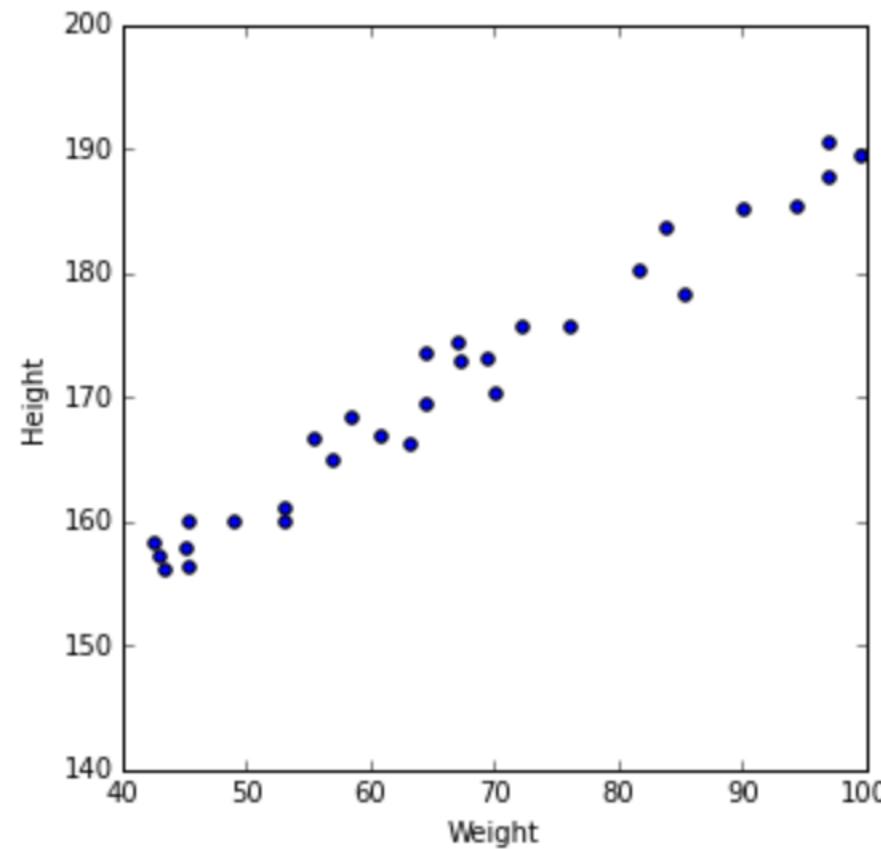
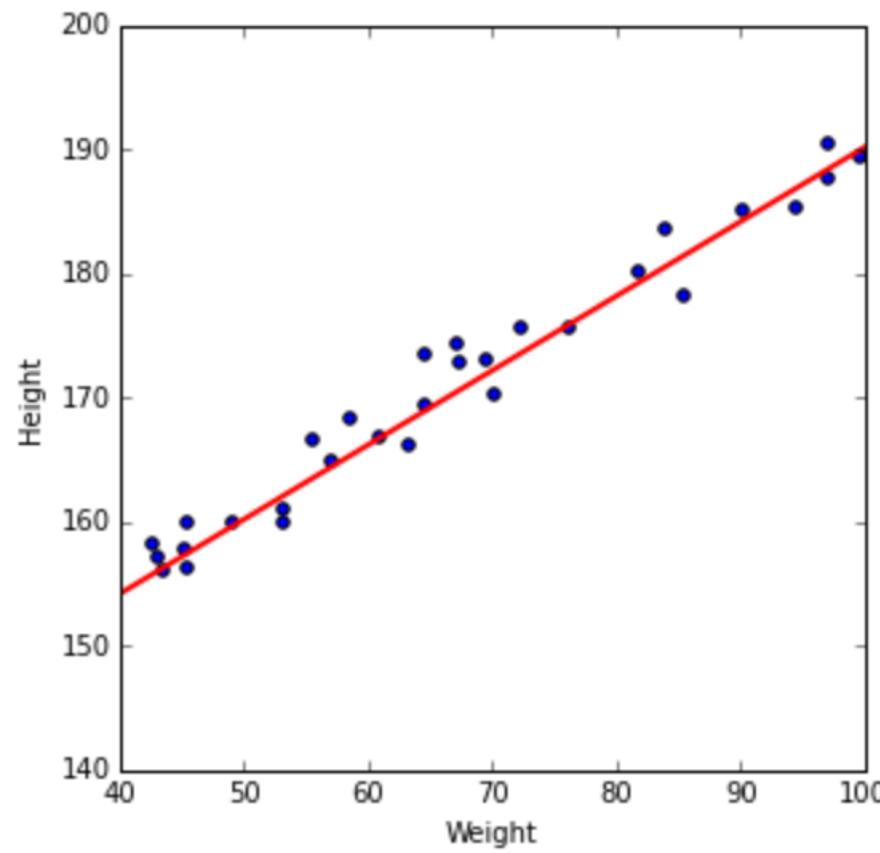


# Линейная регрессия

# Парная регрессия



# Парная регрессия



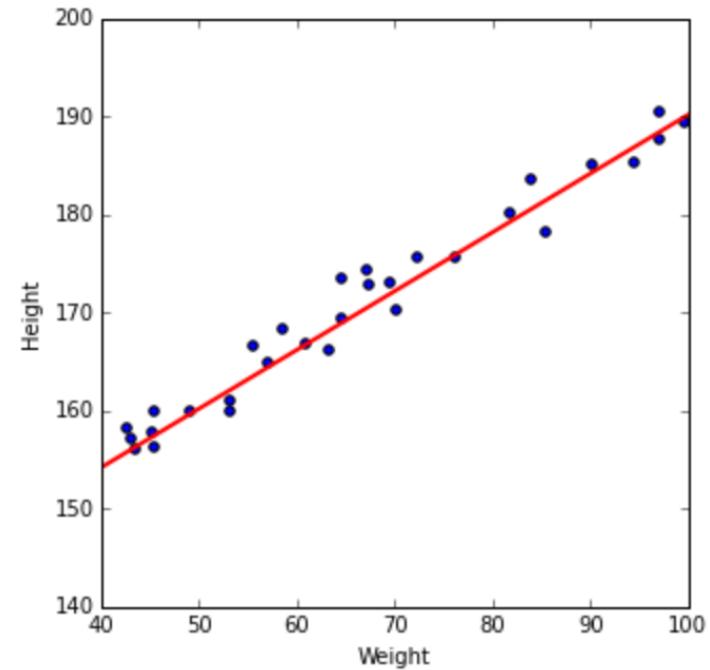
# Парная регрессия

- Простейший случай: один признак
- Модель:  $a(x) = w_1 x + w_0$
- Два параметра:  $w_1$  и  $w_0$
- $w_1$  — тангенс угла наклона
- $w_0$  — где прямая пересекает ось ординат

# Почему модель линейная?

$$a(x) = 2x + 1$$

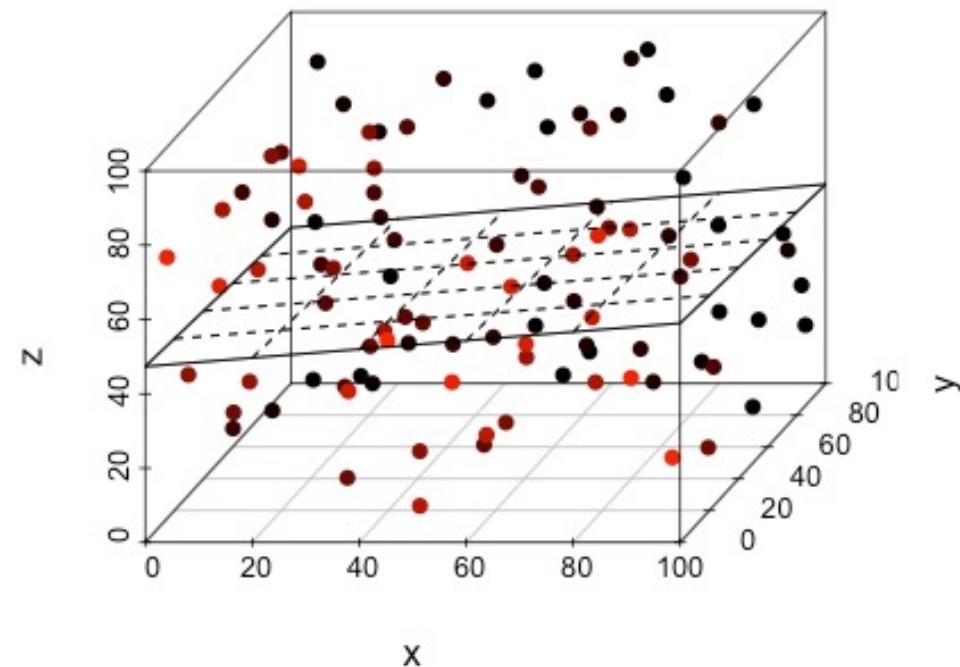
- $x = 1, a(x) = 3$
- $x = 2, a(x) = 5$
- $x = 10, a(x) = 21$
- $x = 20, a(x) = 41$



# Два признака

- Чуть более сложный случай: два признака
- Модель:  $a(x) = w_0 + w_1 x_1 + w_2 x_2$
- Три параметра

# Два признака



# Много признаков

- Общий случай:  $d$  признаков
- Модель

$$a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d$$

- Количество параметров:  $d + 1$

# Много признаков

- Общий случай:  $d$  признаков
- Модель

$$a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d$$

Свободный коэффициент/сдвиг/bias

Веса/коэффициенты

The diagram illustrates the components of a linear regression model. The equation is  $a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d$ . Three blue arrows point from labels below the equation to specific terms: one arrow points from 'Свободный коэффициент/сдвиг/bias' to the term  $w_0$ ; another arrow points from 'Веса/коэффициенты' to the term  $w_d x_d$ ; and a third arrow points from 'Веса/коэффициенты' to the term  $w_1 x_1$ .

