Машинное обучение и моделирование экспериментальных данных

д.ф.-м.н. К. В. Воронцов (voron@forecsys.ru)
к.ф.-м.н. В. В. Стрижов (http://www.strijov.com)

отдел «Интеллектуальные системы» ВЦ РАН
каф. «Интеллектуальные системы» ФУПМ МФТИ
каф. «Математические методы прогнозирования» ВМК МГУ
ЗАО «Форексис»

18 февраля 2012

Содержание

Задача обучения по прецедентам (восстановления зависимости по эмпирическим данным)

$\mathbb{X}$ — множество объектов; $\mathbb{Y}$ — множество ответов;
$\exists y: \mathbb{X} \to \mathbb{Y}$ — неизвестная зависимость.

Дано: $D = (x_n, y_n)_{n=1}^N$ — *обучающая выборка*, $y_n = y(x_n)$

$$\begin{pmatrix} f_1(x_1) & \ldots & f_P(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_N) & \ldots & f_P(x_N) \end{pmatrix} \to \begin{pmatrix} y_1 \\ \ldots \\ y_N \end{pmatrix}$$

Модель зависимости — семейство функций $\mathcal{F} = \{f: \mathbb{X} \to \mathbb{Y}\}$

Найти: функцию $f \in \mathcal{F}$, приближающую $y(x)$ на всём $\mathbb{X}$
и сделать прогнозы для контрольной выборки $D' = (x_n)_{n=1}^K$:

$$\begin{pmatrix} f_1(x'_1) & \ldots & f_P(x'_1) \\ \ldots & \ldots & \ldots \\ f_1(x'_K) & \ldots & f_P(x'_K) \end{pmatrix} \to \begin{pmatrix} ?_1 \\ \ldots \\ ?_K \end{pmatrix}$$

Примеры прикладных задач обучения по прецедентам

- Распознавание, классификация, принятие решений ($|\mathbb{Y}| < \infty$):
  - x — пациент; y — диагноз, рекомендуемая терапия;
  - x — заёмщик; y — вероятность дефолта;
  - x — абонент; y — вероятность ухода к другому оператору;
  - x — текстовое сообщение; y — спам / не спам;
  - x — документ; y — категория в рубрикаторе;
  - x — фрагмент белка; y — тип вторичной структуры;
  - x — фрагмент ДНК; y — функция: промотор / ген;
  - x — фотопортрет; y — идентификатор личности;
- Регрессия и прогнозирование ($\mathbb{Y} = \mathbb{R}$ или $\mathbb{R}^m$):
  - x — история продаж; y — прогноз объёма продаж;
  - x — пара $\langle$клиент,товар$\rangle$; y — рейтинг товара;
  - x — параметры технолог. процесса; y — свойство продукции;
  - x — структура хим. соединения; y — его свойство;
  - x — характеристики недвижимости; y — цена;

Особенности реальных задач обучения по прецедентам

- Особенности исходных данных:
  - неполнота данных (пропуски);
  - неточность данных (погрешности, выбросы);
  - разнородность (сложные «сырые» данные);
  - несбалансированность классов;
  - малые выборки;
  - сверхбольшие выборки;
  - потоковые данные;
  - нестандартные критерии качества;
  - наличие дополнительной непрецедентной информации.
- Требования к методам восстановления зависимостей:
  - обобщающая способность;
  - вычислительная эффективность;
  - простота и интерпретируемость модели;
  - визуализация и контроль промежуточных данных;
  - динамическое дообучение по потоковым данным;

Важные классы методов (типизация условная и неполная)

- Методы на основе функций сходства объектов
  - kNN — метод ближайших соседей
  - RBF — метод потенциальных функций
- Минимизация и регуляризация эмпирического риска
  - SVM — метод опорных векторов;
  - RLR — логистическая регрессия с регуляризацией;
- Логические методы, индукция правил
  - CART, C5.0, ADT, ODT — решающие деревья;
  - КОРА, ТЕМП, ТЕСТ — комбинаторно-логические алгоритмы;
- Композиции
  - ANN — искусственные нейронные сети;
  - Boosting, MatrixNet — взвешенное голосование;
  - MVR Composer — индуктивное порождение моделей;
- Отбор признаков и понижение размерности
  - МГУА — метод группового учёта аргументов;
  - PCL — метод главных компонент;
  - NMF — неотрицательные матричные разложения;

Задача построения разделяющей поверхности

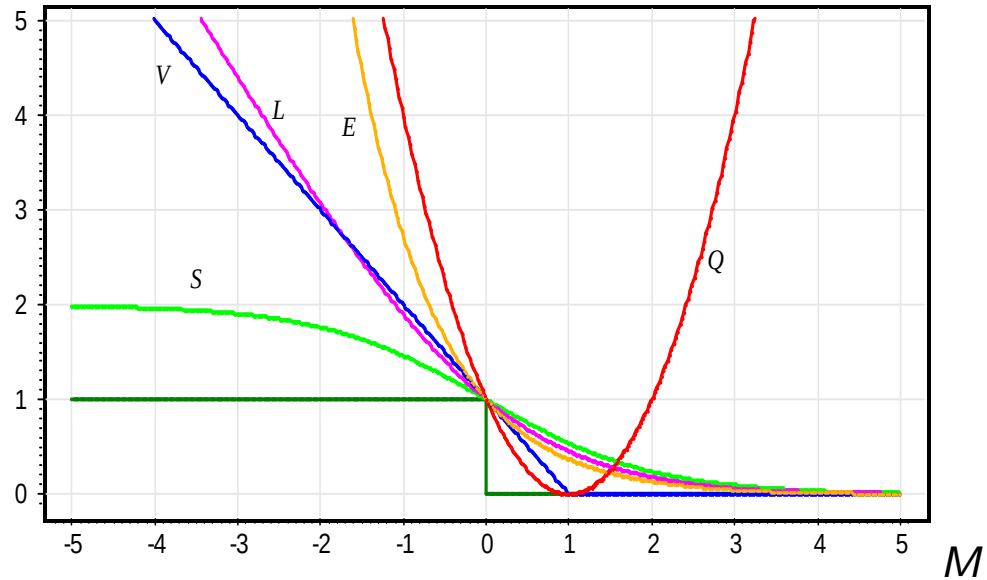
- Задача классификации с двумя классами, $\mathbb{Y} = \{-1, +1\}$:
  $f(w, x) = \operatorname{sign} g(w, x)$, где
  $g(w, x)$ — *дискриминантная функция*,
  w — вектор параметров.
- $g(w, x) = 0$ — разделяющая поверхность;
  $M_n(w) = y_n g(w, x_n)$ — *отступ* (margin) объекта x_n;
  $M_n(w) < 0 \iff$ классификатор $f(w, x)$ ошибается на x_n.
  Принцип оптимальности: чем больше отступ, тем лучше.
- *Минимизация сглаженного эмпирического риска*:

$$Q(w) = \sum_{n=1}^{N} \big[M_n(w) < 0\big] \quad \leqslant \quad \widetilde{Q}(w) = \sum_{n=1}^{N} \mathscr{L}\big(M_n(w)\big) \to \min_w;$$

функция потерь $\mathscr{L}(M)$ невозрастающая, неотрицательная.

Непрерывные аппроксимации пороговой функции потерь

Часто используемые функции потерь $\mathscr{L}(M)$:



$Q(M) = (1-M)^2$ — квадратичная (ЛДФ);
$V(M) = (1-M)_+$ — кусочно-линейная (SVM);
$S(M) = 2(1+e^M)^{-1}$ — сигмоидная (нейронные сети);
$L(M) = \log_2(1+e^{-M})$ — логарифмическая (LR);
$E(M) = e^{-M}$ — экспоненциальная (AdaBoost).

Метод опорных векторов (Support Vector Machine, SVM)

Модель — линейный классификатор:

$$f(x) = \text{sign}\big(\langle w, x\rangle - w_0\big), \quad w, x \in \mathbb{R}^P,\ w_0 \in \mathbb{R}.$$

Задача максимизации зазора между классами приводит к *регуляризации* сглаженного эмпирического риска:

$$Q(w, w_0) = \sum_{n=1}^{N} \big(1 - M_n(w, w_0)\big)_+ + \frac{1}{2C}\|w\|^2 \ \to \ \min_{w, w_0}.$$

Свойства решения этой задачи:

- Максимизация зазора + регуляризация вместе повышают обобщающую способность
- *разреженность*: w зависит только от *опорных объектов* x_n;
- $f(x)$ зависит только от $\langle x, x_n\rangle \Rightarrow$ замена $\langle x, x_n\rangle$ на любое неотрицательно определённое *ядро* $K(x, x_n)$ приводит к нелинейному обобщению SVM.

Обобщение: байесовская регуляризация

$p(x, y|w)$ — вероятностная модель данных;
$p(w; \gamma)$ — априорное распределение параметров модели;
γ — вектор *гиперпараметров*;

Теперь не только появление выборки D,
но и появление модели w также полагается случайным.

Совместное правдоподобие данных и модели:

$$p(D, w) = p(D|w)\, p(w; \gamma).$$

Принцип максимума совместного правдоподобия:

$$L(w, D) = \ln p(D, w) = \sum_{n=1}^{N} \ln p(x_n, y_n|w) + \underbrace{\ln p(w; \gamma)}_{\text{регуляризатор}} \to \max_{w,\gamma}.$$

Композиции алгоритмов классификации

$b_t : \mathbb{X} \to R$ — базовые классификаторы;
$F : R^T \to \mathbb{Y}$ — корректирующая операция;
$a(x) = F\big(b_1(x), \ldots, b_T(x)\big)$ — композиция.

Пример: *взвешенное голосование* (для случая $\mathbb{Y} = \{-1, +1\}$):

$$a(x) = \text{sign} \sum_{t=1}^{T} \alpha_t b_t(x);$$

Методы построения композиций (некоторые вехи):

- Метод комитетов [В.Д.Мазуров, 1971]
- Алгоритмы вычисления оценок [Ю.И.Журавлёв, 1971]
- Алгебраический подход [Ю.И.Журавлёв, 1977]
- Boosting [Y.Freund, R.Shapire, 1995]
- Bagging, Arcing [L.Breiman, 1995]
- MatrixNet [Яндекс, 2009]

Композиции логических закономерностей

Взвешенное голосование (для случая $\mathbb{Y} = \{-1, +1\}$):

$$a(x) = \text{sign} \sum_{t=1}^{T} \alpha_t b_t(x).$$

Бустинг основан на двух эвристиках:

- сглаженная функция потерь ($\mathscr{L}(M) = e^{-M}$ — AdaBoost);
- последовательное «жадное» построение $\alpha_t b_t$ при фиксации всех предыдущих $\alpha_1 b_1, \ldots, \alpha_{t-1} b_{t-1}$.

