Featurebox . Symbol

v0.083

Statement and Acknowledgement :

This tool is deeply customized and copied by **deap** and **sympy.**
This tool is advised to personal and non-commercial use.
This tool is with GNU license and with NO WARRANTY.

Email: 986798607@qq.com

Cite:

Introduction

This tool is a symbol regression tool with dimension calculation, which is aimed at establish expressions with physical limitation.

Features:

- coefficient fitting and addition
- dimension calculation
- accumulative operation and free custom
- characteristics feedback
- high efficiency parallelism

Contents

Example

```python
if __name__ == "__main__":
    from featurebox.symbol.base import SymbolSet
    from featurebox.symbol.dim import dless, Dim
    from featurebox.symbol.flow import BaseLoop
    # data
    data = load_boston()
    x = data["data"]
    y = data["target"]
    c = [6, 3, 4]

    # unit
    from sympy.physics.units import kg
    x_u = [kg] * 13
    y_u = kg
    c_u = [dless, dless, dless]
   # Dim,    the dim also could get by Dim(numpy.array([****])) directly.
    x, x_dim = Dim.convert_x(x, x_u, target_units=None, unit_system="SI")
    y, y_dim = Dim.convert_xi(y, y_u)
    c, c_dim = Dim.convert_x(c, c_u)

    # symbolset
    pset0 = SymbolSet()
    pset0.add_features(x, y, x_dim, y_dim, group=[[1, 2], [4, 5]])
    pset0.add_constants(c, dim=c_dim, prob=None)
    pset0.add_operations(power_categories=(2, 3, 0.5),
                         categories=("Add", "Mul", "Neg", "Abs"),
                         self_categories=None)
    #  run
    bl = BaseLoop(pset=pset0, gen=8, pop=500, hall=2, batch_size=50, n_jobs=10,
                  re_Tree=0, store=False)
    bl.run()
```

Data import

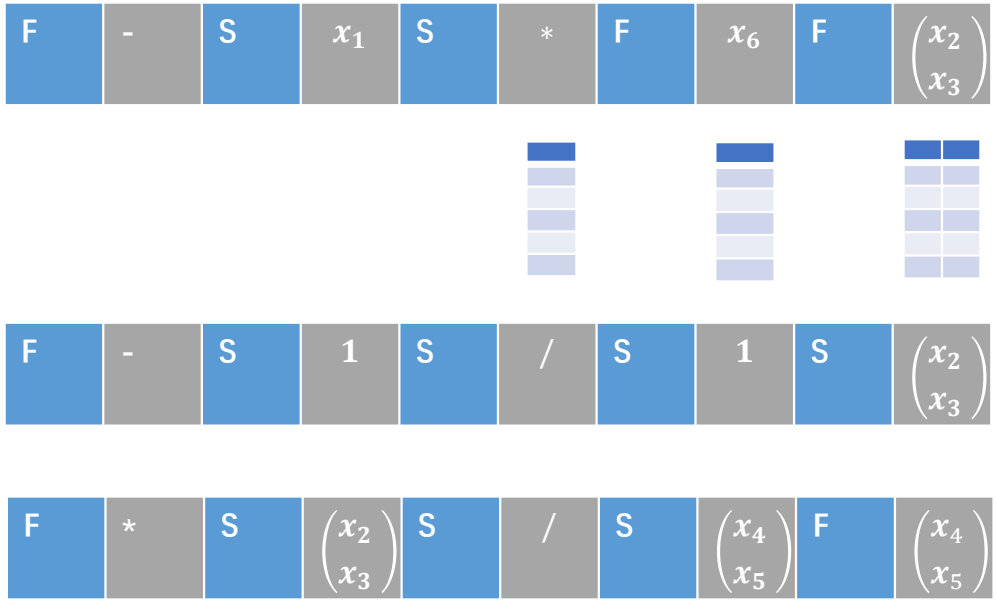
Unit
(optional)

Dim
(optional)

Preparation set
add features, constants
and operations.

