О построении сплайн-регрессионной модели

Д.т.н., проф., Шумейко А.А.

ОКВУЗ «Институт предпринимательства «Стратегия»»,

Метод наименьших квадратов

Пусть дана экспериментальная таблица

x_i	x_0	x_1	$\cdots$	x_n
t_i	t_0	t_1	$\cdots$	t_n

Поставим ей в соответствие функцию вида

$$F\left(\{a_i\}\Big|_{i=0}^{n}, t\right) = a_0\varphi_0(t) + a_1\varphi_1(t) + \ldots + a_m\varphi_m(t)$$

где $\varphi_k(t)$, *k = 1,…,N* - базисные функции, a_k - коэффициенты, подлежащие определению. В частности, если в качестве базисных функций использовать степенные $\varphi_k(t) = t^k$, задача сводится к поиску полинома степени m *(m<n),*приближающего исходную таблицу.

С целью определения коэффициентов a_k будем искать такую функцию $F\left(\{a_i\}\Big|_{i=0}^{n}, t\right)$, отклонение значений которой от

заданных таблицей значений x_i минимально в некотором среднеинтегральном смысле

Метод наименьших квадратов

В дискретном методе наименьших квадратов строится функционал

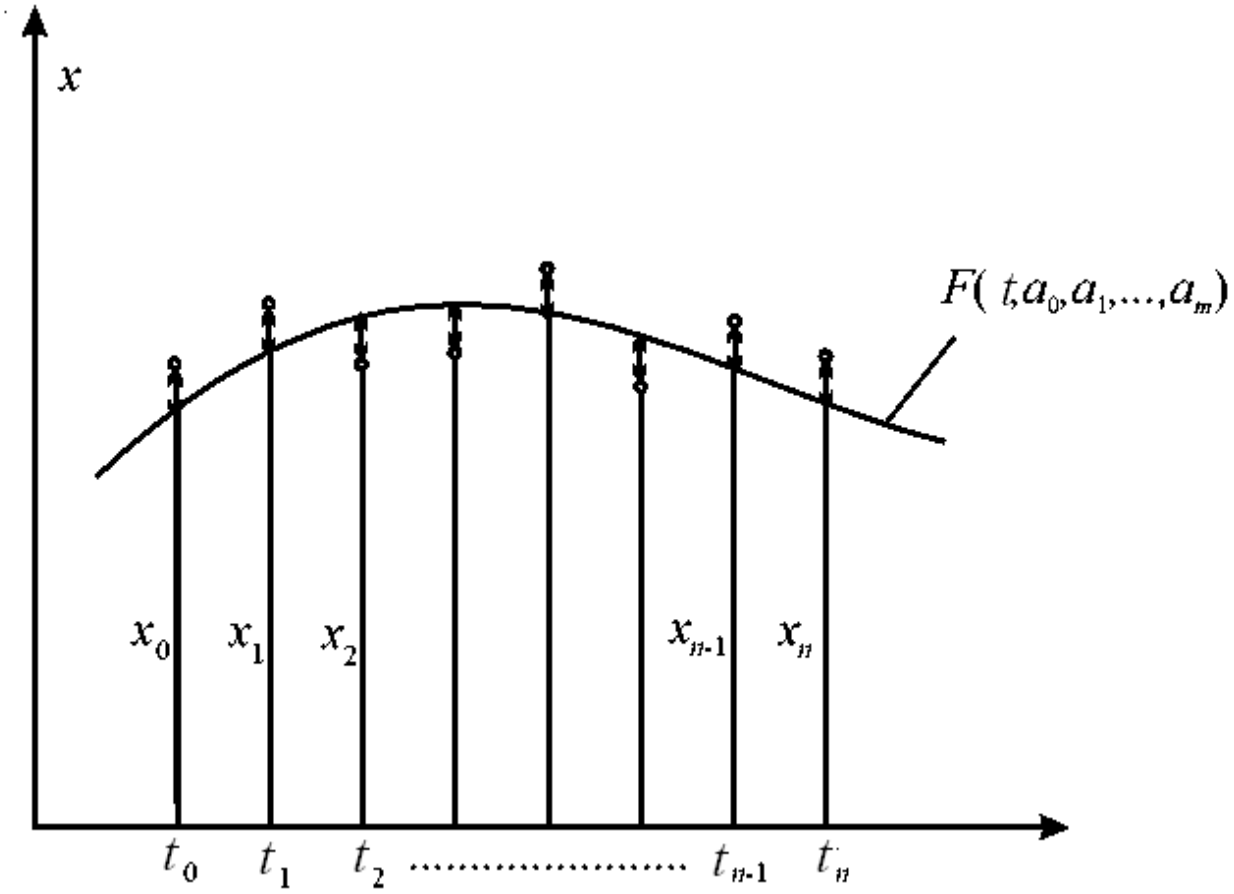
$$S(a_0, a_0, \ldots, a_m) = \sum_{i=0}^{n} \left(F(a_0, a_0, \ldots, a_m, t_i) - x_i\right)^2$$

который геометрически представляет собой сумму квадратов отклонений значений x_i от значений аппроксимирующей функции в точках $t_i\ i = 0, \ldots, n.$

Необходимым (а в данном случае и достаточным) условием минимума функции многих переменных является равенство нулю ее частных производных первого порядка по независимым переменным.