Закономерность класса $y \in \mathbb{Y}$ — это предикат $b_t : \mathbb{X} \to \{0, y\}$

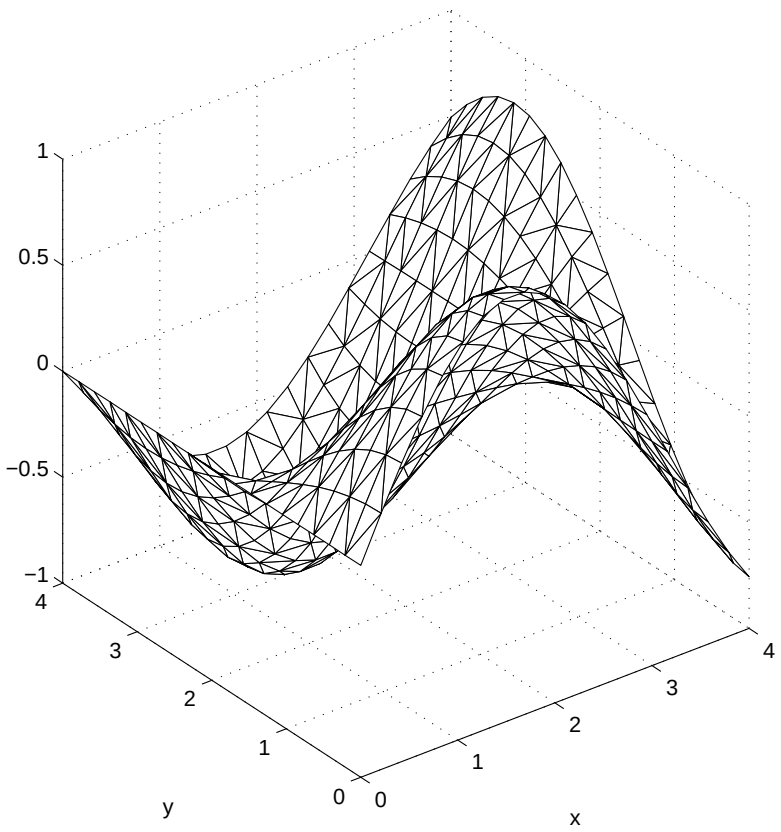
$b_t(x) = \bigwedge_{j \in \omega} \left[\alpha_j \leqslant f_j(x) \leqslant \beta_j\right]$ — пороговые конъюнкции

$b_t(x) = \left[\sum_{j \in \omega} \left[\alpha_j \leqslant f_j(x) \leqslant \beta_j\right] \geqslant w_0\right]$ — синдромные правила

MVR Composer: Процедура порождения моделей

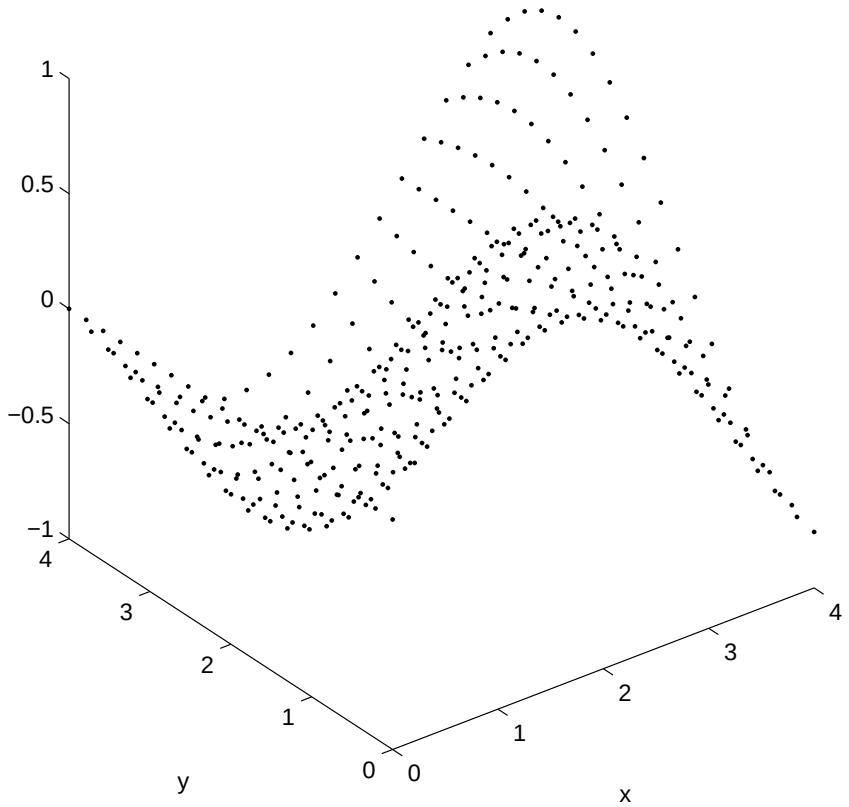
Пример порождения моделей

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



Исходные данные

Выборка состоит из 380 элементов.



Заданы порождающие функции

Функция	Описание	Параметры
	$g(\mathbf{b}, x_1, x_2)$	
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

Рассмотрим набор $\mathcal{F} = \{f_i\}$ моделей — допустимых суперпозиций порождающих функций $G = \{g\}$.

Экспертная информация

Эксперты задают набор начальных моделей, например,

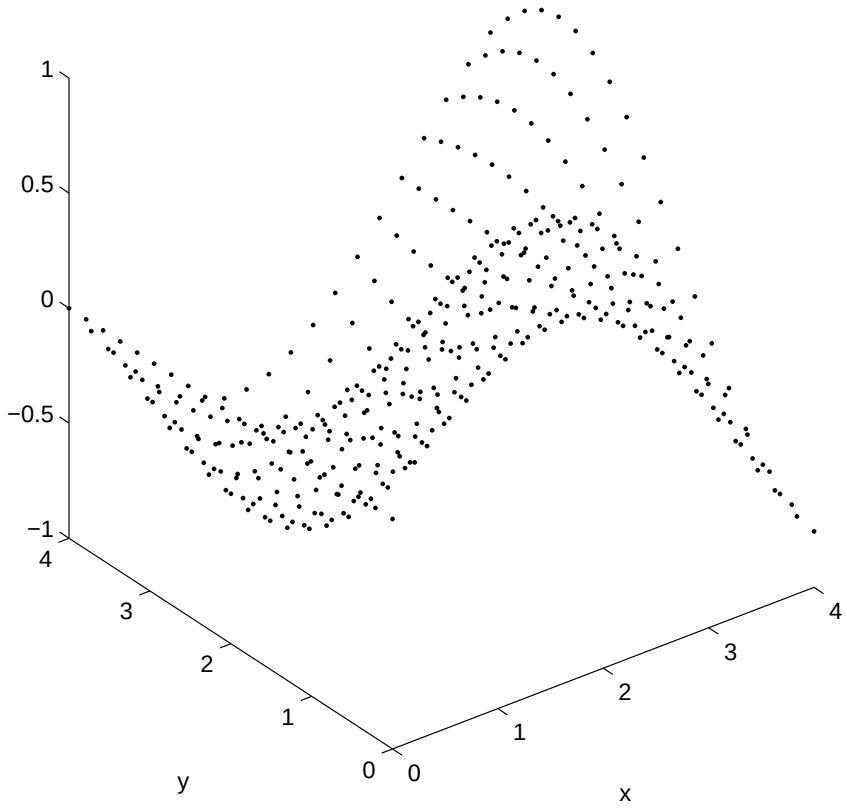
$$\begin{aligned} f_1: &\quad y = \text{linear}(x_1), \\ f_2: &\quad y = \text{normal}(x_2) \end{aligned}$$

и условия порождения моделей:

1. сложность моделей
   $\begin{cases} \text{число элементов суперпозиции } g \text{ не более } 8, \\ \text{число параметров } w \text{ не более } 10; \end{cases}$
2. целевая функция — сумма квадратов регрессионных остатков.

