Линейная и нелинейная регрессия

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

19 марта 2018

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):

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

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

Многомерная линейная регрессия

$f_1(x), \ldots, f_n(x)$ — числовые признаки;
Модель многомерной линейной регрессии:

$$f(x, \alpha) = \sum_{j=1}^{n} \alpha_j f_j(x), \qquad \alpha \in \mathbb{R}^n.$$

Матричные обозначения:

$$\underset{\ell \times n}{F} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}, \quad \underset{\ell \times 1}{y} = \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}, \quad \underset{n \times 1}{\alpha} = \begin{pmatrix} \alpha_1 \\ \ldots \\ \alpha_n \end{pmatrix}.$$

Функционал квадрата ошибки:

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

Нормальная система уравнений

Необходимое условие минимума в матричном виде:

$$\frac{\partial Q}{\partial \alpha}(\alpha) = 2F^{\mathsf{T}}(F\alpha - y) = 0,$$

откуда следует *нормальная система* задачи МНК:

$$F^{\mathsf{T}}F\alpha = F^{\mathsf{T}}y,$$

где $F^{\mathsf{T}}F$ — матрица размера $n \times n$.

Решение системы: $\alpha^* = (F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}y = F^{+}y$.

Значение функционала: $Q(\alpha^*) = \|P_F y - y\|^2$,

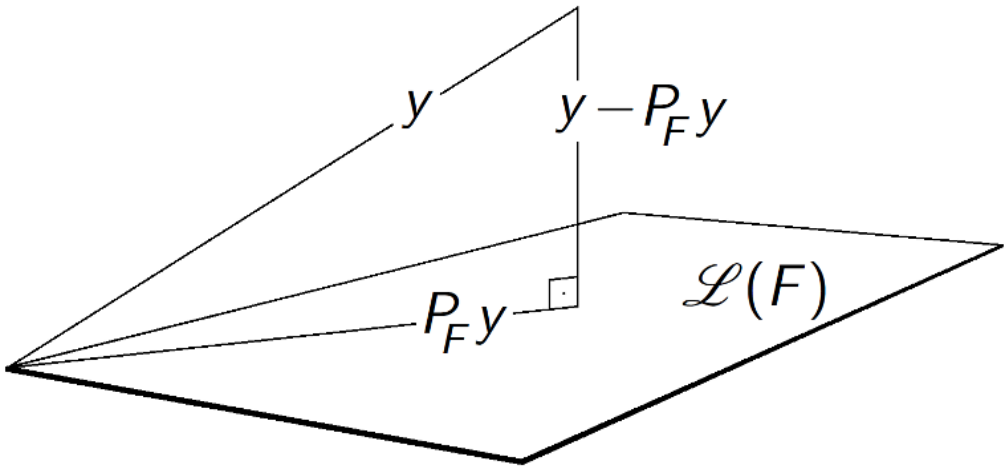
где $P_F = FF^{+} = F(F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}$ — *проекционная матрица*.

Геометрическая интерпретация МНК

Линейная оболочка столбцов матрицы $F = (f_1, \ldots, f_n)$, $f_j \in \mathbb{R}^\ell$:

$$\mathscr{L}(F) = \left\{ \sum_{j=1}^{n} \alpha_j f_j \;\middle|\; \alpha \in \mathbb{R}^n \right\}$$

$P_F = F(F^\mathsf{T}F)^{-1}F^\mathsf{T}$ — проекционная матрица
$P_F y$ — проекция вектора $y \in \mathbb{R}^\ell$ на подпространство $\mathscr{L}(F)$
$(I_\ell - P_F)y$ — проекция y на его ортогональное дополнение



МНК — это опускание перпендикуляра в $\mathbb{R}^\ell$ из y на $\mathscr{L}(F)$

Сингулярное разложение

Произвольная $\ell \times n$-матрица представима в виде *сингулярного разложения* (singular value decomposition, SVD):

$$F = VDU^{\mathsf{T}}.$$

Основные свойства сингулярного разложения:

1. $\ell \times n$-матрица $V = (v_1, \ldots, v_n)$ ортогональна, $V^{\mathsf{T}}V = I_n$, столбцы v_j — собственные векторы матрицы FF^{T};
2. $n \times n$-матрица $U = (u_1, \ldots, u_n)$ ортогональна, $U^{\mathsf{T}}U = I_n$, столбцы u_j — собственные векторы матрицы $F^{\mathsf{T}}F$;
3. $n \times n$-матрица D диагональна, $D = \operatorname{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\right)$, $\lambda_j \geqslant 0$ — собственные значения матриц $F^{\mathsf{T}}F$ и FF^{T}.