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 2\sum\limits_{i=0}^{n} \left(F(a_0, a_0, \ldots, a_m, t_i) - x_i\right)\varphi_0(t_i) = 0, \\ \dfrac{\partial S}{\partial a_1} = 2\sum\limits_{i=0}^{n} \left(F(a_0, a_0, \ldots, a_m, t_i) - x_i\right)\varphi_1(t_i) = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial S}{\partial a_m} = 2\sum\limits_{i=0}^{n} \left(F(a_0, a_0, \ldots, a_m, t_i) - x_i\right)\varphi_m(t_i) = 0. \end{cases}$$

Метод наименьших квадратов



Кусочно-линейная регрессия

Рассмотрим в качестве регрессионной модели ломаную.

Тогда задача нахождения кусочно-регрессионной модели с фиксированными узлами пример вид

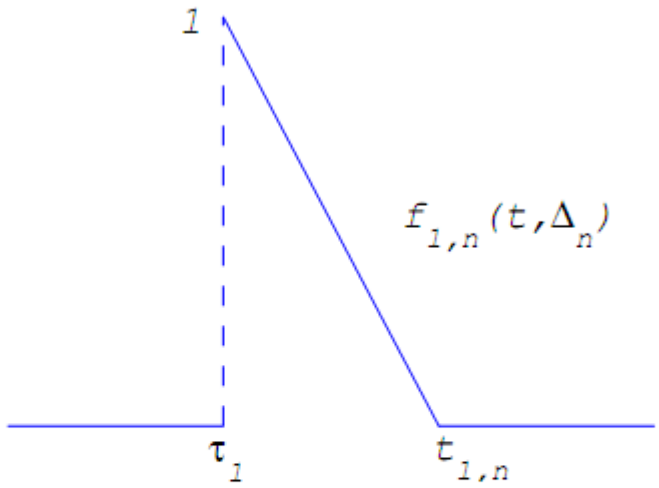
$$\inf_{\Delta_n}\left(\sum_{i=0}^{k}|x_i - P_{1,n}(\tau_i, x, \Delta_n)|^2\right) = \sum_{i=0}^{k}|x_i - P_{1,n}(\tau_i, x, \Delta_n^*)|^2.$$

где

$$P_{1,n}(t, \Delta_n, b_1, \ldots, b_n) = \sum_{i=1}^{n} b_i f_{i,n}(t, \Delta_n),$$

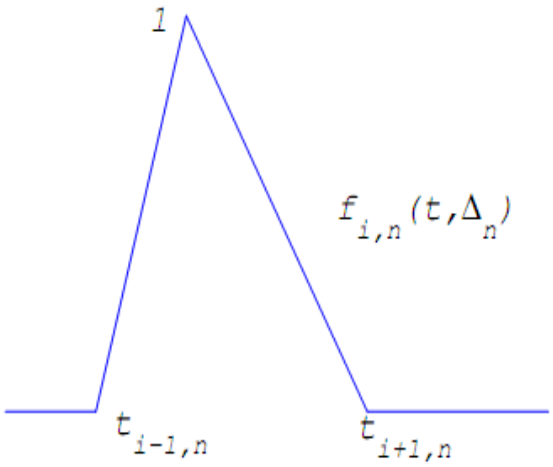
и $f_{i,n}(t,\Delta_n)$ $i=1,\ldots,n$ базисные функции, которые можно записать в следующем виде

Базисные функции



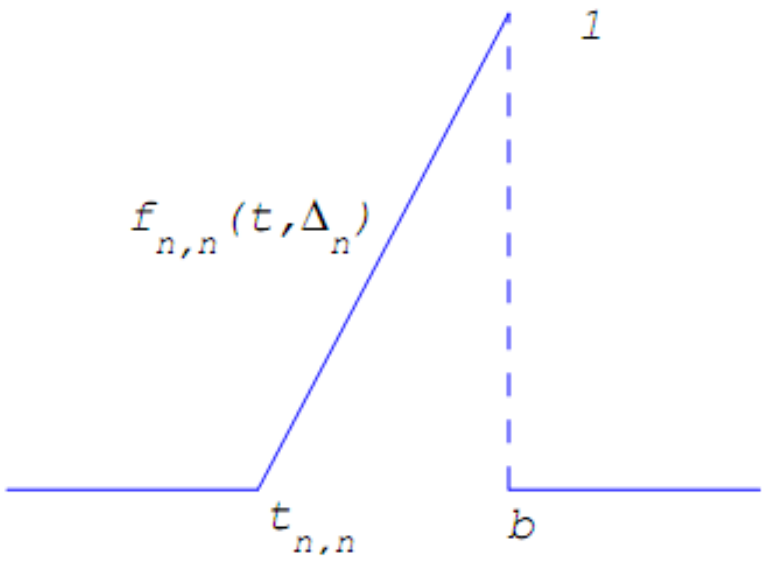
$$f_{1,n}(t,\Delta_n)=\begin{cases}(t_1-t)(t_1-a)^{-1}, & t\in[a,t_1],\\ 0, & sonst.\end{cases},$$

