

# Основы машинного обучения

## Лекция 7

### Линейная регрессия и градиентный спуск

Евгений Соколов

[esokolov@hse.ru](mailto:esokolov@hse.ru)

НИУ ВШЭ, 2024

Градиентный спуск

# Градиентный спуск

1. Начальное приближение:  $w^0$

2. Повторять:

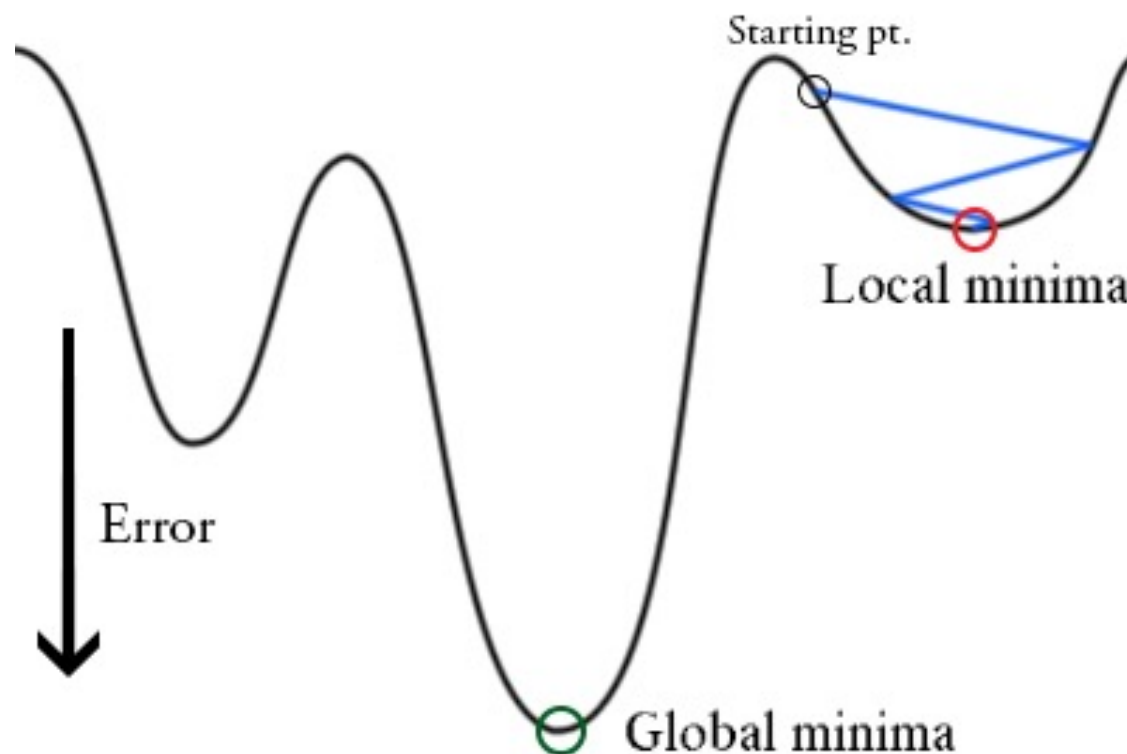
$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

3. Останавливаемся, если

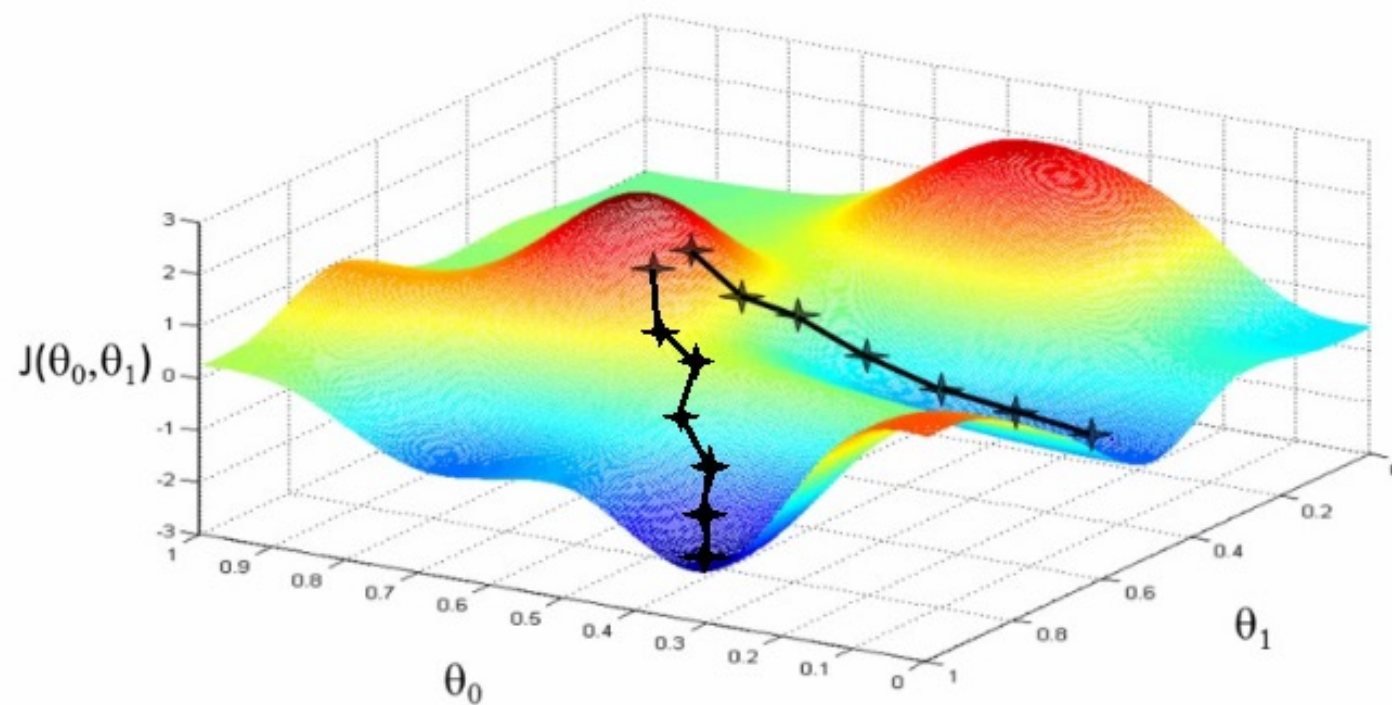
$$\|w^t - w^{t-1}\| < \varepsilon$$

# Локальные минимумы

- Градиентный спуск находит только локальные минимумы



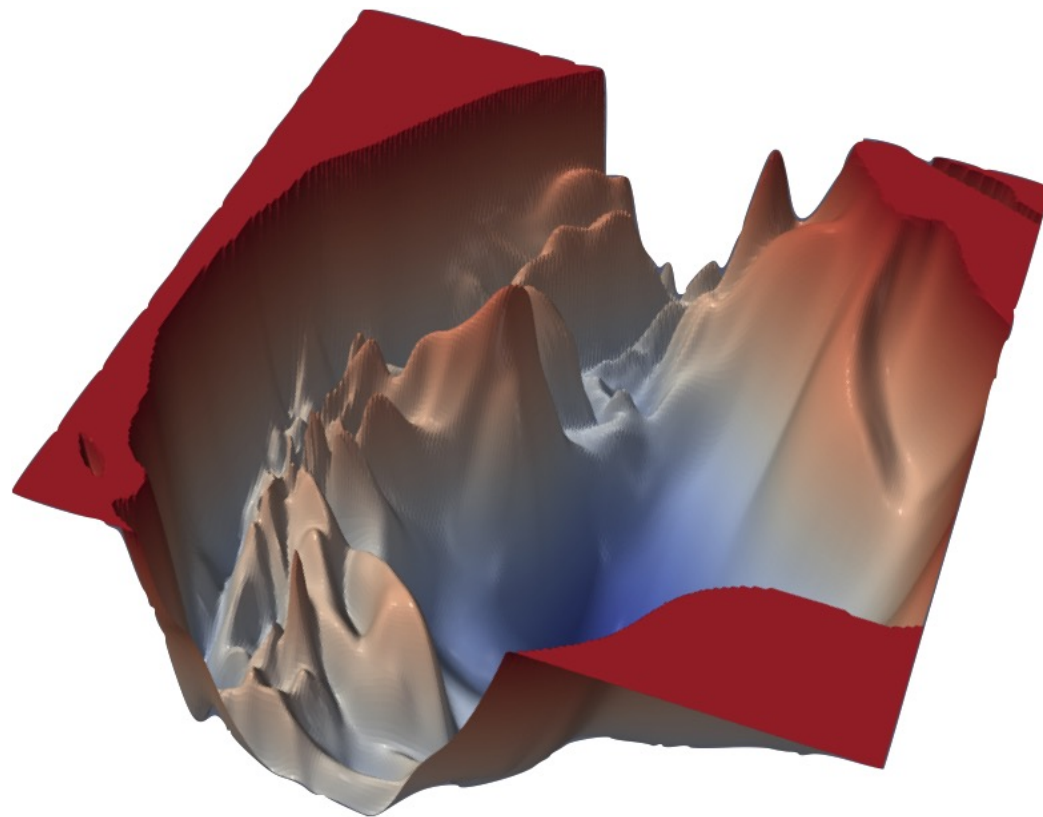
# Локальные минимумы



# Локальные минимумы

- Градиентный спуск находит **локальный минимум**
- Мультистарт — запуск градиентного спуска из разных начальных точек
- Может улучшить результат

