# Методы машинного обучения. Обобщения линейных моделей регрессии и классификации

Воронцов Константин Вячеславович
www.MachineLearning.ru/wiki?title=User:Vokov
вопросы к лектору: k.vorontsov@iai.msu.ru

материалы курса:
github.com/MSU-ML-COURSE/ML-COURSE-25-26
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 21 октября 2025

# Содержание


1. **Нелинейная регрессия**
   - Нелинейная модель регрессии
   - Логистическая регрессия
   - Обобщённая аддитивная модель
2. **Обобщённая линейная модель**
   - Экспоненциальное семейство распределений
   - Максимизация правдоподобия для GLM
   - Логистическая регрессия как частный случай GLM
3. **Неквадратичные функции потерь**
   - Квантильная регрессия
   - Робастная регрессия
   - SVM-регрессия

# Нелинейная модель регрессии

**Дано:** обучающая выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$
$y_i = y(x_i)$, $y\colon X \to Y$ — неизвестная регрессионная зависимость
**Найти:** параметры $\alpha \in \mathbb{R}^p$ модели регрессии $f(x, \alpha)$
**Критерий:** метод наименьших квадратов (МНК)

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

**Метод Ньютона–Рафсона:** итерационный процесс стартует из начального приближения $\alpha^0 = (\alpha_1^0, \ldots, \alpha_p^0)$:

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

$\nabla Q(\alpha^t)$ — градиент функционала $Q$ в точке $\alpha^t$, вектор из $\mathbb{R}^p$
$Q''(\alpha^t)$ — гессиан функционала $Q$ в точке $\alpha^t$, матрица из $\mathbb{R}^{p\times p}$
$h_t$ — величина шага (можно полагать $h_t = 1$).

# Метод Ньютона-Рафсона

Компоненты градиента:

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

Компоненты гессиана:

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

Не хотелось бы обращать гессиан на каждой итерации...

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

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

# 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;

$f_t = \left(f(x_i, \alpha^t)\right)_{\ell \times 1}$ — вектор значений $f$.

Формула $t$-й итерации метода Ньютона–Гаусса:

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

$\beta$ — 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 линейные методы.

# Напоминание. Двухклассовая логистическая регрессия

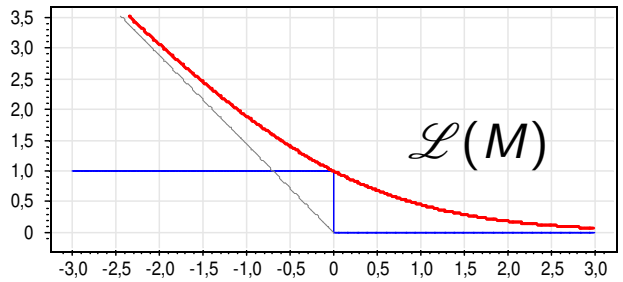
**Дано:** обучающая выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$

**Найти:** параметр $w \in \mathbb{R}^n$ линейной модели $a(x, w) = \text{sign}(w^\intercal x)$

$M = (w^\intercal x)y$ — отступ (margin)

*Логарифмическая функция потерь:*

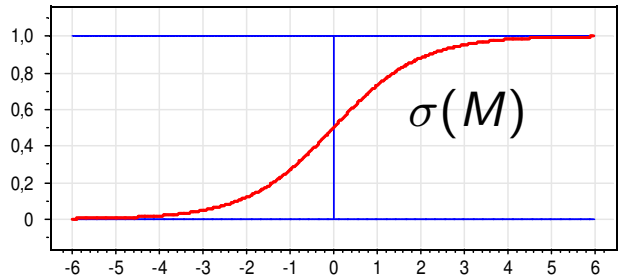
$$\mathscr{L}(M) = \ln\left(1 + e^{-M}\right)$$




*Модель условной вероятности:*

$$P(y|x, w) = \sigma(M) = \frac{1}{1+e^{-M}},$$

$\sigma(M)$ — сигмоидная функция,




**Критерий** максимума log правдоподобия, без регуляризации:

$$Q(w) = \sum_{i=1}^{\ell} \ln P(y_i|x_i, w) = \sum_{i=1}^{\ell} \ln\left(1 + \exp(-w^\intercal x_i y_i)\right) \to \min_w$$