- Количество параметров:  $d + 1$

# Много признаков

Запишем через скалярное произведение:

$$\begin{aligned}a(x) &= w_0 + w_1 x_1 + \cdots + w_d x_d = \\&= w_0 + \langle w, x \rangle\end{aligned}$$

Будем считать, что есть признак, всегда равный единице:

$$\begin{aligned}a(x) &= w_1 x_1 + \cdots + w_d x_d = \\&= w_1 * 1 + w_2 x_2 + \cdots + w_d x_d = \\&= \langle w, x \rangle\end{aligned}$$

# Применимость линейной регрессии

# Модель линейной регрессии

$$a(x) = w_1x_1 + \cdots + w_dx_d = \langle w, x \rangle$$

- Нет гарантий, что целевая переменная именно так зависит от признаков
- Надо формировать признаки так, чтобы модель подходила

# Предсказание стоимости квартиры

- Признаки: площадь, район, расстояние до метро
- Целевая переменная: рыночная стоимость квартиры
- Линейная модель:

$$a(x) = w_0 + w_1 * (\text{площадь})$$

$$+ w_2 * (\text{район})$$

$$+ w_3 * (\text{расстояние до метро})$$

# Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * \text{(площадь)}$$

$$+ w_2 * \text{(район)}$$

$$+ w_3 * \text{(расстояние до метро)}$$

# Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * \text{(площадь)}$$

$$+ w_2 * \text{(район)}$$

$$+ w_3 * \text{(расстояние до метро)}$$

- За каждый квадратный метр добавляем  $w_1$  к прогнозу

# Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$

$$+ w_2 * (\text{район})$$

$$+ w_3 * (\text{расстояние до метро})$$

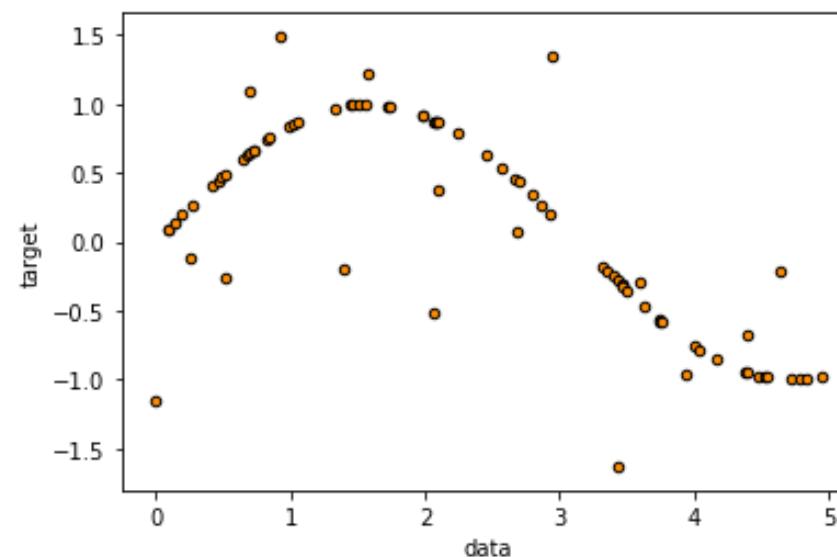
- Что-то странное

# Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$

$$+ w_2 * (\text{район})$$

$$+ w_3 * (\text{расстояние до метро})$$



# Кодирование категориальных признаков

- Значения признака «район»:  $U = \{u_1, \dots, u_m\}$
- Новые признаки вместо  $x_j$ :  $[x_j = u_1], \dots, [x_j = u_m]$
- One-hot кодирование

# Кодирование категориальных признаков

The diagram illustrates the process of encoding categorical features. On the left, a vertical list of categories is shown:

Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

An arrow points from this list to a binary matrix on the right, representing the one-hot encoding of these categories:

ЦАО	ЮАО	САО
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0

# Кодирование категориальных признаков

The diagram illustrates the process of encoding categorical variables. On the left, a vertical list of districts is shown:

Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

An arrow points from this list to a binary matrix on the right, which represents the one-hot encoding of these categories:

ЦАО	ЮАО	САО
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0

$$a(x) = w_0 + w_1 * (\text{площадь})$$

+  $w_2 * (\text{квартира в ЦАО?})$

+  $w_3 * (\text{квартира в ЮАО?})$

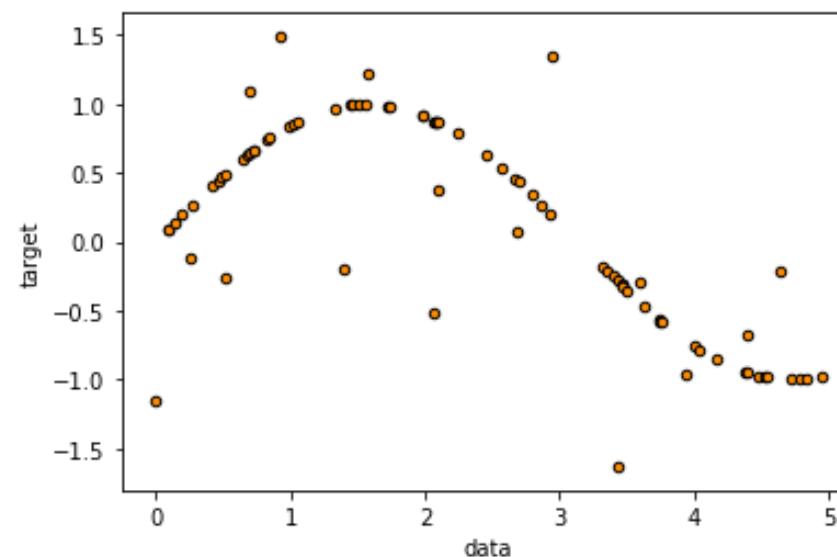
+  $w_4 * (\text{квартира в САО?})$

# Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$

$$+ w_2 * (\text{район})$$

$$+ w_3 * (\text{расстояние до метро})$$

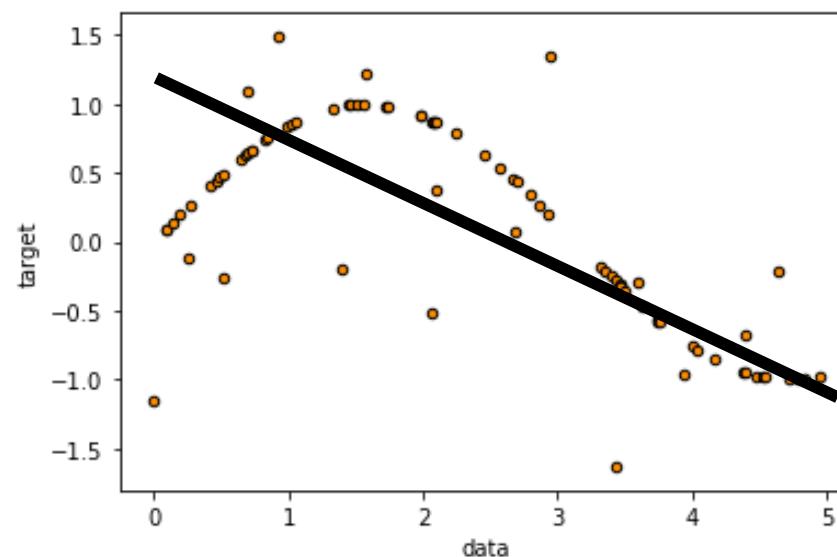


# Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$

$$+ w_2 * (\text{район})$$

$$+ w_3 * (\text{расстояние до метро})$$

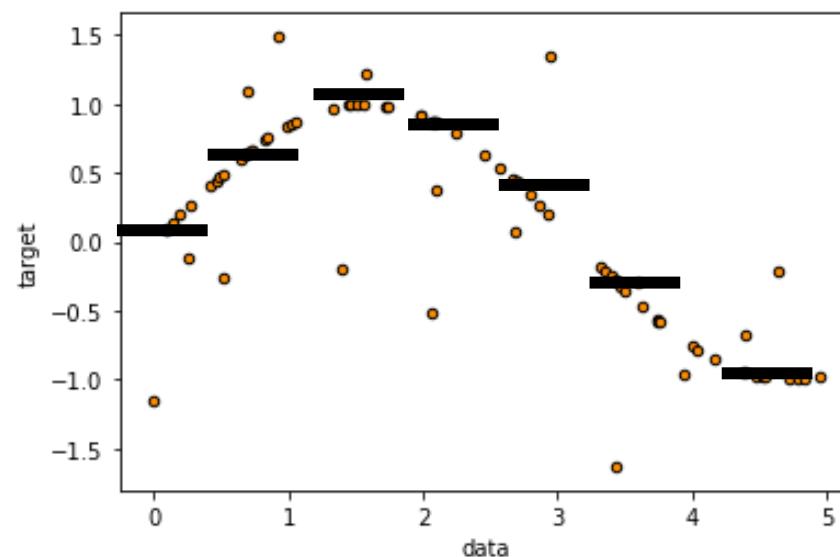


# Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$

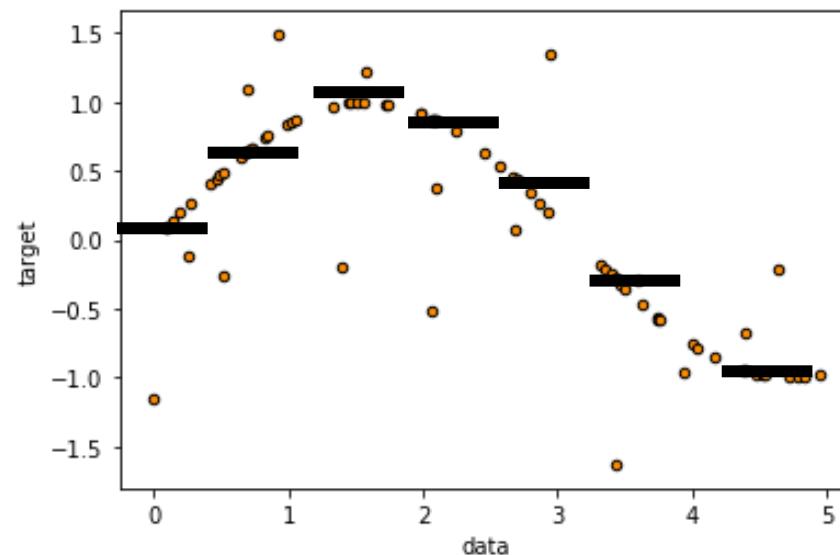
$$+ w_2 * (\text{район})$$

$$+ w_3 * (\text{расстояние до метро})$$



# Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь}) \\ + w_2 * (\text{район}) \\ + w_3 * [t_0 \leq x_3 < t_1] + \dots + w_{3+n} [t_{n-1} \leq x_3 < t_n]$$



# Нелинейные признаки

- Линейная модель с полиномиальными признаками:

$$\begin{aligned}a(x) = & w_0 + w_1 * (\text{площадь}) + w_2 * (\text{этаж}) \\& + w_3 * (\text{расстояние до метро}) + w_4 * (\text{площадь})^2 \\& + w_5 * (\text{этаж})^2 + w_6 * (\text{расстояние до метро})^2 \\& + w_7 * (\text{площадь}) * (\text{этаж}) + \dots\end{aligned}$$

# Линейные модели

- Модель линейной регрессии хороша, если признаки сделаны специально под неё
- Пример: one-hot кодирование категориальных признаков или бинаризация числовых признаков

Линейная регрессия в  
векторном виде

# Модель линейной регрессии

$$a(x) = \langle w, x \rangle$$

- Среднеквадратичная ошибка и задача обучения:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_w$$

# Матрицы

- Матрица — таблица с числами (для простоты)
- Матрица «объекты-признаки»:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix} \in \mathbb{R}^{\ell \times d}$$

# Матрицы

- Матрица — таблица с числами (для простоты)
- Матрица «объекты-признаки»:

объект и его признаки

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

# Матрицы

- Матрица — таблица с числами (для простоты)
- Матрица «объекты-признаки»:

$$\begin{pmatrix} x_{11} & \boxed{x_{12}} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & \boxed{x_{\ell 2}} & \cdots & x_{\ell d} \end{pmatrix}$$

значения признака на всех объектах

# Векторы

- Вектор размера  $d$  — тоже матрица
- Вектор-строка:  $w = (w_1, \dots, w_d) \in \mathbb{R}^{1 \times d}$
- Вектор-столбец:  $w = \begin{pmatrix} w_1 \\ \dots \\ w_d \end{pmatrix} \in \mathbb{R}^{d \times 1}$

# Матричное умножение

- Только для матриц  $A \in \mathbb{R}^{m \times k}$  и  $B \in \mathbb{R}^{k \times n}$
- Результат:  $AB = C \in \mathbb{R}^{m \times n}$
- Правило:

$$c_{ij} = \sum_{p=1}^k a_{ip} b_{pj}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ \quad \\ \quad \end{pmatrix}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ & \end{pmatrix}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 4 \\ 0 & 0 & 2 \end{pmatrix}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 4 \\ 0 & 0 & 2 \end{pmatrix}$$

# Применение линейной модели

- $a(x) = \langle w, x \rangle = w_1x_1 + \cdots + w_dx_d$
- Как применить модель к обучающей выборке?

