Гипотеза порождения данных в задачах регрессионного анализа

Стрижов В.В.
Вычислительный центр им. А.А. Дородницына РАН

Регрессионный анализ — задача оптимизации

Дано:
$\{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ — конечная выборка,
$G = \{g_j : \mathbf{x} \mapsto g_j(\mathbf{x})\}_{j \in J}$ — множество порождающих функций,
$f(\mathbf{w}, \mathbf{x})$ — модель (параметрическое семейство функций),
суперпозиция порождающих функций g,
s — критерий оптимальности.

Найти:
параметры $\mathbf{w}$, доставляющие минимум критерию s для

$$y_i = f(\mathbf{w}, \mathbf{x}_i) + \varepsilon_i.$$

Пример: задача линейной регрессии

Модель регрессии:

$$y_i = \sum_{j \in I} w_j g_j(\mathbf{x}_i) + \varepsilon_i = \langle \mathbf{g}(\mathbf{x}_i), \mathbf{w} \rangle + \varepsilon_i,$$

в других обозначениях

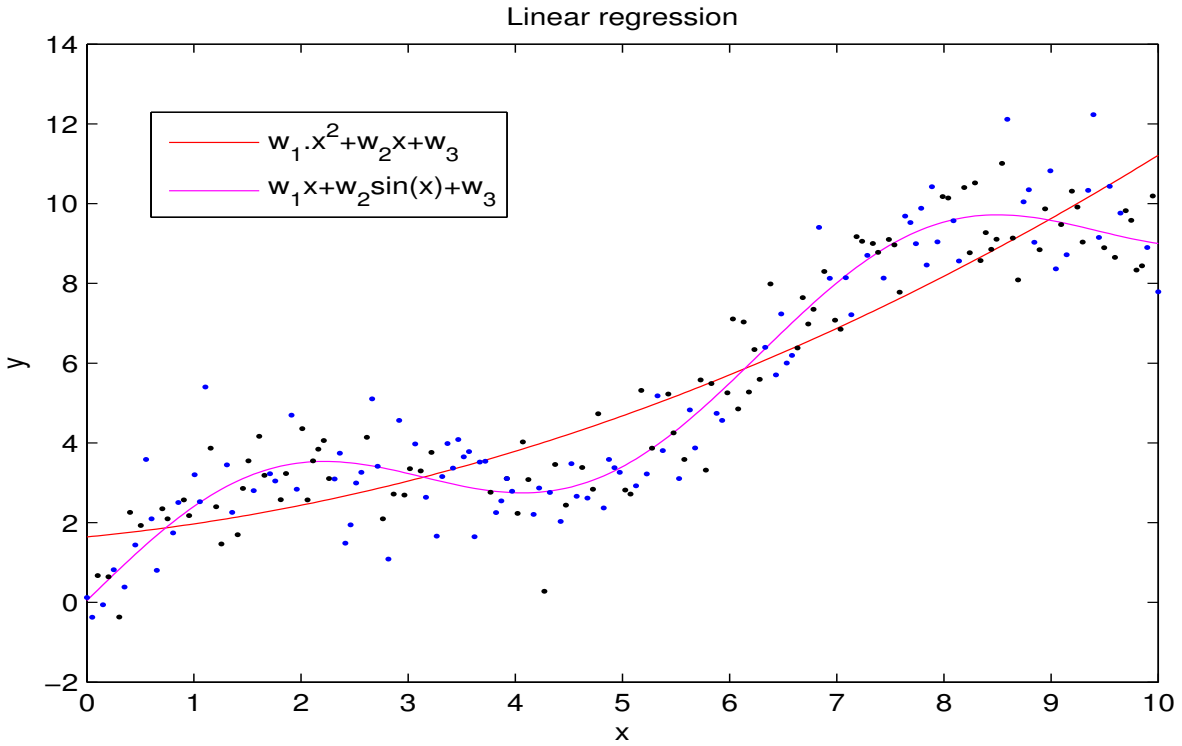
$$\mathbf{y} = X\mathbf{w} + \mathbf{e}.$$

Критерий:

$$s = \|\mathbf{e}\| \longrightarrow \min,$$

$$\mathbf{w} = \arg\min \|\mathbf{y} - X\mathbf{w}\| = (X^T X + \gamma^2 I)^{-1} X^T \mathbf{y}.$$

Пример: линейная регрессия



По оси ординат — свободная переменная,
по оси абсцисс — зависимая.

