ПО для порождения регрессионных моделей

Стрижов В. В.

Вычислительный центр РАН

Интеллектуализация обработки информации
9–13 июня 2008 г.

Алушта

Цель исследования

Регрессионная модель (параметрическое семейство функций) может быть построена:

1. на основе знаний о предметной области;
2. на основе измеряемых данных и выбранной универсальной модели;

Требуется построить алгоритм порождающий

1. несложные интерпретируемые модели,
2. использующий данные и экспертные знания о структуре модели.

Программа MVR Composer (Multi Variate Regression)

Синтез и анализ структуры моделей

Использованы работы:

1. Метод группового учета аргументов
   Ивахненко, А. Г. (1985), Malada, H. R. (1994)
2. Генетическое программирование
   Koza, J. R. (2002), Zelinka, I. (2006)
3. Optimal Brain Surgery
   LeCun, Y., Solla, S. A. (1990)
4. Выбор моделей и связанный байесовский вывод
   MacKay, D. (2003), Nabney, J. (2004)

Дано

Регрессионная выборка:

$\{\mathbf{x}_1, ..., \mathbf{x}_N | \mathbf{x} \in \mathbb{R}^P\}$ — независимые переменные,

$\{y_1, ..., y_N | y \in \mathbb{R}\}$ — зависимые переменные,

$$D = \{(\mathbf{x}_i, y_i)\}.$$

Порождающие функции:

$G = \{g | g : \mathbb{R} \times ... \times \mathbb{R} \longrightarrow \mathbb{R}\}$,

$g = g(\mathbf{b}, \cdot, \cdot, ..., \cdot)$.

G индуктивно задает $\mathcal{F} = \{f_i\}$ —
множество допустимых суперпозиций элементов g,

$f_i = f_i(\mathbf{w}, \mathbf{x})$,

где $\mathbf{w} = \mathbf{b}_1 \vdots \mathbf{b}_2 \vdots ... \vdots \mathbf{b}_r$.

Начальные модели:

$\{f\} \in \mathcal{F}$.

Найти

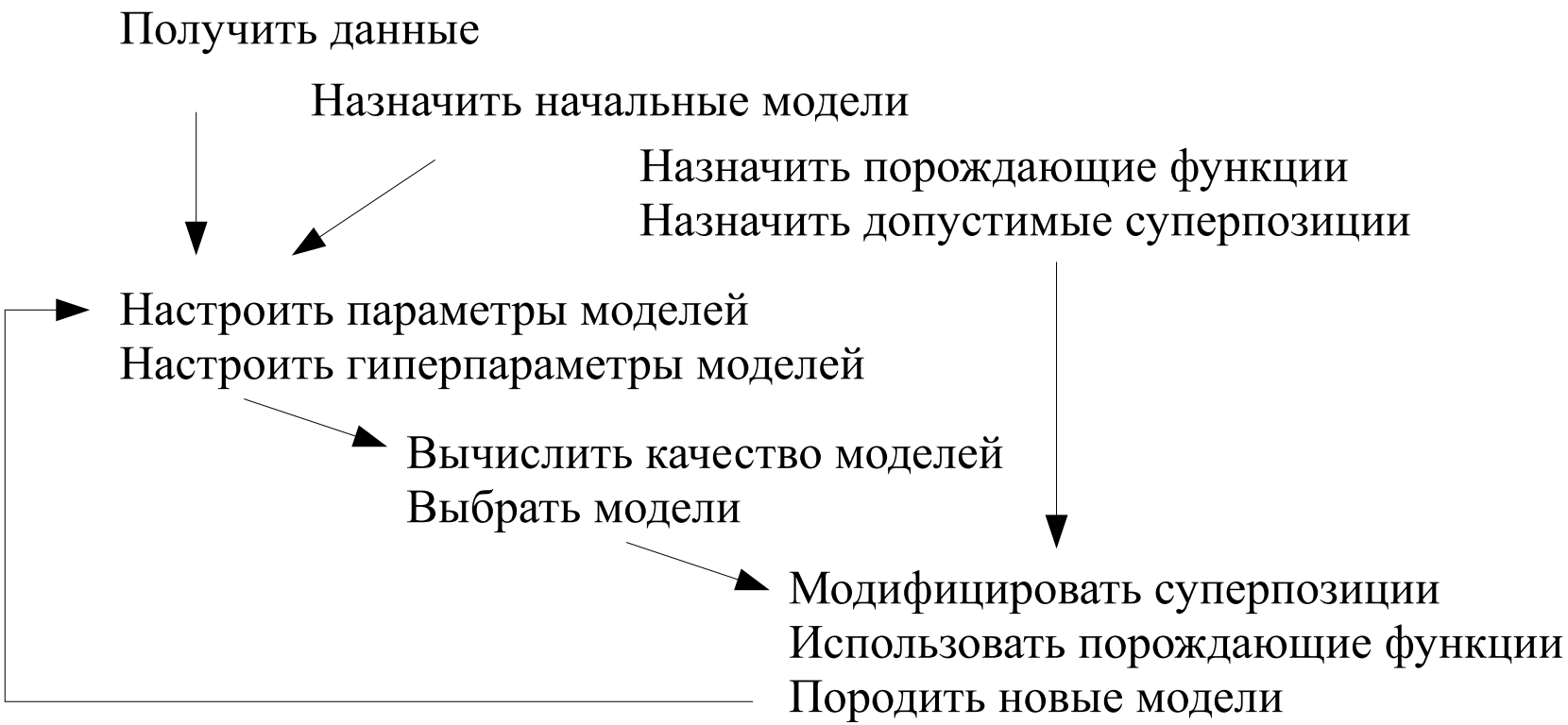
$$y = f_i(\mathbf{w}, \mathbf{x}) + \nu$$

модель оптимальной структуры $f_i \in \mathcal{F}$, доставляющую максимум заданному критерию $P(\mathbf{w}|D, \alpha, \beta, f_i)$.