Модели-претенденты

Задана выборка

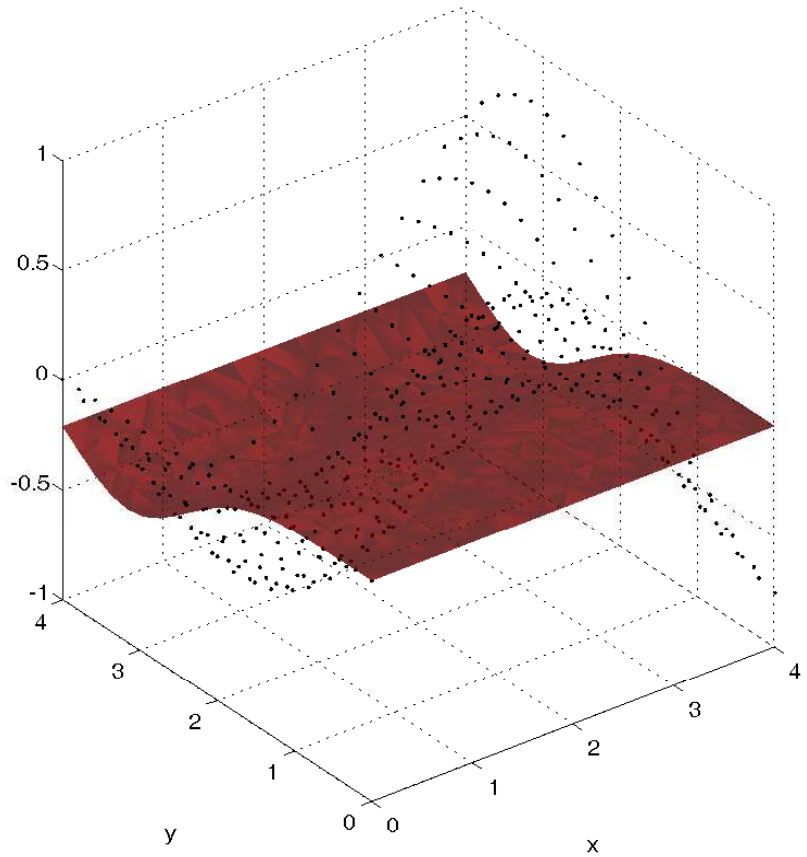


Модели-претенденты

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

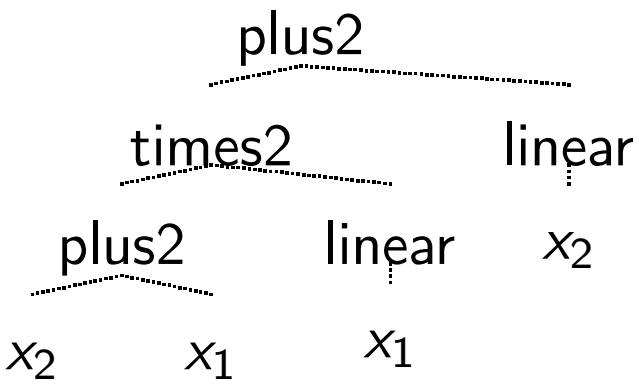
normal

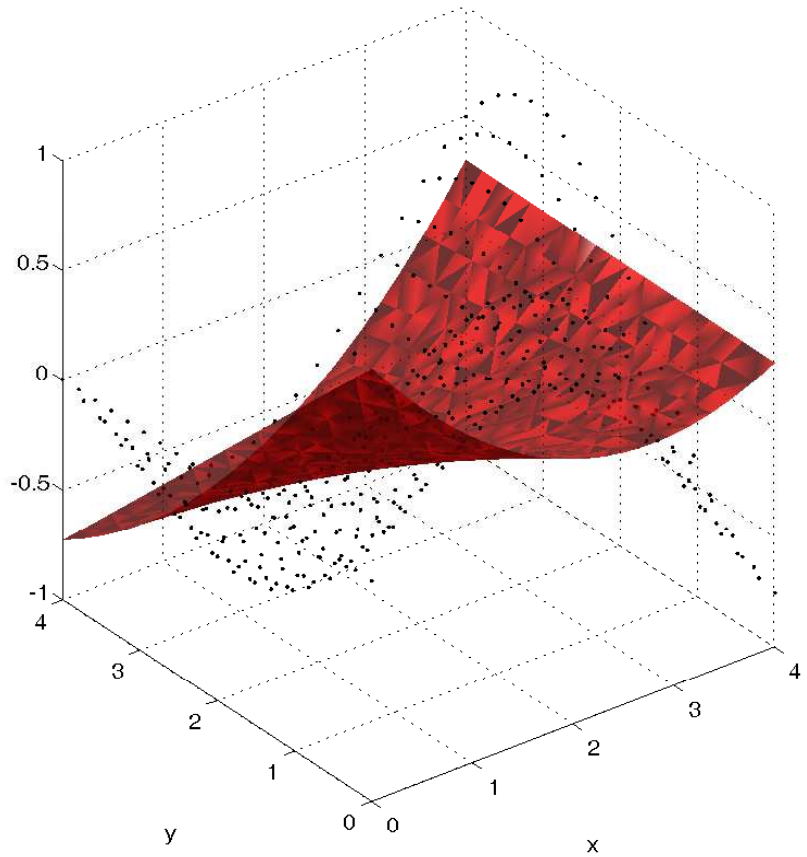
x_2



Модели-претенденты

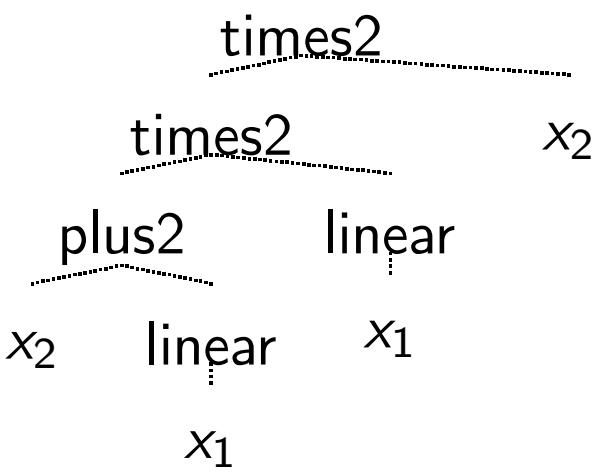
$$\text{plus2}(\varnothing,\text{times2}(\varnothing,\text{plus2}(\varnothing,x_2,x_1),\text{linear}(w_{1:2},x_1)),\text{linear}(w_{3:4},x_2))$$

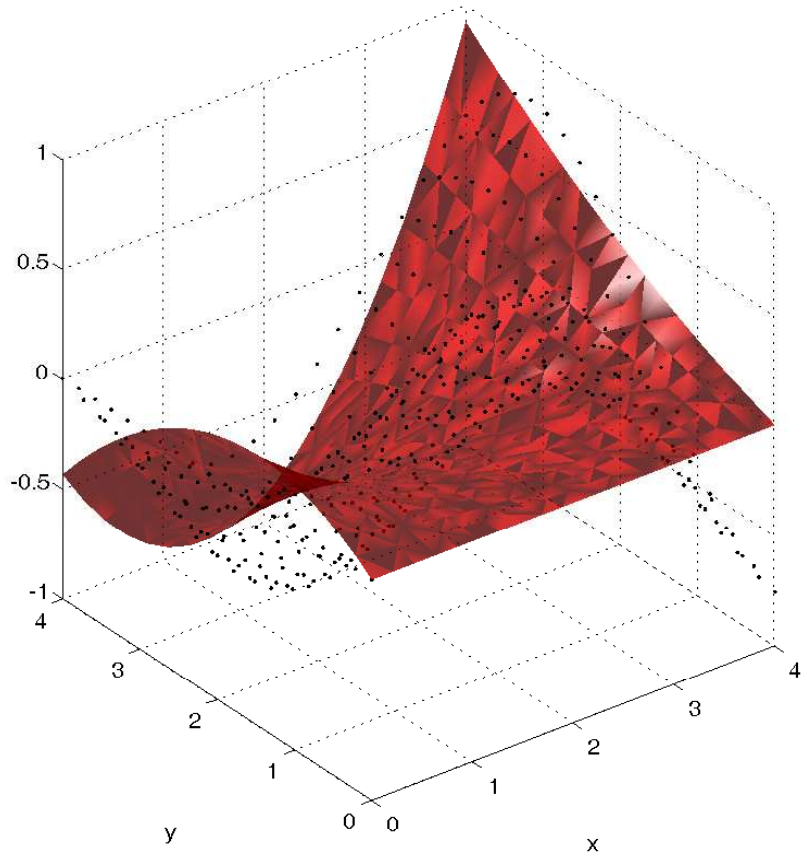




Модели-претенденты

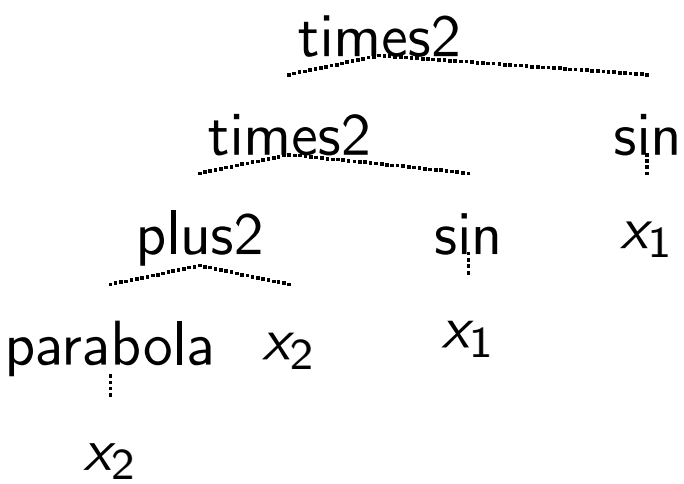
$$\text{times2}(\varnothing,\text{times2}(\varnothing,\text{plus2}(\varnothing,x_2,\text{linear}(w_{1:2},x_1)),\text{linear}(w_{3:4},x_1)),x_2)$$

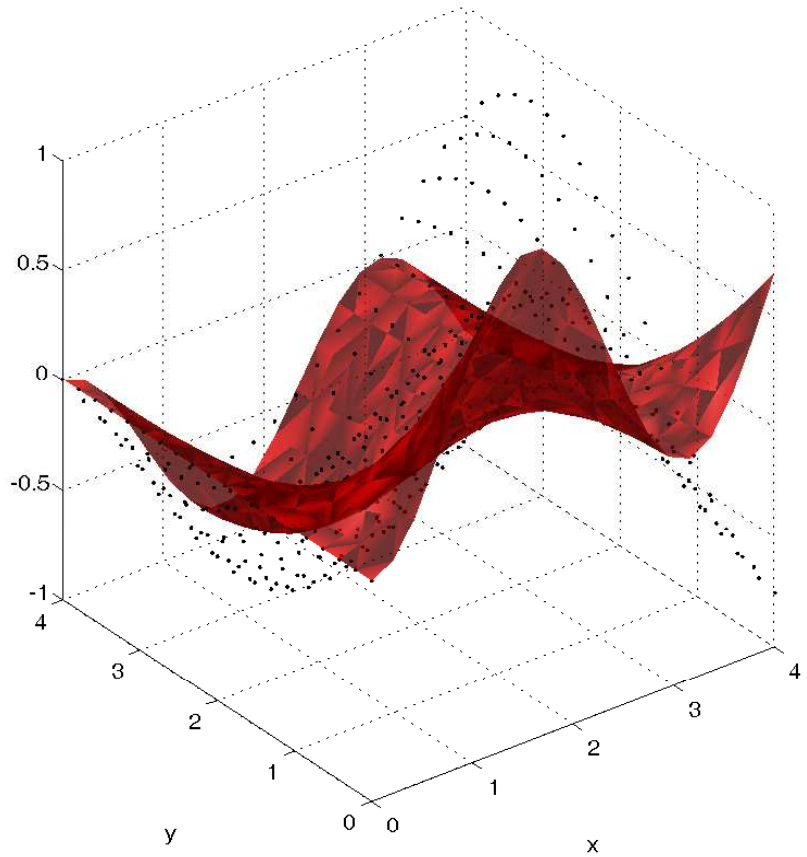




Модели-претенденты

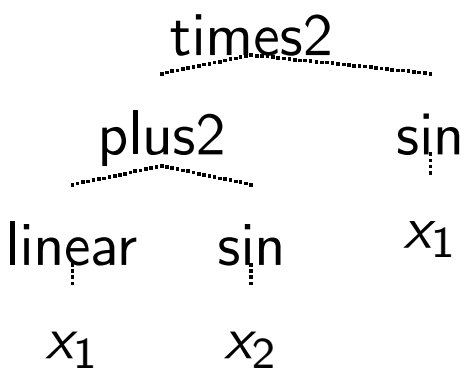
$$\text{times2}(\varnothing,\text{times2}(\varnothing,\text{plus2}(\varnothing,\text{parabola}(w_{1:3},x_2),x_2),\sin(\varnothing,x_1)),\sin(\varnothing,x_1))$$