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix} \xrightarrow{\quad} \begin{pmatrix} \sum_{i=1}^d w_i x_{1i} \\ \sum_{i=1}^d w_i x_{2i} \\ \vdots \\ \sum_{i=1}^d w_i x_{\ell i} \end{pmatrix}$$
$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \xrightarrow{\quad}$$

# Модель линейной регрессии

- Среднеквадратичная ошибка и задача обучения:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_w$$

# Вычисление ошибки

- Отклонения прогнозов от ответов:

$$Xw - y = \begin{pmatrix} \langle w, x_1 \rangle - y_1 \\ \vdots \\ \langle w, x_\ell \rangle - y_\ell \end{pmatrix}$$

# Вычисление ошибки

- Евклидова норма:

$$\|z\| = \sqrt{\sum_{j=1}^n z_j^2}$$

$$\|z\|^2 = \sum_{j=1}^n z_j^2$$

# Вычисление ошибки

- Отклонения прогнозов от ответов:

$$Xw - y = \begin{pmatrix} \langle w, x_1 \rangle - y_1 \\ \vdots \\ \langle w, x_\ell \rangle - y_\ell \end{pmatrix}$$

- Среднеквадратичная ошибка:

$$\frac{1}{\ell} \|Xw - y\|^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2$$

# Обучение линейной регрессии

$$\frac{1}{\ell} \|Xw - y\|^2 \rightarrow \min_w$$

- Вычисление MSE в NumPy:

```
np.square(X.dot(w) - y).mean()
```

# Обучение линейной регрессии

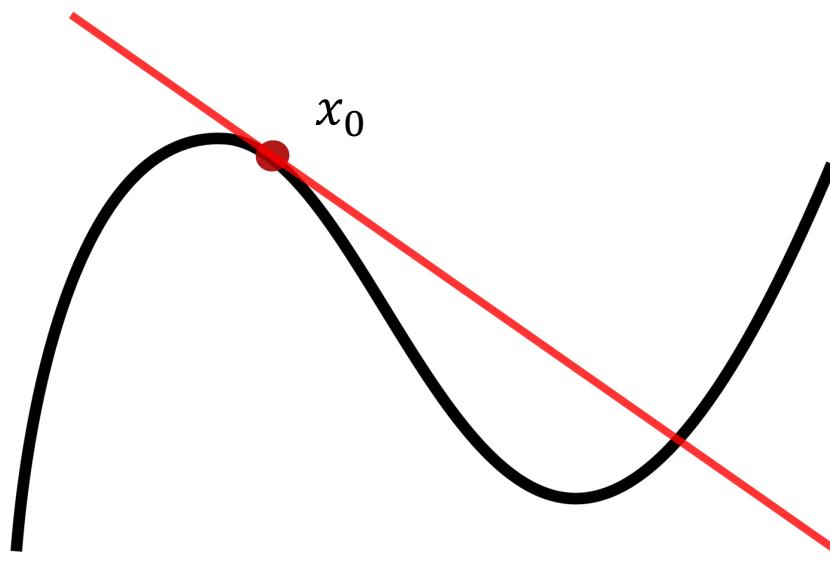
# Среднеквадратичная ошибка

- MSE для линейной регрессии:

$$Q(w_1, \dots, w_d) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\textcolor{red}{w_1}x_1 + \dots + \textcolor{red}{w_d}x_d - y_i)^2$$

# Производная

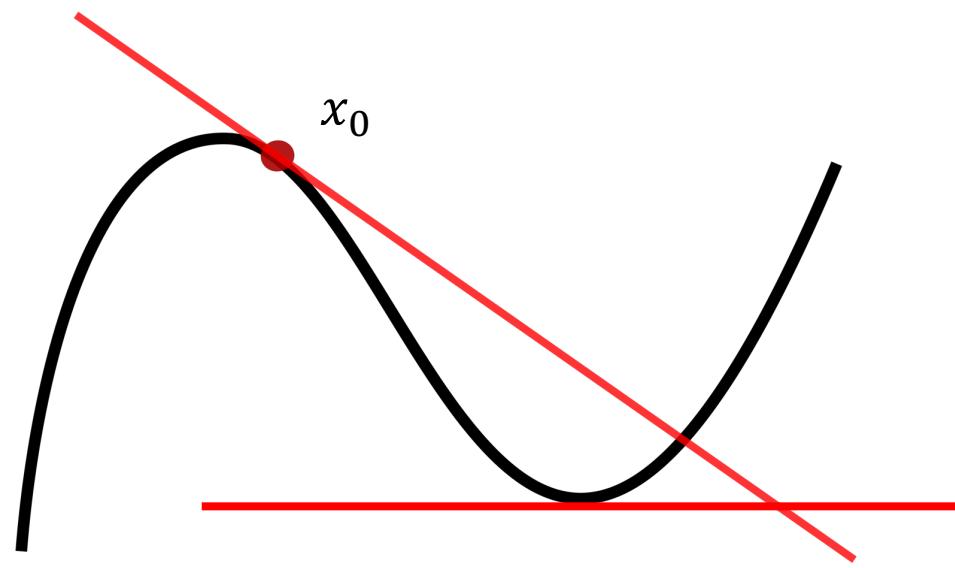
$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$



# Производная

- Если точка  $x_0$  — экстремум и в ней существует производная, то

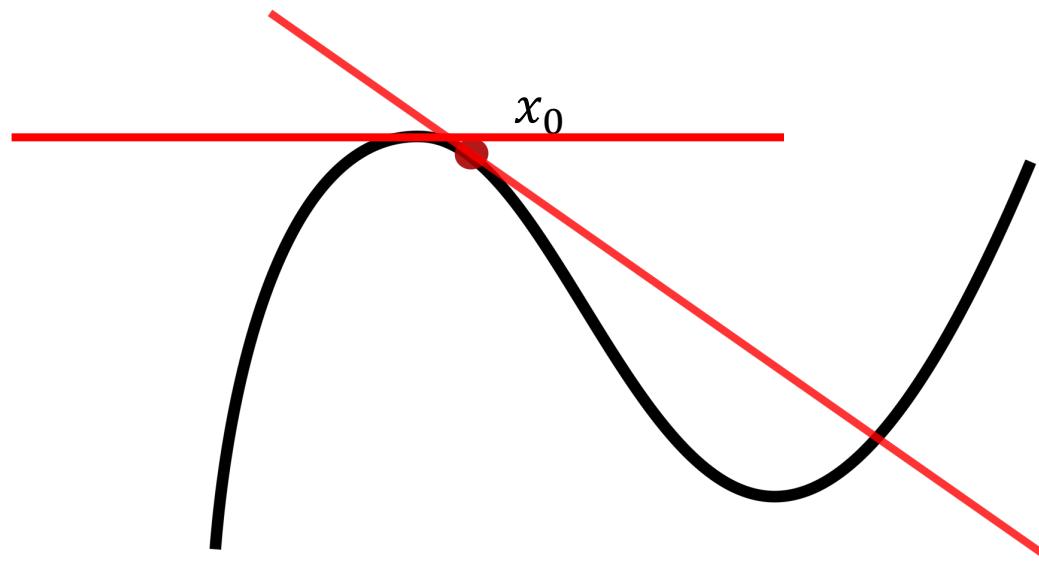
$$f'(x_0) = 0$$



# Производная

- Если точка  $x_0$  — экстремум и в ней существует производная, то

$$f'(x_0) = 0$$



# Градиент

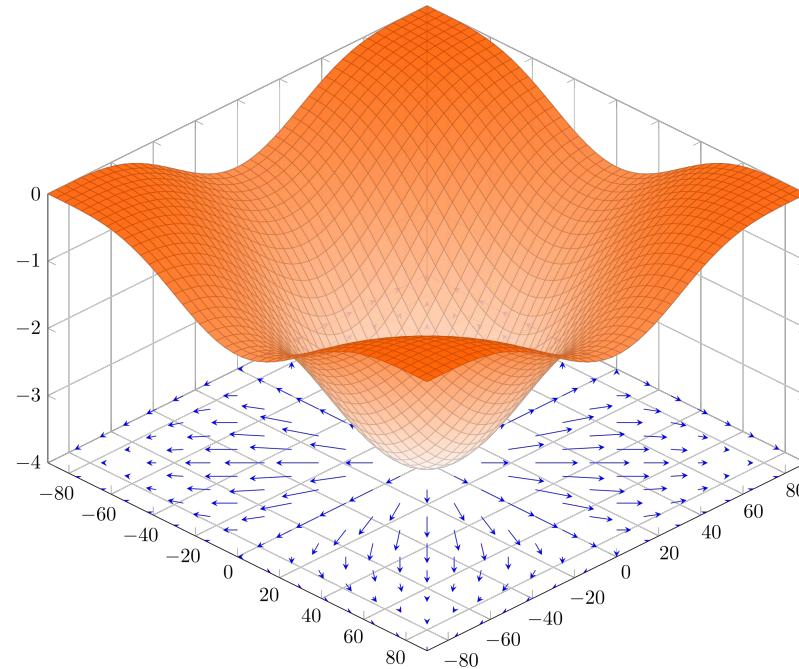
- Градиент — вектор частных производных

$$\nabla f(x) = \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_d} \right)$$

- У градиента есть важное свойство!

# Важное свойство

- Зафиксируем точку  $x_0$
- В какую сторону функция быстрее всего растёт?



# Важное свойство

- Зафиксируем точку  $x_0$
- В какую сторону функция быстрее всего растёт?
- В направлении градиента!
- Если градиент равен нулю, то это экстремум

# Условие экстремума

- Если точка  $x_0$  — экстремум и в ней существует производная, то

$$\nabla f(x_0) = 0$$

# Условие экстремума

- Если точка  $x_0$  — экстремум и в ней существует производная, то

$$\nabla f(x_0) = 0$$

- Если функция выпуклая, то экстремум один
- MSE для линейной регрессии — выпуклая!
  - (при некоторых условиях)

# Обучение линейной регрессии

- Можно посчитать градиент MSE:

$$\nabla \frac{1}{\ell} \|Xw - y\|^2 = \frac{2}{\ell} X^T(Xw - y)$$

- Приравниваем нулю и решаем систему линейных уравнений:

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

# Аналитическое решение

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

- Если матрица  $X^T X$  вырожденная, то будут проблемы
- Даже если она почти вырожденная, всё равно будут проблемы
- Если признаков много, то придётся долго ждать

# Градиент и его свойства

# Среднеквадратичная ошибка

- MSE для линейной регрессии:

$$Q(w_1, \dots, w_d) = \sum_{i=1}^{\ell} (\textcolor{red}{w_1}x_1 + \dots + \textcolor{red}{w_d}x_d - y_i)^2$$

# Градиент

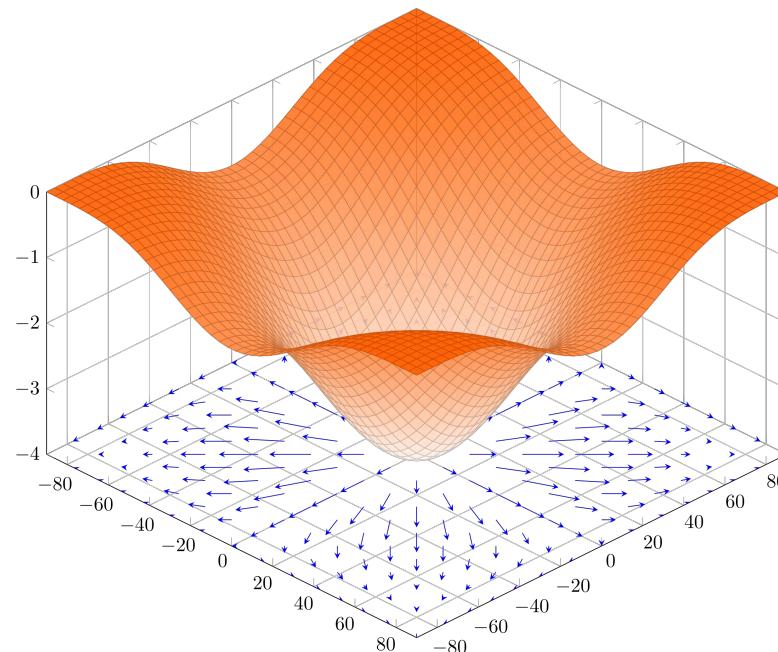
- Градиент — вектор частных производных

$$\nabla f(x) = \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_d} \right)$$

- У градиента есть важное свойство!

