

Машинное обучение

Лекция 2
Линейная регрессия

Михаил Гущин
mhushchyn@hse.ru

НИУ ВШЭ, 2024



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

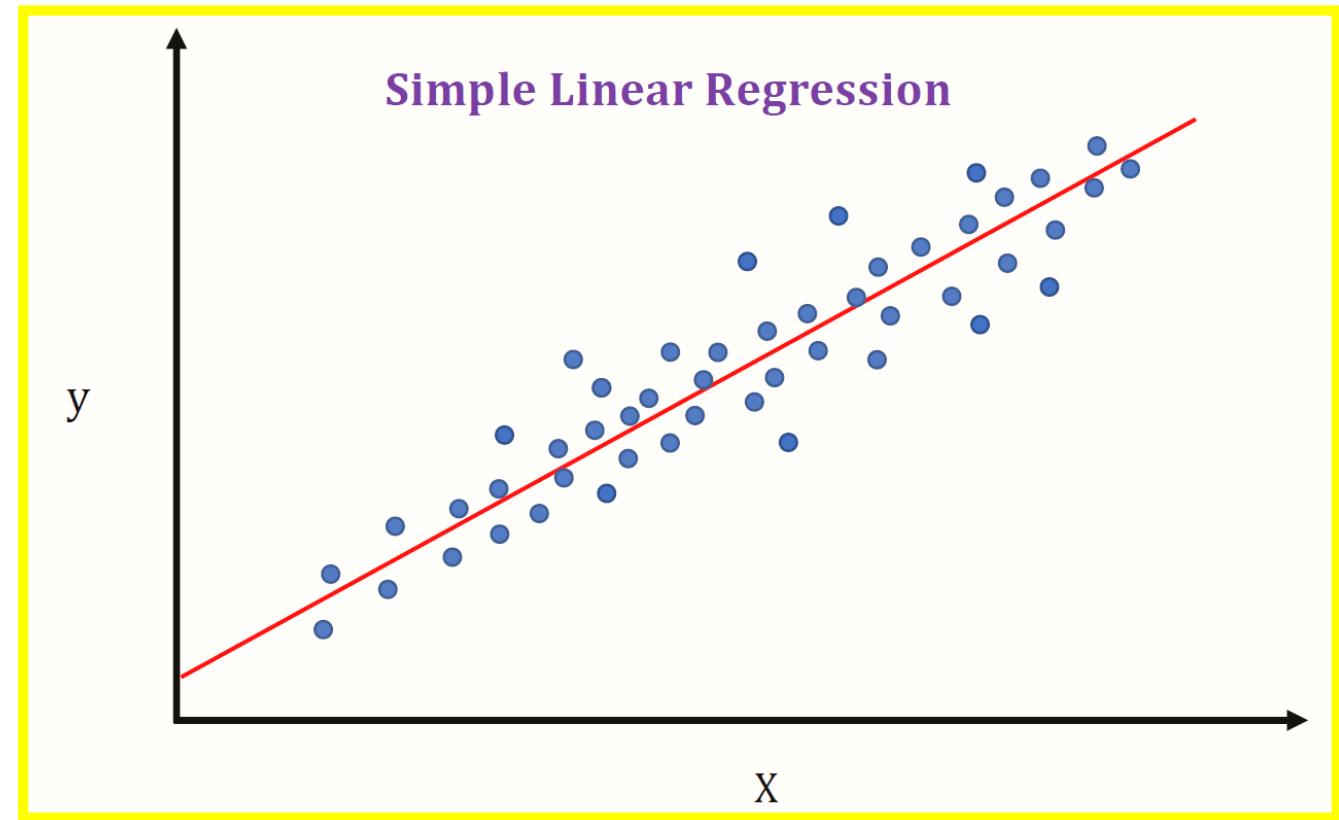
**Обеспокоенные родители:
А если бы все твои
друзья пошли с моста
прыгать, ты бы тоже
прыгнул?**

**Алгоритм машинного обучения (kNN)
Да.**

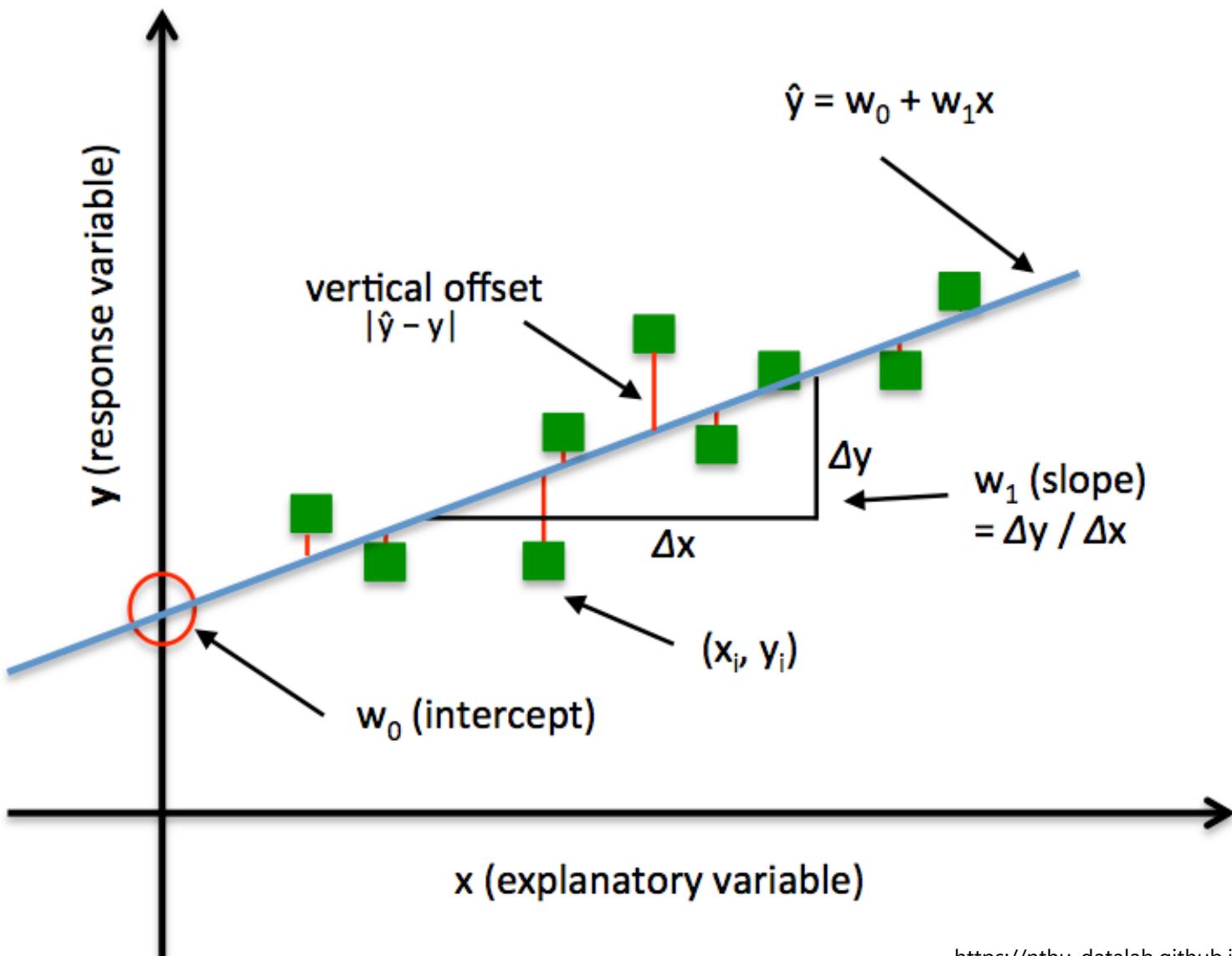
Линейная регрессия. Аналитическое решение.

