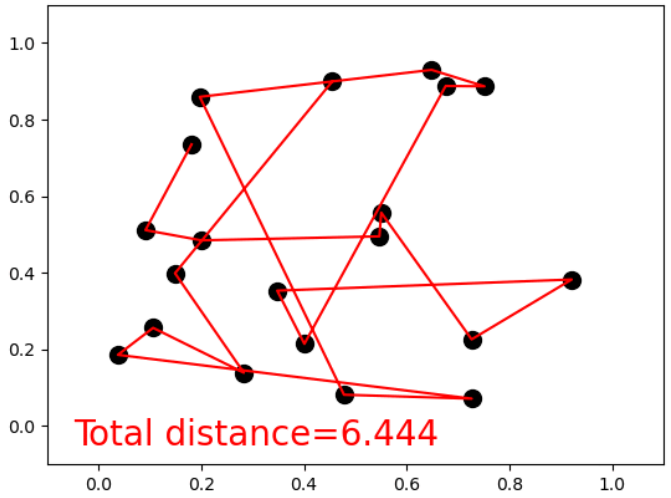
03. 实现分析

1. 旅行商问题实现

（4）结果分析

程序最开始生成不同城市，通过不断地交叉变异，选择不同的路径搭配，来搜索最短的距离。

每次均在100-200代之间趋于稳定，最终结果并一定是最优的，是短时间内大概最优的一个选择，可以通过调节代数、交叉率、变异率等进一步得到更优结果。

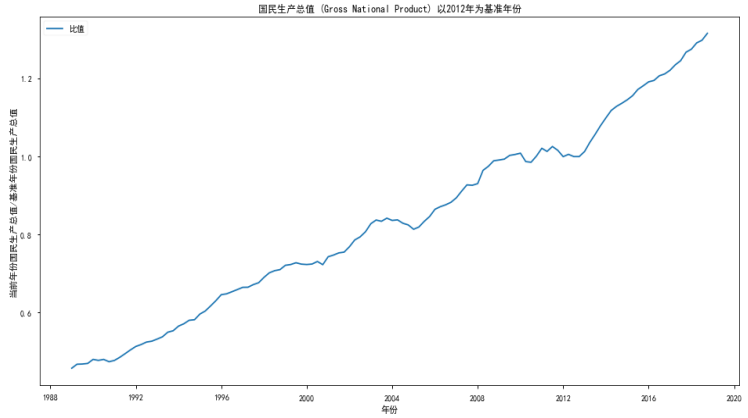



```
Gen: 101 | best fit: 10783.81
Gen: 102 | best fit: 11174.58
Gen: 103 | best fit: 11174.58
Gen: 104 | best fit: 11174.58
Gen: 105 | best fit: 13679.56
Gen: 106 | best fit: 11882.84
Gen: 107 | best fit: 11882.84
Gen: 108 | best fit: 11882.84
Gen: 109 | best fit: 11882.84
Gen: 110 | best fit: 11882.84
Gen: 111 | best fit: 11882.84
Gen: 112 | best fit: 13679.56
Gen: 113 | best fit: 13679.56
Gen: 114 | best fit: 13679.56
Gen: 115 | best fit: 13679.56
Gen: 116 | best fit: 13679.56
Gen: 117 | best fit: 13679.56
```

