遗传算法详解

作者：https://blog.csdn.net/u010451580/article/details/51178225

本文是去年课题组周报中的一个专题讲解，详细讲了 GA，由于是周报，所以十分详细。很适合初学者入门。文中也简单提及了模拟退火算法。文章综合参考了一些互联网资料。发博客以备忘！

三：遗传算法

照例先给出科学定义：

遗传算法（Genetic Algorithm, GA）起源于对生物系统所进行的计算机模拟研究。它是模仿自然界生物进化机制发展起来的随机全局搜索和优化方法，借鉴了达尔文的进化论和孟德尔的遗传学说。其本质是一种高效、并行、全局搜索的方法，能在搜索过程中自动获取和积累有关搜索空间的知识，并自适应地控制搜索过程以求得最佳解。

再给出相关术语：（各位看看就好，后面都会涉及到，再细说）

基因型(genotype)：性状染色体的内部表现；

表现型(phenotype)：染色体决定的性状的外部表现，或者说，根据基因型形成的个体的外部表现；

进化(evolution)：种群逐渐适应生存环境，品质不断得到改良。生物的进化是以种群的形式进行的。

适应度(fitness)：度量某个物种对于生存环境的适应程度。

选择(selection)：**以一定的概率**从种群中选择若干个个体。一般，选择过程是一种**基于适应度**的优胜劣汰的过程。

复制(reproduction)：细胞分裂时，遗传物质 DNA 通过复制而转移到新产生的细胞中，新细胞就继承了旧细胞的基因。

交叉(crossover)：两个染色体的某一相同位置处 DNA 被切断，前后两串分别交叉组合形成两个新的染色体。也称基因重组或杂交；

变异(mutation)：复制时可能（很小的概率）产生某些复制差错，变异产生新的染色体，表现出新的性状。

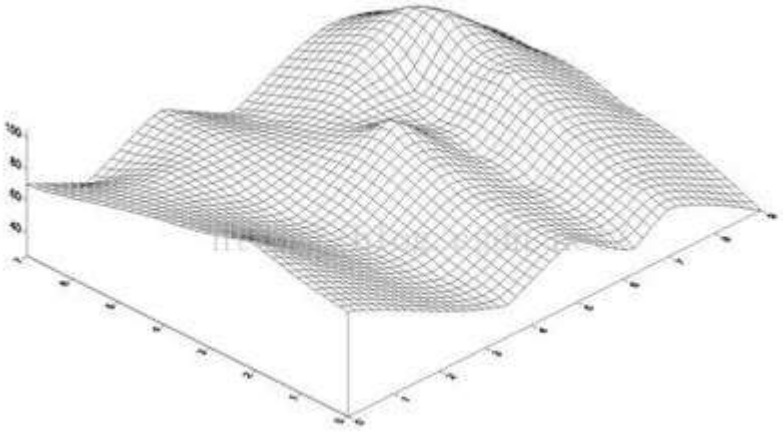
编码(coding)：DNA 中遗传信息在一个长链上按一定的模式排列。遗传编码可看作从表现型到基因型的映射。

解码(decoding)：基因型到表现型的映射。

个体（individual）：指染色体带有特征的实体；
种群（population）：个体的集合，该集合内个体数称为种群的大小。

遗传算法的有趣应用很多，诸如寻路问题，8 数码问题，囚犯困境，动作控制，找圆心问题（在一个不规则的多边形中，寻找一个包含在该多边形内的最大圆圈的圆心），TSP 问题，生产调度问题，人工生命模拟等。下面我以袋鼠为例子讲讲遗传算法。（因为袋鼠会跳）

遗传算法中每一条染色体，对应着遗传算法的一个解决方案，一般我们用适应性函数（fitness function）来衡量这个解决方案的优劣。所以从一个基因组到其解的适应度形成一个映射。可以把遗传算法的过程看作是一个在多元函数里面求最优解的过程。**可以这样想象，这个多维曲面里面有数不清的“山峰”，而这些山峰所对应的就是局部最优解。而其中也会有一个“山峰”的海拔最高的，那么这个就是全局最优解。而遗传算法的任务就是尽量爬到最高峰，而不是陷落在一些小山峰。**（另外，值得注意的是遗传算法不一定要找“最高的山峰”，如果问题的适应度评价越小越好的话，那么全局最优解就是函数的最小值，对应的，遗传算法所要找的就是“最深的谷底”）



