Нелинейная регрессия и нестандартные функции потерь

Воронцов Константин Вячеславович
vokov@forecsys.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

Видеолекции: http://shad.yandex.ru/lectures

2 апреля 2015

Содержание

Метод наименьших квадратов

- X — объекты (часто $\mathbb{R}^n$); Y — ответы (часто $\mathbb{R}$, реже $\mathbb{R}^m$);
  $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка;
  $y_i = y(x_i)$, $y\colon X \to Y$ — неизвестная зависимость;
- $a(x) = f(x, \alpha)$ — модель зависимости,
  $\alpha \in \mathbb{R}^p$ — вектор параметров модели.
- Метод наименьших квадратов (МНК):

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} w_i \bigl(f(x_i, \alpha) - y_i\bigr)^2 \to \min_\alpha,$$

где w_i — вес, степень важности i-го объекта.

$Q(\alpha^*, X^\ell)$ — *остаточная сумма квадратов* (residual sum of squares, RSS).

Нелинейная модель регрессии

Нелинейная модель регрессии $f(x, \alpha)$, $\alpha \in \mathbb{R}^p$.
Функционал среднеквадратичного отклонения:

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} \bigl(f(x_i, \alpha) - y_i\bigr)^2 \to \min_{\alpha}.$$

Метод Ньютона–Рафсона:

1. Начальное приближение $\alpha^0 = (\alpha_1^0, \dots, \alpha_p^0)$.
2. Итерационный процесс

$$\alpha^{t+1} := \alpha^t - h_t \bigl(Q''(\alpha^t)\bigr)^{-1} Q'(\alpha^t),$$

$Q'(\alpha^t)$ — градиент функционала Q в точке α^t,
$Q''(\alpha^t)$ — гессиан функционала Q в точке α^t,
h_t — величина шага (можно полагать $h_t = 1$).

Метод Ньютона-Рафсона

Компоненты градиента:

$$\frac{\partial Q(\alpha)}{\partial \alpha_j} = 2\sum_{i=1}^{\ell}\bigl(f(x_i,\alpha) - y_i\bigr)\frac{\partial f(x_i,\alpha)}{\partial \alpha_j}.$$

Компоненты гессиана:

$$\frac{\partial^2 Q(\alpha)}{\partial \alpha_j \partial \alpha_k} = 2\sum_{i=1}^{\ell}\frac{\partial f(x_i,\alpha)}{\partial \alpha_j}\frac{\partial f(x_i,\alpha)}{\partial \alpha_k} \underbrace{-2\sum_{i=1}^{\ell}\bigl(f(x_i,\alpha) - y_i\bigr)\frac{\partial^2 f(x_i,\alpha)}{\partial \alpha_j \partial \alpha_k}}_{\text{при линеаризации полагается } = 0}.$$

Не хотелось бы обращать гессиан на каждой итерации...

Линеаризация $f(x_i,\alpha)$ в окрестности текущего α^t:

$$f(x_i,\alpha) = f(x_i,\alpha^t) + \sum_{j=1}^{p}\frac{\partial f(x_i,\alpha_j)}{\partial \alpha_j}(\alpha_j - \alpha_j^t) + o(\alpha_j - \alpha_j^t).$$

Метод Ньютона-Гаусса

Матричные обозначения:
$F_t = \left(\frac{\partial f}{\partial \alpha_j}(x_i, \alpha^t)\right)_{\ell \times p}$ — матрица первых производных;
$f_t = \big(f(x_i, \alpha^t)\big)_{\ell \times 1}$ — вектор значений f.

Формула t-й итерации метода Ньютона–Гаусса:

$$\alpha^{t+1} := \alpha^t - h_t \underbrace{(F_t^\top F_t)^{-1} F_t^\top (f_t - y)}_{\beta}.$$

β — это решение задачи многомерной линейной регрессии

$$\|F_t \beta - (f_t - y)\|^2 \to \min_\beta.$$

Нелинейная регрессия сведена к серии линейных регрессий.

Скорость сходимости — как и у метода Ньютона–Рафсона, но для вычислений можно применять стандартные методы.

