Постановка задач и выбор моделей в машинном обучении

Вадим Викторович Стрижов

Московский физико-технический институт

Осенний семестр 2019

Модели как суперпозиции

Экспертно задано множество G порождающих функций $g(\mathbf{w}, \mathbf{x})$.
Для g_i определены области аргументов $\mathbf{w} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^n$ и значений $g_i(x) \in \mathbb{R}^1$.
Множество G пополняется функциями $\mathrm{id}(\mathbf{x})$ и const.
Искомая модель f будет искаться среди множества $\mathfrak{F}$ суперпозиций функций $g \in G$.

Список порождающих функций

Description	In	N in	Out	N out	Comm	Param
Nominal to binary	nom	1	bin	1–4	-	Yes
Ordinal to binary	ord	1	bin	1–4	-	Yes
Linear to linear segments	lin	1	lin	1–4	-	Yes
Linear segments to binary	lin	1	bin	1–4	-	Yes
Get one column of n-matrix	bin	1–4	bin	1	-	Yes
Conjunction	bin	2–6	bin	1	Yes	-
Dijsunction	bin	2–6	bin	1	Yes	-
Negate binary	bin	1	bin	1	-	-
Logarithm	lin	1	lin	1	-	-
Hyperbolic tangent sigmiod	lin	1	lin	1	-	-
Logistic sigmoid	lin	1	lin	1	-	-
Sum	lin	2–3	lin	1	Yes	-
Difference	lin	2	lin	1	No	-
Multiplication	lin,bin	2–3	lin	1	Yes	-
Division	lin	2	lin	1	No	-
Inverse	lin	1	lin	1	-	-
Polynomial transformation	lin	1	lin	1	-	Yes
Radial basis function	lin	1	lin	1	-	Yes
Monomials: $x\sqrt{x}$, etc.	lin	1	lin	1	-	-

Задача порождения признаков

Даны

- измеряемые признаки $\Xi = \{\xi\}$,
- заданные экспертами порождающие функции $G = \{g(\mathbf{b}, \xi)\}$,
$$g : \xi \mapsto x;$$
- правила порождения: $\mathcal{G} \supset G$, где суперпозиция $g_k \circ g_l \in \mathcal{G}$ построена с учетом ограничений на число типы входных и выходных переменных ;
- правила упрощения суперпозиций: g_u не принадлежит $\mathcal{G}$, если существует правило
$$r : g_u \mapsto g_v \in \mathcal{G}.$$

Результат

набор «композитных» признаков $X = \{\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_n\}$.

Внимание! Число порожденных признаков может превосходить число клиентов!

Примеры композитных признаков

- **Frac**(Period of residence, Undeclared income)
- **Frac**(**Seg**(Period of employment), Term of contract)
- **And**(Income confirmation, Bank account)
- **Times**(**Seg**(Score hour), **Frac**(**Seg**(Period of employment), Salary))

Допустимые суперпозиции

$$f = (g_1 \circ \ldots \circ g_V)(\mathbf{x})$$

Определение

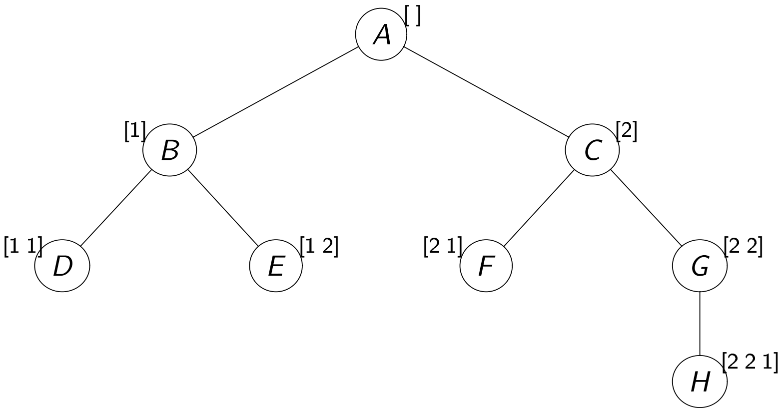
Допустимыми называют суперпозиции, удовлетворяющие следующим требованиям

- *Элементами* s_i *суперпозици* f *могут являться только порождающие функции* g_j *и свободные переменные* x_i*.*
- *Количество аргументов элемента суперпозиции* s_i *равно арности соответствующей ему функции* g_j*.*
- *Порядок аргументов элемента* s_i *суперпозиции* f *соотвествует порядку аргументов соответствующей функции* g_j*.*
- *Для элемента* s_i*, аргументом которого является элемент* s_j*, для соответствующих порождающих функций* $dom(g_i) \supset cod(g_j)$*;*