Flow and loop
set parameters to run

Expression and Tree Code



$$x_1 - x_6 * (x_2 + x_3)$$

$$1 - (\frac{1}{x_4} + \frac{1}{x_5})$$

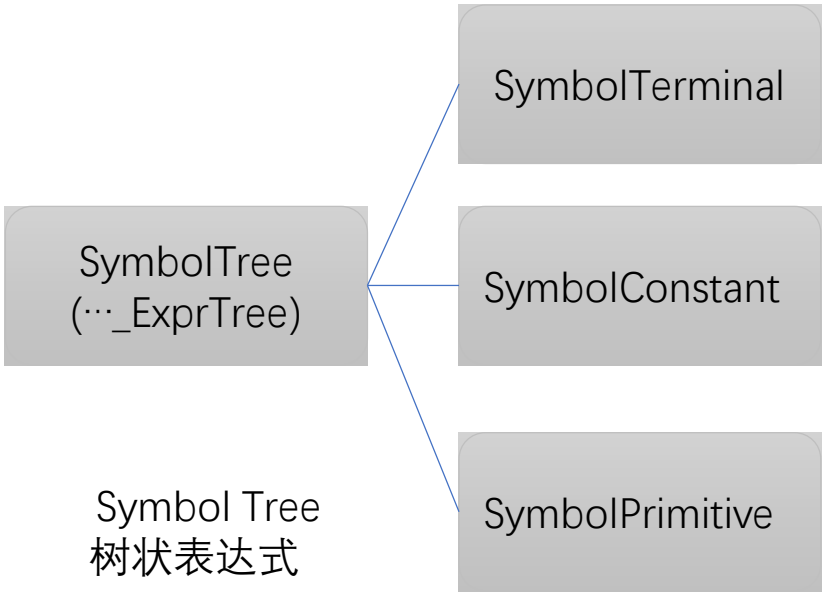
$$\frac{x_2 x_4 + x_3 * x_5}{x_4 + x_5}$$

F : madd $\sum$
sympy function: undefinition
np function: np.sum(axis=1)
dim function: the same as +

S : self
sympy function: lambda x:x
np function: lambda x:x
dim function: lambda x:x

Others: mmul $\prod$
msub
mdiv

Contain



Symbol Tree
树状表达式

Method

.generate(SymbolSet)

Produce the Tree from symbolset

.depart(SymbolSet)

depart the Tree to subtree

. capsule(SymbolSet)
get a short type of tree only contain name

Attribute

self.p_name

self.dim

self.pre_y

self.expr

Contain

SymbolSet

Preparation set
组件合集

Method

.add_operations

Add operations in Preparation set

.add_accumulative_operation

Add accumulative operations in Preparation set

. add_features

Add features in Preparation set

. add_constants

Add constants in Preparation set

. add_tree_to_features

Add SymbolTree in Preparation set back as a terminals,and assign a new name

. compress

Delet details for zip. Use after add all.

Method

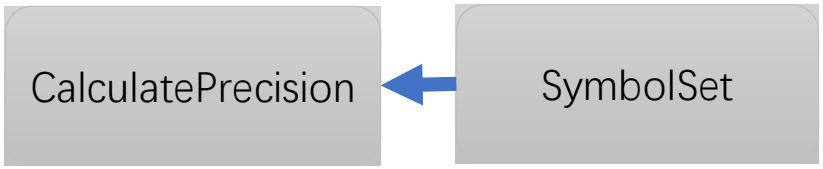
.set_personal_maps

Set personal preferences on features probability, by single point.
see also: preamp.set_sigle_point

.bonding_personal_maps

Set personal preferences on features probability,by cut others point
see also: premap.down_other_point

Contain



CalculatePrecision
数值计算，量纲计算

Method

. Calculate_simple()

calulate the Tree from symbolset
Return the SymbolTree self,but resite the attribute

. Calculate_detial()

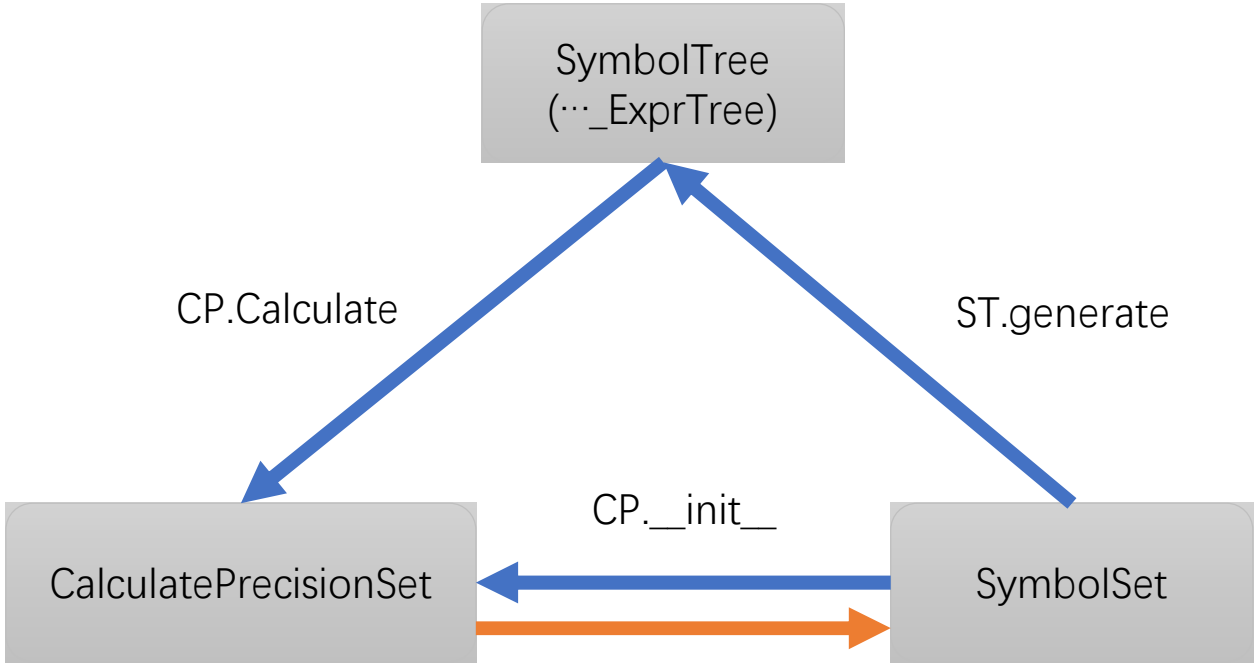
calulate the Tree from symbolset
Return the SymbolTree self,but resite the attribute