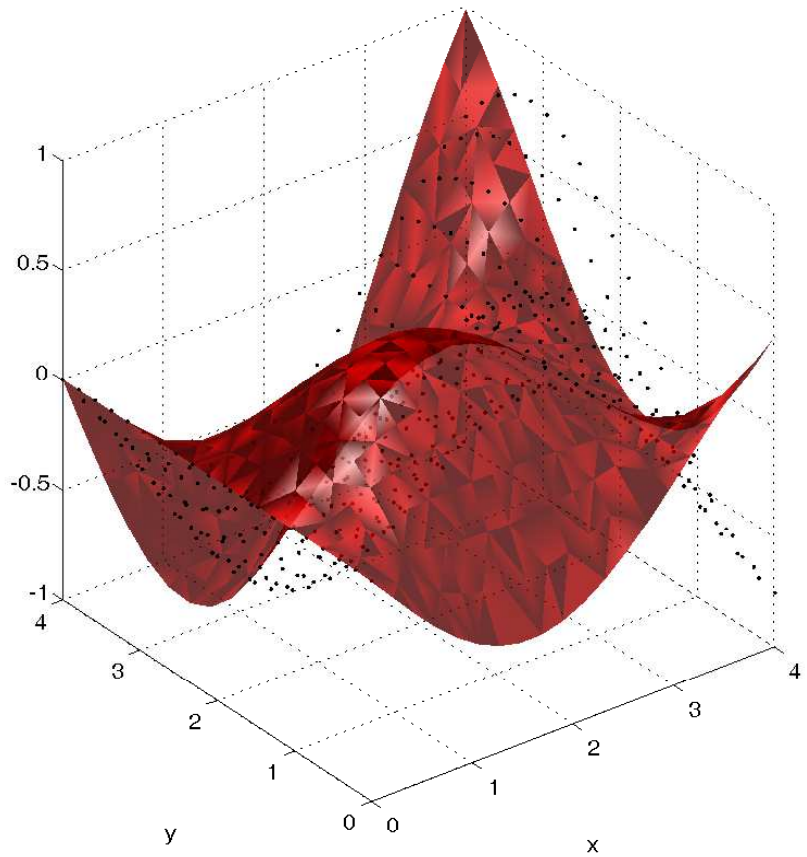




Модели-претенденты

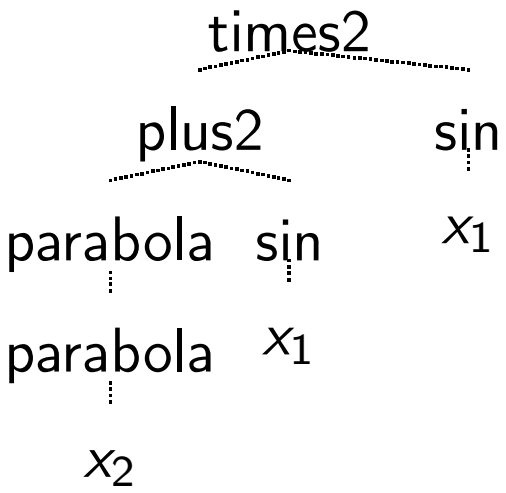
$$\text{times2}(\varnothing,\text{plus2}(\varnothing,\text{linear}(w_{1:2},x_1),\sin(\varnothing,x_2)),\sin(\varnothing,x_1))$$

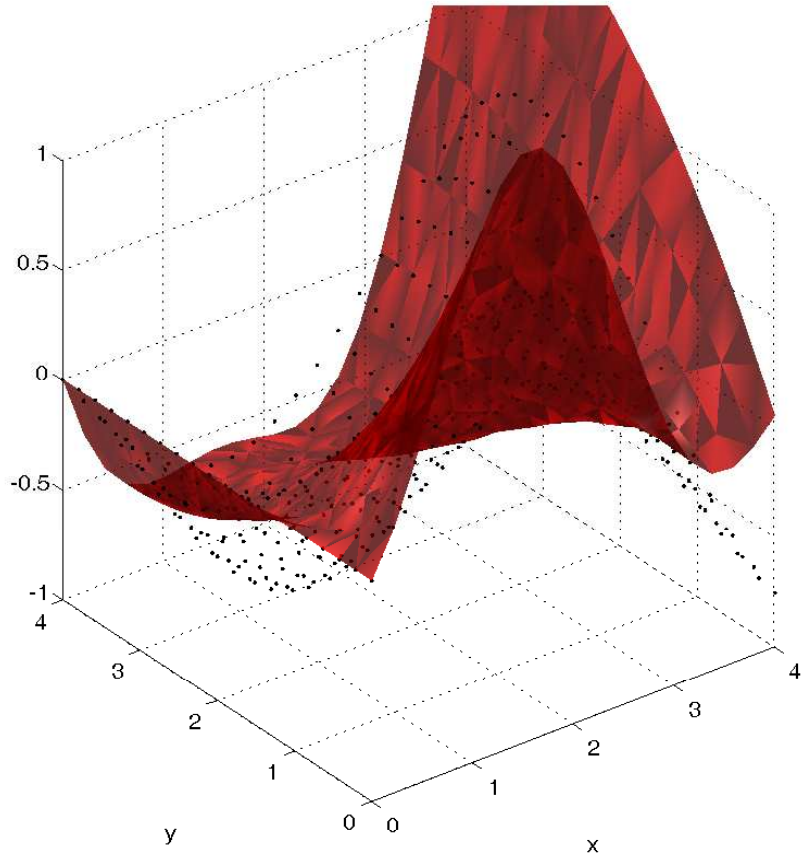




Модели-претенденты

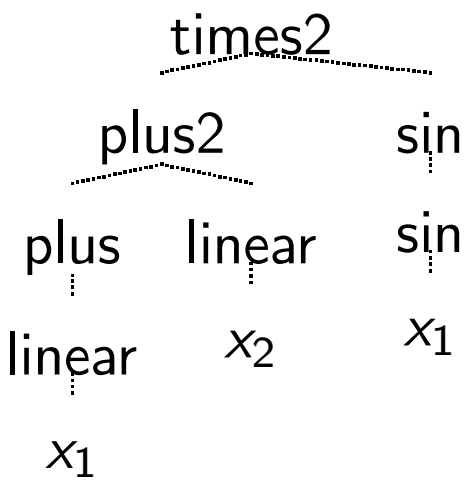
$$\text{times2}(\varnothing,\text{plus2}(\varnothing,\text{parabola}(w_{1:3},\text{parabola}(w_{4:6},x_2)),\sin(\varnothing,x_1)),\sin(\varnothing,x_1))$$

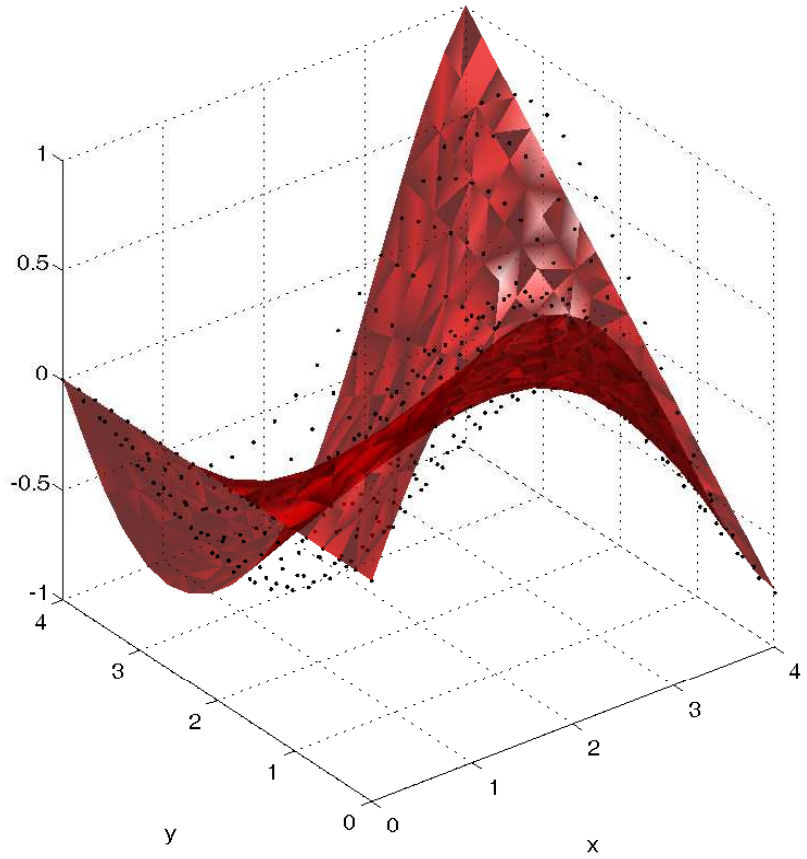




Модели-претенденты

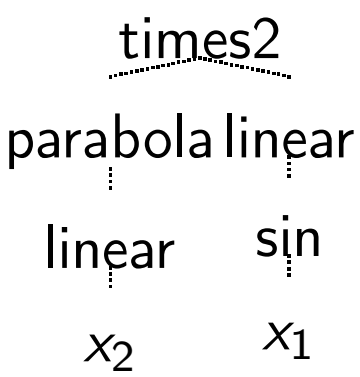
$$\text{times2}(\varnothing,\text{plus2}(\varnothing,\text{plus}(w_1,\text{linear}(w_{2:3},x_1)),\text{linear}(w_{4:5},x_2)),\sin(\varnothing,\sin(\varnothing,x_1))$$

