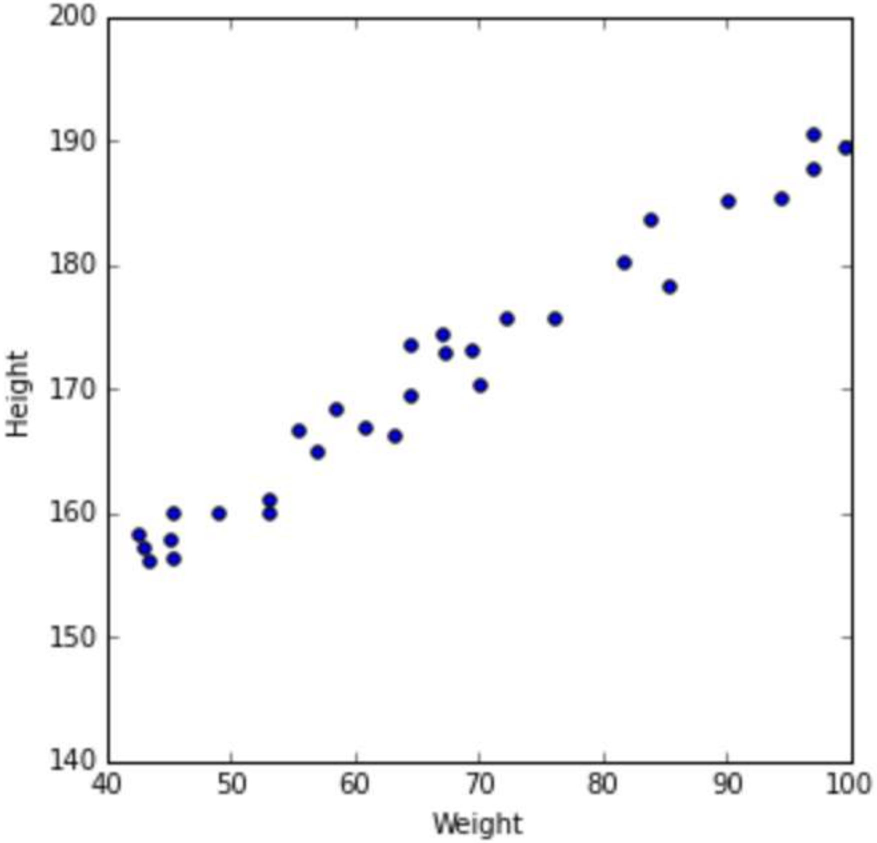
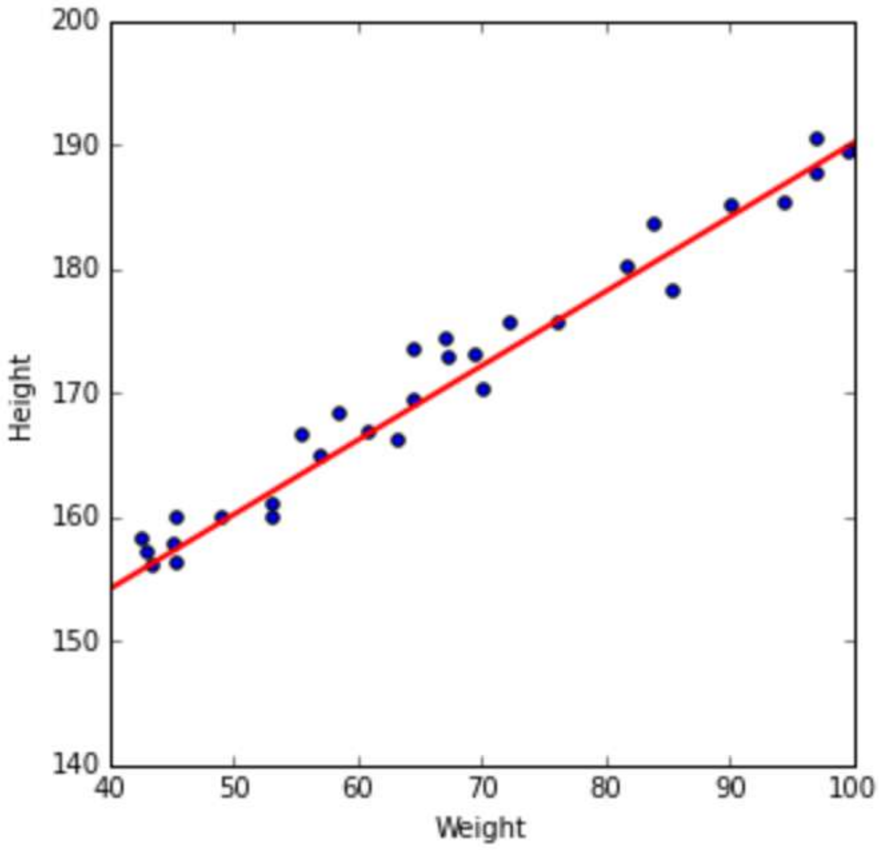
Введение в анализ данных