. Calculate_parallize()

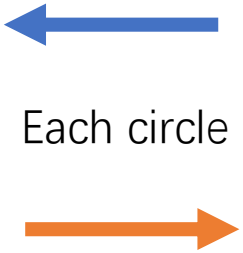
calulate the Tree from symbolset
Return the score, dimension, and dim_score

Attribute

```
        self.pset = pset
        self.terminals = pset.terminals +
pset.constants
        # list of sympy.Symbol, features and
constants
        self.dim_x=
pset.get_values(pset.dim_ter_con)
  # list of dims
        self.data_x = pset.data_x  # list of xi
        self.dim_map = pset.dim_map
        self.np_map = pset.np_map

        self.y = pset.y  # list of y
        self.filter_warning = filter_warning
        self.scoring = scoring
```



Flow relationship between 3 base object



Each circle

2+ circle
if add trees as new
features

Functions Type

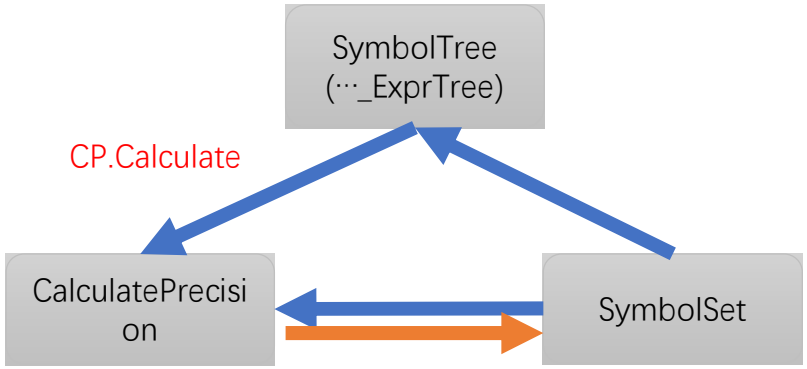
```
String_name     =     repr(SymbolTree)

Sympy_expr      =     compile_(SymbolTree,…)

Func            =     lambdify_(Sympy_expr,…)
Y_pre           =     Func(x)
Y               =     score(Y_true,Y_pre)


Func            =     lambdify_(Sympy_expr,…)
Y_dim           =     Func(x_dim)
```

“func_name”

func_map	Sympy function
np_map	Numpy function
Dim_map	Dim function

Calculation rules
计算规则

1.universal operation are with 3 function.
常见内置规则同时定义3个函数。
2.
self operation is without sympy function, which is constructed automatedly.
自定义规则不定义 sympy function, 默认使用Sympy.
Function 创建



Dimension calculation

Example:

a: d_a : dimension 1
b: d_b : dimension 2
c1: d_1 : dimensionless
d: d_{nan} : without dimension

Operation	Dimension calculation	Result
+	$d_a + d_a$	d_a
	$d_a + d_b$	d_{nan}
	$d_a + d_1$	d_a
	$d_1 + d_1$	d_1
-	the same with +	
*	$d_a * d_a$	$2 * d_a$
	$d_a * d_b$	$d_a + d_b$
	$d_a * d_1$	d_a
	$d_1 * d_1$	d_1
/	d_a / d_a	d_1
	d_a / d_b	$d_a - d_b$
	d_a / d_1	d_a
	d_1 / d_1	d_1

Operation	Dimension calculation	Result
$\sum$ madd	the same with +, could accept 1 input and return it	
- re-name sub to msub	the same with -, could accept 1 input and return it invalid if accept more than 2.	
$\prod$ mmul	the same with *, could accept 1 input and return it	
/ re-name div to mdiv	the same with /, could accept 1 input and return it invalid if accept more than 2.	

