

# Введение в анализ данных

## Лекция 7

### Градиентный спуск и линейная классификация

Евгений Соколов

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

НИУ ВШЭ, 2021

О проверочной

# Способ проверки модели 1

- Обучающая выборка: подбираем параметры
- Валидационная выборка: подбираем гиперпараметры и сравниваем разные модели
- Тестовая выборка: проверяем итоговую модель



# Способ проверки модели 2

- Кросс-валидация по обучающей выборке: подбираем параметры и гиперпараметры, сравниваем модели
- Тестовая выборка: проверяем итоговую модель



# Обучение линейной регрессии

$$Q(w_1, \dots, w_d) = \sum_{i=1}^{\ell} (\textcolor{red}{w}_1 x_1 + \dots + \textcolor{red}{w}_d x_d - y_i)^2 \rightarrow \min_{w_1, \dots, w_d}$$

И аналитическая формула, и градиентный спуск решают **одну и ту же задачу!**

# Обучение линейной регрессии

Что может пойти не так:

- Градиентный спуск может не дойти до минимума
  - Слишком рано остановим
  - Слишком медленные шаги
- У задачи много решений (случается, если есть линейно зависимые признаки)
  - По аналитической формуле нельзя будет посчитать веса
  - Градиентный спуск придёт в один из локальных минимумов

# Параметры и гиперпараметры

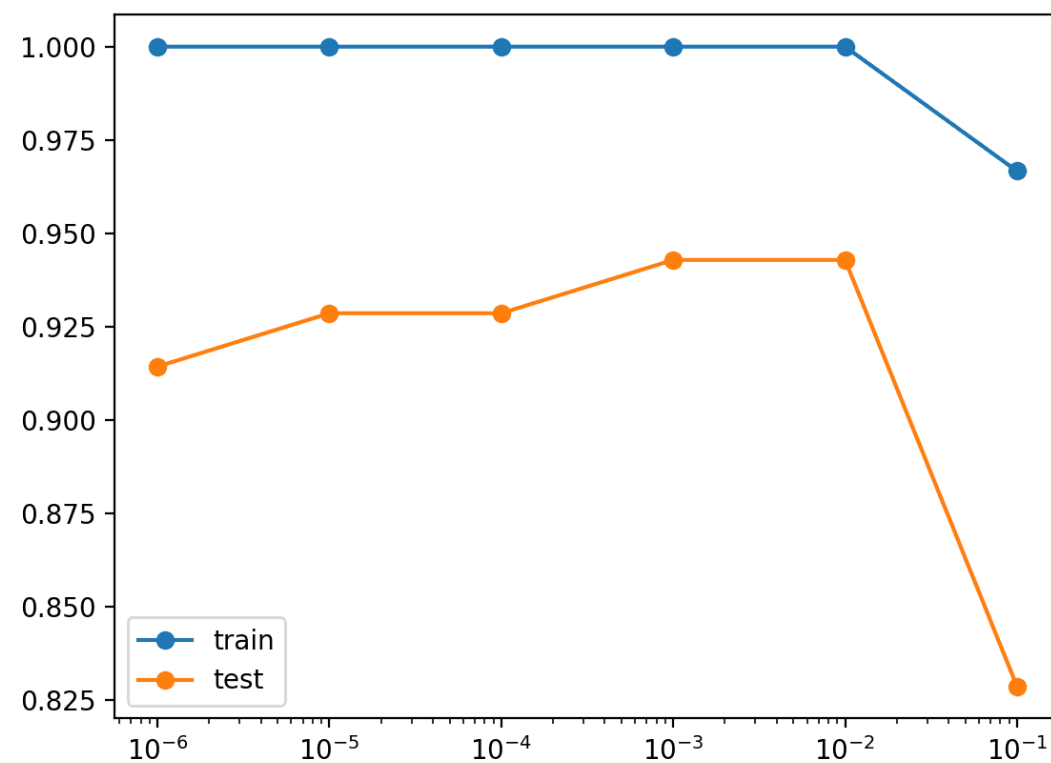
- Параметры нужны, чтобы подогнать модель под данные
  - Выбираем так, чтобы ошибка на обучающей выборке была как можно меньше
- Гиперпараметры нужны, чтобы контролировать сложность модели
  - С точки зрения обучающей выборки они только мешают
  - Чем сильнее регуляризация, тем хуже модель на обучении
  - Зато они позволяют бороться с переобучением
  - Подбираются на кросс-валидации или по отложенной выборке

# Поиск гиперпараметров

- Хотим найти коэффициент регуляризации
- Выбираем сетку: [0.01, 0.1, 1, 10, 100]
- Для каждого значения обучаем модель и считаем ошибку на отложенной выборке (или про кросс-валидации)
- Выбираем вариант с наименьшей ошибкой
  
- Grid search
- Есть и другие подходы



# Поиск гиперпараметров



# Градиентный спуск

- Подбирает веса в линейной модели
- Если заменить движение по антиградиенту на движение в случайную сторону, то получится непонятно что

