Машинное обучение

Лекция 3

Линейная регрессия и градиентный спуск

Ковалев Евгений

ekovalev@hse.ru

НИУ ВШЭ, 2020

Обучение линейной регрессии

Среднеквадратичная ошибка

- MSE для линейной регрессии:

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

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

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

Регуляризация

- Регуляризованный функционал

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

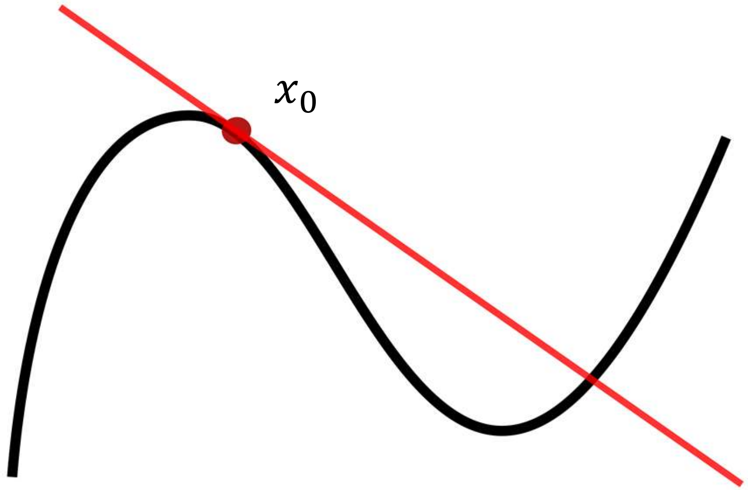
- Аналитическое решение:

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

- Гребневая регрессия (Ridge regression)

Производная

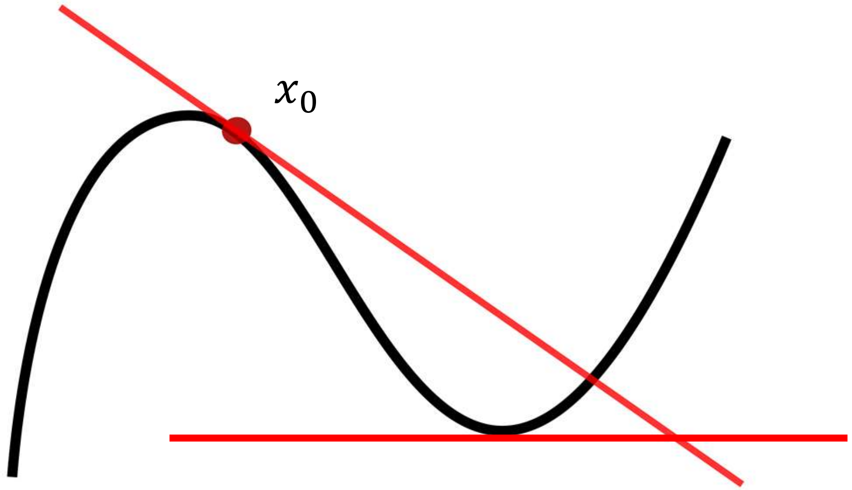
$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$



Производная

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

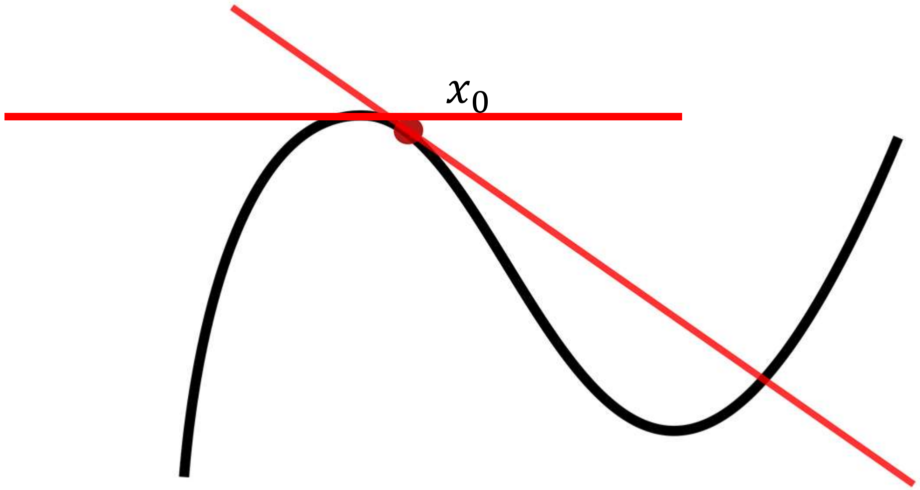
$$f'(x_0) = 0$$



Производная

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

$$f'(x_0) = 0$$



Градиент

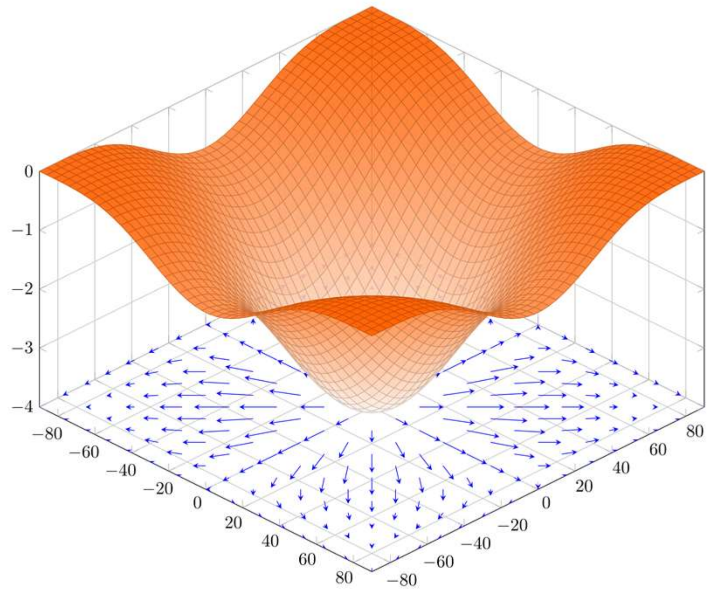
- Градиент — вектор частных производных

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

- 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!
- А быстрее всего убывает в сторону антиградиента

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

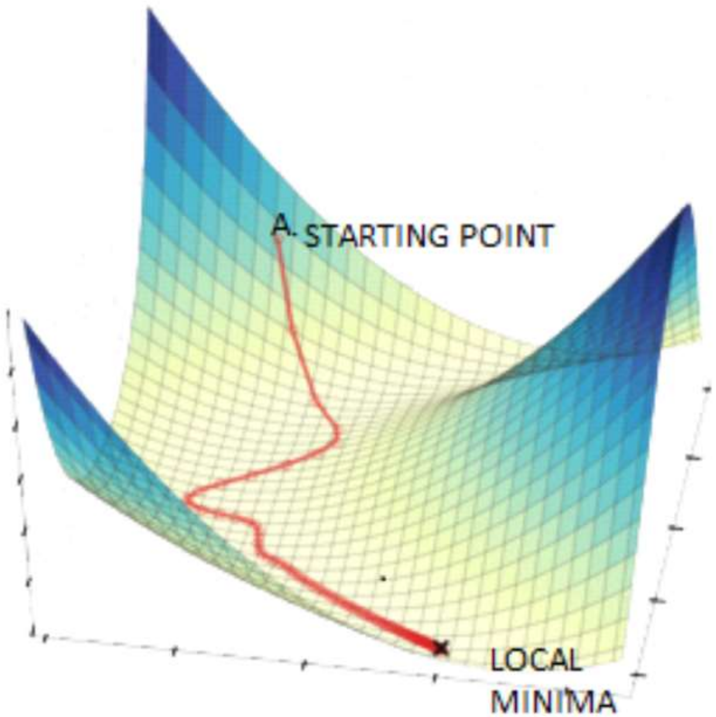
$$\nabla f(x_0) = 0$$

- Если функция выпуклая, то экстремум один
- MSE для линейной регрессии — выпуклая!

Градиентный спуск

Как это пригодится?

Как это пригодится?