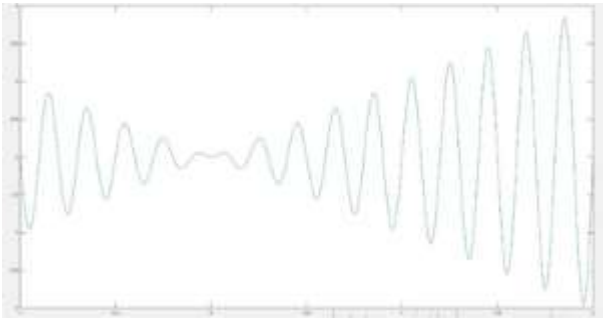
问题的提出与解决方案：

让我们先来考虑考虑下面这个问题的解决办法。

已知一元函数：

$$f(x) = x\sin(10\pi x) + 2 \qquad x \in [-1, 2]$$

现在要求在既定的区间内找出函数的最大值



“袋鼠跳”问题

既然我们把函数曲线理解成一个一个山峰和山谷组成的山脉。那么我们可以设想所得到的每一个解就是一只袋鼠，我们希望它们不断的向着更高处跳去，直到跳到最高的山峰（尽管袋鼠本身不见得愿意那么做）。所以求最大值的过程就转化成一个“袋鼠跳”的过程。

作为对比下面简单介绍“袋鼠跳”的几种方式。

1. 爬山法（最速上升爬山法）：

从搜索空间中随机产生邻近的点，从中选择对应解最优的个体，替换原来的个体，不断重复上述过程。因为爬山法只对“邻近”的点作比较，所以目光比较“短浅”，常常只能收敛到离开初始位置比较近的局部最优解上面。对于存在很多局部最优点的问题，通过一个简单的迭代找出全局最优解的机会非常渺茫。（在爬山法中，袋鼠最有希望到达最靠近它出发点的山顶，但不能保证该山顶是珠穆朗玛峰，或者是一个非常高的山峰。因为一路上它只顾上坡，没有下坡。）

2. 模拟退火：

这个方法来自金属热加工过程的启发。在金属热加工过程中，当金属的温度超过它的熔点（Melting Point）时，原子就会激烈地随机运动。与所有的其它的物理系统相类似，原子的这种运动趋向于寻找其能量的极小状态。在这个能量的变迁过程中，开始时，温度非常高， 使得原子具有很高的能量。随着温度不断降低，金属逐渐冷却，金属中的原子的能量就越来越小，最后达到所有可能的最低点。利用模拟退火的时候，让算法从较大的跳跃开始，使到它有足够的“能量”逃离可能“路过”的局部最优解而不至于限制在其中，当它停在全局最优解附近的时候，逐渐的减小跳跃量，以便使其“落脚 ”到全局最优解上。（在模拟退火中，袋鼠喝醉了，而且随机地大跳跃了很长时间。运气好的话，它从一个山峰跳过山谷，到了另外一个更高的山峰上。但最后，它渐渐清醒了并朝着它所在的峰顶跳去。）

3. 遗传算法：

模拟物竞天择的生物进化过程，通过维护一个潜在解的群体执行了多方向的搜索，并支持这些方向上的信息构成和交换。是以面为单位的搜索，比以点为单位的搜索，更能发现全局最优解。（在遗传算法中，有很多袋鼠，它们降落到喜玛拉雅山脉的任意地方。**这些袋鼠并不知道它们的任务是寻找珠穆朗玛峰。**但每过几年，就在一些海拔高度较低的地方射杀一些袋鼠，并希望存活下来的袋鼠是多产的，在它们所处的地方生儿育女。）（或者换个说法。从前，有一大群袋鼠，它们被莫名其妙的零散地遗弃于喜马拉雅山脉。于是只好在那里艰苦的生活。海拔低的地方弥漫着一种无色无味的毒气，海拔越高毒气越稀薄。可是可怜的袋鼠们对此**全然不觉**，还是习惯于活蹦乱跳。于是，不断有袋鼠死于海拔较低的地方，而越是在海拔高的袋鼠越是能活得更久，也越有机会生儿育女。就这样经过许多年，这些袋鼠们竟然都不自觉地聚拢到了一个个的山峰上，可是在所有的袋鼠中，只有聚拢到珠穆朗玛峰的袋鼠被带回了美丽的澳洲。）