Operation	Dimension calculation	Result
-x (negative particular case -)	$-d_a$	d_a
	$-d_1$	d_1
1/x (negative particular case /)	$1/d_a$	$-d_a$
	$1/d_1$	d_1
ln exp sin cos	$f(d_a)$	d_{nan}
	$f(d_1)$	d_1
x^n	d_a^n	$n * d_a$
	d_1^n	d_1
n^x	n^{d_a}	d_{nan}
n^x	n^{d_1}	d_{nan}
abs	abs(d)	d
self	$self(d)$	d

$$d_1 = function(d_1) \quad for\ any\ function$$
$$g(d_{other}) = function(d_1, d_{other}) \quad for\ any\ function$$

$$d_{nan} = function(d_{nan}) \quad for\ any\ function$$
$$d_{nan} = function(d_{nan}, d_{other}) \quad for\ any\ function$$

Contain

Dim

0
0
0
0
0
0
0

Dim
量纲

Method

.__xx__

operator overloading

.isinteger

.is_same_base

is_same_base,such as m, m^3

.convert_to_Dim

Get scale and Dim from sympy.physics.unit

.convert_x

Get scaled x and Dim from x and sympy.physics.unit

.inverse_convert

Get sympy.physics.unit from Dim

.inverse_convert_x

Get scaled x and sympy.physics.unit from Dim

Attribute

self.unit_map = {'meter': "m", 'kilogram': "kg", 'second': "s", 'ampere': "A", 'mole': "mol", 'candela': "cd", 'kelvin': "K"}

self.unit = SI._base_units

self.dim = ['length', 'mass', 'time', 'current', 'amount_of_substance', 'luminous_intensity', 'temperature']

Default is SI system, with 7 member.
Can be set as other system with less than 7,such as MKS system ('meter': "m", 'kilogram':)

Coefficient and Constant

1. Physical recognized Coefficient and Constant (e= 1.602×10^{-19} **C =** 1.602×10^{-19} **A*S**)

   Add the **value** and its **SI dimension** to input as a new feature, Put it **random site** in the expression.

2. Common number (2,1,3,1/2……)

   Add the **value** and **dimensionless** to input as a new feature, Put it in **random site** of the expression.

3. Fit coefficient and constant

   >Firstly, get expression.

   >Insert the placeholders to expression.

   >Fit the placeholder to coefficient and constant.

   (1) Value, fitted

   (2) Site,

   Locked in the outer sphere of expression. $y = a * f(x) + b * g(x) + c$ or $y = a * f(x) + c$

   (3)**Dimension,**

   The dimension are **automatic acquired.**

Coefficient and Constant

3. Fit coefficient and constant

>Firstly, get expression.

>Insert the placeholders to expression.

>Fit the placeholder to coefficient and constant.

(1) Value, fitted

(2) Site,

Locked in the outer sphere of expression. $y = a * f(x) + b * g(x) + c$ or $y = a * f(x) + c$

(3)**Dimension,**

The dimension are **automatic acquired.**

Rule:

1.the fitted Coefficients in one expression don't change the Dimensional calculation results of inner $f(x)$.

2. the fitted Coefficients are with same dimension, means that the $f(x)$ **and** $g(x)$ **of** meaningful formula are same. The dimensions of fitted Coefficients are get by $\dim_a = \dim_b = \frac{\dim_y}{\dim_{f(x)}} = \frac{\dim_y}{\dim_{g(x)}}$**.**

3.The fitted constant are the same with $\dim_y$**.**

Probability and Control

probability maps are used to control the choice terminals(features and constant)
:
2D Diagonal maps is used to choice
terminals at the affect no other terminals.
:
1D probability map is used to choice terminals when there is no other terminals.

Note:
the summary of all the probability maybe not 1.
Just relative size makes sense

when relative proportion is 1, the two features is bonding forcedly.

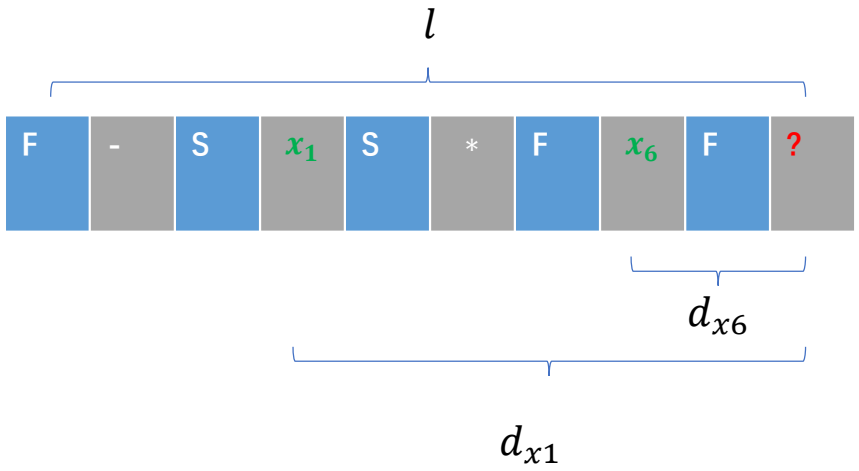
.	x1	x2	x3	x4	x5	x6	x7
x1	0.01	0.2	0.4	0.2	0.3	0.1	0.3
x2	0.2	···	···	···	···	···	···
x3	0.4	···	···	···	···	···	···
x4	0.2	···	···	···	···	···	···
x5	0.3	···	···	···	···	···	···
x6	0.1	···	···	···	···	···	···
x7	0.3	···	···	···	···	···	···