Частный случай, целевая функция SSE:

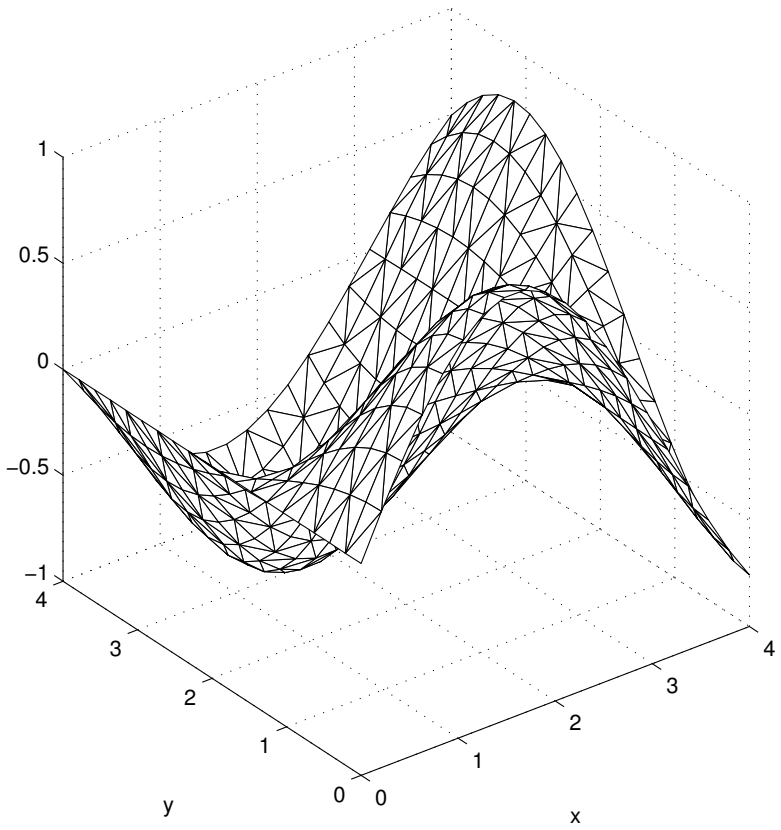
$$S = \beta E_D + \alpha E_W.$$

Структурная схема алгоритма



Модельная задача

Искомая модель: $y = f(\mathbf{w}, \mathbf{x}) = sin(x_1) * sin(w_1 x_2 + w_2)$



Регрессионая выборка, 380 точек: $\mathbb{R}^2 \ni x_i \longrightarrow y_i \in \mathbb{R}$

Порождающие функции

Функция	Описание	Параметры
	$g(\mathbf{b}, x_1, \ldots, x_r)$ — один и более аргументов	
plus2	$y = x_1 + x_2$	–
times2	$y = x_1 \times x_2$	–
	$g(\mathbf{b}, x_1)$ — один аргумент	
invert	$y = 1/x$	–
negate	$y = -x$	–
multiply	$y = ax$	a
add	$y = x + a$	a
gaussian	$y = \frac{\lambda}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) + a$	λ, σ, ξ, a
linear	$y = ax + b$	a, b
parabolic	$y = ax^2 + bx + c$	a, b, c
cubic	$y = ax^3 + bx^2 + cx + d$	a, b, c, d
logsig	$y = \frac{\lambda}{1+\exp(-\sigma(x-\xi))} + a$	λ, σ, ξ, a

Модели начального приближения

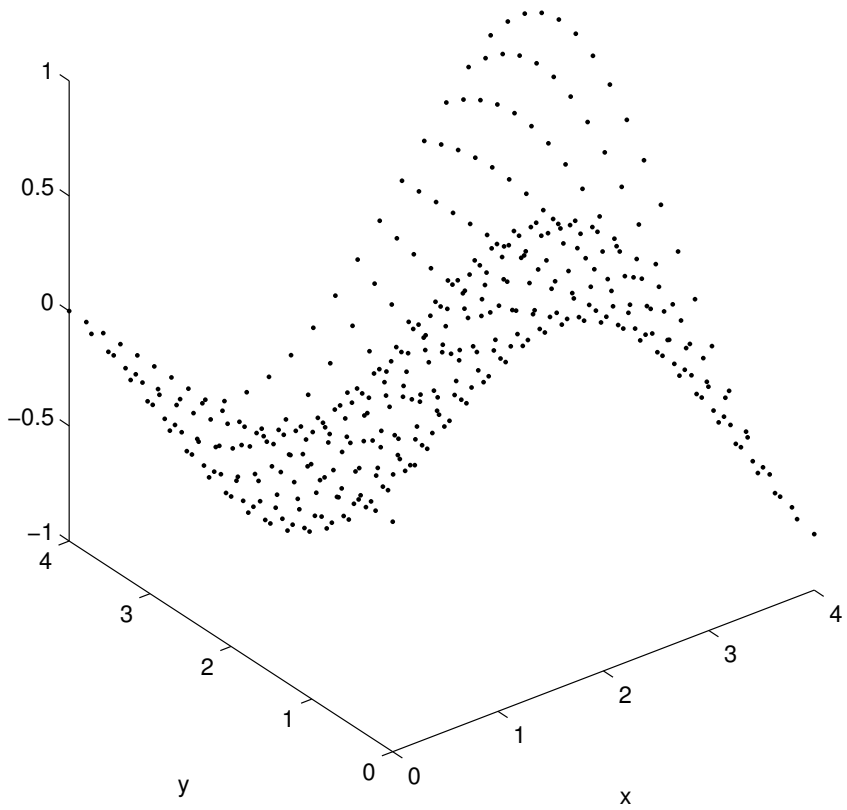
$$f_1: \quad y = \text{linear}(x_1),$$
$$f_2: \quad y = \text{normal}(x_2).$$

Условия порождения модели:

1. сложность моделей: $\begin{cases} \text{число элементов } g \text{ не более 8,} \\ \text{число параметров } w \text{ не более 10;} \end{cases}$
2. целевая функция — сумма квадратов остатков, SSE;
3. выборка на обучение-контроль не разделяется.

Модели-претенденты

Исходные данные

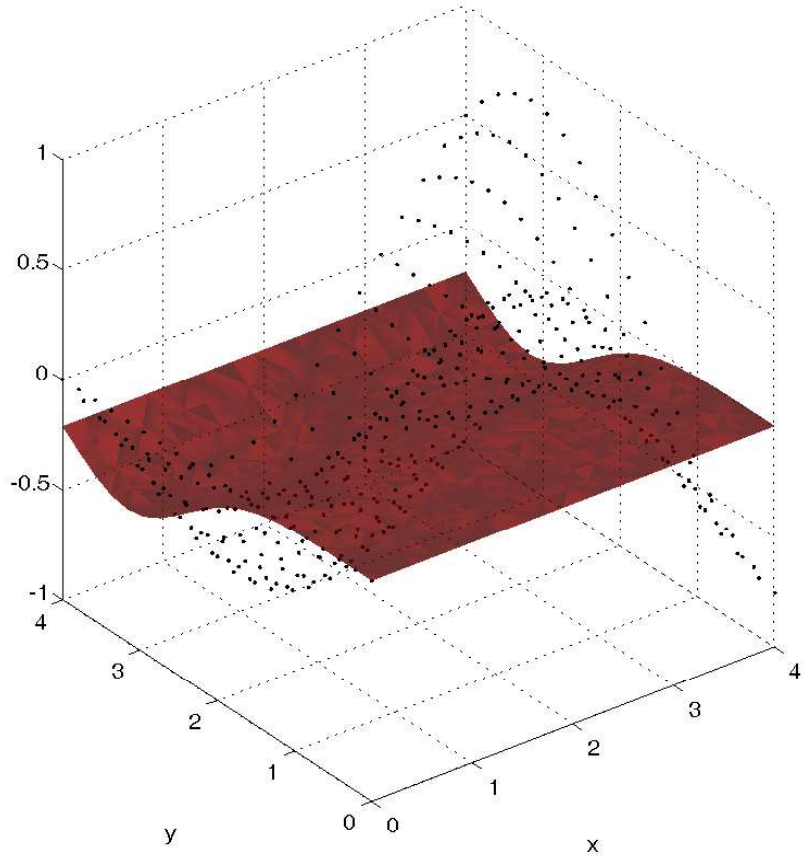


Модели-претенденты

normal(w(1:3),x(:,2))