Градиентный спуск

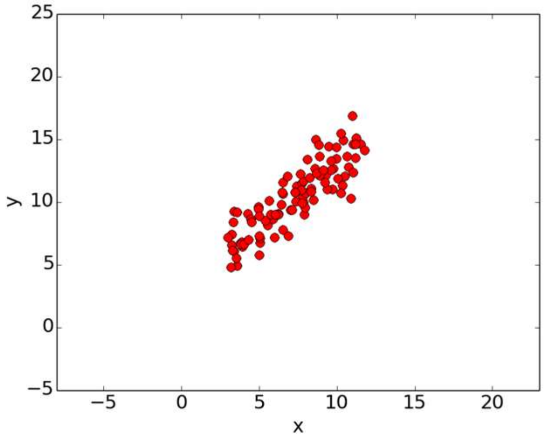
- Стартуем из случайной точки
- Сдвигаемся по антиградиенту
- Повторяем, пока не окажемся в точке минимума

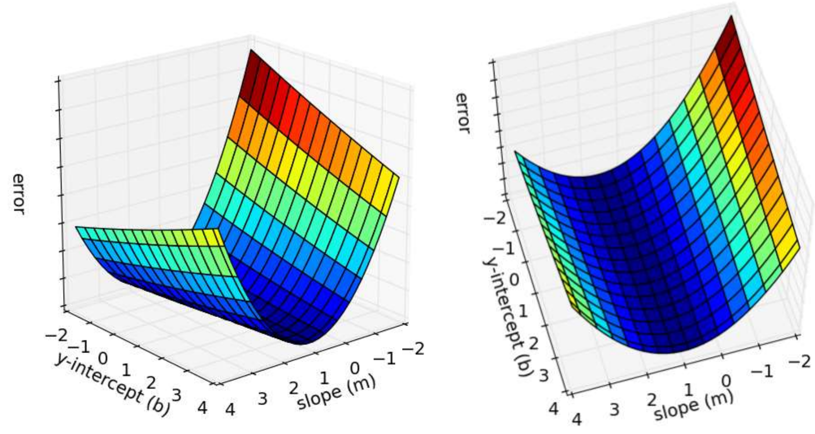
Парная регрессия

- 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

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

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

Линейная регрессия

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

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

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

Длина шага

Градиент в предыдущей точке

Длина шага

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

- Позволяет контролировать скорость обучения

Сходимость

- Останавливаем процесс, если

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

- Другой вариант:

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

- Другой вариант:

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

Градиентный спуск

1. Начальное приближение: w^0

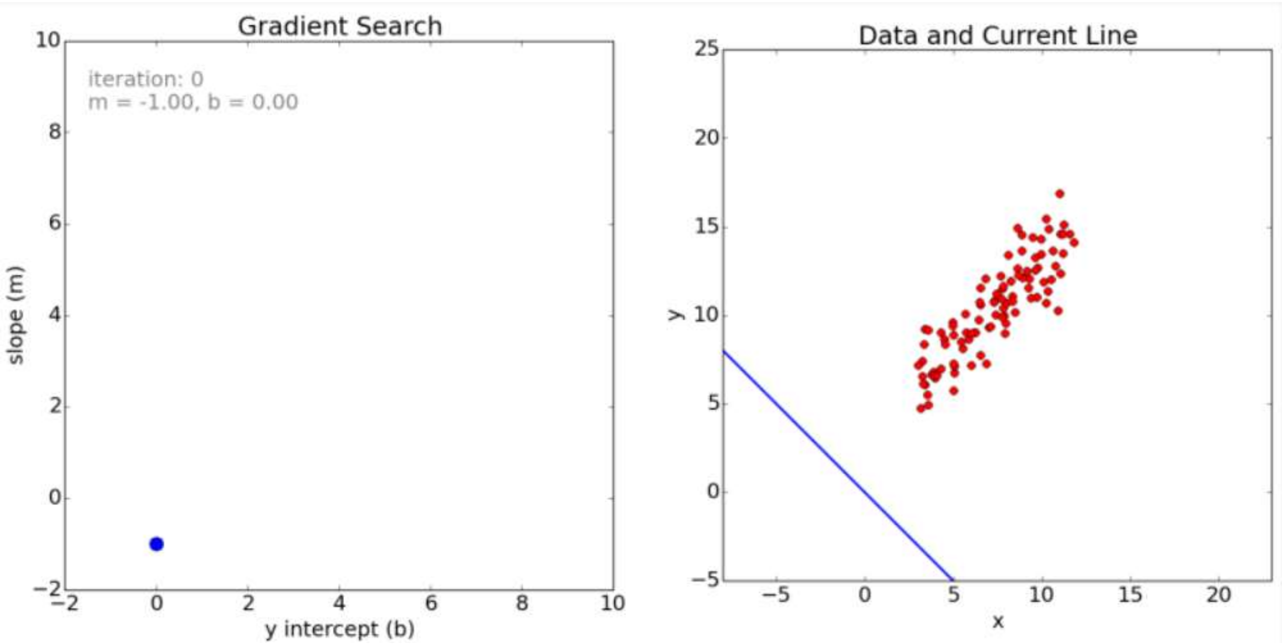
2. Повторять:

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

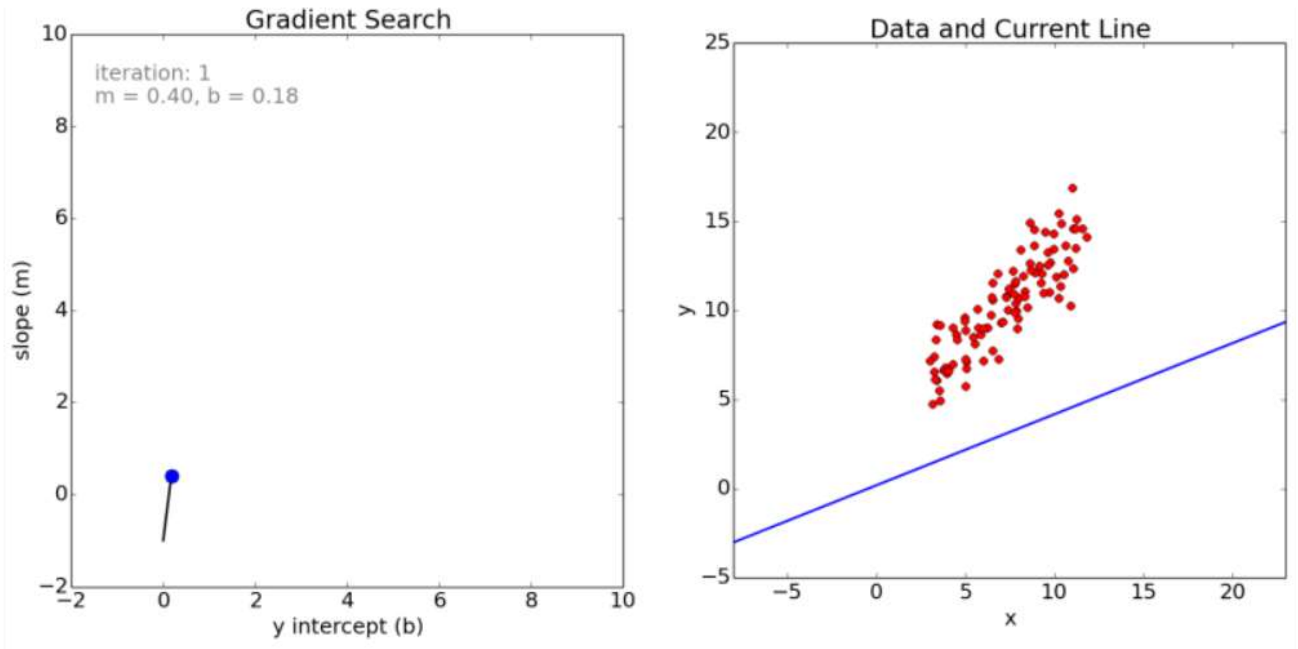
3. Останавливаемся, если

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

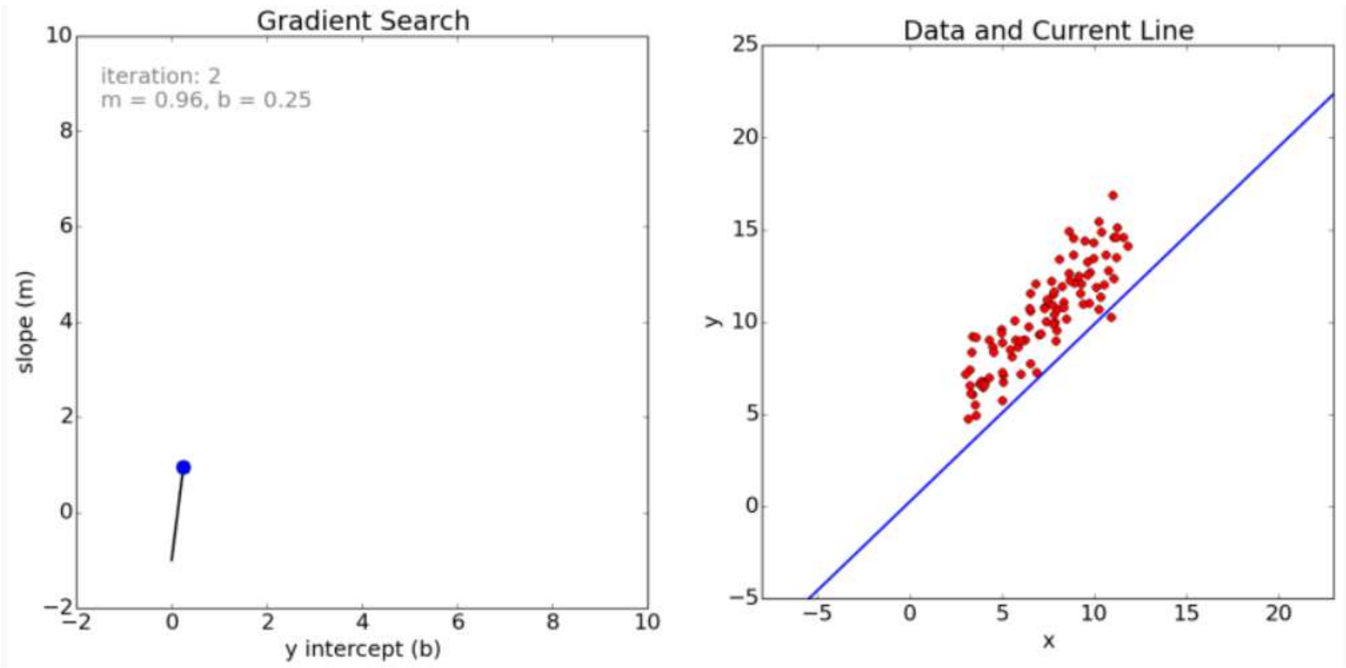
Парная регрессия



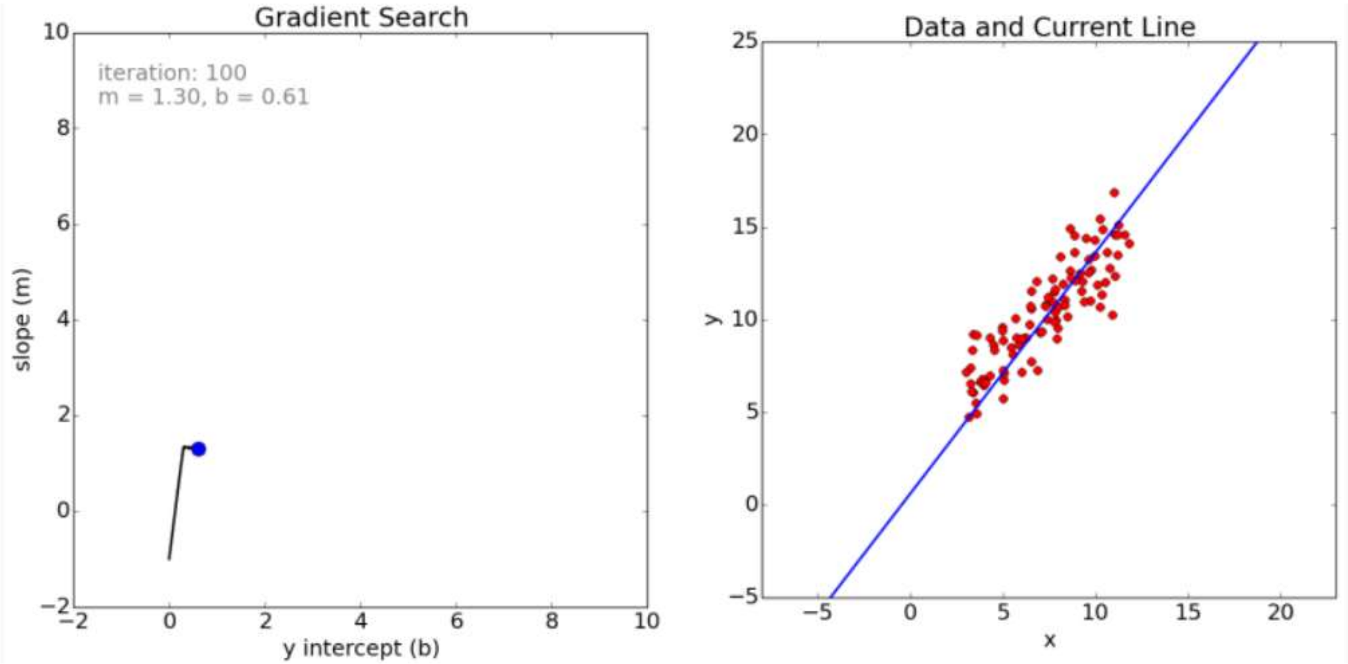
Парная регрессия



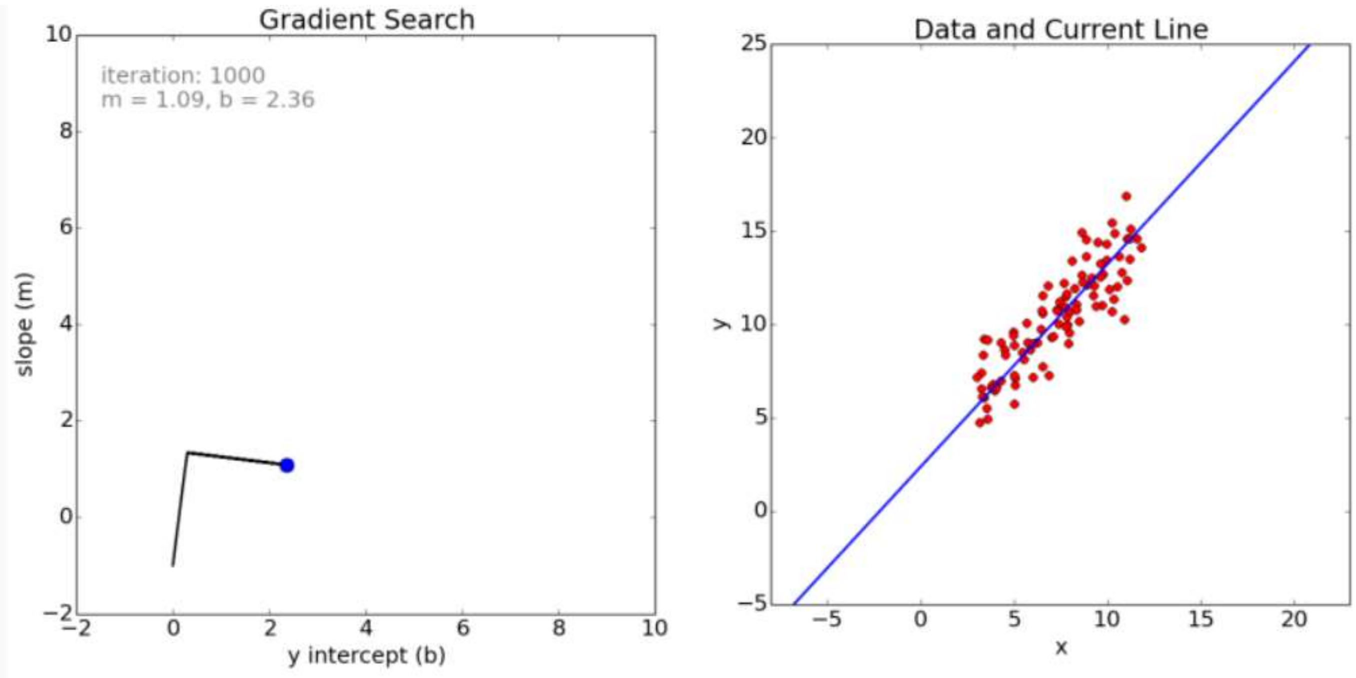
Парная регрессия



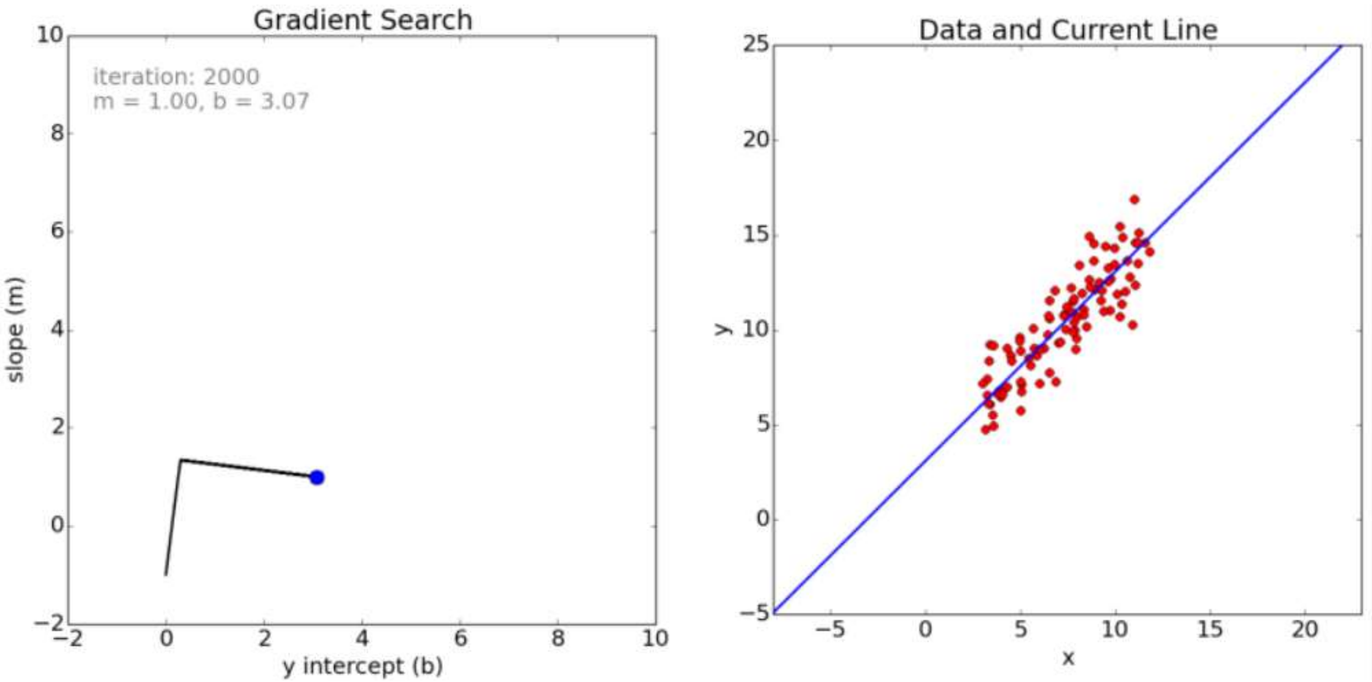