# 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  $i_t$ :

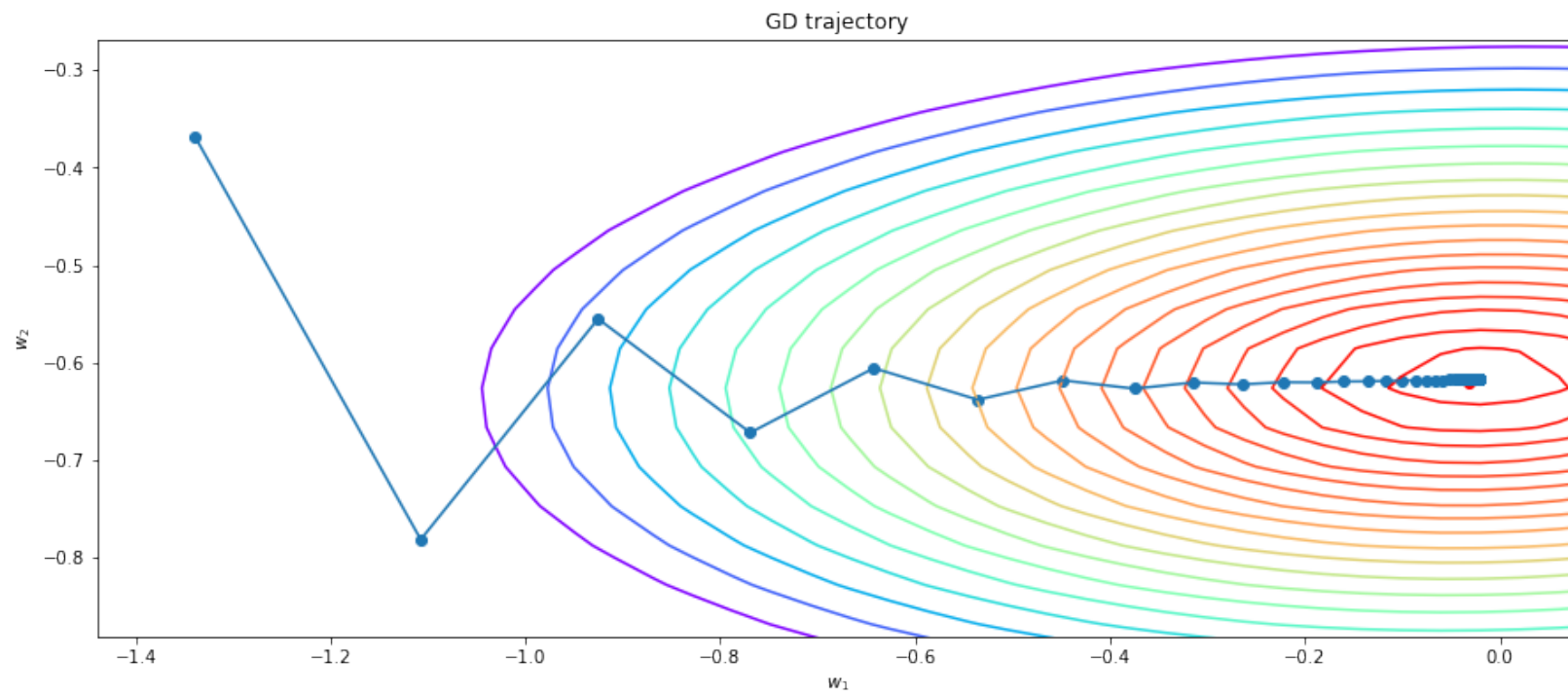
$$w^t = w^{t-1} - \eta \nabla L(y_{i_t}, a(x_{i_t}))$$