# Локальные минимумы



# Длина шага

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

- Позволяет контролировать скорость обучения



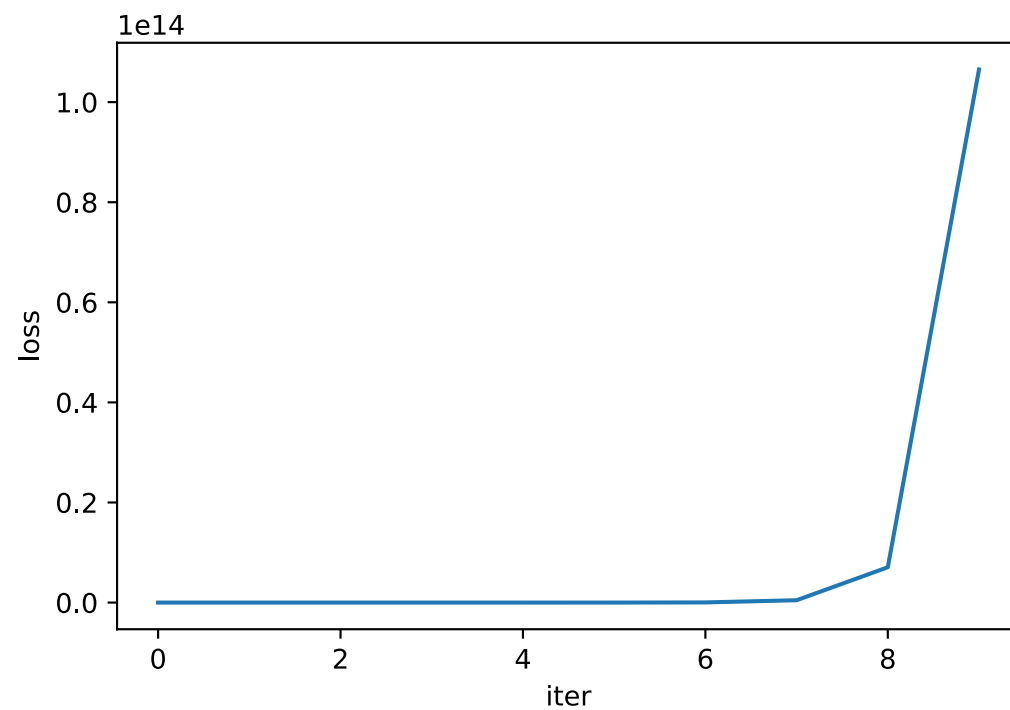
# Длина шага

```
[[ 0.8194022 -11.97609413 -34.41655678  0.98167246 -34.14405489]
 [ -2.83614512  17.19489715  3.29562399  63.8054227  39.70301275]
 [  3.10906179  11.26049837  0.51404712  22.64032379 -28.62078735]
 ...,
 [ -3.61976507  17.63933655  31.65890573  22.5124188 -75.6386039 ]
 [ -1.98472285  3.98588887  29.6135414 -11.11816  33.98746403]
 [ -3.34136103 -12.81955782 -19.5542601  12.62435442  50.24876879]]
```

# Длина шага

Градиент на первом шаге:

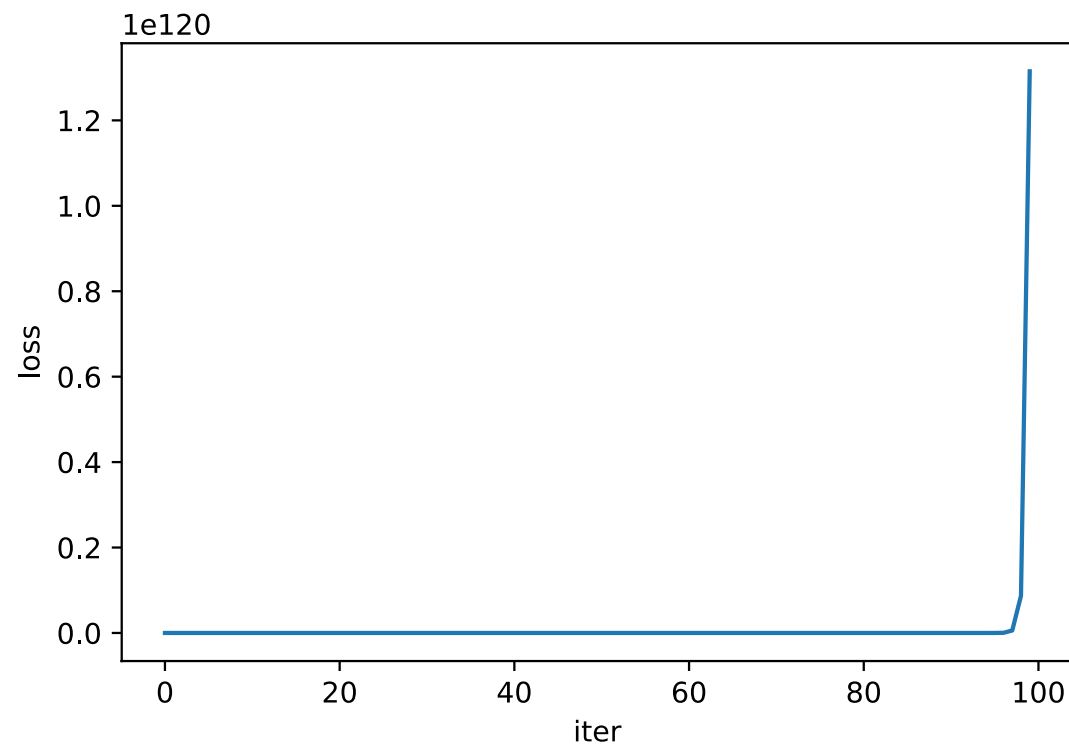
[ 26.52, 564.80, 682.90, 5097.71, 12110.87]



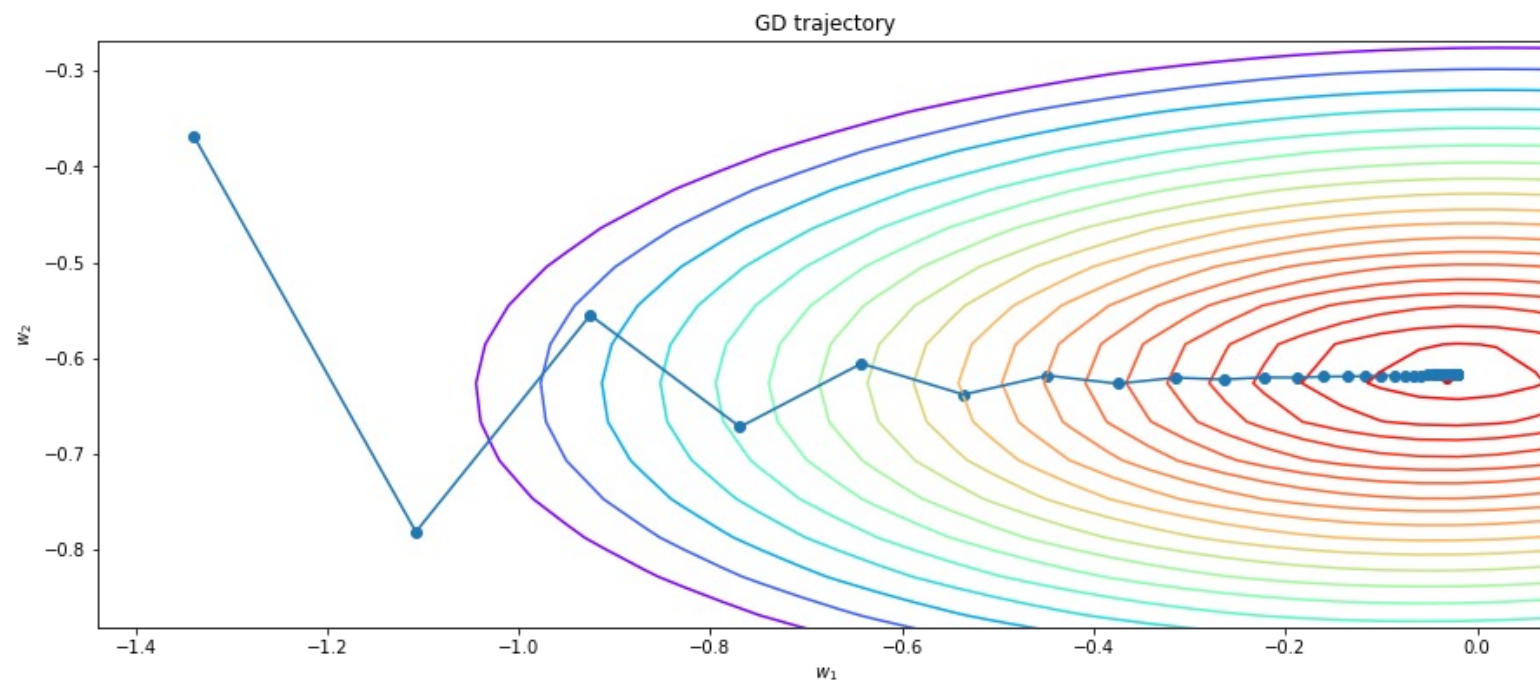
# Длина шага

Градиент на первом шаге:

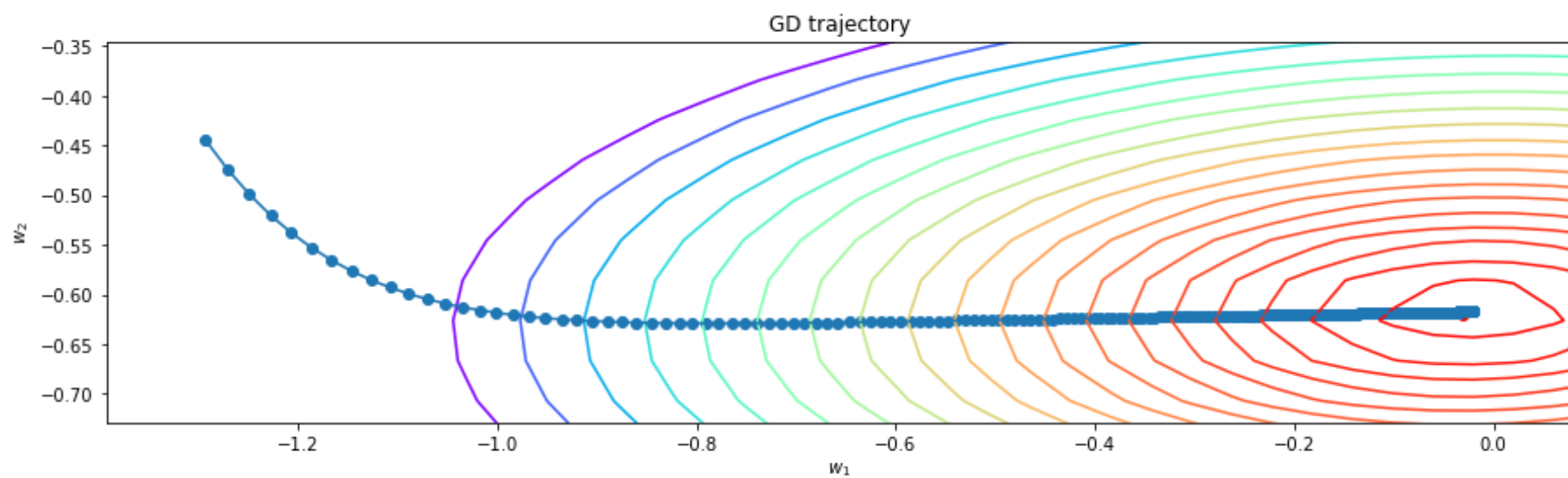
[ 26.52, 564.80, 682.90, 5097.71, 12110.87]



# Длина шага



# Длина шага



# Длина шага

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

- Позволяет контролировать скорость обучения
- Если сделать длину шага недостаточно маленькой, градиентный спуск может разойтись
- Длина шага — параметр, который нужно подбирать

# Переменная длина шага

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

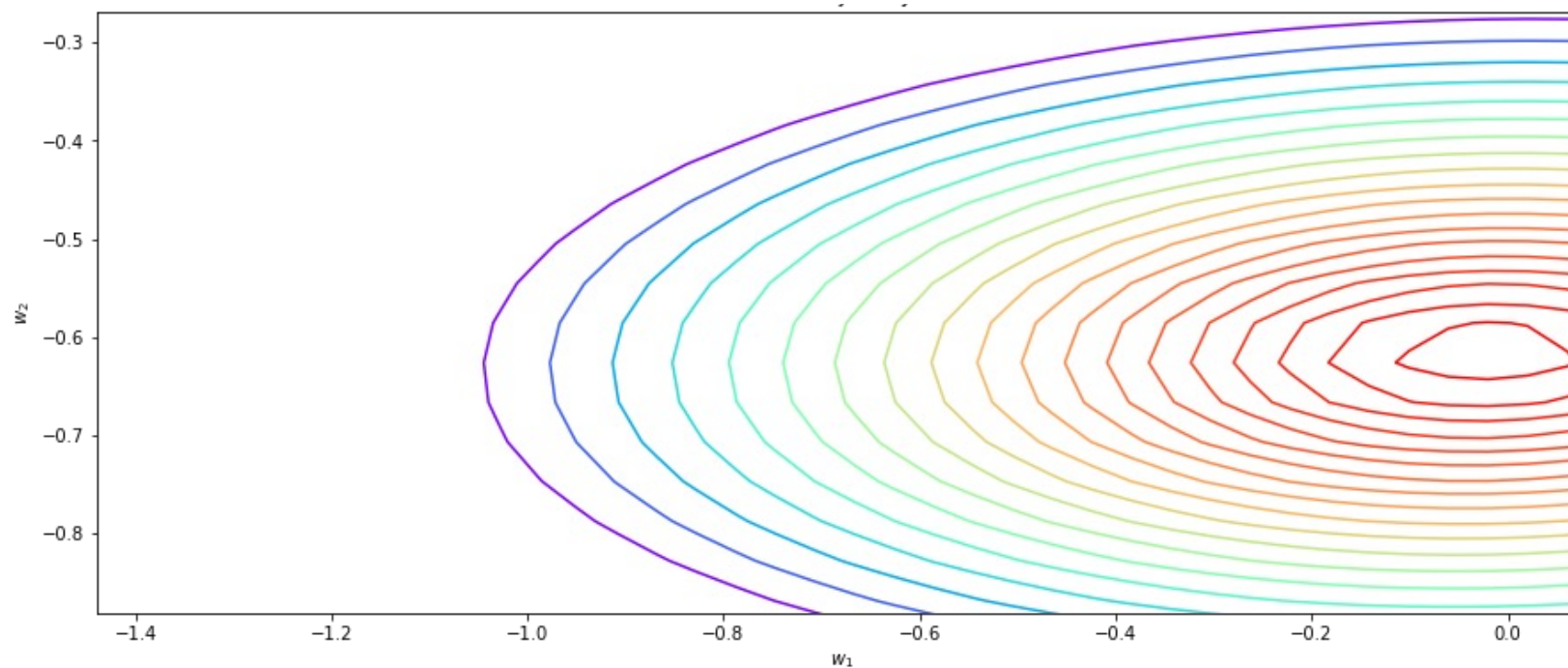
- Длину шага можно менять в зависимости от шага
- Например:  $\eta_t = \frac{1}{t}$
- Ещё вариант:  $\eta_t = \lambda \left( \frac{s}{s+t} \right)^p$

# Масштабирование признаков

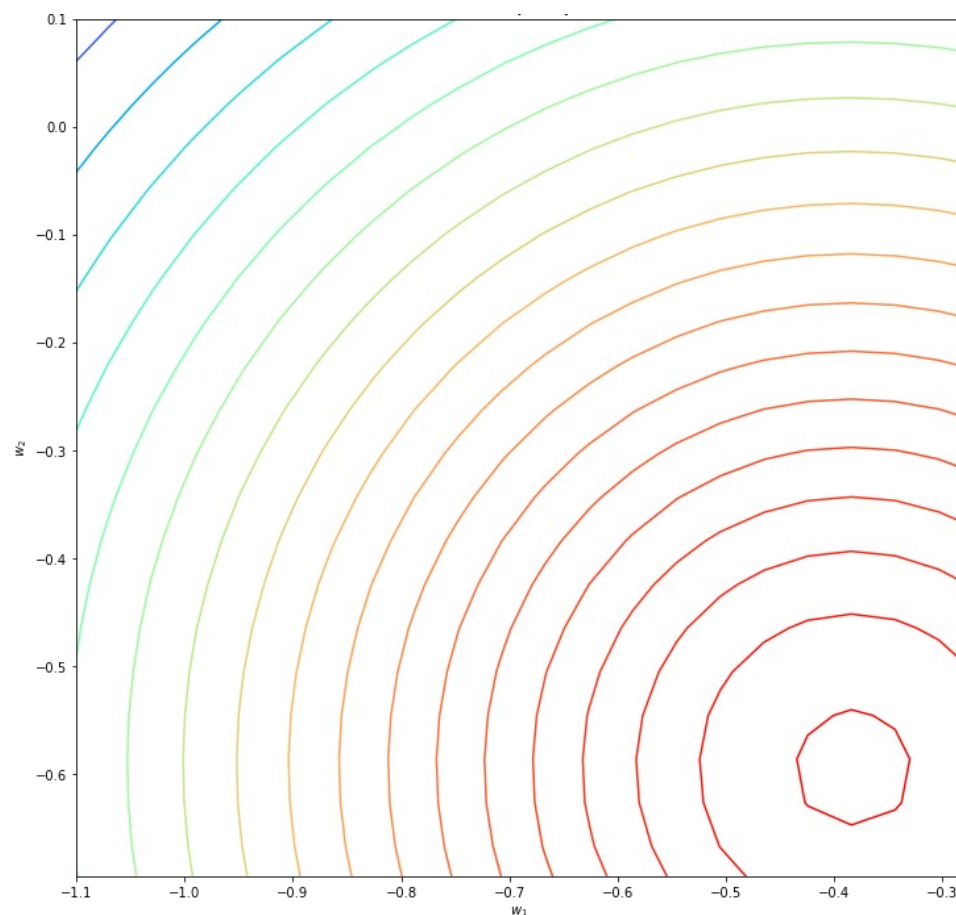
```
[[ 0.8194022 -11.97609413 -34.41655678  0.98167246 -34.14405489]
 [ -2.83614512  17.19489715  3.29562399  63.8054227  39.70301275]
 [ 3.10906179  11.26049837  0.51404712  22.64032379 -28.62078735]
 ...,
 [ -3.61976507  17.63933655  31.65890573  22.5124188 -75.6386039 ]
 [ -1.98472285  3.98588887  29.6135414 -11.11816  33.98746403]
 [ -3.34136103 -12.81955782 -19.5542601  12.62435442  50.24876879]]
```