Парная регрессия



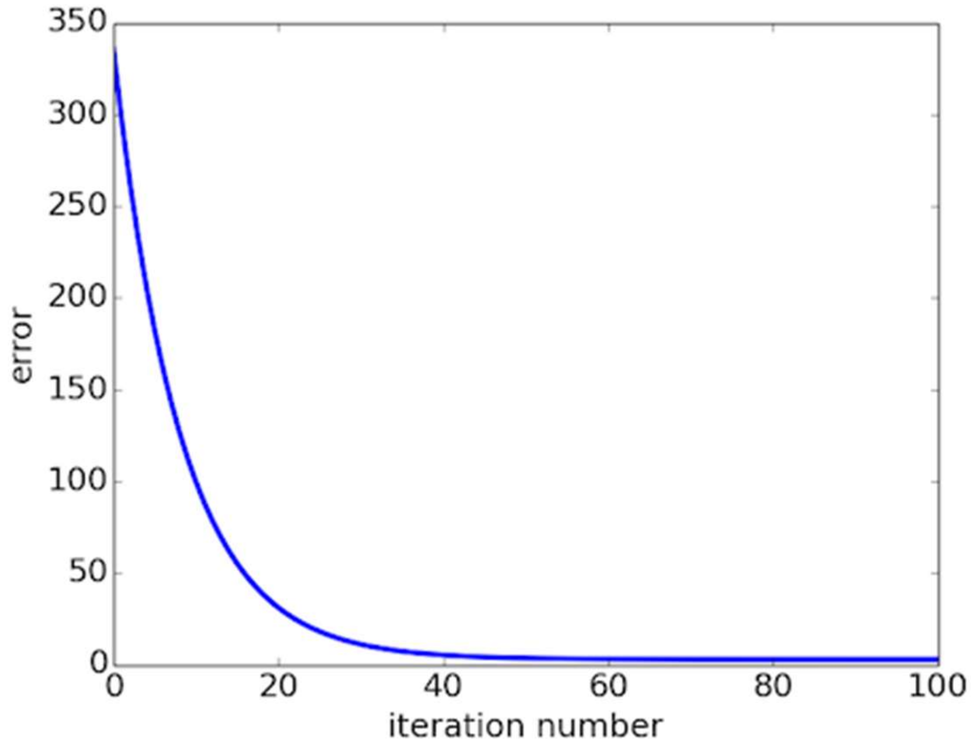
Парная регрессия



Парная регрессия

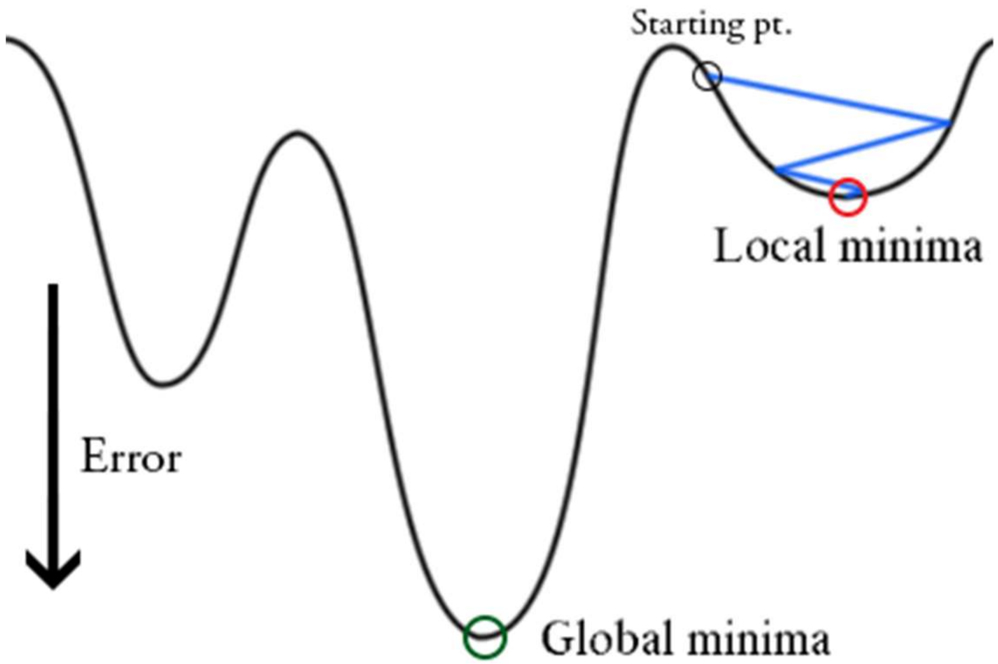


Функционал ошибки

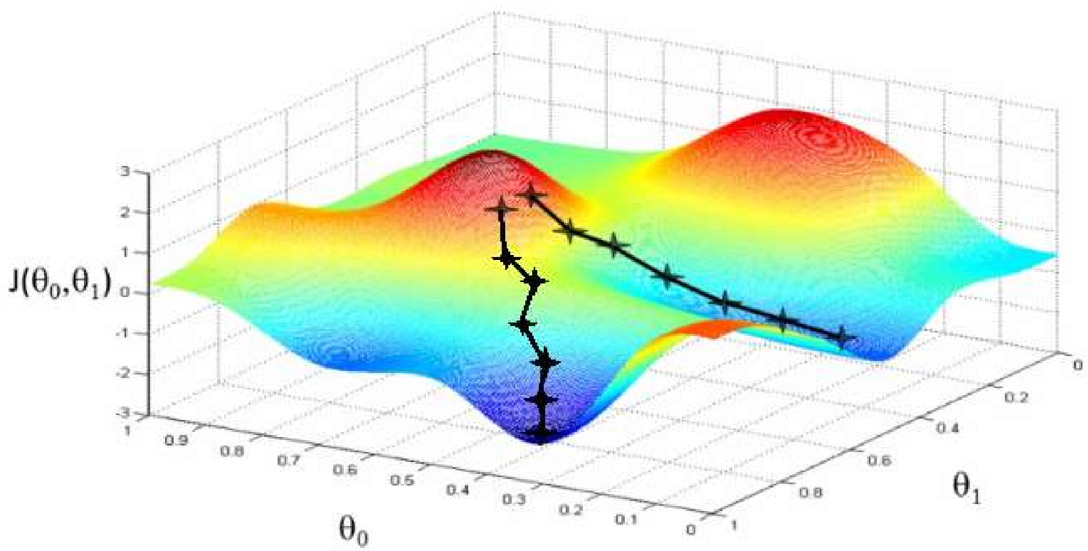


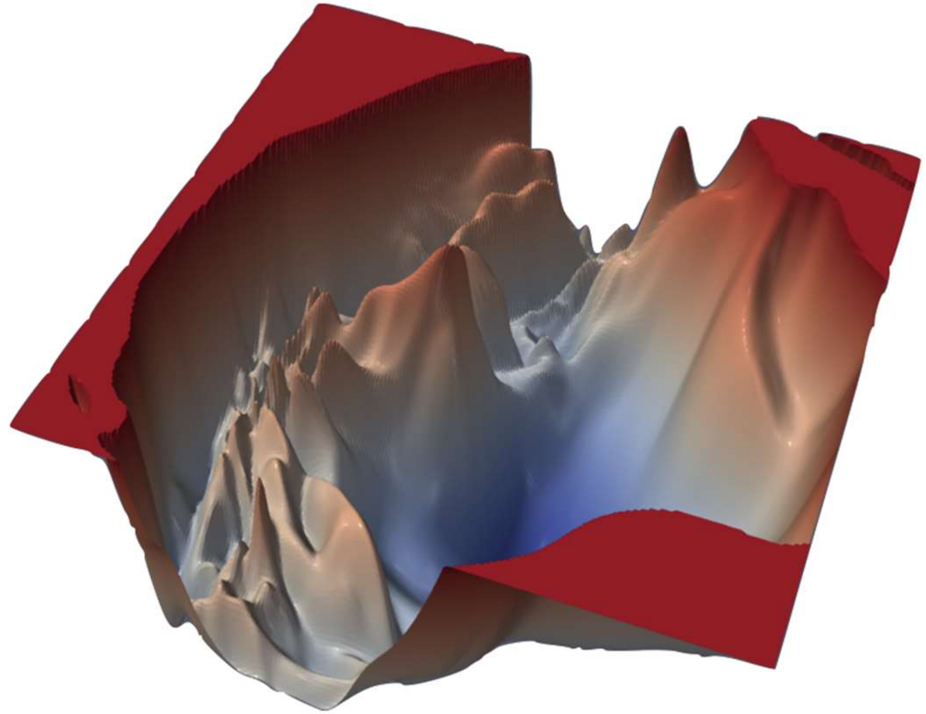
Локальные минимумы

- Градиентный спуск находит только локальные минимумы



Локальные минимумы





Длина шага