Вторая постановка задачи — индуктивный вывод модели оптимальной сложности

$y_i = w^{(0)} + \sum_{p=1}^{n} w^{(p)} x_i^{(p)} + \sum_{p=1}^{n} \sum_{q=1}^{n} w^{(pq)} x_i^{(p)} x_i^{(q)} + \ldots$ — базовая модель,

$\mathbf{x} = \{x^{(p)}\}_{p=1}^{n}$ — вектор свободных переменных, $\mathbf{x} \mapsto g(\mathbf{x})$,

$\mathbf{w} = \{w^{(p)}, w^{(pq)}, \ldots\}_{p,q,\ldots=1}^{n}$ — параметры.

Внешние критерии:

$\Delta^2 = \|\mathbf{y}_B - X_B \mathbf{w}_A\|^2$ — регулярность,

$\eta^2 = \|X_W \mathbf{w}_A - X_W \mathbf{w}_B\|^2$ — непротиворечивость,

$V^2 = (X_W \mathbf{w}_A - X_W \mathbf{w}_W)^T (X_W \mathbf{w}_W - X_W \mathbf{w}_B)$ — помехоустойчивость.

Множества $A, B \subset W = \{1, \ldots, \ell\}, A \cup B = W, A \cap B = \emptyset$ — индексы векторов-строк матрицы X.

Вторая постановка задачи

Дано:

множество функций (настроенных моделей) $\{\bar{f}_k\}$ и один внешний критерий, например, Δ_k^2, η_k^2, V_k^2 или

$$S_k = \alpha^T \mathbf{s}_k, \qquad \|\alpha\| = 1,$$

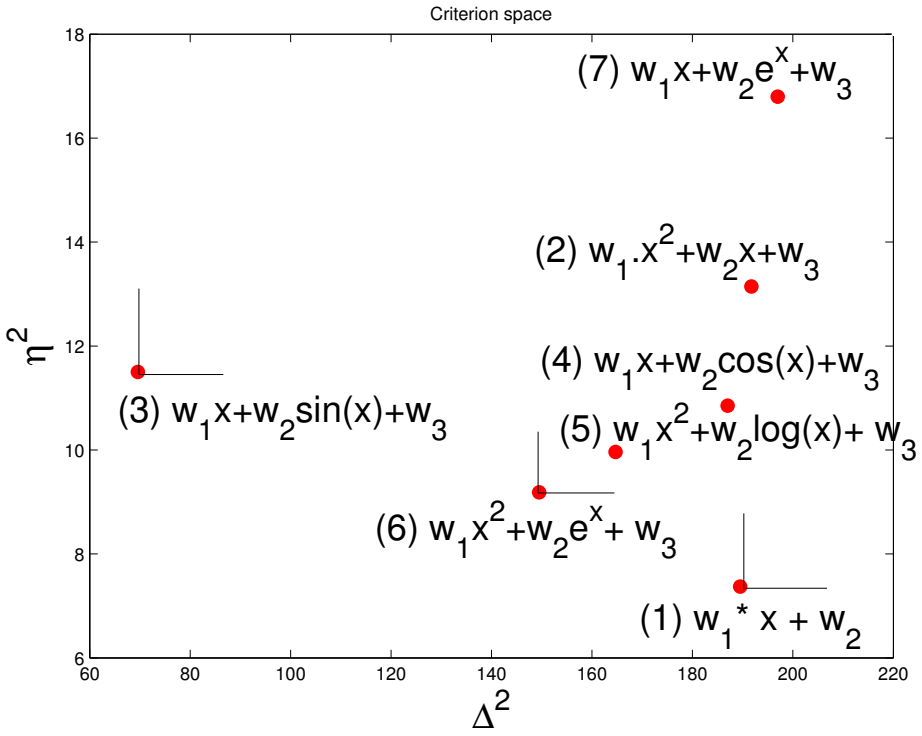
$\mathbf{s}_k$ — вектор критериев k-й модели,

α_i — назначенный вектор весов критериев.