# Метода Ньютона-Рафсона

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

$$w^{t+1} := w^t - h_t\big(Q''(w^t)\big)^{-1}\nabla Q(w^t)$$

Элементы градиента — вектора первых производных $\nabla Q(w^t)$:

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

$\sigma_i = \sigma(y_i w^{\mathsf{T}} x_i) = \mathsf{P}(y_i|x_i, w)$ — вероятность верной классификации

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

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

# Снова сведение к многомерной линейной регрессии

Матричные обозначения: $F = \big(f_j(x_i)\big)_{\ell \times n}$, $D = \operatorname{diag}\big(\sqrt{(1-\sigma_i)\sigma_i}\,\big)$

$$-\big(Q''(w)\big)^{-1}\nabla Q(w) = \underbrace{\big(F^\intercal DDF\big)^{-1}F^\intercal D}_{\tilde{F}^+ = (\tilde{F}^\intercal\tilde{F})^{-1}\tilde{F}^\intercal}\Big(\underbrace{y_i\sqrt{\tfrac{1-\sigma_i}{\sigma_i}}}_{\tilde{y}_i}\Big)_{\ell\times 1}$$

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

$$\|\tilde{F}w - \tilde{y}\|^2 = \sum_{i=1}^{\ell}(1-\sigma_i)\sigma_i\left(w^\intercal x_i - \frac{y_i}{\sigma_i}\right)^2 \to \min_w$$

Интерпретация:

- чем выше вероятность ошибки, тем больше $\frac{1}{\sigma_i}$
- чем ближе $x_i$ к границе, тем больше вес $(1-\sigma_i)\sigma_i$

Таким образом, на каждой итерации происходит более точная настройка на «наиболее трудных» объектах.

# МНК с итерационным перевзвешиванием объектов

## Метод IRLS — Iteratively Reweighted Least Squares

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

---

1: $w := (F^\intercal F)^{-1} F^\intercal y$ — нулевое приближение, обычный МНК;
2: **для** $t := 1, 2, 3, \ldots$
3: $\quad \sigma_i = \sigma(y_i w^\intercal x_i)$ для всех $i = 1, \ldots, \ell$;
4: $\quad \tilde{F} := \operatorname{diag}\left(\sqrt{(1-\sigma_i)\sigma_i}\right) F$;
5: $\quad \tilde{y}_i := y_i \sqrt{\frac{1-\sigma_i}{\sigma_i}}$ для всех $i = 1, \ldots, \ell$;
6: $\quad$ выбрать градиентный шаг $h_t$;
7: $\quad w := w + h_t (\tilde{F}^\intercal \tilde{F})^{-1} \tilde{F}^\intercal \tilde{y}$;
8: $\quad$ **если** $\{\sigma_i\}$ мало изменились **то** выйти из цикла;

# Обобщённая аддитивная модель (Generalized Additive Model)

**Дано:** обучающая выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

**Найти:** нелинейные преобразования признаков $\varphi_j(f_j, \alpha_j)$
(в частности, при $\varphi_j(f_j(x), \alpha_j) = \alpha_j f_j(x)$ это линейная модель):

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

**Критерий:** метод наименьших квадратов

**Идея 1:** поочерёдно уточнять $\varphi_j$ по выборке $\big(f_j(x_i), z_i\big)_{i=1}^{\ell}$:

$$\sum_{i=1}^{\ell} \Big( \varphi_j(f_j(x_i), \alpha_j) - \underbrace{\big(y_i - \sum_{k \neq j} \varphi_k(f_k(x_i), \alpha_k)\big)}_{z_i} \Big)^2 + \tau R(\alpha_j) \;\to\; \min_{\alpha_j}$$

**Идея 2:** используя в качестве $\varphi_j$ ядерное сглаживание или сплайны, постепенно уменьшать $\tau$ у регуляризатора гладкости

$$R(\alpha_j) = \int \big(\varphi_j''(\zeta, \alpha_j)\big)^2 \, d\zeta$$

# Метод backfitting [Хасти, Тибширани, 1986]

Многомерная задача сводится к серии одномерных задач.

**Вход:** $F, y$ — матрица «объекты–признаки» и вектор ответов;
**Выход:** $\varphi_j(f_j, \alpha_j)$ — обучаемые преобразования признаков.