Лекция 3. Линейная регрессия

Парная регрессия



Парная регрессия



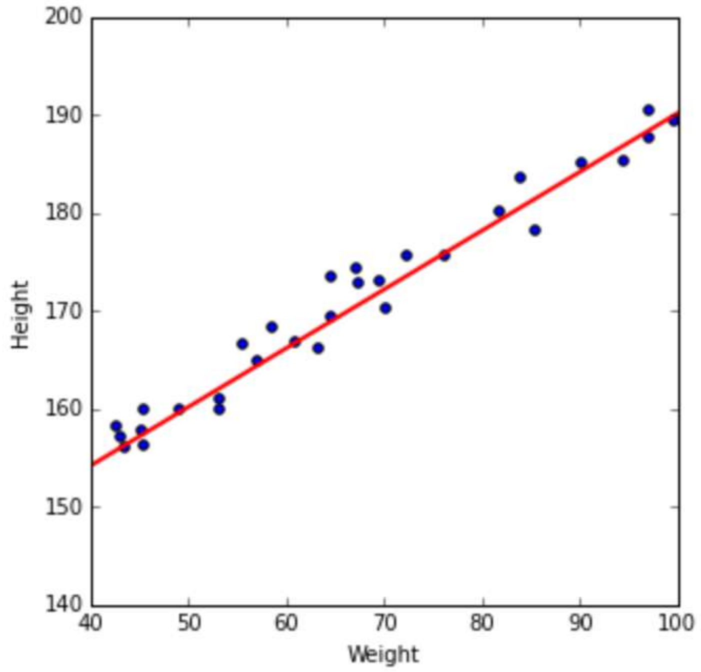
Парная регрессия

- 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*?

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

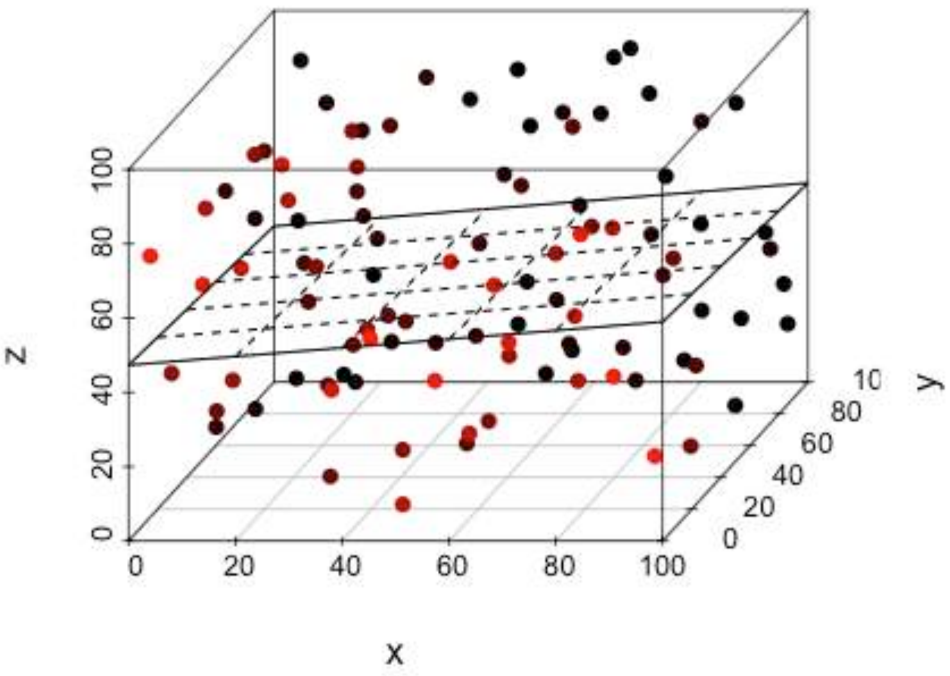
- $x = 1, a(x) = 3$
- $x = 2, a(x) = 5$
- $x = 10, a(x) = 21$
- $x = 20, a(x) = 41$