# Важное свойство

- Зафиксируем точку  $x_0$
- В какую сторону функция быстрее всего растёт?



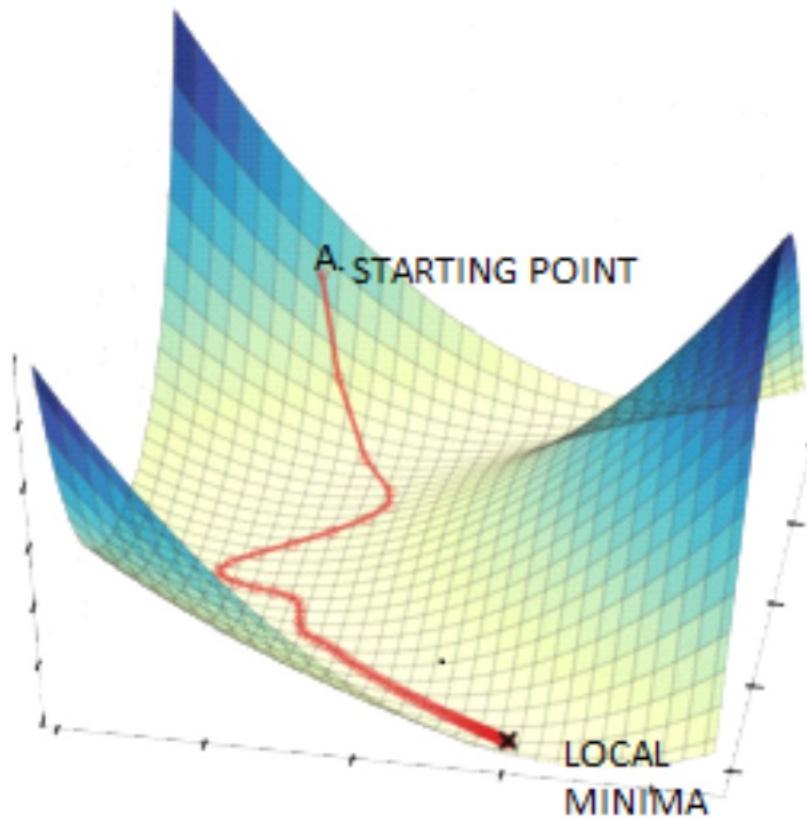
# Важное свойство

- Зафиксируем точку  $x_0$
- В какую сторону функция быстрее всего растёт?
- В направлении градиента!
- А быстрее всего убывает в сторону антиградиента

Как это пригодится?



Как это пригодится?



# Градиентный спуск

# Градиентный спуск

1. Начальное приближение:  $w^0$

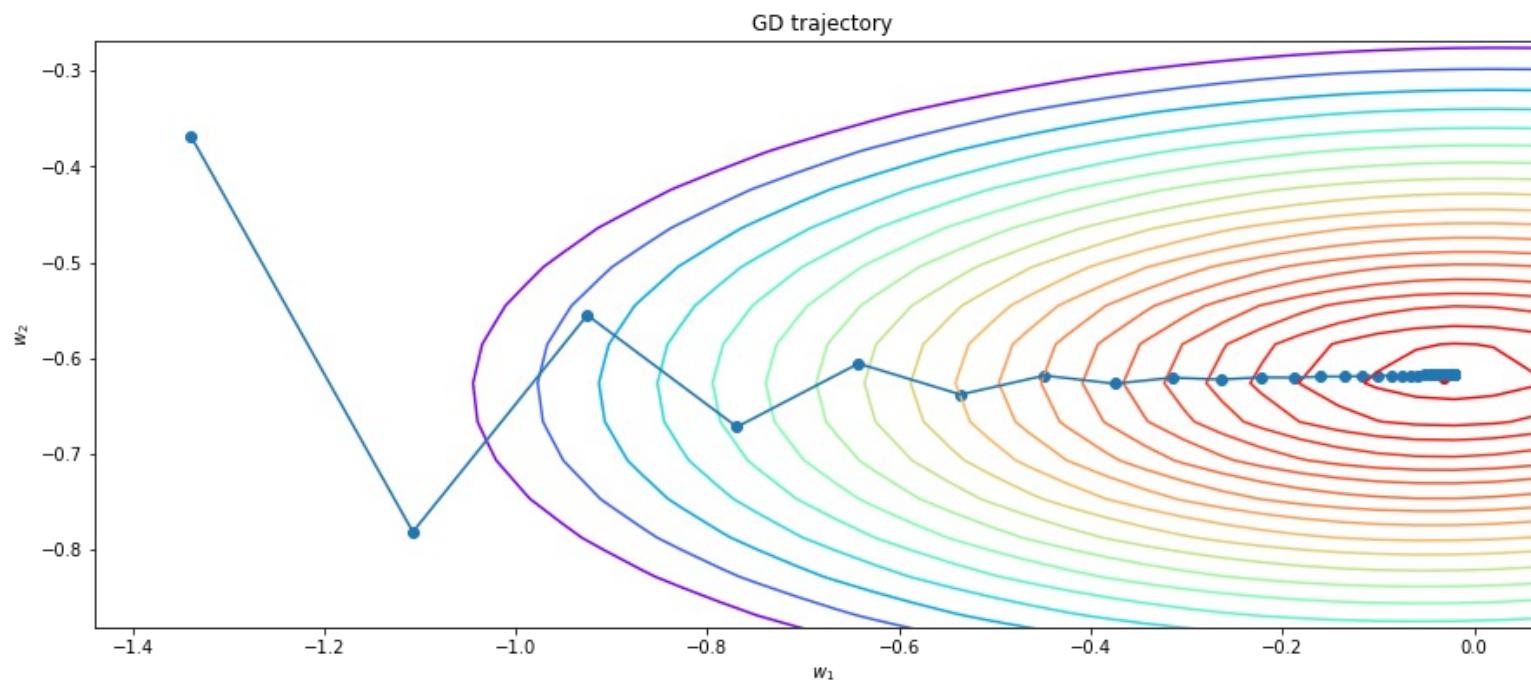
2. Повторять:

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

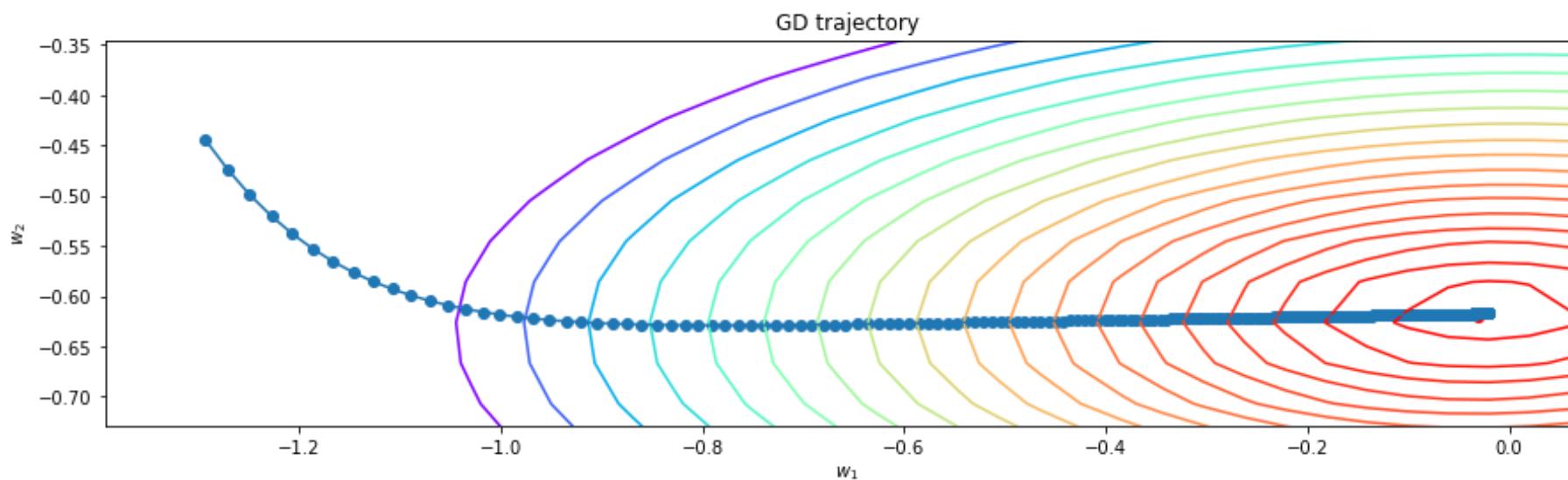
3. Останавливаемся, если

$$\|w^t - w^{t-1}\| < \varepsilon$$

# Длина шага



# Длина шага



# Стохастический градиентный спуск

# Градиентный спуск

1. Начальное приближение:  $w^0$

2. Повторять:

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

3. Останавливаемся, если

$$\|w^t - w^{t-1}\| < \varepsilon$$

# Линейная регрессия

$$Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x \rangle - y_i)^2$$

- $\frac{\partial Q}{\partial w_1} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_{i1} (\langle w, x \rangle - y_i)$
- ...
- $\frac{\partial Q}{\partial w_d} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_{id} (\langle w, x \rangle - y_i)$
- $\nabla Q(w) = \frac{2}{\ell} X^T (Xw - y)$

# Сложности градиентного спуска

- Для вычисления градиента, как правило, надо просуммировать что-то по всем объектам
- И это для одного маленького шага!