Базисные функции



$$f_{i,n}(t,\Delta_n) = \begin{cases} h_{i,n}^{-1}(t-t_{i-1}), & t \in [t_{i-1},t_i], \\ h_{i+1,n}^{-1}(t_{i+1}-t), & t \in [t_i,t_{i+1}], \\ 0 & , sonst. \end{cases} , \quad (i=2,\ldots,n-1),$$

Базисные функции



$$f_{n,n}(t,\Delta_n)=\begin{cases}(b-t)(b-t_n)^{-1}, & t\in[t_n,b],\\ 0, & sonst.\end{cases}.$$

Кусочно-линейная регрессия

Выпишем функцию цели

$$F(b_1, \ldots, b_n, \Delta_n) = \sum_{i=1}^{k} \left(x_i - \sum_{j=1}^{n} b_j f_{j,n}(\tau_i, \Delta_n) \right)^2.$$

и найдем решение задачи

$$\inf_{b_1,\ldots,b_n} F(b_1, \ldots, b_n, \Delta_n) = F(b_1^*, \ldots, b_n^*, \Delta_n).$$

Необходимое и достаточное условие экстремума будет иметь вид

$$\left. \frac{\partial F(b_1, \ldots, b_n, \Delta_n)}{\partial b_i} \right|_{b_1^*,\ldots,b_n^*} = 0, \quad i = 1, \ldots, n.$$

через $(\psi, \phi) = \sum_{i=0}^{k} \psi(\tau_i)\phi(\tau_i)$ обозначим скалярное произведение

Кусочно-линейная регрессия

Нахождение экстремума сводится к решению системы уравнений

$$\begin{pmatrix} \langle f_{1,n}, f_{1,n}\rangle & \langle f_{1,n}, f_{2,n}\rangle & \dots & \langle f_{1,n}, f_{n,n}\rangle \\ \langle f_{2,n}, f_{1,n}\rangle & \langle f_{2,n}, f_{2,n}\rangle & \dots & \langle f_{2,n}, f_{n,n}\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle f_{n,n}, f_{1,n}\rangle & \langle f_{n,n}, f_{2,n}\rangle & \dots & \langle f_{n,n}, f_{n,n}\rangle \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} \langle x, f_{1,n}\rangle \\ \langle x, f_{2,n}\rangle \\ \vdots \\ \langle x, f_{n,n}\rangle \end{pmatrix}$$

где

$$\langle x, f_{j,n}\rangle = \sum_{i=0}^{k} x_i f_{j,n}(\tau_i, \Delta_n) \ , \ j = 1, \dots, n.$$