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

- Позволяет контролировать скорость обучения

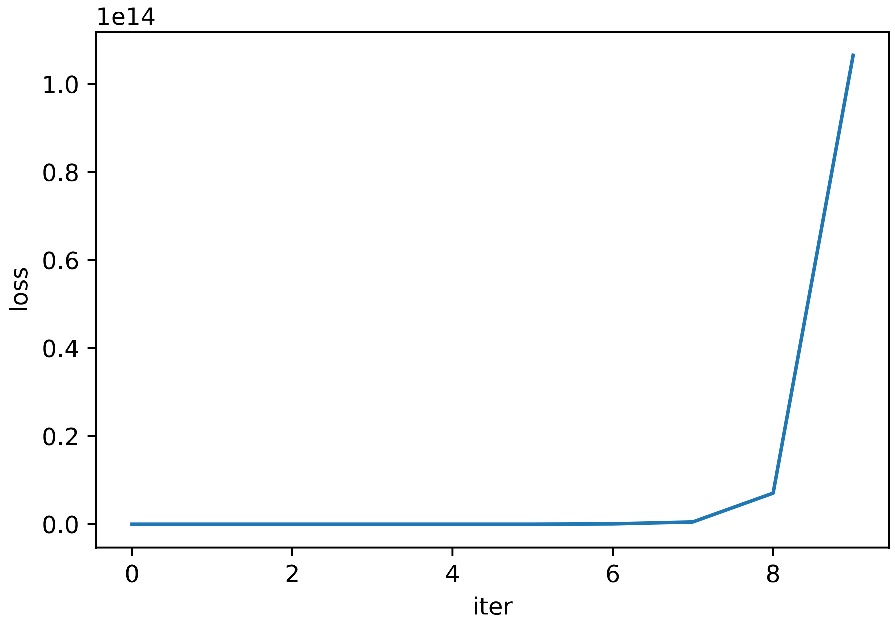
- Если сделать длину шага недостаточно маленькой, градиентный спуск может разойтись

- Длина шага — гиперпараметр, который нужно подбирать

Длина шага

Градиент на первом шаге:

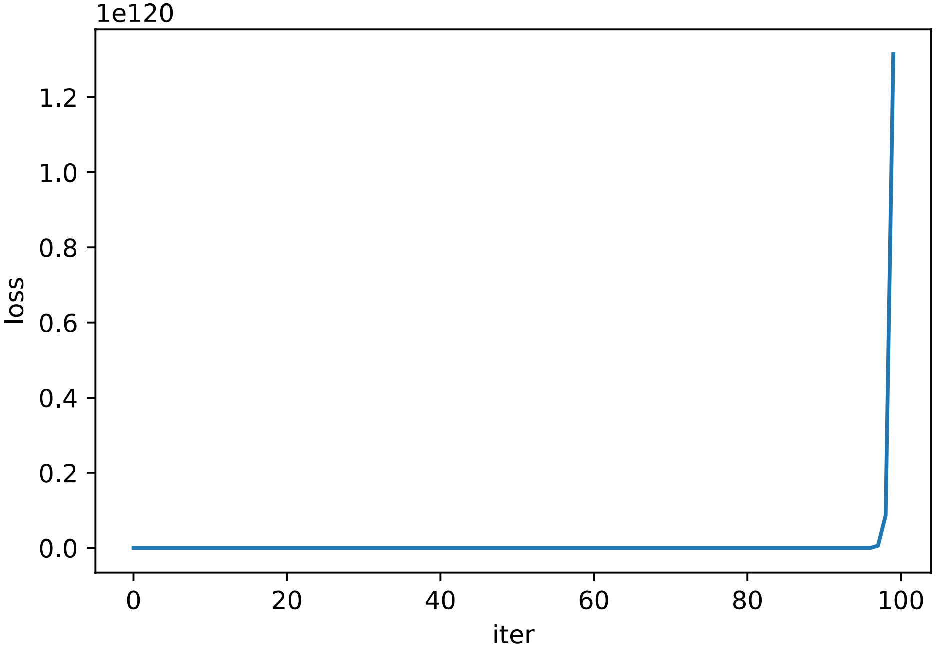
[26.52, 564.80, 682.90, 5097.71, 12110.87]