3. Останавливаемся, если

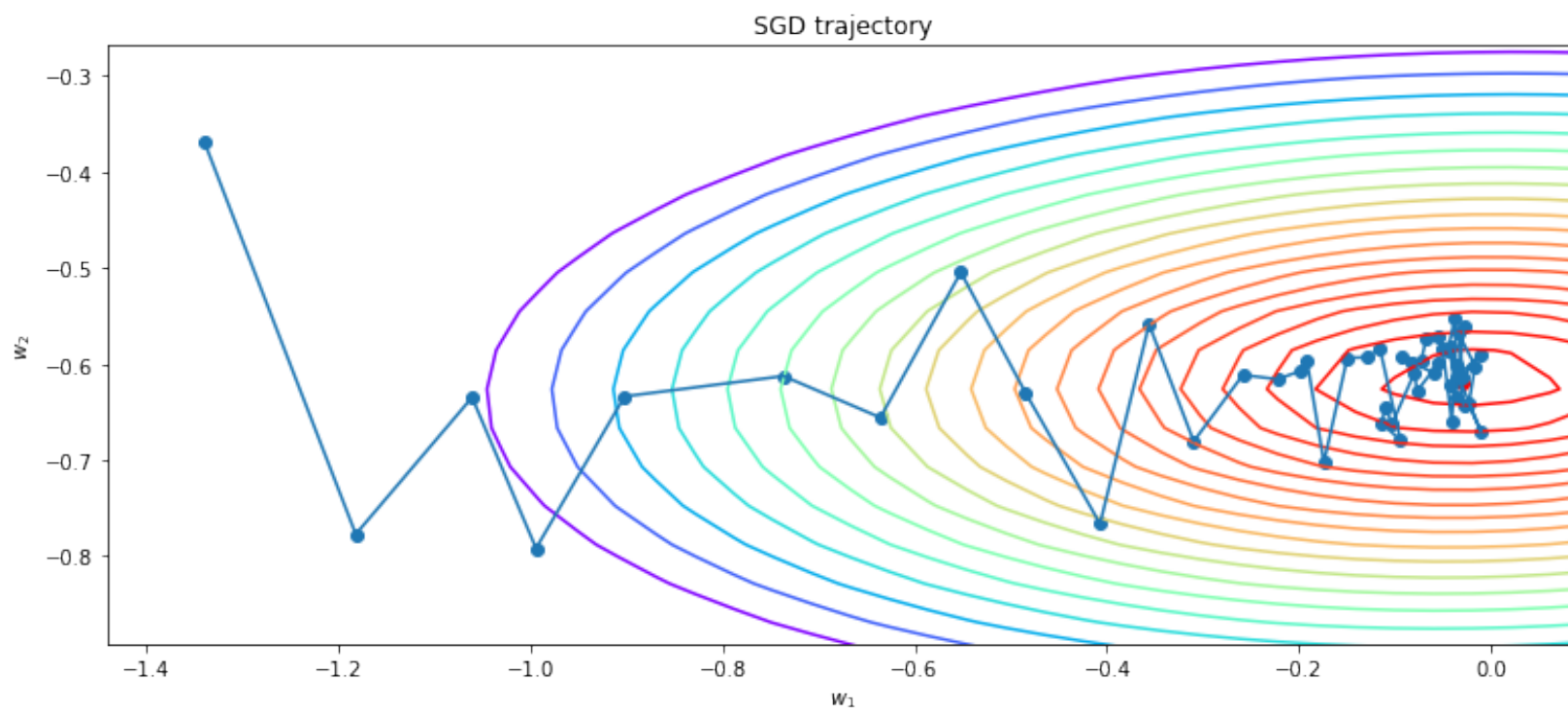
$$\|w^t - w^{t-1}\| < \varepsilon$$



# Градиентный спуск



# Стохастический градиентный спуск



# Стохастический градиентный спуск

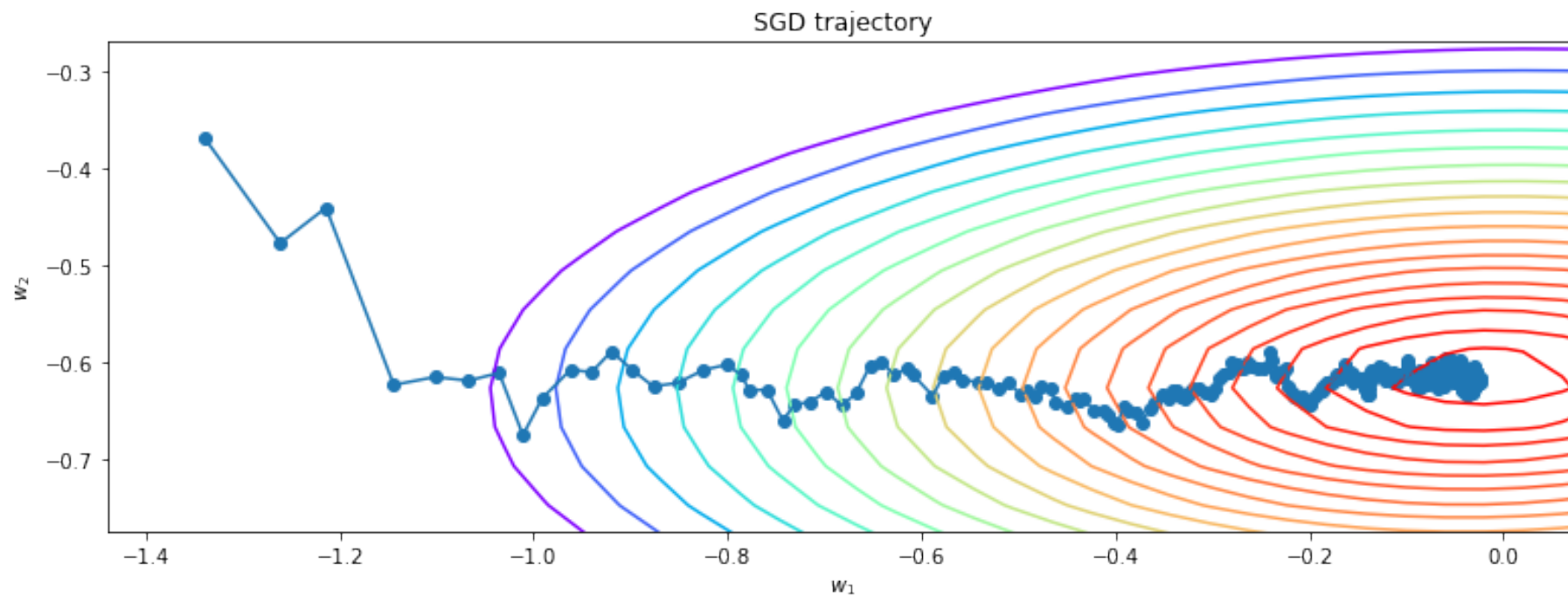
1. Начальное приближение:  $w^0$
2. Повторять, каждый раз выбирая случайный объект  $i_t$ :

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

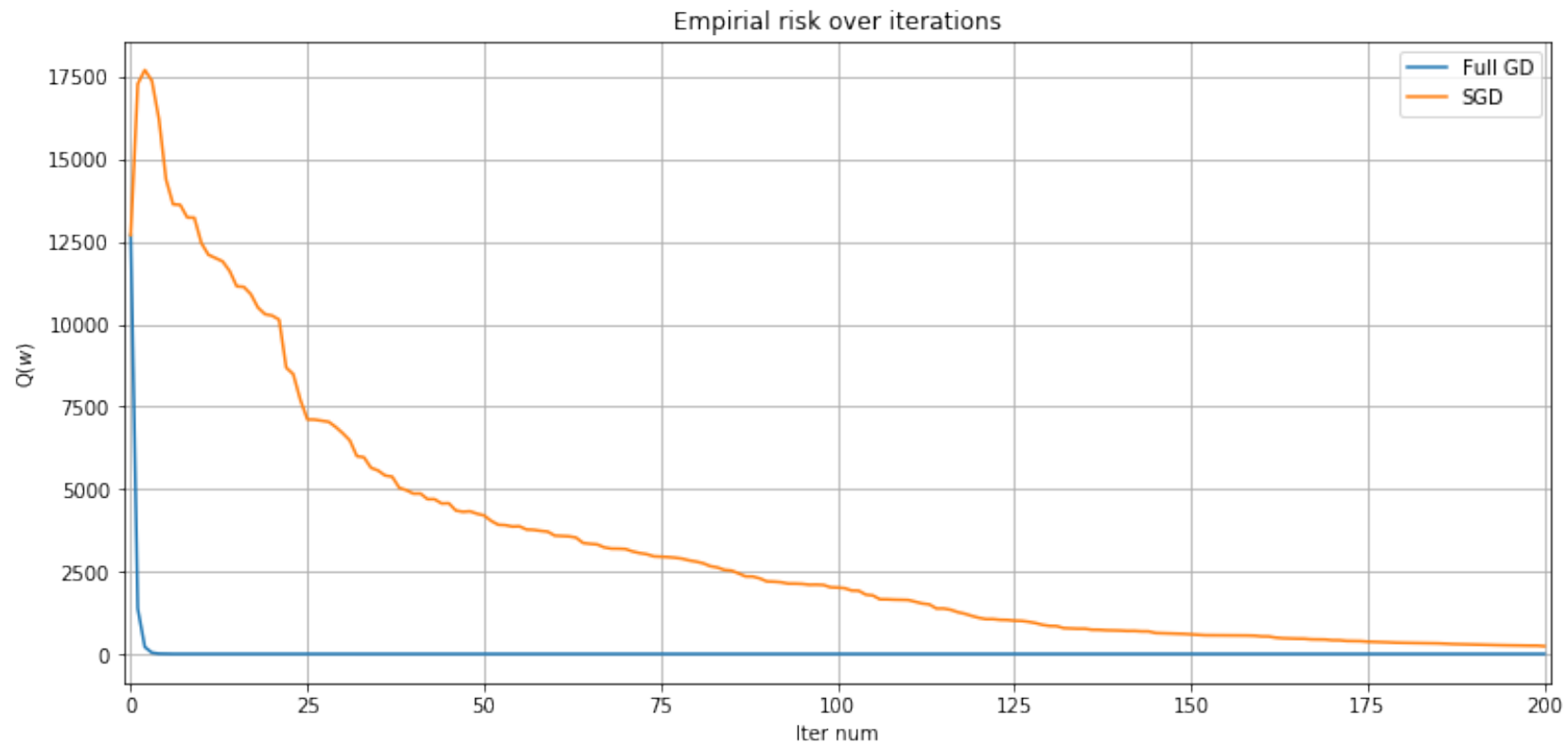
3. Останавливаемся, если ошибка на валидационной выборке перестала падать

