Методы машинного обучения. Нелинейная регрессия, обобщённые линейные модели, нестандартные функции потерь

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-24-25
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 12 ноября 2024

Содержание

Нелинейная модель регрессии

$X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$
$y_i = y(x_i)$, $y\colon X \to Y$ — неизвестная регрессионная зависимость
$a(x, w)$ — нелинейная модель регрессии, $w \in \mathbb{R}^p$

Метод наименьших квадратов (МНК):

$$Q(w, X^\ell) = \sum_{i=1}^{\ell} \bigl(a(x_i, w) - y_i\bigr)^2 \to \min_w .$$

Метод Ньютона–Рафсона:

1. Начальное приближение $w^0 = (w_1^0, \dots, w_p^0)$.
2. Итерационный процесс

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

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

Метод Ньютона-Рафсона

Компоненты градиента:

$$\frac{\partial Q(w)}{\partial w_j} = 2\sum_{i=1}^{\ell}\bigl(a(x_i,w) - y_i\bigr)\frac{\partial a(x_i,w)}{\partial w_j}$$

Компоненты гессиана:

$$\frac{\partial^2 Q(w)}{\partial w_j \partial w_k} = 2\sum_{i=1}^{\ell}\frac{\partial a(x_i,w)}{\partial w_j}\frac{\partial a(x_i,w)}{\partial w_k} \underbrace{-2\sum_{i=1}^{\ell}\bigl(a(x_i,w)-y_i\bigr)\frac{\partial^2 a(x_i,w)}{\partial w_j \partial w_k}}_{\text{при линеаризации полагается } = 0}$$

Не хотелось бы обращать гессиан на каждой итерации...

Линеаризация $a(x_i,w)$ в окрестности текущего w^t:

$$a(x_i,w) = a(x_i,w^t) + \sum_{j=1}^{p}\frac{\partial a(x_i,w_j^t)}{\partial w_j}\bigl(w_j - w_j^t\bigr) + o\bigl(w_j - w_j^t\bigr)$$

Метод Ньютона-Гаусса

Матричные обозначения:
$F_t = \left(\frac{\partial a}{\partial w_j}(x_i, w^t)\right)_{\ell\times p}$ — матрица первых производных;
$a_t = \bigl(a(x_i, w^t)\bigr)_{\ell\times 1}$ — вектор значений a.

Формула t-й итерации метода Ньютона–Гаусса:

$$w^{t+1} := w^t - h_t \underbrace{(F_t^{\mathsf{T}} F_t)^{-1} F_t^{\mathsf{T}} (a_t - y)}_{u}.$$

u — это решение задачи многомерной линейной регрессии

$$\|F_t u - (a_t - y)\|^2 \to \min_u.$$

Нелинейная регрессия сведена к серии линейных регрессий.

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

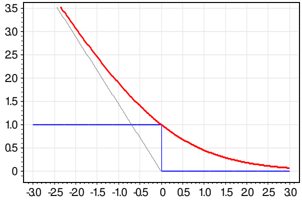
Задача классификации. Логистическая регрессия

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

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 $Q'(w^t)$:

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

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

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

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

Снова сведение к задаче линейной регрессии

В матричных обозначениях $F = \bigl(f_j(x_i)\bigr)_{\ell\times n}$, $D = \mathrm{diag}((1-\sigma_i)\sigma_i)$

$$\bigl(Q''(w)\bigr)^{-1}Q'(w) = -\bigl(F^{\mathsf T}DF\bigr)^{-1}F^{\mathsf T}\bigl(\tfrac{y_i}{\sigma_i}\bigr).$$

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

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

Интерпретация (как будут доказано далее, слайды 20–21):

- $\sigma_i = \mathsf{P}(y_i|x_i)$ — 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 $\frac{1}{\sigma_i}$

ВЫВОД: 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 модели.