2 D Diagonal matrix (probability map)

x1	x2	x3	x4	x5	x6	x7
0.3	0.2	0.4	0.2	0.3	0.1	0.3

1 D probability map

Example individual



the ? would consider the presented x1,x6

$$\vec{p}_{point} = f(d_{x1})\vec{p}_{x1} + f(d_{x6})\vec{p}_{x6}$$

$f(x)$ is a decreasing function.
such as:
$f(x) = l - d$

Contain

PreMap
(np.ndarray)

2 D Diagonal matrix
(probability map)
probability of choice
at others already existed

prob

1 D probability map
probability of single choice

Method

. down_other_point

decrease the value of others and increase target point

. set_sigle_point

set the single couple point but don't change others.

. set_ratio_point

set the ratio on value of point

. set_ratio

set the ratio of summary value on point

Method

get_indexes_value

get the values list of one index

update

update values by individuals

get_one_node_value

get the probability values list of one point of individual weight by distance to the point

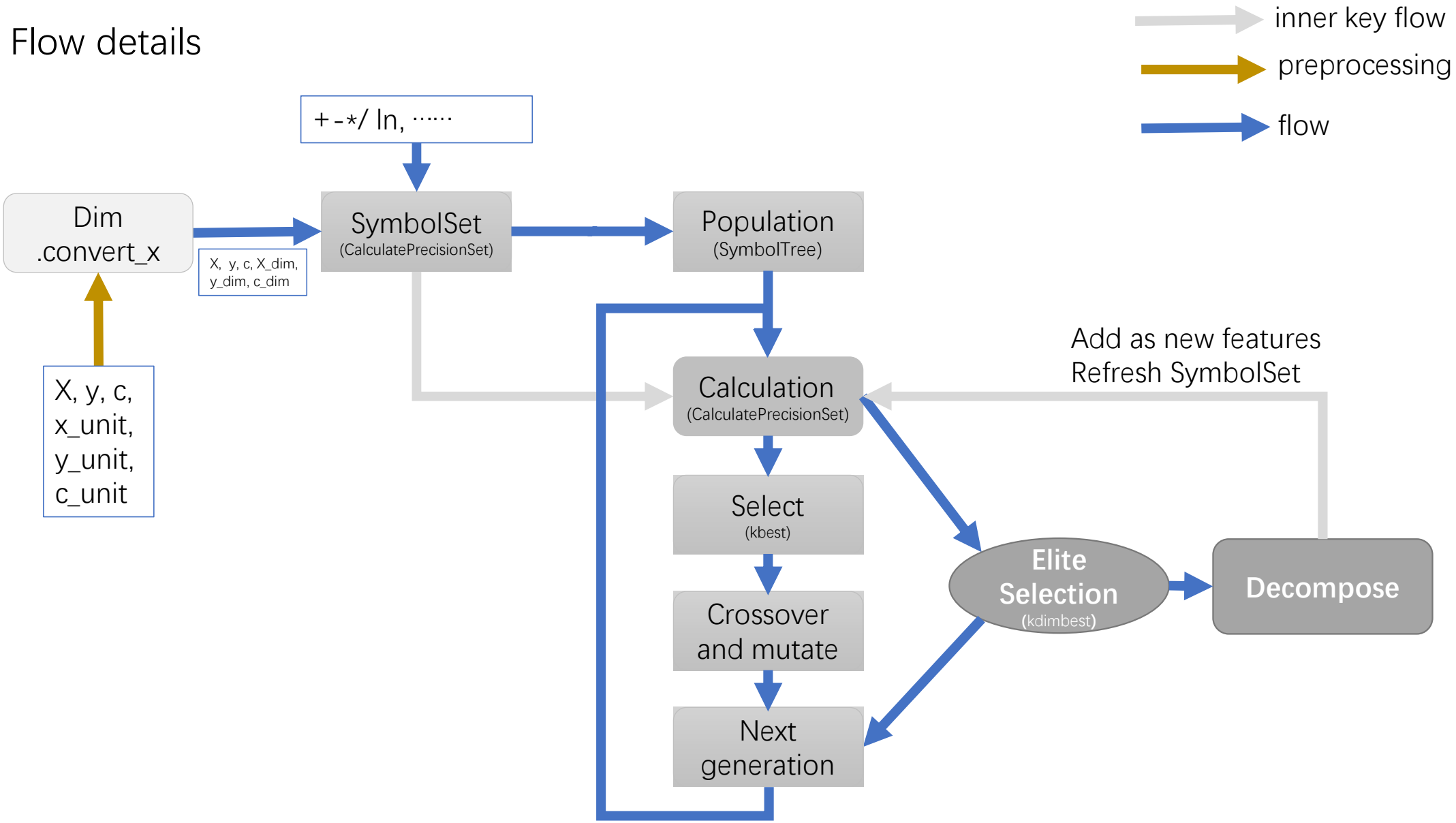
get_nodes_value

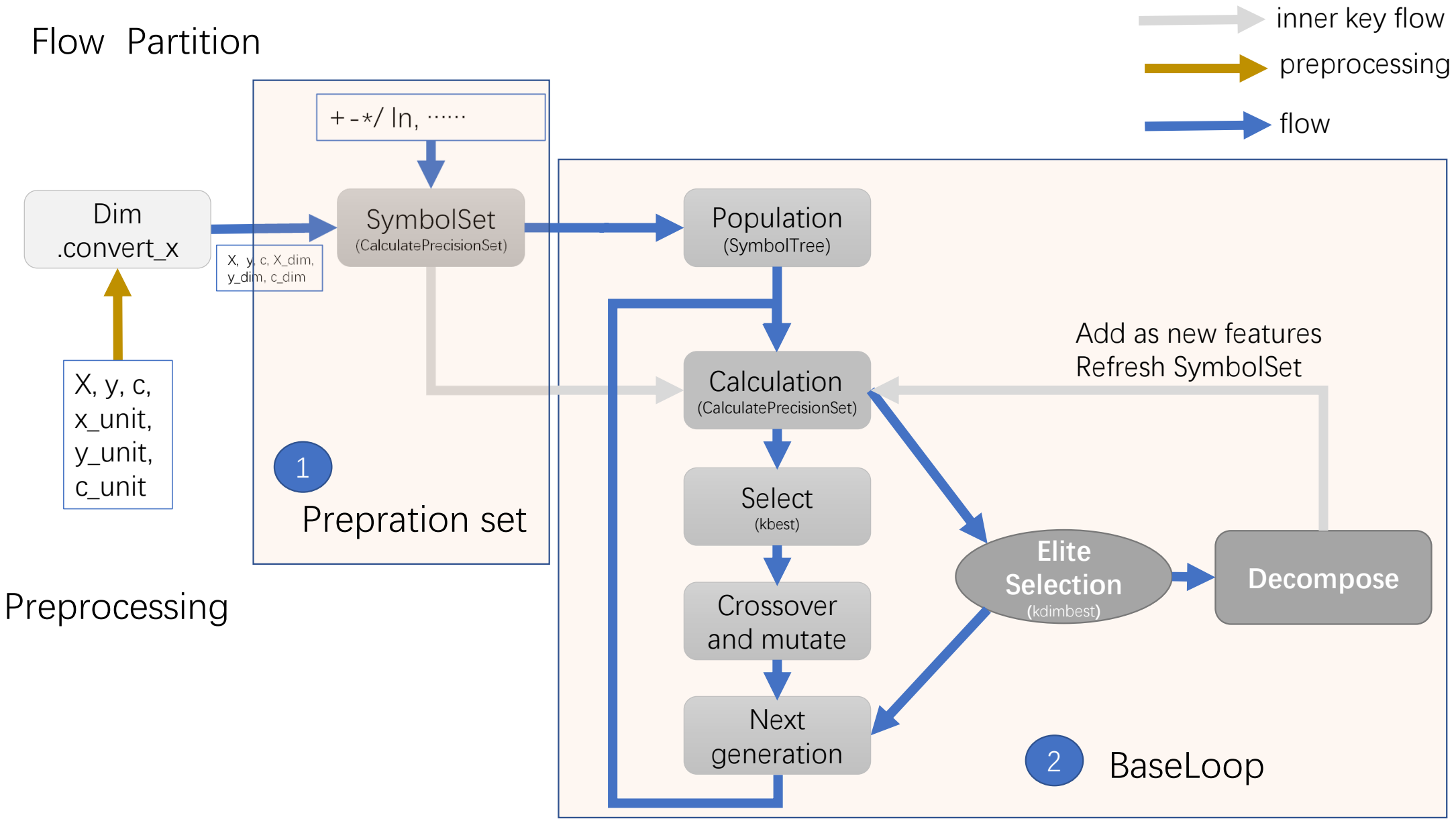
get the probability values list of a few points of individual weight by distance to the point

