Основы машинного обучения

Лекция 6

Линейная регрессия и градиентный спуск

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2022

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

$$a(x) = 10 * (\text{площадь в кв. см.})$$
$$+ 500.000 * (\text{число магазинов рядом})$$
$$+ 100 * (\text{средний доход жильцов дома})$$

- 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

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

- 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:

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

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

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

Пример

- 1000 объектов
- Два признака
- Первый принимает значения от 0 до 1
- Второй равен единице на 10 объектах и нулю на 990 объектах
- $y = x_1 + 2x_2$

Пример

```
[0.3175037 , 1.        ],
[0.59558502, 1.        ],
[0.48660609, 1.        ],
[0.69255463, 1.        ],
[0.81968981, 1.        ],
[0.48844247, 1.        ],
[0.13426702, 1.        ],
[0.850628  , 1.        ],
[0.57499033, 1.        ],
[0.73993748, 1.        ],
[0.70466465, 0.        ],
[0.96821177, 0.        ],
[0.29530732, 0.        ],
[0.70530677, 0.        ],
[0.36567633, 0.        ],
[0.39541072, 0.        ],
[0.23059464, 0.        ],
[0.34401018, 0.        ],
[0.94829675, 0.        ],
[0.29257085, 0.        ],
[0.24599061, 0.        ],
[0.58313798, 0.        ],
```

Пример

$$a(x) = x_1 + 2x_2$$

- Удаляем первый признак, получаем MSE = 0.08
- Удаляем второй признак, получаем MSE = 0.04

- Правильнее удалить признак и посмотреть, как сильно растёт ошибка без него

Градиент и его свойства

Среднеквадратичная ошибка

- MSE для линейной регрессии:

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

Градиент

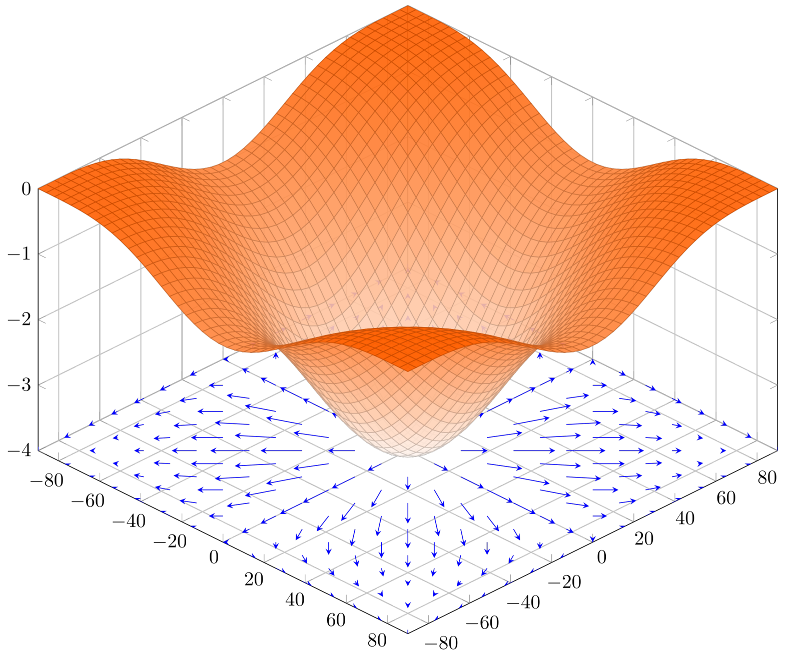
- Градиент — вектор частных производных

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

- У градиента есть важное свойство!

Важное свойство

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

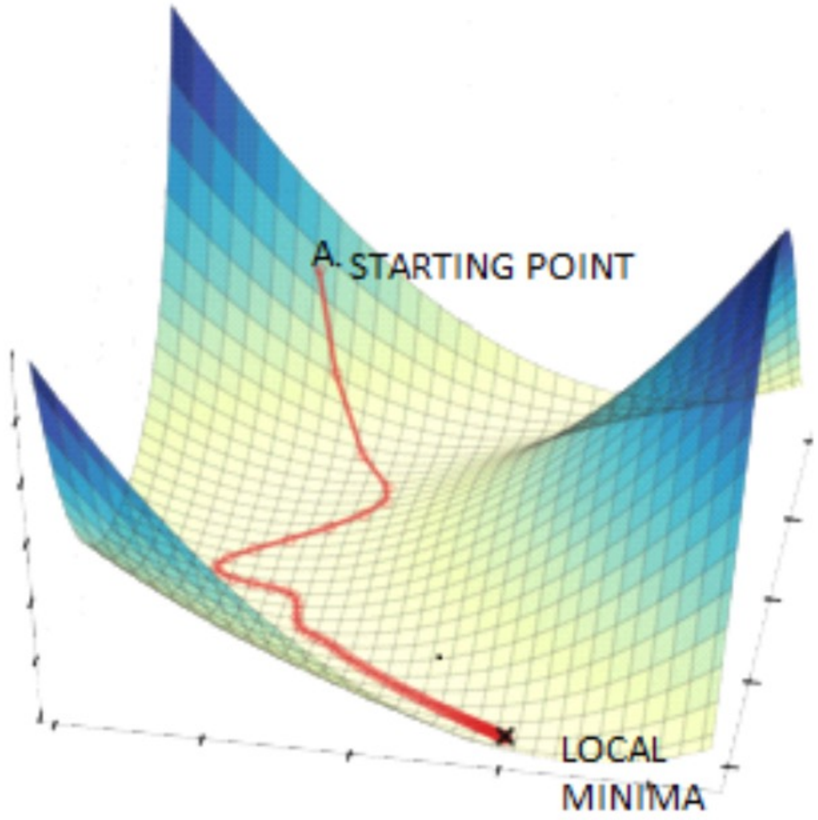


Важное свойство

- 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

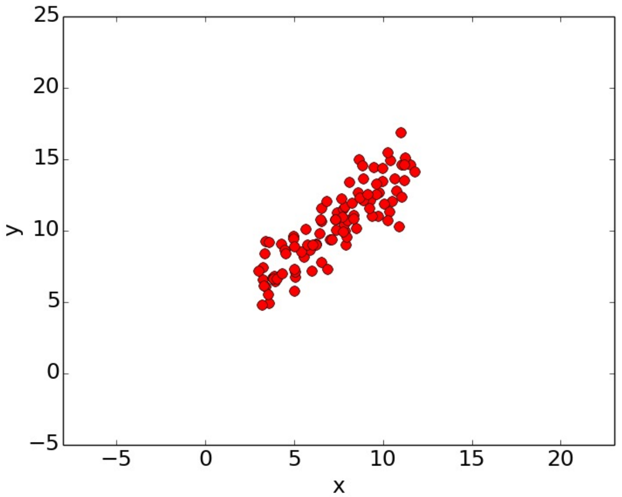
- Стартуем из случайной точки
- Сдвигаемся по антиградиенту
- Повторяем, пока не окажемся в точке минимума

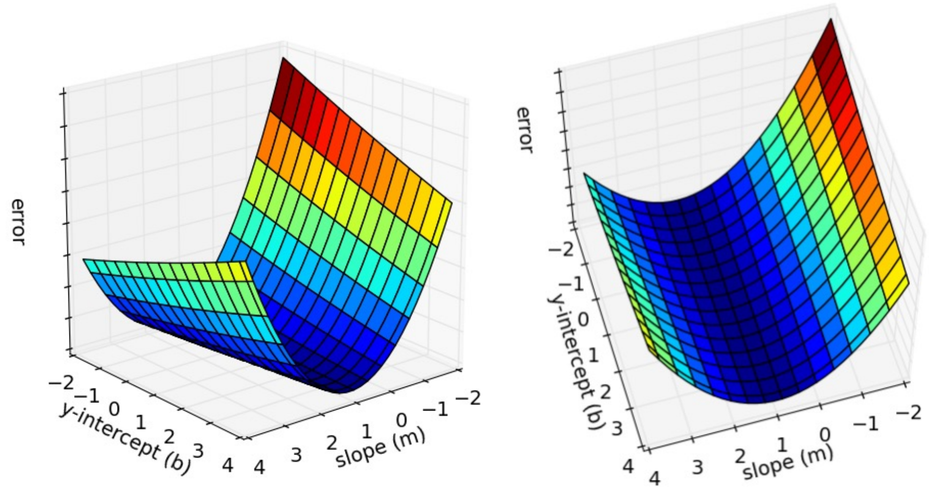
Парная регрессия

- Простейший случай: один признак
- Модель: $a(x) = w_1 x + w_0$
- Два параметра: w_1 и w_0
- Функционал:

$$Q(w_0, w_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$$

Парная регрессия





Выборка

Функционал ошибки

https://spin.atomicobject.com/2014/06/24/gradient-descent-linear-regression/

Парная регрессия

$$Q(w_0, w_1) = \frac{1}{\ell}\sum_{i=1}^{\ell}(w_1 x_i + w_0 - y_i)^2$$

- $\frac{\partial Q}{\partial w_1} = \frac{2}{\ell}\sum_{i=1}^{\ell} x_i(w_1 x_i + w_0 - y_i)$
- $\frac{\partial Q}{\partial w_0} = \frac{2}{\ell}\sum_{i=1}^{\ell}(w_1 x_i + w_0 - y_i)$
- $\nabla Q(w) = \left(\frac{2}{\ell}\sum_{i=1}^{\ell} x_i(w_1 x_i + w_0 - y_i), \frac{2}{\ell}\sum_{i=1}^{\ell}(w_1 x_i + w_0 - y_i)\right)$