$w := (F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}y$ — 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 $i = 1, \ldots, \ell$;
  $\tilde{F} := \operatorname{diag}(\gamma_1, \ldots, \gamma_\ell)F$;
  $\tilde{y}_i := y_i\sqrt{(1-\sigma_i)/\sigma_i}$ 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)

Регрессия с нелинейными преобразованиями признаков φ_j:

$$a(x,w) = \sum_{j=1}^{n} \varphi_j(f_j(x), w_j)$$

В частности, при $\varphi_j(f_j(x), w_j) = w_j f_j(x)$ это линейная модель

Идея 1: поочерёдно уточнять φ_j по выборке $\bigl(f_j(x_i), z_i\bigr)_{i=1}^{\ell}$:

$$\sum_{i=1}^{\ell}\Bigl(\varphi_j(f_j(x_i), w_j) - \underbrace{\bigl(y_i - \sum_{k\neq j}\varphi_k(f_k(x_i), w_k)\bigr)}_{z_i}\Bigr)^2 + \tau R(w_j) \;\to\; \min_{w_j}$$

Идея 2: постепенно уменьшать τ у регуляризатора гладкости

$$R(w_j) = \int \bigl(\varphi_j''(\zeta, w_j)\bigr)^2\, d\zeta$$

В качестве φ_j использовать сплайны или ядерное сглаживание

Метод backfitting [Хасти, Тибширани, 1986]

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

1: начальное приближение:
   $w := (F^\intercal F)^{-1} F^\intercal y$ — линейная регрессия;
   $\varphi_j(f_j, w_j) := w_j f_j(x),\ \ j = 1, \dots, n$;
2: **повторять**
3: **для** $j = 1, \dots, n$
4: $$z_i := y_i - \sum_{k=1, k\neq j}^{n} \varphi_k(f_k(x_i), w_k),\ \ i = 1, \dots, \ell;$$
5: $$w_j := \arg\min_{w_j} \sum_{i=1}^{\ell} \bigl(\varphi(f_j(x_i), w_j) - z_i\bigr)^2 + \tau R(w_j);$$
6: уменьшить коэффициент регуляризации τ;
7: **пока** $Q(w, X^\ell)$ и/или $Q(w, X^k)$ заметно уменьшаются;

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

Связь МНК с методом максимума правдоподобия

Модель данных с некоррелированным гауссовским шумом:

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

Эквивалентная запись: $y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $\mu_i = \mathsf{E} y_i = a(x_i, w)$.

МНК эквивалентен методу максимума правдоподобия (ММП):

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

$$-\ln L(\varepsilon_1, \dots, \varepsilon_\ell | w) = \mathrm{const}(w) + \frac{1}{2} \sum_{i=1}^{\ell} \frac{1}{\sigma_i^2} \bigl(a(x_i, w) - y_i\bigr)^2 \to \min_w;$$

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

Обобщённая линейная модель (Generalized Linear Model, GLM)

Нормальная линейная модель для математического ожидания:

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

Обобщённая линейная модель для математического ожидания:

$$y_i \sim \mathrm{Exp}(\theta_i, \phi_i), \qquad \theta_i = x_i^\top w = g(\mathsf{E} y_i) \text{ — почему?}$$

Exp — *экспоненциальное семейство распределений* с параметрами θ_i, ϕ_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. $y_i \sim \mathrm{Exp}(\theta_i, \phi_i)$:

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

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

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

Примеры распределений из экспоненциального семейства

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

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

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

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

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

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

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

Примеры распределений из экспоненциального семейства

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

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

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

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

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

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

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

Примеры распределений из экспоненциального семейства

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

Контр-примеры не экспоненциальных распределений:

- t-распределение Стьюдента, Коши, гипергеометрическое

Максимизация правдоподобия для GLM

Принцип максимума правдоподобия:

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

где θ_i линейно зависит от w: $\theta_i = x_i^{\mathsf{T}} w = \sum_{j=1}^{n} w_j f_j(x_i)$.

Метод Ньютона-Рафсона: $w^{t+1} := w^t + h_t \bigl(L''(w^t)\bigr)^{-1} L'(w^t)$.