Найти:

такую модель f_k (оптимальной сложности) для которой выбранный внешний критерий $s_k^2 \longrightarrow \min$.

Многокритериальность и Парето-оптимальный фронт



POF в пространстве критериев Δ^2, η^2

Парето-оптимальный фронт

Вектор $\mathbf{s}_k = \{s_k^{(\tau)}\}_{\tau=1}^{T}$ называется недоминируемым, если не найдется ни одного вектора $\mathbf{s}_l$ такого, что

$$s_k^{(\tau)} \leqslant s_l^{(\tau)}, \qquad l = 1, ..., \ell, \qquad \tau = 1, ..., T, \qquad \mathbf{s}_k \neq \mathbf{s}_l.$$

Искомое множество моделей $\{f_k\}$ определено индексами $k : \mathbf{s}_k \in \mathrm{POF}(\{f\}, \{\mathbf{s}\})$.

Обобщенная линейная регрессия

Линейная регрессия предполагает нормальное распределение остатков и диагональную ковариационную матрицу

$$V(\mathbf{e}) = \sigma^2 I,$$

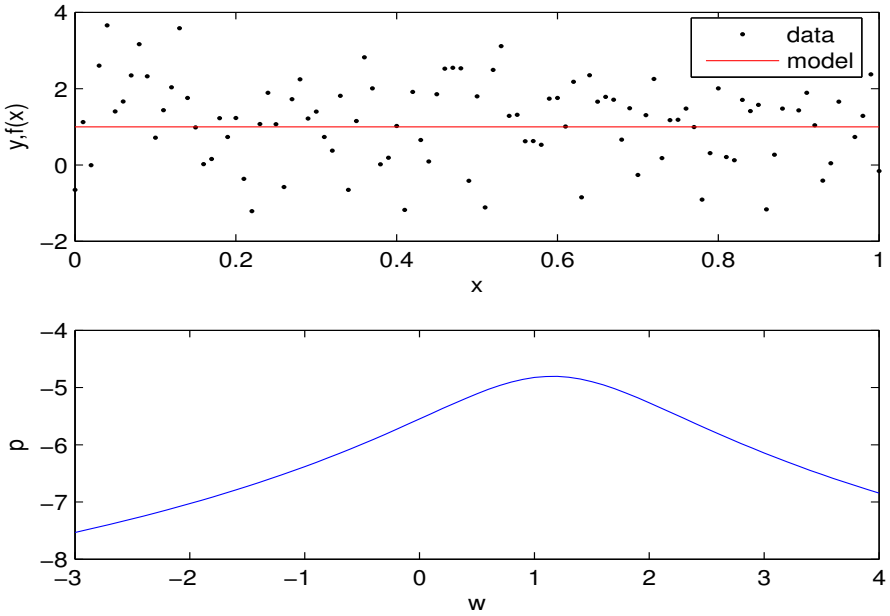
в противном случае

$$\mathsf{E}(\mathbf{y}) = \mu(X\mathbf{w}).$$

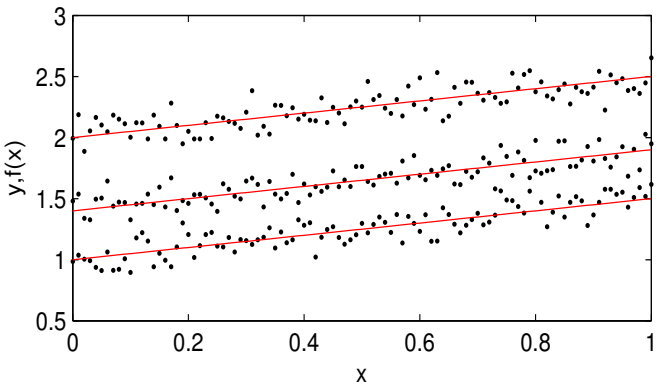
Экспоненциальное семейство:

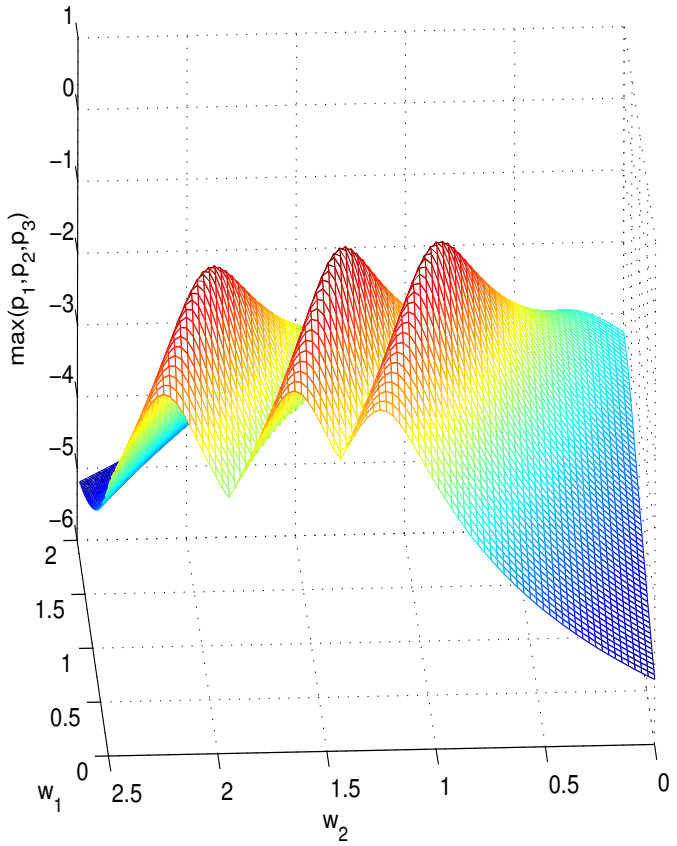
Распределение	Функция $\mu =$
Нормальное	$X\mathbf{w}$
Экспоненциальное	$X\mathbf{w}^{-1}$
Обратное гауссово	$X\mathbf{w}^{-\frac{1}{2}}$
Пуассоновское	$\exp(X\mathbf{w})$
Биномиальное	$\exp(X\mathbf{w})(1-\exp(X\mathbf{w}))^{-1}$

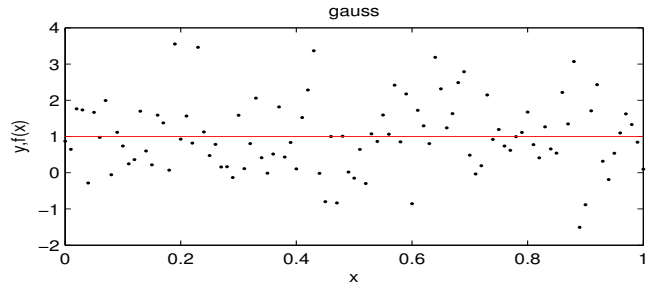
Пространство данных и пространство параметров