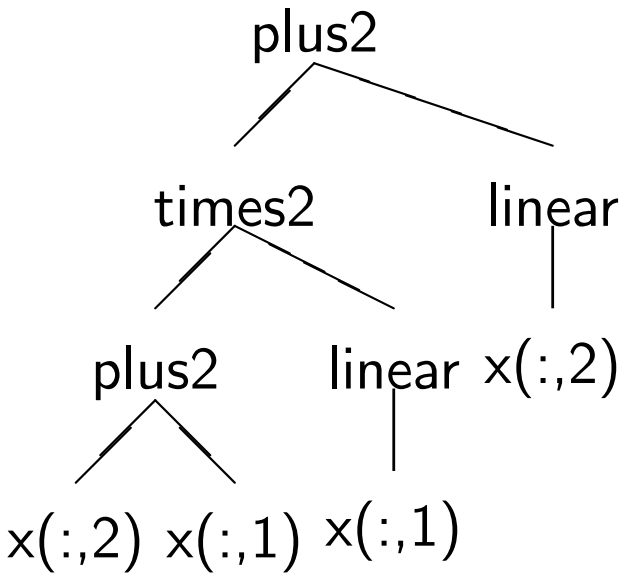
normal

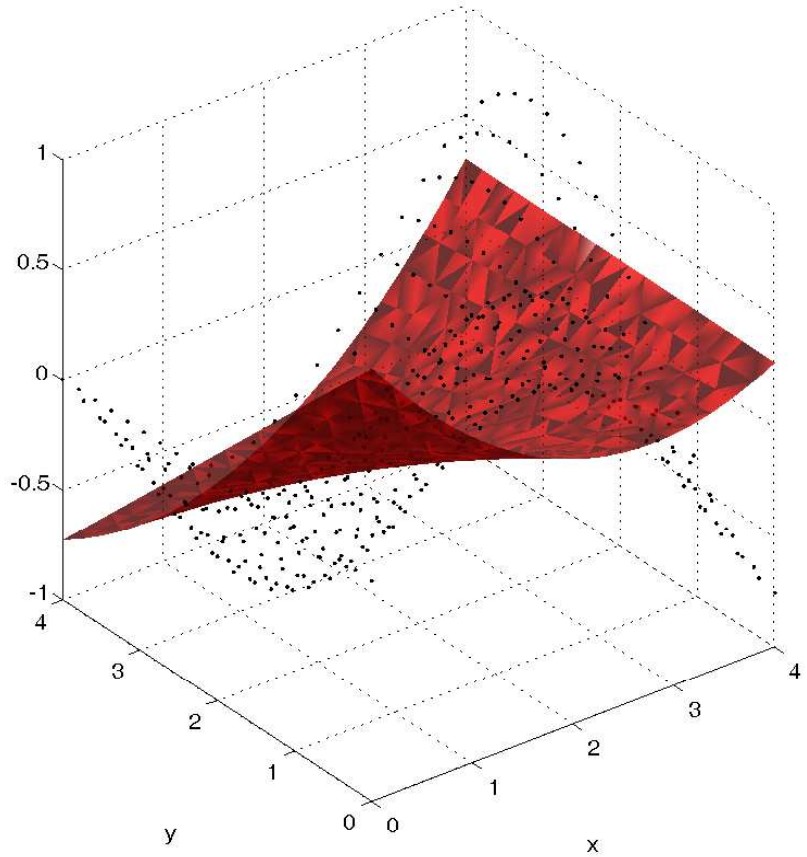
x(:,2)



Модели-претенденты

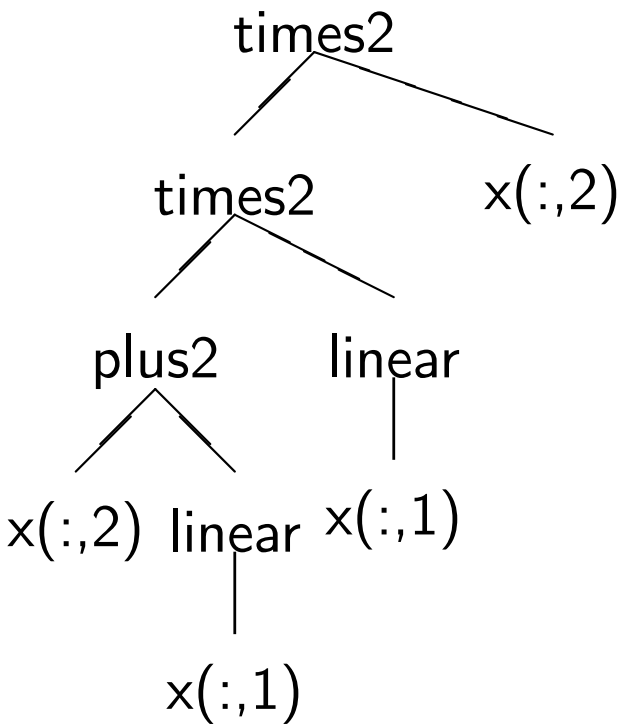
plus2([],times2([],plus2([],x(:,2),x(:,1)),linear(w(1:2),x(:,1))),linear(w(3:4),