遗传算法的实现过程

遗传算法的实现过程实际上就像自然界的进化过程那样。首先寻找一种对问题潜在解进行“数字化”编码的方案。（建立表现型和基因型的映射关系）然后用随机数初始化一个种群（那么第一批袋鼠就被随意地分散在山脉上），种群里面的个体就是这些数字化的编码。接下来，通过适当的解码过程之后（得到袋鼠的位置坐标），用适应性函数对每一个基因个体作一次适应度评估（袋鼠爬得越高，越是受我们的喜爱，所以适应度相应越高）。用选择函数按照某种规定择优选择（我们要每隔一段时间，在山上射杀一些所在海拔较低的袋鼠，以保证袋鼠总体数目持平。）。让个体基因变异（让袋鼠随机地跳一跳）。然后产生子代（希望存活下来的袋鼠是多产的，并在那里生儿育女）。遗传算法并不保证你能获得问题的最优解，但是使用遗传算法的最大优点在于你不必去了解和操心如何去“找”最优解。（你不必去指导袋鼠向那边跳，跳多远。）而只要简单的“否定”一些表现不好的个体就行了。（**把那些总是爱走下坡路的袋鼠射杀，这就是遗传算法的精粹！**）

所以我们总结出遗传算法的一般步骤：

开始循环直至找到满意的解。

1.评估每条染色体所对应个体的适应度。

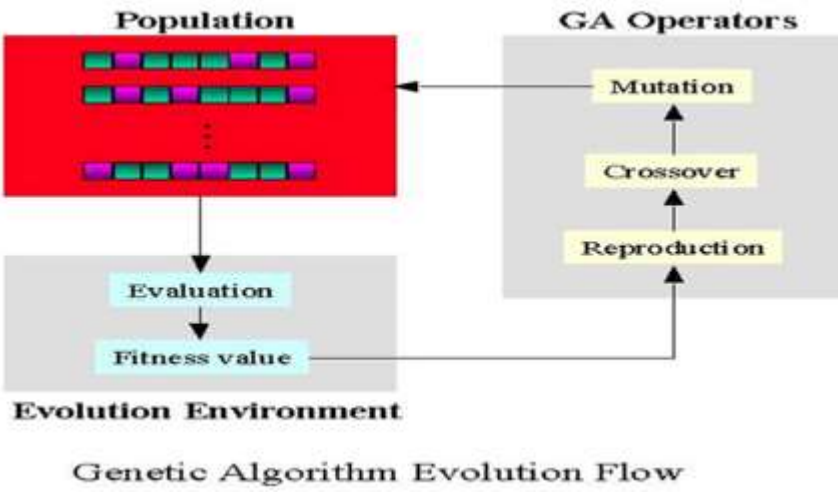
2.遵照适应度越高，选择概率越大的原则，从种群中选择两个个体作为父方和母方。

3.抽取父母双方的染色体，进行交叉，产生子代。

4.对子代的染色体进行变异。

5.重复 2，3，4 步骤，直到新种群的产生。

结束循环。



Genetic Algorithm Evolution Flow

接下来，我们将详细地剖析遗传算法过程的每一个细节。

编制袋鼠的染色体----基因的编码方式

受到人类染色体结构的启发，我们可以设想一下，假设目前只有“0”，“1”两种碱基，我们也用一条链条把他们有序的串连在一起，因为每一个单位都能表现出 1 bit 的信息量，所以一条足够长的染色体就能为我们勾勒出一个个体的所有特征。**这就是二进制编码法**，染色体大致如下：