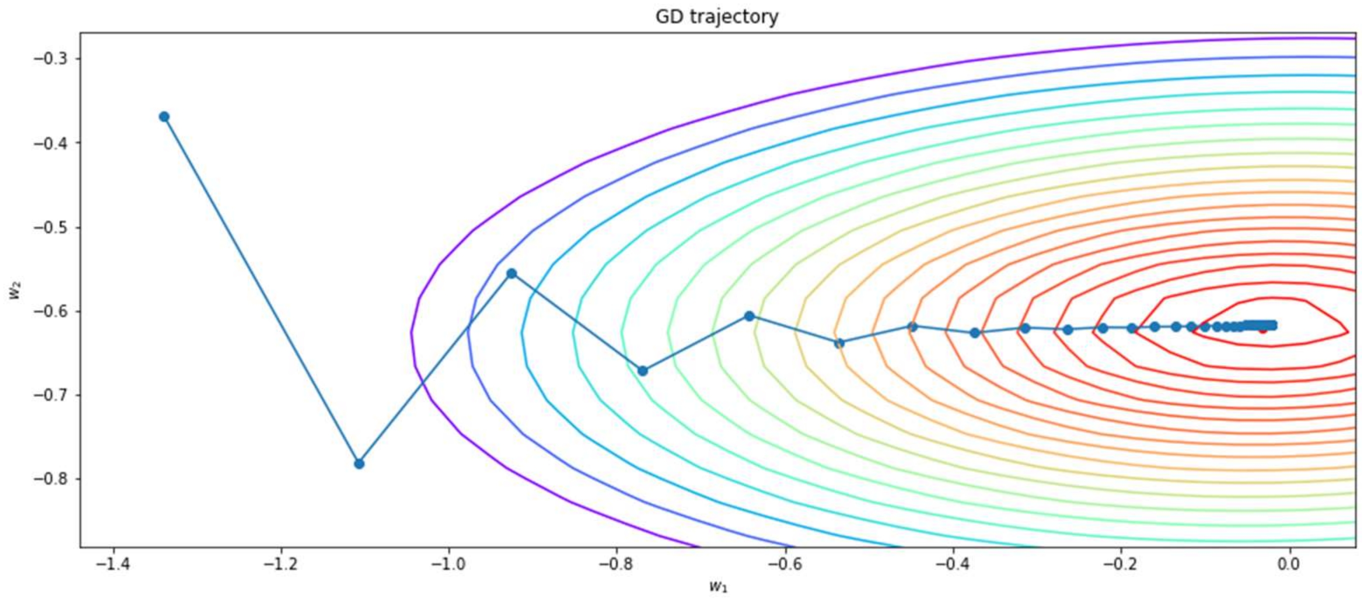
Длина шага

Градиент на первом шаге:

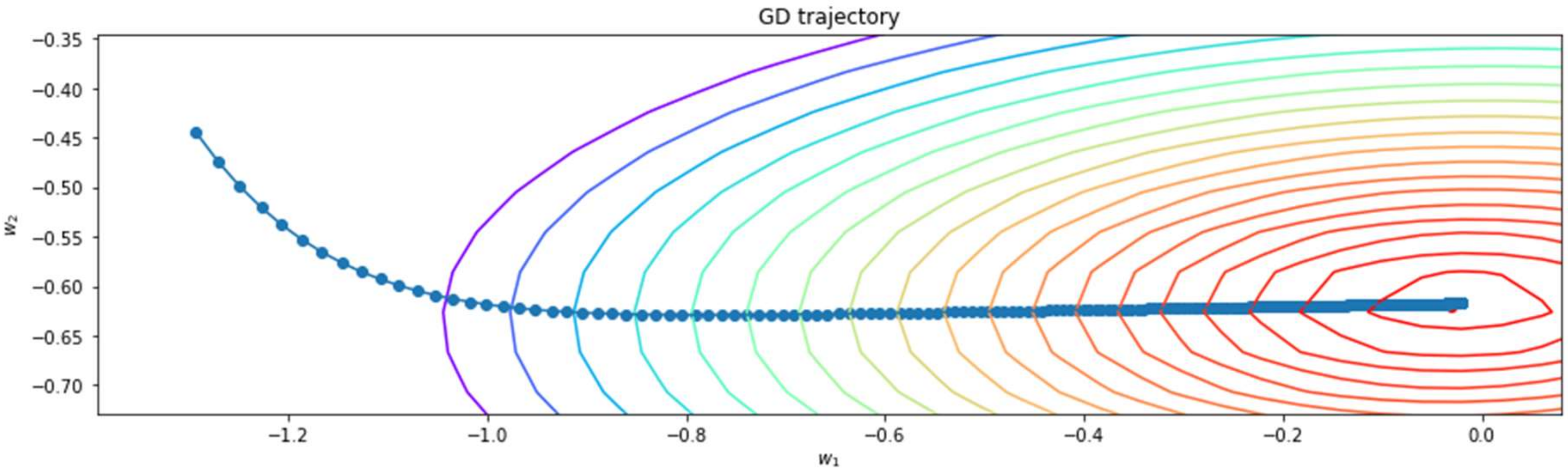
[26.52, 564.80, 682.90, 5097.71, 12110.87]



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

Переменная длина шага

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

- Длину шага можно менять в зависимости от шага
- Например: $\eta_t = \frac{1}{t}$
- Шаг наискорейшего спуска:

$$\eta_t = \arg\min_\eta Q(w^t) = \arg\min_\eta Q\left(w^{t-1} - \eta \nabla Q(w^{t-1})\right)$$

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

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

- Градиент:

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

- Может, оценить градиент одним слагаемым?

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

Стохастический градиентный спуск

1. Начальное приближение: w^0

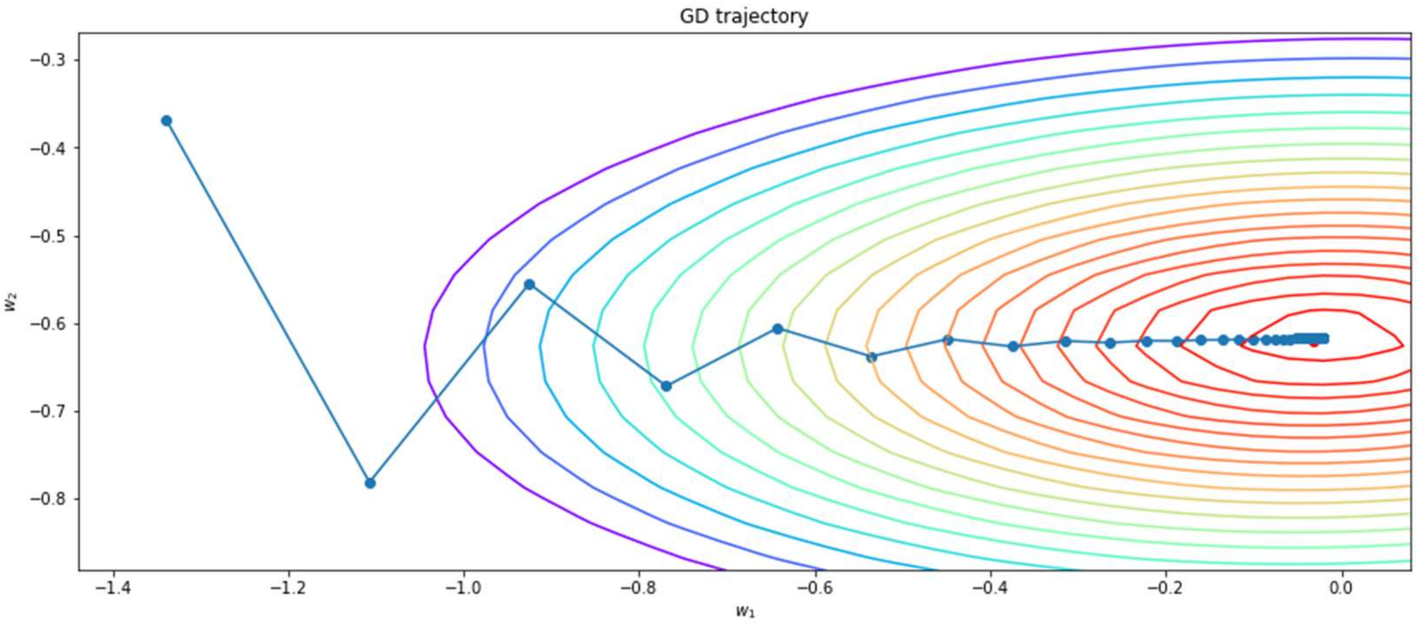
2. Повторять, каждый раз выбирая случайный объект i_t:

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

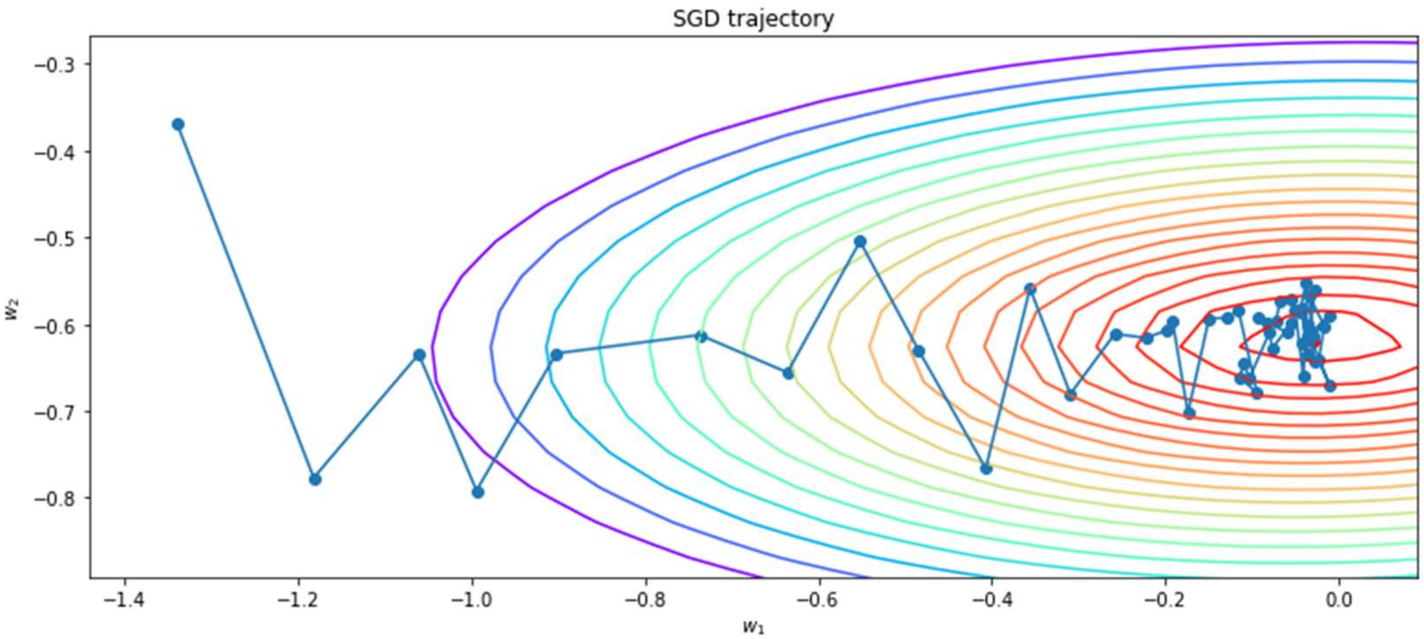
3. Останавливаемся, если

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

Градиентный спуск



Стохастический градиентный спуск



Стохастический градиентный спуск

1. Начальное приближение: w^0

2. Повторять, каждый раз выбирая случайный объект i_t:

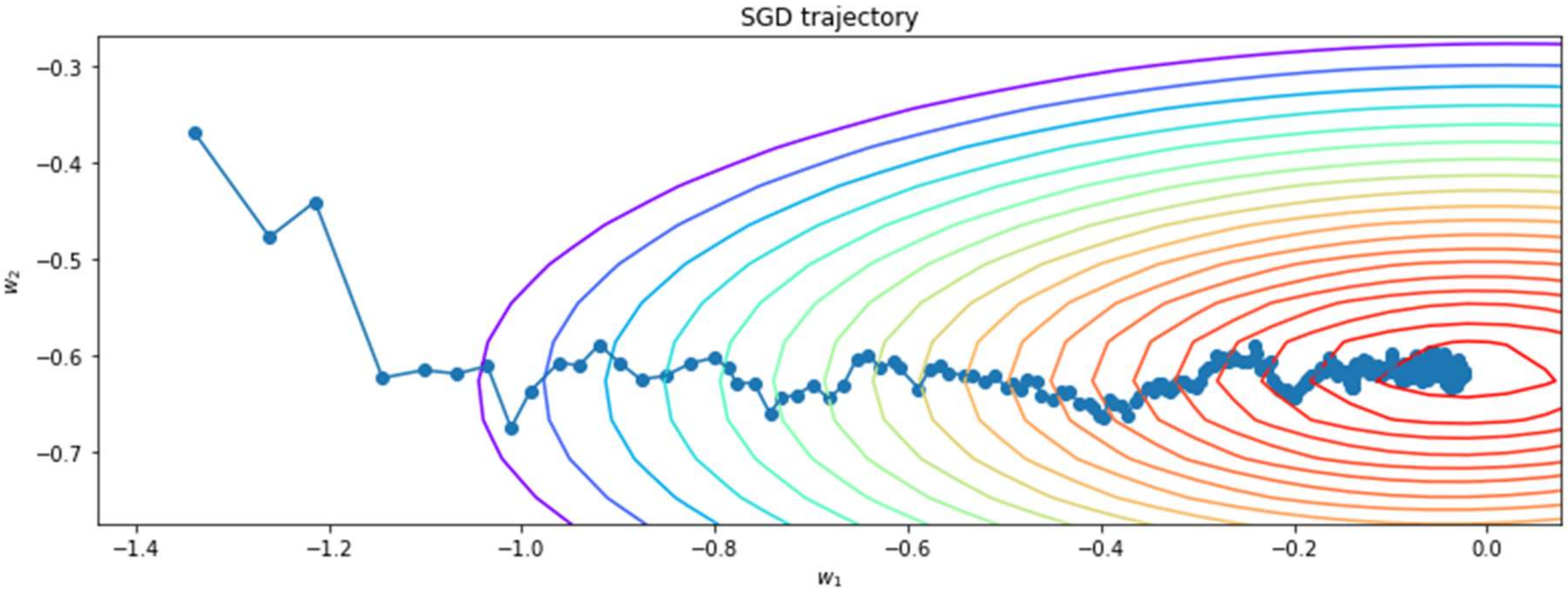
$$w^t = w^{t-1} - \eta_t \nabla L\left(y_{i_t}, a(x_{i_t})\right)$$