Решение МНК через сингулярное разложение

Псевдообратная F^+, вектор МНК-решения α^*,
МНК-аппроксимация целевого вектора $F\alpha^*$:

$$F^+ = (UDV^\intercal VDU^\intercal)^{-1} UDV^\intercal = UD^{-1}V^\intercal = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j v_j^\intercal;$$

$$\alpha^* = F^+ y = UD^{-1}V^\intercal y = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^\intercal y);$$

$$F\alpha^* = P_F y = (VDU^\intercal) UD^{-1}V^\intercal y = VV^\intercal y = \sum_{j=1}^{n} v_j (v_j^\intercal y);$$

$$\|\alpha^*\|^2 = \|UD^{-1}V^\intercal y\|^2 = \|D^{-1}V^\intercal y\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j} (v_j^\intercal y)^2.$$

Регуляризация L_2 (гребневая регрессия)

Штраф за увеличение нормы вектора весов $\|\alpha\|$:

$$Q_\tau(\alpha) = \|F\alpha - y\|^2 + \tfrac{1}{\sigma}\|\alpha\|^2,$$

где $\tau = \frac{1}{\sigma}$ — неотрицательный *параметр регуляризации*.

Вероятностная интерпретация: априорное распределение вектора α — гауссовское с ковариационной матрицей σI_n.

Модифицированное МНК-решение (τI_n — «гребень»):

$$\alpha^*_\tau = (F^\intercal F + \tau I_n)^{-1} F^\intercal y.$$

Преимущество сингулярного разложения:
можно подбирать параметр τ, вычислив SVD только один раз.

Регуляризованный МНК через сингулярное разложение

Вектор регуляризованного МНК-решения α_τ^*
и МНК-аппроксимация целевого вектора $F\alpha_\tau^*$:

$$\alpha_\tau^* = U(D^2 + \tau I_n)^{-1} D V^\intercal y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^\intercal y);$$

$$F\alpha_\tau^* = VDU^\intercal \alpha_\tau^* = V \operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^\intercal y = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} v_j (v_j^\intercal y);$$

$$\|\alpha_\tau^*\|^2 = \|(D^2 + \tau I_n)^{-1} D V^\intercal y\|^2 = \sum_{j=1}^{n} \frac{\lambda_j}{(\lambda_j + \tau)^2} (v_j^\intercal y)^2.$$

$F\alpha_\tau^* \neq F\alpha^*$, но зато решение становится гораздо устойчивее.

Выбор параметра регуляризации τ

Контрольная выборка: $X^k = (x'_i, y'_i)_{i=1}^k$;

$$\underset{k\times n}{F'} = \begin{pmatrix} f_1(x'_1) & \ldots & f_n(x'_1) \\ \ldots & \ldots & \ldots \\ f_1(x'_k) & \ldots & f_n(x'_k) \end{pmatrix}, \quad \underset{k\times 1}{y'} = \begin{pmatrix} y'_1 \\ \ldots \\ y'_k \end{pmatrix}.$$

Вычисление функционала Q на контрольных данных T раз потребует $O(kn^2 + knT)$ операций:

$$Q(\alpha^*_\tau, X^k) = \|F'\alpha^*_\tau - y'\|^2 = \Big\| \underbrace{F'U}_{k\times n} \operatorname{diag}\Big(\frac{\sqrt{\lambda_j}}{\lambda_j+\tau}\Big) \underbrace{V^{\mathsf{T}}y}_{n\times 1} - y' \Big\|^2.$$

Зависимость $Q(\tau)$ обычно имеет характерный минимум.

Регуляризация сокращает «эффективную размерность»

Сжатие (shrinkage) или *сокращение весов* (weight decay):

$$\|\alpha_\tau^*\|^2 = \sum_{j=1}^{n} \frac{\lambda_j}{(\lambda_j+\tau)^2}(v_j^\mathsf{T} y)^2 \quad < \quad \|\alpha^*\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j}(v_j^\mathsf{T} y)^2.$$

Почему говорят о *сокращении эффективной размерности*?

Роль размерности играет след проекционной матрицы:

$$\operatorname{tr} F(F^\mathsf{T}F)^{-1}F^\mathsf{T} = \operatorname{tr}(F^\mathsf{T}F)^{-1}F^\mathsf{T}F = \operatorname{tr} I_n = n.$$

При использовании регуляризации:

$$\operatorname{tr} F(F^\mathsf{T}F + \tau I_n)^{-1}F^\mathsf{T} = \operatorname{tr}\operatorname{diag}\left(\frac{\lambda_j}{\lambda_j+\tau}\right) = \sum_{j=1}^{n}\frac{\lambda_j}{\lambda_j+\tau} < n.$$

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:

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

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.