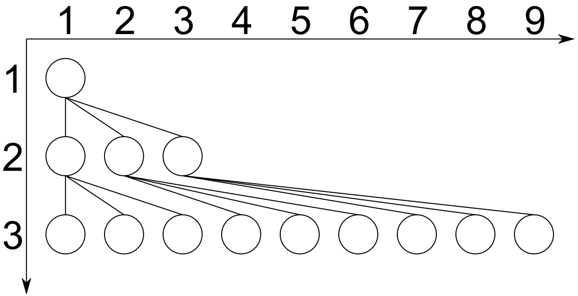
Каждой суперпозиции f однозначно соответствует дерево Γ_f:

- В вершинах v_i дерева Γ_f находятся соответствующие функции g_j.
- Число дочерних вершин у v_i равно арности соответствующей функции g_j.
- Порядок дочерних вершин v_i соотвествует порядку аргументов функции g_j.
- Листьями дерева Γ_f являются x_i или const.

Естественным образом каждой вершине может быть сопоставлена координатная строка.



Наиболее удобным способом задания координат является прямоугольная система, в которой ограничена арность вершин $a_{\max}$.



Расстояние между моделями

Определение

Дерево $\Gamma'(V')$ *называется поддеревом дерева* $\Gamma(V)$, *если его множество вершин* V' *является подмножеством множества* V.

Определение

Два дерева Γ_1 *и* Γ_2 *называются изоморфными, если между их множествами вершин существует взаимно однозначное отображение, сохраняющее метки вершин.*

Определение

Дерево Γ_0 *называется общим поддеревом деревьев* Γ_1 *и* Γ_2, *если в них существуют поддеревья* Γ_1' *и* Γ_2', *изоморфные дереву* Γ_0.

Определение

Общее поддерево двух деревьев называется наибольшим, если в нем содержится наибольшее число вершин среди других общих поддеревьев. Наибольшее поддерево деревьев $\Gamma_i(V_i)$ *и* $\Gamma_j(V_j)$ *обозначается как* $\Gamma_{ij}(V_{ij})$*. Символом p будет обозначаться количество элементов в множестве* V*:* $|V_i| = p_i$*,* $|V_j| = p_j$*,* $|V_{ij}| = p_{ij}$*.*

Рассмотрим абсолютную функцию расстояния r между Γ_i и Γ_j, зависящую от их размеров и наибольшего общего подграфа,

$$r_{ij} = p_i + p_j - p_{ij},$$

Доказательство аксиом метрики

Рассмотрим неравенство треугольника. Пусть даны графы Γ_1, Γ_2 и Γ_3. Необходимо доказать неравенство:

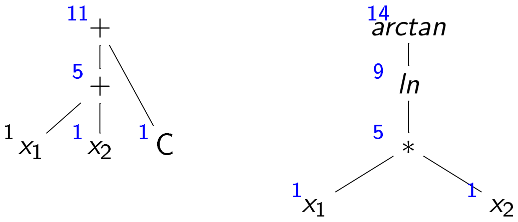
$$r_{12} + r_{23} \geq r_{13},$$

$$p_1 + p_2 - 2p_{12} + p_2 + p_3 - 2p_{23} \geq p_1 + p_3 - 2p_{13},$$

$$p_2 - p_{12} - p_{23} + p_{13} \geq 0.$$

Описание суперпозиции f, представленной в виде дерева Γ, определяется следующими параметрами.

1. Число вершин v_i в дереве Γ.
2. Количество порождающих функций g_i, не являющихся терминальными вершинами x_i.