Компоненты вектора градиента $L'(w)$:

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

Компоненты матрицы Гессе $L''(w)$:

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

Матричные обозначения

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

$\tilde F = D_t F,\ \ D_t = \mathrm{diag}\Bigl(\sqrt{\frac{1}{\phi_i}c''(\theta_i)}\Bigr)$ — веса объектов, $\theta_i = x_i^{\mathsf T} w^t$;

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

Тогда метод Ньютона-Рафсона снова приводит к IRLS:

$$w^{t+1} := w^t - h_t \underbrace{\bigl(F^{\mathsf T} D_t D_t F\bigr)^{-1} F^{\mathsf T} D_t}_{(\tilde F^{\mathsf T}\tilde F)^{-1}\tilde F^{\mathsf T}} \Bigl(\underbrace{\sqrt{\frac{\phi_i}{c''(\theta_i)}}\,\frac{y_i - c'(\theta_i)}{\phi_i}}_{\tilde y_i}\Bigr)_{\ell\times 1}.$$

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

$$Q(w) = \|\tilde F w - \tilde y\|^2 \to \min_w .$$

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 модели.

$w := (F^\intercal F)^{-1} F^\intercal y$ — нулевое приближение, обычный МНК;
для $t := 1, 2, 3, \ldots$
  $\theta_i = x_i^\intercal w^t$ для всех $i = 1, \ldots, \ell$;
  $\gamma_i := \sqrt{\frac{1}{\phi_i} c''(\theta_i)}$ для всех $i = 1, \ldots, \ell$;
  $\tilde{F} := \mathrm{diag}(\gamma_1, \ldots, \gamma_\ell) F$;
  $\tilde{y}_i := \frac{y_i - c'(\theta_i)}{\phi_i \gamma_i}$ для всех $i = 1, \ldots, \ell$;
  выбрать градиентный шаг h_t;
  $w := w + h_t (\tilde{F}^\intercal \tilde{F})^{-1} \tilde{F}^\intercal \tilde{y}$;
  если $\{\theta_i\}$ мало изменились **то** выйти из цикла;

Логистическая регрессия как частный случай GLM

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

$$\theta_i = g(\mu_i) = \ln \tfrac{\mu_i}{1-\mu_i} \qquad \mu_i = g^{-1}(\theta_i) = \tfrac{1}{1+\exp(-\theta_i)} \equiv \sigma(\theta_i)$$

Апостериорная вероятность классов, $\tilde{y}_i = 2y_i - 1 \in \{-1, +1\}$:

$$\left.\begin{aligned} \mathsf{P}(y_i{=}1|x_i) &= \mathsf{E}y_i = \mu_i = \sigma(\theta_i) \\ \mathsf{P}(y_i{=}0|x_i) &= 1-\mu_i = \sigma(-\theta_i) \end{aligned}\right\} \qquad p(y_i|x_i) = \sigma(\underbrace{\tilde{y}_i x_i^{\mathsf{T}} w}_{\text{margin}})$$

Максимум правдоподобия $\iff$ минимум критерия log-loss:

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

$$\iff \sum_{i=1}^{\ell} \ln\bigl(1 + \exp(-\tilde{y}_i x_i^{\mathsf{T}} w)\bigr) \;\to\; \min_w$$

Логистическая регрессия как частный случай GLM

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

$$\theta_i = g(\mu_i) = \ln \frac{\mu_i}{1-\mu_i} \qquad \mu_i = g^{-1}(\theta_i) = \frac{1}{1+\exp(-\theta_i)} \equiv \sigma(\theta_i)$$

Всего лишь из двух предположений:

- y_i — бернуллиевские случайные величины с $\mathsf{E}y_i = \mu_i$
- модель линейна и μ_i монотонно зависит от $\theta_i = x_i^{\mathsf{T}} w$

вытекают все основные свойства логистической регрессии:

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

$$x_i^{\mathsf{T}} w = \theta_i = \ln \frac{\mu_i}{1-\mu_i} = \ln \frac{\mathsf{P}(y_i = 1|x_i)}{\mathsf{P}(y_i = 0|x_i)}.$$

Метод наименьших модулей (Least Absolute Deviation Regression)

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

$Q = \sum\limits_{i=1}^{\ell} \mathscr{L}(\varepsilon_i) \to \min\limits_{w}$ — 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, $\mathscr{L}(\varepsilon) = |\varepsilon|$:

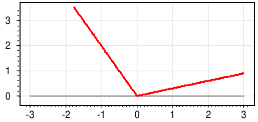
$$\sum_{i=1}^{\ell} |a - y_i| \to \min_{a} \quad \Rightarrow \quad a = \mathrm{median}\{y_1, \ldots, y_\ell\} = y^{(\ell/2)},$$

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

Медиана более устойчива к редким большим выбросам y_i.