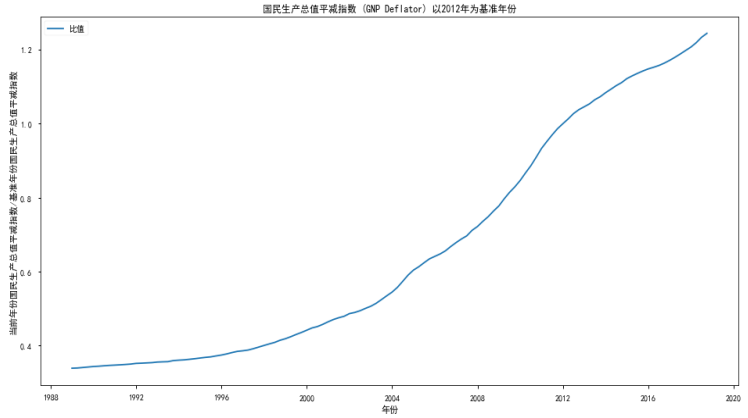
03. 实现分析

2. 基于遗传算法的符号回归问题实现

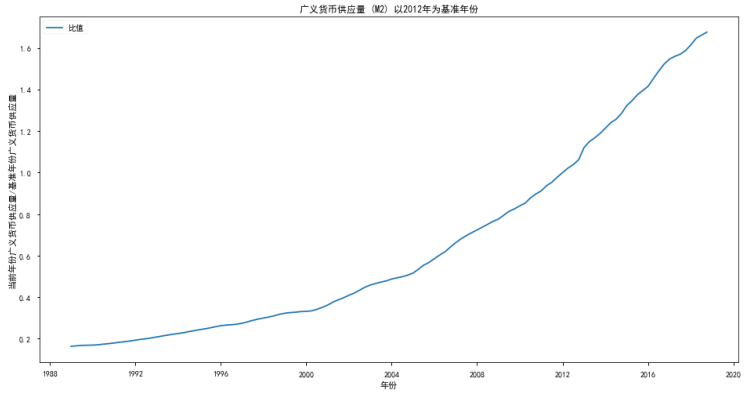
（1）国民生产总值数据图



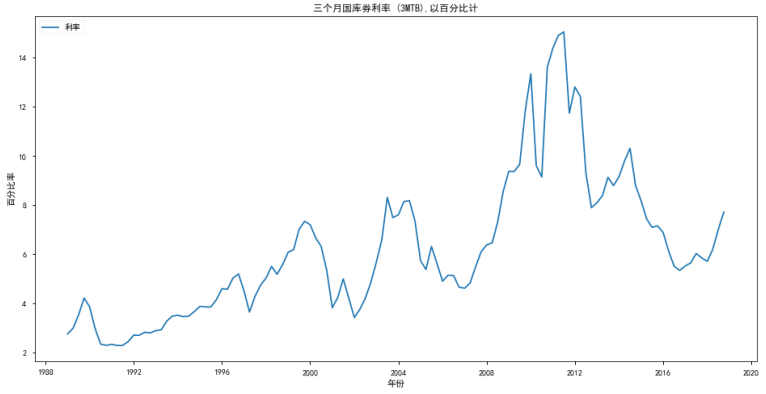
（2）国民生产总值平减指数数据图



（3）广义货币供应量数据图

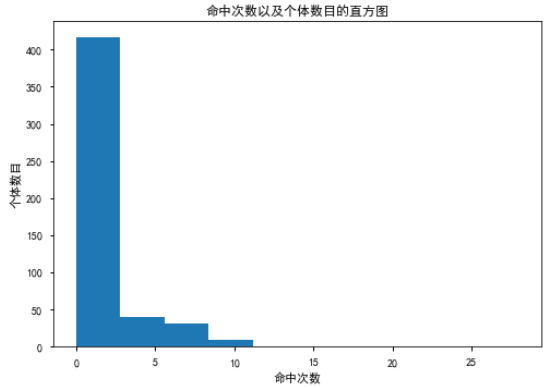


（4）三个月国库券利率数据图

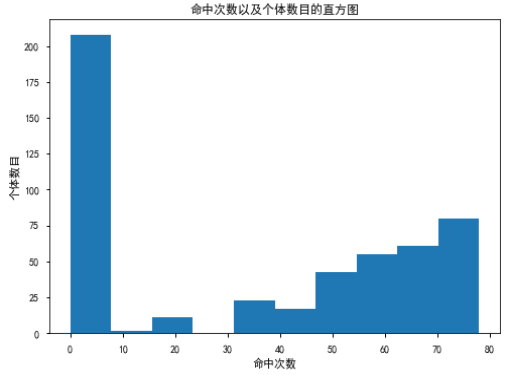


03. 实现分析

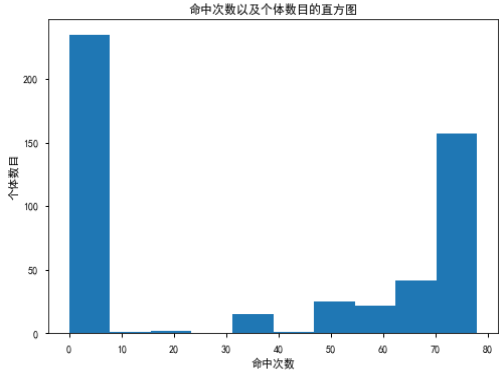
2. 基于遗传算法的符号回归问题实现



Gen=1



Gen=10



