ГУАП

Основы машинного обучения

Поляк Марк Дмитриевич

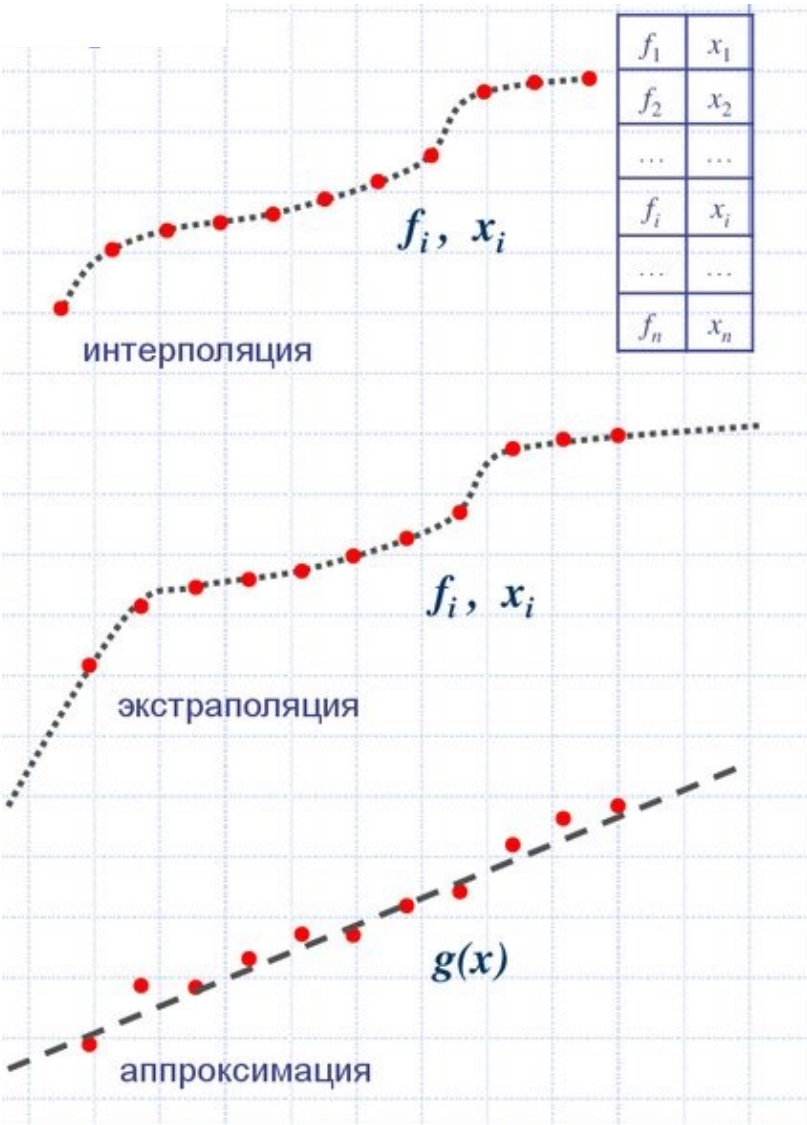
2025

Регрессионный анализ

Лекция 3

Интерполяция, экстраполяция, аппроксимация

- Интерполяция – определение промежуточных значений функции по известному дискретному набору значений
- Экстраполяция – определение значений функции за пределами первоначально известного интервала
- Аппроксимация – определение в явном виде параметров функции, описывающей распределение точек



Пример Рунге. Аппроксимация функции полиномом

- Функция $y = \frac{1}{1+25x^2}$ на отрезке $x \in [-2, 2]$.
- Признаковое описание объекта $x := (1, x^1, x^2, \dots, x^n)$
- Модель полиномиальной регрессии (полином степени *n*):

$$g(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x + \cdots + \theta_n x^n$$

- Обучение с помощью МНК:

$$Q(\boldsymbol{\theta}, X^\ell) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \cdots + \theta_n x_i^n - y_i)^2 \to \min_{\theta_0, \dots, \theta_n}$$