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 $R(\alpha_j)$:

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

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

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

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;

нулевое приближение:

$\alpha :=$ решение задачи МЛР с признаками $f_j(x)$;

$\varphi_j(f_j, \alpha_j) := \alpha_j f_j(x)$, для всех признаков $j = 1, \ldots, n$;

повторять

  для всех признаков $j = 1, \ldots, n$

$$z_i := y_i - \sum_{k=1, k\neq j}^{n} \varphi_k(f_k(x_i), \alpha_k), \quad i = 1, \ldots, \ell;$$

$$\alpha_j := \arg\min_{\alpha} \sum_{i=1}^{\ell} \big(\varphi(f_j(x_i), \alpha) - z_i\big)^2 + \tau R(\alpha);$$

  уменьшить коэффициент регуляризации τ;

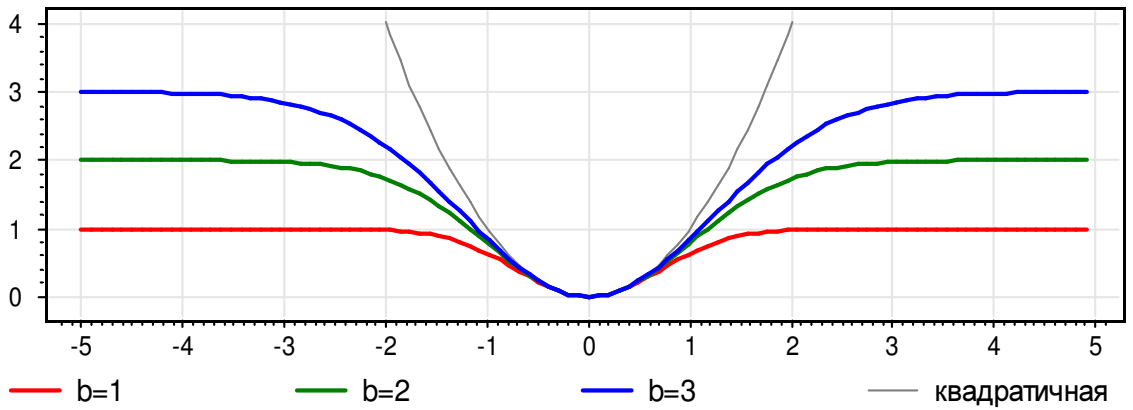
пока $Q(\alpha, X^\ell)$ и/или $Q(\alpha, X^k)$ заметно уменьшаются;

Робастная регрессия

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

Неквадратичная функция потерь:

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



Постановка задачи:

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

Задача также решается методом Ньютона-Рафсона.

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

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

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.

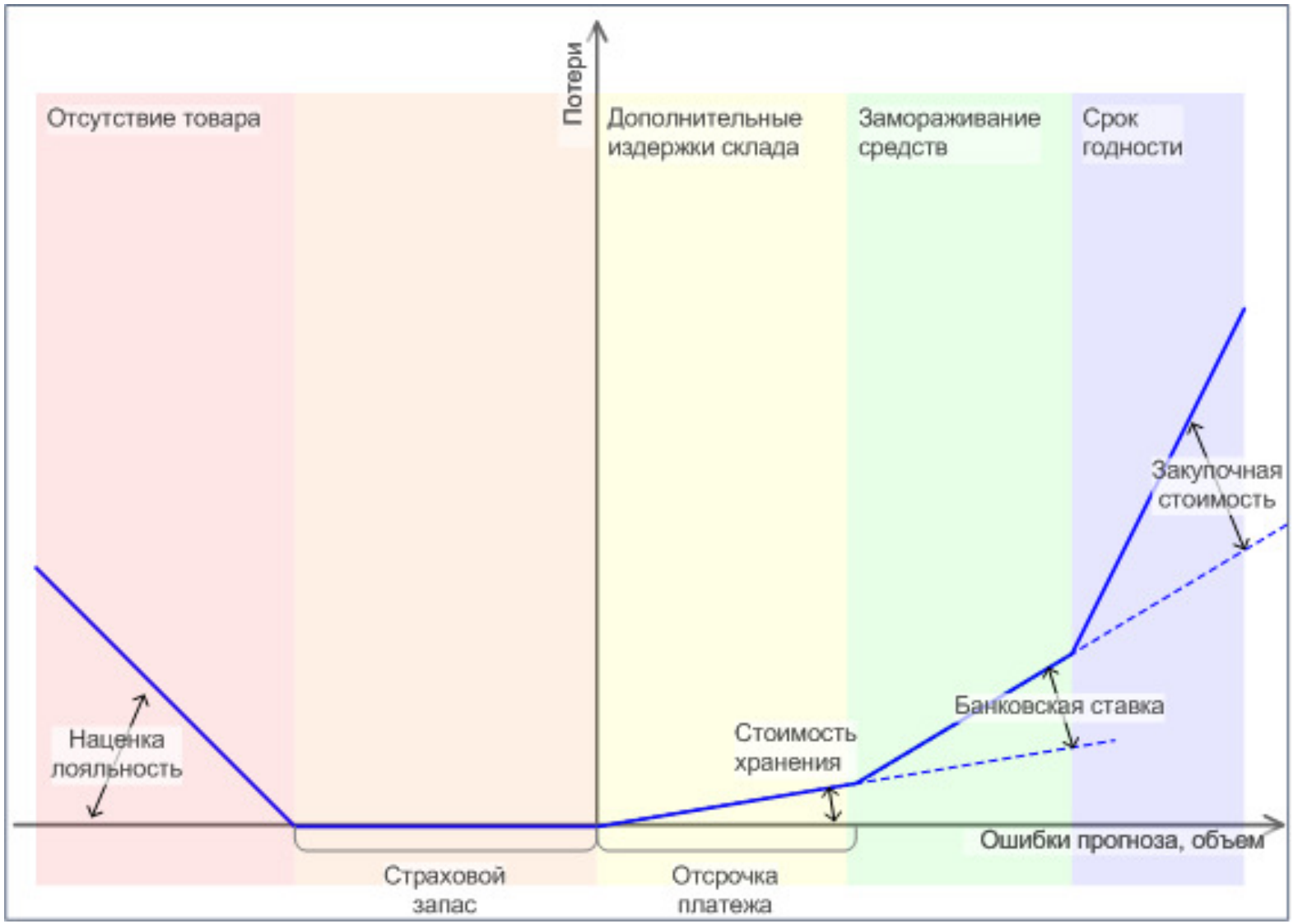
Линейная модель регрессии: $a(x_i) = \langle x_i, w \rangle$.