Начальное приближение

- w^0 — инициализация весов
- Например, из стандартного нормального распределения

Градиентный спуск

- Повторять до сходимости:

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

Новая точка

Размер шага

Градиент в предыдущей точке

Сходимость

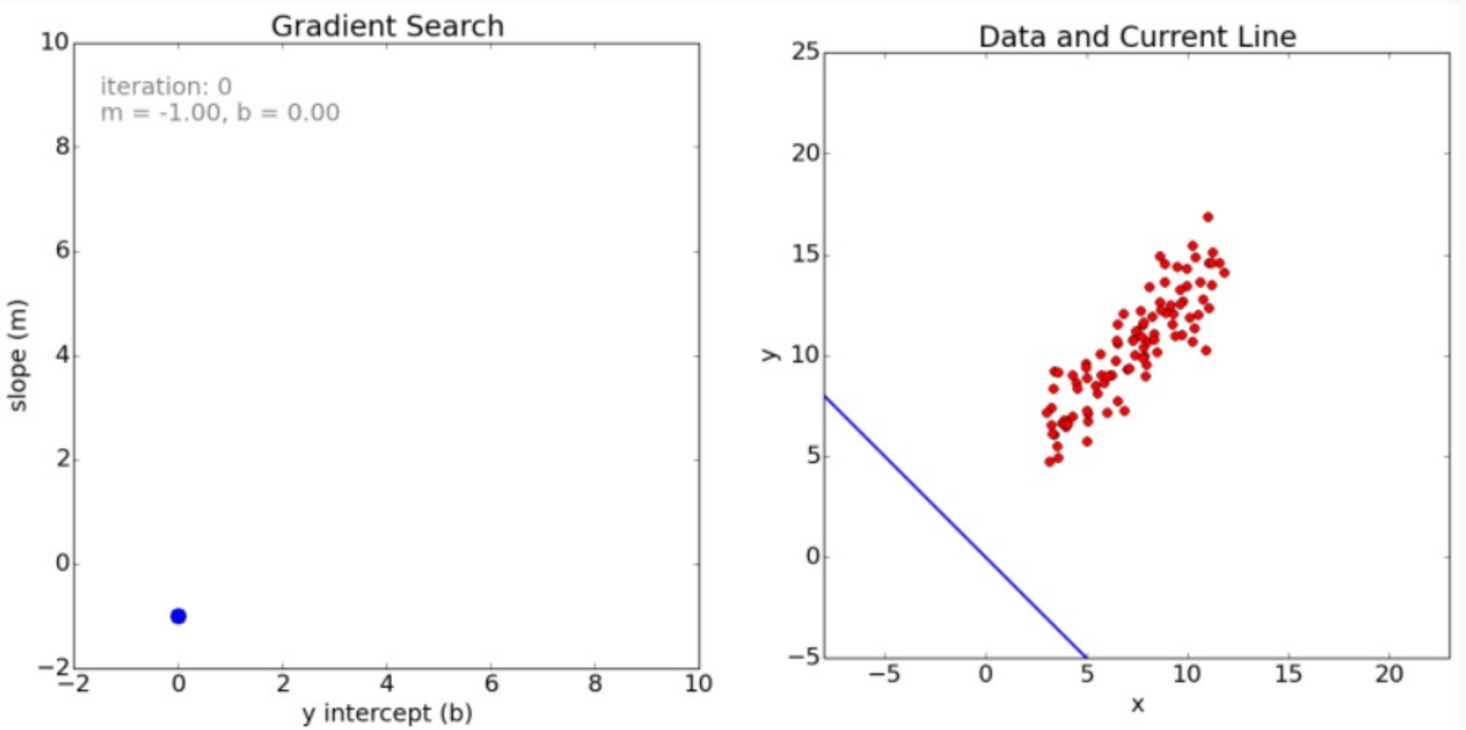
- Останавливаем процесс, если

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

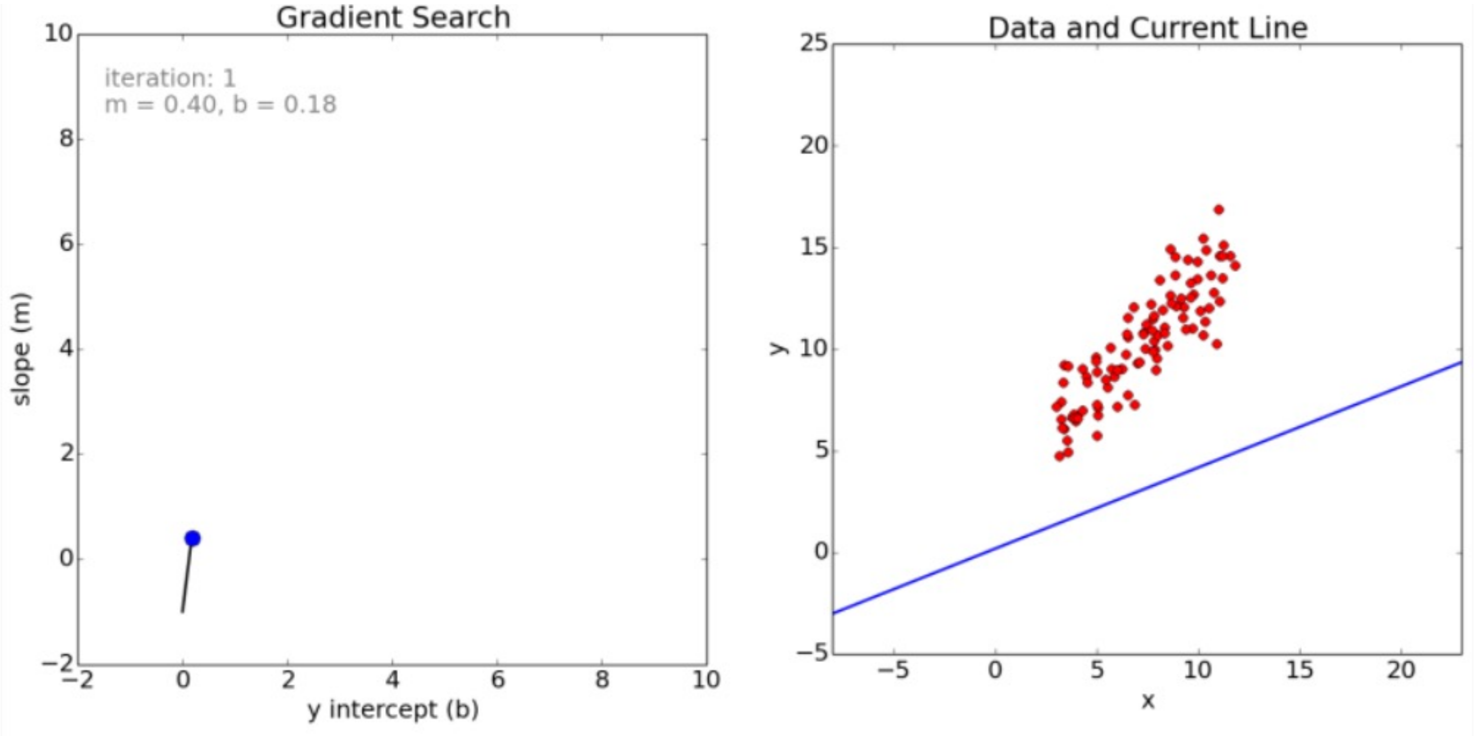
- Другой вариант:

$$\|\nabla Q(w^t)\| < \varepsilon$$

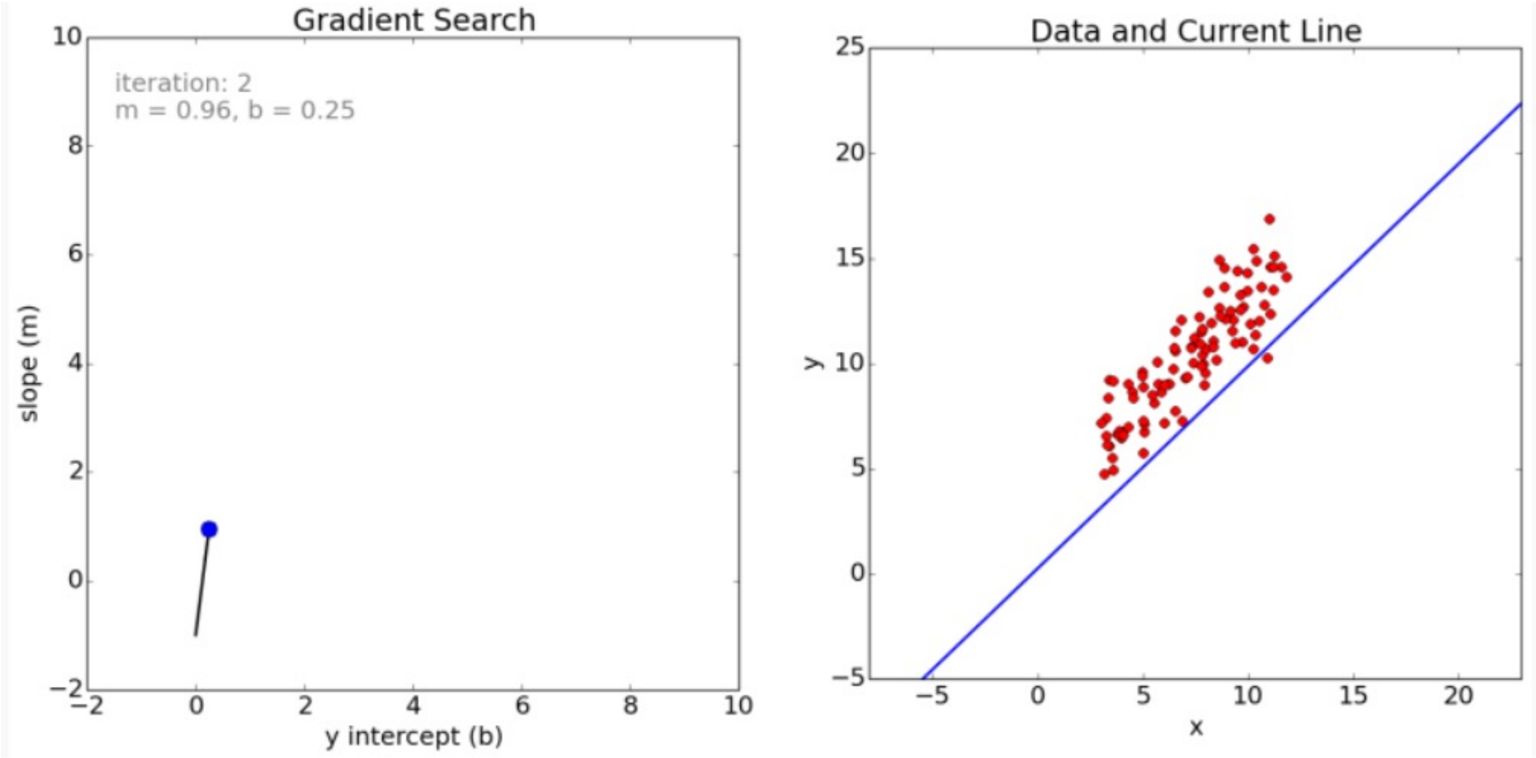
Парная регрессия



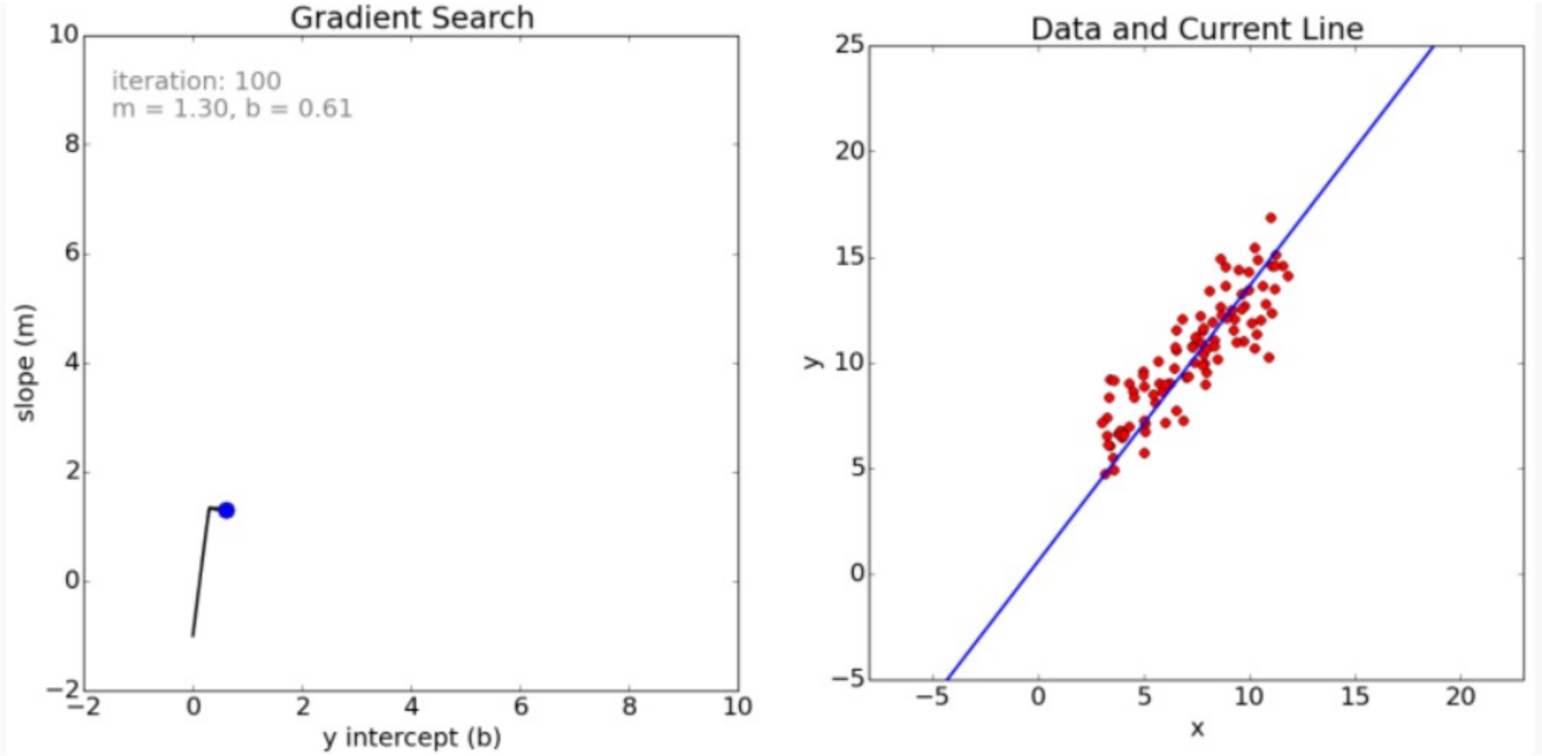
Парная регрессия



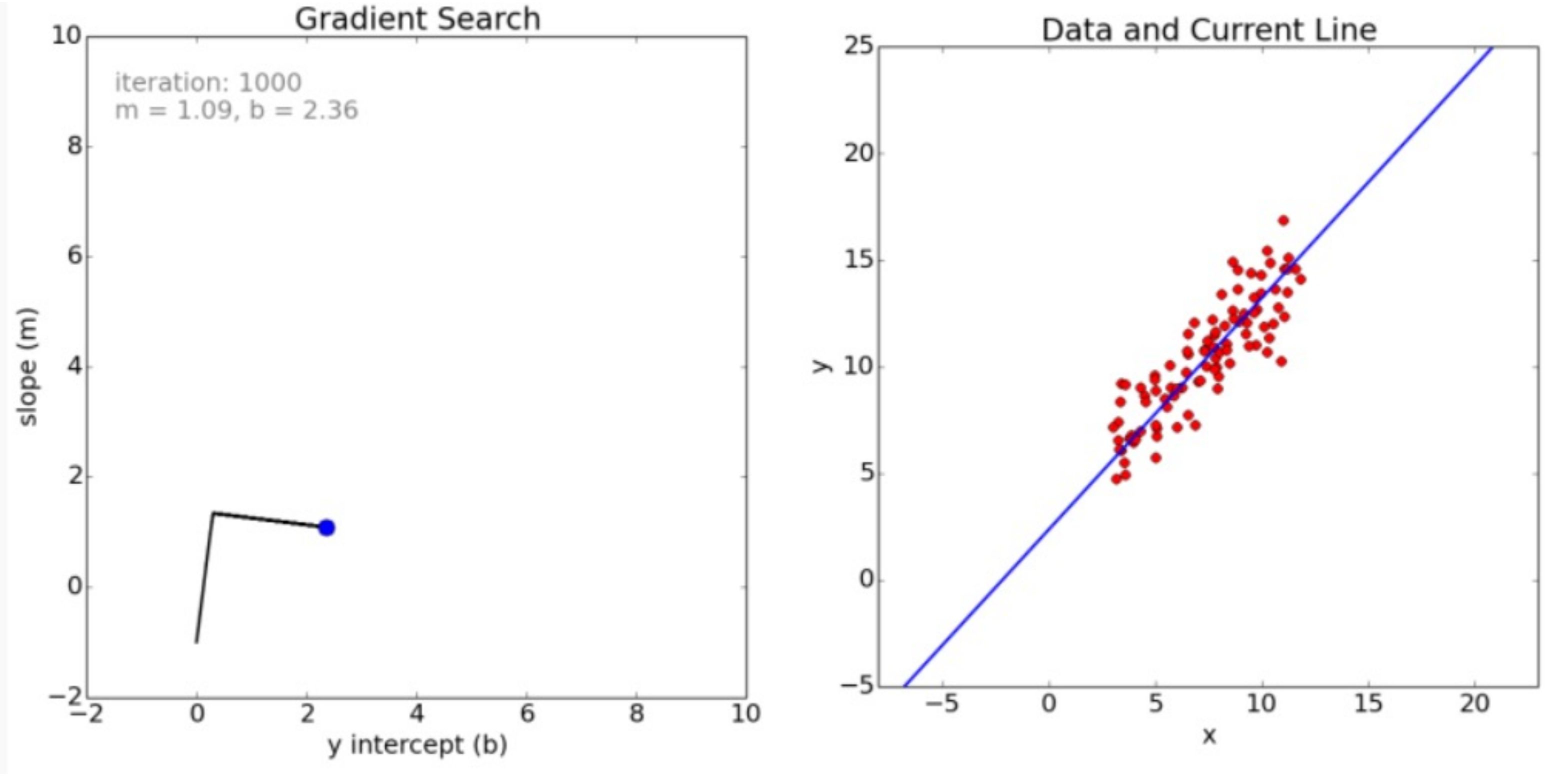
Парная регрессия



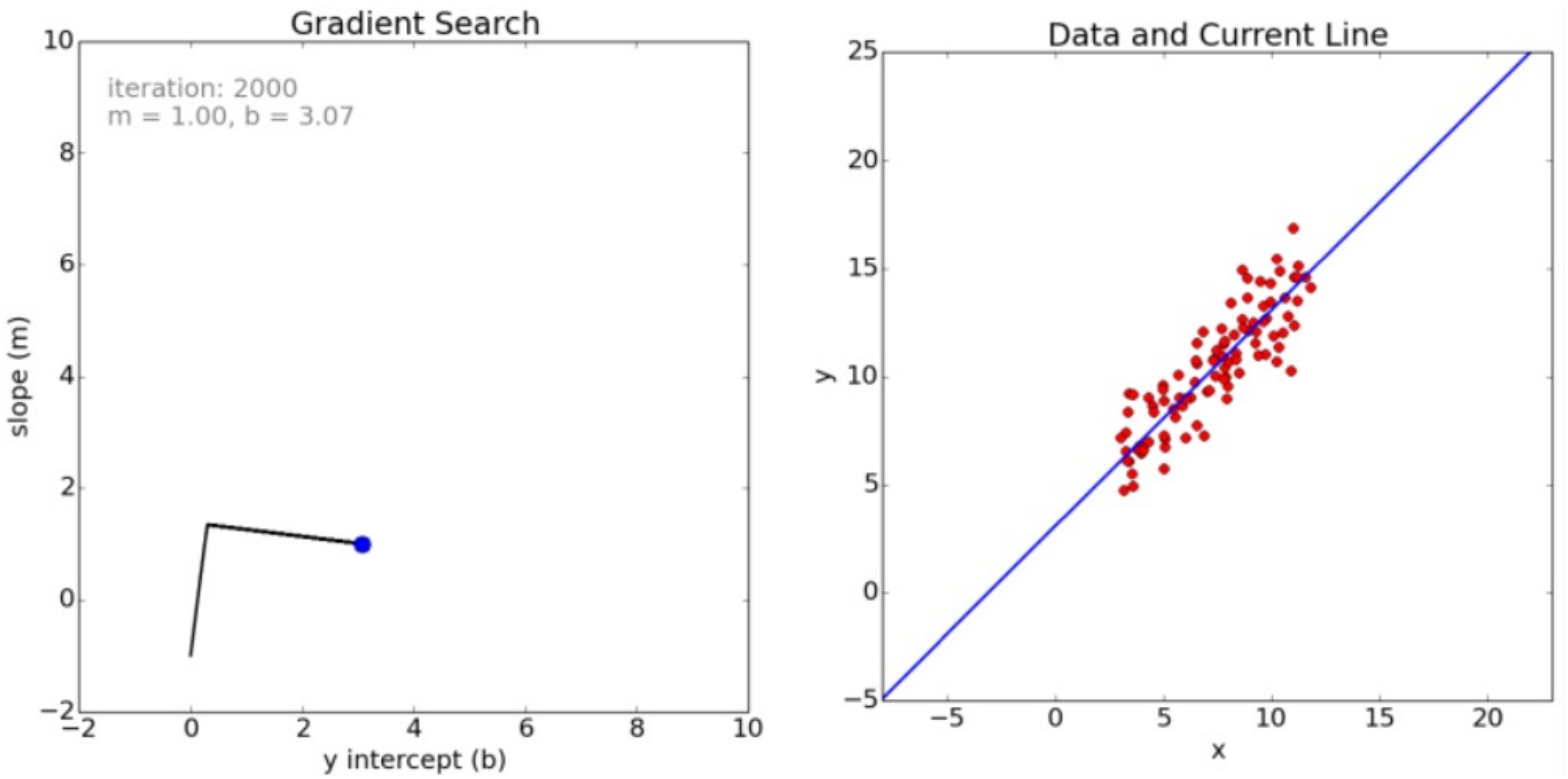
Парная регрессия

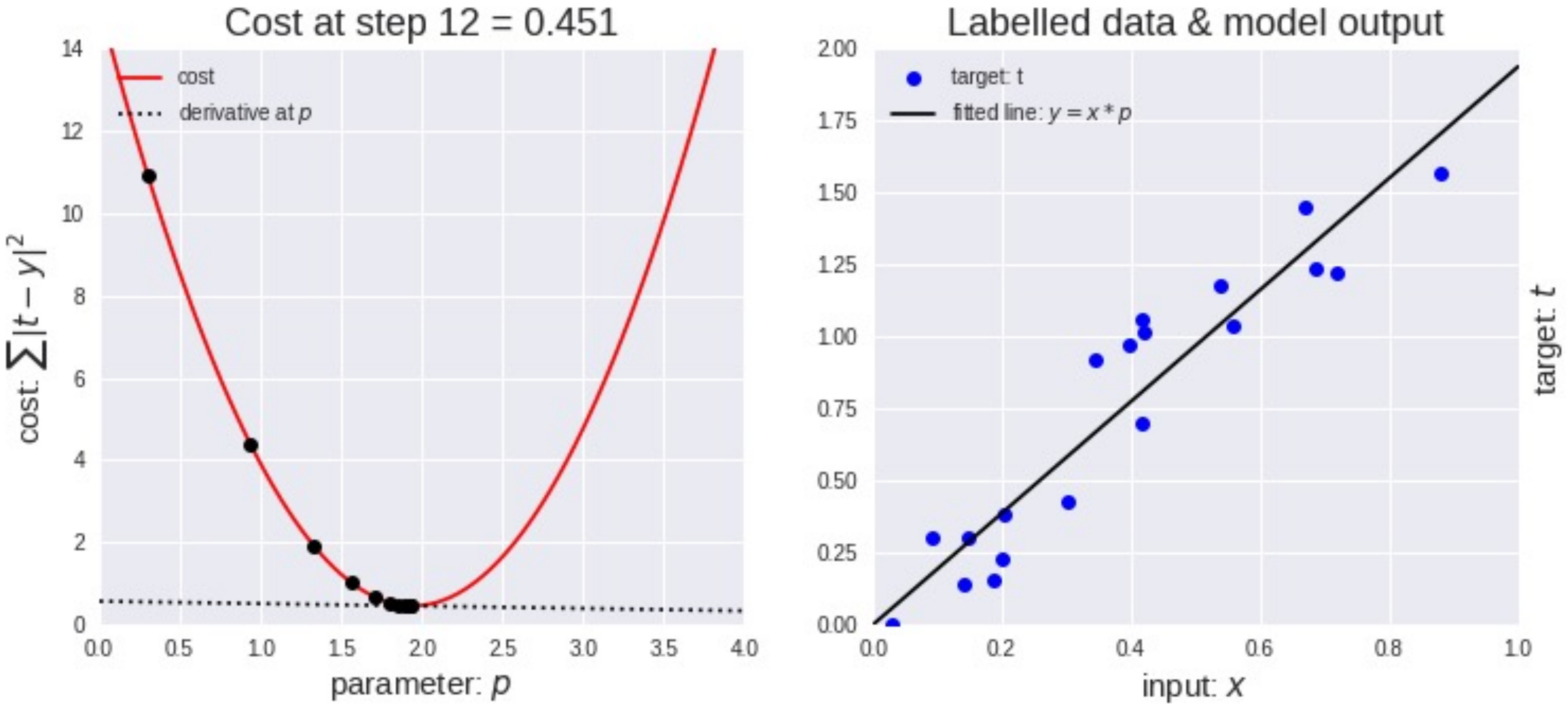


Парная регрессия