Сложность суперпозиции

Сложность суперпозиции равна f сложности корня v_0 и определяется индуктивно с помощью следующей процедуры:

1. Сложность C суперпозиции, дерево которой представляет собой одну вершину, соответствующую константной порождающей функции, равна 0:
$$C(\text{const}) = 0.$$
2. Сложность C суперпозиции f, дерево Γ которой представляет собой одну свободную переменную x_i равна 1:
$$C(x_i) = 1.$$
3. Сложность суперпозиции $g_i(f)$, где $C(f) = K$, а g_i - порождающая функция, равна сумме количества элементов в $g_i(f)$ и сложности f:
$$C(g_i(f)) = C(f) + |g_i(f)|$$

Построение всех допустимых суперпозиций

Перед первым шагом построим начальное значения множества моделей $\mathcal{F}_0$:

$$\mathcal{F}_0 = \{X, C\},$$

На каждом шаге для $\mathcal{F}_i$ построим множество U_i из суперпозиций, полученных при применении функций $g_i \in G$ к элементам $f_j \in \mathcal{F}_{i-1}$:

$$U_i = \{g_i(f_{j_1}, \dots f_{j_k}), \mid g_i \in G_u, f \in \mathcal{F}_{i-1}\}.$$

Тогда множество F_i

$$\mathcal{F}_i = \mathcal{F}_{i-1} \cup U_i.$$

Количество возможных суперпозиций

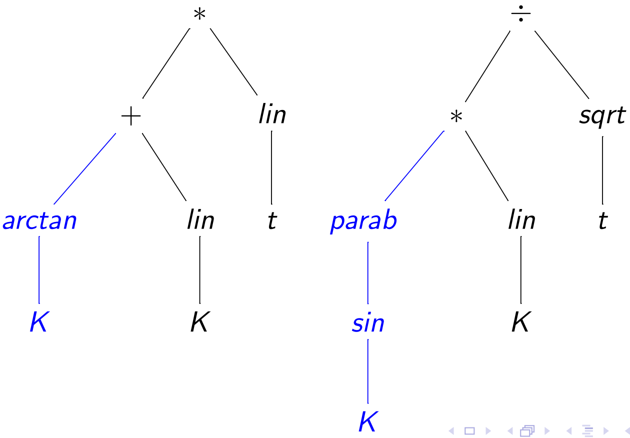
Пусть в множестве порождающих функций G содержится l_p функций максимальной арности $p > 1$, и имеется $n > 1$ независимых переменных.
Справедлива следующая оценка количества суперпозиций, порожденных алгоритмом $\mathfrak{A}$ после k-ой итерации:

$$|\mathcal{F}_k| = \mathcal{O}(l_p^{(p^k-1)/(p-1)} n^{p^k}).$$

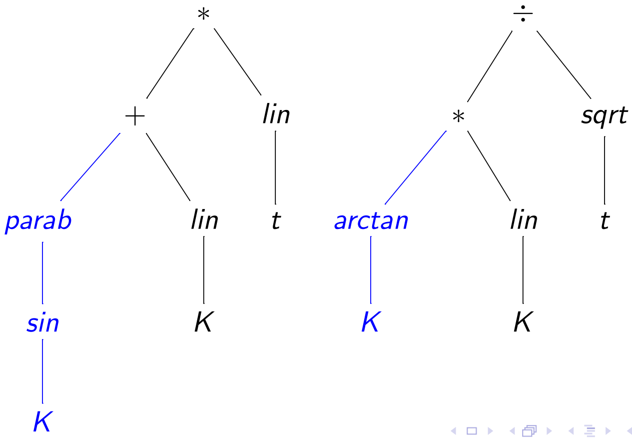
Алгоритм последовательного порождения моделей

0. Экспертно задается множество исходных моделей.
1. Выполняется обмен поддеревьями.
2. Выполняется модфикация суперпозиции
3. В соответствии с критерием качества отбираются лучшие модели и далее происходит следующая итерация процесса.

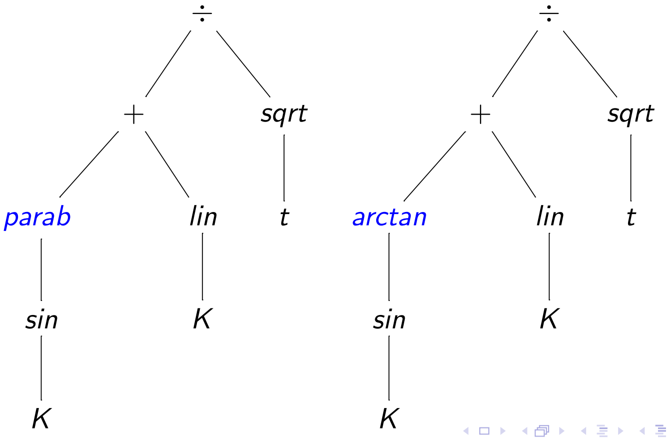
1. Выполняется обмен поддеревьями:
   1. случайно выбирается пара суперпозиций f_i и f_j.
   2. в деревьях Γ_i и Γ_j выбираются вершины v_i и v_j
   3. порождаются новые модели f_i' и f_j' путем обмена поддеревьями с корнями в выбранных вершинах.