# Масштабирование признаков



# Масштабирование признаков



# Масштабирование признаков

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

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

# 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

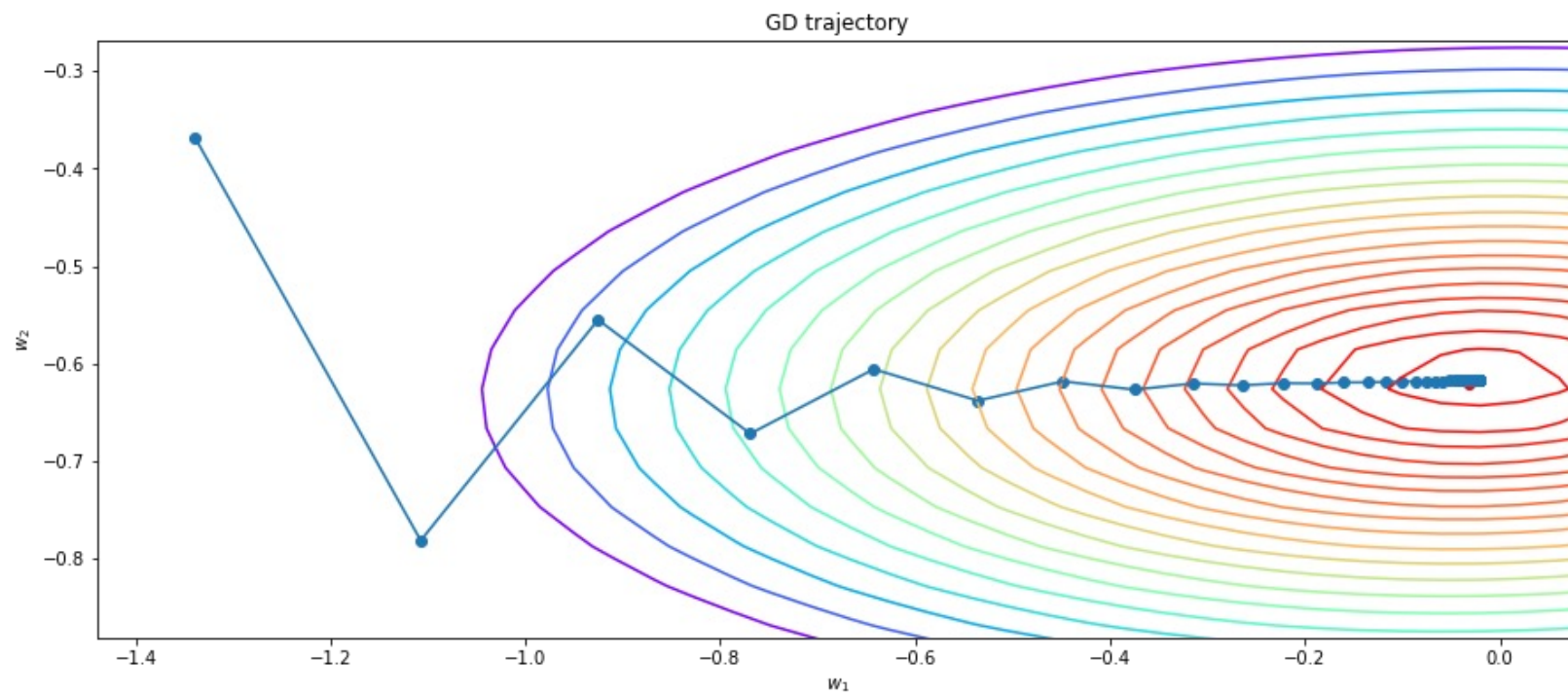
1. Начальное приближение:  $w^0$
2. Повторять, каждый раз выбирая случайный объект  $i_t$ :