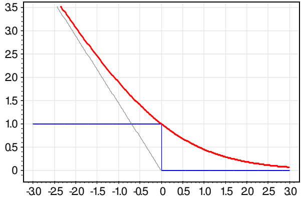
Задача классификации. Логистическая регрессия

$Y = \{-1, +1\}$ — два класса, $a(x, w) = \operatorname{sign}(w^\top x)$, $x, w \in \mathbb{R}^n$.

Функционал аппроксимированного эмпирического риска:

$$Q(w) = \sum_{i=1}^{\ell} \left[M_i(w) < 0\right] \leqslant \sum_{i=1}^{\ell} \mathscr{L}\left(w^\top x_i y_i\right) \to \min_w,$$

где $\mathscr{L}(M) = \log\left(1 + e^{-M}\right)$ — логарифмическая функция потерь



$M_i = w^\top x_i y_i$

Метода Ньютона-Рафсона

Метода Ньютона-Рафсона для минимизации функционала $Q(w)$:

$$w^{t+1} := w^t - h_t \bigl(Q''(w^t)\bigr)^{-1} Q'(w^t),$$

Элементы градиента — вектора первых производных $Q'(w^t)$:

$$\frac{\partial Q(w)}{\partial w_j} = -\sum_{i=1}^{\ell} (1-\sigma_i) y_i f_j(x_i), \quad j = 1, \ldots, n.$$

Элементы гессиана — матрицы вторых производных $Q''(w^t)$:

$$\frac{\partial^2 Q(w)}{\partial w_j \partial w_k} = \sum_{i=1}^{\ell} (1-\sigma_i) \sigma_i f_j(x_i) f_k(x_i), \quad j, k = 1, \ldots, n,$$

где $\sigma_i = \sigma(y_i w^{\mathsf{T}} x_i)$, $\sigma(z) = \frac{1}{1+e^{-z}}$ — сигмоидная функция.

Матричные обозначения

$F = \big(f_j(x_i)\big)_{\ell\times n}$ — матрица «объекты–признаки»;
$\Gamma = \mathrm{diag}\big(\sqrt{(1-\sigma_i)\sigma_i}\,\big)$ — диагональная $\ell\times\ell$-матрица;
$\tilde F = \Gamma F$ — взвешенная матрица «объекты–признаки»;
$\tilde y_i = y_i\sqrt{(1-\sigma_i)/\sigma_i}$, $\tilde y = (\tilde y_i)_{\ell\times 1}$ — взвешенный вектор ответов.

Тогда в методе Ньютона-Рафсона:

$$\big(Q''(w)\big)^{-1}Q'(w) = -(F^{\mathsf T}\Gamma^2 F)^{-1}F^{\mathsf T}\Gamma\tilde y = -(\tilde F^{\mathsf T}\tilde F)^{-1}\tilde F^{\mathsf T}\tilde y = -\tilde F^{+}\tilde y.$$

Это совпадает с МНК-решением линейной задачи регрессии со взвешенными объектами и модифицированными ответами:

$$Q(w) = \|\tilde F w - \tilde y\|^2 = \sum_{i=1}^{\ell}(1-\sigma_i)\sigma_i\Big(w^{\mathsf T}x_i - \frac{y_i}{\sigma_i}\Big)^2 \to \min_w .$$

Интерпретация

На каждом шаге метода Ньютона-Рафсона решается задача многомерной линейной регрессии:

$$Q(w) = \sum_{i=1}^{\ell} (1-\sigma_i)\sigma_i \left(w^{\mathsf{T}} x_i - \frac{y_i}{\sigma_i} \right)^2 \to \min_w .$$

Интерпретация:

- $\sigma_i = \mathsf{P}(y_i | x_i)$ — вероятность правильной классификации x_i
- чем ближе x_i к границе, тем больше вес $(1-\sigma_i)\sigma_i$
- чем выше вероятность ошибки, тем больше y_i/σ_i

ВЫВОД: на каждой итерации происходит более точная настройка на «наиболее трудных» объектах.