а замечая, что

$$\langle f_{i,n}, f_{j,n}\rangle = 0 \ , \ \forall i, j : \ |i - j| \geq 2,$$

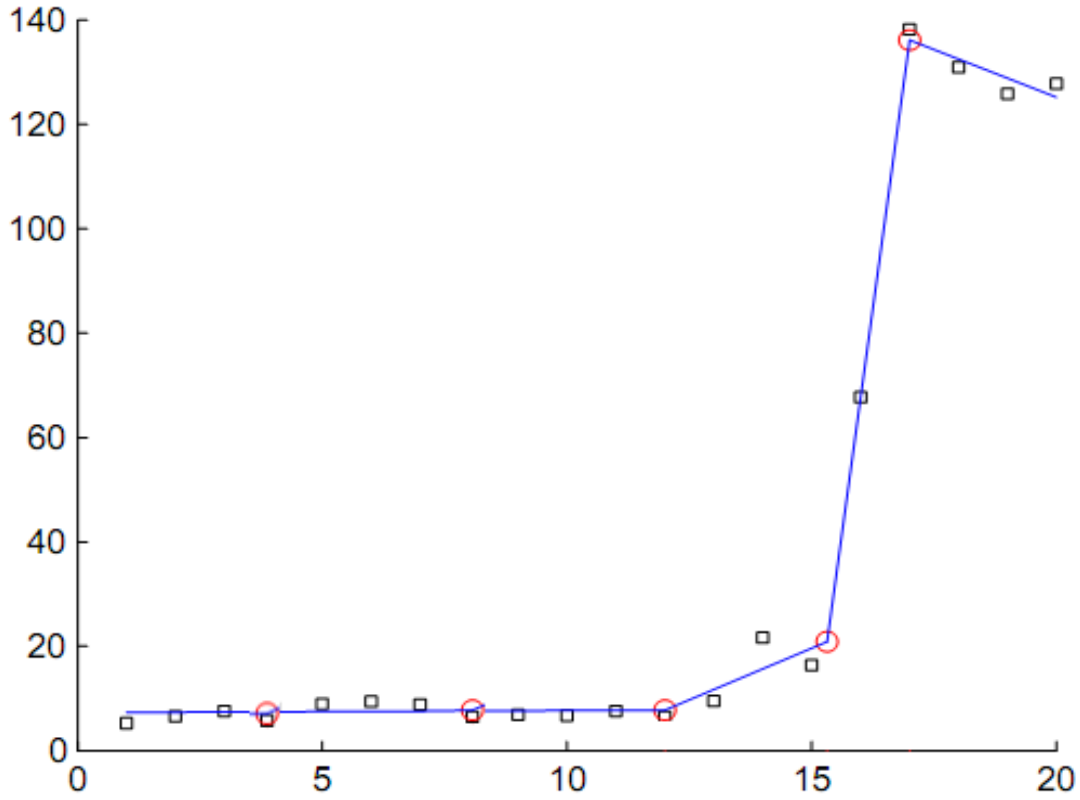
Кусочно-линейная регрессия

Получаем систему уравнений с трехдиагональной матрицей

$$A = \begin{pmatrix} \langle f_{1,n}, f_{1,n} \rangle & \langle f_{1,n}, f_{2,n} \rangle & 0 & \dots & 0 \\ \langle f_{2,n}, f_{1,n} \rangle & \langle f_{2,n}, f_{2,n} \rangle & \langle f_{2,n}, f_{3,n} \rangle & \dots & 0 \\ 0 & \langle f_{3,n}, f_{2,n} \rangle & \langle f_{3,n}, f_{3,n} \rangle & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \langle f_{n,n}, f_{n,n} \rangle \end{pmatrix}.$$

Применяя метод прогонки, получаем эффективный алгоритм нахождения уравнения кусочно-линейной регрессии с фиксированными узлами.

Пример



Сплайн- регрессионная модель

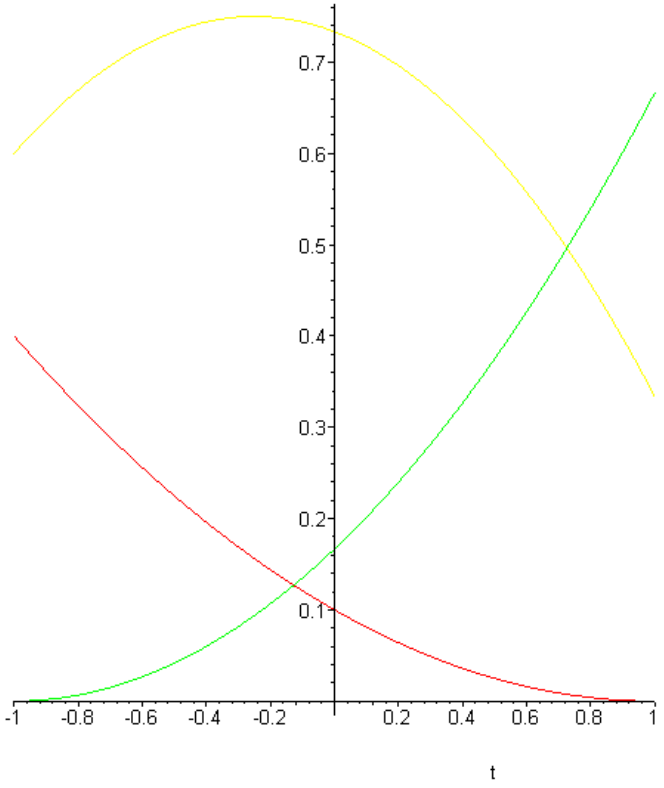
Рассмотрим в качестве регрессионной модели параболический сплайн минимального дефекта

$$s_2(x) = \sum_i C_{i+1/2} B_2\left(\frac{1}{h_{i+1/2}}\left(x - \frac{x_i + x_{i+1}}{2}\right)\right) =$$