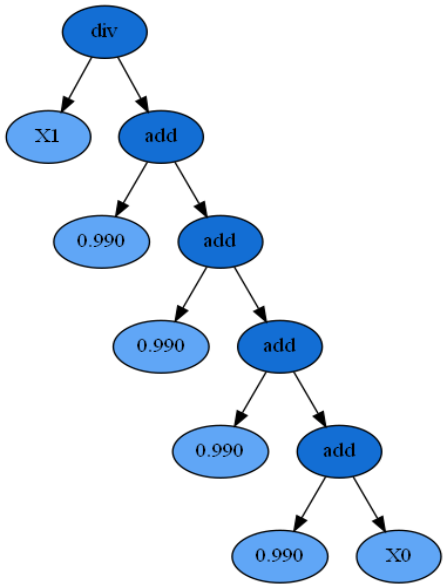
Gen=20

随着迭代次数的增加，高命中次数（适应度高）个体逐渐增多，种群分布随着个体的变化发生大幅度的改变

03. 实现分析

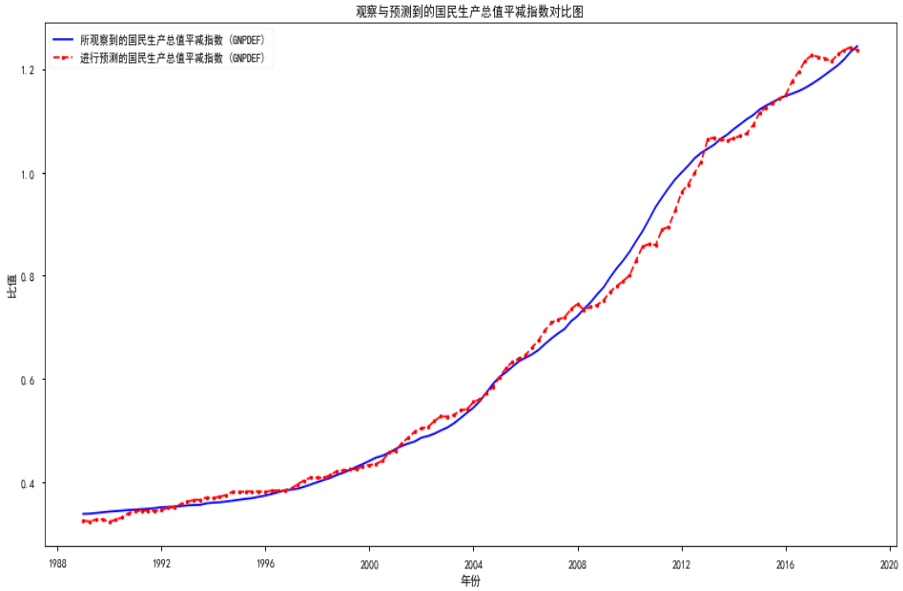
2. 基于遗传算法的符号回归问题实现

（5）二叉树表达式图



根据表达式进行预测

（7）观察以及预测模型对比图



注：红线为预测，蓝线为观察

03. 实现分析

2. 基于遗传算法的符号回归问题实现

（7）预测模型得分及平方和误差

A．实现结果

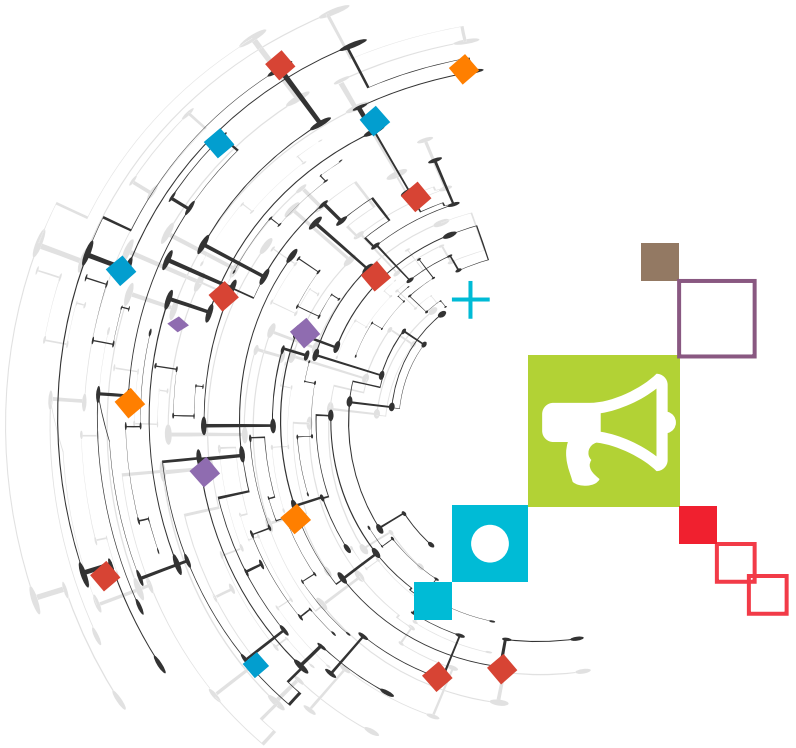
数据范围	1—120	1—80	81—120
预测得分	0.99447	0.98916	0.92959
平方和误差	0.06143	0.09123	0.15823