---

1: начальное приближение: $\alpha := (F^\intercal F)^{-1} F^\intercal y$;
$\varphi_j(f_j, \alpha_j) := \alpha_j f_j(x), \;\; j = 1, \ldots, n$;
2: **повторять**
3: **для** $j = 1, \ldots, n$
4: $z_i := y_i - \sum\limits_{k=1, k\neq j}^{n} \varphi_k(f_k(x_i), \alpha_k), \;\; i = 1, \ldots, \ell$;
5: $\alpha_j := \arg\min\limits_{\alpha} \sum\limits_{i=1}^{\ell} \big(\varphi(f_j(x_i), \alpha) - z_i\big)^2 + \tau R(\alpha)$;
6: уменьшить коэффициент регуляризации $\tau$;
7: **пока** $Q(\alpha, X^\ell)$ и/или $Q(\alpha, X^k)$ заметно уменьшаются;

---

*T.J.Hastie, R.J.Tibshirani.* Generalized Additive Models. 1990.

# Напоминание. Вероятностная постановка задачи регрессии

**Дано:** выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in X$, $y_i \in \mathbb{R}$

**Найти:** параметр $w$ модели регрессии $y_i = a(x_i, w) + \varepsilon_i$, где шум $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ гауссовский, некоррелированный $\mathsf{E}\varepsilon_i\varepsilon_k = 0$:

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i = \mathsf{E}y_i = a(x_i, w), \quad i = 1, \ldots, \ell$$

**Критерий** максимума правдоподобия эквивалентен МНК:

$$p(\varepsilon_1, \ldots, \varepsilon_\ell | w) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}\varepsilon_i^2\right) \to \max_w$$

$$-\ln p(\varepsilon_1, \ldots, \varepsilon_\ell | w) = \text{const} + \frac{1}{2}\sum_{i=1}^{\ell} \frac{1}{\sigma_i^2}\big(a(x_i, w) - y_i\big)^2 \to \min_w$$

Что использовать вместо метода наименьших квадратов, если $y_i$ не гауссовские, в частности, если $y_i$ дискретнозначные?

# Обобщение: экспоненциальное распределение шума

**Дано:** выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in X$, $y_i \in \mathbb{R}$
**Найти:** параметр $w$ при более общем предположении о шуме:

$$y_i \sim \mathrm{Exp}(\theta_i, \phi_i), \qquad \theta_i = g(\mathrm{E}y_i) = a(x_i, w), \quad i = 1, \ldots, \ell$$

Exp — *экспоненциальное семейство распределений*
с параметрами $\theta_i$, $\phi_i$ и параметрами-функциями $c(\theta)$, $h(y, \phi)$:

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

Математическое ожидание и дисперсия с.в. $y_i \sim \mathrm{Exp}(\theta_i, \phi_i)$:

$$\mu_i = \mathrm{E}y_i = c'(\theta_i) \quad \Rightarrow \quad \theta_i = [c']^{-1}(\mu_i) = g(\mathrm{E}y_i)$$
$$\mathrm{D}y_i = \phi_i c''(\theta_i)$$

$g(\mu) = [c']^{-1}(\mu)$ — монотонная *функция связи* (link function)

# Примеры распределений из экспоненциального семейства

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

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

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

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

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

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

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

# Примеры распределений из экспоненциального семейства

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

$$p(y_i|\mu_i, n_i) = C_{n_i}^{y_i} \left(\frac{\mu_i}{n_i}\right)^{y_i} \left(1 - \frac{\mu_i}{n_i}\right)^{n_i - y_i} =$$

$$= \exp\left(y_i \ln \frac{\mu_i}{n_i - \mu_i} + n_i \ln(n_i - \mu_i) + \ln C_{n_i}^{y_i} - n_i \ln n_i\right);$$

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

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

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

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

# Примеры распределений из экспоненциального семейства

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

**Контр-примеры** не экспоненциальных распределений:

- $t$-распределение Стьюдента, Коши, гипергеометрическое

# Обобщённая линейная модель (Generalized Linear Model, GLM)

**Дано:** выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

**Найти:** $w \in \mathbb{R}^n$ линейной модели $\theta_i = \langle x_i, w \rangle = \sum_{j=1}^{n} w_j f_j(x_i)$

**Критерий** максимума правдоподобия для оценивания $w$:

$$Q(w) = \ln \prod_{i=1}^{\ell} p(y_i|\theta_i, \phi_i) = \sum_{i=1}^{\ell} \frac{y_i\theta_i - c(\theta_i)}{\phi_i} \to \max_w,$$