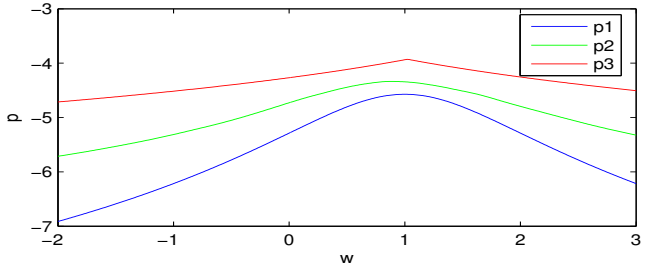


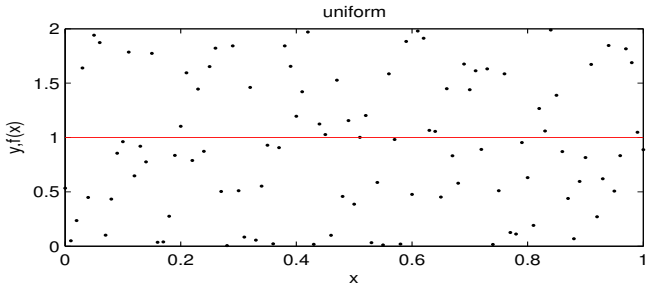
Модель $f(\mathbf{x}) = \mathbf{w} = \text{const}$, $\mathbf{w}_{\text{ML}} = \arg\max(p(\mathbf{w}|D, f))$

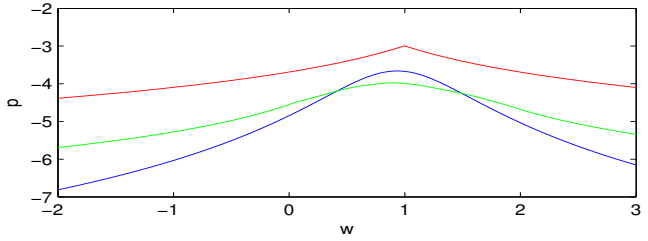

y,f(x)
x
p
w

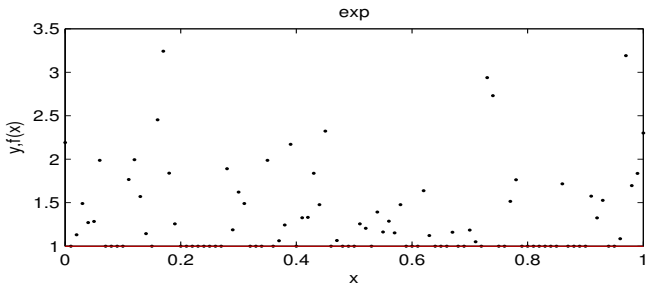

max(p1,p2,p3)
w1
w2

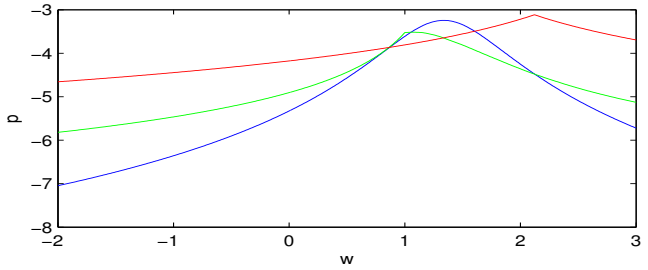

gauss
y,f(x)
x


p1
p2
p3
p
w


uniform
y,f(x)
x


p
w


exp
y,f(x)
x


p
w

Гипотеза порождения данных

$y = f_k(\mathbf{w}, \mathbf{x}) + \nu$ — модель с аддитивным Гауссовским шумом с дисперсией σ_ν, $\beta = \frac{1}{\sigma_\nu^2}$.

Плотность вероятности появления данных

$$p(D|\mathbf{w}, \beta, f_k) = \frac{\exp(-\beta E_D(D|\mathbf{w}, f_k))}{Z_D(\beta)},$$

где

$$E_D = \frac{\beta}{2} \sum_{n=1}^{N} (f_k(\mathbf{x}_n) - y_n)^2.$$

Гипотеза порождения параметров

α — регуляризующий параметр, $\alpha = \frac{1}{\sigma^2_{\mathbf{w}}}$.