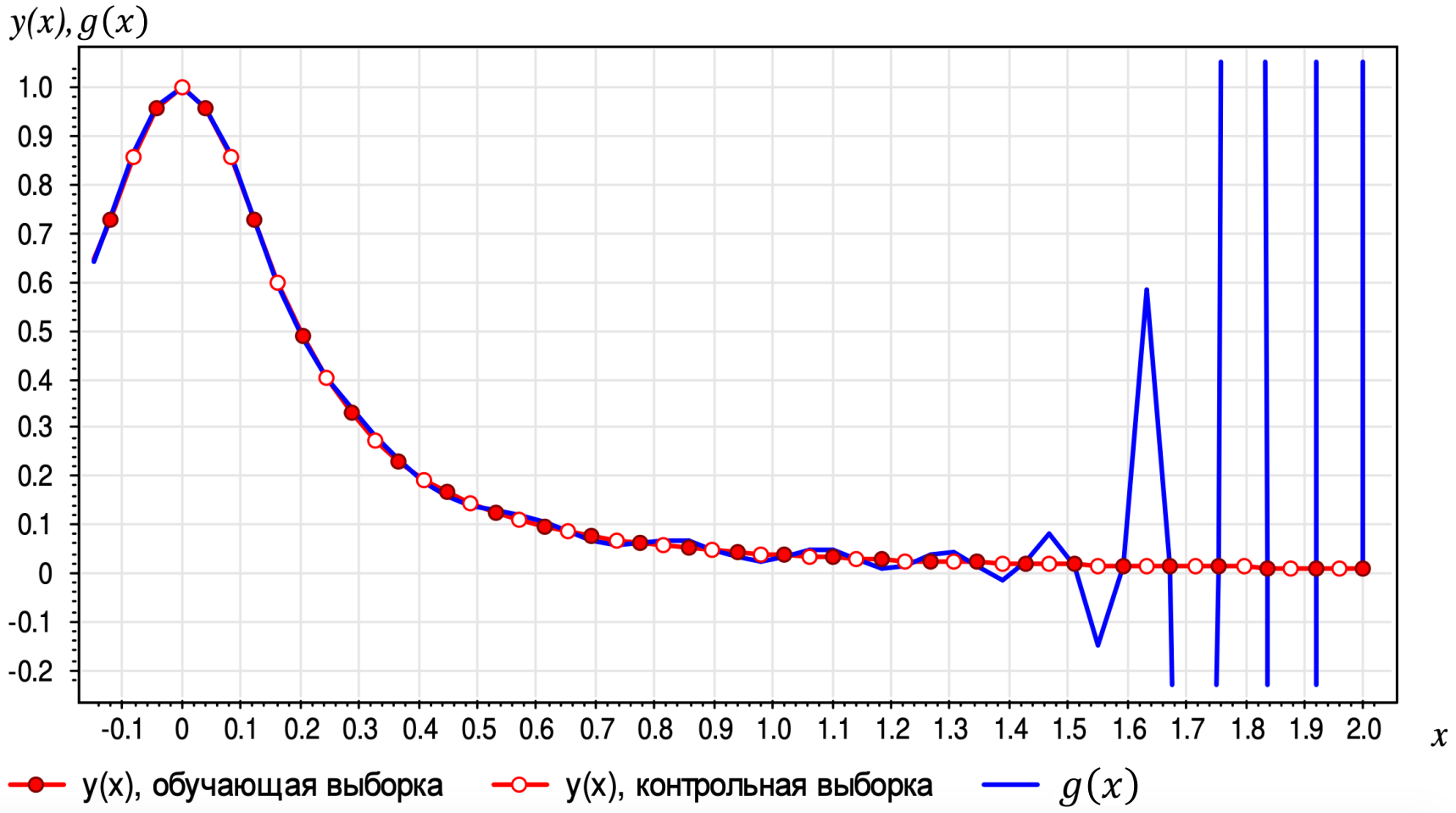
- Обучающая выборка: $X^\ell = \left\{ x_i = 4\frac{i-1}{\ell-1} - 2 \,\middle|\, i = 1, \dots, \ell \right\}$
- Контрольная выборка: $X^k = \left\{ x_i = 4\frac{i-0.5}{\ell-1} - 2 \,\middle|\, i = 1, \dots, \ell - 1 \right\}$