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

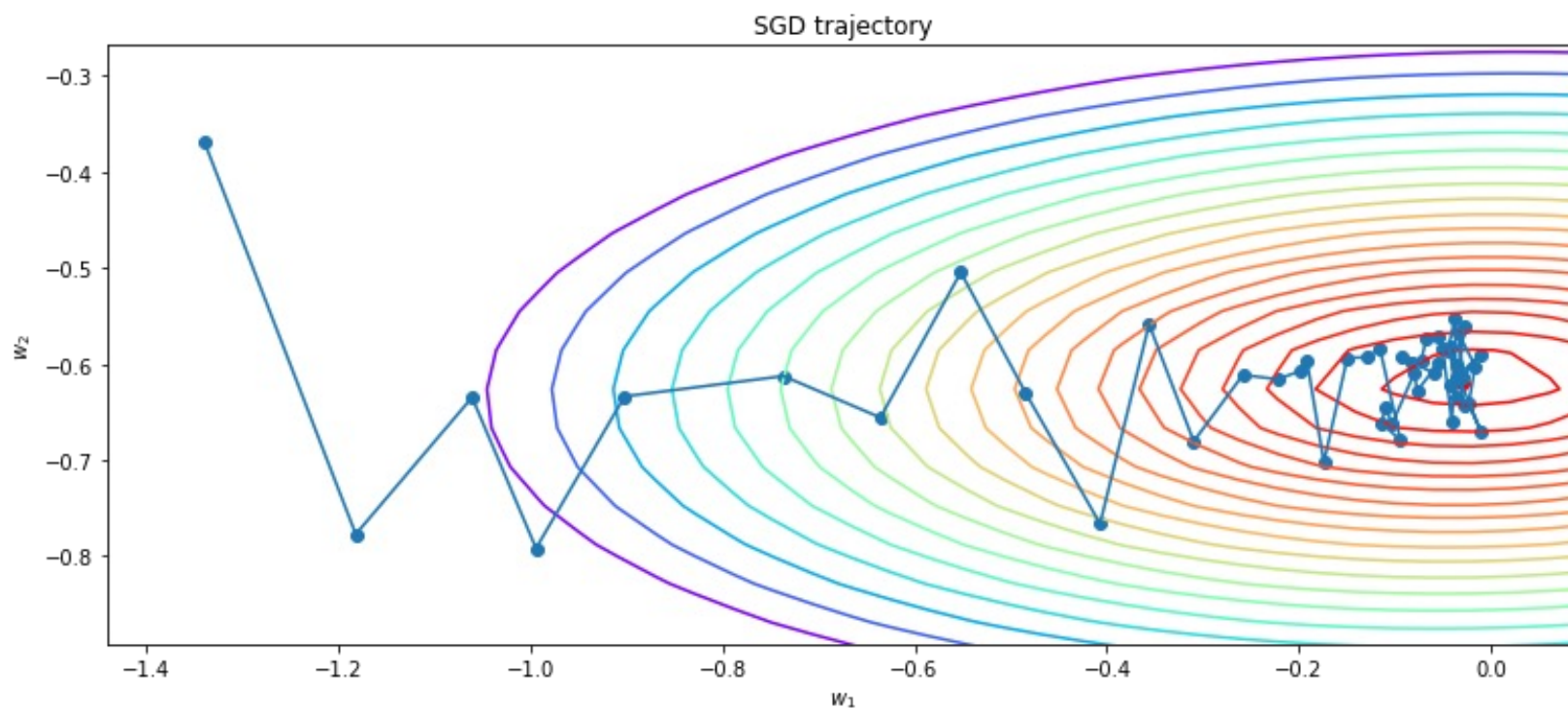
3. Останавливаемся, если

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

# Градиентный спуск



# Стохастический градиентный спуск



# Стохастический градиентный спуск

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

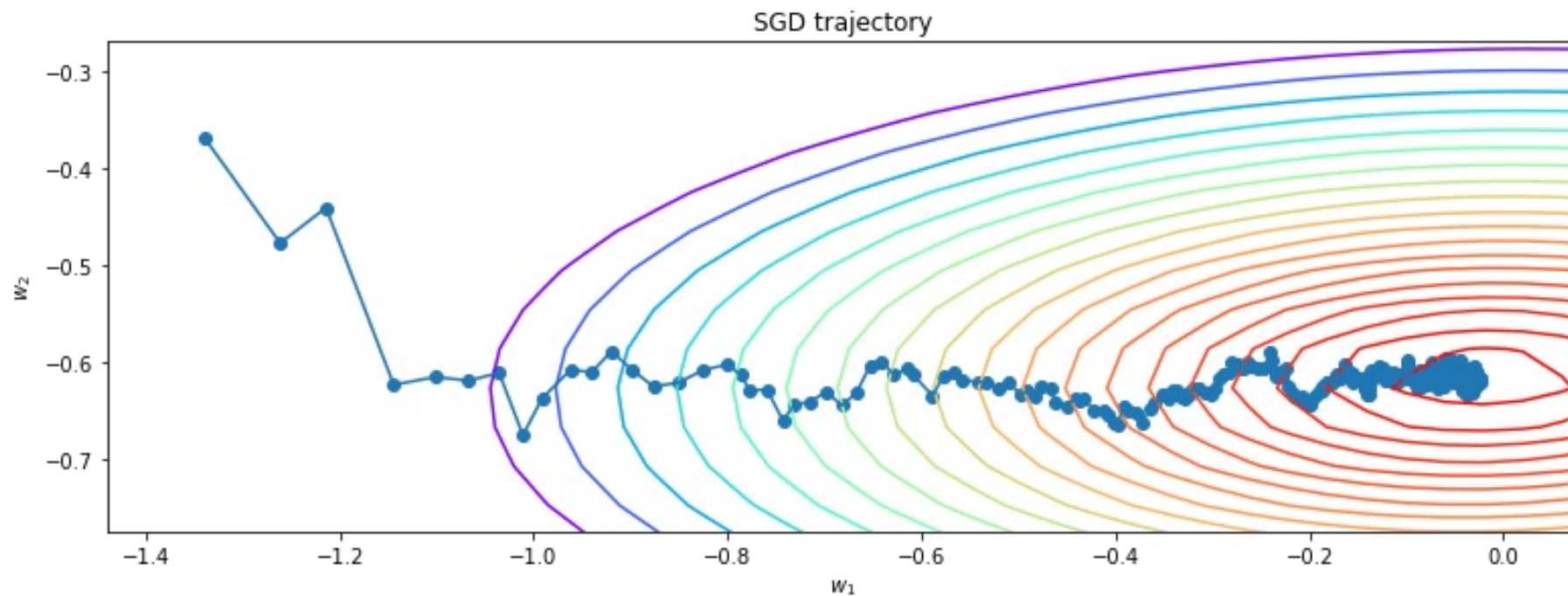
$$w^t = w^{t-1} - \eta_t \nabla L(y_{i_t}, a(x_{i_t}))$$

3. Останавливаемся, если

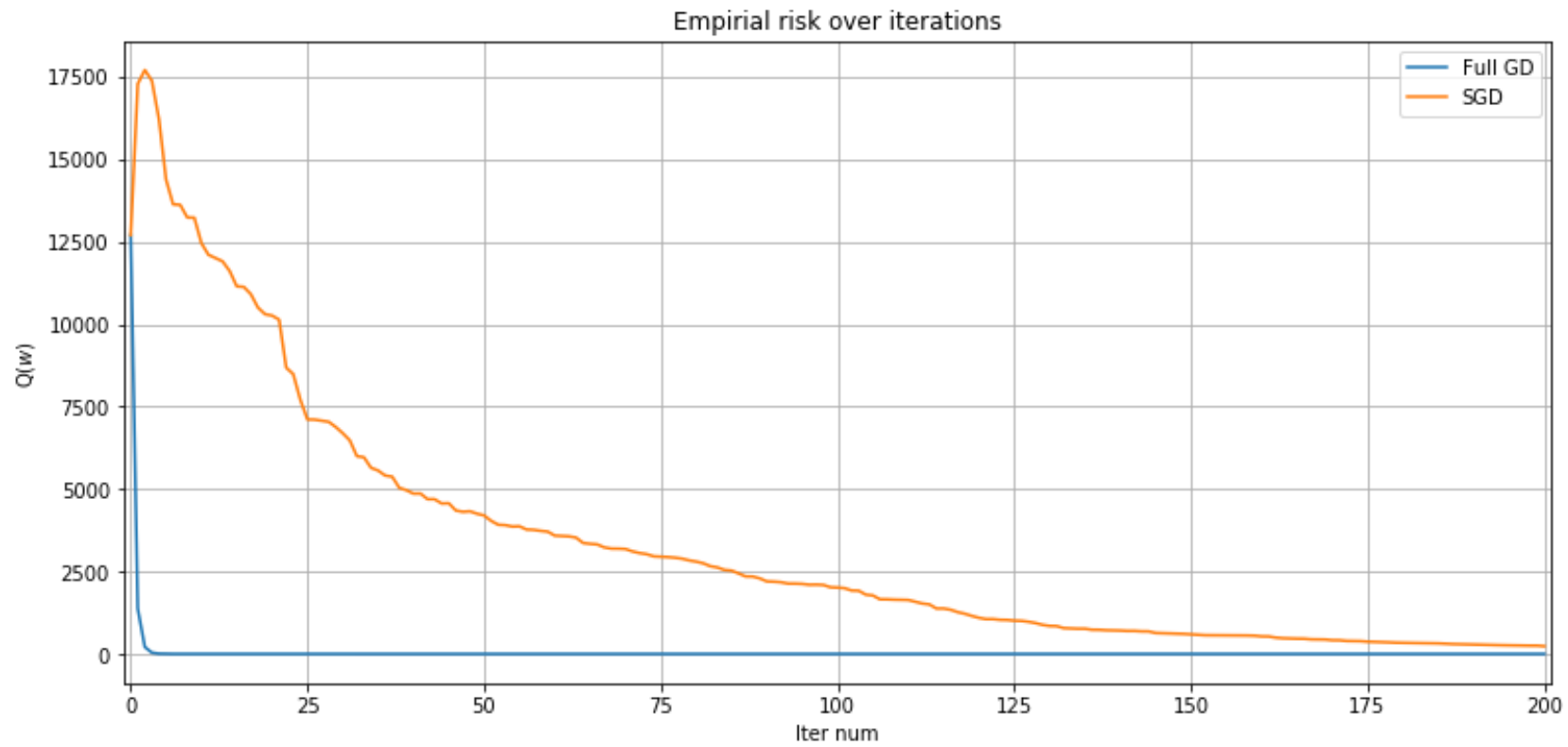
$$\|w^t - w^{t-1}\| < \varepsilon$$

# Стохастический градиентный спуск

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



# Стохастический градиентный спуск



# Mini-batch

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

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

3. Останавливаемся, если

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

# Функции потерь в задачах регрессии



# Среднеквадратичная ошибка

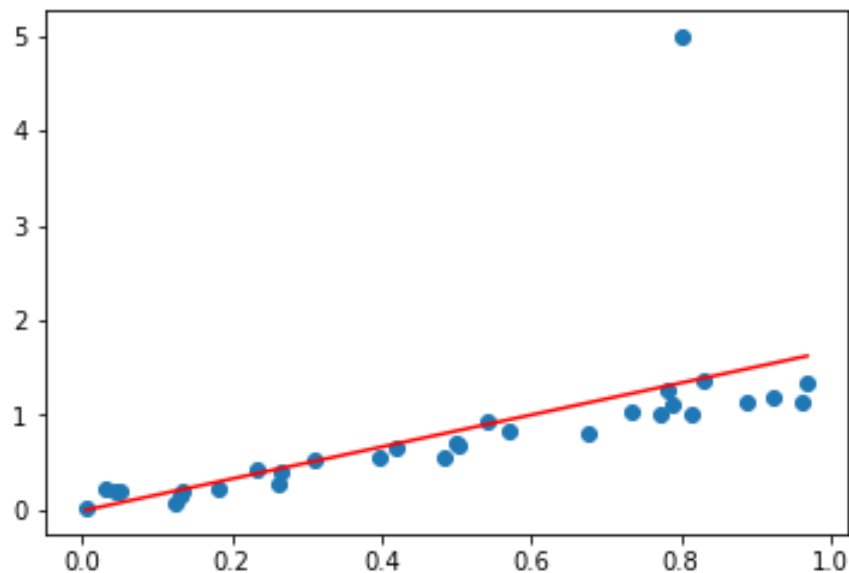
- Частый выбор — квадратичная функция потерь