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

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

- Количество параметров: $d + 1$

Много признаков

- Общий случай: d признаков
- Модель

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

Веса/коэффициенты

Свободный коэффициент/сдвиг/bias

- Количество параметров: $d + 1$

Много признаков

Запишем через скалярное произведение:

$$a(x) = w_0 + w_1 x_1 + \dots + w_d x_d =$$
$$= w_0 + \langle w, x \rangle$$

Будем считать, что есть признак, всегда равный единице:

$$a(x) = w_1 x_1 + \dots + w_d x_d =$$
$$= w_1 * 1 + w_2 x_2 + \dots + w_d x_d =$$
$$= \langle w, x \rangle$$

Применимость линейной регрессии

Модель линейной регрессии

$$a(x) = w_1 x_1 + \cdots + w_d x_d = \langle w, x \rangle$$

- 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

$$a(x) = w_0 + w_1 * (\text{площадь})$$
$$+ w_2 * (\text{район})$$
$$+ w_3 * (\text{расстояние до метро})$$

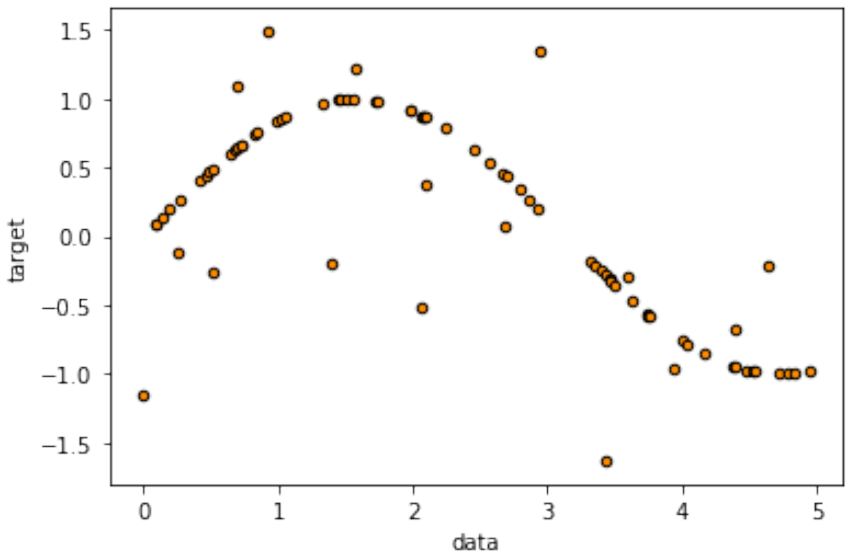
- Что-то странное

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

Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

→

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