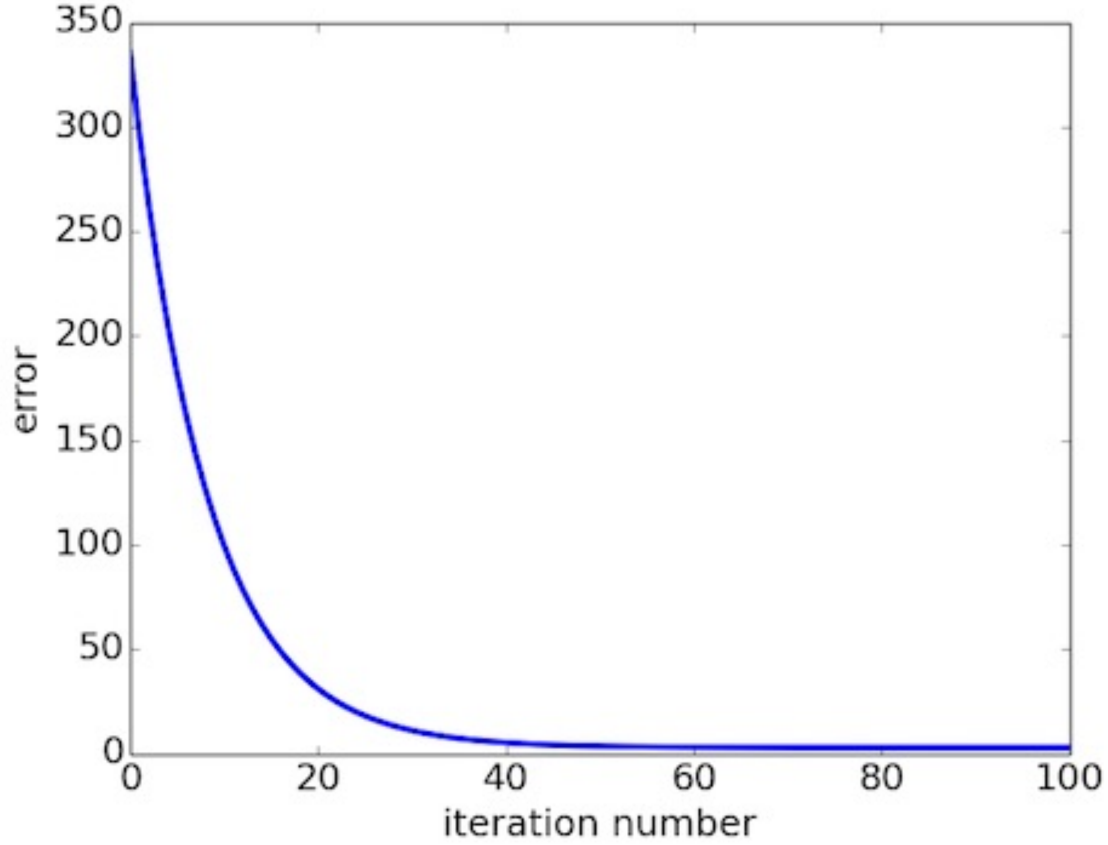
Парная регрессия





Cost at step 12 = 0.451
cost
derivative at p
cost: $\sum|t-y|^2$
parameter: p
0
2
4
6
8
10
12
14
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Labelled data & model output
target: t
fitted line: $y = x * p$
input: x
target: t
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0.0