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

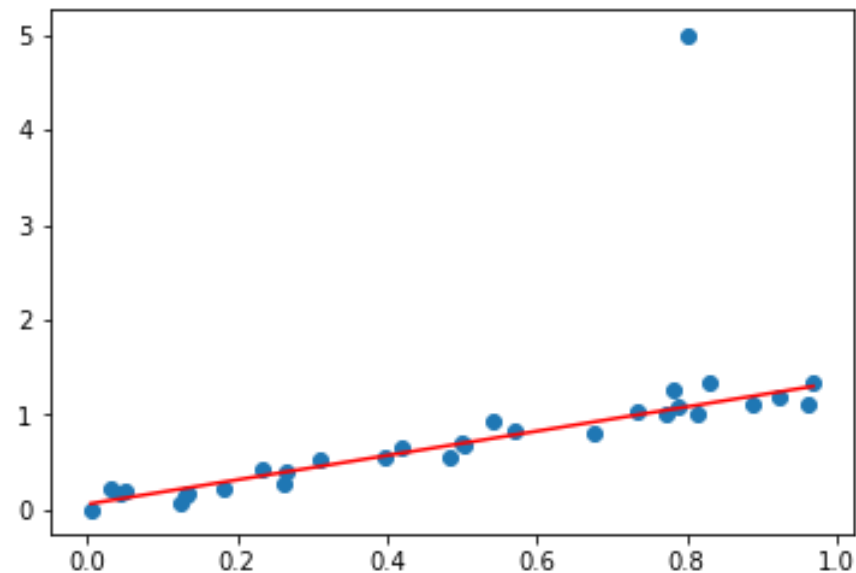
- Функционал ошибки — среднеквадратичная ошибка (mean squared error, MSE)

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

# 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

$$MSE \approx 1164$$

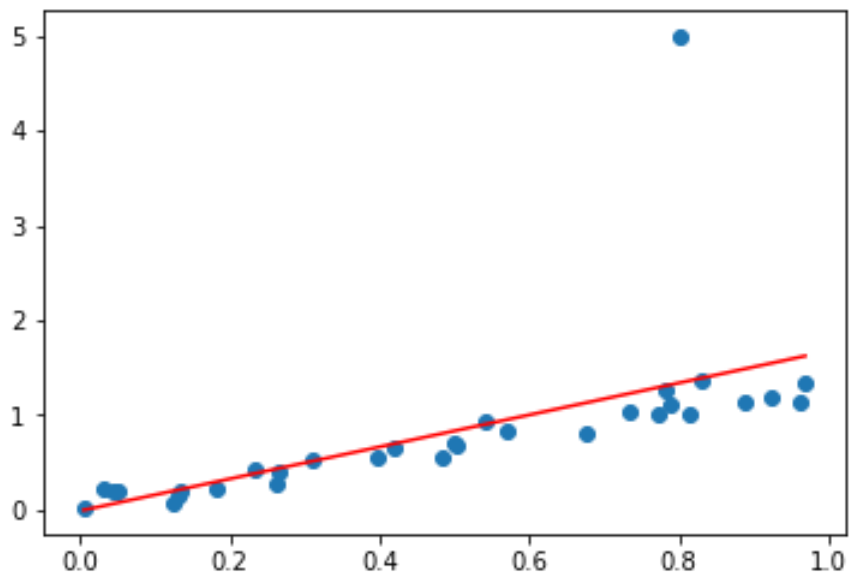
# Средняя абсолютная ошибка

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

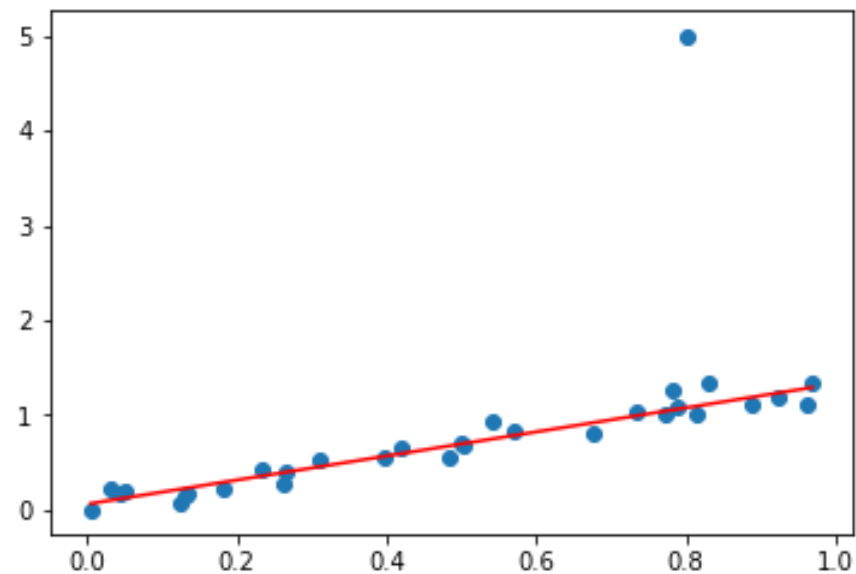
- Функционал ошибки — средняя абсолютная ошибка (mean absolute error, MAE)

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

# 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

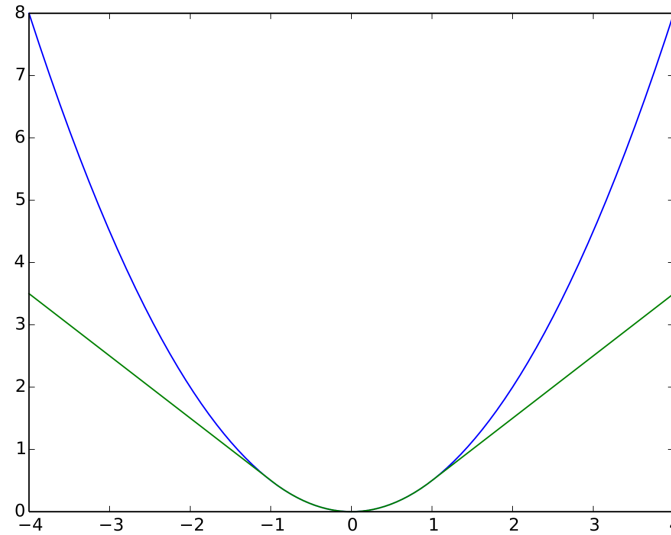
$$L_H(y, a) = \begin{cases} \frac{1}{2}(y - a)^2, & |y - a| < \delta \\ \delta \left( |y - a| - \frac{1}{2}\delta \right), & |y - a| \geq \delta \end{cases}$$

- 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)

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

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

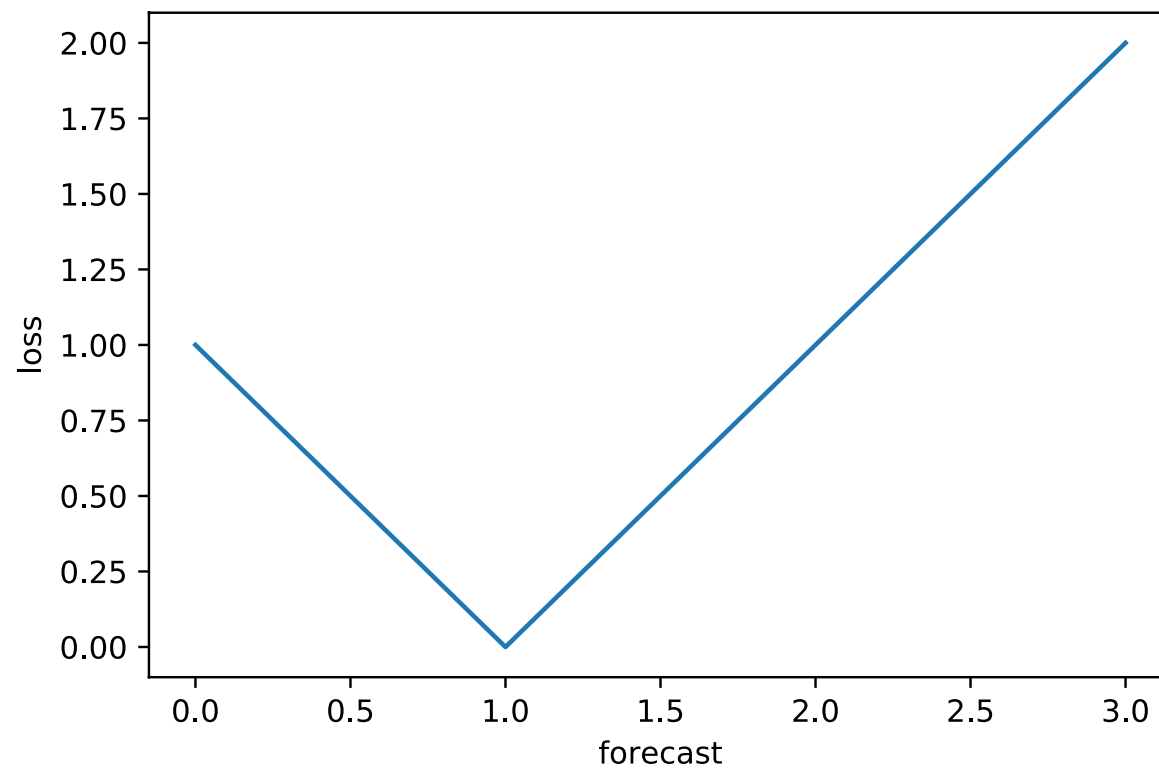
# MAPE

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

- Особенности (при  $a \geq 0$ ):
- Недопрогноз штрафует 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