1. Выполняется обмен поддеревьями:
   1. случайно выбирается пара суперпозиций f_i и f_j.
   2. в деревьях Γ_i и Γ_j выбираются вершины v_i и v_j
   3. порождаются новые модели f_i' и f_j' путем обмена поддеревьями с корнями в выбранных вершинах.

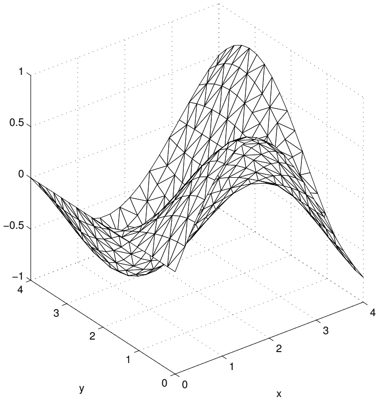


2. Выполняется модификация полученных моделей.
    1. В дереве Γ_j выбирается вершина v_k;
    2. Из элементов G, имеющих число аргументов, как у g_k, выбирается g_s.
    3. Порождающая функция g_k заменяется функцией g_s.



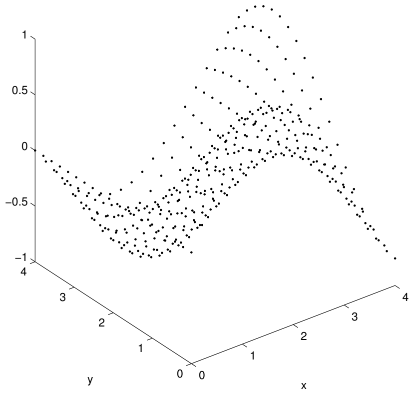
Think of a model

Let it be $y = f(\mathbf{w}, \mathbf{x}) = sin(x_1) * sin(w_1 x_2 + w_2)$.



Given data

The corresponded sample set is shown; it has 380 samples.



Given primitive functions

Function	Description	Parameters
	$g(\mathbf{b}, x_1, x_2)$	
plus	$y = x_1 + x_2$	–
times	$y = x_1 x_2$	–
	$g(\mathbf{b}, x_1)$	
divide	$y = 1/x$	–
multiply	$y = ax$	a
add	$y = x + a$	a
normal	$y = \frac{\lambda}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) + a$	λ, σ, ξ, a
linear	$y = ax + b$	a, b
parabolic	$y = ax^2 + bx + c$	a, b, c
sin	$y = sin(x)$	–
logsig	$y = \frac{\lambda}{1+\exp(-\sigma(x-\xi))} + a$	λ, σ, ξ, a

Set of the generated models

Let the generated models $\mathcal{F} = \{f_i\}$ be a set
of admissible superpositions
of the primitive functions $G = \{g\}$.

Expert information

Experts assign the initial models

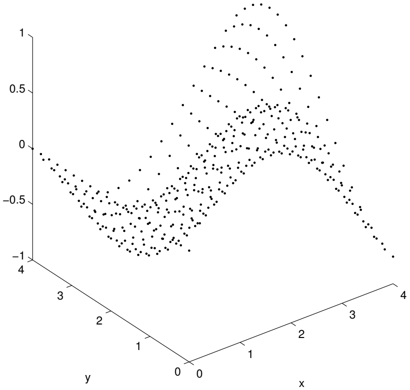
$$
\begin{aligned}
f_1: \quad & y = \text{linear}\,(x_1), \\
f_2: \quad & y = \text{normal}\,(x_2).
\end{aligned}
$$

And the initial conditions

1. the model complexity:
   $$\begin{cases} \text{number of primitives in a superposition } g \text{ no more than } 8, \\ \text{number of parameters } w \text{ no more than } 10; \end{cases}$$
2. the target function is sum of squared errors, SSE.