МНК с итерационным перевзвешиванием объектов
IRLS — Iteratively Reweighted Least Squares

Вход: F, y — матрица «объекты–признаки» и вектор ответов;
Выход: w — вектор коэффициентов линейной комбинации.

$w := (F^\mathsf{T} F)^{-1} F^\mathsf{T} y$ — нулевое приближение, обычный МНК;
для $t := 1, 2, 3, \ldots$
  $\sigma_i = \sigma(y_i w^\mathsf{T} x_i)$ для всех $i = 1, \ldots, \ell$;
  $\gamma_i := \sqrt{(1 - \sigma_i)\sigma_i}$ для всех $i = 1, \ldots, \ell$;
  $\tilde{F} := \mathrm{diag}(\gamma_1, \ldots, \gamma_\ell) F$;
  $\tilde{y}_i := y_i \sqrt{(1 - \sigma_i)/\sigma_i}$ для всех $i = 1, \ldots, \ell$;
  выбрать градиентный шаг h_t;
  $w := w + h_t (\tilde{F}^\mathsf{T} \tilde{F})^{-1} \tilde{F}^\mathsf{T} \tilde{y}$;
  если $\{\sigma_i\}$ мало изменились **то** выйти из цикла;

Обобщённая аддитивная модель (Generalized Additive Model)

Модель регрессии с нелинейными преобразованиями исходных признаков $\varphi_j\colon \mathbb{R} \to \mathbb{R}$:

$$f(x, \alpha) = \sum_{j=1}^{n} \varphi_j(f_j(x), \alpha).$$

В частности, при $\varphi_j(f_j(x), \alpha) = \alpha_j f_j(x)$ это линейная регрессия.

ИДЕЯ: поочерёдно уточнять функции φ_j по обучающей выборке $\bigl(f_j(x_i), z_i\bigr)_{i=1}^{\ell}$, постепенно ослабляя регуляризирующее требование гладкости $R(\varphi_j) = \int \bigl(\varphi_j''(\zeta)\bigr)^2 \, d\zeta \to \min$:

$$Q(\varphi, X^\ell) \sum_{i=1}^{\ell} \biggl(\varphi_j(f_j(x_i)) - \underbrace{\Bigl(y_i - \sum_{k \neq j} \varphi_k(f_k(x_i))\Bigr)}_{z_i}\biggr)^2 + \tau R(\varphi_j) \to \min_{\varphi_j}$$

Метод backfitting [Хасти, Тибширани, 1986]

Вход: F, y — матрица «объекты–признаки» и вектор ответов;
Выход: $\varphi_j(x)$ — все функции преобразования признаков.

1: нулевое приближение:
   $\alpha :=$ решение задачи МЛР с признаками $f_j(x)$;
   $\varphi_j(x) := \alpha_j f_j(x),\ \ j = 1, \dots, n$;
2: **повторять**
3:    **для** $j = 1, \dots, n$
4: $$z_i := y_i - \sum_{k=1, k\neq j}^{n} \varphi_k(f_k(x_i)),\ \ i = 1, \dots, \ell;$$
5: $$\varphi_j := \arg\min_{\varphi} \sum_{i=1}^{\ell} \bigl(\varphi(f_j(x_i)) - z_i\bigr)^2 + \tau R(\varphi_j);$$
6:    уменьшить коэффициент регуляризации τ;
7: **пока** $Q(\varphi, X^\ell)$ и/или $Q(\varphi, X^k)$ заметно уменьшаются;

Напоминание: связь ММП и МНК

Линейная модель данных с гауссовским шумом:

$$y_i = x_i^\mathsf{T}\alpha + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad i = 1, \dots, \ell.$$

Метод максимума правдоподобия (ММП) совпадает с МНК:

$$L(\varepsilon_1, \dots, \varepsilon_\ell | \alpha) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}\varepsilon_i^2\right) \to \max_\alpha;$$

$$-\ln L(\varepsilon_1, \dots, \varepsilon_\ell | \alpha) = \mathrm{const}(\alpha) + \frac{1}{2}\sum_{i=1}^{\ell} \frac{1}{\sigma_i^2}\left(x_i^\mathsf{T}\alpha - y_i\right)^2 \to \min_\alpha;$$

Как использовать линейные модели, если y_i не гауссовские, в частности, если y_i дискретнозначные?

