Нелинейная регрессия. Непараметрическая регрессия. Нестандартные функции потерь

К. В. Воронцов
vokov@forecsys.ru

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

27 марта 2014

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;
  $y_i = y(x_i)$, $y : X \to Y$ — 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

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

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

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:

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

Метод Ньютона-Гаусса

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

Формула t-й итерации метода Ньютона–Гаусса:

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

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.

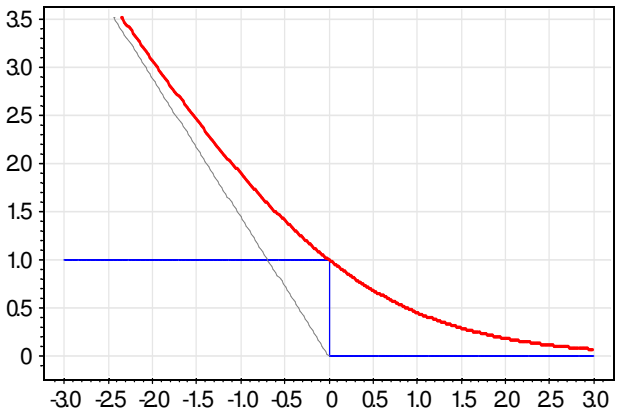
Задача классификации. Логистическая регрессия

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

Функционал аппроксимированного эмпирического риска:

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

где $\mathscr{L}(M) = \log\big(1 + e^{-M}\big)$ — 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.

$\sigma_i = \mathsf{P}(y_i|x_i)$ — вероятность правильной классификации x_i.

Матричные обозначения

$F_{\ell\times n} = \big(f_j(x_i)\big)$ — матрица «объекты–признаки»;
$\Gamma_{\ell\times\ell} = \operatorname{diag}\big(\sqrt{(1-\sigma_i)\sigma_i}\,\big)$ — диагональная матрица;
$\tilde{F} = \Gamma F$ — взвешенная матрица «объекты–признаки»;
$\tilde{y}_i = y_i\sqrt{(1-\sigma_i)/\sigma_i}$, $\tilde{y} = (\tilde{y}_i)_{i=1}^{\ell}$ — 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:

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

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.

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: $\quad \gamma_i := \sqrt{(1-\sigma_i)\sigma_i}$ для всех $i = 1, \ldots, \ell$;
5: $\quad \tilde{F} := \operatorname{diag}(\gamma_1, \ldots, \gamma_\ell) F$;
6: $\quad \tilde{y}_i := y_i \sqrt{(1-\sigma_i)/\sigma_i}$ для всех $i = 1, \ldots, \ell$;
7: $\quad$ выбрать градиентный шаг h_t;
8: $\quad w := w + h_t (\tilde{F}^\intercal \tilde{F})^{-1} \tilde{F}^\intercal \tilde{y}$;
9: $\quad$ **если** $\{\sigma_i\}$ мало изменились **то** выйти из цикла;

Обобщение линейной модели регрессии

Пусть $\varphi_j : \mathbb{R} \to \mathbb{R}$ — некоторые нелинейные преобразования исходных признаков. Модель регрессии:

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

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

ИДЕЯ: будем по очереди уточнять функции φ_j по обучающей выборке $\left(f_j(x_i), z_i\right)_{i=1}^{\ell}$:

$$Q(\varphi_j, X^\ell) = \sum_{i=1}^{\ell} \Bigg(\varphi_j(f_j(x_i)) - \underbrace{\Big(y_i - \sum_{k=1, k\neq j}^{n} \varphi_k(f_k(x_i)) \Big)}_{z_i = \text{const}(\varphi_j)} \Bigg)^2 \to \min_{\varphi_j}.$$

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 $f_j(x)$;
   $\varphi_j(x) := \alpha_j f_j(x), \ \ j = 1, \ldots, n$;
2: **повторять**
3: **для** $j = 1, \ldots, n$
4: $z_i := y_i - \sum\limits_{k=1, k\neq j}^{n} \varphi_k(f_k(x_i)), \ \ i = 1, \ldots, \ell$;
5: $\varphi_j := \arg\min\limits_{\varphi} \sum\limits_{i=1}^{\ell} \big(\varphi(f_j(x_i)) - z_i\big)^2$; — одномерная регрессия
6: $Q_j := \sum\limits_{i=1}^{\ell} \big(\varphi_j(f_j(x_i)) - z_i\big)^2$;
7: **пока** значения Q_j не стабилизируются