Задача регрессии

- ▶ Есть объекты (X)
- ▶ Нужно предсказать некоторую величину (y)
- ▶ Функция, которая описывает зависимость y от X - **модель регрессии**



Линейная регрессия



<https://nchu-datalab.github.io>

Векторная форма

- ▶ Пусть дан набор из n точек: $\{x_i, y_i\}_{i=1}^n$, где
 - $x_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$ - вектор из d признаков объекта;
 - y_i - скалярная величина, которую хотим предсказать для объекта.

- ▶ Модель линейной регрессии:

$$\hat{y}_i = w_0 + \sum_{j=1}^d w_j x_{ij}$$

- w_j - веса модели;
 - \hat{y}_i - прогноз для объекта;
- ▶ Квадрат ошибки прогноза модели для объекта: $(\hat{y}_i - y_i)^2$

Матричная форма

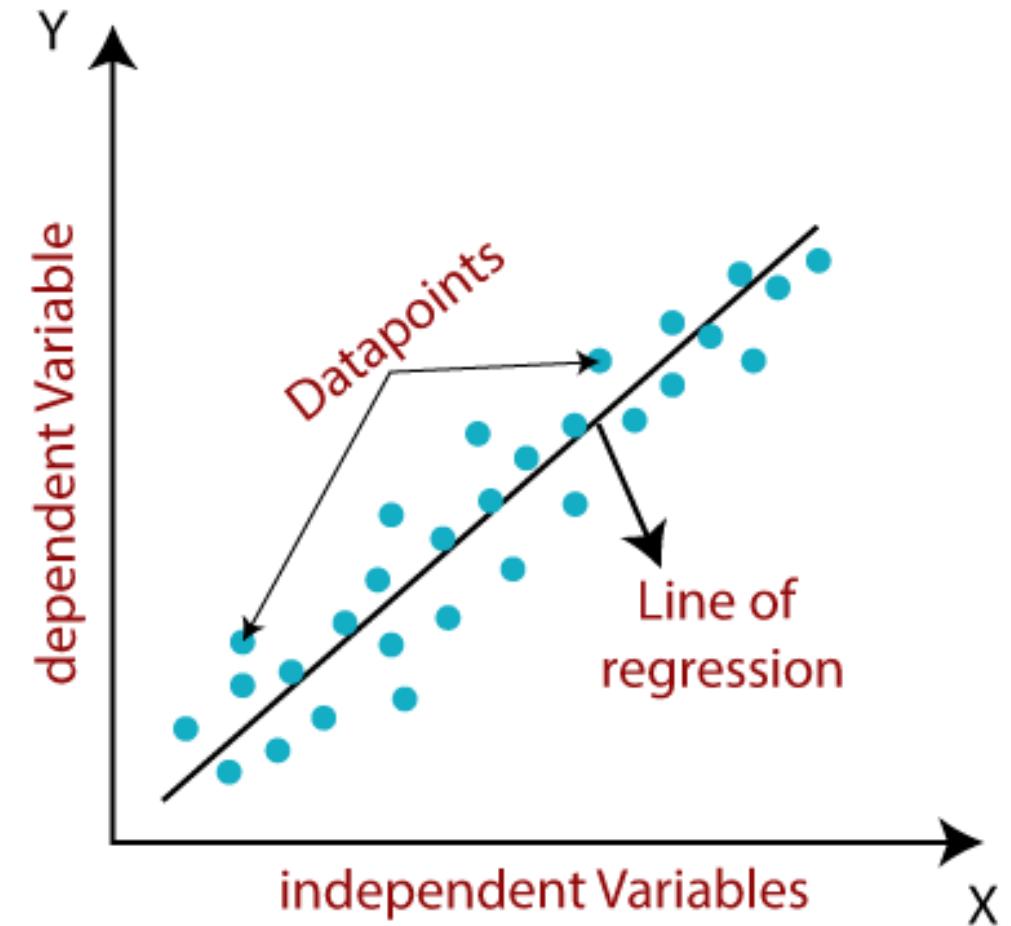
- Модель линейной регрессии:

$$\hat{y} = Xw$$

- $X = \begin{pmatrix} \mathbf{1} & x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & & \vdots \\ \mathbf{1} & x_{n1} & \cdots & x_{nd} \end{pmatrix}$ - матрица признаков объектов $(n, d + 1)$;
 - $w = (w_0, w_1, \dots, w_d)^T$ - вектор $(d + 1)$ весов модели;
 - $\hat{y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)^T$ - вектор прогнозов модели для (n) объектов;
-
- Вектор квадратов ошибок прогнозов модели: $(\hat{y} - y)^2$

Задача

- ▶ Хотим, чтобы средняя квадратичная ошибка прогнозов $(\hat{y} - y)^2$ была минимальной
- ▶ **Как найти** оптимальные веса w модели?