Метод Ньютона-Рафсона: $w^{t+1} := w^t + h_t \big(Q''(w^t)\big)^{-1} \nabla Q(w^t)$

Компоненты вектора градиента $\nabla Q(w)$:

$$\frac{\partial Q(w)}{\partial w_j} = \sum_{i=1}^{\ell} \frac{y_i - c'(\theta_i)}{\phi_i} f_j(x_i).$$

Компоненты матрицы Гессе $Q''(w)$:

$$\frac{\partial^2 Q(w)}{\partial w_j \partial w_k} = -\sum_{i=1}^{\ell} \frac{c''(\theta_i)}{\phi_i} f_j(x_i) f_k(x_i).$$

# Матричные обозначения

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

$\tilde{F} = DF, \quad D = \operatorname{diag}\left(\sqrt{\frac{1}{\phi_i} c''(\theta_i)}\right)$ — веса объектов, $\theta_i = \langle x_i, w^t \rangle$

$\tilde{y} = \big(\tilde{y}_i\big)_{\ell \times 1}, \quad \tilde{y}_i = \frac{y_i - c'(\theta_i)}{\sqrt{\phi_i c''(\theta_i)}}$ — модифицированный вектор ответов

Тогда метод Ньютона-Рафсона снова приводит к IRLS:

$$w^{t+1} := w^t - h_t \underbrace{\big(F^\intercal D D F\big)^{-1} F^\intercal D}_{\tilde{F}^+ = (\tilde{F}^\intercal \tilde{F})^{-1} \tilde{F}^\intercal} \Bigg( \underbrace{\sqrt{\frac{\phi_i}{c''(\theta_i)}} \frac{y_i - c'(\theta_i)}{\phi_i}}_{\tilde{y}_i} \Bigg)_{\ell \times 1}$$

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

$$\|\tilde{F} w - \tilde{y}\|^2 \to \min_w$$

# МНК с итерационным перевзвешиванием объектов

## Метод IRLS — Iteratively Reweighted Least Squares

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

---

начальное приближение: $w := (F^\mathsf{T}F)^{-1}F^\mathsf{T}y$;
**для** $t := 1, 2, 3, \ldots$
  $\theta_i = \langle x_i, w^t \rangle$ для всех $i = 1, \ldots, \ell$;
  $\tilde{F} := \operatorname{diag}\left(\sqrt{\frac{1}{\phi_i}c''(\theta_i)}\right)F$;
  $\tilde{y}_i := \frac{y_i - c'(\theta_i)}{\sqrt{\phi_i c''(\theta_i)}}$ 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** $\{\theta_i\}$ мало изменились **то** выйти из цикла;

# Двухклассовая логистическая регрессия

Распределение Бернулли, $y_i \in \{0,1\}$: $p(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i}$

$$\theta_i = g(\mu_i) = \ln\frac{\mu_i}{1-\mu_i} = \langle x_i, w\rangle, \quad \mu_i = g^{-1}(\theta_i) = \frac{1}{1+e^{-\theta_i}} \equiv \sigma(\theta_i) = \mathsf{E}y_i$$

**Дано:** выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \{0,1\} \sim p(y_i|\mu_i)$

**Найти:** вероятностную модель $\mathsf{E}(y|x) = \mathsf{P}(y{=}1|x) = \sigma\big(\langle x, w\rangle\big)$

**Критерий:** максимум log-правдоподобия (log-loss)

$$Q(w) = \sum_{i=1}^{\ell} \ln p(y_i|\mu_i) = \sum_{i=1}^{\ell} y_i \ln\mu_i + (1-y_i)\ln(1-\mu_i) \to \max_w$$

Альтернативная кодировка: $y_i \in \{0,1\} \;\to\; \tilde{y}_i = 2y_i - 1 \in \{\pm 1\}$

$$-\sum_{i=1}^{\ell} \ln p(\tilde{y}_i|x_i) = \sum_{i=1}^{\ell} \ln\big(1 + \exp(-\underbrace{\langle w, x_i\rangle \tilde{y}_i}_{\text{margin}})\big) \;\to\; \min_w$$