Формула Надарая–Ватсона

Приближение константой $a(x) = \alpha$ в окрестности $x \in X$:

$$Q(\alpha; X^\ell) = \sum_{i=1}^{\ell} w_i(x)(\alpha - y_i)^2 \to \min_{\alpha \in \mathbb{R}};$$

где $w_i(x) = K\left(\frac{\rho(x,x_i)}{h}\right)$ — веса объектов x_i относительно x;
$K(r)$ — *ядро*, невозрастающее, ограниченное, гладкое;
h — ширина окна сглаживания.

Формула *ядерного сглаживания* Надарая–Ватсона:

$$a_h(x; X^\ell) = \frac{\sum_{i=1}^{\ell} y_i w_i(x)}{\sum_{i=1}^{\ell} w_i(x)} = \frac{\sum_{i=1}^{\ell} y_i K\left(\frac{\rho(x,x_i)}{h}\right)}{\sum_{i=1}^{\ell} K\left(\frac{\rho(x,x_i)}{h}\right)}.$$

Обоснование формулы Надарая–Ватсона

Теорема

Пусть выполнены следующие условия:

1) *выборка* $X^\ell = (x_i, y_i)_{i=1}^\ell$ *простая, из распределения* $p(x, y)$;
2) *ядро* $K(r)$ *ограничено:* $\int_0^\infty K(r)\,dr < \infty$, $\lim\limits_{r\to\infty} rK(r) = 0$;
3) *зависимость* $\mathsf{E}(y|x)$ *не имеет вертикальных асимптот:*
   $\mathsf{E}(y^2|x) = \int_Y y^2 p(y|x)\,dy < \infty$ *при любом* $x \in X$;
4) *последовательность* h_ℓ *убывает, но не слишком быстро:*
   $\lim\limits_{\ell\to\infty} h_\ell = 0$, $\lim\limits_{\ell\to\infty} \ell h_\ell = \infty$.

Тогда имеет место сходимость по вероятности:

$$a_{h_\ell}(x; X^\ell) \xrightarrow{P} \mathsf{E}(y|x) \text{ в любой точке } x \in X,$$

в которой $\mathsf{E}(y|x)$, $p(x)$ *и* $\mathsf{D}(y|x)$ *непрерывны и* $p(x) > 0$.

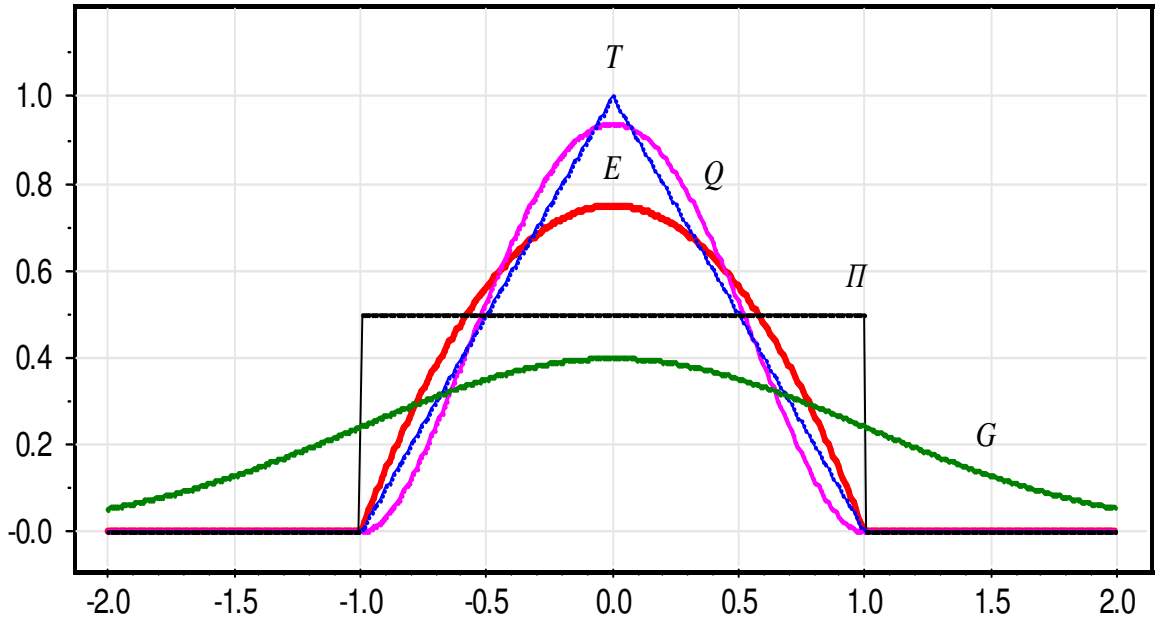
- Ядро $K(r)$
  — существенно влияет на гладкость функции $a_h(x)$,
  — слабо влияет на качество аппроксимации.
- Ширина окна h
  — существенно влияет на качество аппроксимации.
- При неравномерной сетке $\{x_i\}$ — переменная ширина окна:

$$w_i(x) = K\left(\frac{\rho(x, x_i)}{h(x)}\right),$$

  где $h(x) = \rho(x, x^{(k+1)})$, $x^{(k+1)}$ — k-й сосед объекта x.
- Оптимизация ширины окна по скользящему контролю:

$$\mathrm{LOO}(h, X^\ell) = \sum_{i=1}^{\ell} \Big(a_h\big(x_i; X^\ell \setminus \{x_i\}\big) - y_i\Big)^2 \to \min_h.$$

Часто используемые ядра $K(r)$



$E(r) = (1 - r^2)\big[|r| \leqslant 1\big]$ — оптимальное (Епанечникова);
$Q(r) = (1 - r^2)^2\big[|r| \leqslant 1\big]$ — квартическое;
$T(r) = \big(1 - |r|\big)\big[|r| \leqslant 1\big]$ — треугольное;
$G(r) = \exp\big(-\frac{1}{2}r^2\big)$ — гауссовское;
$\Pi(r) = \big[|r| \leqslant 1\big]$ — прямоугольное.

Локально взвешенное сглаживание (LOWESS — LOcally WEighted Scatter plot Smoothing)

Основная идея:
чем больше величина ошибки $\varepsilon_i = \left|a_h\left(x_i; X^\ell \backslash \{x_i\}\right) - y_i\right|$, тем в большей степени прецедент (x_i, y_i) является выбросом, и тем меньше должен быть его вес $w_i(x)$.

Эвристика:
домножить веса $w_i(x)$ на коэффициенты $\gamma_i = \tilde{K}(\varepsilon_i)$,
где $\tilde{K}$ — ещё одно ядро, вообще говоря, отличное от $K(r)$.

Рекомендация:
квартическое ядро $\tilde{K}(\varepsilon) = K_Q\left(\frac{\varepsilon}{6\,\mathrm{med}\{\varepsilon_i\}}\right)$,
где $\mathrm{med}\{\varepsilon_i\}$ — медиана вариационного ряда ошибок.

Алгоритм LOWESS

Вход: X^ℓ — обучающая выборка;
Выход: коэффициенты γ_i, $i = 1, \ldots, \ell$;

1: инициализация: $\gamma_i := 1$, $i = 1, \ldots, \ell$;
2: **повторять**
3: **для всех** объектов $i = 1, \ldots, \ell$
4: вычислить оценки скользящего контроля:

$$a_i := a_h\big(x_i; X^\ell \setminus \{x_i\}\big) = \frac{\sum\limits_{j=1,\, j\neq i}^{\ell} y_j \gamma_j K\big(\frac{\rho(x_i,x_j)}{h(x_i)}\big)}{\sum\limits_{j=1,\, j\neq i}^{\ell} \gamma_j K\big(\frac{\rho(x_i,x_j)}{h(x_i)}\big)};$$