0.2
0.4
0.6
0.8
1.0

Функционал ошибки



Линейная регрессия

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

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

Градиентный спуск

1. Начальное приближение: w^0

2. Повторять:

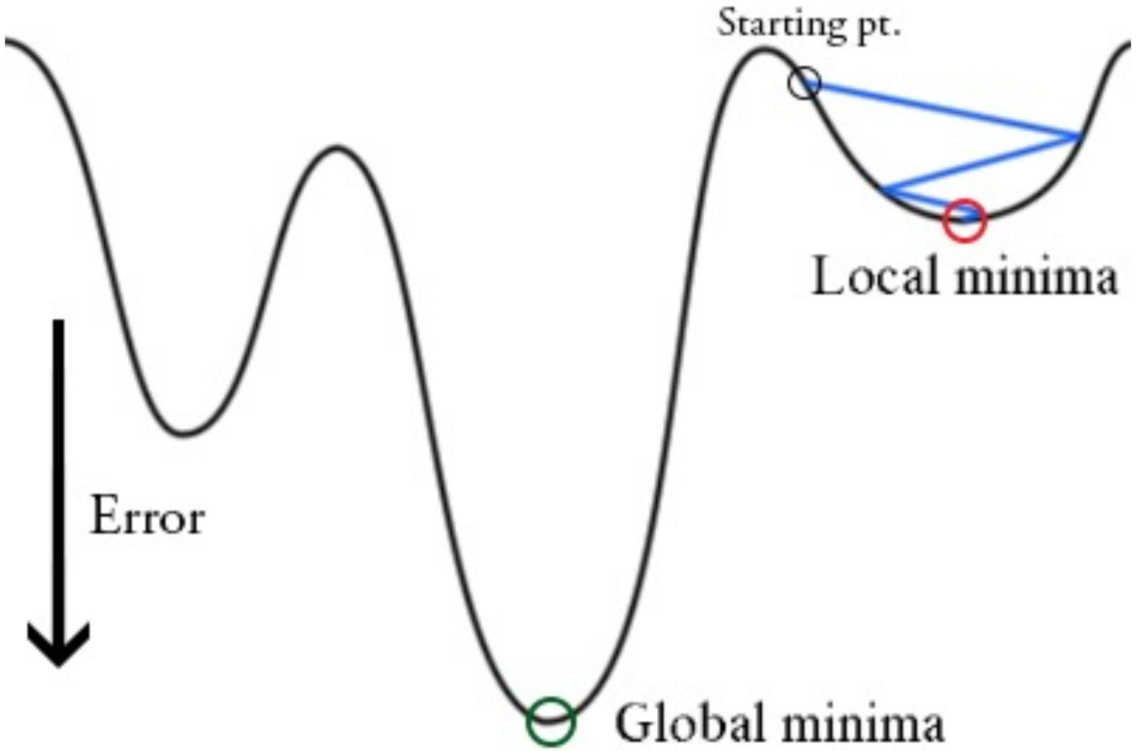
$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