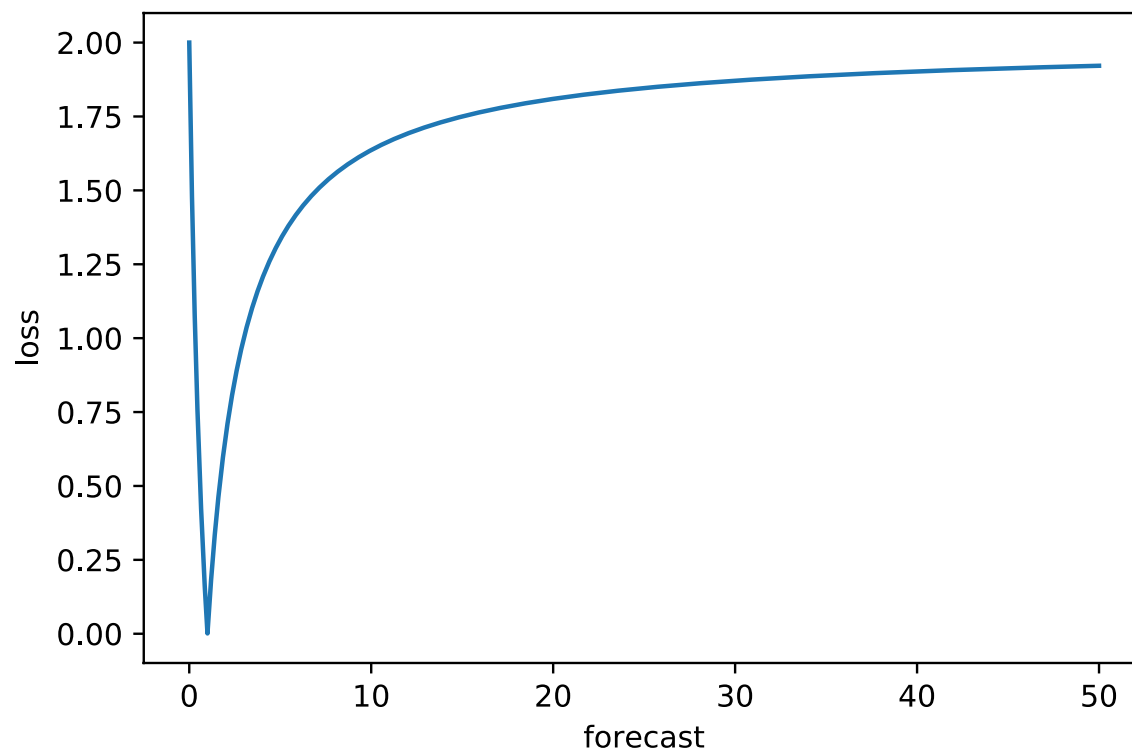
- Symmetric Mean Absolute Percentage Error (симметричный средний модуль относительной ошибки)

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

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

# SMAPE

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



# Модель линейной классификации



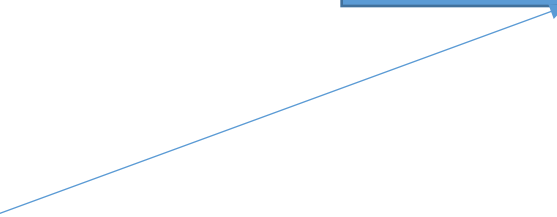
# Классификация

- $\mathbb{Y} = \{-1, +1\}$
- $-1$  — отрицательный класс
- $+1$  — положительный класс
- $a(x)$  должен возвращать одно из двух чисел

# Линейная регрессия

$$a(x) = w_0 + \sum_{j=1}^d w_j x_j$$

Вещественное  
число!



# Линейный классификатор

$$a(x) = \text{sign} \left( w_0 + \sum_{j=1}^d w_j x_j \right)$$

# Линейный классификатор

$$a(x) = \text{sign} \left( w_0 + \sum_{j=1}^d w_j x_j \right)$$

Свободный  
коэффициент

Веса

Признаки

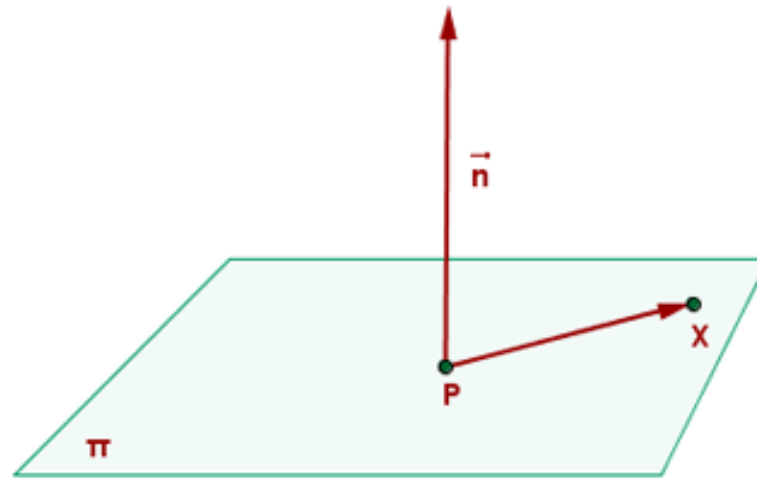
# Линейный классификатор

- Будем считать, что есть единичный признак

$$a(x) = \text{sign} \sum_{j=1}^d w_j x_j = \text{sign} \langle w, x \rangle$$

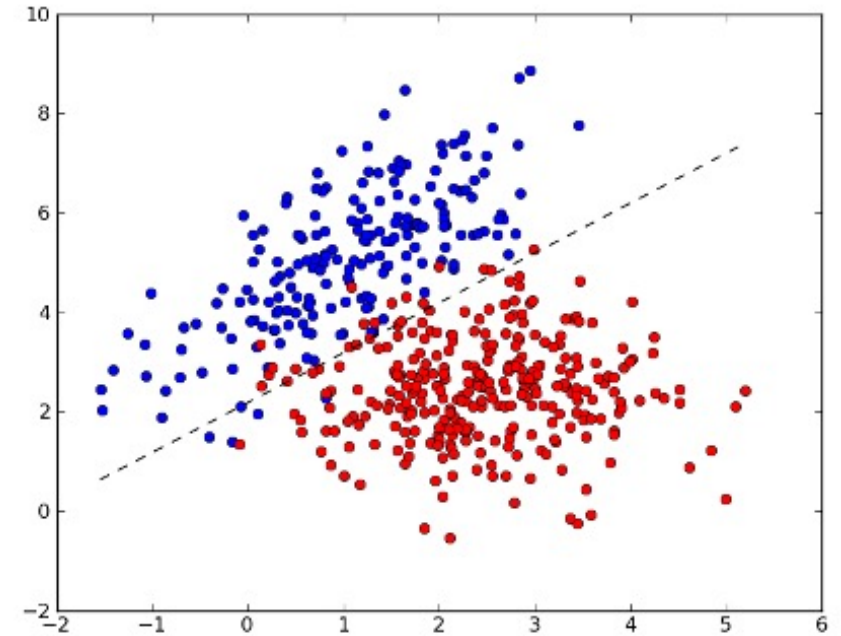
# Геометрия линейного классификатора

Уравнение гиперплоскости:  $\langle w, x \rangle = 0$



# Геометрия линейного классификатора

- Линейный классификатор проводит гиперплоскость
- $\langle w, x \rangle < 0$  — объект «слева» от неё
- $\langle w, x \rangle > 0$  — объект «справа» от неё