$$C_{i-1/2}\frac{(x_{i+1} - x)^2}{(h_{i+1/2} + h_{i-1/2})h_{i+1/2}} +$$

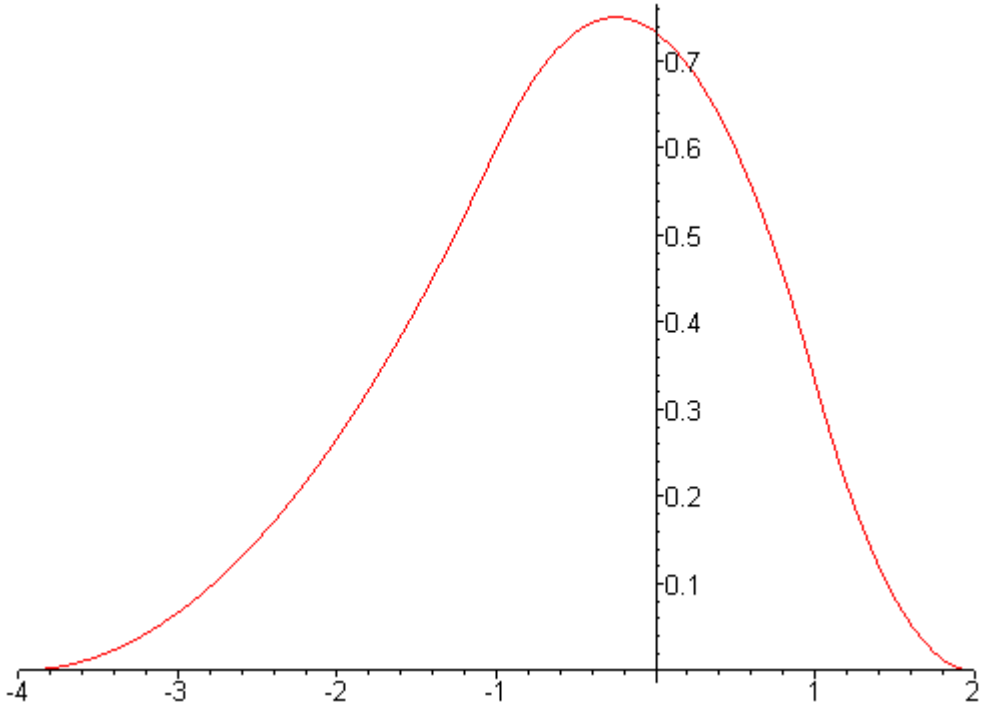
$$C_{i+1/2}\left(\frac{(x_{i+1} - x)(x - x_{i-1})}{(h_{i+1/2} + h_{i-1/2})h_{i+1/2}} + \frac{(x_{i+2} - x)(x - x_i)}{(h_{i+3/2} + h_{i+1/2})h_{i+1/2}}\right) +$$

$$C_{i+3/2}\frac{(x_i - x)^2}{(h_{i+3/2} + h_{i+1/2})h_{i+1/2}}$$



B-сплайн

Параболический B-сплайн имеет вид



Сплайн- регрессионная модель

Для фиксированного набора узлов $\{x_i\}_{i=0}^{n}$ коэффициенты $\{C_{i+1/2}\}_{i=0}^{n}$ будем искать из условия

$$\Phi\left(\{C_{i+1/2}\}_{i=1}^{n-1}\right) = \int_{x_0}^{x_n}\left(f(x) - \sum_i C_{i+1/2}B_2\left(\frac{1}{h_{i+1/2}}\left(x - \frac{x_i + x_{i+1}}{2}\right)\right)\right)^2 dx \to \min$$

Условия экстремума будут иметь вид

$$\frac{\partial\Phi\left(\{C_{i+1/2}\}_{i=1}^{n-1}\right)}{\partial C_{\nu+1/2}} = -2\int_{x_0}^{x_n}\left(f(x) - \sum_i C_{i+1/2}B_2\left(\frac{1}{h_{i+1/2}}\left(x - \frac{x_i + x_{i+1}}{2}\right)\right)\right)B_2\left(\frac{1}{h_{\nu+1/2}}\left(x - \frac{x_\nu + x_{\nu+1}}{2}\right)\right)dx = 0.$$

Учитывая локальность носителя B-сплайна, получаем систему уравнений

$$MC = F,$$

где

Сплайн- регрессионная модель

$$M = \begin{pmatrix} b_{0,0} & b_{0,1} & b_{0,2} & 0 & 0 & \cdots & 0 \\ b_{1,0} & b_{1,1} & b_{1,2} & b_{1,3} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & b_{n,n} \end{pmatrix} \quad C = \begin{pmatrix} C_{1/2} \\ C_{3/2} \\ \vdots \\ C_{n+1/2} \end{pmatrix} \quad F = \begin{pmatrix} \varphi_0 \\ \varphi_1 \\ \vdots \\ \varphi_n \end{pmatrix}$$