3. Останавливаемся, если

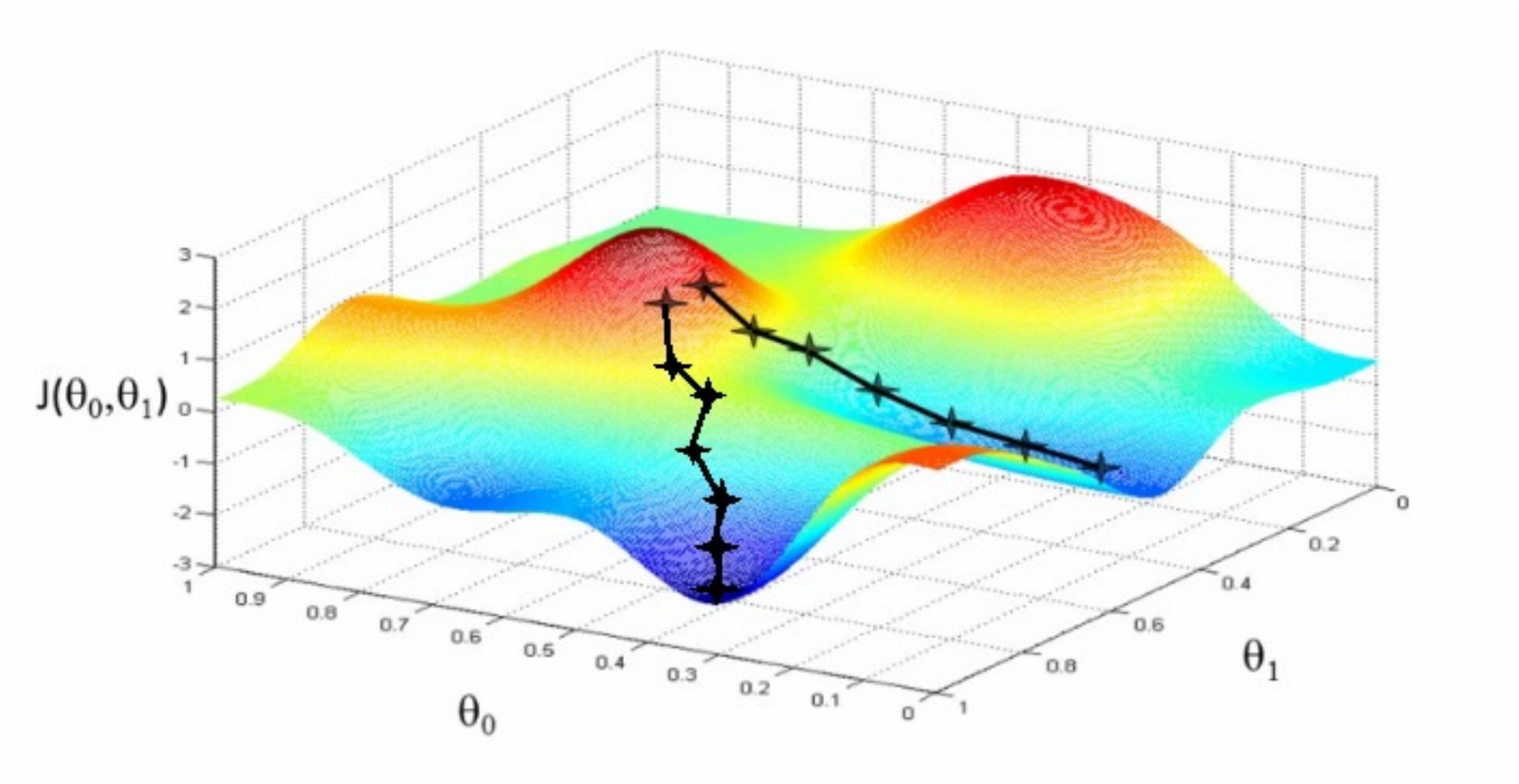
$$\|w^t - w^{t-1}\| < \varepsilon$$

Локальные минимумы

- Градиентный спуск находит только локальные минимумы



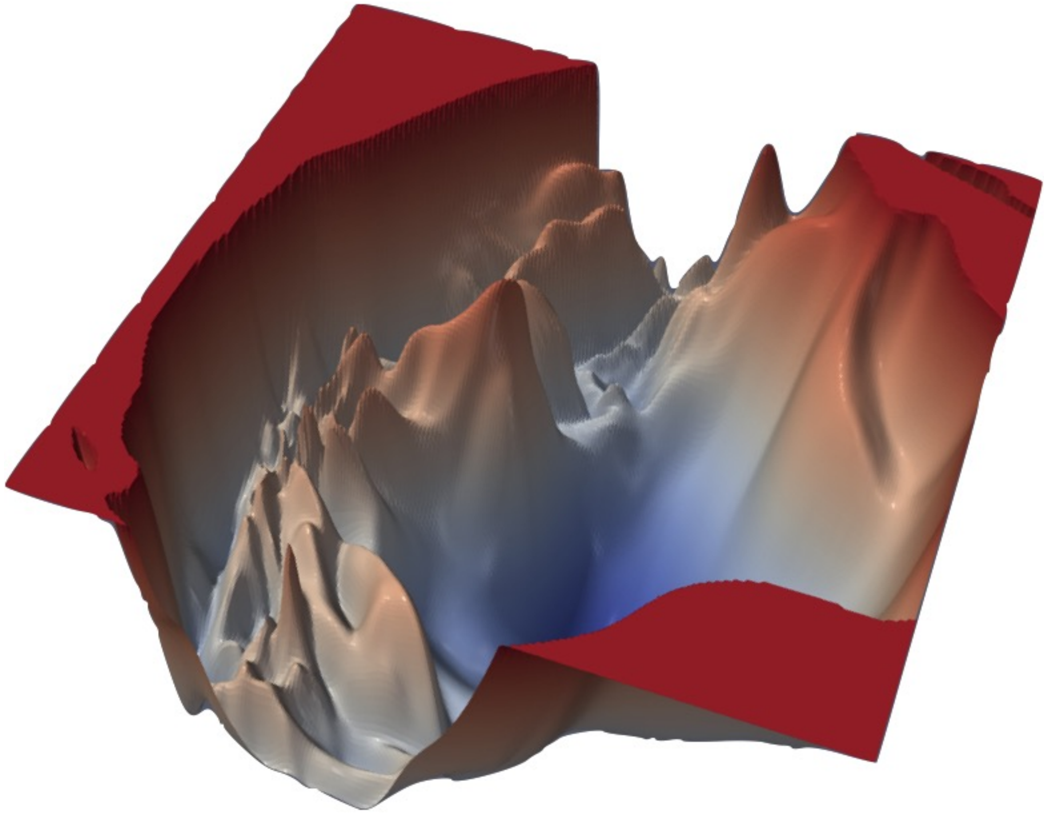
Локальные минимумы



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

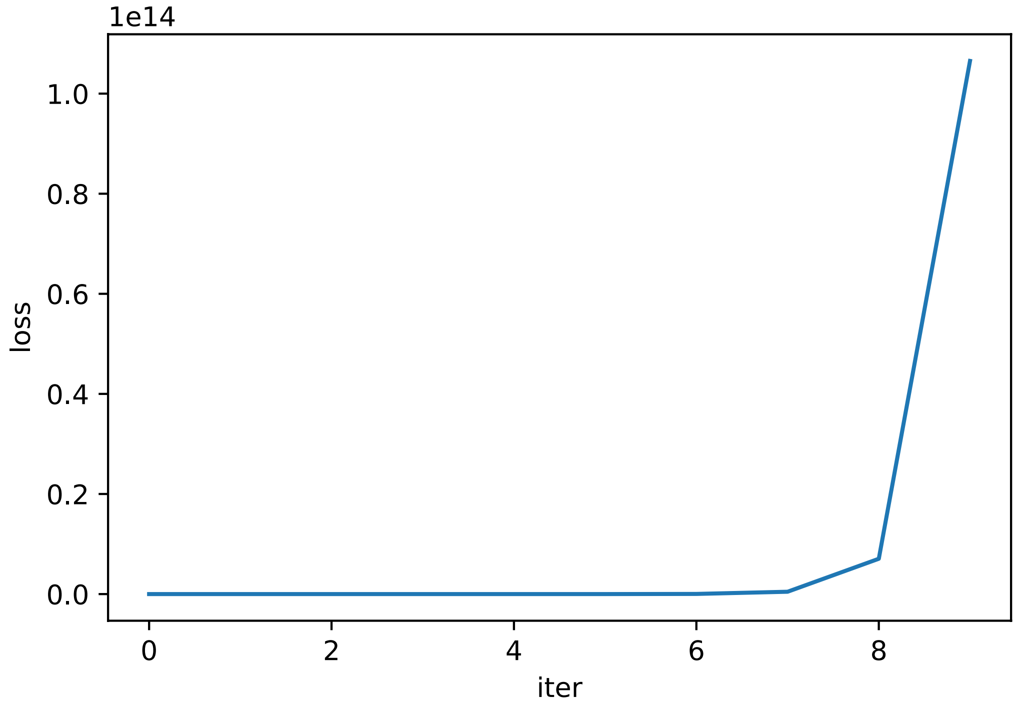
Длина шага

```
[[  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:

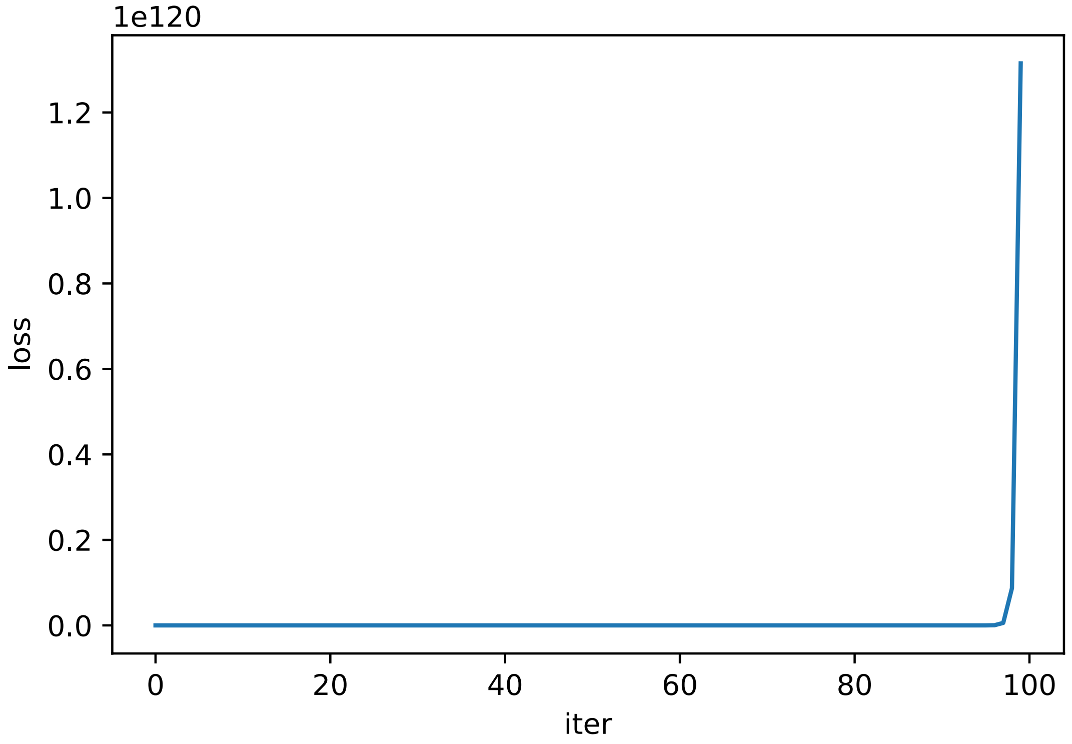
[26.52, 564.80, 682.90, 5097.71, 12110.87]



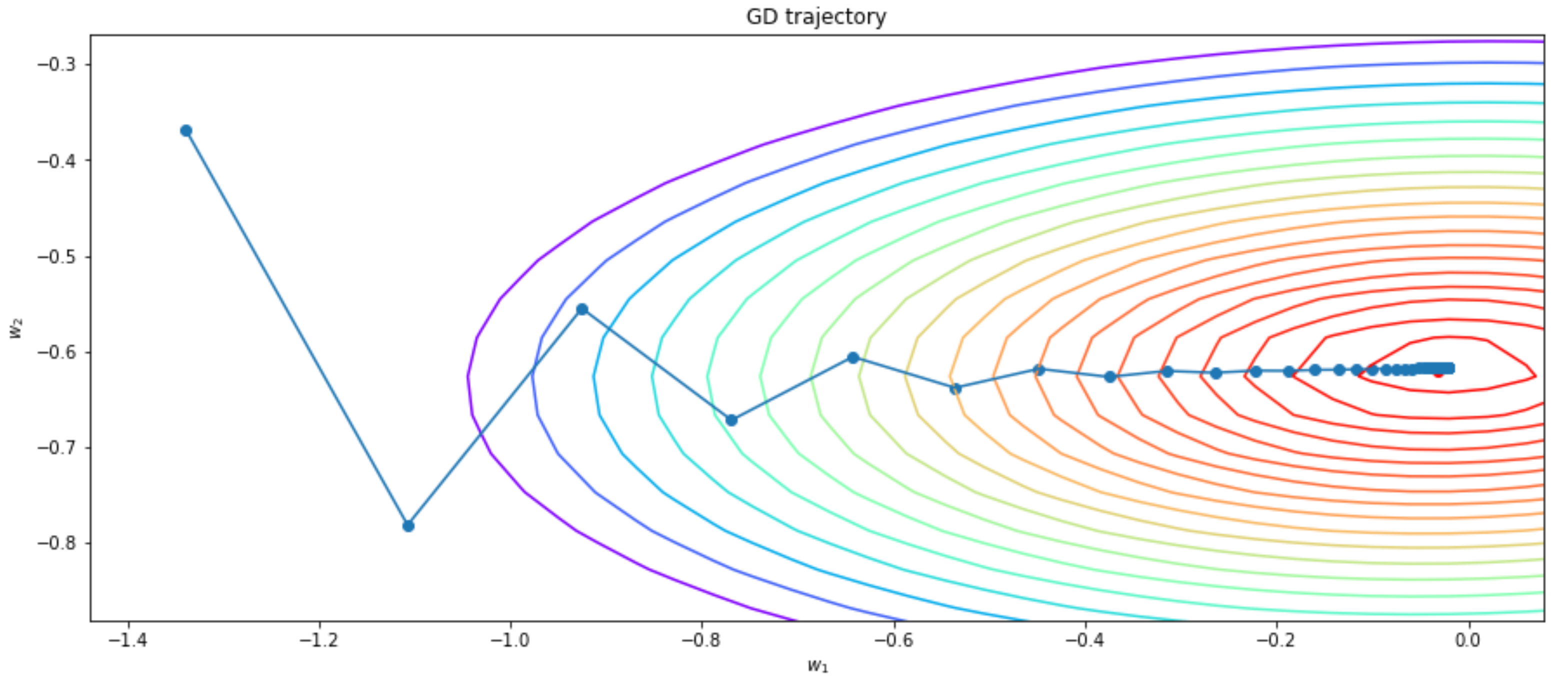
Длина шага

Градиент на первом шаге:

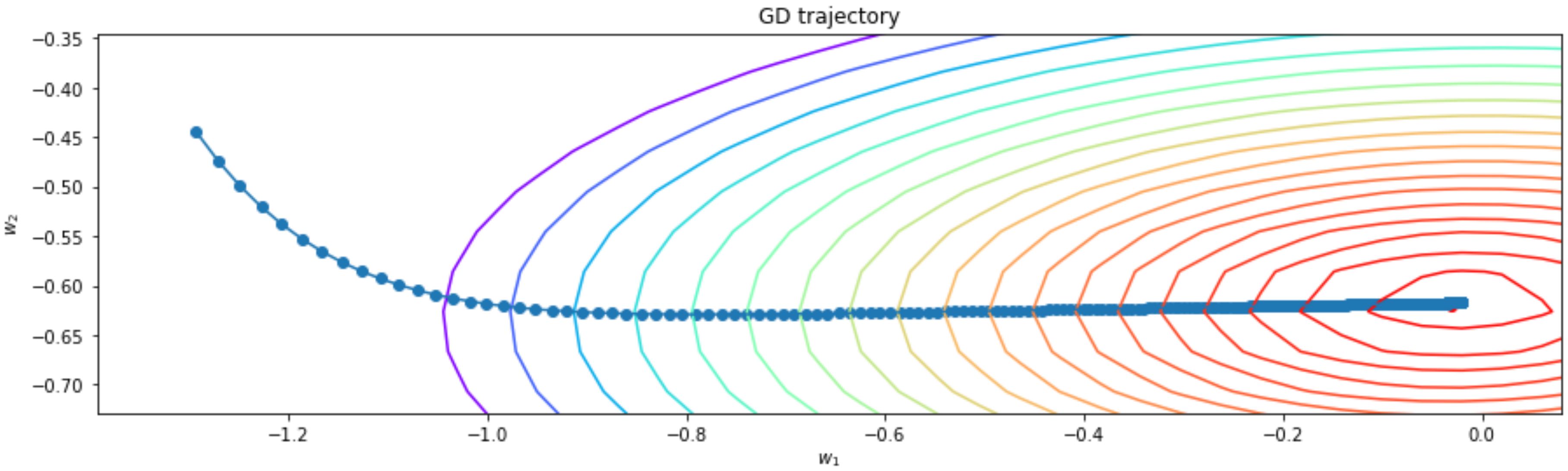
[26.52, 564.80, 682.90, 5097.71, 12110.87]



Длина шага



Длина шага



GD trajectory
−0.35
−0.40
−0.45
−0.50
−0.55
−0.60
−0.65
−0.70
w_2
−1.2
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