Competitive models

Given data

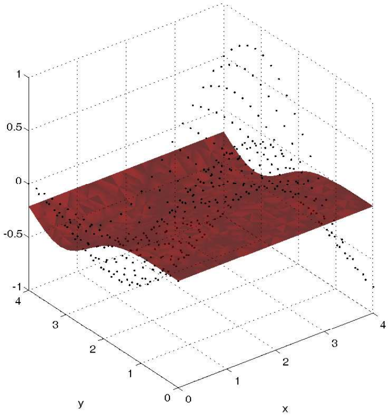


Competitive models

normal($w_{1:3}$,x_2)

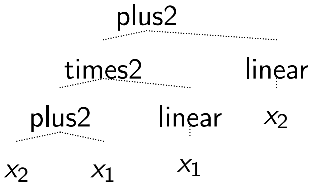
normal

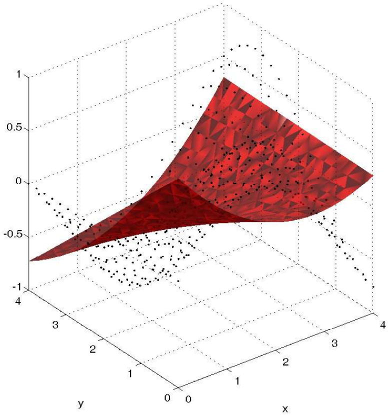
x_2



Competitive models

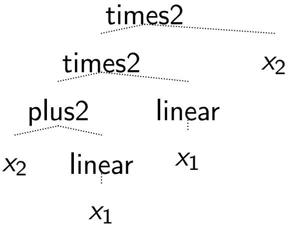
$$\text{plus2}(\emptyset,\text{times2}(\emptyset,\text{plus2}(\emptyset,x_2,x_1),\text{linear}(w_{1:2},x_1)),\text{linear}(w_{3:4},x_2))$$

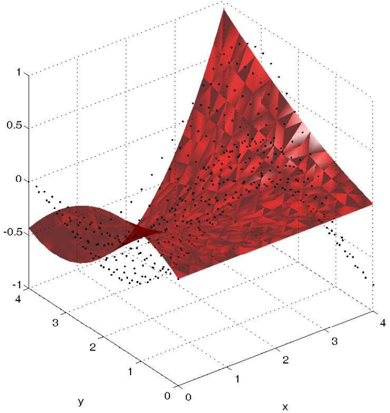




Competitive models

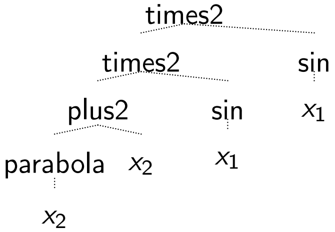
$$\text{times2}(\emptyset,\text{times2}(\emptyset,\text{plus2}(\emptyset,x_2,\text{linear}(w_{1:2},x_1)),\text{linear}(w_{3:4},x_1)),x_2)$$

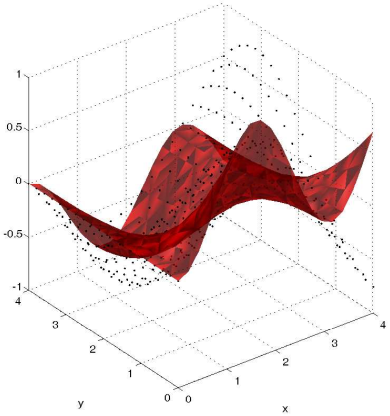




Competitive models

$$\text{times2}(\emptyset,\text{times2}(\emptyset,\text{plus2}(\emptyset,\text{parabola}(w_{1:3},x_2),x_2),\sin(\emptyset,x_1)),\sin(\emptyset,x_1))$$





Competitive models

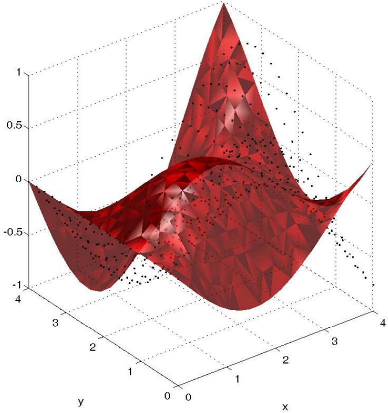
$$\text{times2}(\emptyset,\text{plus2}(\emptyset,\text{linear}(w_{1:2},x_1),\sin(\emptyset,x_2)),\sin(\emptyset,x_1))$$