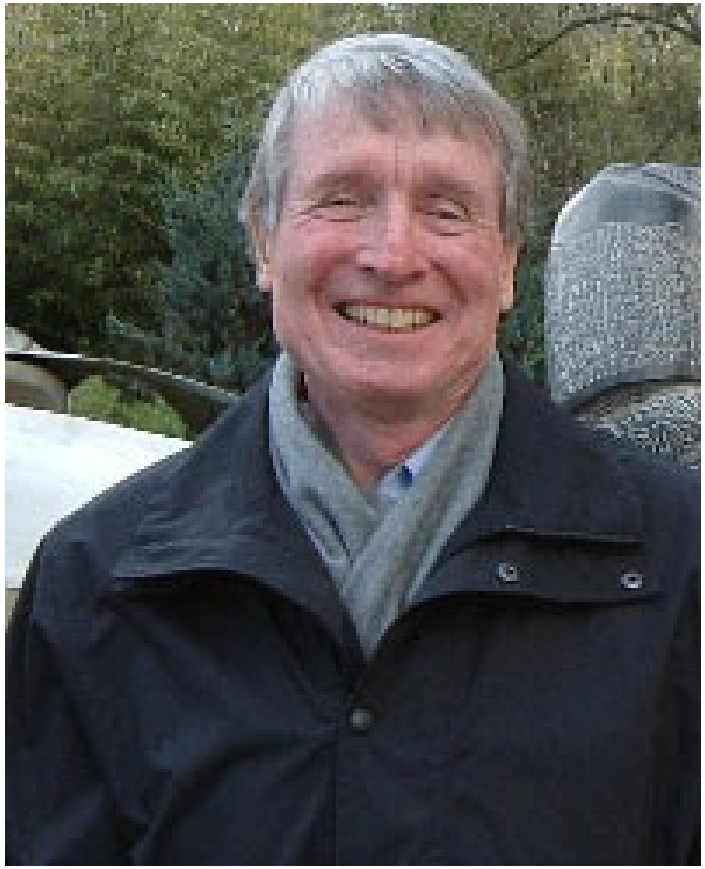
здесь

$$b_{i,j} = \int_{x_0}^{x_n} B_2\left(\frac{1}{h_{i+1/2}} \left(x - \frac{x_i + x_{i+1}}{2} \right) \right) B_2\left(\frac{1}{h_{j+1/2}} \left(x - \frac{x_j + x_{j+1}}{2} \right) \right) dx$$

и

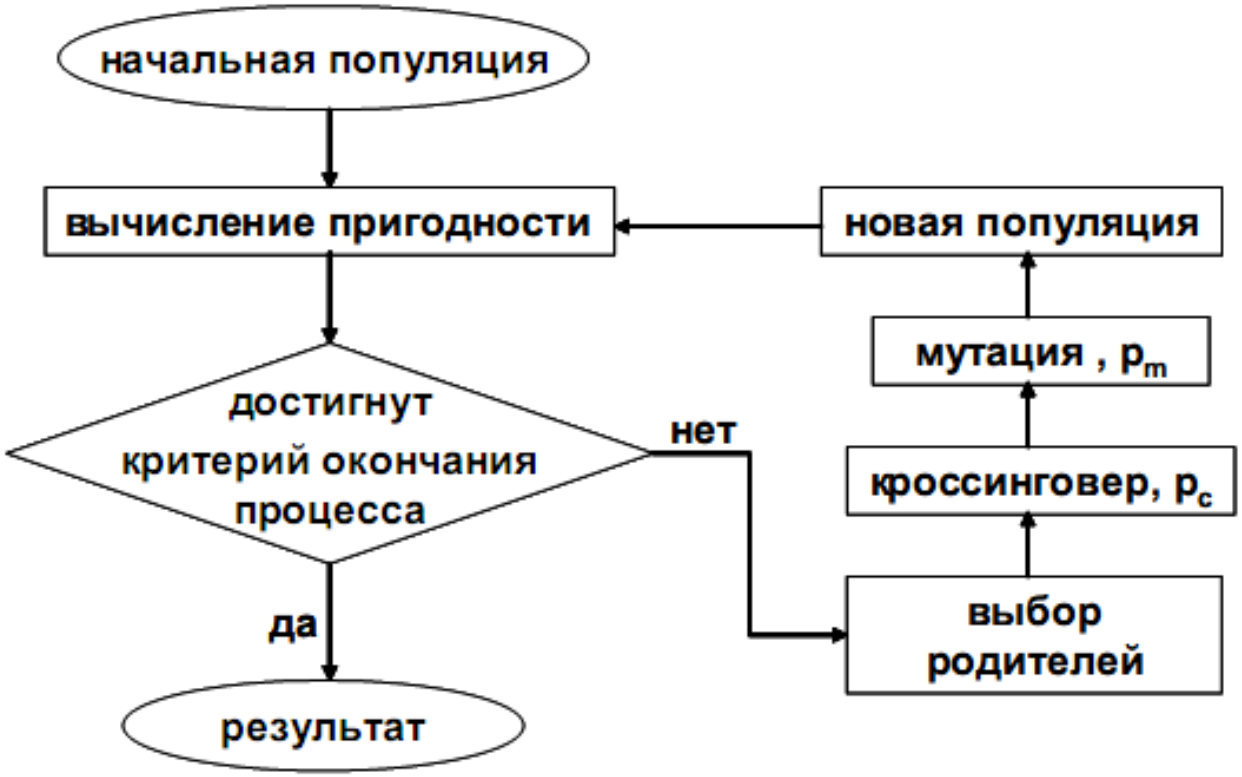
$$\varphi_i = \int_{x_0}^{x_n} f(x) B_2\left(\frac{1}{h_{i+1/2}} \left(x - \frac{x_i + x_{i+1}}{2} \right) \right) dx$$

John Holland



«Отцом-основателем» генетических алгоритмов считается Джон Холланд (*John Holland*), книга которого «Адаптация в естественных и искусственных системах» (*Adaptation in Natural and Artificial Systems*) является основополагающим трудом в этой области исследований.