Плотность вероятности параметров

$$p(\mathbf{w}|\alpha, f_k) = \frac{\exp(-\alpha E_W(\mathbf{w}|f_k))}{Z_W(\alpha)},$$

где

$$E_W = \frac{\alpha}{2}||\mathbf{w}||^2.$$

Критерий с учетом гиперпараметров

При заданных значениях гиперпараметров α, β и фиксированной функции f_k

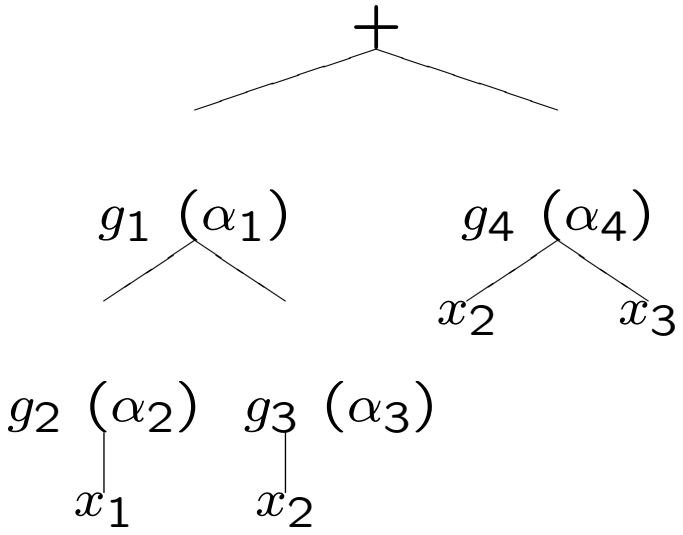
$$p(\mathbf{w}|D, \alpha, \beta, f_k) = \frac{p(D|\mathbf{w}, \beta, f_k)p(\mathbf{w}|\alpha, f_k)}{p(D|\alpha, \beta)} =$$

$$= \frac{\exp(-s(\mathbf{w}|f_k))}{Z_s(\alpha, \beta)}.$$

$s(\mathbf{w}|f_k) = \alpha E_W + \beta E_D$ — критерий (целевая функция).

Каждому параметру или группе параметров можно поставить в соответствие гиперпараметр α; каждой подвыборке можно поставить в соответствие гиперпараметр β.