# Оценка градиента

$$Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a(x_i))$$

- Градиент:

$$\nabla Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} \nabla L(y_i, a(x_i))$$

- Может, оценить градиент одним слагаемым?

$$\nabla Q(w) \approx \nabla L(y_i, a(x_i))$$

# Стохастический градиентный спуск

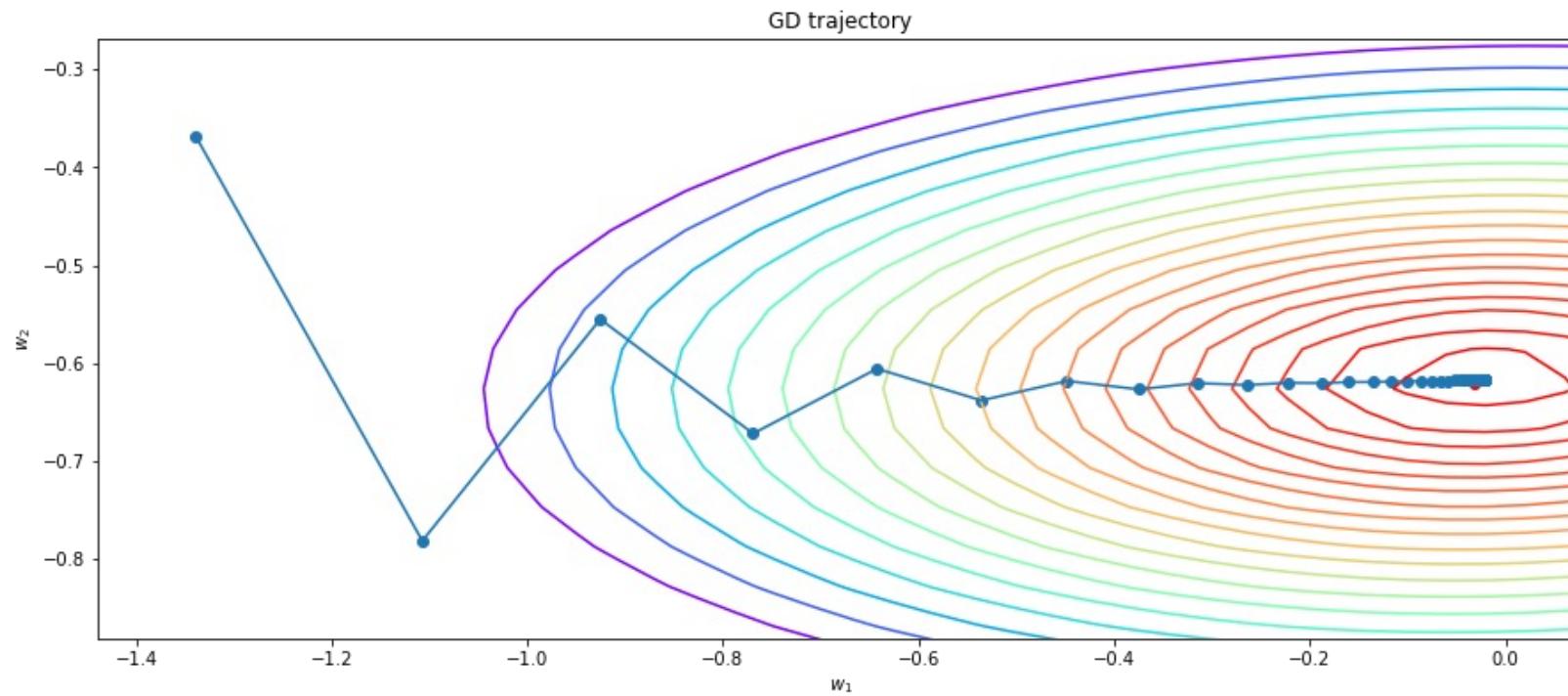
1. Начальное приближение:  $w^0$
2. Повторять, каждый раз выбирая случайный объект  $i_t$ :

$$w^t = w^{t-1} - \eta \nabla L(y_{i_t}, a(x_{i_t}))$$

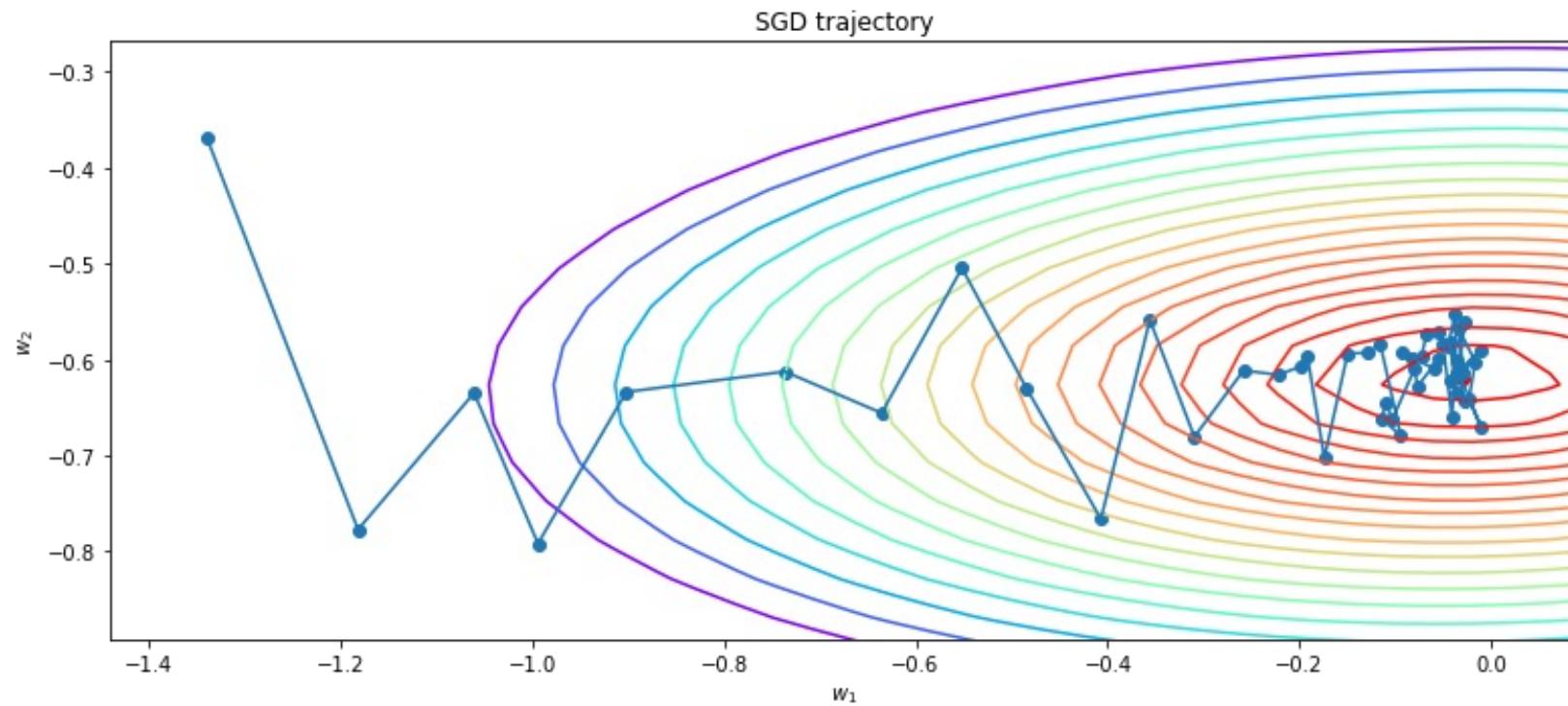
3. Останавливаемся, если

$$\|w^t - w^{t-1}\| < \varepsilon$$

# Градиентный спуск



# Стохастический градиентный спуск



# Стохастический градиентный спуск

1. Начальное приближение:  $w^0$
2. Повторять, каждый раз выбирая случайный объект  $i_t$ :

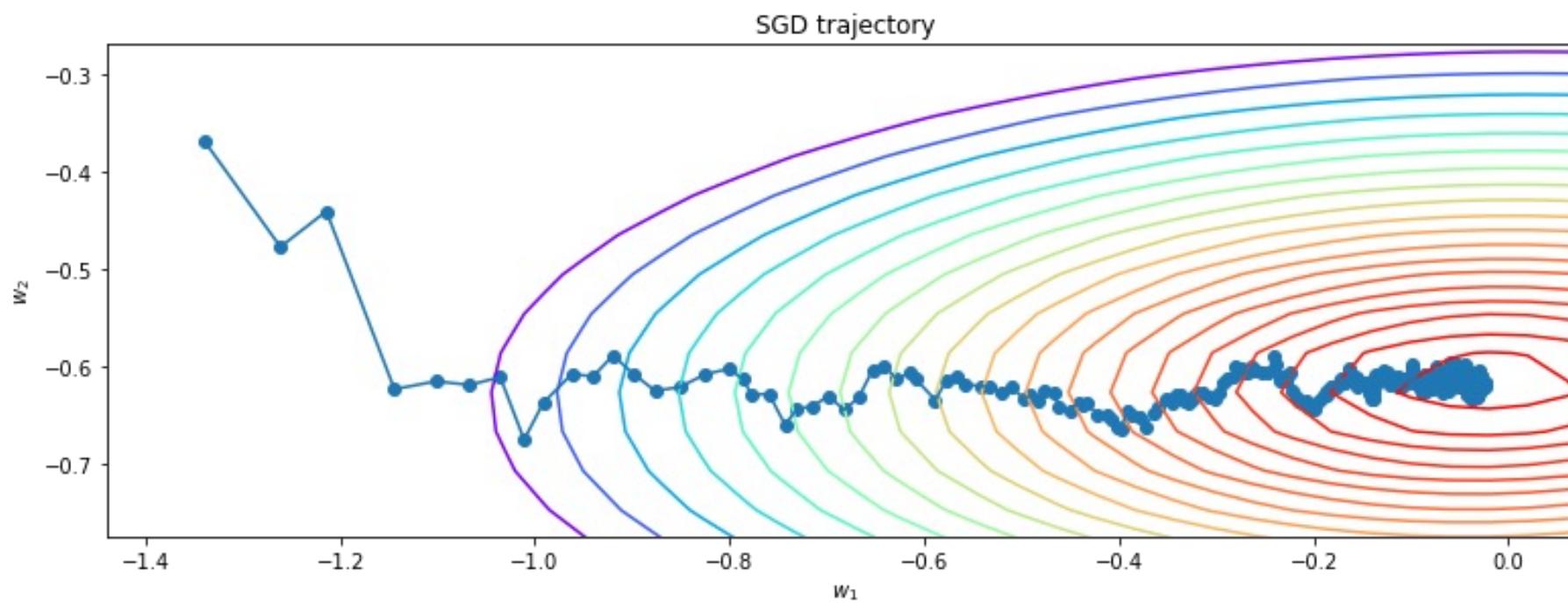
$$w^t = w^{t-1} - \eta_t \nabla L(y_{i_t}, a(x_{i_t}))$$