B．论文结果

数据范围	1—120	1—80	81—120
预测得分	0.99348	0.99795	0.99061
平方和误差	0.07539	0.00927	0.06611

C．结论

虽然在总体的训练中精度略高于论文，但是在局部的训练中低于论文，还需要继续学习与探索。



04. 参考资料展示

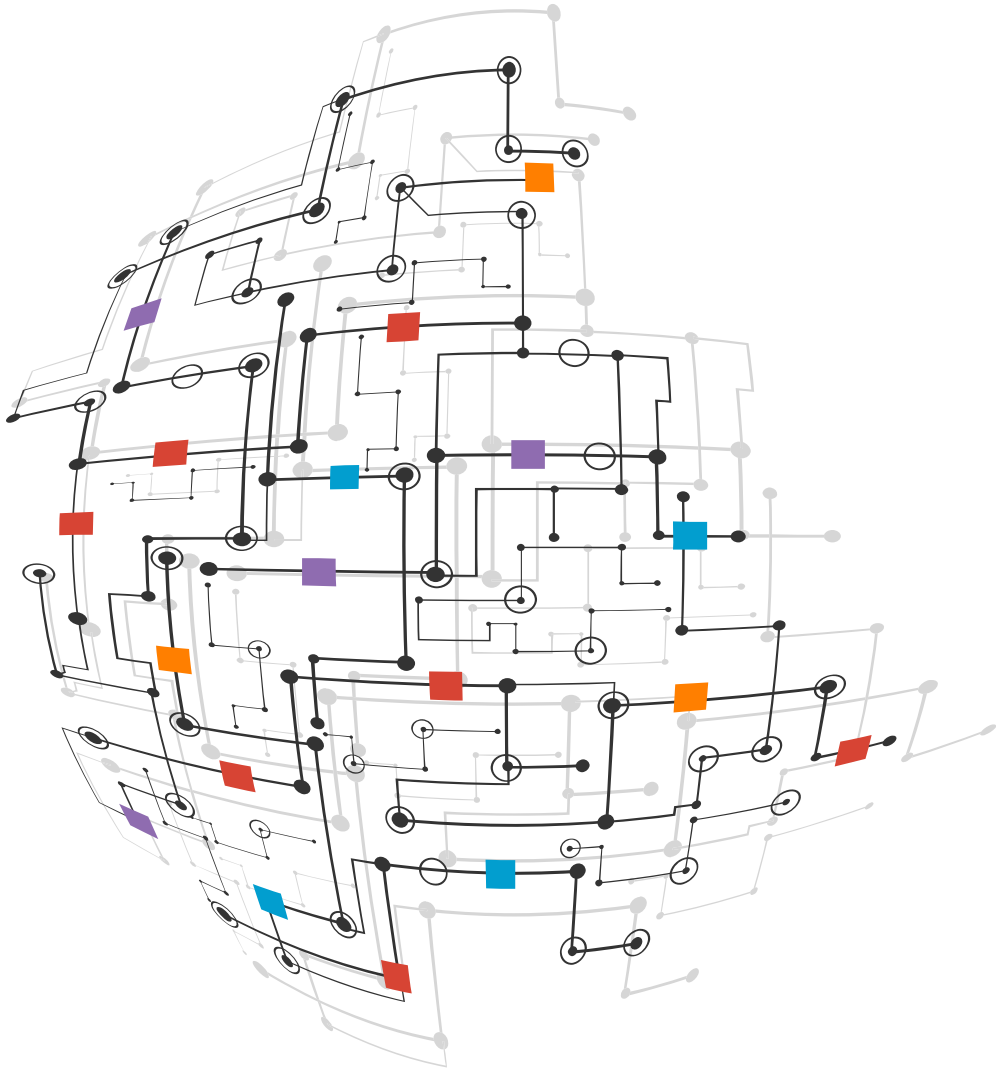
04. 参考资料

论文：

[1] [CN]遗传算法中适应度函数的研究
[2] [CN]用遗传算法求解旅行商问题及其代码设计
[3] [EN]基于标签传播的局部搜索遗传算法检测动态社区
[4] [EN]遗传编程是一种对计算机编程的自然选择的方法
[5] [CN]基于遗传算法的移动机器人路径规划研究_崔建军
[6] [CN]用遗传算法求解TSP问题_任昊南

网页：

[1] 基于遗传编程(Genetic Programming)的符号回归(Symbolic Regression)简介 [EB/OL]. https://blog.csdn.net/likehightime/article/details/5275264 -2021/01/02
[2] 遗传编程的示例（二次多项式的符号回归） [EB/OL]. http://www.genetic-programming.com/gpquadraticexample.html -2021/01/02
[3] 遗传算法求解旅行商问题 [EB/OL]. https://zhuanlan.zhihu.com/p/137351343 -2021/01/02
[4] 百度百科----遗传算法 [EB/OL]. https://baike.baidu.com/item/遗传算法 -2021/01/02
[5] 遗传算法 (Genetic Algorithm) [EB/OL].https://mofanpy.com/tutorials/machine-learning/evolutionary-algorithm/intro-genetic-algorithm/ -2021/01/02
[6] 遗传算法 (Genetic Algorithm) 例子 旅行商人问题 (TSP) [EB/OL].https://mofanpy.com/tutorials/machine-learning/evolutionary-algorithm/genetic-algorithm-travel-sales-problem/ -2021/01/02
[7] 目标优化智能算法之遗传算法 [EB/OL]. https://tianle.me/2017/04/19/GA/ -2021/01/02
[8] 人工智能8—遗传算法实验 [EB/OL]. https://blog.csdn.net/qq_43653930/article/details/103142930 -2021/01/02