# Логистическая регрессия как частный случай GLM

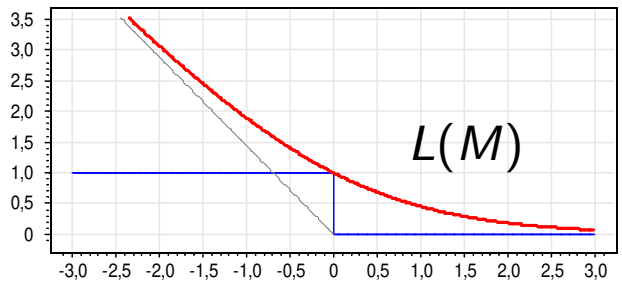
Всего лишь из двух предположений:

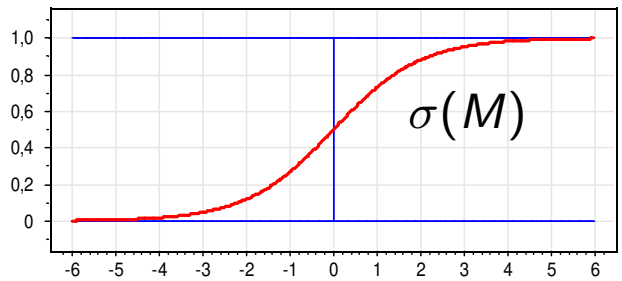
- $y_i$ — бернуллиевские случайные величины с $\mathsf{E}y_i = \mu_i$
- параметр связан с линейной моделью: $\theta_i = g(\mu_i) = \langle x_i, w \rangle$

следуют важнейшие свойства логистической регрессии:

- логарифмическая функция потерь $\ln\bigl(1 + \exp(-\langle x_i, w\rangle \tilde{y}_i)\bigr)$;
- сигмоидная функция связи $\mathsf{P}(y_i|x_i) = \sigma(\langle x_i, w\rangle \tilde{y}_i)$;
- связь линейной модели с *отношением шансов* (odds ratio):

$$\langle x_i, w\rangle = \ln \frac{\mu_i}{1-\mu_i} = \ln \frac{\mathsf{P}(y_i=1|x_i)}{\mathsf{P}(y_i=0|x_i)}$$



# Многоклассовая логистическая регрессия

Категориальное (дискретное) распределение, $y_i \in Y$, $|Y| < \infty$:

$$p(y_i|\mu_i) = \prod_{y \in Y} \mu_{yi}^{[y=y_i]} = \exp\Big( \sum_{y \in Y} [y=y_i] \ln \mu_{yi} \Big), \quad \mu_i = (\mu_{yi})_{y \in Y}$$

$$\theta_{yi} = g(\mu_{yi}) = \ln \mu_{yi} = \langle x_i, w_y \rangle, \quad \sum_y \mu_{yi} = 1, \quad \mu_{yi} > 0, \quad \varphi_i = 1$$

$$\mu_{yi} = g^{-1}(\theta_{yi}) = \frac{\exp(\theta_{yi})}{\sum_{z \in Y} \exp(\theta_{zi})} = \operatorname*{SoftMax}_{y \in Y} \theta_{yi} = \mathrm{P}(y|x_i)$$

**Дано:** выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in Y \sim p(y_i|\mu_i)$

**Найти:** линейную вероятностную модель классификации

$$P(y|x,w) = \operatorname*{SoftMax}_{y \in Y} \langle w_y, x \rangle, \quad a(x) = \arg\max_{y \in Y} \langle w_y, x \rangle, \quad w = (w_y)_{y \in Y}$$

**Критерий:** максимум log-правдоподобия (log-loss)

$$Q(w) = \sum_{i=1}^{\ell} \ln P(y_i|x_i,w) \;\to\; \max_w$$

# Метод наименьших модулей (Least Absolute Deviation Regression)

$\mathscr{L}(\varepsilon_i)$ — функция потерь; $\varepsilon_i = \big(a(x_i, w) - y_i\big)$ — ошибка;

$Q = \sum_{i=1}^{\ell} \mathscr{L}(\varepsilon_i) \to \min_w$ — критерий обучения модели по выборке.