Основные принципы ГА



Использование ГА для оптимизации узлов кусочно-линейной регрессии

Для отбора родителей использовался оператор селекции, а рекомбинация генов потомков определялась из условия

Потомок=Родитель(1)+α(Родитель(2)-Родитель(1)),

где значение параметра α выбиралось случайным образом из промежутка [-d,1+d], $d \geq 0$. В нашем случае использовалось значение d=0.25.

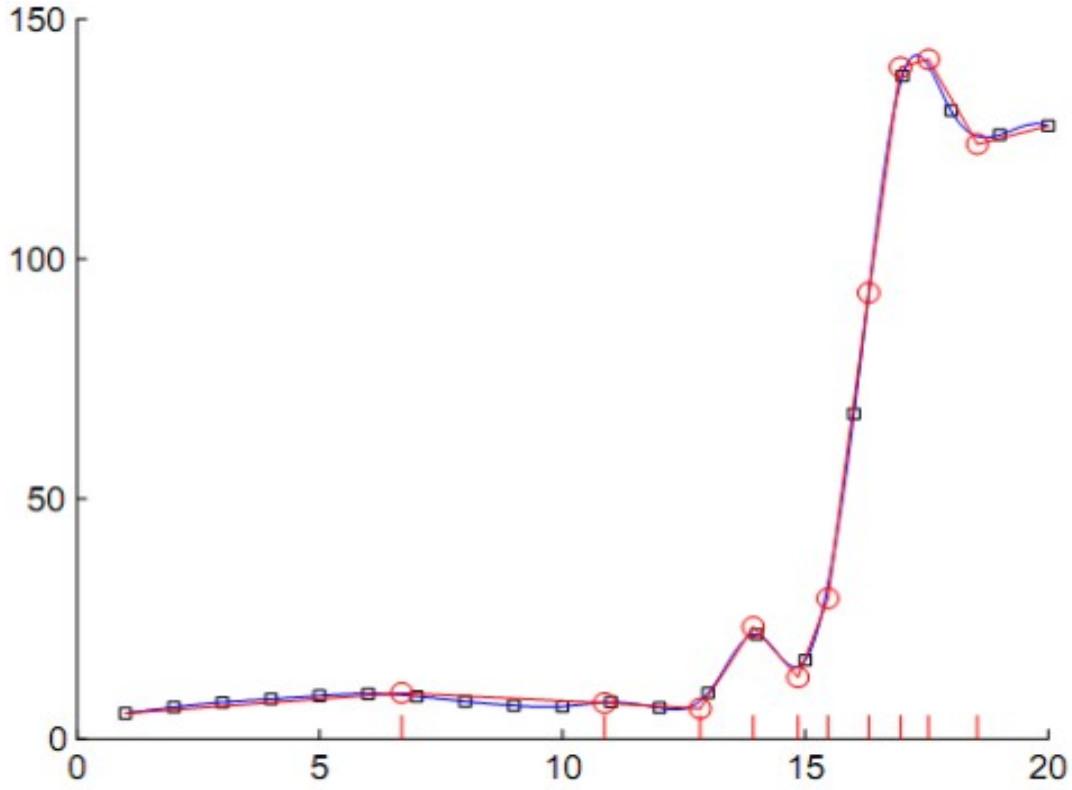
Для каждого гена потомка выбиралось свое значение параметра α.

Гены могут мутировать согласно правилу

Новое значение=Старое значение $\pm \delta \times h$,

где $0<\delta<0.5$ и h выбиралось равным текущему значению гена.

Пример



Спасибо за внимание!

Оптимизация узлов для сплайнов

Пусть - $\Delta_{n[a,b]} = \{a = t_{0,n} < \ldots < t_{n,n} = b\}$ произвольное разбиение отрезка *[a,b]* и - $S_{r,k}(\Delta_{n[a,b]})$ множество всех сплайнов порядка r дефекта k по разбиению $\Delta_{n[a,b]}$ то есть множество функций с непрерывной *(r-k)*-й производной на *[a,b],* совпадающих на каждом промежутке $(t_{i,n}, t_{i+1,n})(i = 0,1,\ldots,n-1)$ с алгебраическим полиномом степени не выше r.

Через $P(\Delta_{n[a,b]})$ обозначим оператор отображающий $C^{\rho}_{[a.b]}$ в $S_{r,k}(\Delta_{n[a,b]})$

$\rho \geq r - k$.