010010011011011110111110

上面的编码方式虽然简单直观，但明显地，当个体特征比较复杂的时候，需要大量的编码才能精确地描述，相应的解码过程（类似于生物学中的 DNA 翻译过程，就是把基因型映射到表现型的过程。）将过分繁复，为改善遗传算法的计算复杂性、提高运算效率，提出了浮点数编码。染色体大致如下：

1.2 –3.3 – 2.0 –5.4 – 2.7 – 4.3

（注：还有一种编码方式叫符号编码）

那么我们如何利用这两种编码方式来为袋鼠的染色体编码呢？因为编码的目的是建立表现型到基因型的映射关系，而表现型一般就被理解为个体的特征。比如人的基因型是 46 条染色体所描述的却能解码成一个眼，耳，口，鼻等特征各不相同的活生生的人。所以我们要想为“袋鼠”的染色体编码，我们必须先来考虑“袋鼠”的“个体特征”是什么。也许有的人会说，袋鼠的特征很多，比如性别，身长，体重，也许它喜欢吃什么也能算作其中一个特征。但具体在解

决这个问题的情况下，我们应该进一步思考：无论这只袋鼠是长短，肥瘦，黑白只要它在低海拔就会被射杀，同时也没有规定身长的袋鼠能跳得远一些，身短的袋鼠跳得近一些。当然它爱吃什么就更不相关了。**我们由始至终都只关心一件事情：袋鼠在哪里。**因为只要我们知道袋鼠在那里，我们就能做两件必须去做的事情：

（1）通过查阅喜玛拉雅山脉的地图来得知袋鼠所在的海拔高度（通过自变量求适应函数的值。）以判断我们有没必要把它射杀。

（2）知道袋鼠跳一跳（交叉和变异）后去到哪个新位置。

如果我们一时无法准确的判断哪些“个体特征”是必要的，哪些是非必要的，我们常常可以用到这样一种思维方式：比如你认为袋鼠的爱吃什么东西非常必要，那么你就想一想，有两只袋鼠，它们其它的个体特征完全同等的情况下，一只长得黑，另外一只长得不是那么黑。你会马上发现，这不会对它们的命运有丝毫的影响，它们应该有同等的概率被射杀！**只因它们处于同一个地方。**（**值得一提的是**，如果你的基因编码设计中包含了袋鼠黑不黑的信息，这其实不会影响到袋鼠的进化的过程，而那只攀到珠穆朗玛峰的袋鼠黑与白什么的也完全是随机的，但是它所在的位置却是非常确定的。）

以上是对遗传算法编码过程中经常经历的思维过程，必须把具体问题抽象成数学模型，突出主要矛盾，舍弃次要矛盾。只有这样才能简洁而有效的解决问题。

既然确定了袋鼠的位置作为个体特征，具体来说位置就是横坐标。那么接下来，我们就要建立表现型到基因型的映射关系。就是说如何用编码来表现出袋鼠所在的横坐标。由于横坐标是一个实数，所以说透了我们就是要对这个实数编码。回顾我们上面所介绍的两种编码方式，最先想到的应该就是，对于二进制编码方式来说，编码会比较复杂，而对于浮点数编码方式来说，则会比较简洁。恩，正如你所想的，用浮点数编码，仅仅需要一个浮点数而已。而下面则介绍如何建立二进制编码到一个实数的映射。

明显地，一定长度的二进制编码序列，只能表示一定精度的浮点数。譬如我们要求解**精确到六位小数**，由于区间长度为 2 – (-1) = 3 ,为了保证精度要求，至