Квантúльная регрессия (Quantile Regression)

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



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

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

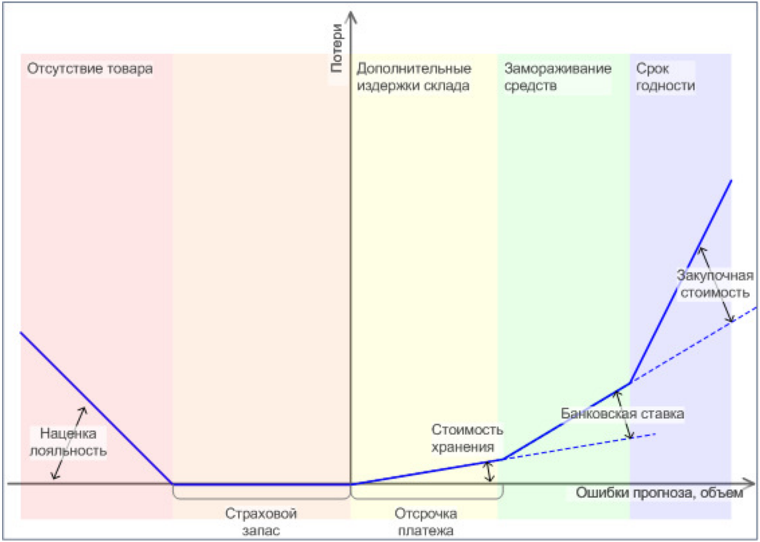
Линейная модель регрессии: $a(x_i, w) = \langle x_i, w \rangle$.

Сведение к задаче линейного программирования:

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

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

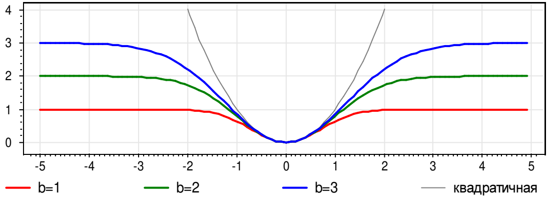
Пример. Задача прогнозирования объёмов продаж



Робастная регрессия (Robust Regression)

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

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



Постановка задачи:

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

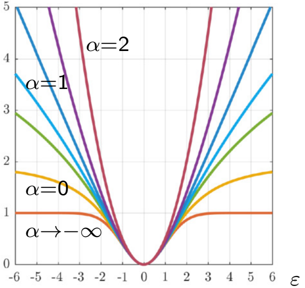
Эта задача также решается методом Ньютона-Рафсона.

Функции потерь для робастной регрессии

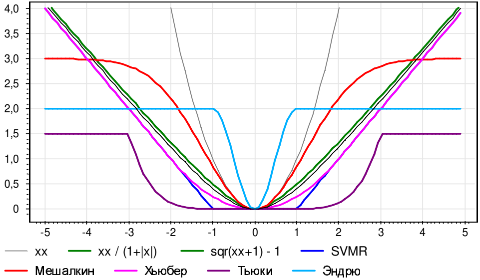
Семейство функций потерь Баррона с параметром α:

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

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



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

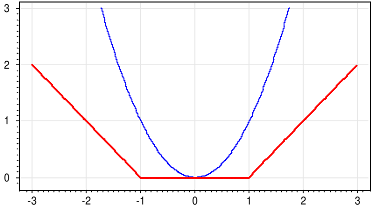


Jonathan T. Barron. A General and Adaptive Robust Loss Function. 2019.

Напоминание: SVM-регрессия. Тоже робастная регрессия

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

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



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
  — сводится к последовательности линейных регрессий
  — метод Ньютона-Рафсона приводит к IRLS
- 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