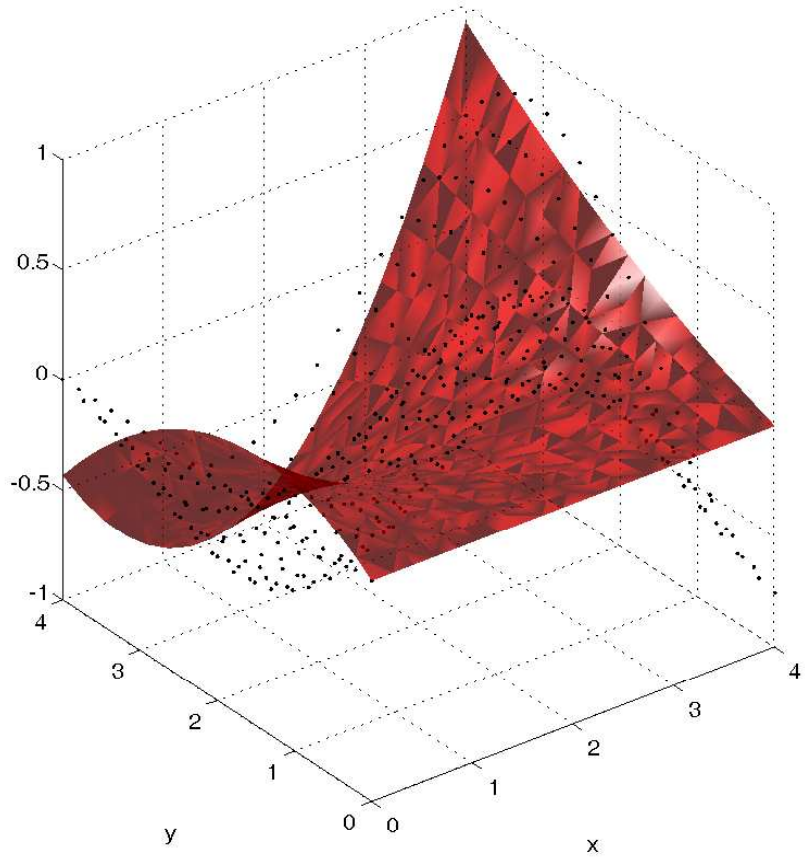




Модели-претенденты

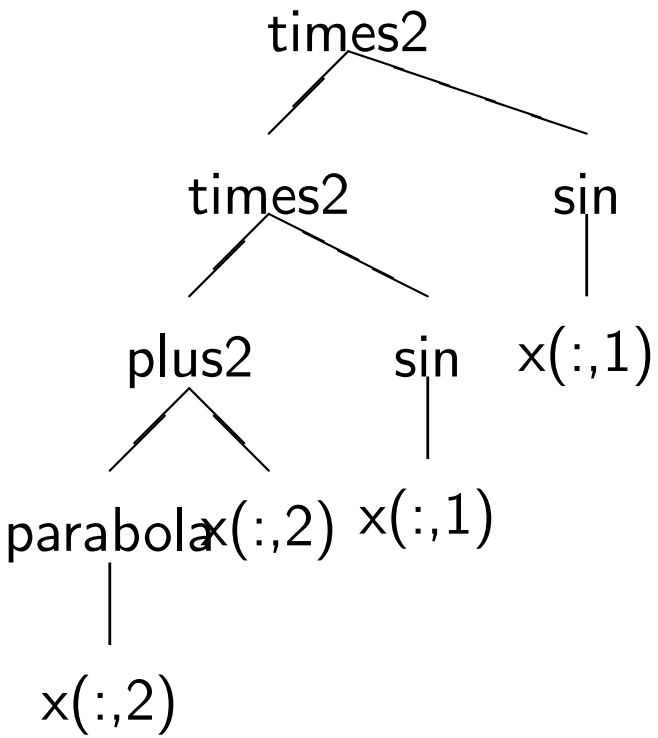
times2([],times2([],plus2([],x(:,2),linear(w(1:2),x(:,1))),linear(w(3:4),x(:,1))

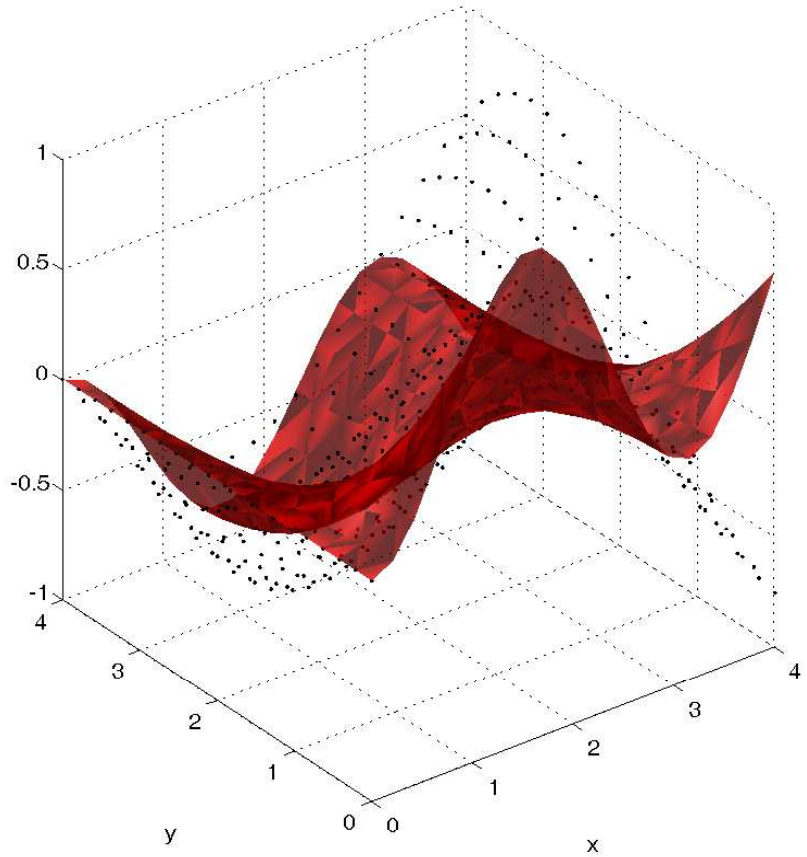




Модели-претенденты

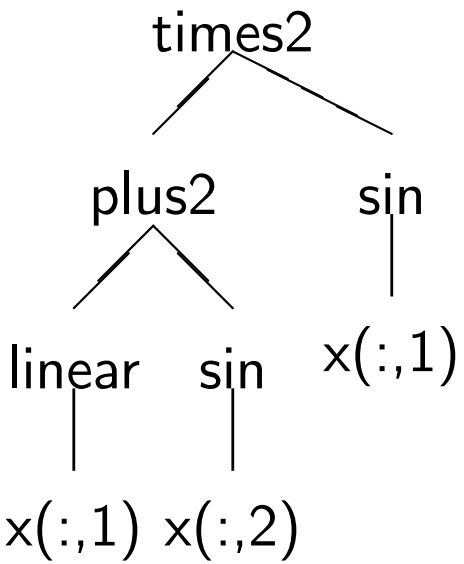
times2([],times2([],plus2([],parabola(w(1:3),x(:,2)),x(:,2)),sin([],x(:,1))),sin