少把区间[-1,2]分为 3×10^6 等份。又因为

$$2097152 = 2^{21} < 3*10^6 < 2^{22} = 4194304$$

所以编码的二进制串至少需要 22 位。

把一个二进制串（b0,b1,....bn)转化位区间里面对应的实数值通过下面两个步骤。

（1）将一个二进制串代表的二进制数转化为 10 进制数：

$$(b_0 \cdots b_{20} b_{21})_2 = (\sum_{i=0}^{21} b_i \cdot 2^i)_{10} = x'$$

（2）对应区间内的实数：

$$x = -1 + x' \frac{(2-(-1))}{2^{22}-1}$$

（像极了模数转换）

例如一个二进制串<1000101110110101000111>表示实数值 0.637197。

$$x' = (1000101110110101000111)_2 = 2288967$$

$$x = -0.1 + 2288967 \cdot \frac{3}{2^{22}-1} = 0.637197$$

（纠正一个错误，这里是-1）

二进制串<00000000000000000000>和<11111111111111111111>则分别表示区间的两个端点值-1 和 2。

好了，目前为止我们把袋鼠的染色体给研究透了，让我们继续跟进袋鼠的进化旅程

物竞天择－－适应性评分与及选择函数。

1.物竞——适应度函数（fitness function）

自然界生物竞争过程往往包含两个方面：生物相互间的搏斗与及生物与客观环境的搏斗过程。但在我们这个实例里面，你可以想象到，袋鼠相互之间是非常友好的，它们并不需要互相搏斗以争取生存的权利。它们的生死存亡更多是取决于你的判断。因为你要衡量哪只袋鼠该杀，哪只袋鼠不该杀，所以你必须制定一个衡量的标准。而对于这个问题，这个衡量的标准比较容易制定：袋鼠所在的海拔高度。（因为你单纯地希望袋鼠爬得越高越好。）所以我们直接用袋鼠的海拔高度作为它们的适应性评分。即适应度函数直接返回函数值就行了。

2.天择——选择函数（selection）

自然界中，越适应的个体就越有可能繁殖后代。但是也不能说适应度越高的就肯定后代越多，只能是从概率上来说更多。（毕竟有些所处海拔高度较低的袋鼠很幸运，逃过了你的眼睛。）那么我们怎么来建立这种概率关系呢？下面我们介绍一种常用的选择方法——轮盘赌（Roulette Wheel Selection）选择法。

比如我们有 5 条染色体，他们所对应的适应度评分分别为：5，7，10，13，15。

所以累计总适应度为：

$$F=\sum_{i=1}^{n} f_i = 5+7+10+13+15 = 50$$

所以各个个体被选中的概率分别为：

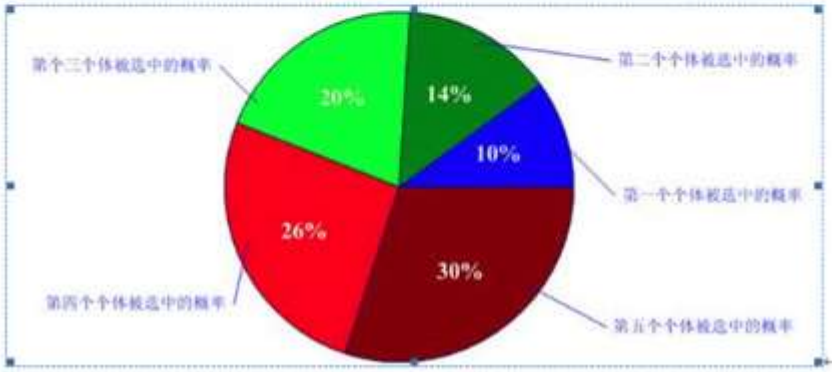
$$P_1 = \frac{f_1}{F} \times 100\% = \frac{5}{50} \times 100\% = 10\%$$

$$P_2 = \frac{f_2}{F} \times 100\% = \frac{7}{50} \times 100\% = 14\%$$

$$P_3 = \frac{f_3}{F} \times 100\% = \frac{10}{50} \times 100\% = 20\%$$

$$P_4 = \frac{f_4}{F} \times 100\% = \frac{13}{50} \times 100\% = 26\%$$

$$P_5 = \frac{f_5}{F} \times 100\% = \frac{15}{50} \times 100\% = 30\%$$



你可以想象一下，我们转动轮盘，轮盘停下来的时候，指针会随机地指向某一个个体所代表的区域，那么非常幸运地，这个个体被选中了。（很明显，适应度评分越高的个体被选中的概率越大。）