Решение

- ▶ **Функция потерь (Loss function)** (скалярная и векторная формы):

$$L = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \frac{1}{n} (\hat{\mathbf{y}} - \mathbf{y})^T (\hat{\mathbf{y}} - \mathbf{y})$$

- ▶ Значение L – среднеквадратичная ошибка (**Mean Squared Error (MSE)**)
- ▶ Мы хотим минимизировать L :

$$L \rightarrow \min_w$$

Аналитическое решение

$$L = (\hat{y} - y)^T (\hat{y} - y) = (\mathbf{X}\mathbf{w} - y)^T (\mathbf{X}\mathbf{w} - y)$$

Чтобы найти минимум L , надо:

$$\frac{\partial L}{\partial \mathbf{w}} = 0$$

Тогда

$$\frac{\partial L}{\partial \mathbf{w}} = 2\mathbf{X}^T(\mathbf{X}\mathbf{w} - y) = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^Ty = 0$$

Получаем оптимальные веса w линейной регрессии:

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

The background of the image is a dense, abstract pattern of numerous thin, curved lines in various colors, primarily shades of red, orange, and pink, set against a dark, almost black, background. These lines create a sense of depth and motion, resembling a complex network or a microscopic view of organic structures.

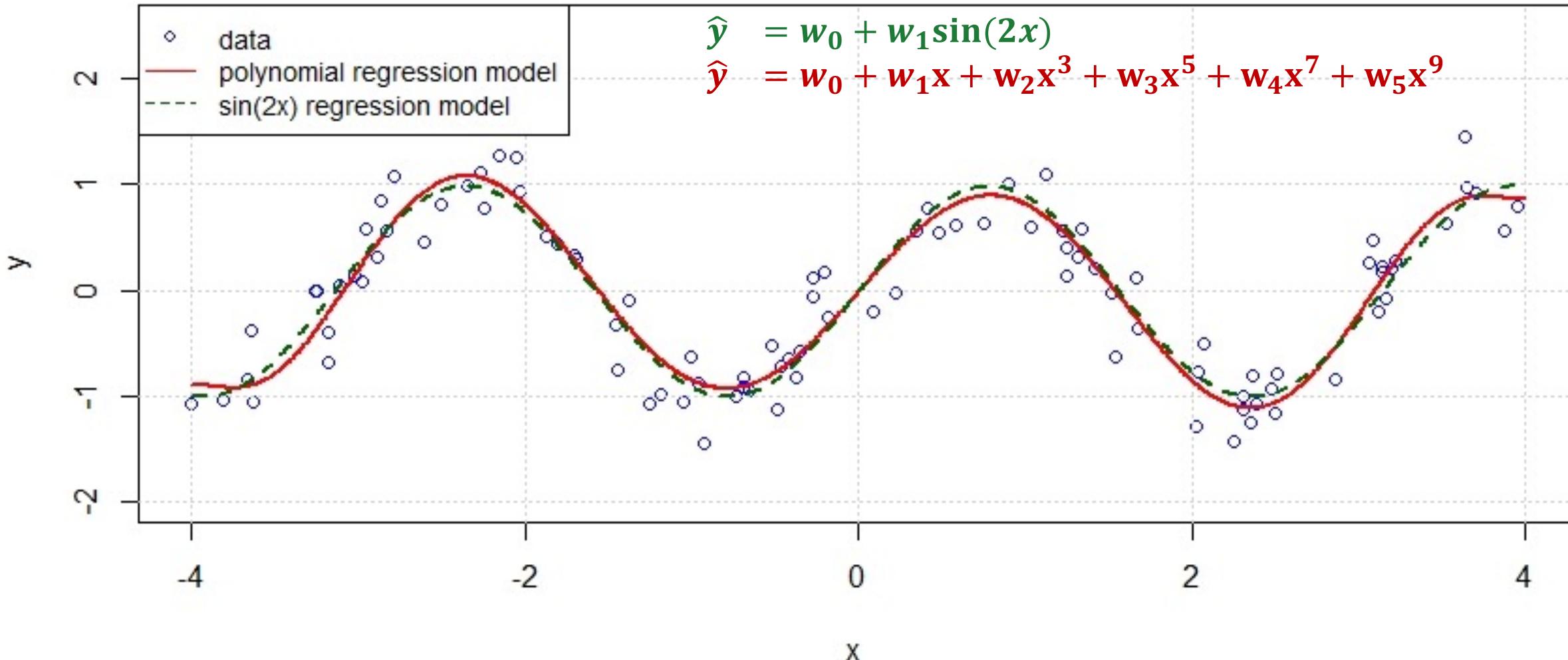
Нелинейные зависимости

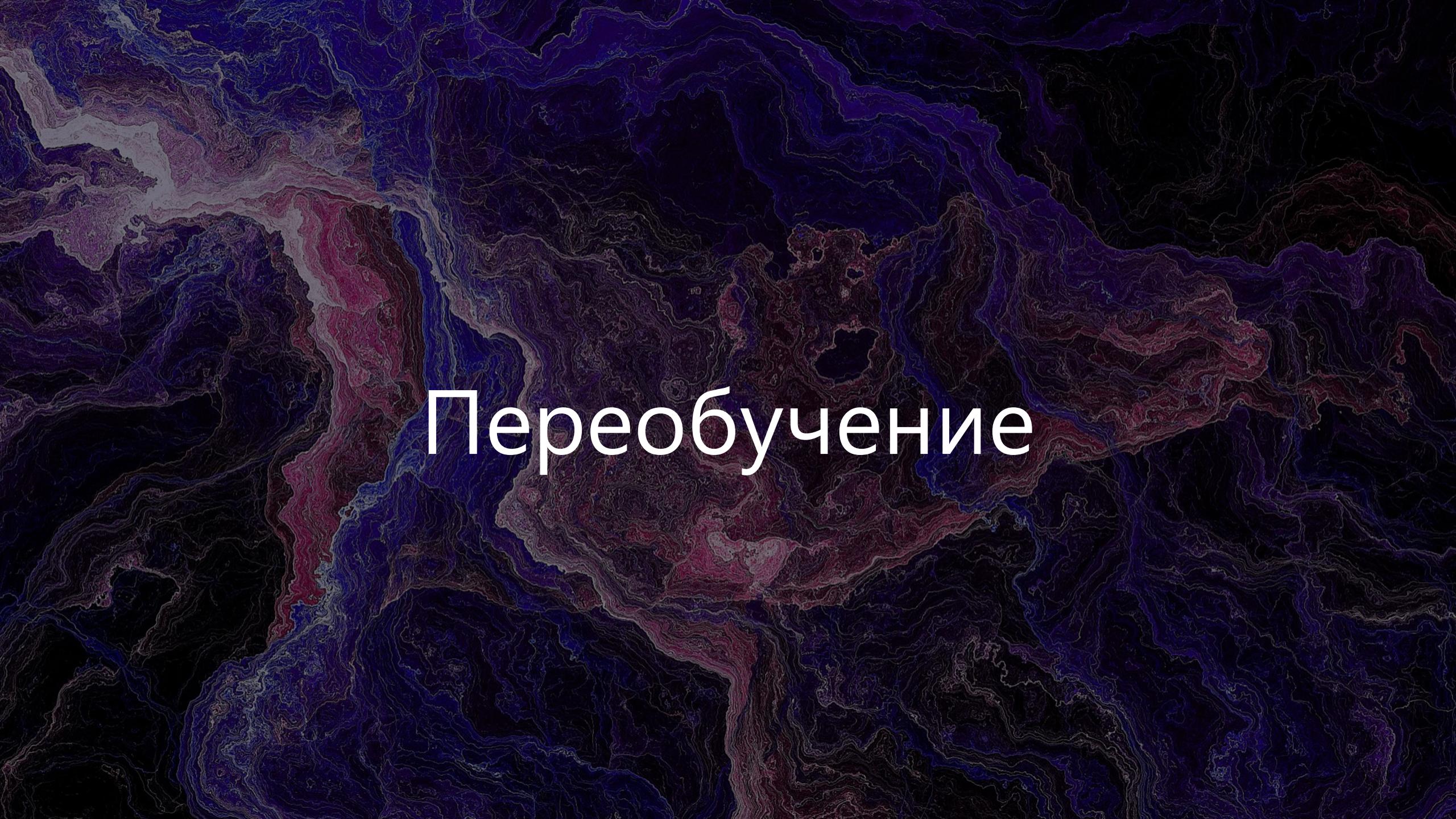
Нелинейные зависимости

- ▶ Пусть дан вектор признаков x_i
- ▶ Выберем любые нелинейные функции от x : $\phi_0(x), \phi_1(x), \dots, \phi_M(x)$
 - Например, $x^2, \sin(x), e^{-x}$
- ▶ Тогда модель линейной регрессии:

$$\hat{y}_i = \sum_{j=0}^{\textcolor{blue}{k}} w_j \phi_j(x_i)$$

Пример





Переобучение

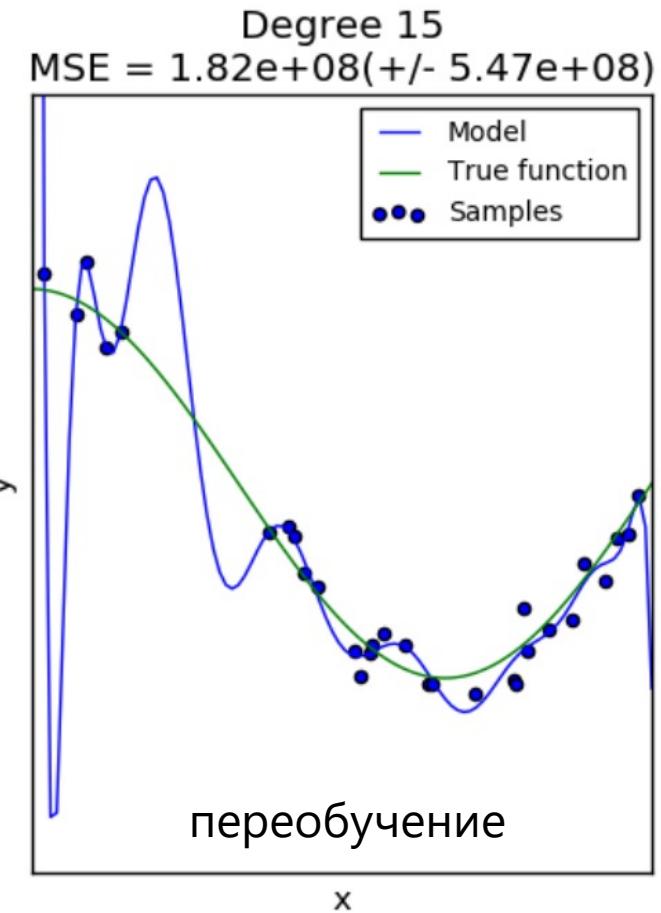
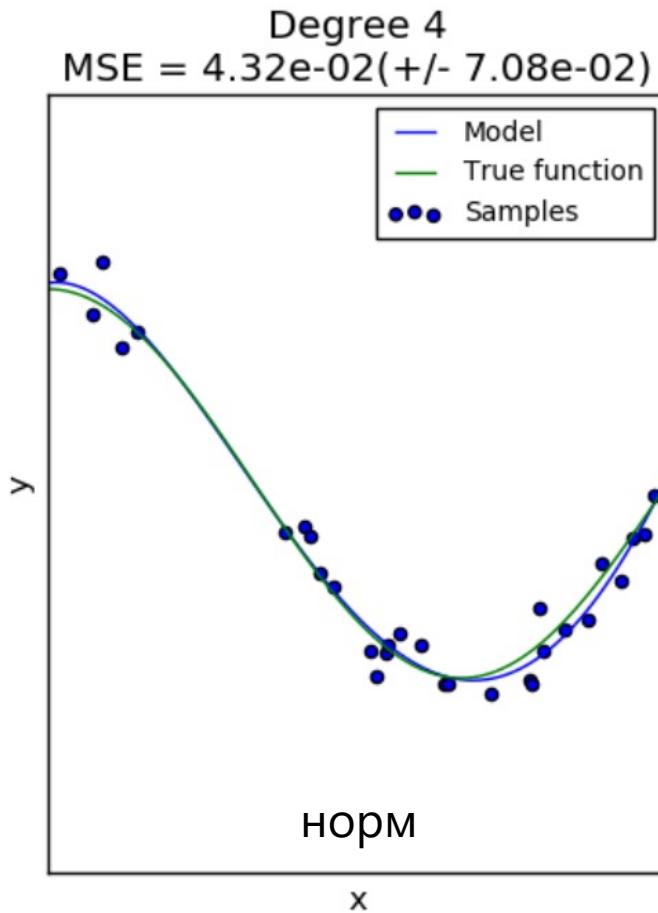
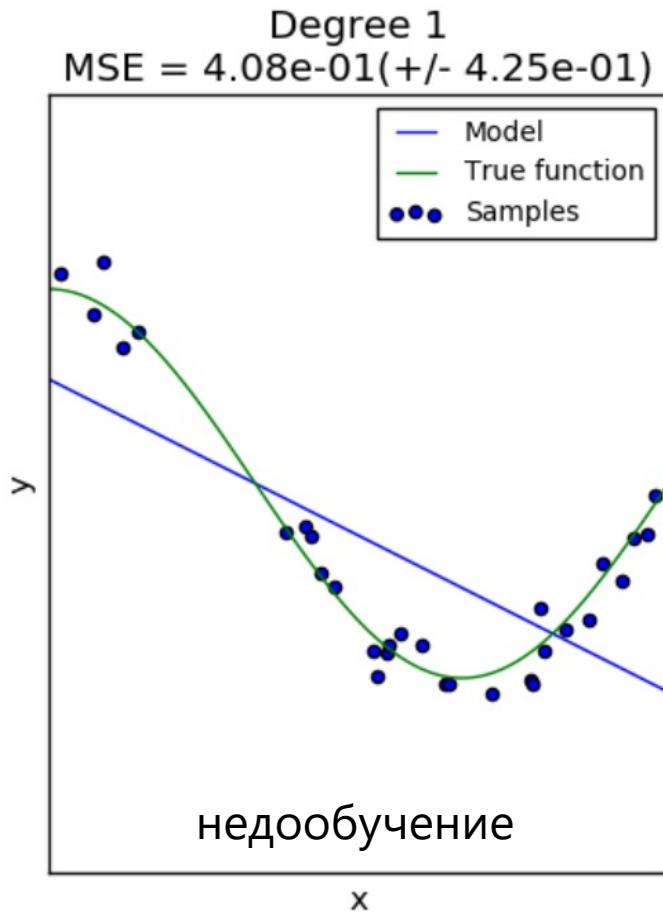
Задача

- ▶ Пусть дан набор из n точек: $\{x_i, y_i\}_{i=1}^n$, где $x_i \in \mathbb{R}^1$