get_ind_value

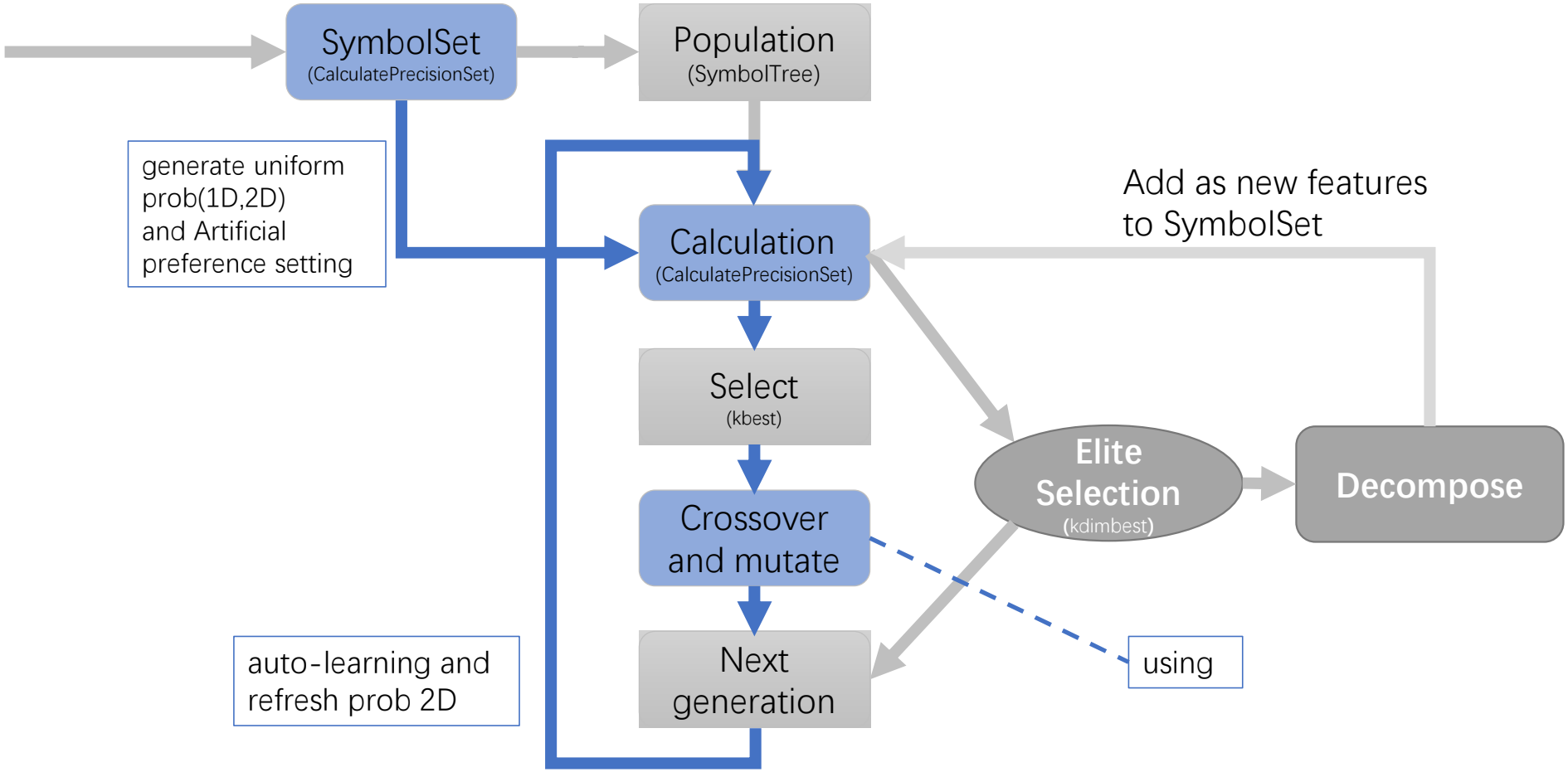
get the probability values of individual average weighted

Flow details




Flow Partition
inner key flow
preprocessing
flow
+ - * / ln, ……
Dim
.convert_x
SymbolSet
(CalculatePrecisionSet)
X, y, c, X_dim, y_dim, c_dim
X, y, c,
x_unit,
y_unit,
c_unit
1
Prepration set
Preprocessing
Population
(SymbolTree)
Calculation
(CalculatePrecisionSet)
Add as new features
Refresh SymbolSet
Select
(kbest)
Crossover
and mutate
Next
generation
Elite
Selection
(kdimbest)
Decompose
2
BaseLoop