注：还有精英选择机制

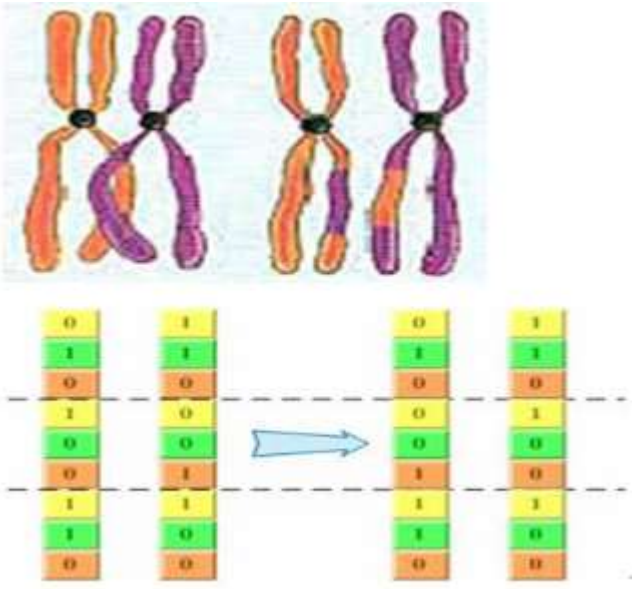
遗传变异——基因重组（交叉）与基因突变。

应该说**这两个步骤就是使得子代不同于父代的根本原因（注意，我没有说是子代优于父代，只有经过自然的选择后，才会出现子代优于父代的倾向。）**。对于这两种遗传操作，二进制编码和浮点型编码在处理上有很大的差异，其中二进制编码的遗传操作过程，比较类似于自然界里面的过程，下面将分开讲述。

1.基因重组/交叉(recombination/crossover)

（1）二进制编码

二进制编码的基因交换过程非常类似高中生物中所讲的同源染色体的联会过程——随机把其中几个位于同一位置的编码进行交换，产生新的个体。



（2）浮点数编码

如果一条基因中含有多个浮点数编码，那么也可以用跟上面类似的方法进行基因交叉，不同的是进行交叉的基本单位不是二进制码，而是浮点数。而如果对于单个浮点数的基因交叉，就有其它不同的重组方式了，比如中间重组：随机产生就能得到介于父代基因编码值和母代基因编码值之间的值作为子代基因编码的值。比如 5.5 和 6 交叉，产生 5.7，5.6。

考虑到“袋鼠跳”问题的具体情况——袋鼠的个体特征仅仅表现为它所处的位置。可以想象，同一个位置的袋鼠的基因是完全相同的，而两条相同的基因进行交叉后，相当于什么都没有做，所以我们不打算在这个例子里面使用交叉这一个遗传操作步骤。（当然硬要这个操作步骤也不是不行的，你可以把两只异地的袋鼠捉到一起，让它们交配，然后产生子代，再把它们送到它们应该到的地方。）

2.基因突变(Mutation)

（1）二进制编码

基因突变过程：基因突变是染色体的某一个位点上基因的改变。基因突变使一个基因变成它的等位基因，并且通常会引起一定的表现型变化。正如上面所说，二进制编码的遗传操作过程和生物学中的过程非常相类似，基因串上的“ 0”或“ 1”有一定几率变成与之相反的“ 1”或“ 0”。例如下面这串二进制编码：