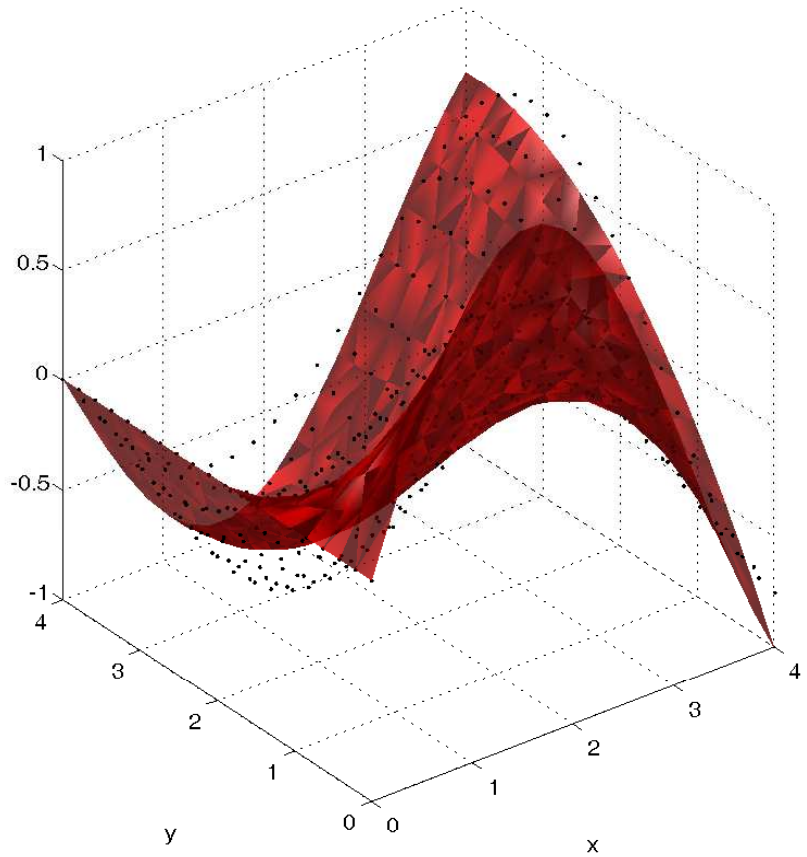




Модели-претенденты

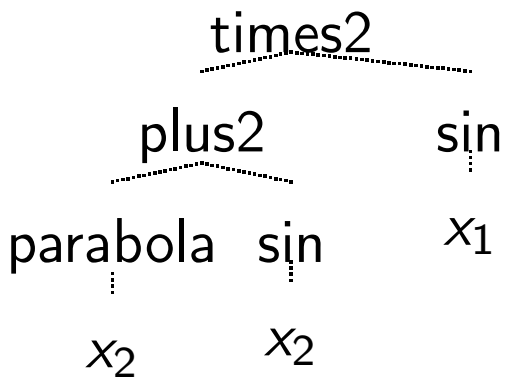
$$\text{times2}(\varnothing, \text{parabola}(w_{1:3}, \text{linear}(w_{4:5}, x_2)), \text{linear}(w_{6:7}, \sin(\varnothing, x_1)))$$

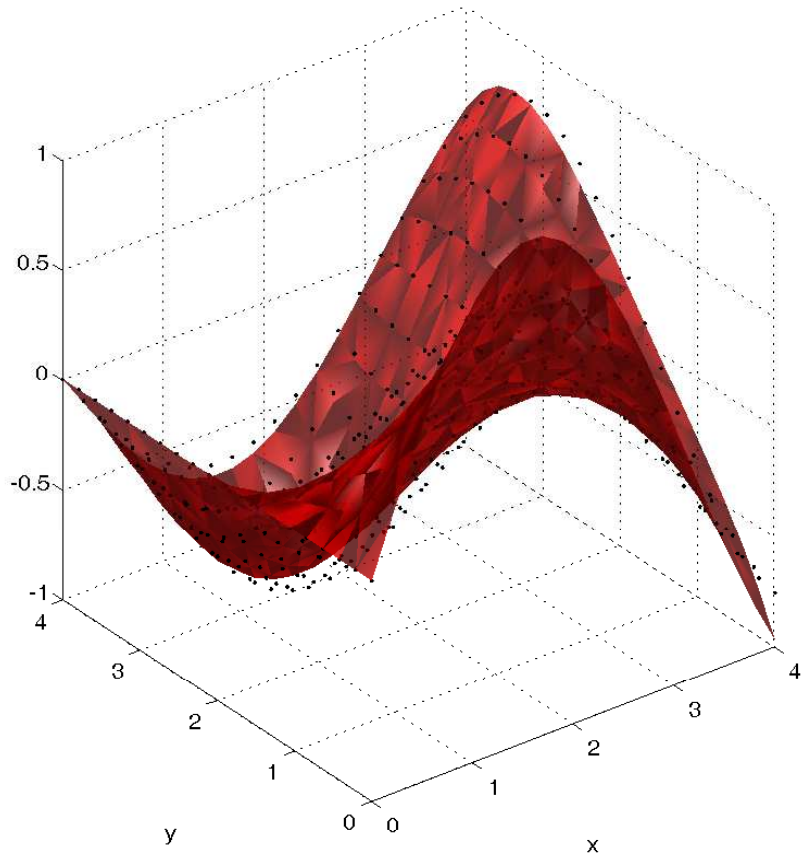




Модели-претенденты

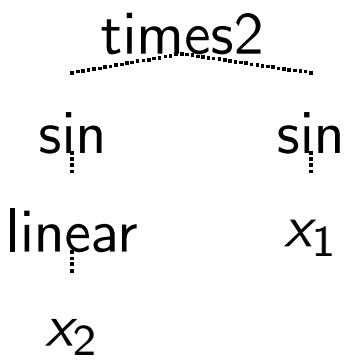
$$\text{times2}(\varnothing,\text{plus2}(\varnothing,\text{parabola}(w_{1:3},x_2),\sin(\varnothing,x_2)),\sin(\varnothing,x_1))$$

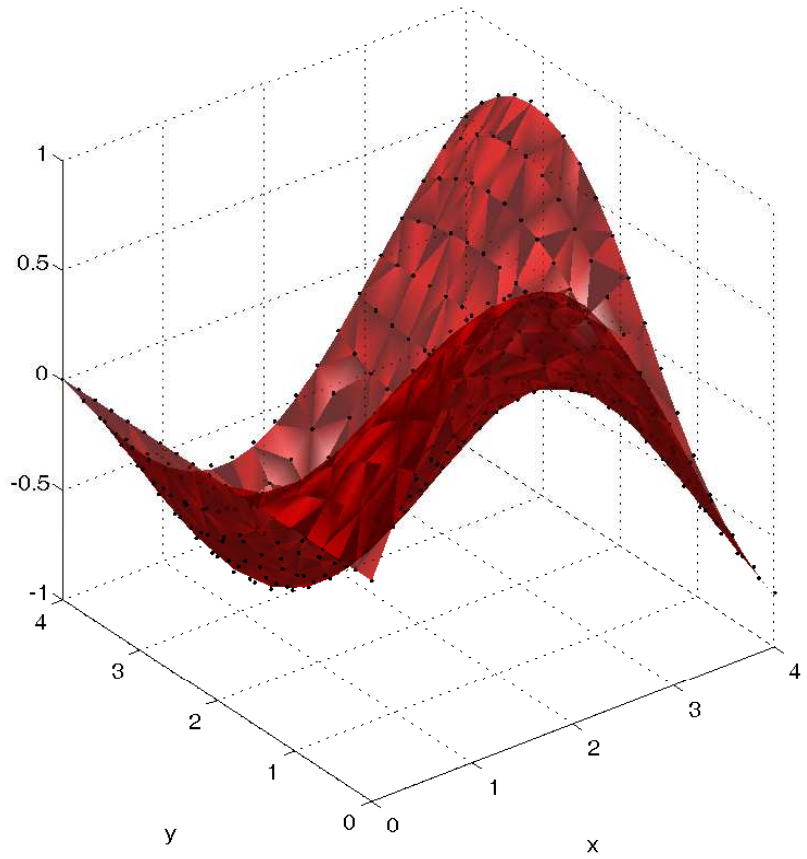




Модели-претенденты

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





Дано

Выборка:

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

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

Обозначим выборку $D = \{(\mathbf{x}_n, y_n)\}$.

Порождающие функции:

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

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

$G \ni g$ определяет набор допустимых суперпозиций $\mathcal{F} = \{f_i\}$ индуктивно;

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

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

Требуется

Требуется найти модель $f_i \in \mathcal{F}$,

$$y_n = f_i(\mathbf{w}, \mathbf{x}_n) + \varepsilon,$$

которая доставляет максимум
целевой функции $p(\mathbf{w}|D, \alpha, \beta, f_i)$,

$f_i \in \mathcal{F}$ — набор моделей-претендентов,

$\mathbf{w}$ — параметры модели,

D — выборка,

α, β — гиперпараметры или регуляризующие параметры.

Порождение моделей — 1

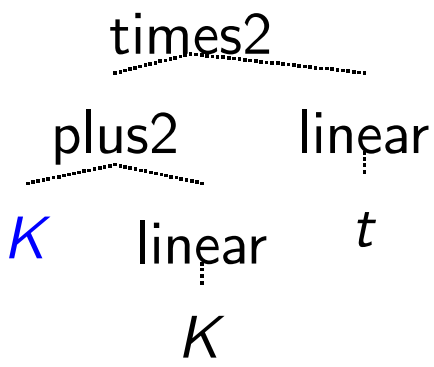
Базовый алгоритм порождения моделей состоит из трех шагов.

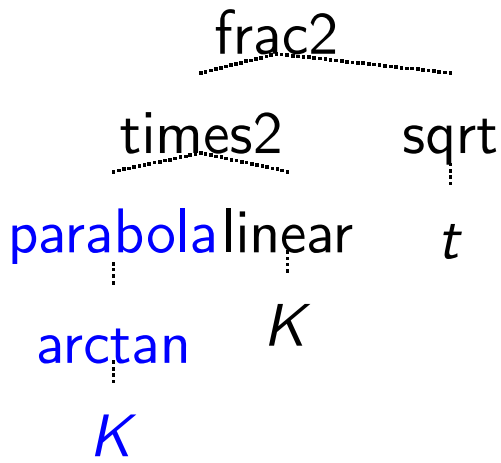
1. Оценить параметры и гиперпараметры каждой модели из порожденного множества $\mathcal{F} = \{f_1, \ldots, f_M\}$:

$$\mathbf{w}_i^{\text{MP}} = \arg\min_{\mathbf{w}} S(\mathbf{w}|D, A, \beta, f_i).$$

Порождение моделей — 2

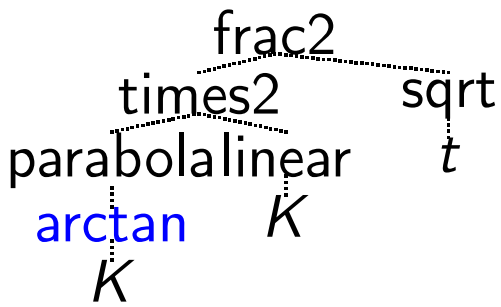
2. Поменять местами элементы пары моделей:
   1. поменять местами элементы g_{ik} and g_{jl} с индексами $i, j \in \{1, \ldots, M\}$ моделей f_i и f_j,
   2. создать новые модели f_i' и f_j'.