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

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

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

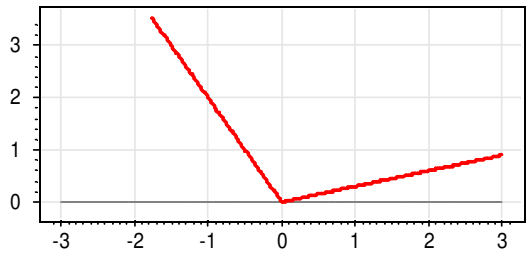
$$\sum_{i=1}^{\ell}|a - y_i| \to \min_a \quad \Rightarrow \quad a = \text{median}\{y_1, \ldots, y_\ell\} = y^{(\ell/2)},$$

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

Медиана более устойчива к редким большим выбросам $y_i$.

# Квантильная регрессия (Quantile Regression)

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




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

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

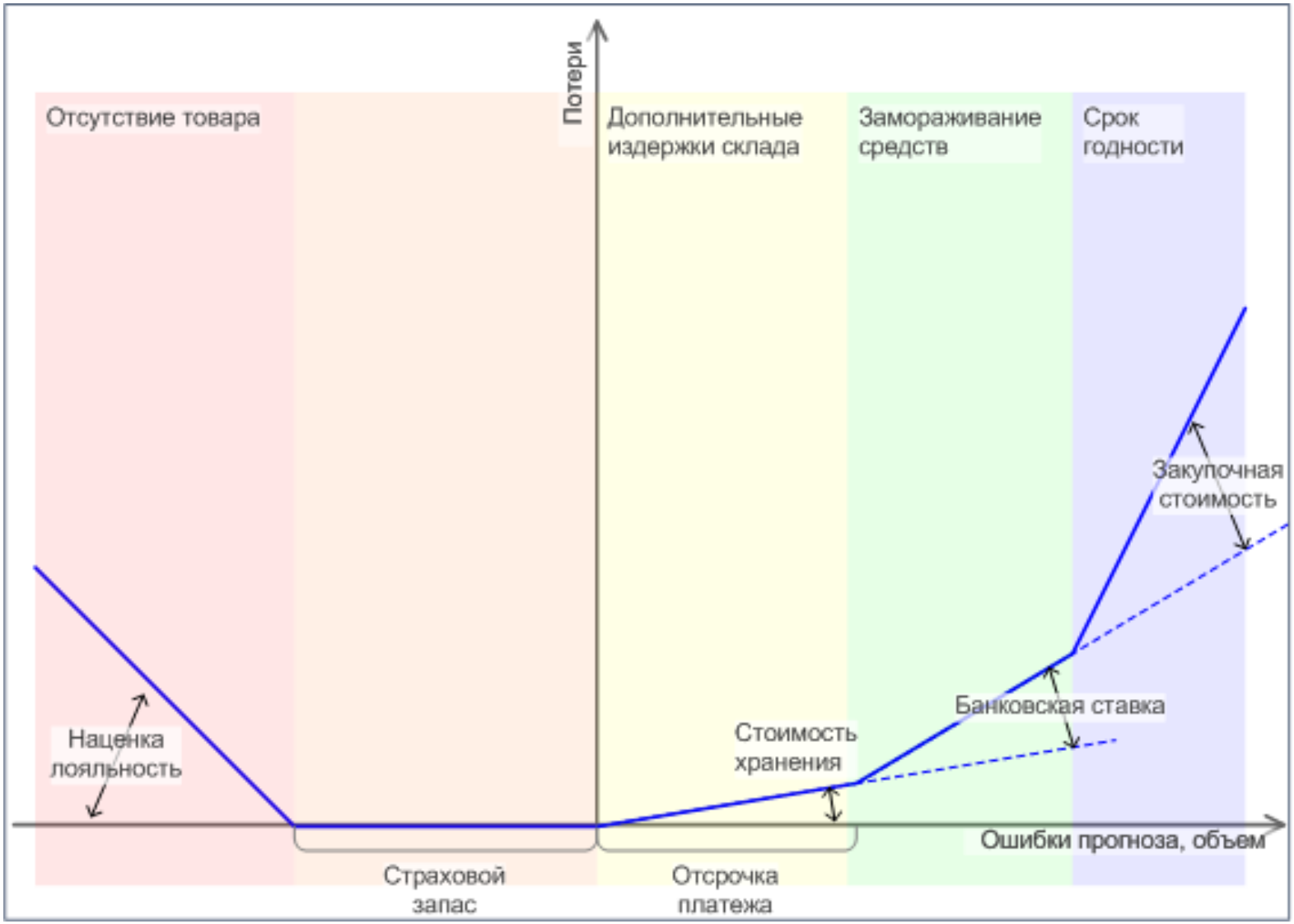
Линейная модель регрессии: $a(x_i, w) = \langle x_i, w \rangle$.