# 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

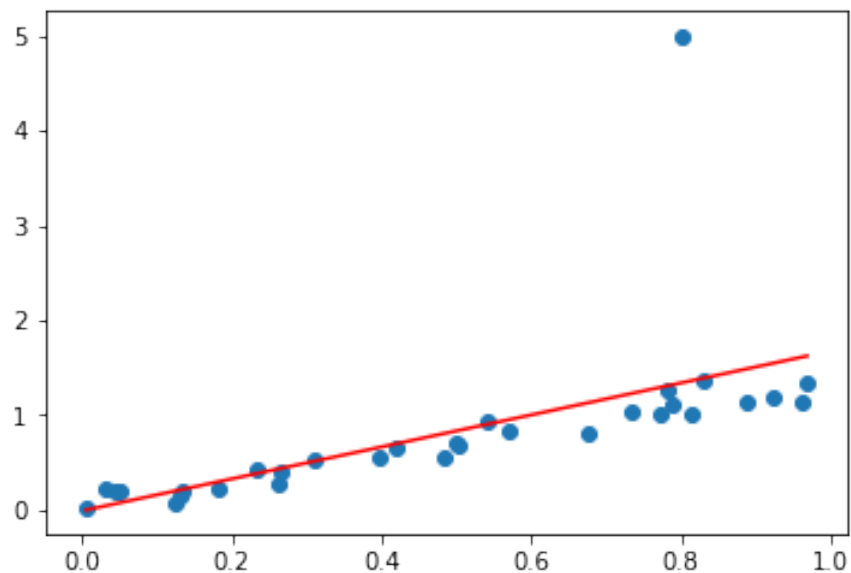
$$L(y, a) = (a - y)^2$$

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

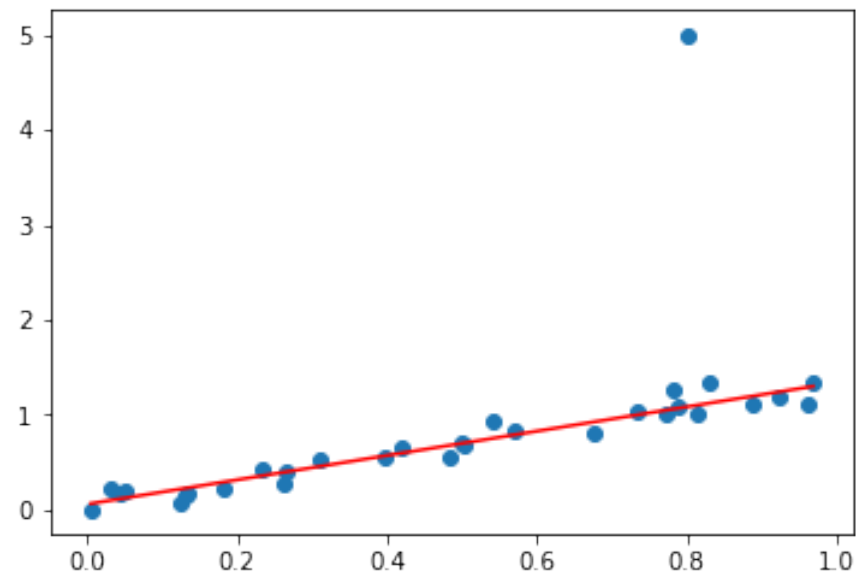
$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$



# 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

$$MSE \approx 1164$$

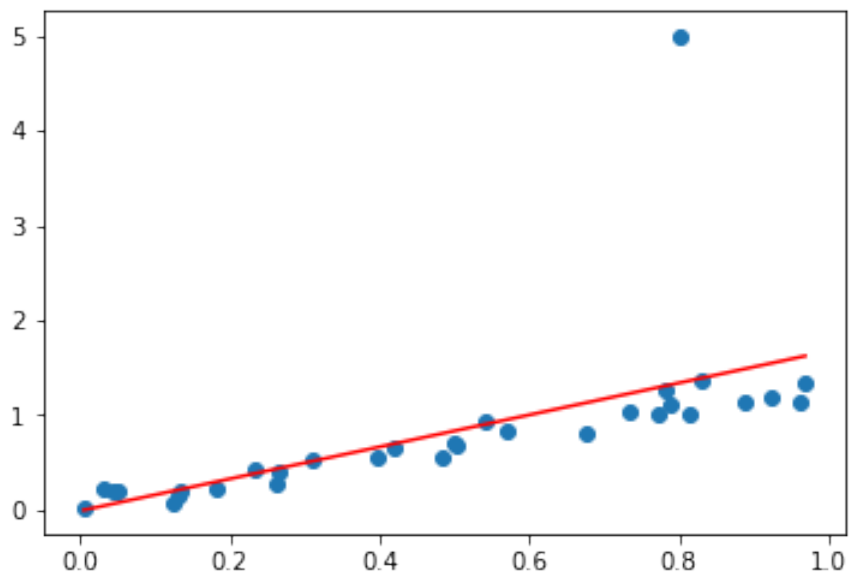
# Средняя абсолютная ошибка

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

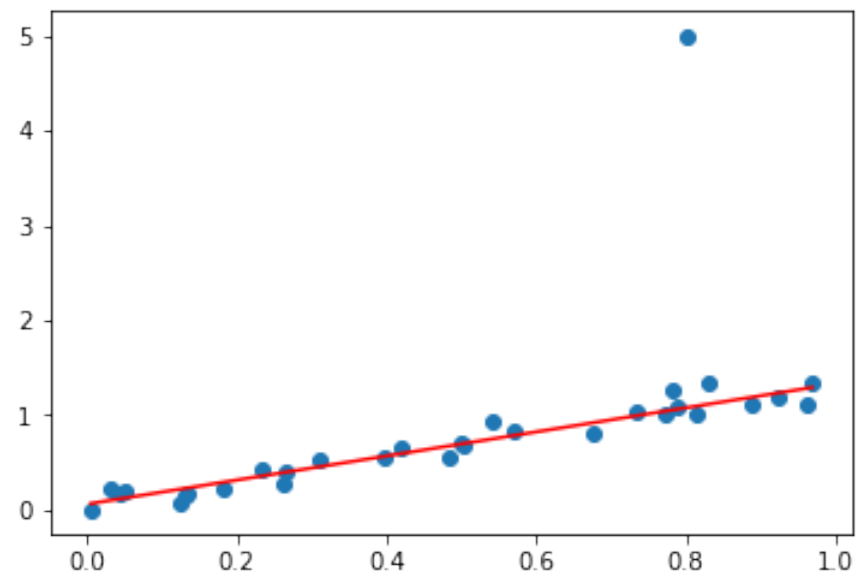
- Функционал ошибки — средняя абсолютная ошибка (mean absolute error, MAE)

