Нелинейная регрессия
Обобщённые линейные модели
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

МФТИ • 17 октября 2020

Содержание

Напоминание: 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;
  $y_i = y(x_i)$, $y : X \to Y$ — неизвестная зависимость;
- $a(x) = f(x, \alpha)$ — модель зависимости,
  $\alpha \in \mathbb{R}^p$ — вектор параметров модели.
- Метод наименьших квадратов (МНК):

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

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:

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

Метод Ньютона–Рафсона:

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

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

$Q'(\alpha^t)$ — градиент функционала Q в точке α^t, вектор из $\mathbb{R}^p$
$Q''(\alpha^t)$ — гессиан функционала Q в точке α^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 α^t:

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

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:

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

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.

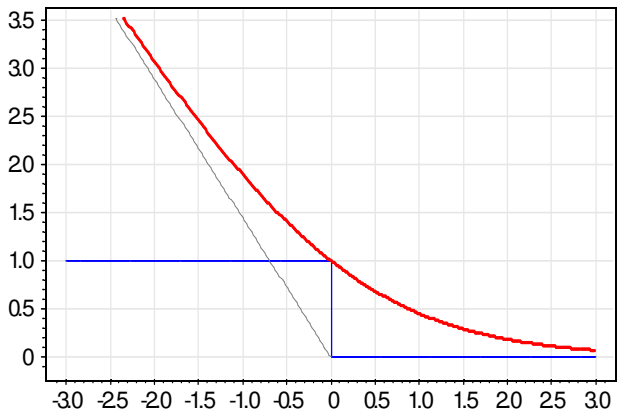
Задача классификации. Логистическая регрессия

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

Функционал аппроксимированного эмпирического риска:

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

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



$M_i = w^\intercal x_i y_i$

Метода Ньютона-Рафсона

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

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

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.

Снова сведение к задаче линейной регрессии

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

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

Это совпадает с МНК-решением линейной задачи 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:

- $\sigma_i = \mathrm{P}(y_i|x_i)$ — 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

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 $i = 1, \ldots, \ell$;
  $\tilde{F} := \operatorname{diag}(\gamma_1, \ldots, \gamma_\ell)F$;
  $\tilde{y}_i := y_i\sqrt{(1-\sigma_i)/\sigma_i}$ 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 функциями признаков $\varphi_j : \mathbb{R} \to \mathbb{R}$:

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

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

ИДЕЯ: поочерёдно уточнять φ_j по выборке $\left(f_j(x_i), z_i\right)_{i=1}^{\ell}$, постепенно ослабляя регуляризатор гладкости $R(\alpha_j)$ (можно использовать сплайны или ядерное сглаживание):

$$Q(\alpha_j) + \tau R(\alpha_j) \to \min_{\alpha_j}$$

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

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

Метод backfitting [Хасти, Тибширани, 1986]

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

1: начальное приближение:
   $\alpha :=$ решение задачи МЛР с признаками $f_j(x)$;
   $\varphi_j(f_j, \alpha_j) := \alpha_j f_j(x), \;\; j = 1, \ldots, n$;
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 τ;
7: **пока** $Q(\alpha, X^\ell)$ и/или $Q(\alpha, X^k)$ заметно уменьшаются;

Напоминание: связь ММП и МНК

Модель данных с некоррелированным гауссовским шумом:

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

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

МНК эквивалентен методу максимума правдоподобия (ММП):

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

$$-\ln L(\varepsilon_1, \ldots, \varepsilon_\ell | \alpha) = \text{const}(\alpha) + \frac{1}{2}\sum_{i=1}^{\ell} \frac{1}{\sigma_i^2}\big(f(x_i, \alpha) - y_i\big)^2 \to \min_\alpha;$$

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

Обобщённая линейная модель (Generalized Linear Model, GLM)

Нормальная линейная модель для математического ожидания:

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

Обобщённая линейная модель для математического ожидания:

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

$g(\mu)$ — монотонная *функция связи* (link function),
Exp — *экспоненциальное семейство распределений*
с параметрами θ_i, ϕ_i и параметрами-функциями $c(\theta)$, $h(y, \phi)$:

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

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, $y_i \in \mathbb{R}$:

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

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

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

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

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

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

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
- 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:

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

где θ_i зависит от α: $\theta_i = x_i^{\mathsf{T}}\alpha = \sum_{j=1}^{n} \alpha_j f_j(x_i)$.