Обобщённая линейная модель (Generalized Linear Model, GLM)

Линейная модель для матожидания:

$$\mu_i = \mathsf{E} y_i = x_i^{\mathsf{T}} \alpha, \qquad y_i \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

Обобщённая линейная модель для матожидания:

$$g(\mu_i) = g(\mathsf{E} y_i) = \theta_i = x_i^{\mathsf{T}} \alpha, \qquad y_i \sim \mathrm{Exp}(\mu_i, \phi_i),$$

$g(\mu)$ — *функция связи* (link function),
Exp — *экспоненциальное семейство распределений*:

$$p(y_i | \theta_i, \phi_i) = \exp\left(\frac{y_i \theta_i - c(\theta_i)}{\phi_i} + h(y_i, \phi_i) \right).$$

Замечательные свойства экспоненциального семейства:

$$\mu_i = \mathsf{E} y_i = c'(\theta_i), \quad \Rightarrow \quad g(\mu) = [c']^{-1}(\mu)$$

$$\mathsf{D} y_i = \phi_i c''(\theta_i).$$

Примеры распределений из экспоненциального семейства

Нормальное (гауссовское) распределение, $y_i \in \mathbb{R}$:

$$p(y_i|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2\sigma_i^2}(y_i - \mu_i)^2\right) =$$

$$= \exp\left(\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma_i^2} - \frac{y_i^2}{2\sigma_i^2} - \frac{1}{2}\ln(2\pi\sigma_i^2)\right);$$

$$\theta_i = g(\mu_i) = \mu_i, \qquad c(\theta_i) = \tfrac{1}{2}\mu_i^2 = \tfrac{1}{2}\theta_i^2, \qquad \phi_i = \sigma_i^2.$$

Пуассоновское распределение, $y_i \in \{0, 1, 2, \dots\}$:

$$p(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp\left(\frac{y_i\ln(\mu_i) - \mu_i}{1} - \ln y_i!\right);$$

$$\theta_i = g(\mu_i) = \ln(\mu_i), \qquad c(\theta_i) = \mu_i = e^{\theta_i}, \qquad \phi_i = 1.$$

Примеры распределений из экспоненциального семейства

Биномиальное распределение, $y_i \in \{0, 1, \dots, n_i\}$:

$$p(y_i|\mu_i, n_i) = C_{n_i}^{y_i} \mu_i^{y_i} (1-\mu_i)^{n_i - y_i} =$$
$$= \exp\Bigl(y_i \ln \tfrac{\mu_i}{1-\mu_i} + n_i \ln(1-\mu_i) + \ln C_{n_i}^{y_i}\Bigr);$$

$$\theta_i = g(\mu_i) = \ln \tfrac{\mu_i}{1-\mu_i}, \qquad c(\theta_i) = -n_i \ln(1-\mu_i) = n_i \ln(1+e^{\theta_i}).$$

Распределение Бернулли, $y_i \in \{0, 1\}$:

$$p(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \exp\Bigl(y_i \ln \tfrac{\mu_i}{1-\mu_i} + \ln(1-\mu_i)\Bigr);$$

$$\theta_i = g(\mu_i) = \ln \tfrac{\mu_i}{1-\mu_i}, \qquad c(\theta_i) = -\ln(1-\mu_i) = \ln(1+e^{\theta_i}).$$

Некоторые распределения из экспоненциального семейства

- нормальное (гауссовское)
- распределение Пуассона
- биномиальное и мультиномиальное
- геометрическое
- ξ^2-распределение
- бета-распределение
- гамма-распределение
- распределение Дирихле
- распределение Лапласа с фиксированным матожиданием

Не экспоненциальные:

t-распределение Стьюдента, Коши, гипергеометрическое

Максимизация правдоподобия для GLM

Принцип максимума правдоподобия:

$$L(\alpha; X^\ell) = \sum_{i=1}^{\ell} \frac{y_i\theta_i - c(\theta_i)}{\phi_i} \to \max_{\alpha}, \qquad \theta_i = x_i^{\mathsf{T}}\alpha = \sum_{j=1}^{n} \alpha_j f_j(x_i)$$