3. Останавливаемся, если

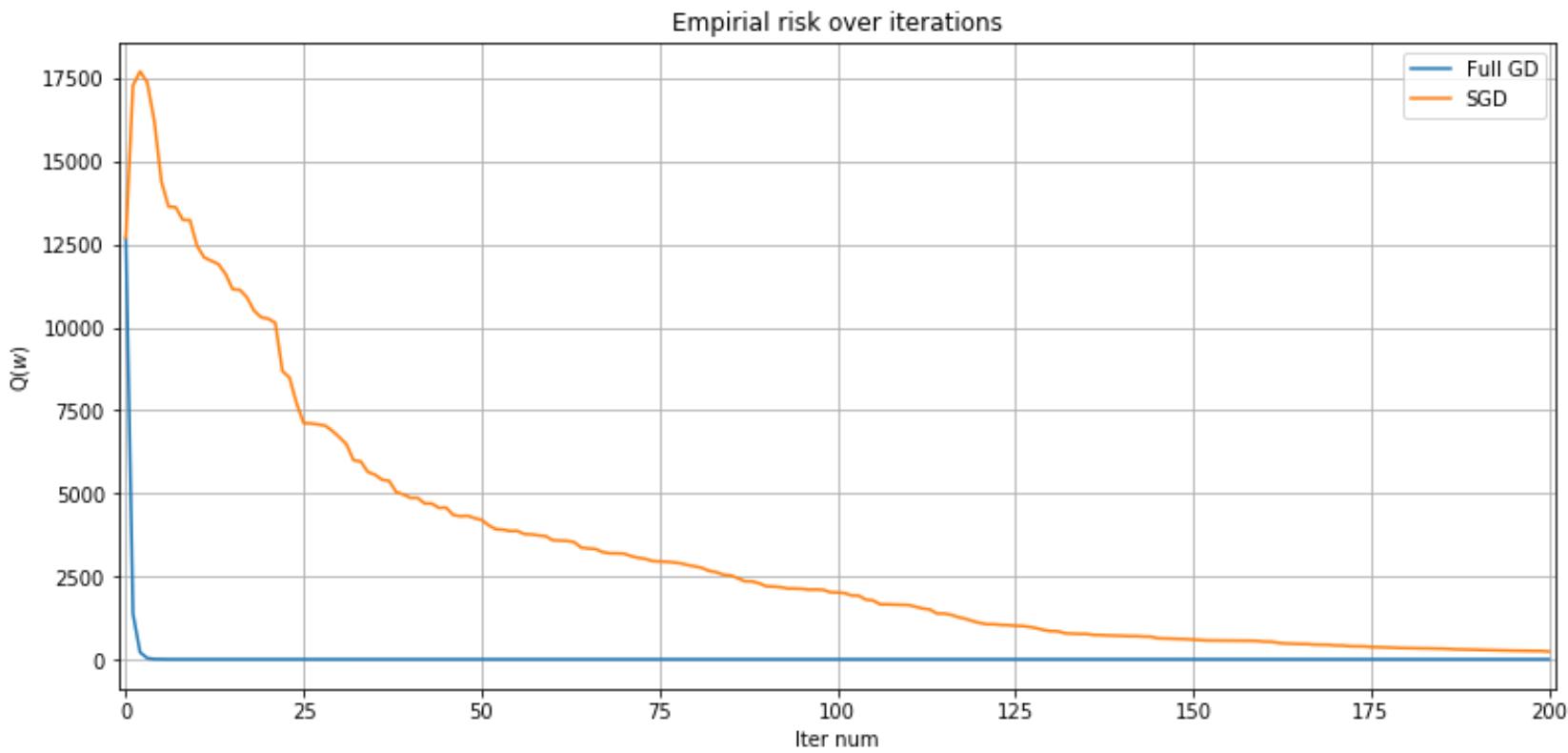
$$\|w^t - w^{t-1}\| < \varepsilon$$

# Стochastic gradient descent

$$\eta_t = \frac{0.1}{t^{0.3}}$$



# Стochastic gradient descent



# Mini-batch

1. Начальное приближение:  $w^0$
2. Повторять, каждый раз выбирая  $m$  случайных объектов  $i_1, \dots, i_m$ :

$$w^t = w^{t-1} - \eta_t \frac{1}{m} \sum_{j=1}^m \nabla L \left( y_{i_j}, a(x_{i_j}) \right)$$

3. Останавливаемся, если

$$\|w^t - w^{t-1}\| < \varepsilon$$

# Резюме

- SGD существенно упрощает каждый шаг
- Траектория в SGD существенно менее стабильная
- Но это может помочь выбираться из локальных минимумов
- Mini-batch — некоторый компромисс между SGD и Full GD

# Функции потерь в задачах регрессии

# Среднеквадратичная ошибка

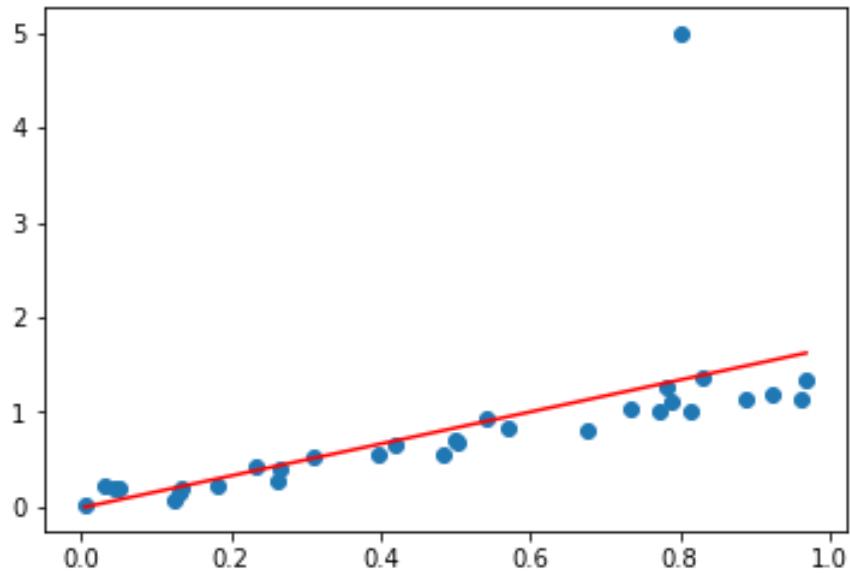
- Частый выбор — квадратичная функция потерь

$$L(y, a) = (a - y)^2$$

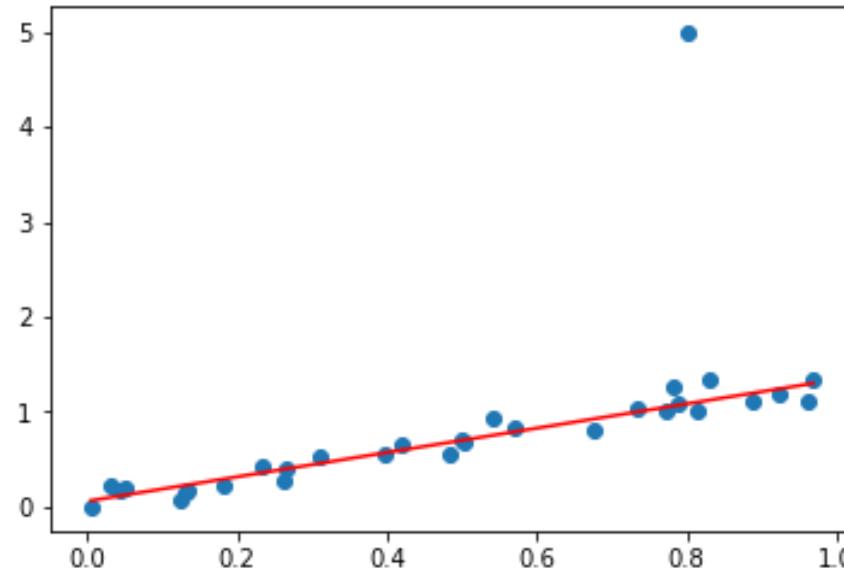
- Функционал ошибки — среднеквадратичная ошибка (mean squared error, MSE)

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

# Выбросы



С учётом выброса



Без учёта выброса

Обучение на среднеквадратичную ошибку

# Выбросы

$a(x)$	$y$	$(a(x) - y)^2$
2	1	1
1	2	1
2	3	1
5	4	1
6	5	1
7	100	8649
6	7	1

$$MSE \approx 1236$$

# Выбросы

$a(x)$	$y$	$(a(x) - y)^2$
4	1	9
5	2	9
6	3	9
7	4	9
8	5	9
10	100	8100
10	7	9

$$MSE \approx 1164$$

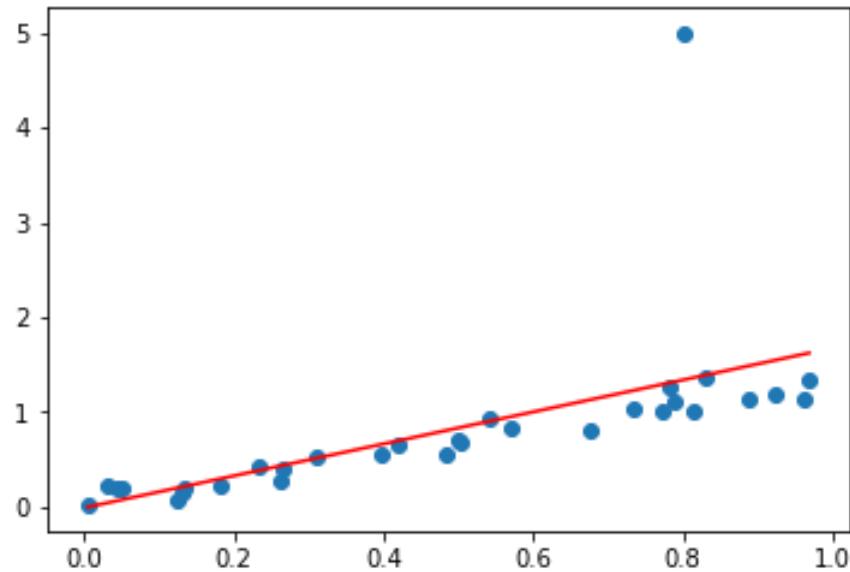
# Средняя абсолютная ошибка

$$L(y, a) = |a - y|$$

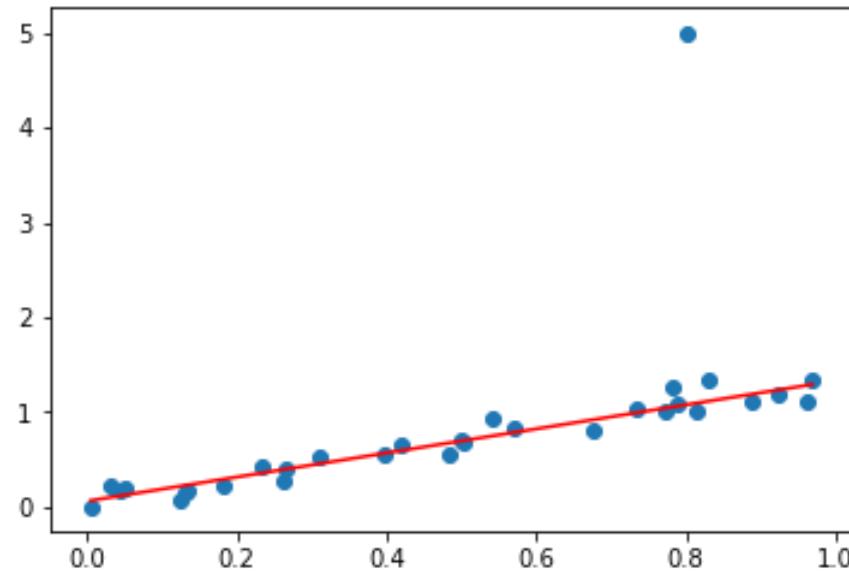
- Функционал ошибки — средняя абсолютная ошибка (mean absolute error, MAE)