Пример Рунге. Переобучение

$$y = \frac{1}{1 + 25x^2}$$

$g(x)$ – полином степени $n = 38$

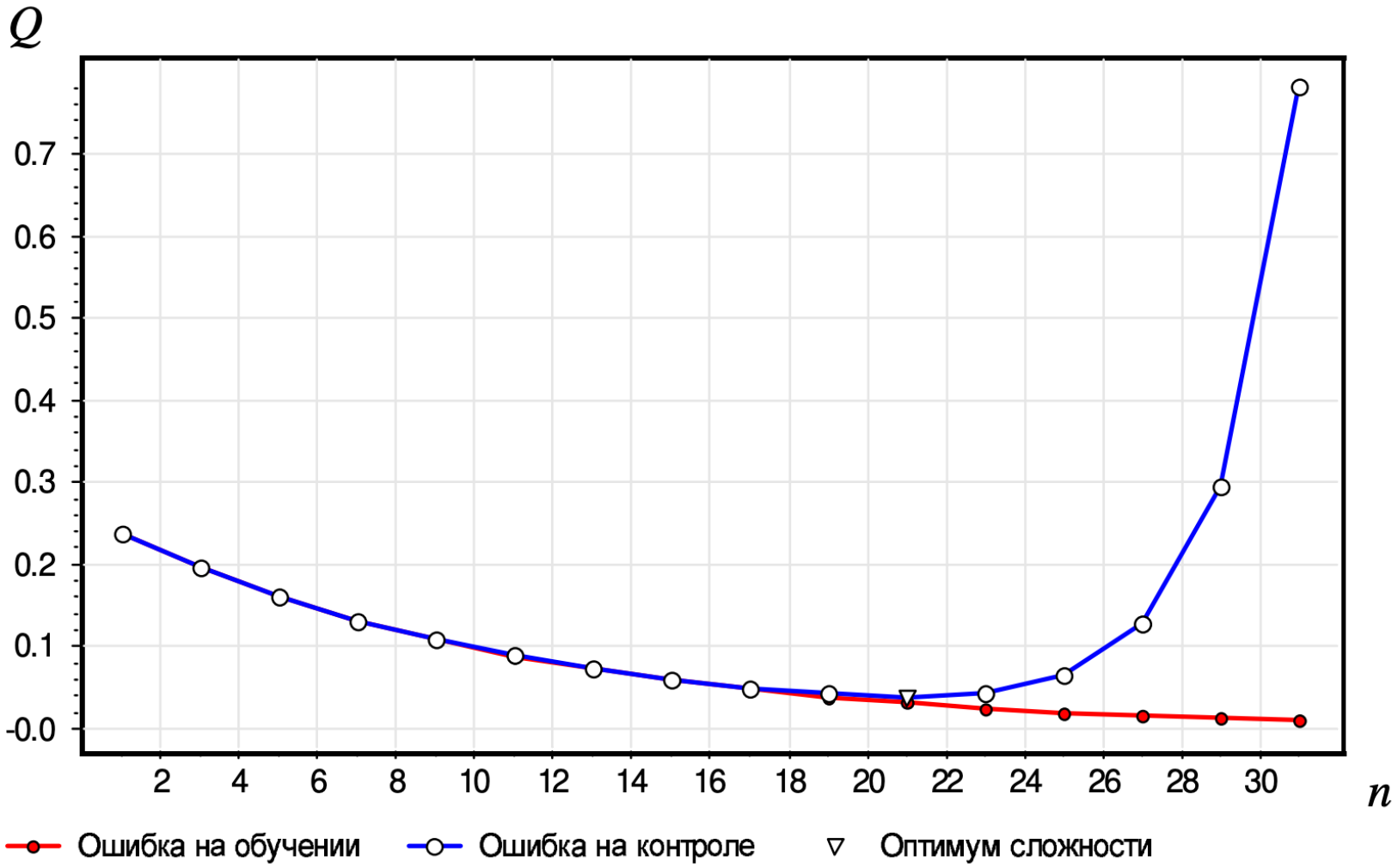
$\ell = 50$



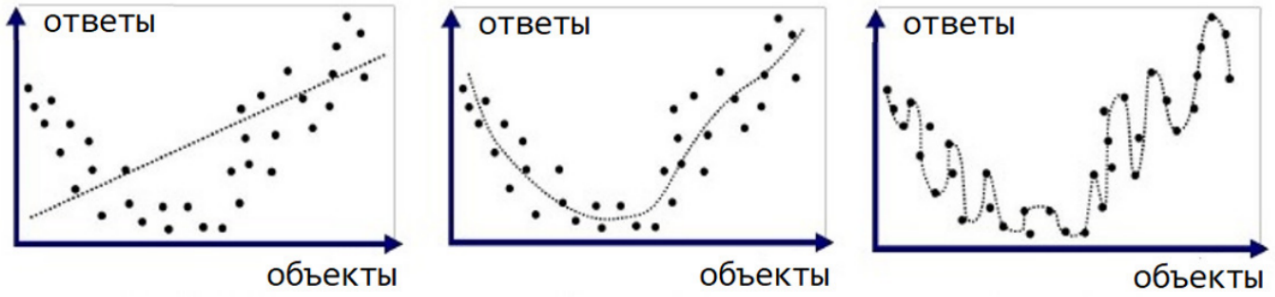
https://ru.wikipedia.org/wiki/Феномен_Рунге