times2

plus2 sin

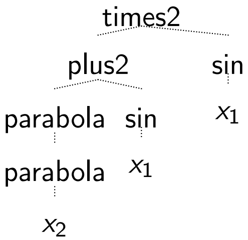
linear sin x_1

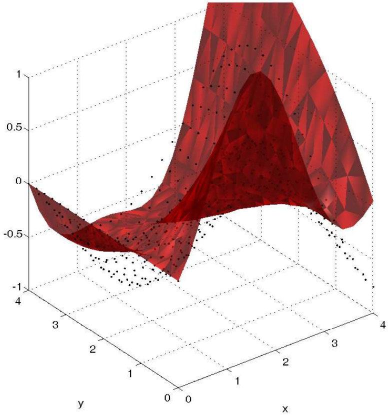
x_1 x_2



Competitive models

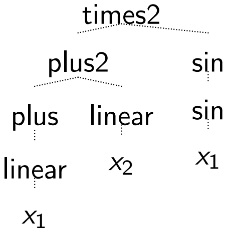
$$\mathrm{times2}(\emptyset,\mathrm{plus2}(\emptyset,\mathrm{parabola}(w_{1:3},\mathrm{parabola}(w_{4:6},x_2)),\sin(\emptyset,x_1)),\sin(\emptyset,x_1))$$

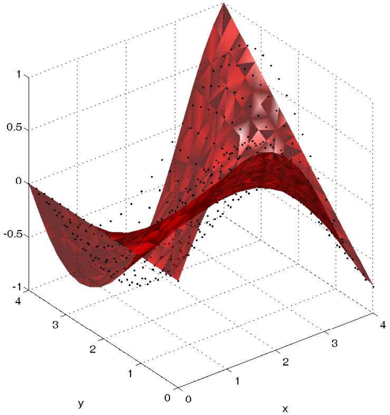




Competitive models

$$\mathrm{times2}(\emptyset,\mathrm{plus2}(\emptyset,\mathrm{plus}(w_1,\mathrm{linear}(w_{2:3},x_1)),\mathrm{linear}(w_{4:5},x_2)),\mathrm{sin}(\emptyset,\mathrm{sin}(\emptyset,x_1)))$$





Competitive models

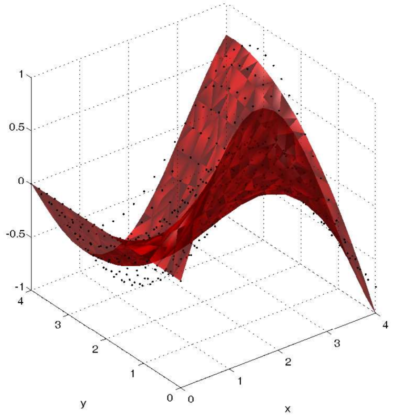
$$\text{times2}(\emptyset,\text{parabola}(w_{1:3},\text{linear}(w_{4:5},x_2)),\text{linear}(w_{6:7},\text{sin}(\emptyset,x_1)))$$

times2

parabola linear

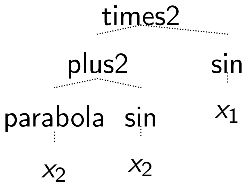
linear sin

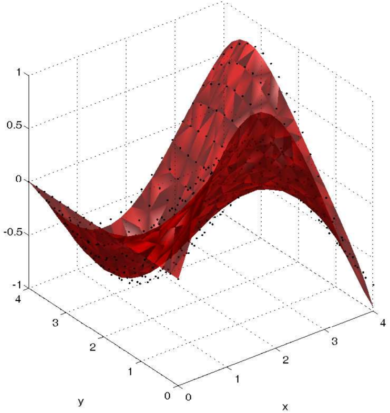
x_2 x_1



Competitive models

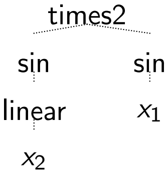
times2($\emptyset$,plus2($\emptyset$,parabola($w_{1:3}$,x_2),sin($\emptyset$,x_2)),sin($\emptyset$,x_1))

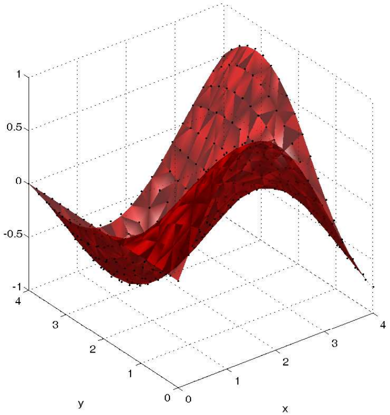




Competitive models

times2($\emptyset$,sin($\emptyset$,linear($w_{1:2}$,x_2)),sin($\emptyset$,x_1))





Процедура построения модели