Кодирование категориальных признаков

Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

$\rightarrow$

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

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

$$+ w_2 * (\text{квартира в ЦАО?})$$

$$+ w_3 * (\text{квартира в ЮАО?})$$

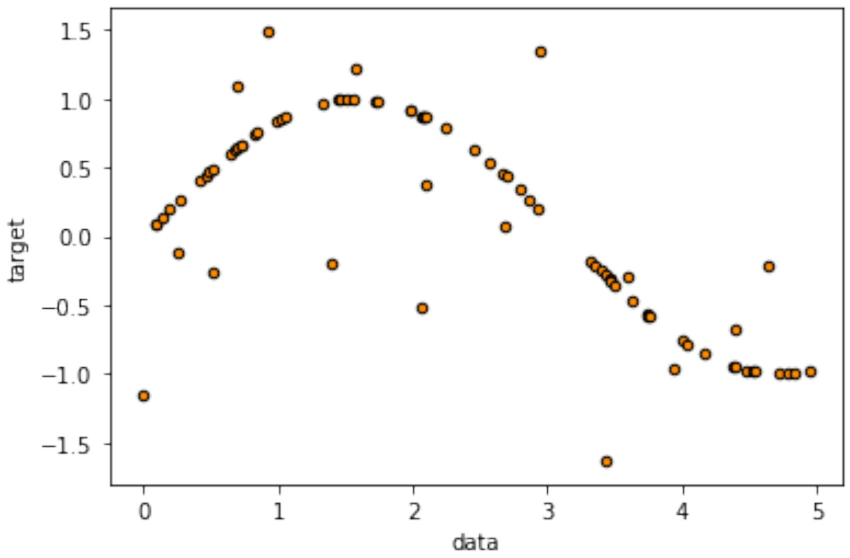
$$+ w_4 * (\text{квартира в САО?})$$

Предсказание стоимости квартиры

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

$$+ w_2 * (\text{район})$$

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

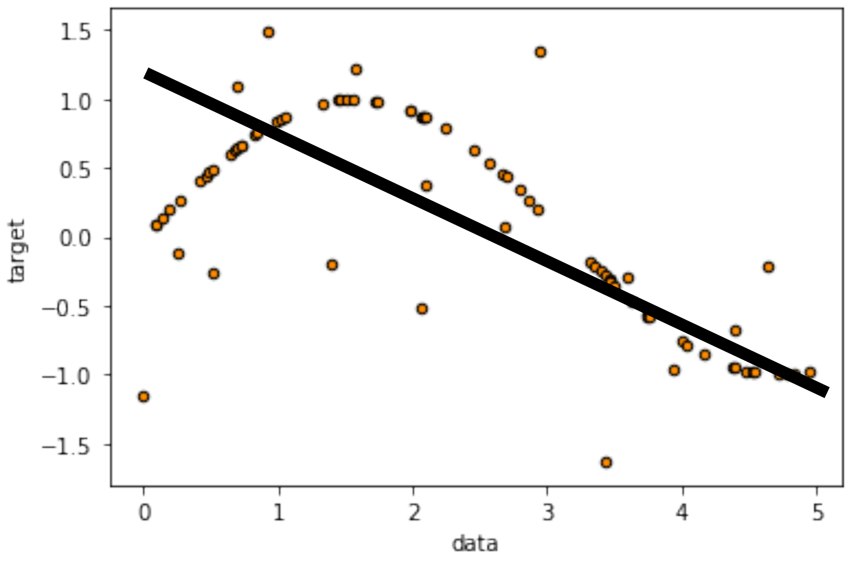


Предсказание стоимости квартиры

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

$$+ w_2 * (\text{район})$$

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

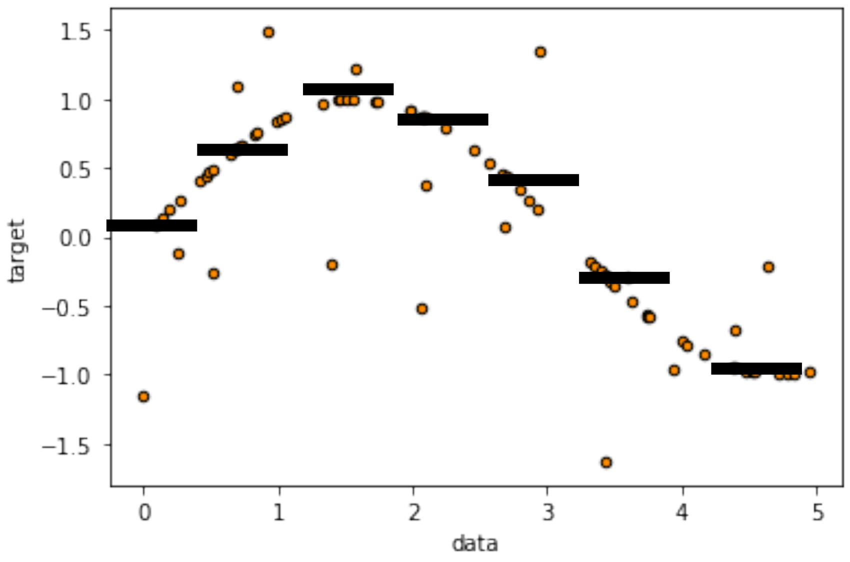


Предсказание стоимости квартиры

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

$$+ w_2 * (\text{район})$$

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

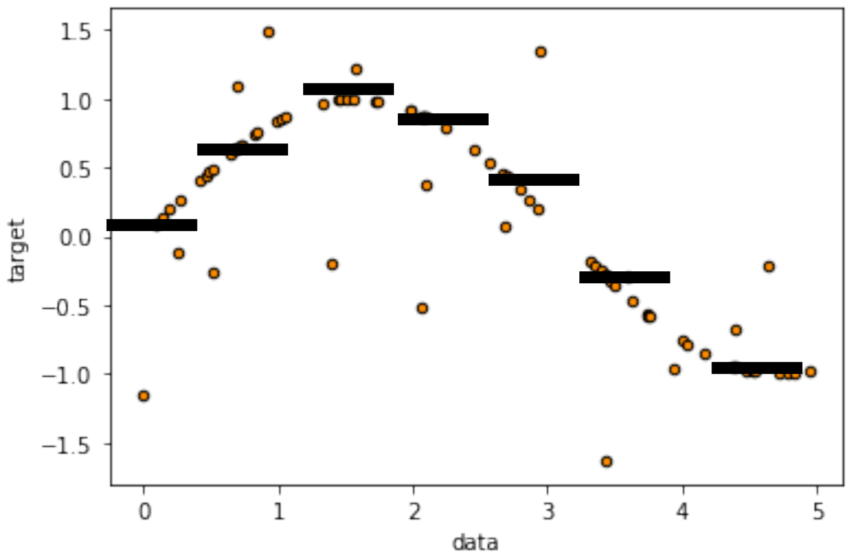


Предсказание стоимости квартиры

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

$$+ w_2 * (\text{район})$$

$$+ w_3 * [t_0 \le x_3 < t_1] + \cdots + w_{3+n}[t_{n-1} \le x_3 < t_n]$$