Пример Рунге. Зависимость Q от степени полинома *n*

Переобучение — это когда $Q\big(\mu(X^\ell), X^k\big) \gg Q\big(\mu(X^\ell), X^\ell\big)$:

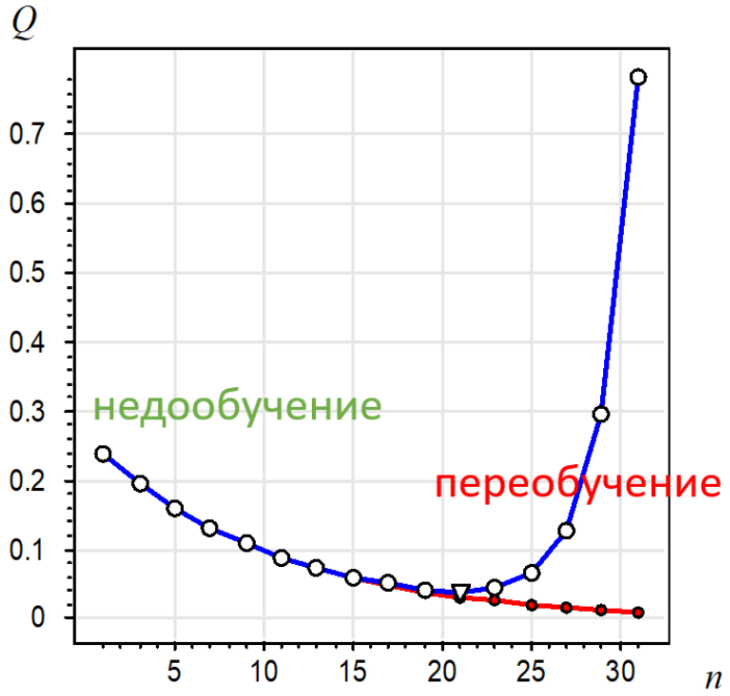


Проблема недообучения и переобучения



недообучение переобучение

- Недообучение (underfitting): данных много, параметров недостаточно, модель простая, негибкая
- Переобучение (overfitting): данных мало, параметров слишком много, модель сложная, избыточно гибкая



Переобучение – ключевая проблема в машинном обучении

- Из-за чего возникает переобучение?
  - Избыточные параметры в модели $g(x, \theta)$ «расходуются» на чрезмерно тонкую подгонку под обучающую выборку.
  - Выбор g из A производится по неполной информации X^{ℓ}
- Как обнаружить переобучение?
  - Эмпирически, путем разбиения выборки на train и test (для test должны быть известны правильные ответы)
- Избавиться от переобучения нельзя. Как его минимизировать?
  - Увеличивать объем обучающих данных (big data).
  - Накладывать ограничения на θ (регуляризация).
  - Минимизировать одну из теоретических оценок.
  - Выбирать модель по оценкам обобщающей способности.

Эмпирические оценки обобщающей способности

- Эмпирический риск на тестовых данных (hold-out):

$$\mathrm{HO}\left(\mu, X^{\ell}, X^{k}\right) = Q\left(\mu\left(X^{\ell}\right), X^{k}\right) \to \min$$

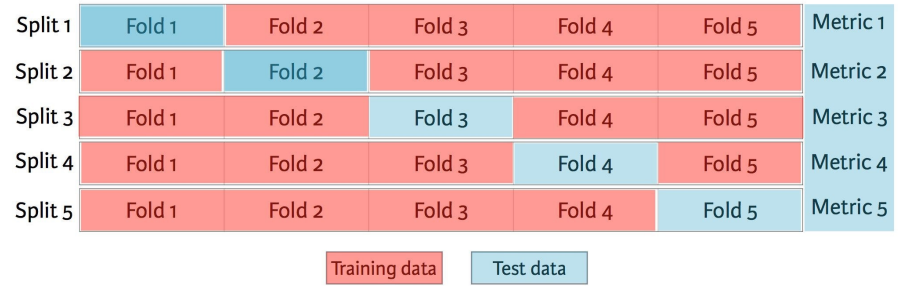
- Скользящий контроль (leave-one-out), $L = \ell + 1$:

$$\mathrm{LOO}(\mu, X^{L}) = \frac{1}{L}\sum_{i=1}^{L} \mathcal{L}(\mu(X^{L}\backslash\{x_i\}), x_i\ ) \to \min$$