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

# Выбросы



Обучение на MSE



Обучение на MAE

# Выбросы

$a(x)$	$y$	$ a(x) - y $
2	1	1
1	2	1
2	3	1
5	4	1
6	5	1
7	100	93
6	7	1

$$MAE \approx 14.14$$

# Выбросы

$a(x)$	$y$	$ a(x) - y $
4	1	3
5	2	3
6	3	3
7	4	3
8	5	3
10	100	90
10	7	3

$$MAE \approx 15.43$$

# ФУНКЦИЯ ПОТЕРЬ ХУБЕРА

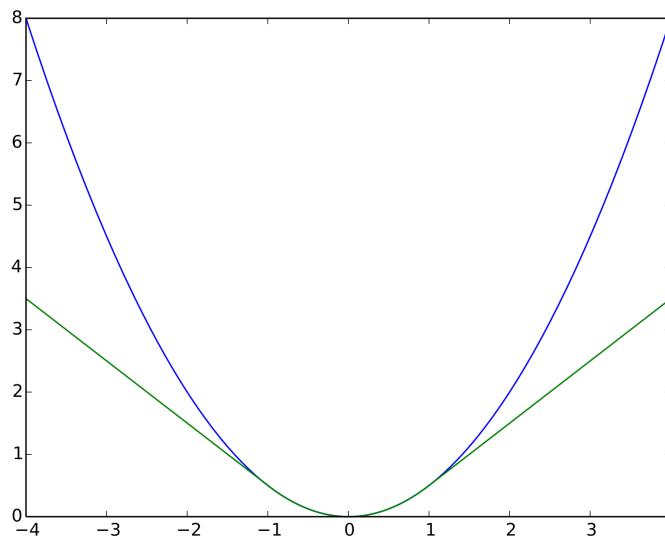
$$L_H(y, a) = \begin{cases} \frac{1}{2}(y - a)^2, & |y - a| < \delta \\ \delta \left( |y - a| - \frac{1}{2}\delta \right), & |y - a| \geq \delta \end{cases}$$

- ФУНКЦИОНАЛ ОШИБКИ:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} L_H(y_i, a(x_i))$$

# Функция потерь Хубера

$$L_H(y, a) = \begin{cases} \frac{1}{2}(y - a)^2, & |y - a| < \delta \\ \delta \left( |y - a| - \frac{1}{2}\delta \right), & |y - a| \geq \delta \end{cases}$$



# MAPE

- Mean Absolute Percentage Error (средний модуль относительной ошибки)

$$L(y, a) = \left| \frac{y - a}{y} \right|$$

$$Q(a, X) = \frac{100\%}{\ell} \sum_{i=1}^{\ell} \left| \frac{a(x_i) - y_i}{y_i} \right|$$

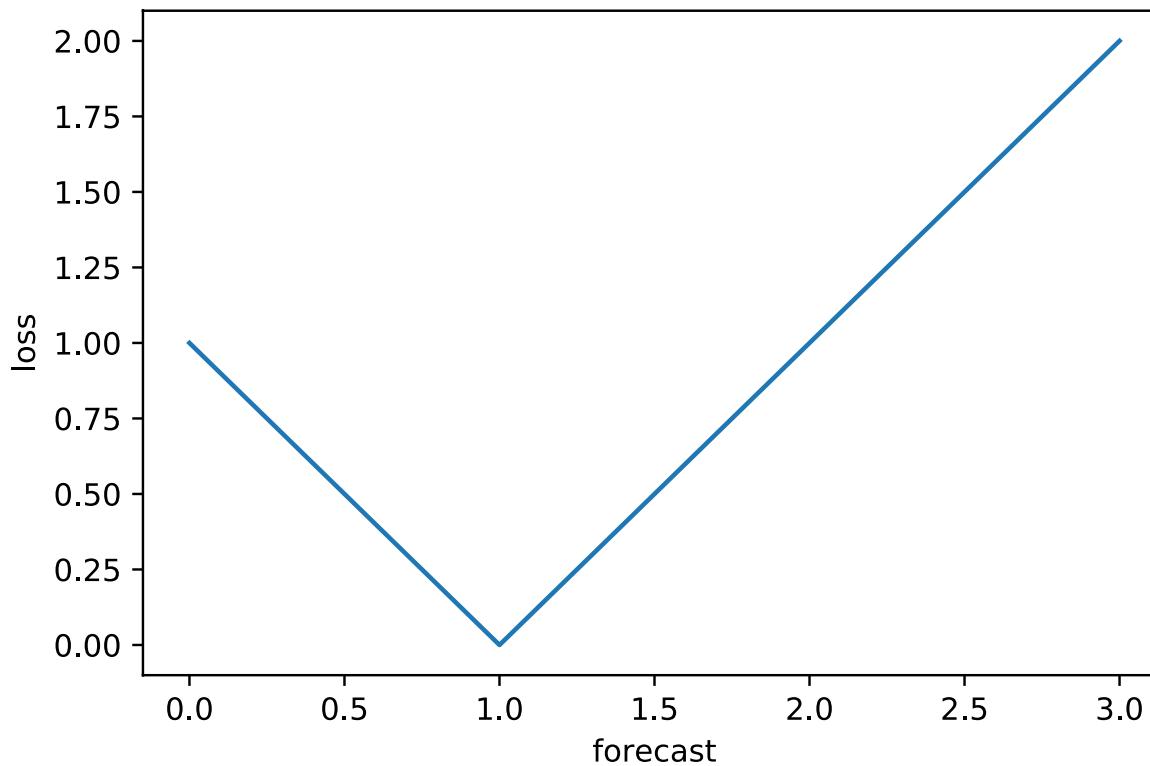
# MAPE

$$L(y, a) = \left| \frac{y - a}{y} \right|$$

- Особенности (при  $a \geq 0$ ):
  - Недопрогноз штрафуется максимум на единицу
  - Перепрогноз может быть оштрафован любым числом
  - Несимметричная функция потерь (отдаёт предпочтение недопрогнозу)

# MAPE

$$L(y, a) = \left| \frac{y - a}{y} \right|$$



# SMAPE

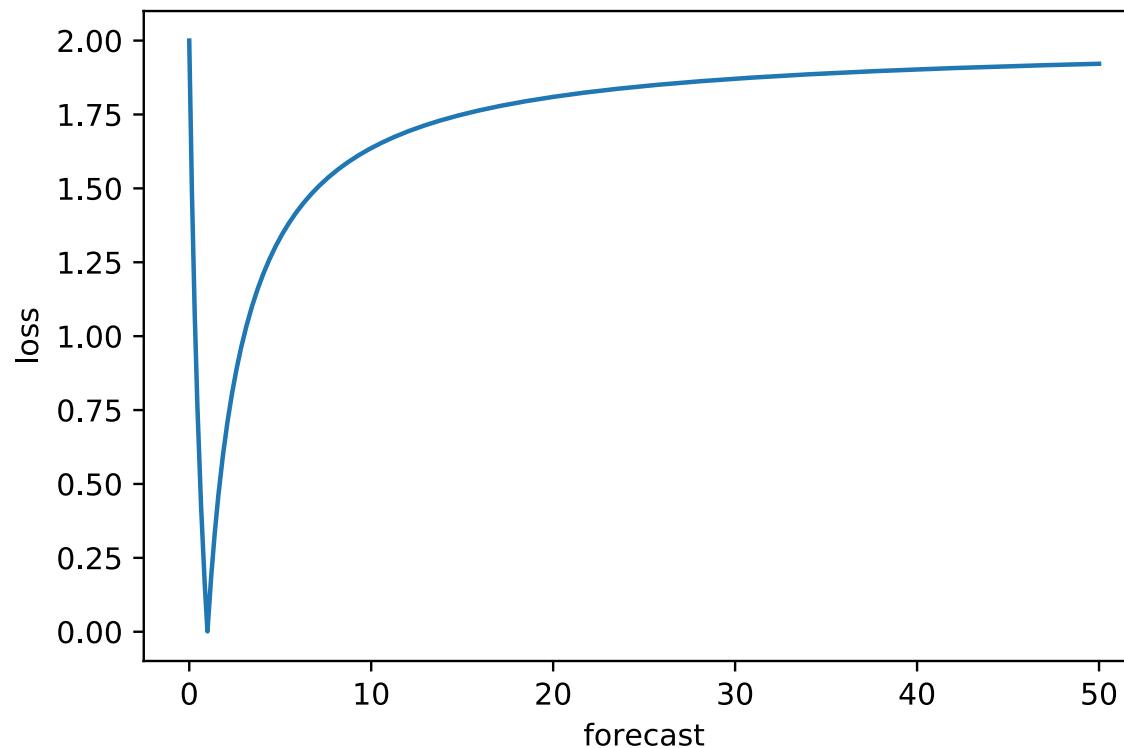
- Symmetric Mean Absolute Percentage Error (симметричный средний модуль относительной ошибки)

$$L(y, a) = \frac{|y - a|}{(|y| + |a|)/2}$$

$$Q(a, X) = \frac{100\%}{\ell} \sum_{i=1}^{\ell} \frac{|y_i - a(x_i)|}{(|y_i| + |a(x_i)|)/2}$$

# SMAPE

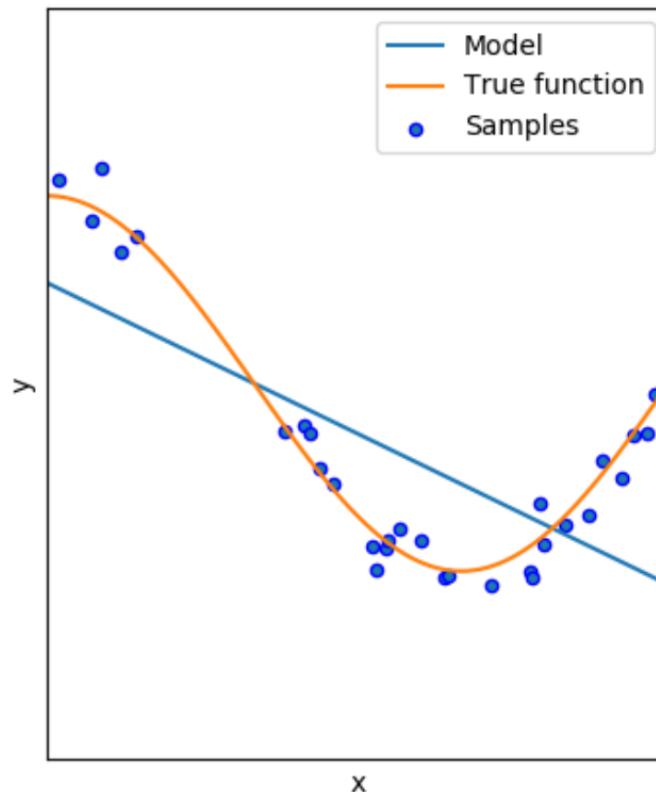
$$L(y, a) = \frac{|y - a|}{(|y| + |a|)/2}$$