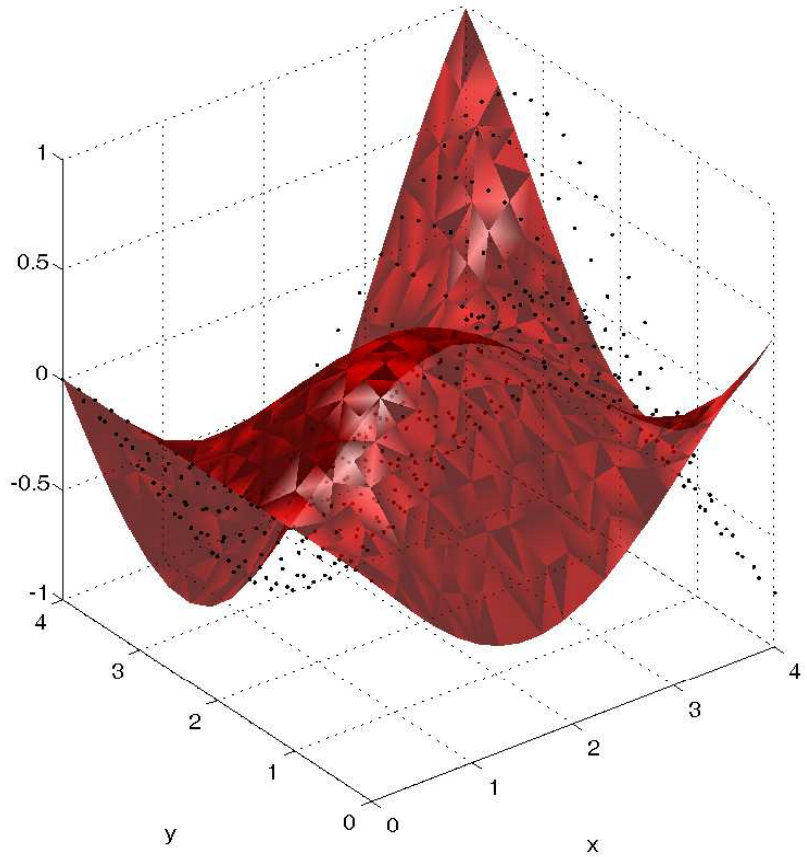




Модели-претенденты

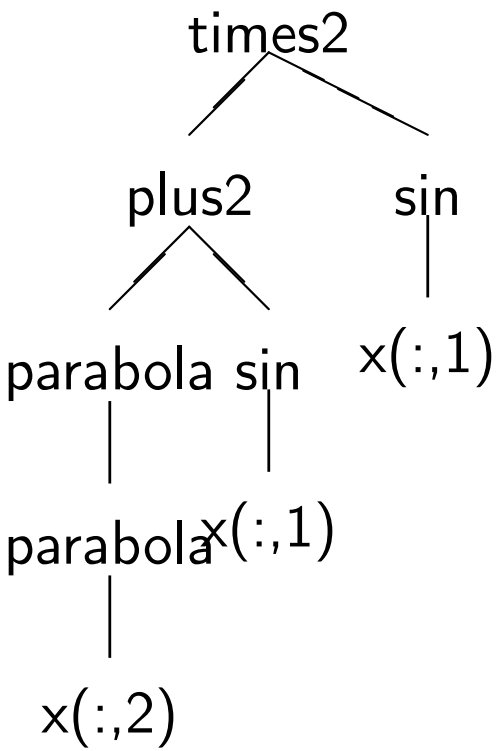
times2([],plus2([],linear(w(1:2),x(:,1)),sin([],x(:,2))),sin([],x(:,1)))

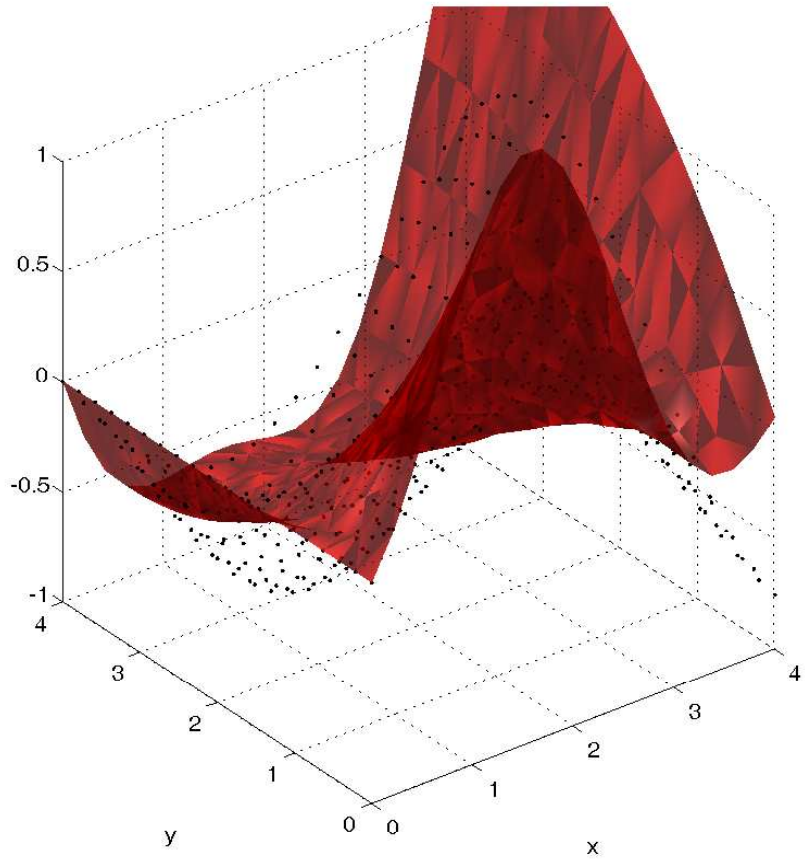




Модели-претенденты

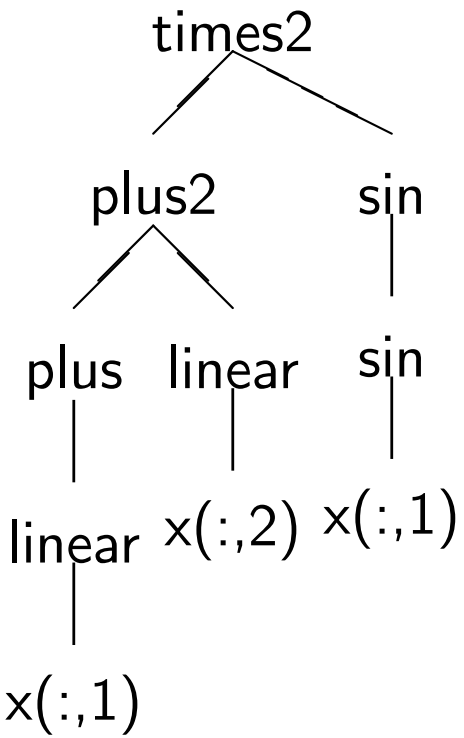
times2([],plus2([],parabola(w(1:3),parabola(w(4:6),x(:,2))),sin([],x(:,1))),si