[9] 经济机器是怎样运行的 [EB/OL]. https://www.youtube.com/watch?v=rFV7wdEX-Mo&feature=emb_rel_end -2021/01/02
[10] 【10分钟算法】遗传算法-带例子和动画/Genetic Algorithm [EB/OL]. https://www.bilibili.com/video/BV1yt4y1a7RY?from=search&seid=3179391201366450659 -2021/01/02
[11] 随机森林算法OOB_SCORE最佳特征选择 [EB/OL]. https://www.cnblogs.com/dinol/p/11614352.html -2021/01/02
[12] 利用 gplearn 进行特征工程 [EB/OL]. https://bigquant.com/community/t/topic/120709 -2021/01/02
[13] 用遗传算法实现符号回归——浅析gplearn [EB/OL]. https://zhuanlan.zhihu.com/p/31185882 -2021/01/02
[14] 使用gplearn自定义特征自动生成模块 [EB/OL]. https://zhuanlan.zhihu.com/p/76047703?from_voters_page=true -2021/01/02
[15] 【算法】超详细的遗传算法(Genetic Algorithm)解析 [EB/OL]. https://www.jianshu.com/p/ae5157c26af9 -2021/01/02
[16] 中国每年国民生产总值和货币供应量的数据 [EB/OL]. http://www.360doc.com/content/19/1108/18/34989057_871927918.shtml -2021/01/02
[17] Graphviz的安装及纠错 [EB/OL]. https://blog.csdn.net/qq_28409193/article/details/79880886 -2021/01/02
[18] STATISTICS AND COMPUTING [EB/OL]. https://www.shengsci.com/sci/6408.html -2021/01/02
[19] 优化算法系列-遗传算法（1）——基本理论枯燥版本 [EB/OL]. https://www.cnblogs.com/haimishasha/p/9816735.html -2021/01/02
[20] TSP问题—启发式遗传算法 [EB/OL]. https://www.jianshu.com/p/b3cd8e674ff0 -2021/01/02



感谢观看

学号：31801341　　姓名：童峻涛