Порождение моделей — 3

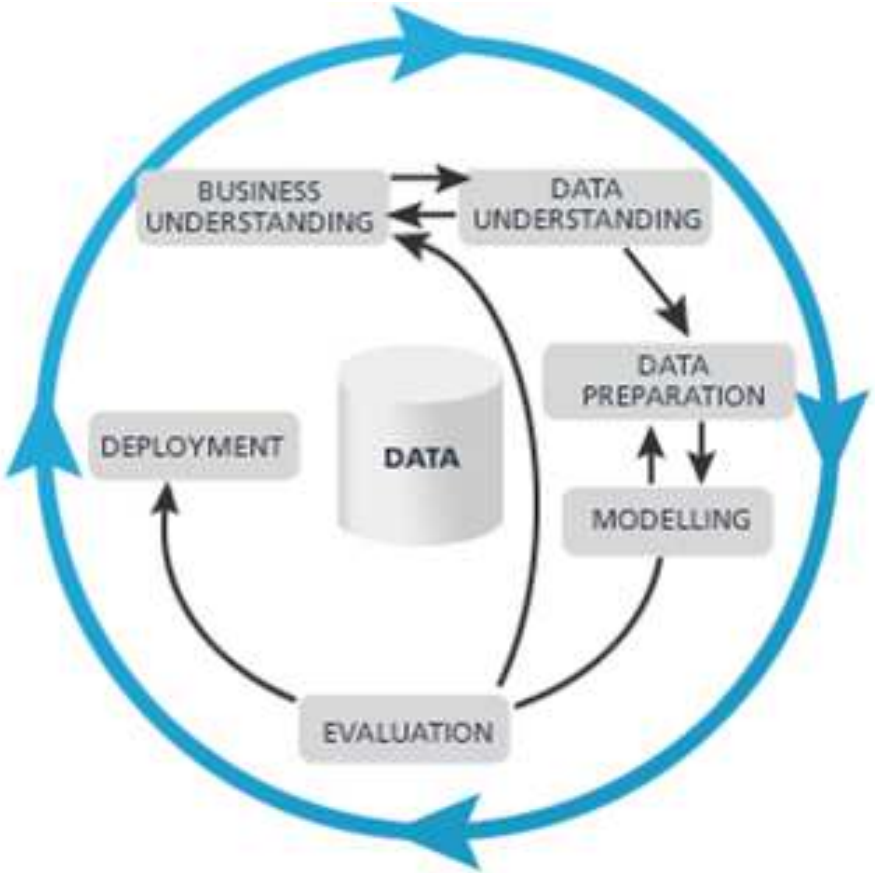
3. Модификация элементов моделей: изменить элементы $\{f_i'\}$:
   1. выбрать элемент g_{ik} модели f_i,
   2. выбрать из множества G элемент g_s (с учетом допустимости замены g_{ik}),
   3. заменить элемент g_{ik} на порождающую функцию g_s.



Межотраслевой стандарт CRISP-DM

Методология ведения проектов Data Mining, предложенная компаниями ISL (SPSS), NCR Corporation, Daimler-Benz, OHRA.

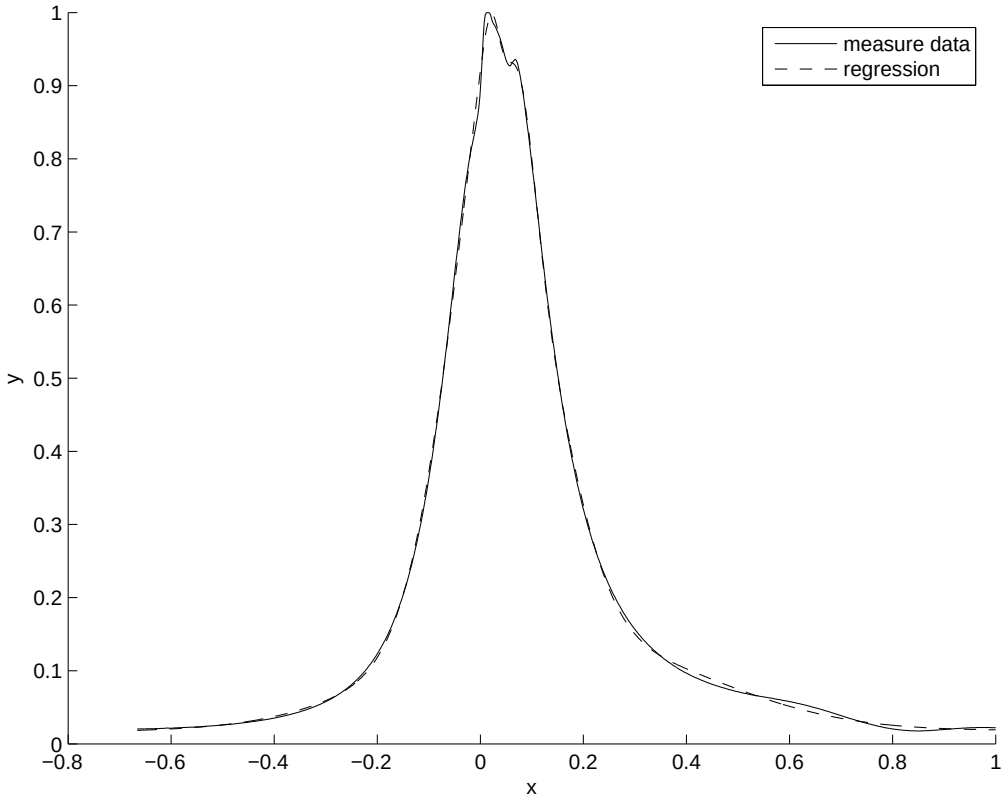
Процесс анализа данных разбивается на 6 основных этапов:



MVR Composer: Примеры приложений

- Моделирование процессов горения в двигателе внутреннего сгорания
  - **Задача:** спрогнозировать концентрацию кислорода в выхлопных газах
  - **Цель:** исключить датчик кислорода из конструкции двигателя
  - **Результат:** точность прогноза значений датчика 0.96%
- Прогнозирование волатильности опционов
  - **Задача:** построить модель зависимости волатильности опциона от его цены и времени до исполнения
  - **Цель:** уточнить справедливую цену опциона
  - **Результат:** предложена более точная модель волатильности
- Прогнозирование цен и объёмов электроэнергии
  - **Задача:** автоматическое формирование прогнозов
  - **Цель:** продажа электроэнергии на свободном рынке
  - **Результат:** точность прогнозов 5% с надёжностью 90%

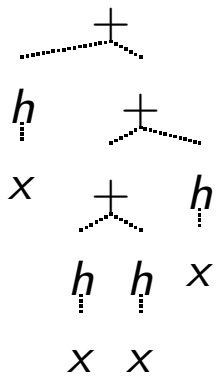
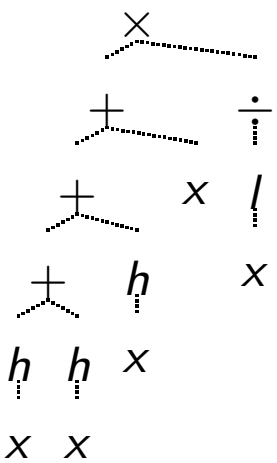
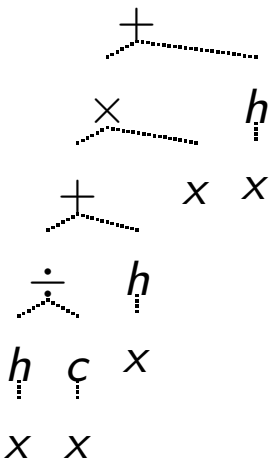
Пример: моделирование процессов горения



Давления в камере внутреннего сгорания дизельного двигателя:

x — угол вращения коленчатого вала (нормирован),
y — давление (нормировано),
выборка содержит 4000 элементов.

Полученные модели

Model 1	Model 2	Model 3
		

Legend: h — gaussian $y = \lambda(2\pi\sigma^{-1/2})\exp(-(x-\xi)^2(2\sigma^{-2}) + a)$,
c — cubic $y = ax^3 + bx^2 + cx + d$, l — linear $y = ax + b$.

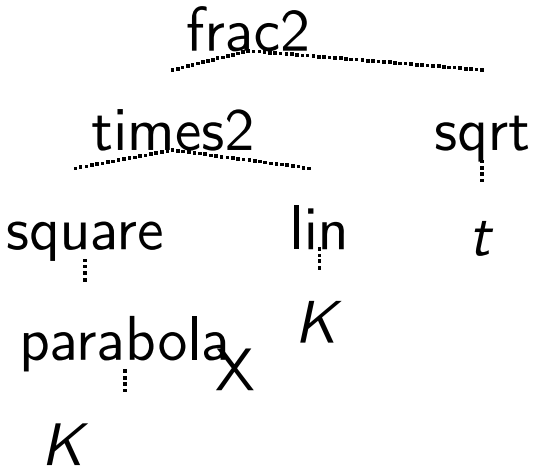
$$f_2 = g_1(g_2(g_3(g_4(g_5(x), g_6(x)), g_7(x)), x), g_8(x)).$$