- Кросс-проверка, кросс-валидация (cross-validation), $L = \ell + k$:

$$\mathrm{CV}(\mu, X^{L}) = \frac{1}{|P|}\sum_{p\in P} Q\left(\mu\left(X_p^{\ell}\right), X_p^{k}\right) \to \min$$

где P – множество разбиений $X^{L} = X_p^{\ell} \sqcup X_p^{k}$



Многомерная линейная регрессия

- 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;
  $y_i = \varphi(x_i),\ \varphi: X \to Y$ – неизвестная зависимость.
- $a(x) = g(x, \theta)$ – модель зависимости,
  $\theta \in \mathbb{R}^p$ – вектор параметров модели.
- Метод наименьших квадратов (МНК):

$$Q\left(\theta, X^\ell\right) = \sum_{i=1}^{\ell} (g(x_i, \theta) - y_i)^2 \to \min_\theta$$

Многомерная линейная регрессия

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

$$g(x, \boldsymbol{\theta}) = \sum_{j=1}^{n} \theta_j f_j(x), \boldsymbol{\theta} \in \mathbb{R}^{\mathrm{n}}$$

- Матричные обозначения:

$$\boldsymbol{F}_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_\ell) & \cdots & f_n(x_\ell) \end{pmatrix}, \qquad \boldsymbol{y}_{\ell \times 1} = \begin{pmatrix} y_1 \\ \vdots \\ y_\ell \end{pmatrix}, \qquad \boldsymbol{\theta}_{n \times 1} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$$

- Функционал квадрата ошибки:

$$Q\left(\theta, X^\ell\right) = \sum_{i=1}^{\ell} (g(x_i, \theta) - y_i)^2 = \|\boldsymbol{F\theta} - \boldsymbol{y}\|^2 \to \min_{\theta}$$

Нормальная система уравнений

- Необходимое условие минимума в матричном виде:

$$\frac{\partial Q(\theta)}{\partial \theta} = 2\boldsymbol{F}^T(\boldsymbol{F\theta} - \boldsymbol{y}) = 0$$

откуда следует нормальная система задачи МНК:

$$\boldsymbol{F}^T\boldsymbol{F\theta} = \boldsymbol{Fy}$$

где $\boldsymbol{F}^T\boldsymbol{F}$ – матрица размера $n \times n$.

- Решение системы: $\boldsymbol{\theta}^* = (\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{y} = \boldsymbol{F}^+\boldsymbol{y}$

Значение функционала: $Q(\boldsymbol{\theta}^*) = \|\boldsymbol{P_F y} - \boldsymbol{y}\|^2$,

где $\boldsymbol{P_F} = \boldsymbol{FF}^+ = (\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T$ – проекционная матрица.

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 $\ell \times \ell$-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 $n \times n$-матрицы $F^{\mathsf{T}}F$;
3. $n \times n$-матрица D диагональна, $D = \operatorname{diag}\bigl(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\bigr)$, $\lambda_j \geqslant 0$ — общие собственные значения матриц $F^{\mathsf{T}}F$ и FF^{T}.

Гребневая регрессия (ridge regression)

Штраф за увеличение L_2-нормы вектора весов $||\boldsymbol{\theta}||$:

$$Q_\tau(\boldsymbol{\theta}) = \|F\boldsymbol{\theta} - y\|^2 + \frac{\tau}{2}\|\boldsymbol{\theta}\|^2,$$

где τ – неотрицательный *параметр регуляризации*.