- ▶ Для каждого x_i создадим дополнительные признаки:
 - $x_i, x_i^2, x_i^3, \dots, x_i^k$
- ▶ Рассмотрим модель полиномиальной линейной регрессии:

$$\hat{y}_i = w_0 + \sum_{j=1}^{\mathbf{k}} w_j x_i^j$$

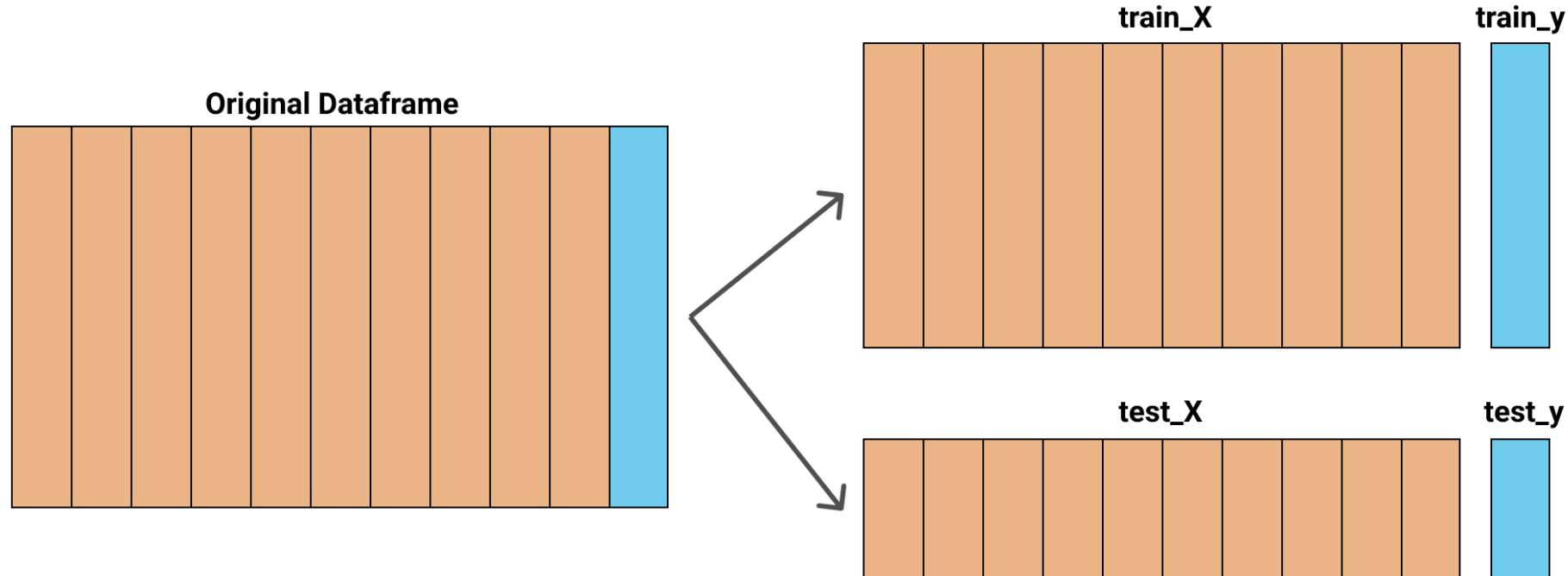
- ▶ Максимальную степень полинома \mathbf{k} будем менять от 1 до 15