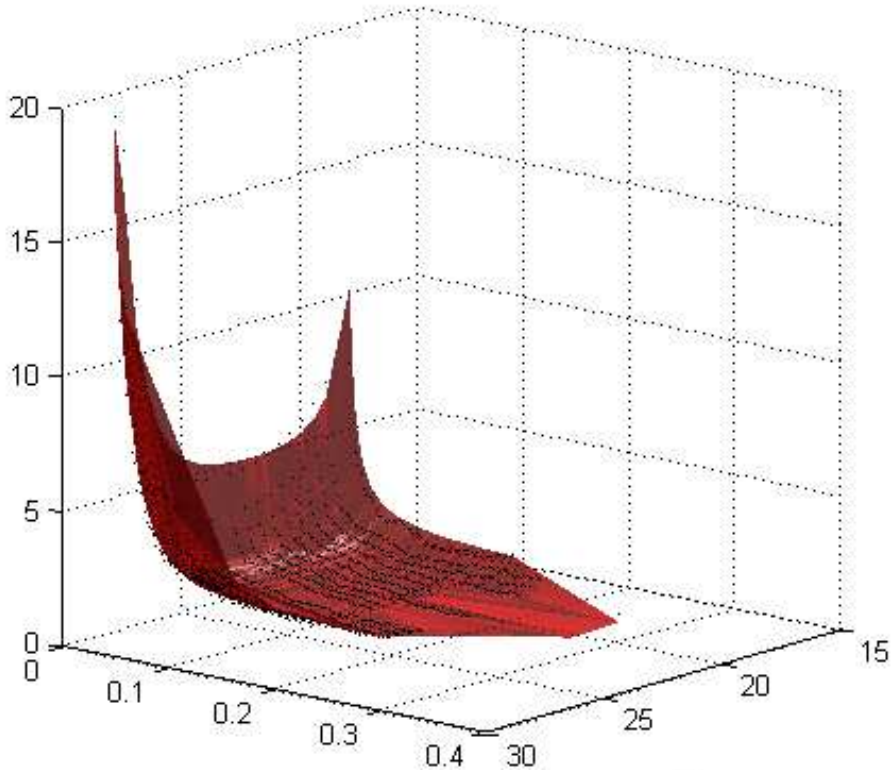
Полная форма записи модели 2:

$$y = (ax+b)^{-1}\left(x + \sum_{i=1}^{3} \frac{\lambda_i}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x-\xi_i)^2}{2\sigma_i^2}\right) + a_i\right).$$

Пример: моделирование волатильности европейских опционов

Европейский опцион на сырую нефть марки Brent.
Подразумеваемая волатильность σ зависит от времени до исполнения опциона t и от цены исполнения K: $\sigma = \frac{(w_1 K + w_2)(w_1 K^2 + w_2 K + w_3)^2}{\sqrt{t}}$.





Почасовое прогнозирование объемов и цен электроэнергии

Данные:

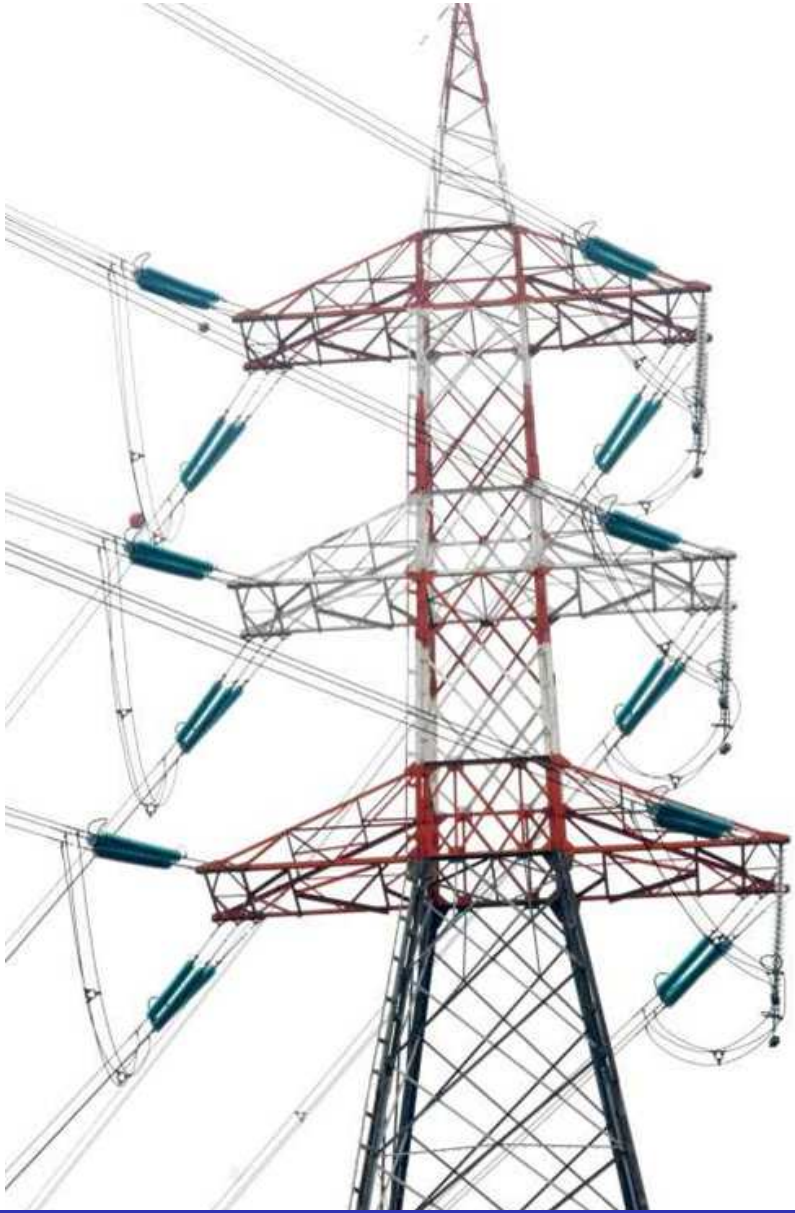
- исторические объемы потребления и цены на электроэнергию, набор временных рядов.

Требуется спрогнозировать:

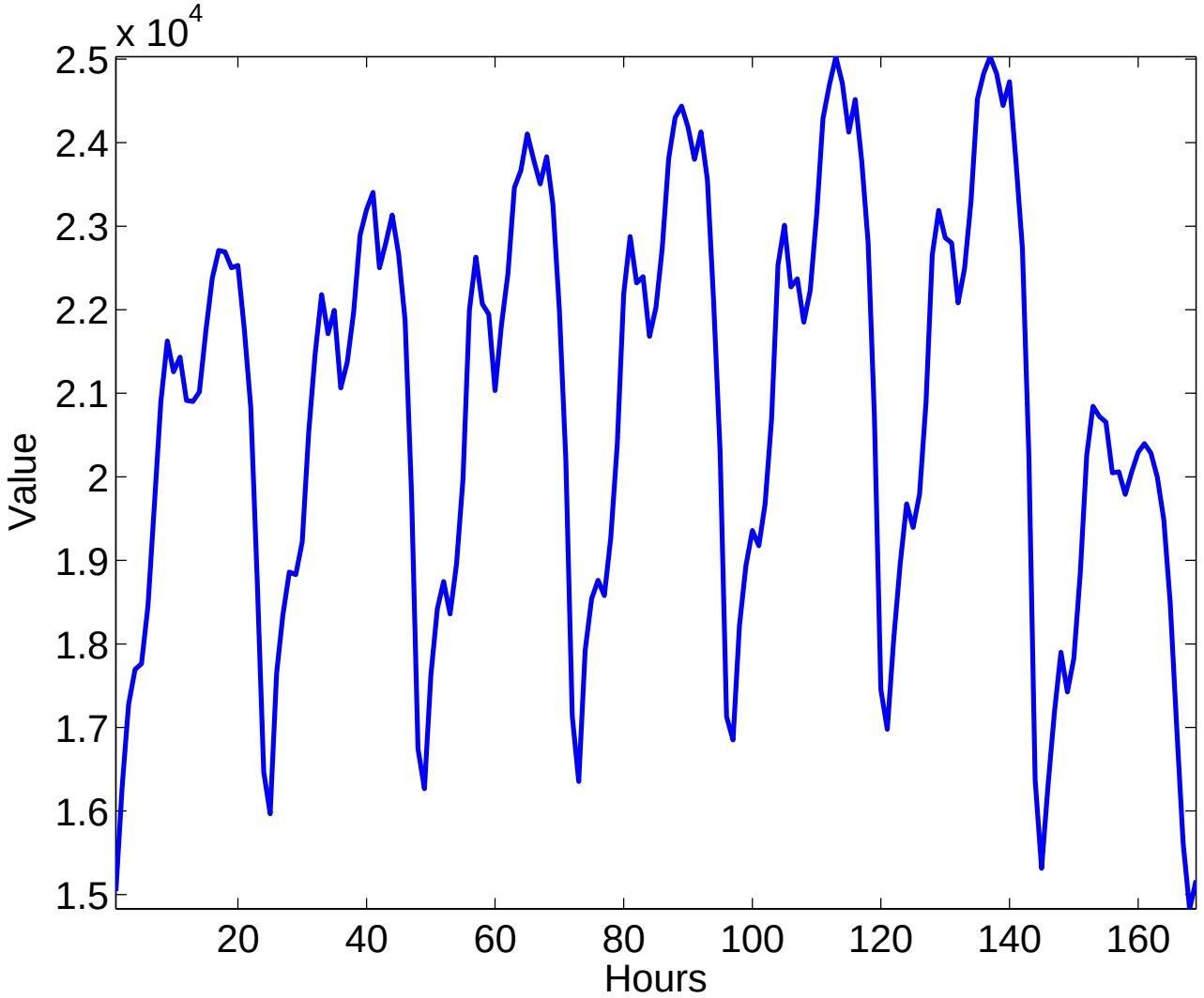
- следующий день на каждый час
  - ✓ объем потребления
  - ✓ цену.

Решение:

- порождение и выбор авторегрессионных моделей.