5: **для всех** объектов $i = 1, \ldots, \ell$
6: $\gamma_i := \tilde{K}\big(|a_i - y_i|\big)$;
7: **пока** коэффициенты γ_i не стабилизируются;

Метод наименьших модулей

$\varepsilon_i = \big(a(x_i) - y_i\big)$ — 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, $\mathscr{L}(\varepsilon) = |\varepsilon|$:

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

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

Квантильная регрессия

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

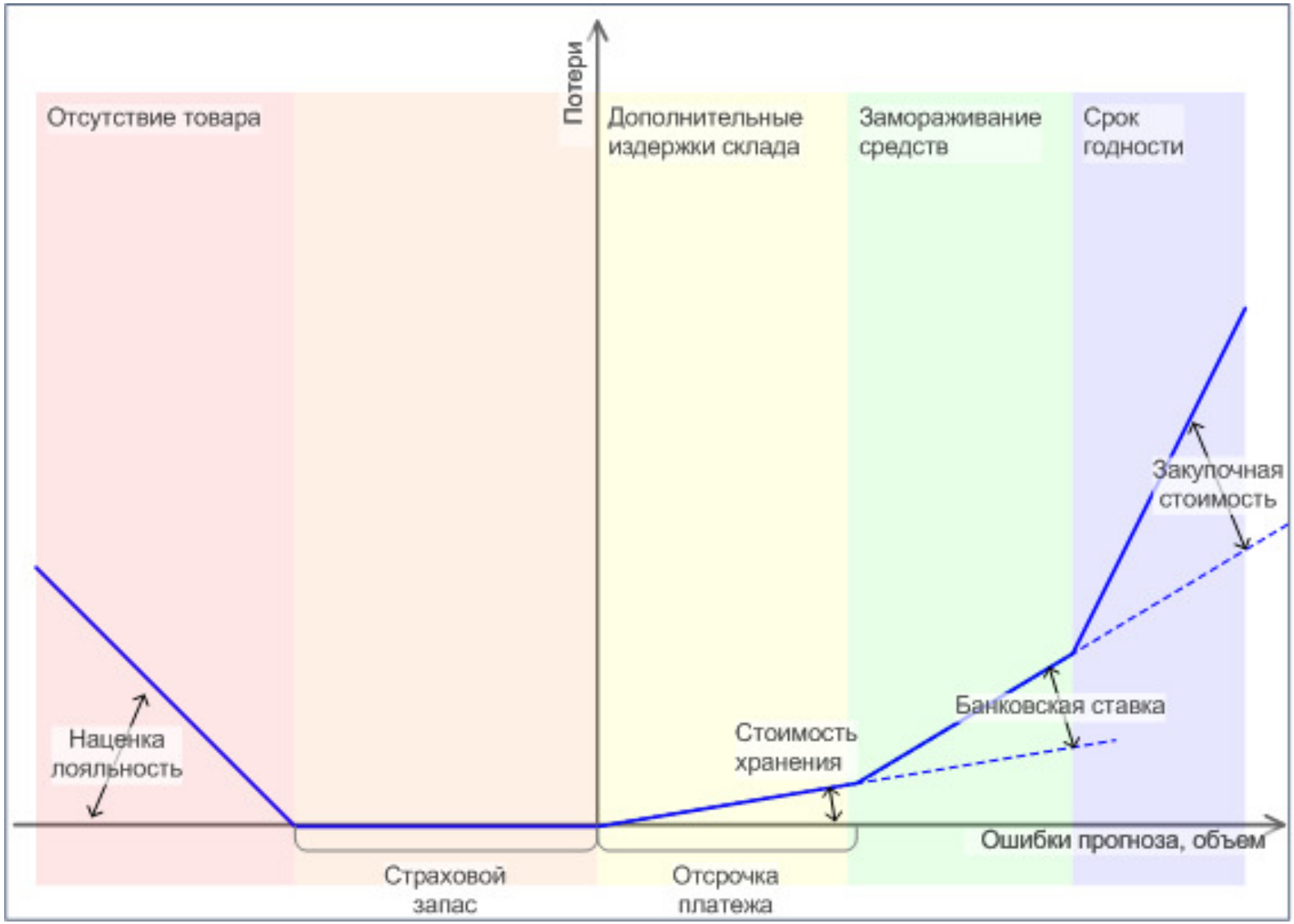
Линейная модель регрессии: $a(x_i) = \langle x_i, w \rangle$.