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

# 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

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

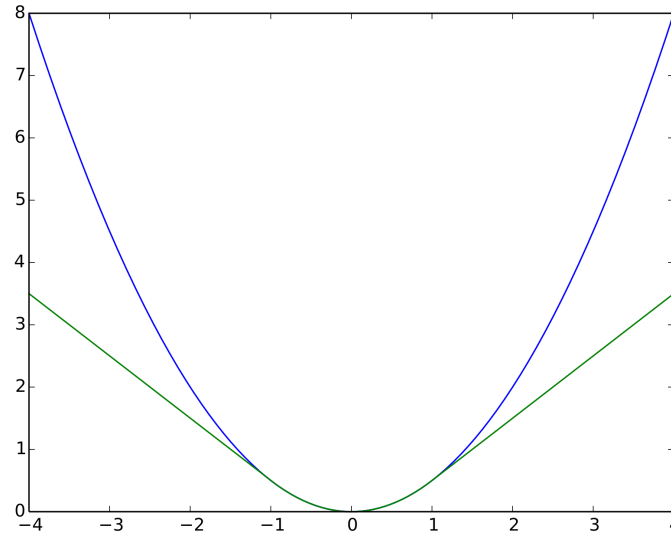
- Функционал ошибки:

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



# 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)

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

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

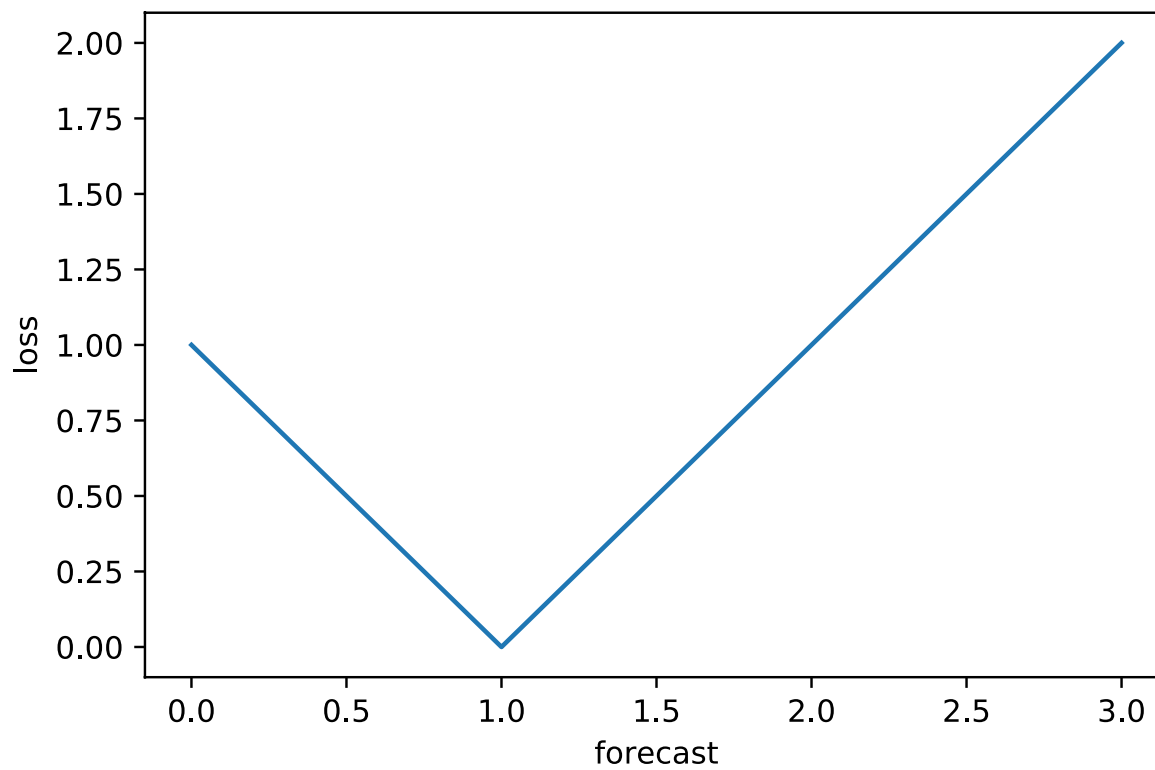
# MAPE

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

- 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

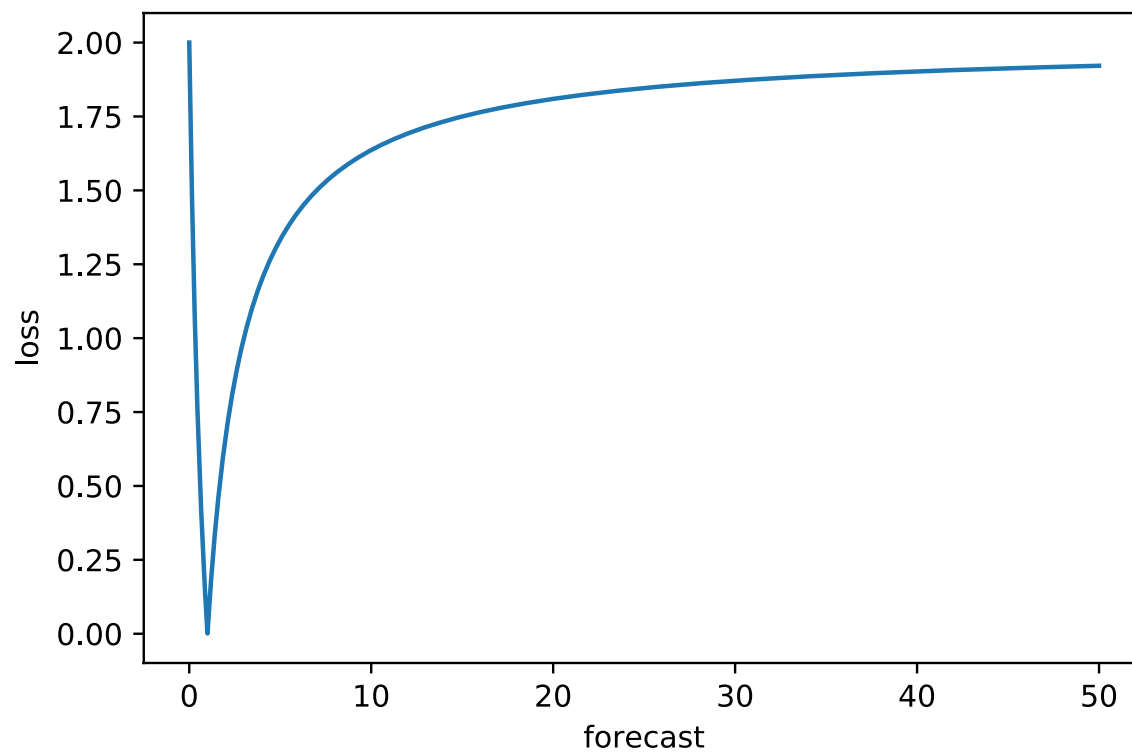