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

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

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

$\tilde{F} = W_t F, \quad W_t = \operatorname{diag}\Big(\sqrt{\frac{1}{\phi_i} c''(\theta_i)}\Big), \quad \theta_i = x_i^{\mathsf{T}} \alpha^t;$

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

Тогда метод Ньютона-Рафсона снова приводит к IRLS:

$$\alpha^{t+1} := \alpha^t - h_t \underbrace{\big(F^{\mathsf{T}} W_t W_t F\big)^{-1} F^{\mathsf{T}} W_t}_{(\tilde{F}^{\mathsf{T}}\tilde{F})^{-1}\tilde{F}^{\mathsf{T}}} \Bigg(\underbrace{\sqrt{\frac{\phi_i}{c''(\theta_i)}} \frac{y_i - c'(\theta_i)}{\phi_i}}_{\tilde{y}_i} \Bigg)_{\ell \times 1}.$$

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

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

Логистическая регрессия как частный случай GLM

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

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

Принцип максимума правдоподобия приводит к log-loss:

$$\sum_{i=1}^{\ell} \ln \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \sum_{i=1}^{\ell} y_i \ln \mu_i + (1-y_i)\ln(1-\mu_i)$$

Выражение для апостериорной вероятности класса $+1$:

$$\mathsf{P}(y_i=1|x_i) = \mathsf{E}y_i = \mu_i = \frac{1}{1+\exp(-\theta_i)} = \sigma(\theta_i) = \sigma(x_i^{\mathsf{T}}\alpha)$$

Линейный классификатор и *отношение шансов* (odds ratio):

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

Метод наименьших модулей

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

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

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

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

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

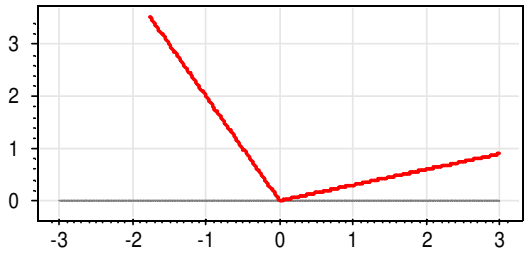
Метод наименьших модулей, $\mathscr{L}(\varepsilon) = |\varepsilon|$:

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

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

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

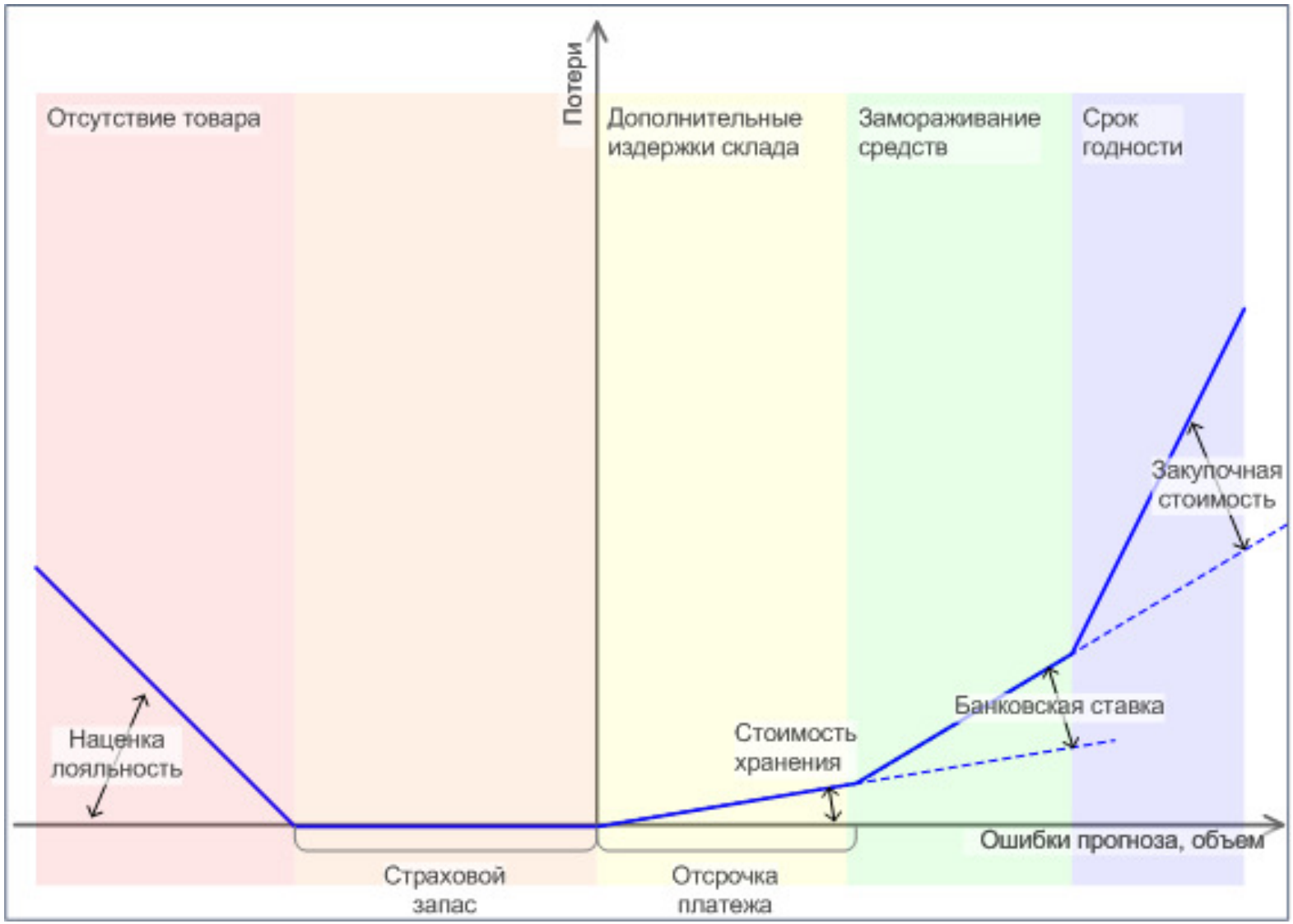
Линейная модель регрессии: $a(x_i) = \langle x_i, w \rangle$.