3. Останавливаемся, если

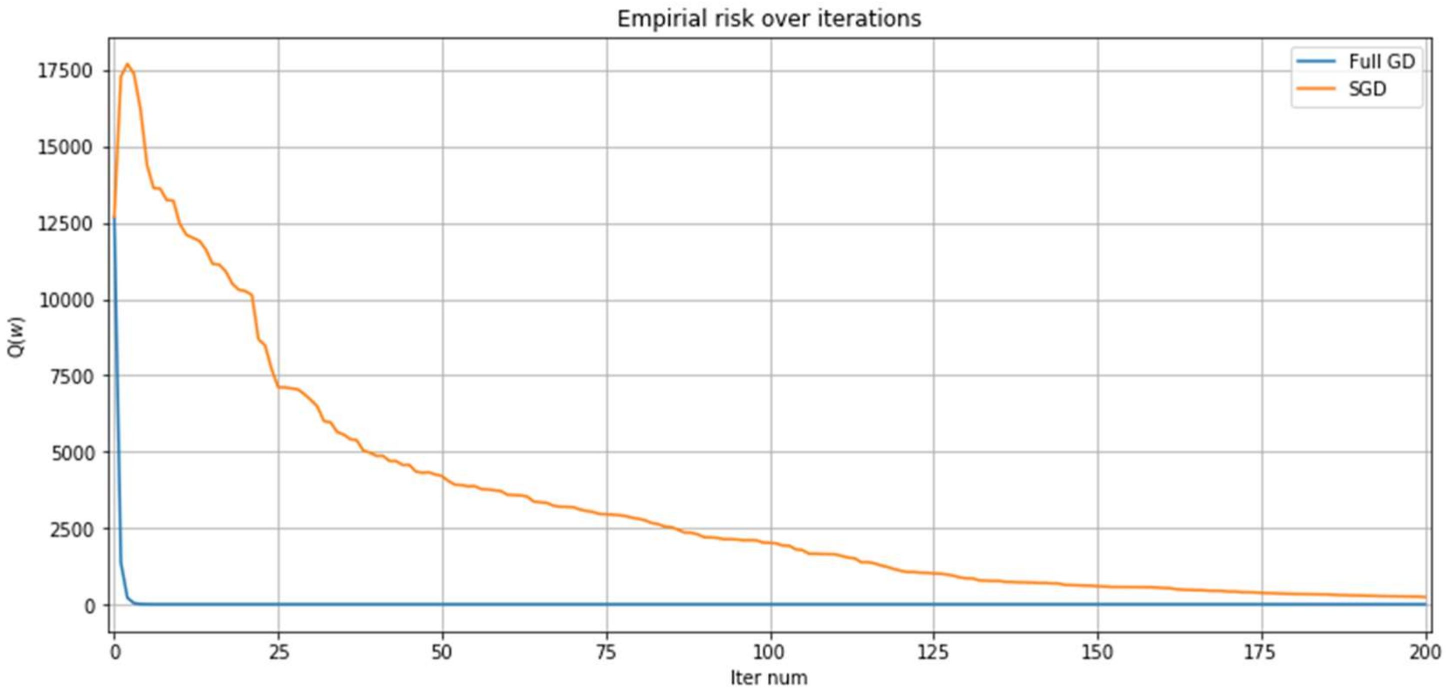
$$\|w^t - w^{t-1}\| < \varepsilon$$

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

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

Масштабирование признаков

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

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

- 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

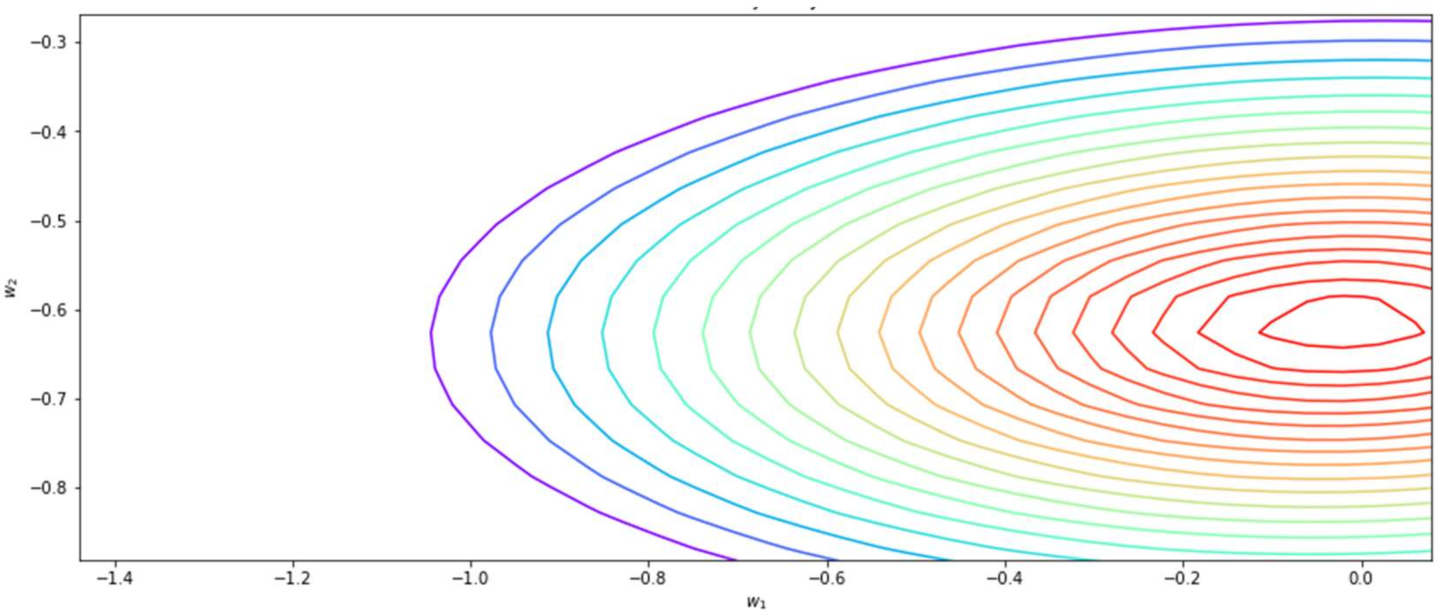
- Standard scaling:

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

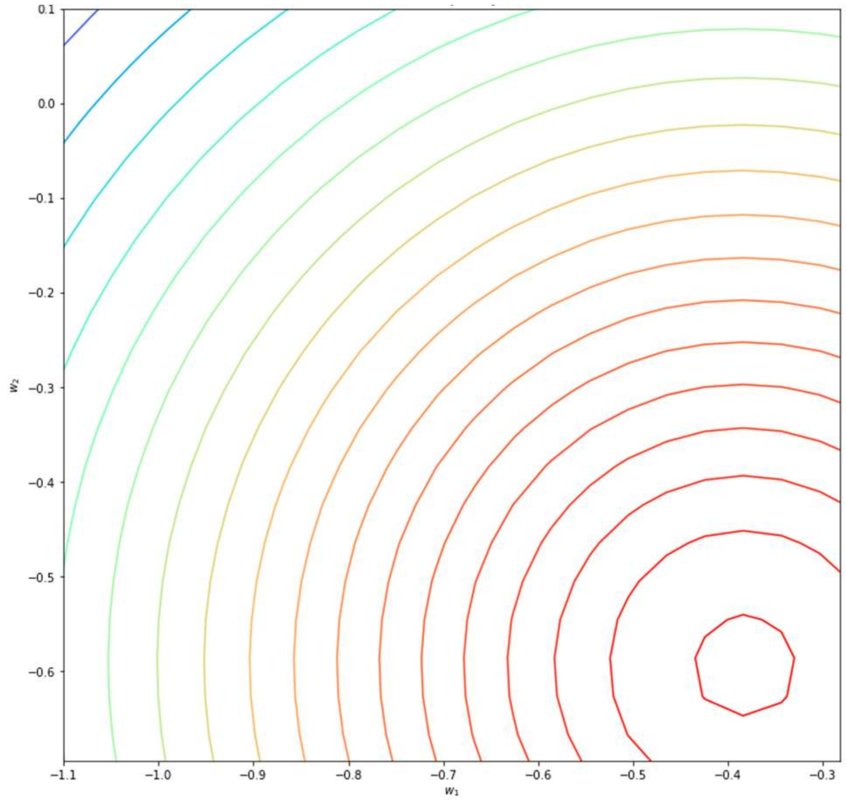
- Min-max scaling:

$$x_i^j := \frac{x_i^j - \min(x^j)}{\max(x^j) - \min(x^j)}$$

Масштабирование признаков



Масштабирование признаков



Масштабирование признаков

- До масштабирования значения градиента, соответствующие большим признакам, преобладают над остальными

- После масштабирования все параметры обновляются в равных пропорциях