- Symmetric Mean Absolute Percentage Error (симметричный средний модуль относительной ошибки)

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

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

# SMAPE

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



# Модель линейной классификации

# Классификация

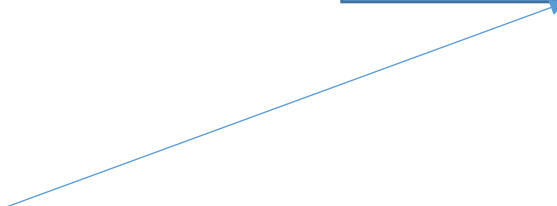
- $\mathbb{Y} = \{-1, +1\}$
- $-1$  — отрицательный класс
- $+1$  — положительный класс
- $a(x)$  должен возвращать одно из двух чисел



# Линейная регрессия

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

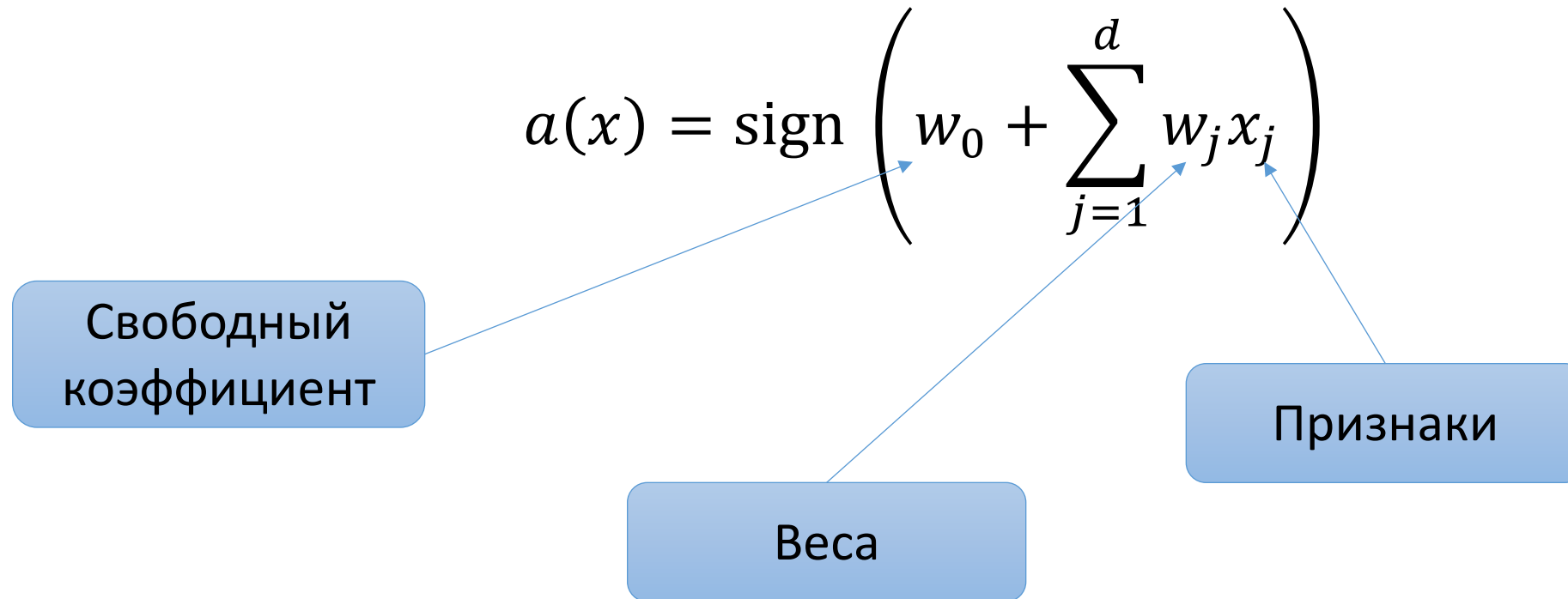
Вещественное  
число!