Решение

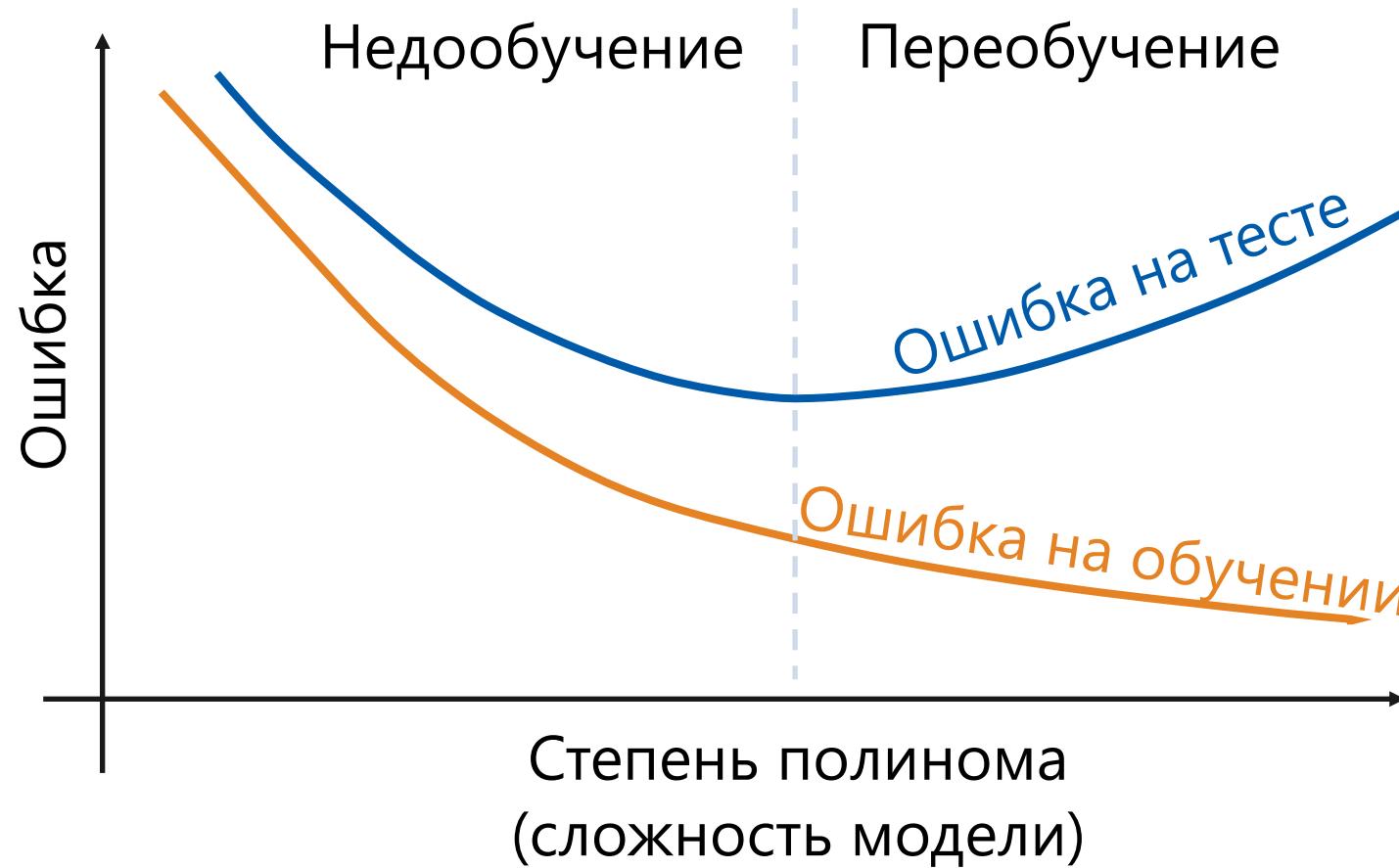


Обучение и тест

- ▶ **Обучающая выборка (train):** для обучения модели
- ▶ **Тестовая (отложенная) выборка (test):** для измерения качества модели



Переобучение



Регуляризация

Проблема переобучения

- ▶ Модель линейной регрессии:

$$\hat{y}_i = w_0 + \sum_{j=1}^d w_j x_{ij}$$

- ▶ Ошибка прогноза модели для объекта: $|\hat{y}_i - y_i|$
- ▶ Пусть значение некоторых весов очень большие по модулю, например $|w_k| > 10^3$
- ▶ Тогда малые изменения dx_{ik} приводят к очень большим изменениям $|d\hat{y}_i| = |w_k dx_{ik}|$

Регуляризация

- ▶ Давайте добавим к функции потерь $L(w)$ **штраф** $R(w)$ на **величину весов** модели:

$$L_\alpha(w) = L(w) + \alpha R(w)$$

- α – коэффициент регуляризации (подбираем сами)
- ▶ Регуляризация не позволяет весам модели принимать слишком большие значения

Виды регуляризации

- ▶ L_1 регуляризация (Lasso):

$$R_1(w) = \sum_{j=1}^d |w_j|$$

- ▶ L_2 регуляризация (Ridge):