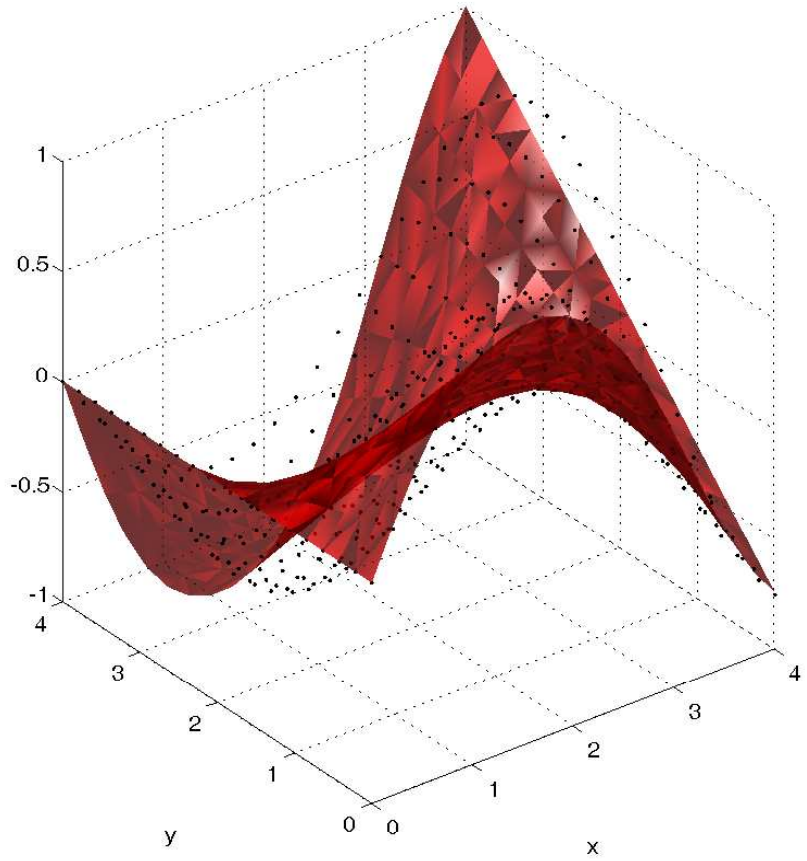




Модели-претенденты

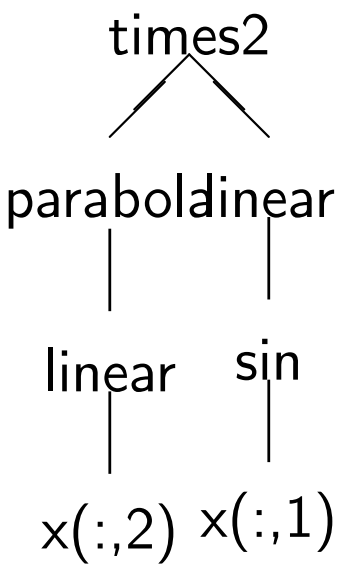
times2([],plus2([],plus(w(1),linear(w(2:3),x(:,1))),linear(w(4:5),x(:,2))),sin(

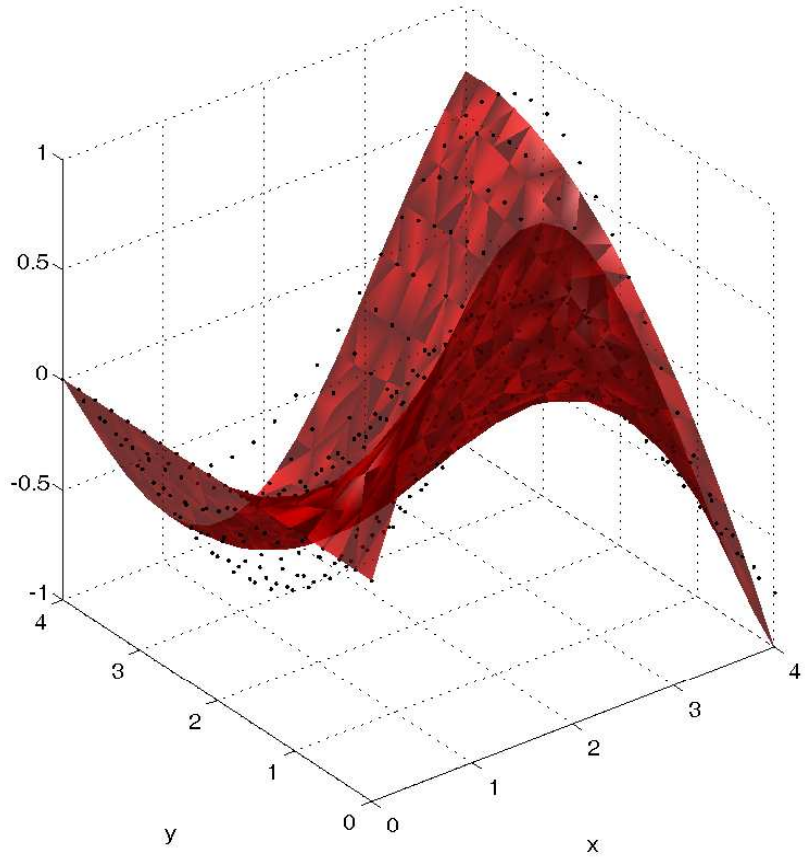




Модели-претенденты

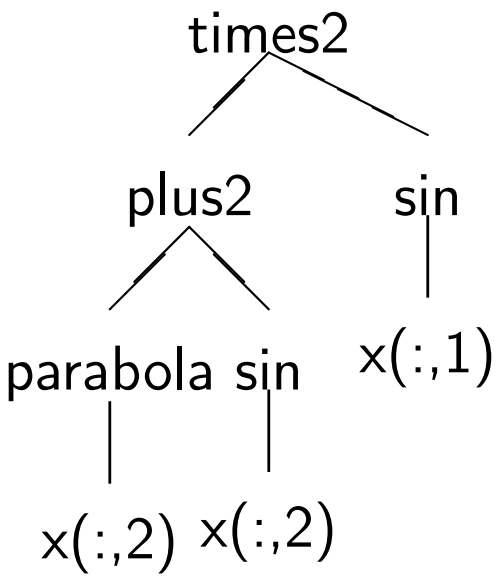
times2([],parabola(w(1:3),linear(w(4:5),x(:,2))),linear(w(6:7),sin([],x(:,1))))