**Сведение к задаче линейного программирования:**

замена переменных $\varepsilon_i^+ = \big(a(x_i, w) - y_i\big)_+$, $\varepsilon_i^- = \big(y_i - a(x_i, w)\big)_+$

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

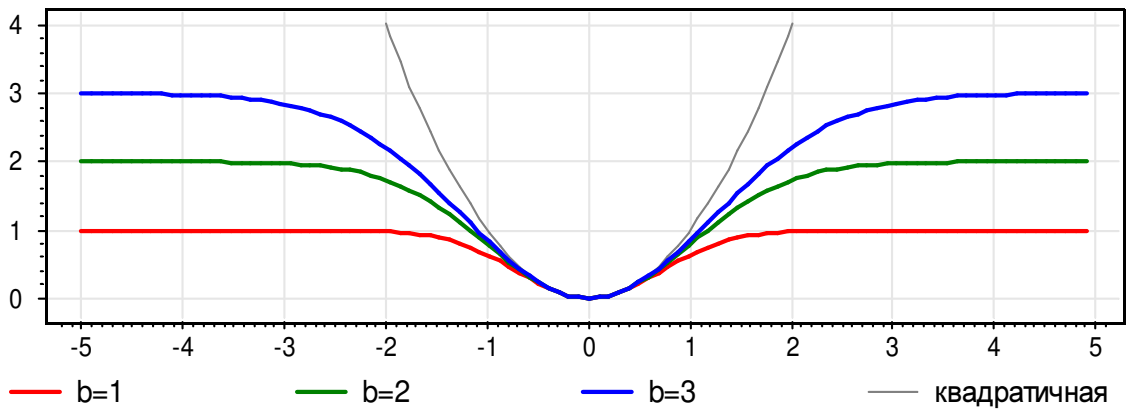
# Пример. Задача прогнозирования объёмов продаж



# Робастная регрессия (Robust Regression)

$a(x, w)$ — модель регрессии; $\varepsilon_i = \big(a(x_i, w) - y_i\big)$ — ошибка;
$\mathscr{L}(\varepsilon)$ — функция потерь, устойчивая к большим выбросам $\varepsilon$

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




**Постановка задачи:**

$$\sum_{i=1}^{\ell} \exp\Big(-\frac{1}{b}\big(a(x_i, w) - y_i\big)^2\Big) \to \max_w .$$

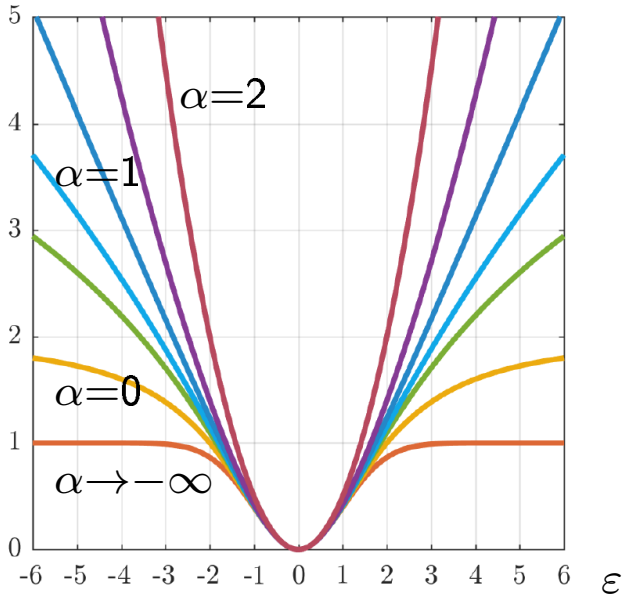
Эта задача также решается методом Ньютона-Рафсона.