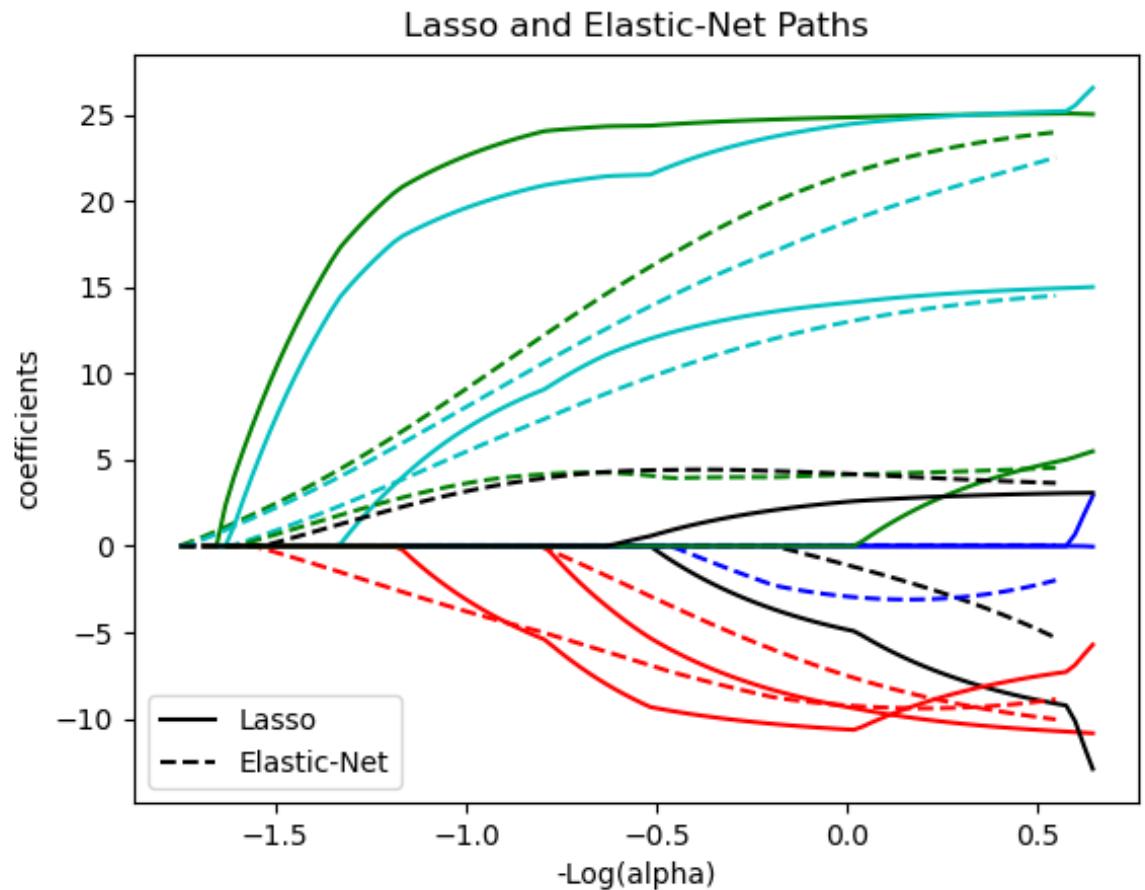
$$R_2(w) = \sum_{j=1}^d w_j^2$$

- ▶ $L_1 + L_2$ регуляризация (Elastic Net):

$$L_\alpha(w) = L(w) + \alpha_1 R_1(w) + \alpha_2 R_2(w)$$

Свойства регуляризации

- ▶ L_2 регуляризация стремится уменьшить веса модели
- ▶ L_1 позволяет проводить **отбор признаков**
- ▶ **L_1 обнуляет веса** для наименее информативных признаков



<https://scikit-learn.org>

Гиперпараметры и параметры

- ▶ Рассмотрим пример функции потерь с регуляризацией:

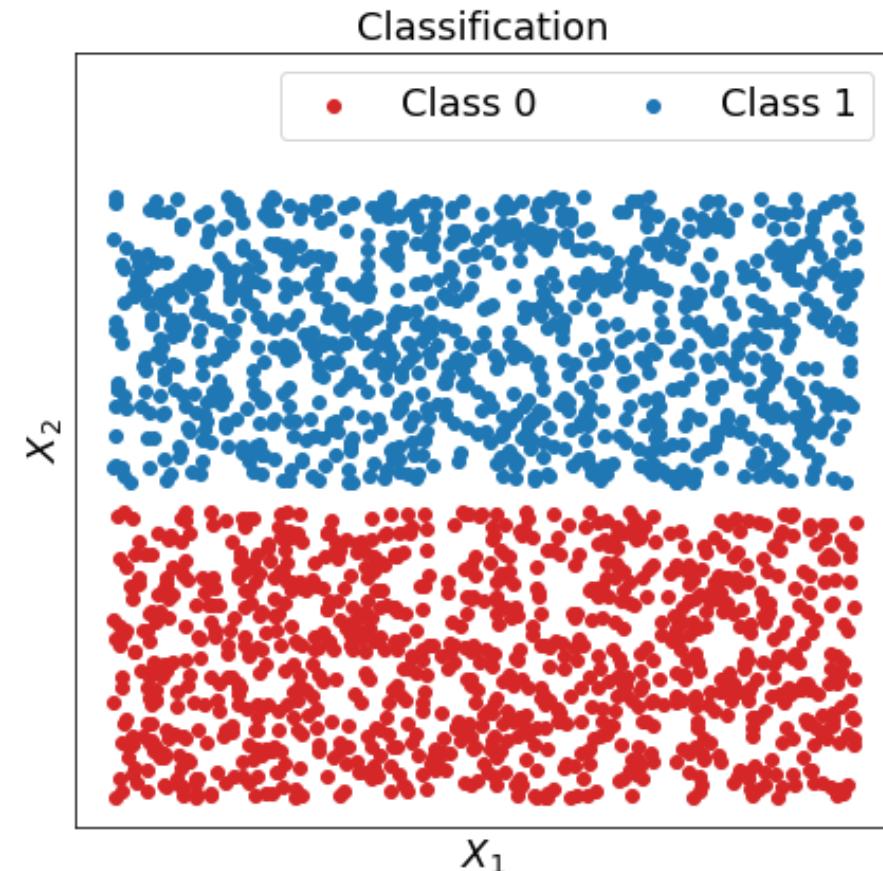
$$L_\alpha(w) = L(w) + \alpha R(w)$$

- ▶ Здесь w – веса нашей модели. Их будем называть **параметрами** модели. Они **определяются в процессе обучения**.
- ▶ α – коэффициент регуляризации. Его **значение задаем мы сами**. Такие параметры будем называть **гиперпараметрами**.