# Геометрия линейного классификатора

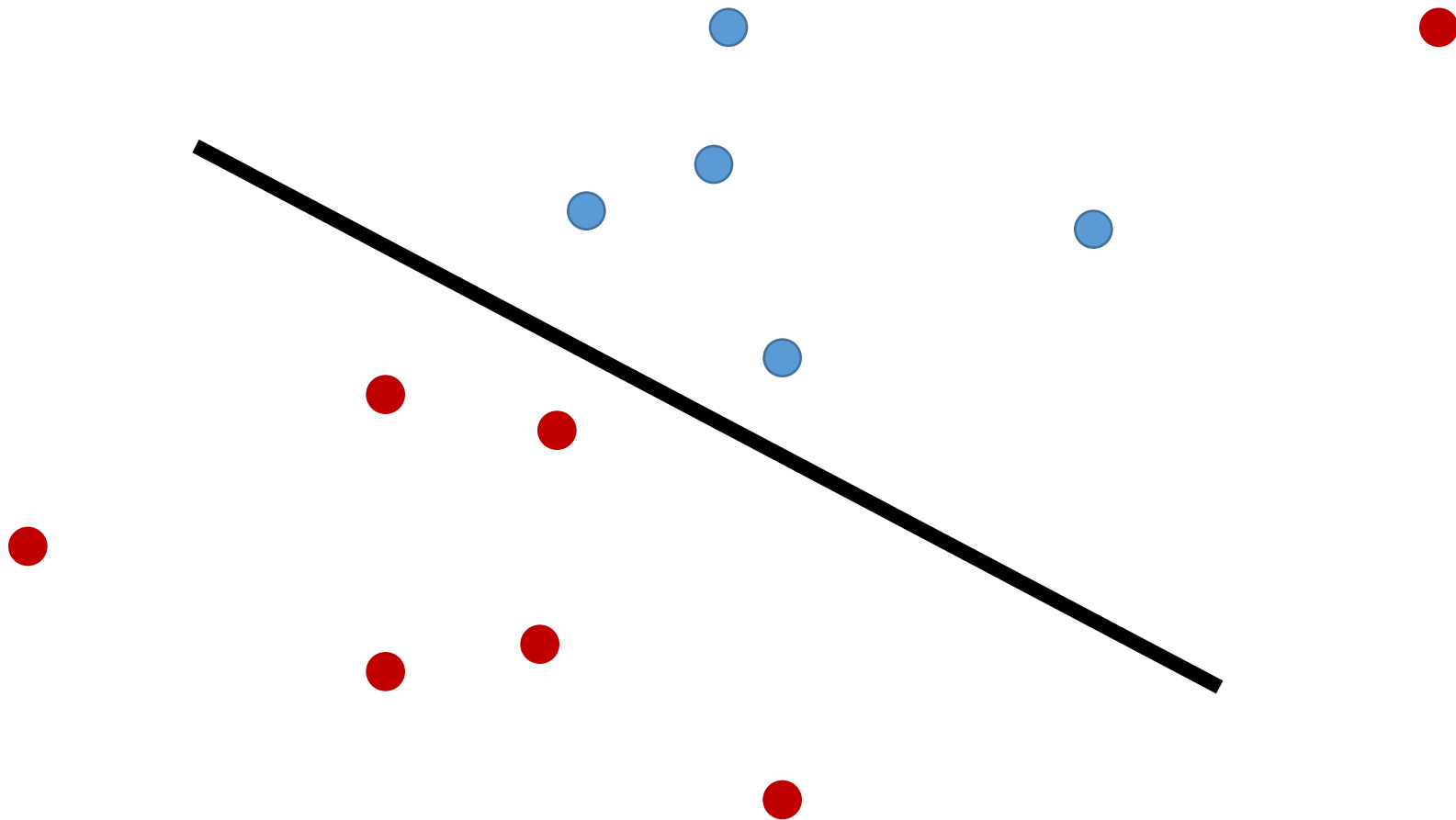
- Расстояние от точки до гиперплоскости  $\langle w, x \rangle = 0$ :

$$\frac{|\langle w, x \rangle|}{\|w\|}$$

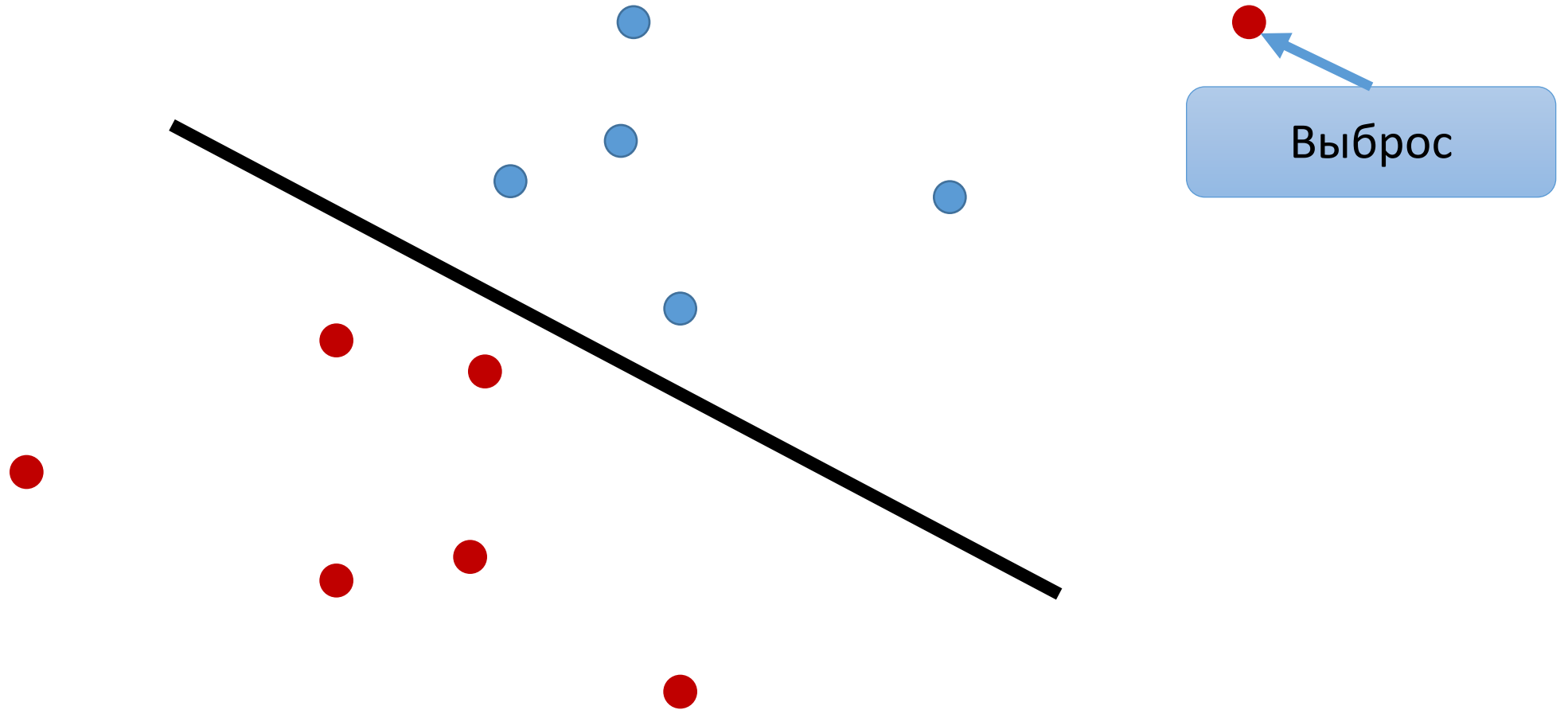
- Чем больше  $\langle w, x \rangle$ , тем дальше объект от разделяющей гиперплоскости



# Геометрия линейного классификатора

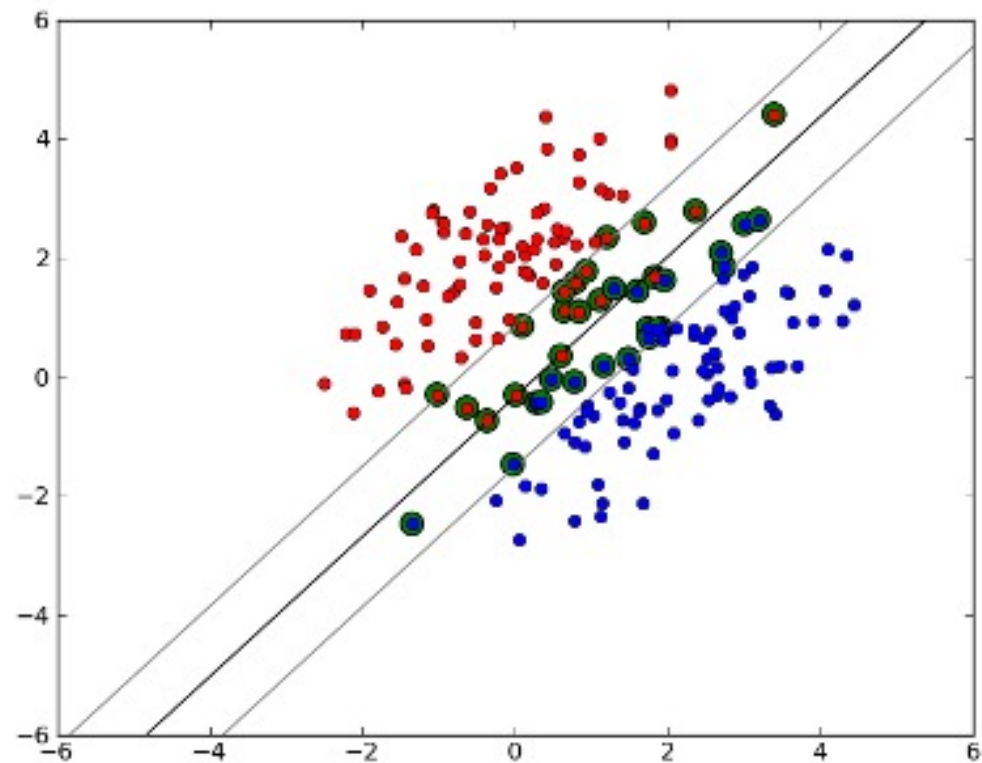


# Геометрия линейного классификатора



# Отступ

- $M_i = y_i \langle w, x_i \rangle$
- $M_i > 0$  — классификатор дает верный ответ
- $M_i < 0$  — классификатор ошибается
- Чем дальше отступ от нуля, тем больше уверенности



# Порог

$$a(x) = \text{sign}(\langle w, x \rangle - t)$$

- $t$  — порог классификатора
- Можно подбирать для оптимизации функции потерь, отличной от использованной при обучении

# Линейный классификатор

- Линейный классификатор разделяет два класса гиперплоскостью
- Чем больше отступ по модулю, тем дальше объект от гиперплоскости
- Знак отступа говорит о корректности предсказания