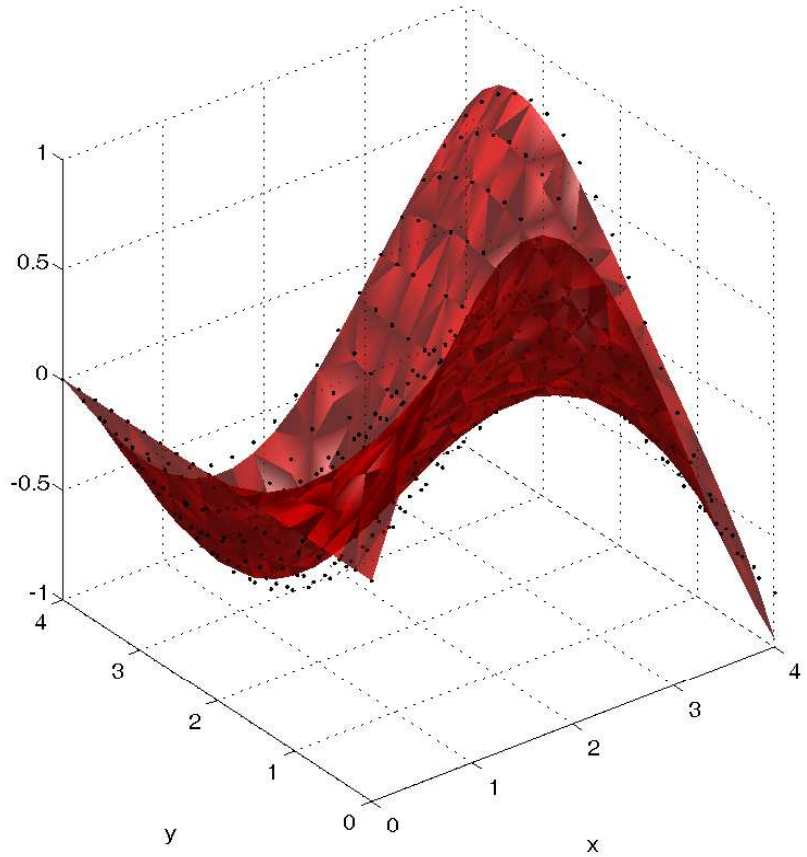




Модели-претенденты

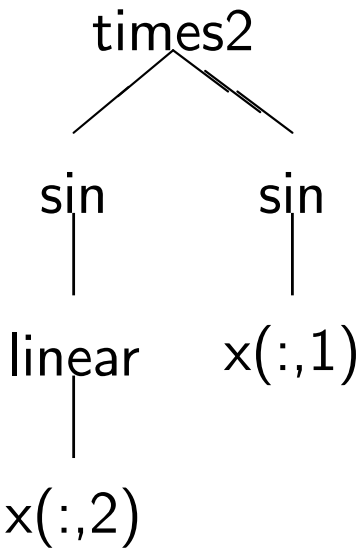
times2([],plus2([],parabola(w(1:3),x(:,2)),sin([],x(:,2))),sin([],x(:,1)))

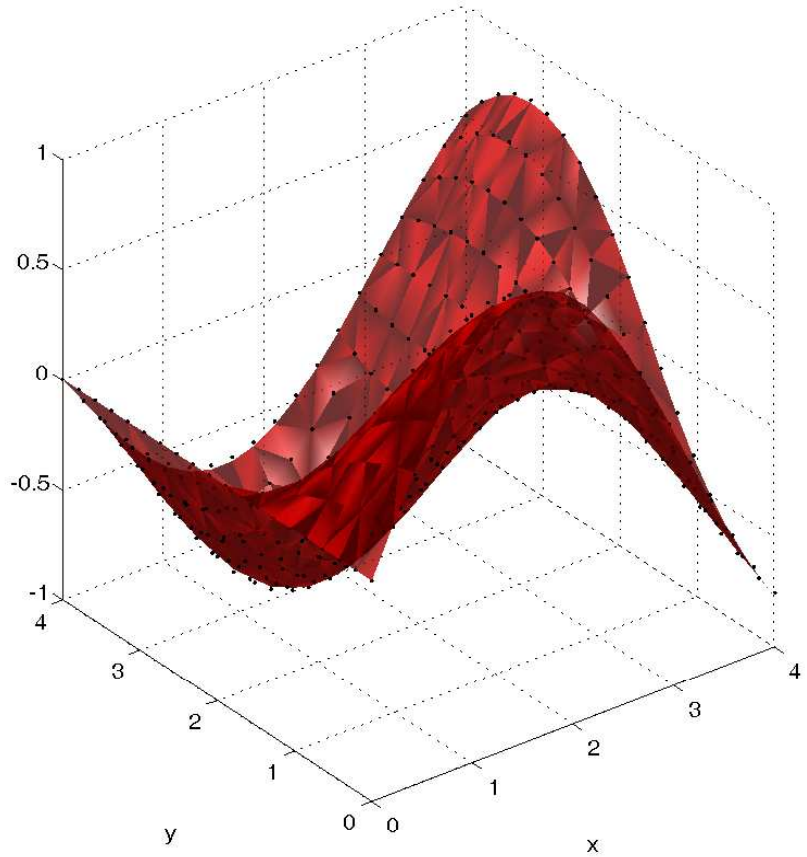




Модели-претенденты

times2([],sin([],linear(w(1:2),x(:,2))),sin([],x(:,1)))





Конструктивное порождение всех допустимых суперпозиций

Функции $g_v \in G$ проиндексированы $v \in \mathcal{V} = \{1, ..., V\}$.

Отображение $\iota : \mathcal{V} \to \mathcal{A}$ — всевозможные сочетания с повторениями из V по K, где $K = 1, \ldots, R$ и

$$|\mathcal{A}| = \sum_{K=1}^{R} \bar{C}_K^V.$$

$\mathcal{A} \ni A_\iota$ — вектор с элементами $a_\iota(k)$ из $\mathcal{V}$;

на $A_\iota \times A_\iota$ задан набор матриц инцидентности $\{\rho_i(A_\iota)\}$, где индекс i матрицы ρ определяет допустимую суперпозицию f_i функций $g \in G$.

Суперпозиция f_i называется допустимой, если

1. число аргументов каждого элемента суперпозиции должно совпадать с числом аргументов соответствующей порождающей функции

$$\sum_{l=1}^{K_\iota} \rho_i(l,k) = s(a_\iota(k)), \quad \text{для всех} \quad k = 1, \ldots, K_\iota$$

(число вершин графа, смежных вершине с номером k, есть число $s(a_\iota(k))$ аргументов функции g_v при $v = a_\iota(k)$);

2. число единиц в матрице ρ_i равно суммарному числу аргументов f_i

$$\sum_{k=1}^{K_\iota} \sum_{l=1}^{K_\iota} \rho_i(l,k) = \sum_{k=1}^{K_\iota} s(a_\iota(k)),$$

где $s = s(v)$ — число аргументов функции g_v;

3. граф ρ_i является ациклическим.

Специализация: Нелинейные модели

Задано множество порождающих функций $\{g_1, \ldots, g_V\}$ — непрерывно-дифференцируемых; задано ограничение на число элементов суперпозиции R.
Модели — множество всех допустимых суперпозиций.