# Линейный классификатор

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

# Линейный классификатор



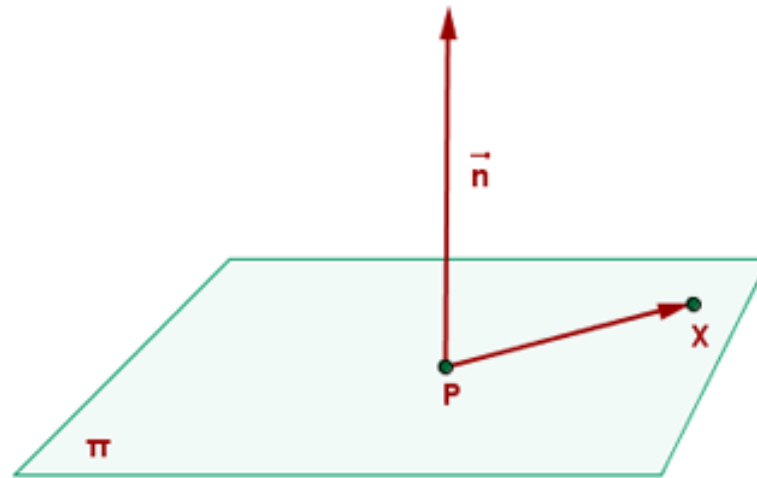
# Линейный классификатор

- Будем считать, что есть единичный признак

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

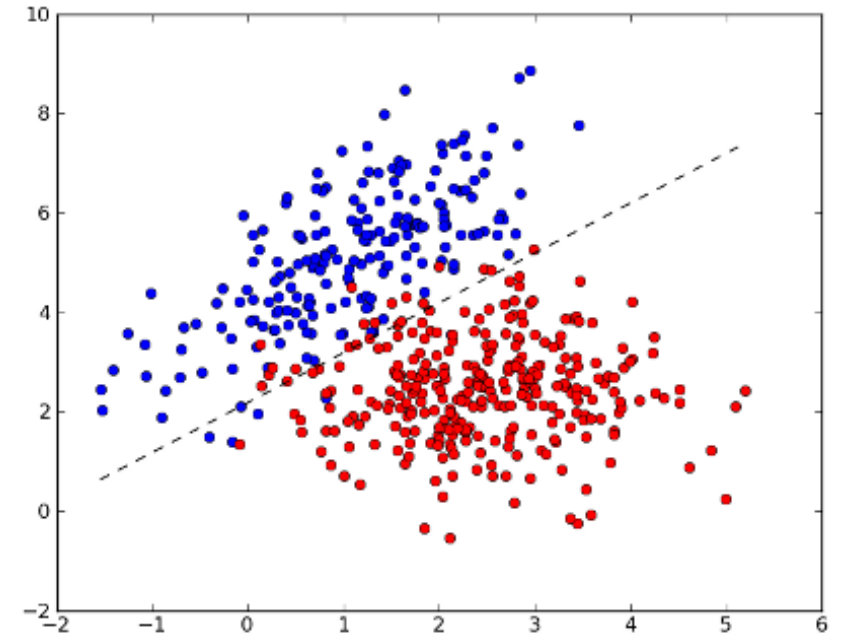
# Геометрия линейного классификатора

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



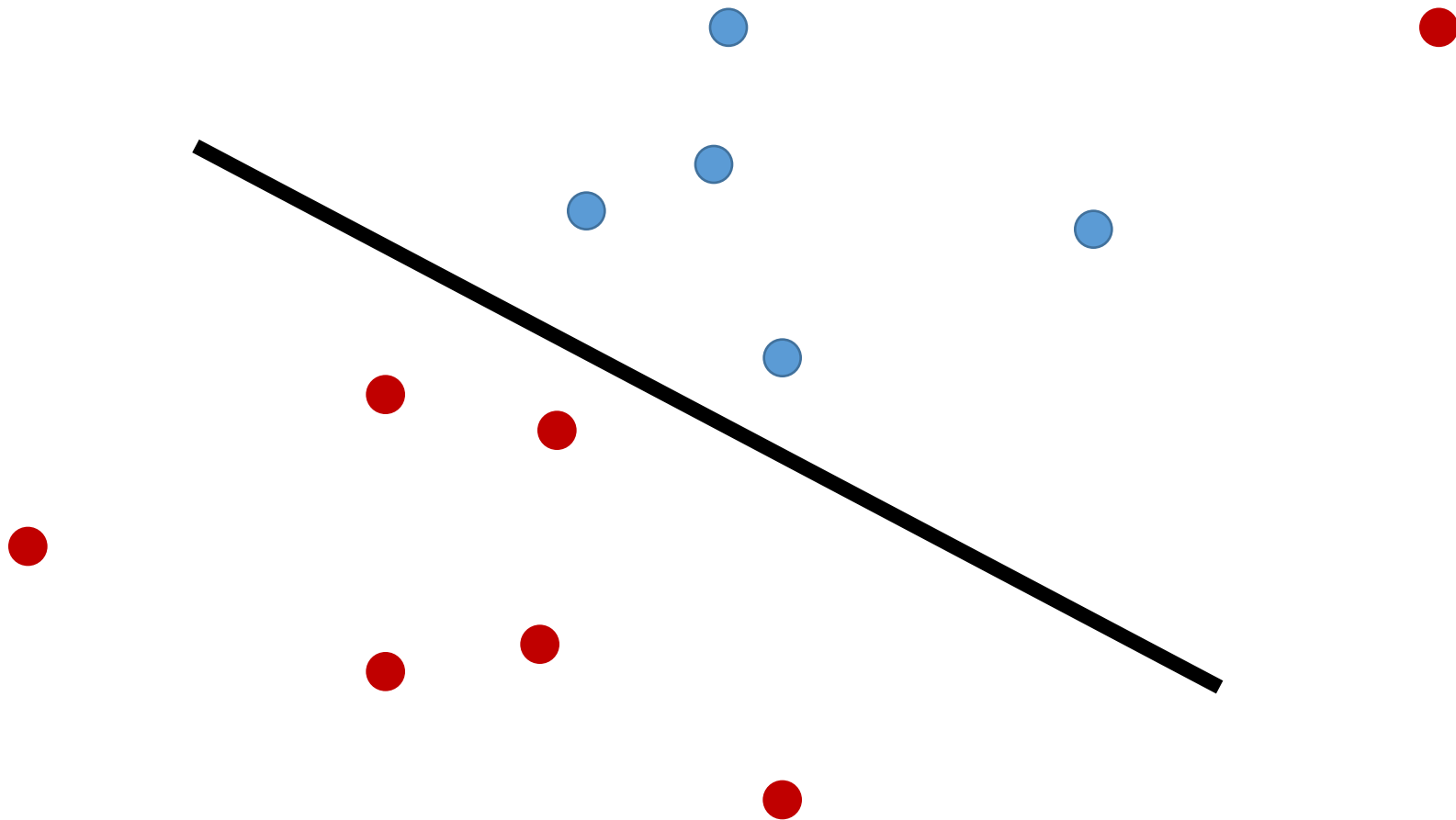