101101001011001

经过基因突变后，可能变成以下这串新的编码：

001101011011001

（2）浮点型编码

浮点型编码的基因突变过程一般是对原来的浮点数增加或者减少一个小随机数。比如原来的浮点数串如下：

1.2,3.4,5.1, 6.0, 4.5

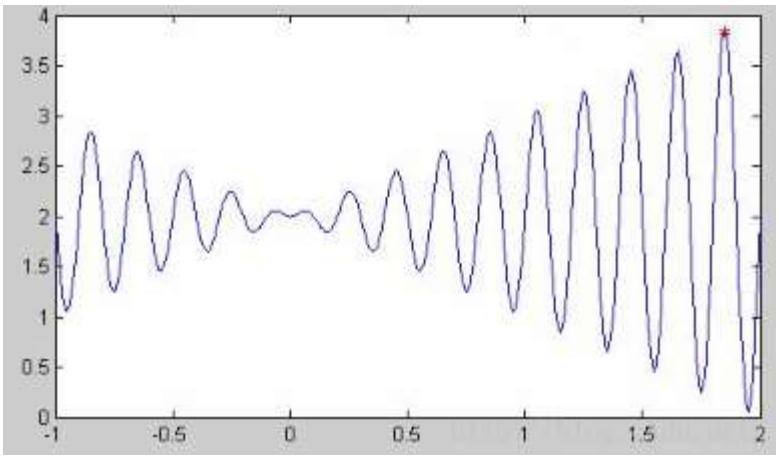
变异后，可能得到如下的浮点数串：

1.3,3.1,4.9, 6.3, 4.4

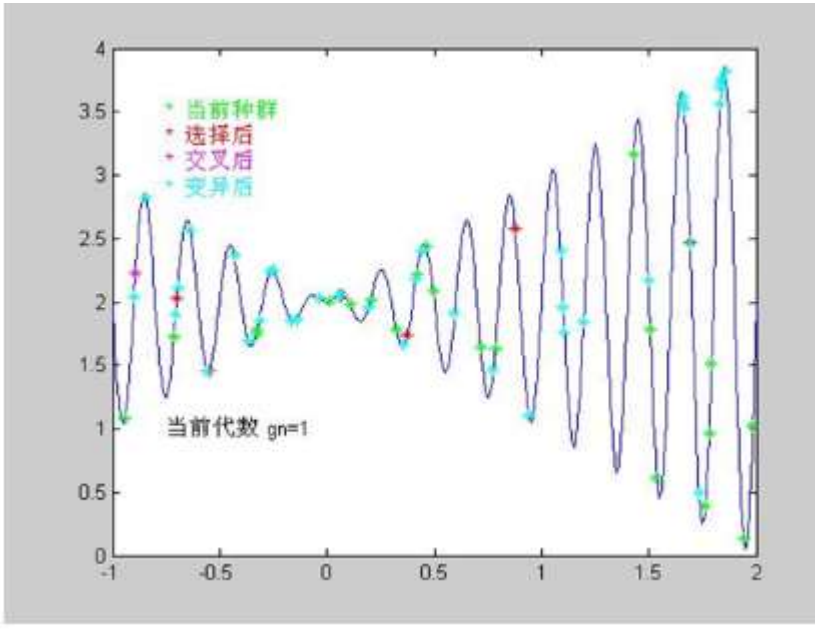
当然，**这个小随机数也有大小之分，我们一般管它叫“步长”。**（想想“袋鼠跳”问题，袋鼠跳的长短就是这个步长。）一般来说**步长越大，开始时进化的速度会比较快，但是后来比较难收敛到精确的点上。**而小步长却能较精确的收敛到一个点上。所以**很多时候为了加快遗传算法的进化速度，而又能保证后期能够比较精确地收敛到最优解上面，会采取动态改变步长的方法。**其实这个过程与前面介绍的模拟退火过程比较相类似。