Пример модели



$$f = g_1\left(g_2(x_1), g_3(x_2)\right) + g_4(x_2, x_3)$$

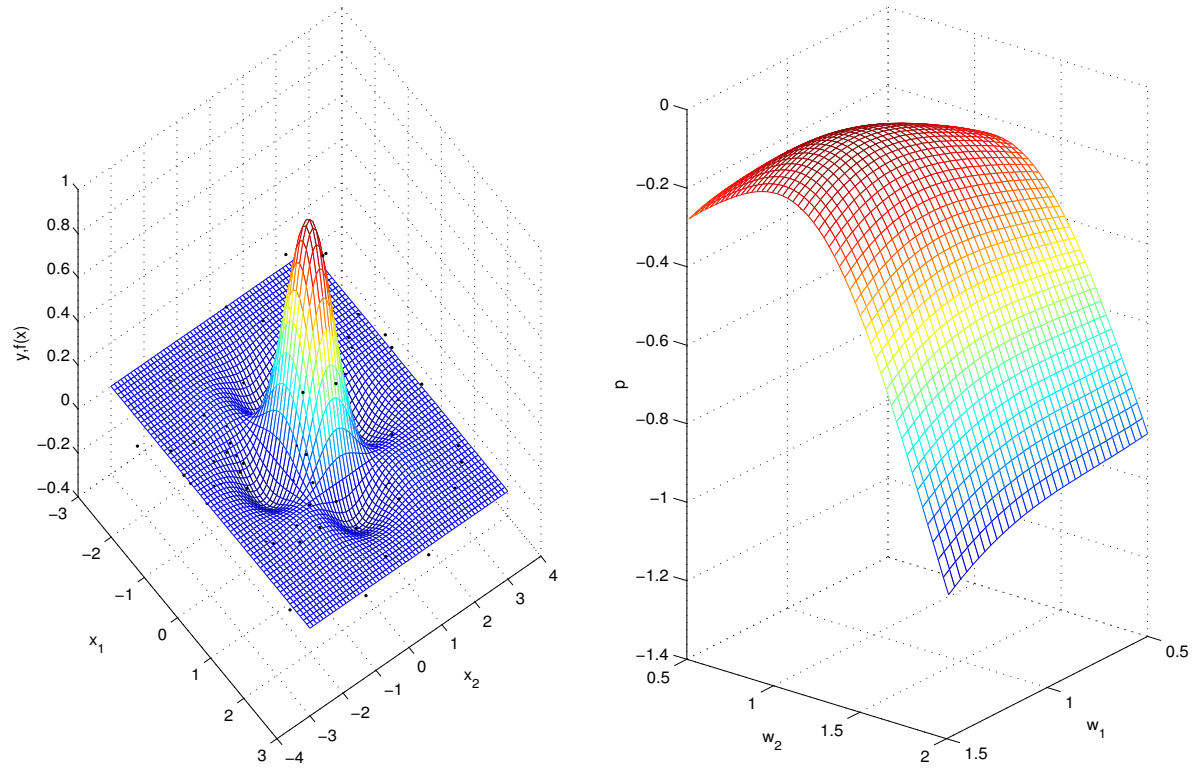
Связный Байесовский вывод

Первый уровень

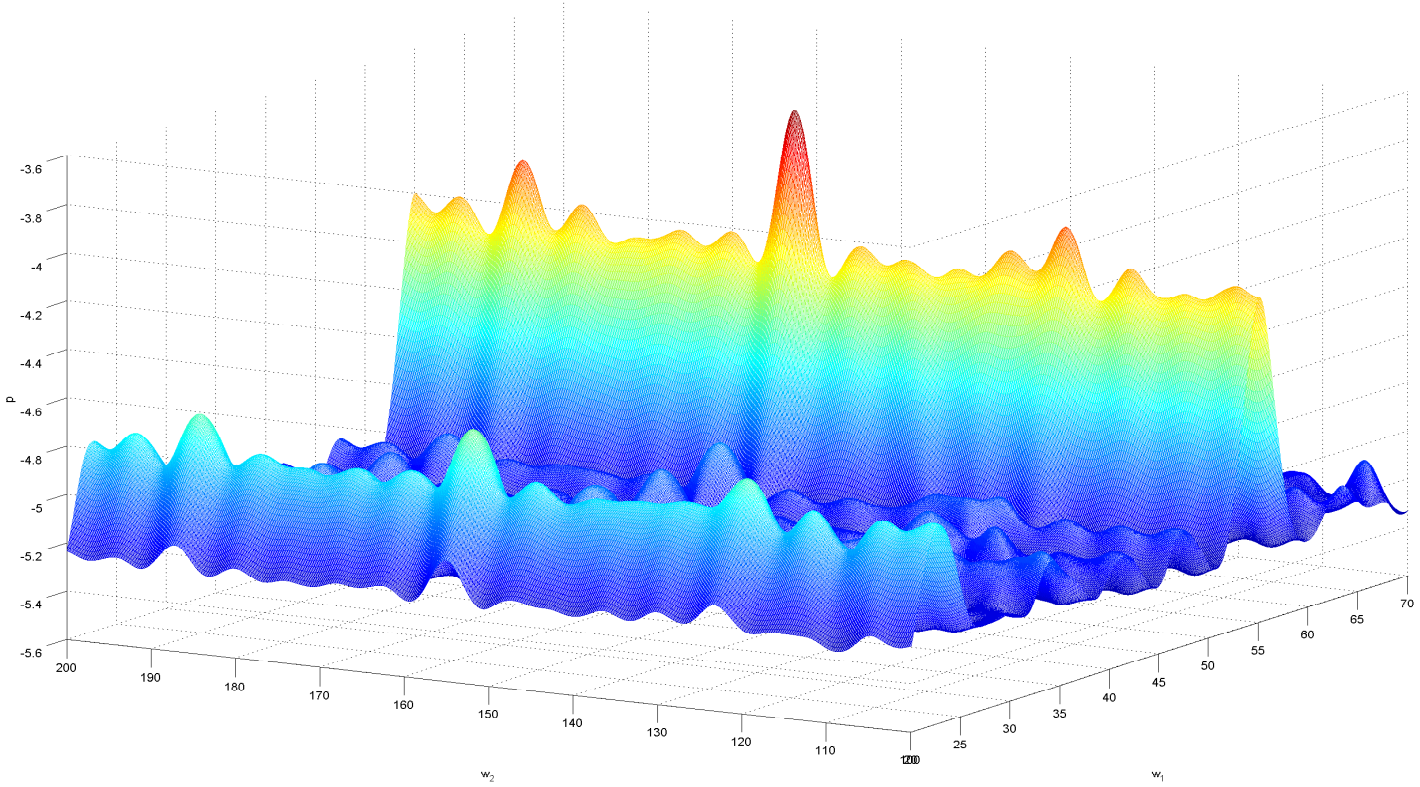
$$p(\mathbf{w}|D, f_k) = \frac{p(D|\mathbf{w}, f_k)p(\mathbf{w}|f_k)}{\int p(D|\mathbf{w}', f_k)p(\mathbf{w}'|f_k)d\mathbf{w}'}$$