Probability of Flow details refresh and auto-learning



1 Prepration set parameter and method

parameter	Type	Doc	Chinese
.add_operations .add_accumulative_operation			
power_categories=None,	list of int	power function	添加幂函数，指数
categories	list of str	function	常见函数
self_categories=None,	list of list of 5 member	self definition	自定义函数 (详情见文档)
power_categories_prob="balance",	int,str	probility of power categories	添加幂函数出现概率
categories_prob="balance",	int,str	probility of categories	常见函数函数出现概率
special_prob=None	dict	specific probility of categories	指定某函数出现概率

1 Prepration set parameter and method

parameter	Type	Doc	Chinese
add_features add_constants			
x	np.ndarray 2D	feature values	特征值
y	np.ndarray	target value	目标值
x_dim	list of Dim	feature dimensions	特征量纲 若为1，默认全部无量纲
y_dim	Dim	target dimensions	目标量纲 若为1，默认无量纲
prob	None, list of float the same size wih x.shape[1]	1D sigle choice probability	单特征出现概率
group	list of list	group features index	特征分组，强制绑定在一起计算
feature_name	None, list of float the same size wih x.shape[1	feature name	初始特征名字，若用，仅用于展示
bonding_personal_maps set_personal_maps			
sv	list of 3 member's list	set value of interaction map such as [[3,4,0.5],[5,6,0.03]]	定制特征互影响概率

2 Flow loop parameter

Parameter	Type	Doc	Chinese
pset	SymbolSet	the feature x and traget y and others should have been added.	已经添加好特征常数，运算符的准备序列
pop	int	number of population	遗传种群大小
gen	int	generation	遗传代数
mutate_prob	float [0,1)	probability of mutate	变异概率
mate_prob	float [0,1)	probability of mate(crossover)	交叉概率
initial_max	int	max initial size of expression	初始个体(表达式)尺寸上限
max_value	int	max size of expression	个体(表达式)尺寸上限
hall	int >=1	number of HallOfFame(elite) to maintain	精英个数
re_hall	None, int>=2	number of HallOfFame to add to next generation.	回馈精英个数
re_Tree	int	number of new features to add to next generation.	每次循环，个体作为新特征个数
personal_map	bool or “auto”	"auto" is using premap and with auto refresh the premap with individual True is just using constant premap False is just use the prob of terminals	是否使用互影响概率（2D），以及是否自动更新

2 Flow loop parameter

Parameter	Type	Doc	Chinese
scoring	list of Callbale	default is [sklearn.metrics.r2_score], scores to evaluate the (y_true, y_calculate)	评分列表，若使用多评分，请保证这些评分可以加权平均
score_pen	tuple of float	>0 : max problem, best is positive, worse -np.inf <0 : min problem, best is negative, worse np.inf if multiply score method, the scores must be turn to same dimension in preprocessing or weight by score_pen. Because the all the selection are stand on the mean(w_i*score_i). default (1,)	最大值问题为（1，） 最小值问题为（-1，） 多评分问题（0.6，0.4）加权平均
add_coef	bool	add coef in expression or not.	是否添加系数项
inter_add	bool	add intercept in expression or not.	是否添加截距项
inner_add:bool	bool	add coef to inner expression or not.	系数项可否到公式内部
cal_dim	bool	the dim calculation	是否计算量纲
dim_type	Dim or None or list of Dim	**more strict** ↓ "coef": af(x)+b. a,b have dimension, f(x) is not dnan. "integer": in af(x)+b. f(x) is interger dimension. [Dim1,Dim2]: f(x) in list. Dim: f(x) ~= Dim. (see fuzzy) Dim: f(x) == Dim. None: f(x) == pset.y_dim	目标量纲
fuzzy	bool	choose the dim with same base with dim_type, such as m,m^2,m^3.	放宽量纲限制到 同底量纲

2 Flow loop parameter

Parameter	Type	Doc	Chinese
stats:	dict	details of logbook to show. default is stats = {“fitness_dim_is_target“: ("mean",), "dim_is_traget": ("sum",)}	记录统计（详情见文档） 此记录与精英记录hall独立。可以查看更多信息，精英hall只对最好个体保存。 两者对应结果，即打印结果和输出结果可能为不同信息。
verbose	bool	print verbose logbook or not	动态打印记录统计
tq	bool	print progress bar or not	进度条
store	bool,or str of path	store	默认存储当前文件夹下。 若为字符串，请保证字符串为路径
filter_warning	bool	filter_warning	是否过滤警告
n_jobs	int default 1, advise 6	paralyze number	并行参数
batch_size	int, default 40, depend of machine	batch_size to calcalation	分批大小
random_state:	None,int	np.random seed	随机种子