Сведение к задаче линейного программирования:

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

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

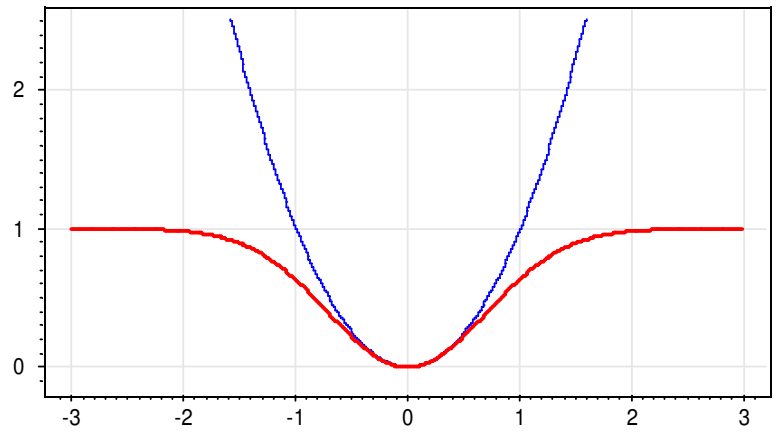
Задача прогнозирования объёмов продаж



Робастная регрессия

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

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



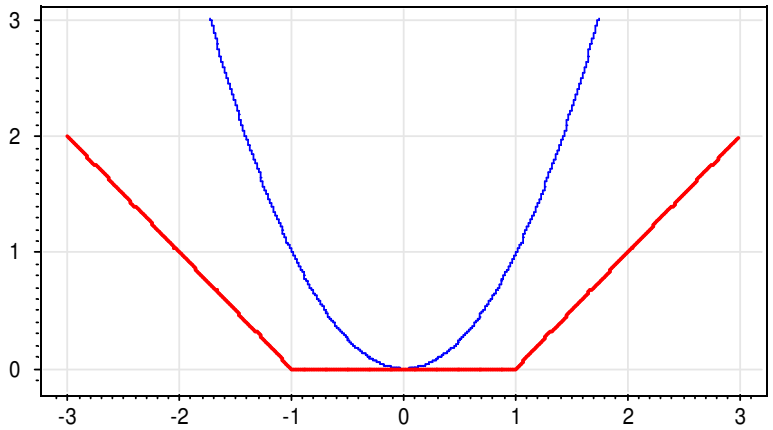
$$\sum_{i=1}^{\ell} \exp\left(-\frac{1}{2\sigma^2}(f(x_i, \alpha) - y_i)^2\right) \to \max_{\alpha}.$$

Задача решается методом Ньютона-Рафсона.

SVM-регрессия

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

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



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

SVM-регрессия

Замена переменных:

$$\xi_i^+ = \big(\langle w, x_i\rangle - w_0 - y_i - \delta\big)_+;$$
$$\xi_i^- = \big(-\langle w, x_i\rangle + w_0 + y_i - \delta\big)_+;$$

Постановка задачи SVM-регрессии:

$$\begin{cases} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}(\xi_i^+ + \xi_i^-) \to \min\limits_{w, w_0, \xi^+, \xi^-}; \\ y_i - \delta - \xi_i^- \leqslant \langle w, x_i\rangle - w_0 \leqslant y_i + \delta + \xi_i^+, \quad i = 1, \ldots, \ell; \\ \xi_i^- \geqslant 0, \quad \xi_i^+ \geqslant 0, \quad i = 1, \ldots, \ell. \end{cases}$$

Это задача квадратичного программирования с линейными ограничениями-неравенствами

Резюме в конце лекции

- Нелинейная регрессия
  — сводится к последовательности линейных
- Непараметрическая регрессия (сглаживание)
  — очень просто, но главное — подобрать ширину окна
- Неквадратичные функции потерь
  — зависят от задачи