# Геометрия линейного классификатора

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

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

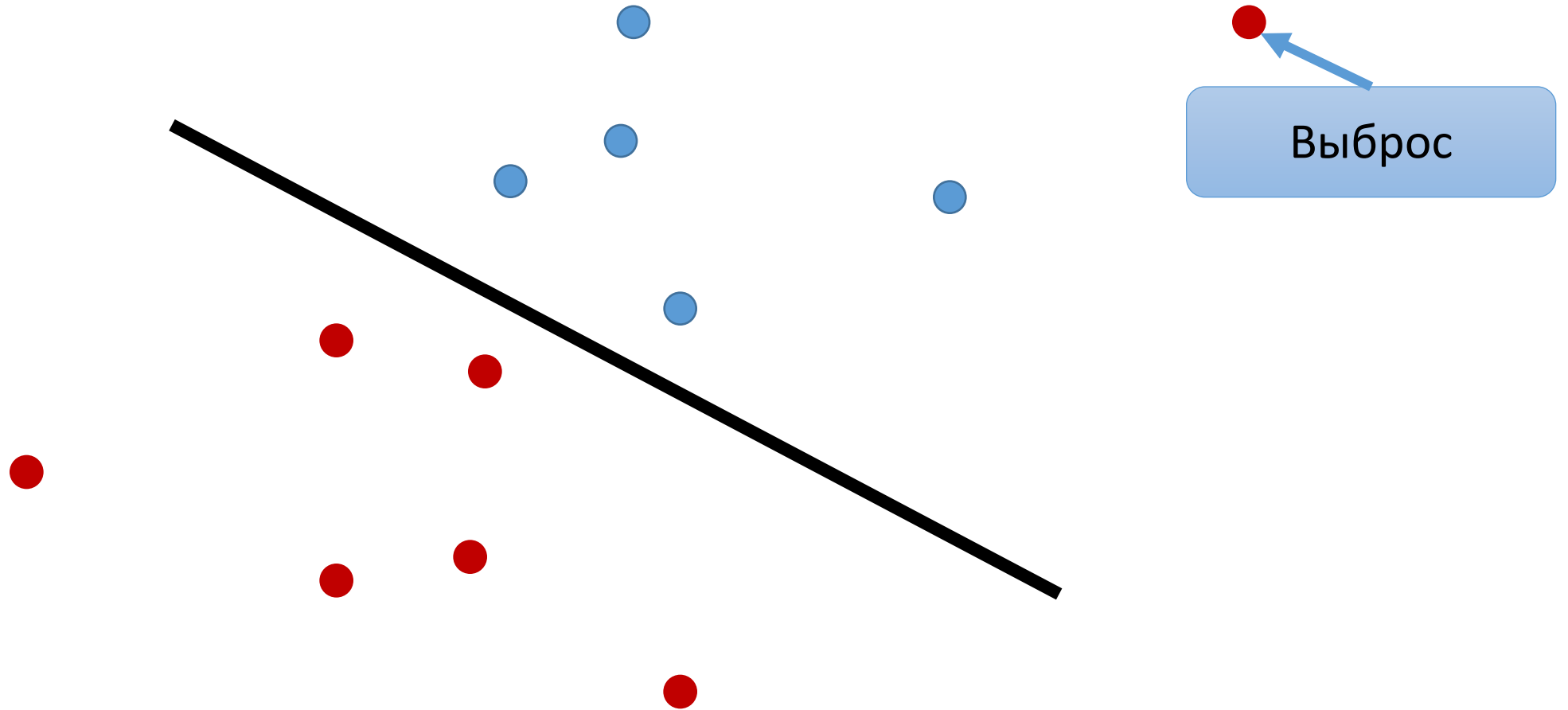
- Чем больше  $\langle w, x \rangle$ , тем дальше объект от разделяющей гиперплоскости

# Геометрия линейного классификатора



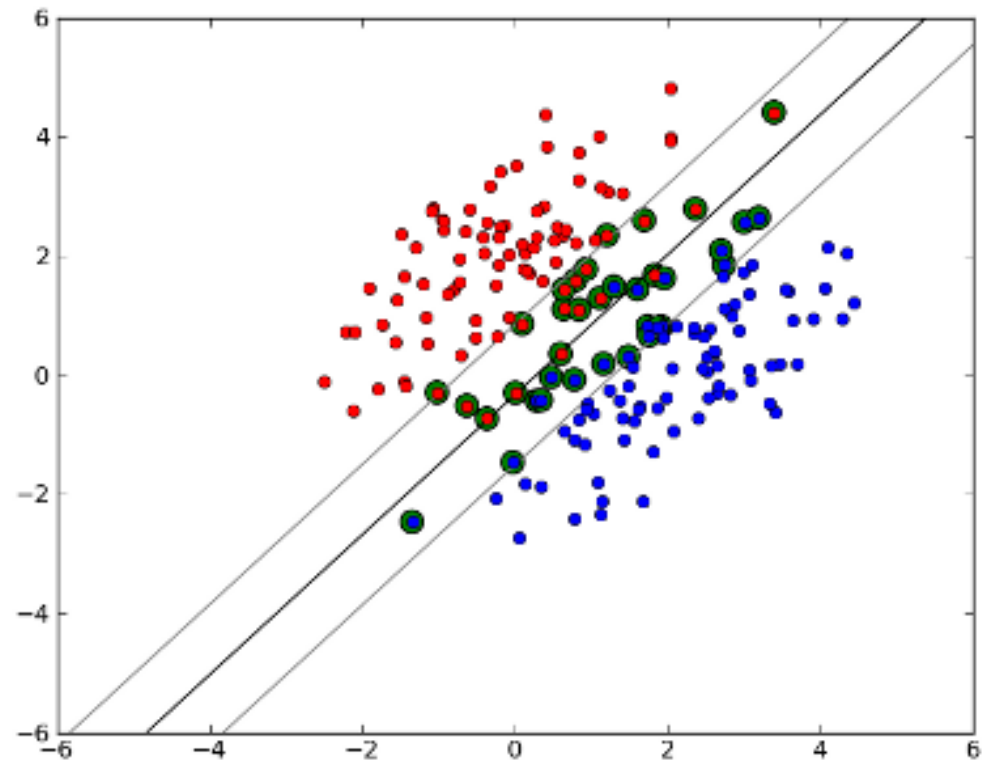


# Геометрия линейного классификатора



# 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