Модифицированное МНК-решение (τI_n – «гребень», ridge):

$$\frac{\partial Q_\tau(\boldsymbol{\theta})}{\partial \theta} = 2F^T(F\boldsymbol{\theta} - y) + 2\tau\theta = 0$$

$$\theta_\tau^* = (F^T F + \tau I_n)^{-1} F^T y.$$

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

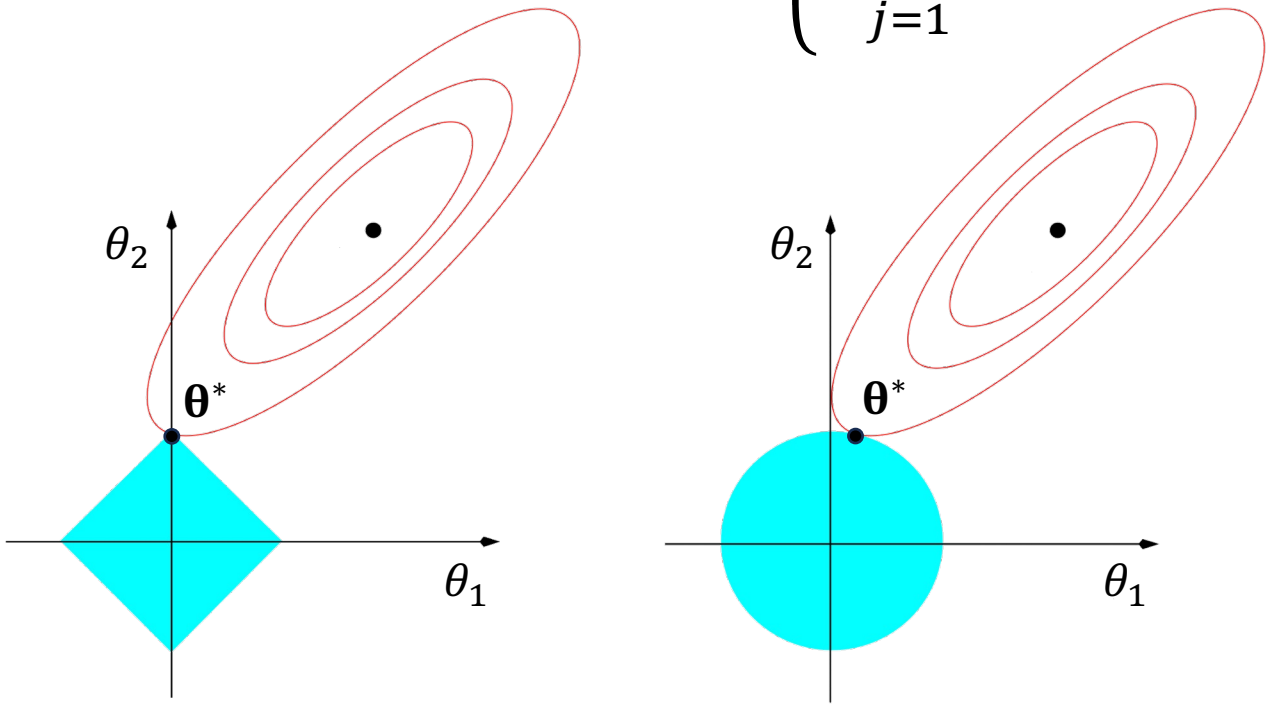
Регуляризация по L_1-норме для отбора признаков

LASSO – Least Absolute Shrinkage and Selection Operator

$$\|F\boldsymbol{\theta} - y\|^2 + \mu \sum_{j=1}^{n} |\theta_j| \to \min_{\boldsymbol{\theta}} \quad \Leftrightarrow \quad \begin{cases} \|F\boldsymbol{\theta} - y\|^2 \to \min\limits_{\boldsymbol{\theta}}; \\ \mu \sum\limits_{j=1}^{n} |\theta_j| \le t; \end{cases}$$

LASSO (L_1):

$$\sum_{j=1}^{n} |\theta_j| \le t$$



Ridge (L_2):