Метод Ньютона-Рафсона:

$$\alpha^{t+1} := \alpha^t - h_t \bigl(L''(\alpha^t)\bigr)^{-1} L'(\alpha^t).$$

Компоненты вектора градиента $L'(\alpha)$:

$$\frac{\partial L(\alpha)}{\partial \alpha_j} = \sum_{i=1}^{\ell} \frac{y_i - c'(x_i^{\mathsf{T}}\alpha)}{\phi_i} f_j(x_i).$$

Компоненты матрицы Гессе $L''(\alpha)$:

$$\frac{\partial^2 L(\alpha)}{\partial \alpha_j \partial \alpha_k} = -\sum_{i=1}^{\ell} \frac{c''(x_i^{\mathsf{T}}\alpha)}{\phi_i} f_j(x_i) f_k(x_i).$$

Матричные обозначения

$F = \bigl(f_j(x_i)\bigr)_{\ell\times n}$ — матрица «объекты–признаки»;

$y = \bigl(y_i\bigr)_{\ell\times 1}$ — вектор ответов.

$\mu^t = \bigl(c'(\theta_i)\bigr)_{\ell\times 1}$ — вектор матожиданий, $\theta_i = x_i^{\mathsf{T}}\alpha^t$.

$W^t = \mathrm{diag}\bigl(\frac{1}{\phi_i}c''(\theta_i)\bigr)$ — диагональная матрица.

Тогда метод Ньютона-Рафсона снова приводит к IRLS:

$$\alpha^{t+1} := \alpha^t - h_t\bigl(F^{\mathsf{T}}W^tF\bigr)^{-1}F^{\mathsf{T}}W^t\frac{1}{c''(\theta_i)}(y-\mu^t).$$

Это совпадает с МНК-решением линейной задачи регрессии со взвешенными объектами и модифицированными ответами:

$$Q(\alpha) = \|\tilde{F}\alpha - (\tilde{y}-\tilde{\mu})\|^2 \to \min_{\alpha}.$$

Метод наименьших модулей

$\varepsilon_i = \bigl(a(x_i) - y_i\bigr)$ — ошибка

$\mathscr{L}(\varepsilon_i)$ — функция потерь

$Q = \sum\limits_{i=1}^{\ell} \mathscr{L}(\varepsilon_i) \to \min\limits_{a}$ — критерий обучения модели по выборке

Метод наименьших квадратов, $\mathscr{L}(\varepsilon) = \varepsilon^2$:

$$\sum_{i=1}^{\ell} (a - y_i)^2 \to \min_a, \qquad a = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i.$$

Метод наименьших модулей, $\mathscr{L}(\varepsilon) = |\varepsilon|$:

$$\sum_{i=1}^{\ell} |a - y_i| \to \min_a, \qquad a = \mathrm{median}\{y_1, \ldots, y_\ell\} = y^{(\ell/2)},$$

где $y^{(1)}, \ldots, y^{(\ell)}$ — вариационный ряд значений y_i

Квантильная регрессия

Квантúльная регрессия, $\mathscr{L}(\varepsilon) = \begin{cases} C_+|\varepsilon|, & \varepsilon > 0 \\ C_-|\varepsilon|, & \varepsilon < 0; \end{cases}$:

$$\sum_{i=1}^{\ell} \mathscr{L}\,(a - y_i) \to \min_a, \qquad a = y^{(q)}, \quad q = \frac{\ell C_-}{C_- + C_+}$$