Пример:

$G = \{\mathrm{id}(x), \sin(x), \mathrm{times}(x, x')\}$, $R = 2$;
Модели:

$$\begin{aligned} f_1 &= x, \\ f_2 &= x^2, \\ f_3 &= x\sin(x), \\ f_4 &= \sin(\sin(x)), \\ f_5 &= \sin^2(x), \\ f_6 &= \sin(x). \end{aligned}$$

Специализация: Линейные модели и МГУА

$$y = w_0 + \sum_{i=1}^{V} w_i g_i + \sum_{i=1}^{V}\sum_{j=1}^{V} w_{ij} g_i g_j + \sum_{i=1}^{V}\sum_{j=1}^{V}\sum_{k=1}^{V} w_{ijk} g_i g_j g_k + \dots$$

Задано множество $\{g_1, \dots, g_V\}$ функций, не имеющих параметров; зафиксирована структура суперпозиции (2 монома многорядного МГУА)

$$f = wsum(prod(g_{v_1}, \dots, g_{v_i}), prod(g_{v_j}, \dots, g_{v_K})).$$

Пример:

$G = \{\mathrm{id}(x), \sin(x)\}$, $R = 7$;

Модели:

$$\begin{aligned} f_1 &= w_0 + w_1 x + w_2 x^2, \\ f_2 &= w_0 + w_1 x + w_2 \sin(x), \\ f_3 &= w_0 + w_1 x + w_2 x \sin(x), \\ &\dots \\ f_{10} &= w_0 + w_1 x^2 + w_2 \sin^2(x). \end{aligned}$$

Специализация: RBF и кусочно-линейные модели

$$y = \sum_{i=1}^{N} w_i \varphi(\|\mathbf{x} - c_i\|), \qquad y = \sum_{i=1}^{N} (a_i + \mathbf{b}_i \cdot (\mathbf{x} - \mathbf{c}_i)) \varphi(\|\mathbf{x} - c_i\|),$$

где $\varphi(\mathbf{x}, \mathbf{c}) = \varphi(\|\mathbf{x}, \mathbf{c}\|)$.
Задано множество $\{g_1, \ldots, g_V\}$ потенциальных функций с параметром $\mathbf{c}$ (и, возможно, дополнительными параметрами), зафиксирована структура суперпозиции

$$f = rbfsum(g_v), \qquad f = rbflin(g_v).$$

Пример:

$G = \{g_1 = \exp(-\beta r^2), g_2 = \sqrt{r^2 + \beta^2}, g_3 = r^2 \log(r)\}$, где $r = \|\mathbf{x} - \mathbf{c}\|_2$, $R = N + 1 = 3$;
Модели:

$$\begin{aligned} f_1 &= w_1 g_1(\mathbf{x}) + w_2 g_1(\mathbf{x}), \\ f_2 &= w_1 g_2(\mathbf{x}) + w_2 g_2(\mathbf{x}), \\ f_3 &= w_1 g_3(\mathbf{x}) + w_2 g_3(\mathbf{x}). \end{aligned}$$

Специализация: Векторные авторегрессионные модели

Прогноз временных рядов — частный случай одномерной регрессии, $f : t \mapsto y$, $t \in \mathbb{N}^1, y \in \mathbb{R}^1$,

$$y = \mathbf{a}_0 + \sum_{\tau=1}^{T} (\mathbf{w}_\tau, \mathbf{g}(\mathbf{x}_{t-\tau})) = \sum_{\tau=1}^{T} \sum_{i=1}^{P} \left(w_{\tau,i}, g_i(x_{(t-\tau),i})\right),$$

где $\mathbf{x}_t$ значения набора временных рядов в момент t.
Зафиксирована структура

$$y = lagsum(g_{v_1}, \ldots, g_{v_P}).$$

Пример:

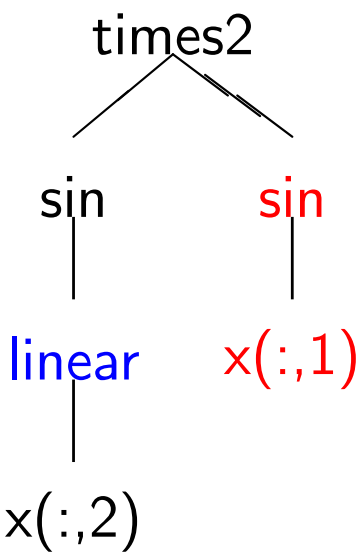
$G = \{x(t), \exp(x), \log(x)\}$, R = P + 1 = 3; T = 2;
Модели:

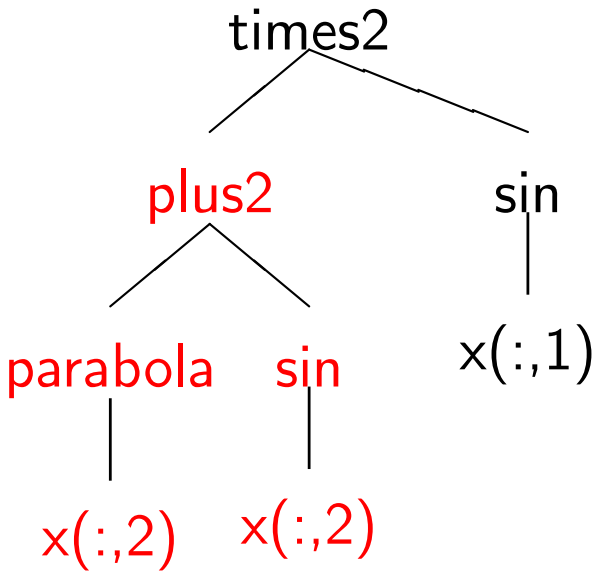
$$\begin{aligned} f_1 = &\quad w_1 x(t-1) + w_2 x(t-2), \\ f_2 = &\quad w_1 exp(x(t-1)) + w_2 exp(x(t-2)), \\ f_3 = &\quad w_1 log(x(t-1)) + w_2 log(x(t-2)). \end{aligned}$$

Заключение

- Разработана программа порождения линейных и нелинейных регрессионных моделей MVR Composer, язык Matlab
- Открытый исходный код: http://sourceforge/projects/mvr.net или http://strijov.com
- Описание теории: http://machinelearning.ru, раздел "Регрессионный анализ"

Модификация суперпозиций





Изоморфные модели

- Коммутирующие аргументы:

$$g(y, x) \mapsto g(x, y).$$

- Приведение к каноническому виду:

$$\frac{\sin(x)}{\cos(x)} \mapsto \tan(x).$$

- Канонические исключения:

$$x + x \mapsto 2 \times x, \quad \text{где } \times \in G, \quad \text{но } const = 2 \notin G.$$

- Неприводимость к каноническому виду при (фактическом) равенстве:

$$f_{ver1} ?\mapsto f_{ver2}, \text{ при } f_{ver1}(w, x) = f_{ver2}(w, x),$$

$$\text{для всех } w, x \in \mathrm{dom}(f_{ver1}) = \mathrm{dom}(f_{ver1}).$$

Алгоритмы оптимизации

- Оптимизация параметров:
  - все модели — генетические оптимизационные алгоритмы;
  - в том числе нелинейные, непрерывно-дифференцируемые — метод сопряженных градиентов или алгоритм Левенберга-Марквардта (при условии целевой функции SSE);
  - в том числе линейные — метод наименьших квадратов (при условии целевой функции SSE).
- Оптимизация структуры:
  - полный перебор моделей начиная от простейших;
  - многорядный направленный поиск МГУА;
  - генетическое программирование.