# Переобучение и регуляризация линейных моделей

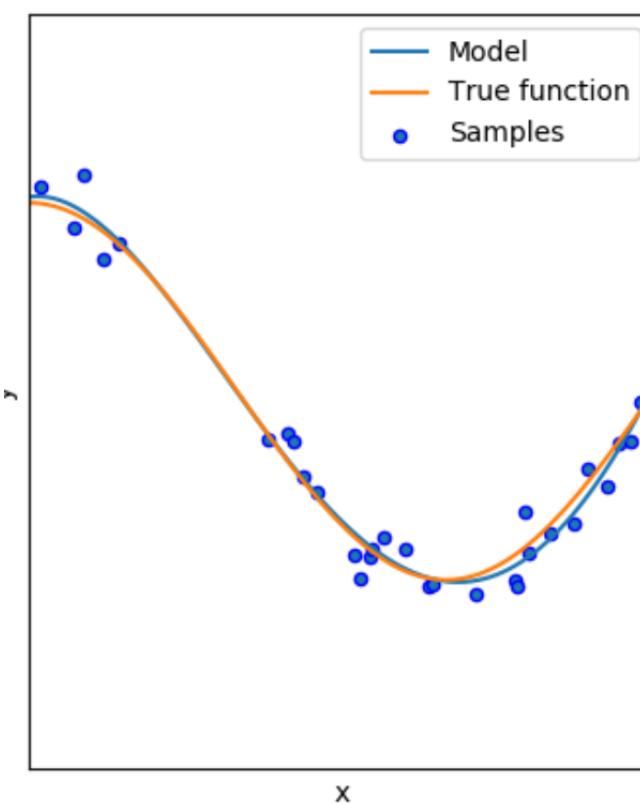
# Нелинейная задача

$$a(x) = w_0 + w_1 x$$



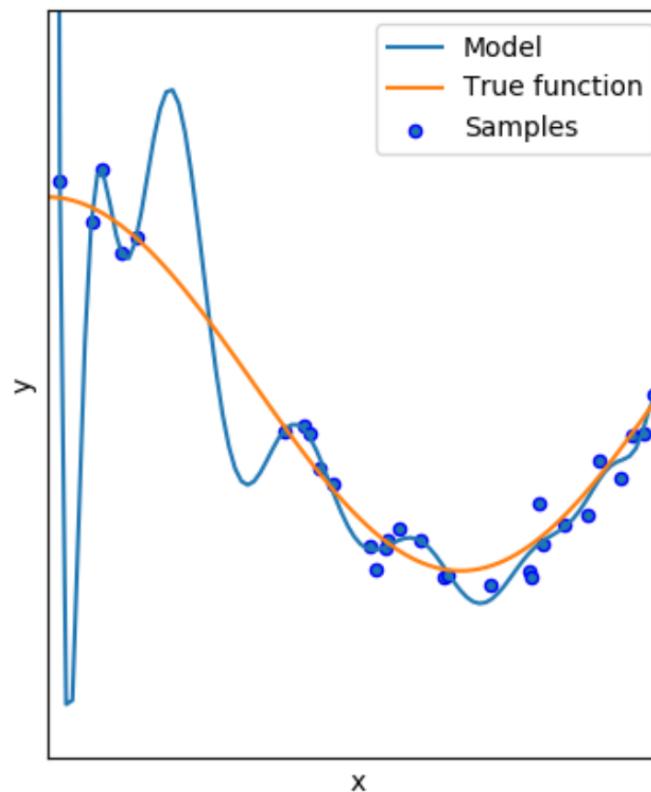
# Нелинейная задача

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$$



# Нелинейная задача

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \cdots + w_{15} x^{15}$$



# Симптом переобучения

$$a(x) = 0.5 + 13458922x - 43983740x^2 + \dots$$

- Большие коэффициенты — симптом переобучения
- Эмпирическое наблюдение

# Симптом переобучения

- Большие коэффициенты в линейной модели — это плохо
- Пример: предсказание роста по весу

$$a(x) = 698x - 41714$$

- Изменение веса на 0.01 кг приведет к изменению роста на 7 см
- Не похоже на правильную зависимость

# Регуляризация

- Будем штрафовать за большие веса!
- Пример функционала:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2$$

- Регуляризатор:

$$\|w\|^2 = \sum_{j=1}^d w_j^2$$

# Регуляризация

- Регуляризованный функционал

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2 \rightarrow \min_w$$

- $\lambda$  — коэффициент регуляризации

# Регуляризация

- Регуляризованный функционал

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2 \rightarrow \min_w$$

- Аналитическое решение:

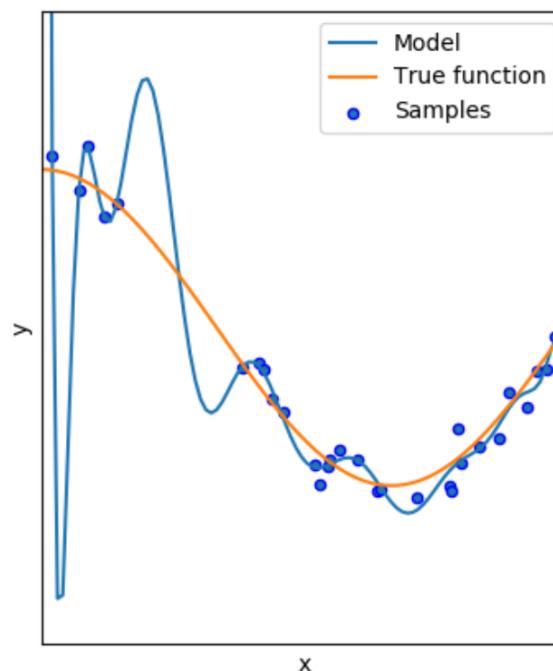
$$w = (X^T X + \lambda I)^{-1} X^T y$$

- Гребневая регрессия (Ridge regression)

# Эффект регуляризации

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \cdots + w_{15} x^{15}$$

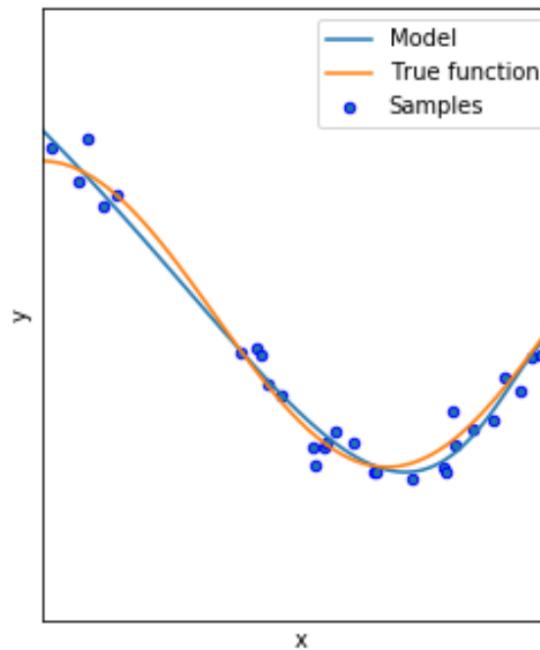
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 \rightarrow \min_w$$



# Эффект регуляризации

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \cdots + w_{15} x^{15}$$

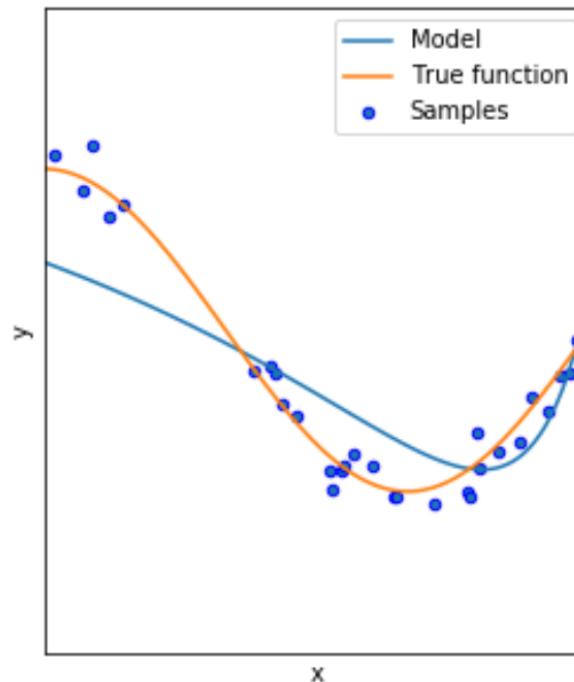
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \textcolor{red}{0.01} \|w\|^2 \rightarrow \min_w$$



# Эффект регуляризации

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \cdots + w_{15} x^{15}$$

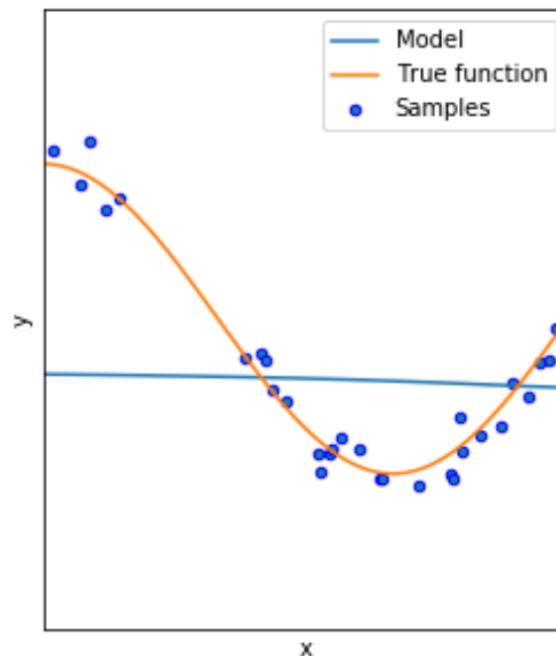
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \textcolor{red}{1} \|w\|^2 \rightarrow \min_w$$



# Эффект регуляризации

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \cdots + w_{15} x^{15}$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \textcolor{red}{100} \|w\|^2 \rightarrow \min_w$$



# Лассо

- Регуляризованный функционал

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^d |w_j| \rightarrow \min_w$$

- LASSO (Least Absolute Shrinkage and Selection Operator)
- Некоторые веса зануляются
- Приводит к отбору признаков

# Регуляризаторы

- $\|z\|_2 = \sqrt{\sum_{j=1}^d z_j^2}$  —  $L_2$ -норма
- $\|z\|_1 = \sum_{j=1}^d |z_j|$  —  $L_1$ -норма

# Интерпретация линейных моделей

# Предсказание стоимости квартиры

$$a(x) = 100.000 * (\text{площадь}) \\ + 500.000 * (\text{число магазинов рядом}) \\ + 100 * (\text{средний доход жильцов дома})$$

# Предсказание стоимости квартиры

$$a(x) = 100.000 * (\text{площадь})$$

$$+ 500.000 * (\text{число магазинов рядом})$$

$$+ 100 * (\text{средний доход жильцов дома})$$

- Чем больше вес, тем важнее признак?

# Предсказание стоимости квартиры

$$a(x) = 100.000 * (\text{площадь в кв. м.}) \\ + 500.000 * (\text{число магазинов рядом}) \\ + 100 * (\text{средний доход жильцов дома})$$

- Чем больше вес, тем важнее признак?

# Предсказание стоимости квартиры

$$a(x) = 10 * (\text{площадь в кв. см.})$$

+ 500.000 \* (число магазинов рядом)

+ 100 \* (средний доход жильцов дома)

- Чем больше вес, тем важнее признак?

# Предсказание стоимости квартиры

$$a(x) = 100.000 * (\text{площадь в кв. м.}) \\ + 500.000 * (\text{число магазинов рядом}) \\ + 100 * (\text{средний доход жильцов дома})$$

- Чем больше вес, тем важнее признак?

# Предсказание стоимости квартиры

$$a(x) = 100.000 * (\text{площадь в кв. м.}) \\ + 500.000 * (\text{число магазинов рядом}) \\ + 100 * (\text{средний доход жильцов дома})$$

- Чем больше вес, тем важнее признак?
- Только если признаки масштабированы!

# Масштабирование признаков

- Отмасштабируем  $j$ -й признак
- Вычисляем среднее и стандартное отклонение признака на обучающей выборке:

$$\mu_j = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i^j$$

$$\sigma_j = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (x_i^j - \mu_j)^2}$$

# Масштабирование признаков

- Вычтем из каждого значения признака среднее и поделим на стандартное отклонение:

$$x_i^j := \frac{x_i^j - \mu_j}{\sigma_j}$$

# Регуляризация

- Если модель переобучается, то веса используются для запоминания обучающей выборки
- Правильнее масштабировать признаки и регуляризовать модель перед изучением весов