где $y^{(1)}, \dots, y^{(\ell)}$ — вариационный ряд значений y_i

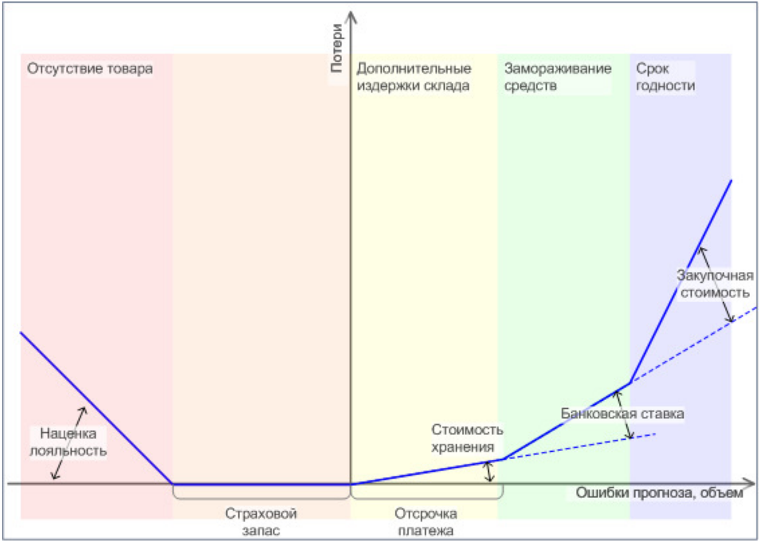
Линейная модель регрессии: $a(x_i) = \langle x_i, w\rangle$.

Сведение к задаче линейного программирования:

замена переменных $\varepsilon_i^+ = \bigl(a(x_i) - y_i\bigr)_+$, $\varepsilon_i^- = \bigl(y_i - a(x_i)\bigr)_+$;

$$\begin{cases} Q = \sum\limits_{i=1}^{\ell} C_+\varepsilon_i^+ + C_-\varepsilon_i^- \to \min\limits_w; \\ \langle x_i, w\rangle - y_i = \varepsilon_i^+ + \varepsilon_i^-; \\ \varepsilon_i^+ \geqslant 0; \quad \varepsilon_i^- \geqslant 0. \end{cases}$$

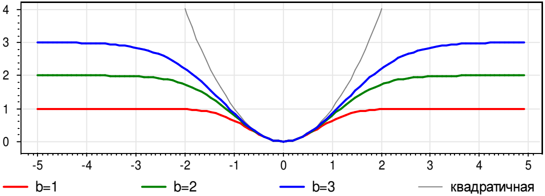
Пример. Задача прогнозирования объёмов продаж



Робастная регрессия

Модель регрессии: $a(x) = f(x, \alpha)$

Функция Мешалкина: $\mathscr{L}(\varepsilon) = b\bigl(1 - \exp\bigl(-\frac{1}{b}\varepsilon^2\bigr)\bigr)$



Постановка задачи:

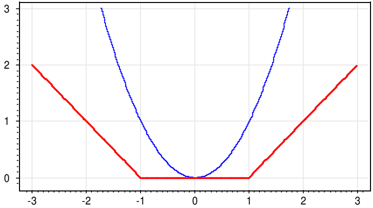
$$\sum_{i=1}^{\ell} \exp\Bigl(-\frac{1}{b}(f(x_i, \alpha) - y_i)^2\Bigr) \to \max_{\alpha}.$$

Задача решается методом Ньютона-Рафсона.

SVM-регрессия (напоминание)

Модель регрессии: $a(x) = \langle x, w\rangle - w_0,\ \ w \in \mathbb{R}^n,\ \ w_0 \in \mathbb{R}.$

Функция потерь: $\mathscr{L}(\varepsilon) = \bigl(|\varepsilon| - \delta\bigr)_+$



Постановка задачи:

$$\sum_{i=1}^{\ell} \bigl(|\langle w, x_i\rangle - w_0 - y_i| - \delta\bigr)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w,w_0}.$$

Задача решается путём замены переменных
и сведения к задаче квадратичного программирования

Резюме в конце лекции

- Нелинейная регрессия
  — основана на методе Ньютона-Рафсона
  — сводится к последовательности линейных регрессий
- Логистическая регрессия
  — не регрессия, а классификация
  — основана на методе Ньютона-Рафсона
- Неквадратичные функции потерь
  — проблемно-ориентированные (зависят от задачи)
  — приводят к разным методам, отличным от МНК