Исходный временной ряд, одна неделя



Периодические компоненты набора временных рядов

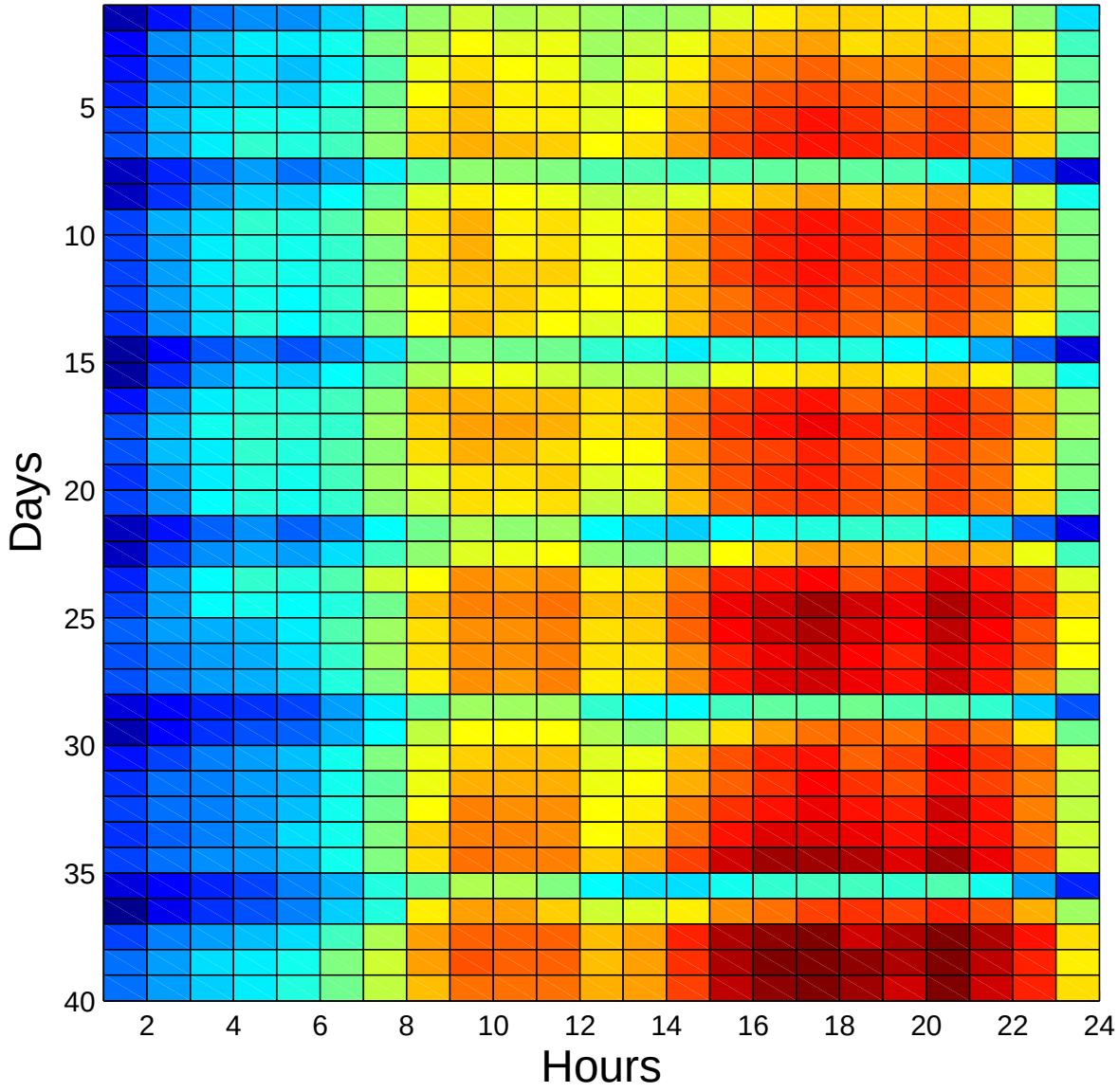
Временные ряды:

- цена на электроэнергию,
- потребление,
- длина светового дня,
- температура,
- влажность,
- скорость ветра,
- расписание праздников.

Периоды:

- годовая сезонность (температура, длина светового дня),
- еженедельная,
- ежедневная (рабочие дни, выходные),
- праздники,
- апериодические события.

Авторегрессионная матрица, пять выходных



Авторегрессионная матрица и линейная модель

$$\underset{(m+1)\times(n+1)}{X^*} = \left(\begin{array}{c|ccc} s_T & s_{T-1} & \cdots & s_{T-\varkappa+1} \\ \hline s_{(m-1)\varkappa} & s_{(m-1)\varkappa-1} & \cdots & s_{(m-2)\varkappa+1} \\ \cdots & \cdots & \cdots & \cdots \\ s_{n\varkappa} & s_{n\varkappa-1} & \cdots & s_{n(\varkappa-1)+1} \\ \cdots & \cdots & \cdots & \cdots \\ s_{\varkappa} & s_{\varkappa-1} & \cdots & s_1 \end{array}\right).$$

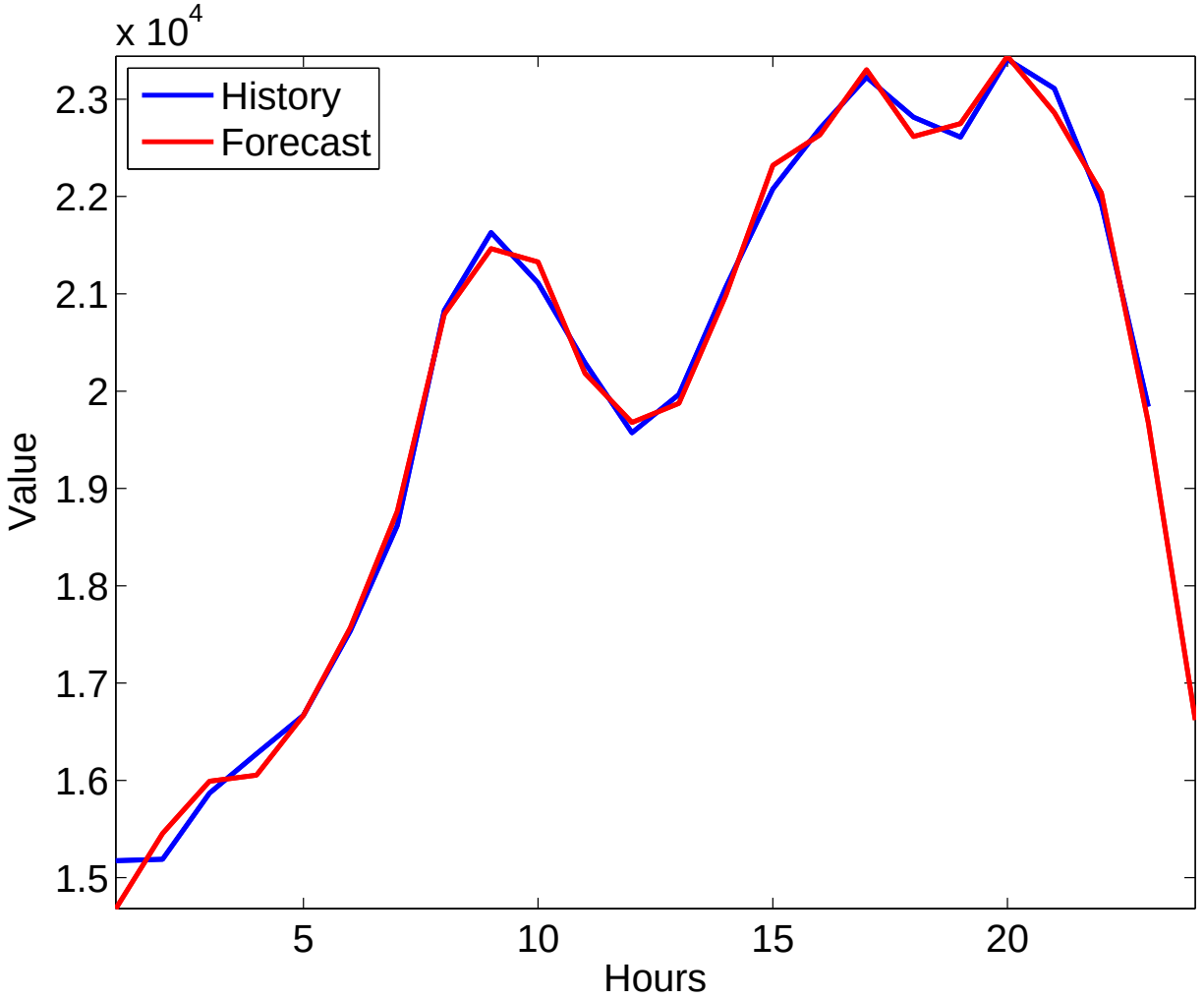
Кратко,

$$X^* = \left[\begin{array}{c|c} \underset{1\times1}{s_T} & \underset{1\times n}{\mathbf{x}_{m+1}} \\ \hline \underset{m\times1}{\mathbf{y}} & \underset{m\times n}{X} \end{array}\right].$$

Решение задачи прогнозирования:

$$\hat{\mathbf{y}} = X\mathbf{w}, \quad \text{прогноз} \quad \hat{y}_{m+1} = \hat{s}_T = \mathbf{w}^{\mathrm{T}}{\mathbf{x}_{m+1}}^{\mathrm{T}}.$$

Прогнозирование на один день, пример



Композиции моделей прогнозирования: Примеры приложений

- Прогнозирование объёмов продаж в торговой сети
  - **Задача:** прогноз для каждого товара в каждом магазине
  - **Цель:** автоматическое формирование заказов в магазины
  - **Результат:** сокращение издержек 0,6–3,8% от оборота
- Прогнозирование оттока клиентов
  - **Задача:** оценивание вероятности ухода абонента
  - **Цель:** удержание выгодных клиентов
  - **Результат:** среди выделенных 10% абонентов 40–60% уходящие
- Прогнозирование спроса в производстве
  - **Задача:** прогноз спроса на пиво
  - **Цель:** формирование плана продаж
  - **Результат:** увеличение точности прогнозов на 18,6%
- Прогнозирование рейтингов рекламы
  - **Задача:** прогноз спроса на эфир по группам населения
  - **Цель:** оптимизация размещения рекламных блоков
  - **Результат:** увеличение точности прогнозов на 4%

Примеры прикладных задач

- Анализ данных жидкостной хроматографии

$$z(t,\lambda) = \sum_i X_i(t) Y_i(\lambda)$$

дано: $z(t,\lambda)$ — выход сканирующего УФ-детектора;
найти: $X_i(t)$ — хроматограмма i-го вещества,
найти: $Y_i(\lambda)$ — спектр i-го вещества.

- Анализ данных ДНК-микрочипов

$$I(p,k) = \sum_g a_{pg} C_{gk}$$

дано: $I(p,k)$ — интенсивность свечения p-й пробы на k-м чипе;
найти: a_{pg} — коэффициент сродства p-й пробы g-му гену,
найти: C_{gk} — концентрация g-го гена на k-м чипе.

- Тематические модели коллекций текстовых документов

$$p(w|d) = \sum_t p(w|t) p(t|d)$$

дано: $p(w|d)$ — частоты слов w в документах d;
найти: $p(w|t)$ — распределения слов w в темах t,
найти: $p(t|d)$ — распределения тем t в документах d.

контакты:

Воронцов Константин Вячеславович
voron@forecsys.ru
www.MachineLearning.ru/wiki, «Участник:Vokov»

Стрижов Вадим Викторович
strijov@forecsys.ru
http://www.strijov.com
www.MachineLearning.ru/wiki, «Участник:Strijov»