Второй уровень

$$P(f_k|D) = \frac{p(D|f_k)P(f_k)}{\sum_{i=1}^{M} p(D|f_k)P(f_k)}$$



Модель с нормальным распределением случайной переменной, унимодальная функция правдоподобия



Регрессионная модель — сумма синусов
мультимодальная функция правдоподобия

Задача совместного индуктивного поиска

Дано:
$\{(\mathbf{x}_i, y_i)\}$ — конечная выборка,
$G = \{g_j : \mathbf{x} \mapsto g_j(\mathbf{x})\}_{j \in J}$ — множество порождающих функций,
G задает множество суперпозиций

$$\mathcal{F} = \{f_k(\mathbf{w}, \mathbf{x})\}_{k<K}$$

путем обобщенного индуктивного определения.
Множество функций $\mathcal{P} = \{p | \int p(\mathbf{w}|D, f)d\mathbf{w} = 1\}$ задает множество критериев $\mathcal{S} = \{s\}$.

Найти:
множество пар $\{(\bar{f}_k, \bar{s}_k)\}$, принадлежащих POF$(\mathcal{F}, \mathcal{S})$, где

$$\bar{s}_k = \max p(\mathbf{w}|D, f_k).$$

Алгоритм индуктивного поиска

Заданы $D, G, \mathcal{P}$ и начальный набор моделей $F_0 = \{f_1, ..., f_M\}$.

Для каждой гипотезы порождения данных из $\mathcal{P}$:
Начало итераций
Для каждой модели f_k минимизируем критерий $s(f_k(\mathbf{w}))$.
Определяем значения гиперпараметров α_{kj} и β_k.
Порождаем производные модели $f'_1, ..., f'_M$.
Выбираем M наилучших моделей.
Конец итераций

Строим Парето-оптимальный фронт в пространстве критериев.

Дивергенция Дженсена-Шеннона

Метризация пространства критериев, введение функции расстояния между функциями плотности распределения p, q случайной переменной соответствующих моделей.

$$JSD(p\|q) = \frac{1}{2}D(p\|v) + \frac{1}{2}D(q\|v),$$

$$\text{где} \quad D = -\int p(\mathbf{z}) \ln \frac{q(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z}$$

есть дивергенция Куллбэка-Лейблера и

$$v = \frac{1}{2}(p + q).$$

Заключение

1. Поставлена задача совместного поиска регрессионных моделей и гипотез порождения данных.

2. Существуют пары «оптимальная модель — гипотеза порождения данных».

3. Известна зависимость структуры модели от гипотезы порождения данных.

4. Создана система автоматического построения оптимальных регрессионных моделей, см. http://strijov.com — открытый код Matlab.