w_1

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

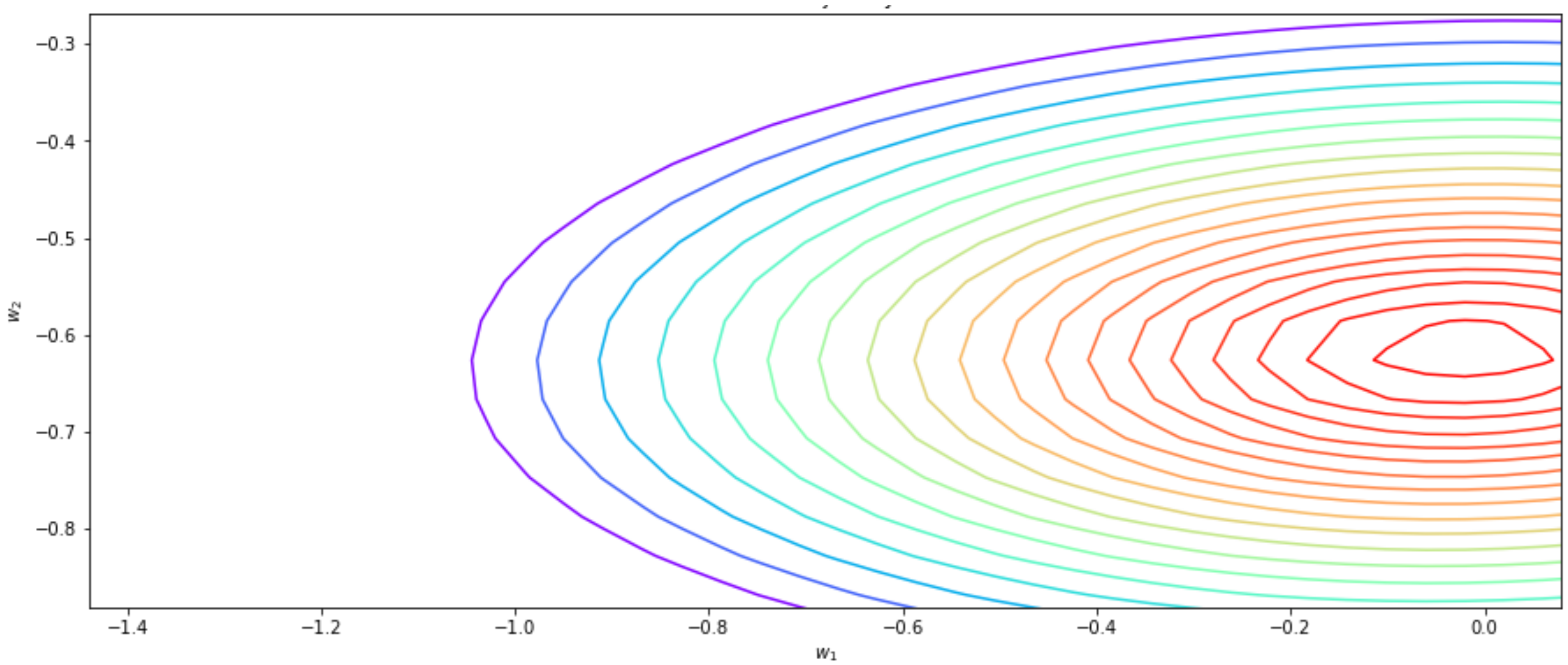
$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1})$$

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

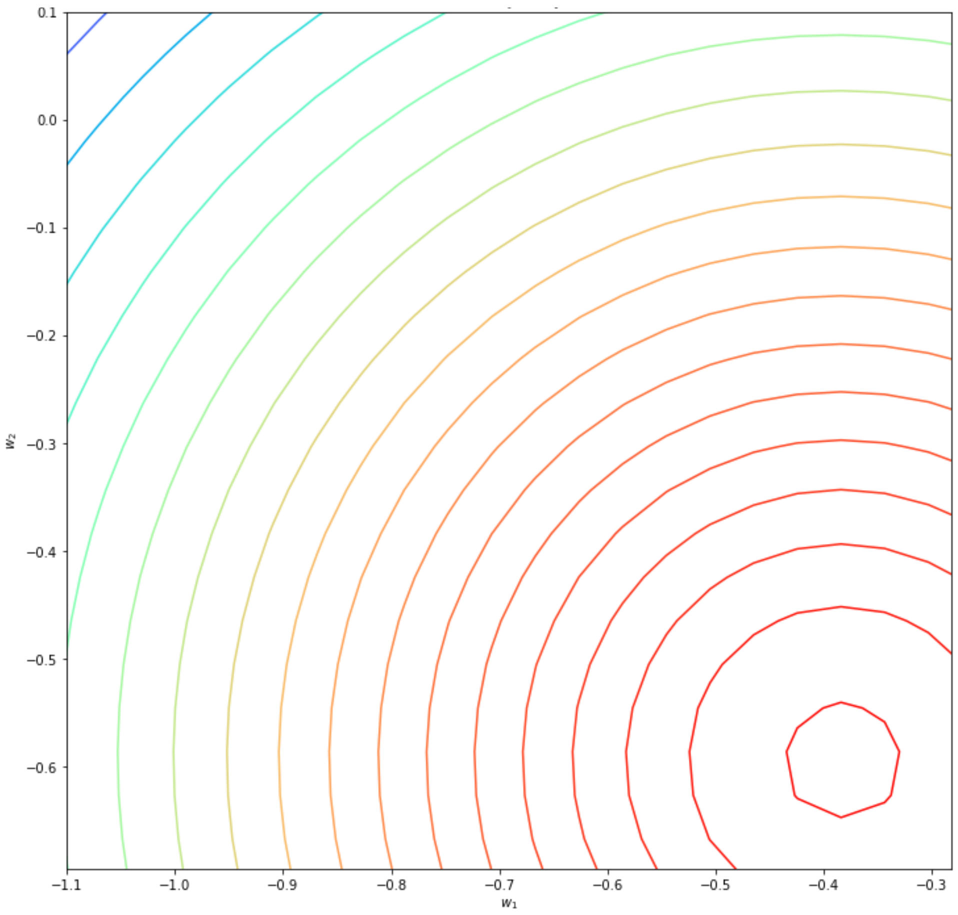
Масштабирование признаков

```
[[  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:

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