Важность признаков (Feature importance)

Интуиция

- ▶ Не все признаки одинаково полезны для решения задачи
- ▶ Некоторые из них более информативны, чем другие
- ▶ Например, X_1 неинформативна для классификации
- ▶ **Цель** – определить **важность каждого признака**



Линейные модели

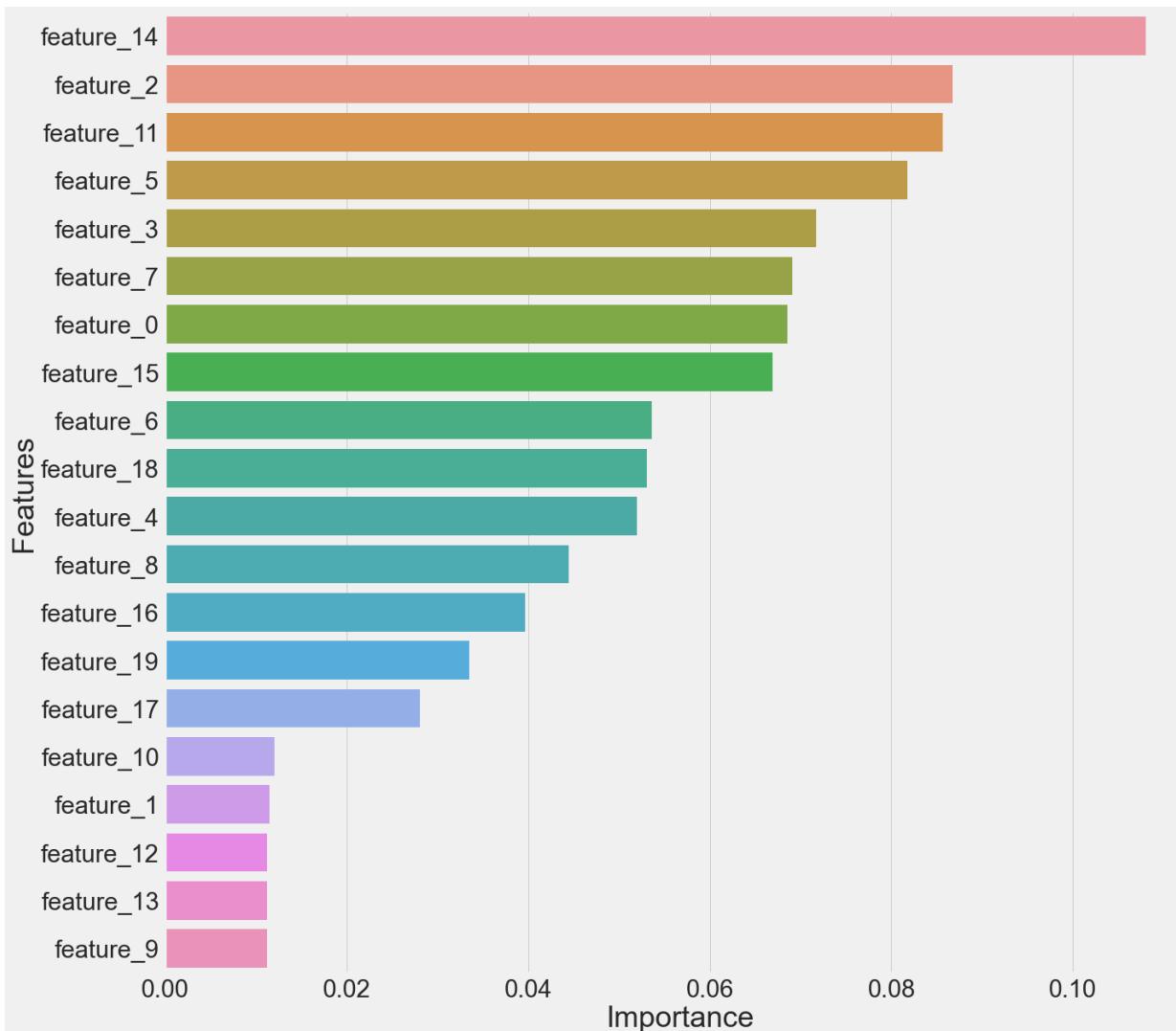
Рассмотрим линейную модель с регуляризацией (L_1 или L_2):

$$\hat{y} = w_0 + w_1 f_1 + w_2 f_2 + \cdots + w_k f_k$$

Если признаки нормированы (значения одного масштаба), то
важность признака f_i равна:

$$Imp(f_i) = |w_i|$$

Пример



Заключение



Вопросы

- ▶ Что такое объект, целевая переменная, признак, модель, функция потерь, функционал ошибки и обучение?
- ▶ Что такое переобучение и недообучение? Как отличить переобучение от недообучения?
- ▶ Что такое кросс-валидация и для чего она используется? Чем применение кросс-валидации лучше, чем разбиение выборки на обучение и контроль?
- ▶ Чем гиперпараметры отличаются от параметров?
- ▶ Запишите формулы для линейной модели регрессии и для среднеквадратичной ошибки. Запишите среднеквадратичную ошибку в матричном виде.
- ▶ Что такое регуляризация? Для чего ее используют в линейных моделях? Запишите L1- и L2-регуляризаторы. Почему L1-регуляризация отбирает признаки?