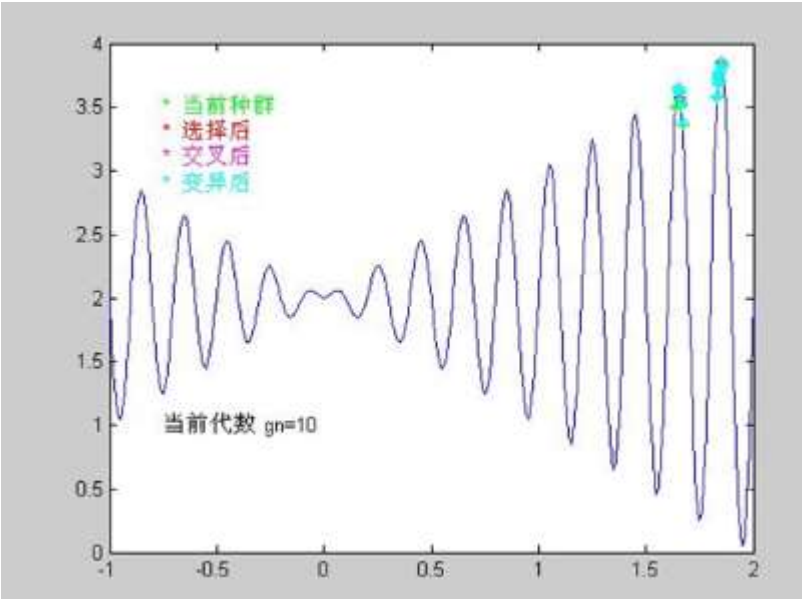
到此为止，**基因编码，基因适应度评估，基因选择，基因变异都一一实现了**，剩下来的就是把这些遗传过程的“零件”装配起来了。（写成代码）

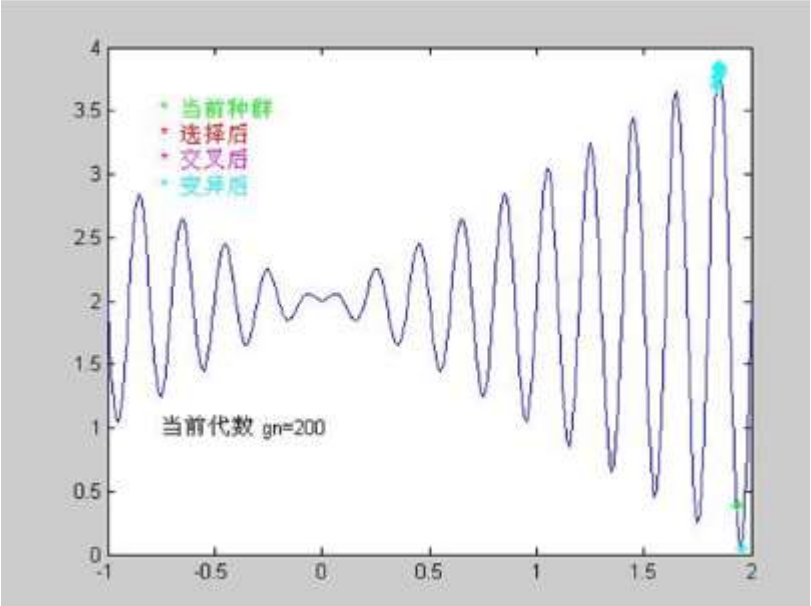
下面是上例的运行结果：



红点代表真实的最大点，由求导法可求的为 f(1.85)=3.85







总结:

编码原则

完备性（completeness）：问题空间的所有解都能表示为所设计的基因型;

健全性（soundness）：任何一个基因型都对应于一个可能解;

非冗余性（non-redundancy）：问题空间和表达空间一一对应。

适应度函数的重要性

适应度函数的选取直接影响遗传算法的收敛速度以及能否找到最优解。一般而言，适应度函数是由目标函数变换而成的。

适应度函数设计不当有可能出现欺骗问题：
（1）进化初期，个别超常个体控制选择过程；
（2）进化末期，个体差异太小导致陷入局部极值。

欺骗问题举例：

还是袋鼠问题，如果低海拔的地方出现毒雾，会杀死袋鼠，只有爬上珠穆朗玛峰顶端的袋鼠才能生存下来。

因为喜马拉雅山脉有很多山峰，我们以高度作为适应度，case（1）：如果不在珠峰的猴子若比在珠峰半山腰的猴子要高，因为种群大小不变，在珠峰的猴子可能就会被淘汰；case（2）：100 只猴子都不在珠峰；

1. 选择的作用：优胜劣汰，适者生存；

2. 交叉的作用：保证种群的稳定性，朝着最优解的方向进化；

3. 变异的作用：保证种群的多样性，避免交叉可能产生的局部收敛。