Сведение к задаче линейного программирования:

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

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

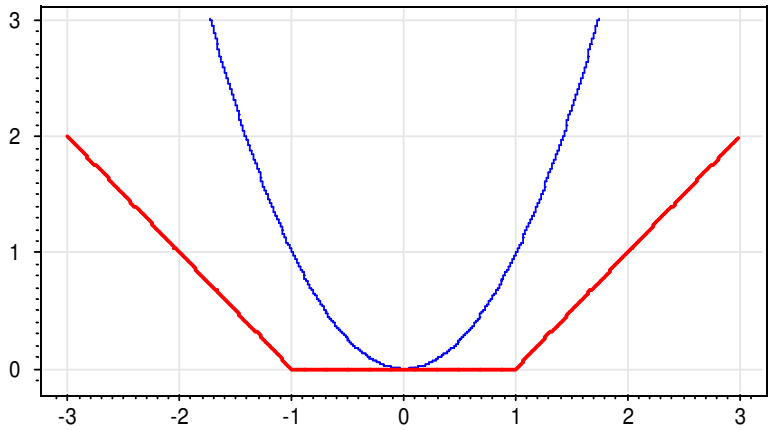
Пример. Задача прогнозирования объёмов продаж



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 также решается путём замены переменных и сведения к задаче математического (квадратичного) программирования

Связь МНК с методом максимума правдоподобия

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:

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

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

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

Замечательные свойства экспоненциального семейства:

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

Примеры распределений из экспоненциального семейства

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

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

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

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

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

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

Примеры распределений из экспоненциального семейства

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

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

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

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

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

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

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:

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

Метод Ньютона-Рафсона:

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

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

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

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

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

Матричные обозначения

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

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

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

$W^t = \operatorname{diag}\big(\frac{1}{\varphi_i} c''(\theta_i)\big)$ — диагональная матрица.

Тогда метод Ньютона-Рафсона приводит к МНК
с итеративным перевзвешиванием объектов
(IRLS, Iteratively Reweighted Least Squares):

$$\alpha^{t+1} := \alpha^t - h_t \big(F^{\mathsf{T}} W^t F\big)^{-1} F^{\mathsf{T}} W^t \operatorname{diag}\big(\tfrac{1}{c''(\theta_i)}\big)(y - \mu^t).$$

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

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

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):

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

Резюме в конце лекции

- Метод наименьших квадратов
  — нормальный некоррелированный шум
- Многомерная линейная регрессия
  — через сингулярное разложение
- Гребневая регрессия
  — тоже через сингулярное разложение
- Нелинейная регрессия
  — сводится к последовательности линейных регрессий
  — используется метод Ньютона-Рафсона
- Логистическая регрессия
  — не регрессия, а классификация
  — используется метод Ньютона-Рафсона
- Обобщённая линейная регрессия
  — обобщает обычную и логистическую регрессию
  — используется метод Ньютона-Рафсона
- Неквадратичные функции потерь
  — проблемно-ориентированные (зависят от задачи)
  — приводят к разным методам, отличным от МНК