При фиксированных r и k последовательность $\{P^*(\Delta^*_{n[a,b]})\}_{n=1}^{\infty}$ будем называть асимптотически наилучшей для функции $x(t)$, если при $n \to \infty$

$$\int_a^b \left(x(t) - P^*(\Delta^*_{n[a,b]}, t)\right)^2 dt = \inf_{\Delta n[a,b]} \inf \left\{ \int_a^b \left(x(t) - P(\Delta_{n[a,b]}, t)\right)^2 dt \,\middle|\, P(\Delta_{n[a,b]}) \in S_{r,k}(\Delta_{n[a,b]}) \right\} (1 + o(1))$$

Сплайн $s_3(x, \Delta_{n[a,b]})$ называется интерполяционным для функции $x(t)$, если $S_{3,1}(\Delta_{n[a,b]})$

$$s_3(x, \Delta_{n[a,b]}, t_{i,n}) = x(t_{i,n}), i = 1,2,\ldots,n-1$$

Основной результат

Доказано, что для любого разбиения $\Delta_{n[a,b]}$

$$\min\left\{\int_a^b \left(x(t) - \ell(t)\right)^2 dt \,\middle|\, \ell(t) \in S_{1,1}\left(\Delta_{n[a,b]}\right)\right\} = \int_a^b \left(x(t) - s_3''\left(X, \Delta_{n[a,b]}, t\right)\right)^2 dt$$

где $X(t)$ - вторая первообразная $x(t)$ такая, что $X(a) = 0.$ Кроме того, если узлы разбиения $\Delta^*_{n[a,b]}$ выбраны из условия

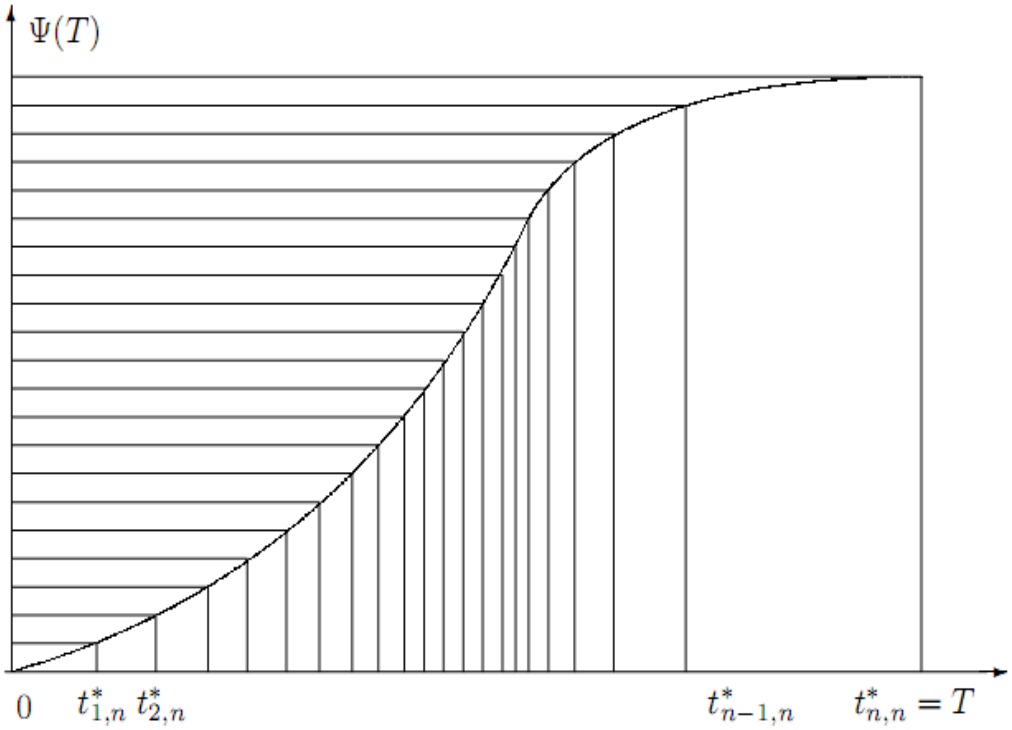
$$\int_a^{t^*_{i,n}} \left(|x''(t)| + n^{-1/4}\right)^{2/5} dt = \frac{i}{n}\int_a^b \left(|x''(t)| + n^{-1/4}\right)^{2/5} dt, i = 0,1,\ldots,n,$$

то при $n \to \infty$

$$\inf_{\Delta n[a,b]} \inf\left\{\int_a^b \left(x(t) - \ell(t)\right)^2 dt \,\middle|\, \ell \in S_{1,1}(\Delta_{n[a,b]})\right\} = \int_a^b \left(x(t) - s_3''\left(X, \Delta^*_{n[a,b]}, t\right)\right)^2 dt(1 + o(1)) =$$

$$= \frac{1}{720n^4}\left\|x''\right\|^2_{2/5[a.b]}\left(1 + o(1)\right).$$

Геометрическая иллюстрация выбора оптимальных узлов кусочно-линейной регрессии



где $$\Psi(\tau) = \int_a^{\tau} \left(|x''(t)| + n^{-1/4} \right)^{2/5} dt$$

Пример