$$\sum_{j=1}^{n} \theta_j^2 \le t$$

T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning, 2017

Конструирование признаков

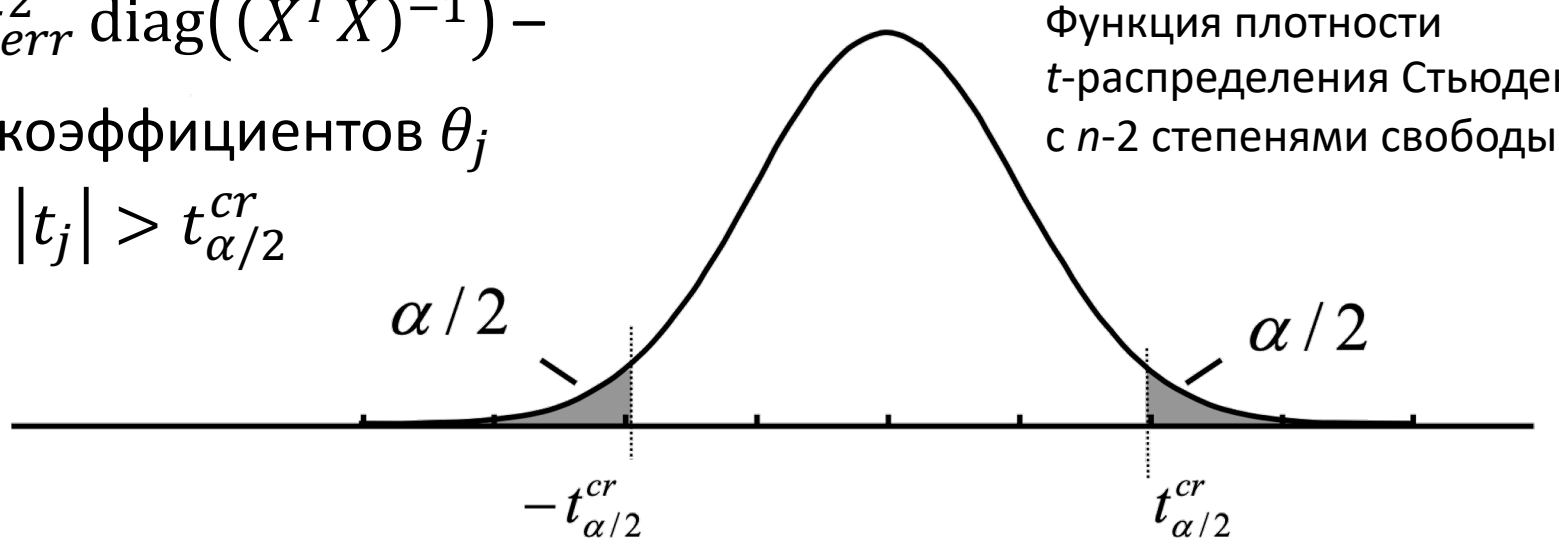
- Использование интуиции для создания новых признаков путем преобразования или комбинирования оригинальных признаков.
  - Пример: предсказание стоимости жилья.
    Признаки: x_1 – площадь квартиры (м. кв.), x_2 – город (категориальный). Модель:
    $$g_1(x, \boldsymbol{\theta}) = \theta_2 x_2 + \theta_1 x_1 + \theta_0$$
    Добавляем новый признак: $x_3 = x_1 x_2$, чтобы напрямую учесть в модели различия стоимости кв. метра в разных регионах. Новая модель:
    $$g_2(x, \boldsymbol{\theta}) = \theta_3 x_3 + \theta_2 x_2 + \theta_1 x_1 + \theta_0$$

- Способы конструирования признаков
  - Полиномиальные признаки: возведение существующих признаков в степень или их комбинирование
  - Агрегация данных: среднее, сумма или медиана по имеющимся признакам
  - Лаги – значения за предыдущие периоды, которые могут влиять на текущие
  - Временны*е* признаки – день недели, месяц или номер квартала, которые учитывают сезонные (периодические) изменения в данных
  - Знания из предметной области

Информативность признаков

- Проверка значимости коэффициентов уравнения регрессии
- Статистический тест (t-критерий Стьюдента): коэффициент является значимым, если он отличен от нуля.
  - $H_0: \boldsymbol{\theta}^2 = 0$ – нулевая гипотеза (ответ не зависит от признаков объекта)
  - $H_1: \boldsymbol{\theta}^2 \neq 0$ – альтернативная гипотеза

  Для каждого признака j и коэффициента θ_j:
  - $t_j = \frac{\theta_j}{\sigma_{\theta_j}}$, где $\sigma_{\boldsymbol{\theta}} = \sqrt{\sigma_{err}^2 \operatorname{diag}\left((X^T X)^{-1}\right)}$ – вектор СКО коэффициентов θ_j
  - H_0 отвергается, если $\left|t_j\right| > t_{\alpha/2}^{cr}$



Функция плотности t-распределения Стьюдента с n-2 степенями свободы