# Функции потерь для робастной регрессии

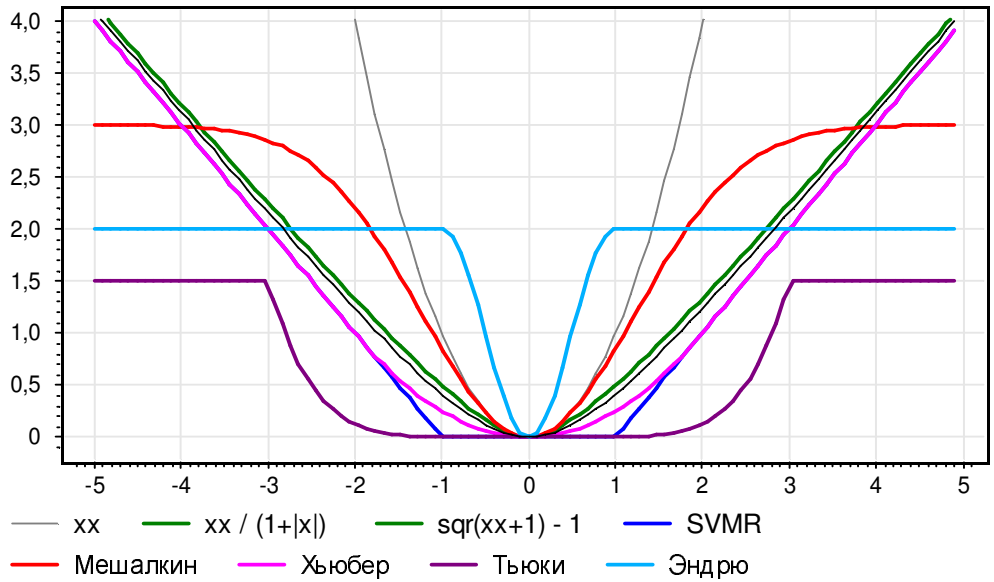
Семейство функций потерь Баррона с параметром $\alpha$:

$$\mathscr{L}_\alpha(\varepsilon) = \frac{|\alpha-2|}{\alpha}\left(\left(\frac{\varepsilon^2}{|\alpha-2|}+1\right)^{\alpha/2}-1\right)$$

$\mathscr{L}_\alpha(\varepsilon)$




другие примеры робастных $\mathscr{L}(\varepsilon)$

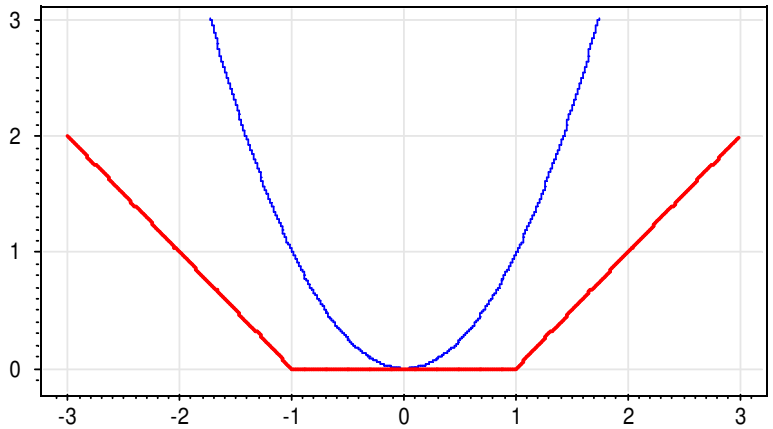



*Jonathan T. Barron.* A General and Adaptive Robust Loss Function. 2019.

# Напоминание: SVM-регрессия. Тоже робастная регрессия

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

$\mathscr{L}(\varepsilon) = \big(|\varepsilon| - \delta\big)_+$ — кусочно-линейная функция потерь




**Постановка задачи:**

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

Задача решается путём замены переменных
и сведения к задаче квадратичного программирования

# Резюме в конце лекции

- Нелинейная регрессия
  — сводится к последовательности линейных регрессий
- Логистическая регрессия
  — не регрессия, а классификация
  — метод Ньютона-Рафсона приводит к IRLS
- Обобщённая линейная модель (GLM)
  — мощно обобщает обычную и логистическую регрессию
  — метод Ньютона-Рафсона приводит к IRLS
- Обобщённая аддитивная регрессия (GAM, backfitting)
  — сводится к серии одномерных сглаживаний
- Неквадратичные функции потерь
  — проблемно-ориентированные (зависят от задачи)
  — в том числе робастная регрессия
  — приводят к разным методам, отличным от МНК
  — в некоторых случаях к методу Ньютона-Рафсона и IRLS