Сведение к задаче линейного программирования:

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

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

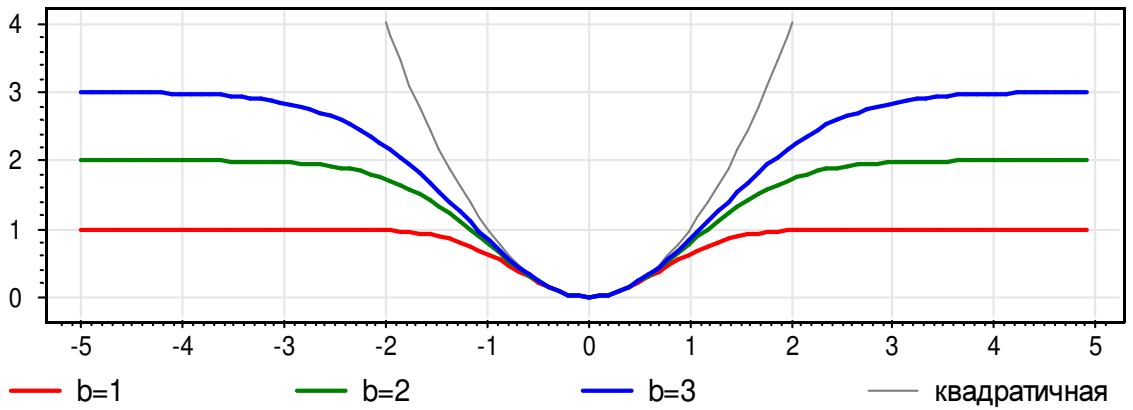
Пример. Задача прогнозирования объёмов продаж



Робастная регрессия

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

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



Постановка задачи:

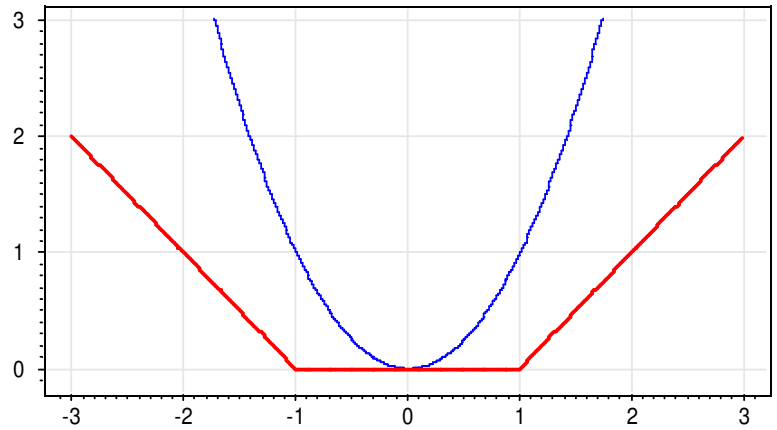
$$\sum_{i=1}^{\ell} \exp\Big(-\frac{1}{b}(f(x_i, \alpha) - y_i)^2\Big) \to \max_{\alpha}.$$

Эта задача также решается методом Ньютона-Рафсона.

SVM-регрессия (напоминание)

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

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



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
  — используется метод Ньютона-Рафсона
- Логистическая регрессия
  — не регрессия, а классификация
  — используется метод Ньютона-Рафсона
- Обобщённая линейная модель (GLM)
  — мощно обобщает обычную и логистическую регрессию
  — используется метод Ньютона-Рафсона
- Обобщённая аддитивная регрессия (GAM, backfitting)
  — сводится к серии одномерных сглаживаний
- Неквадратичные функции потерь
  — проблемно-ориентированные (зависят от задачи)
  — приводят к разным методам, отличным от МНК