Нелинейные признаки

- Линейная модель с полиномиальными признаками:

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

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:

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

Применение линейной модели

- $a(x) = \langle w, x \rangle = w_1 x_1 + \dots + w_d x_d$
- Как применить модель к обучающей выборке?

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

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

$$\begin{pmatrix} \sum_{i=1}^{d} w_i x_{1i} \\ \sum_{i=1}^{d} w_i x_{2i} \\ \vdots \\ \sum_{i=1}^{d} w_i x_{\ell i} \end{pmatrix}$$

Модель линейной регрессии

- Среднеквадратичная ошибка и задача обучения:

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

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

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

- Вычисление MSE в NumPy:

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

Обучение линейной регрессии

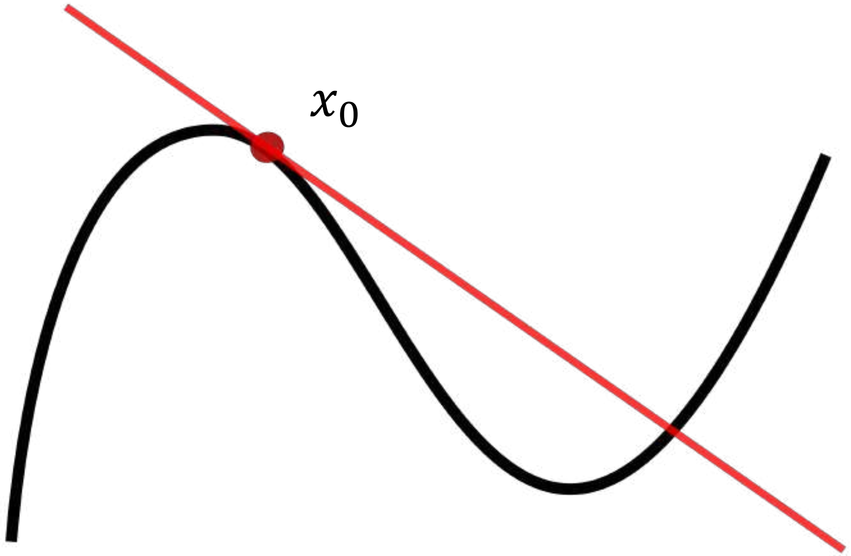
Среднеквадратичная ошибка

- MSE для линейной регрессии:

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

Производная

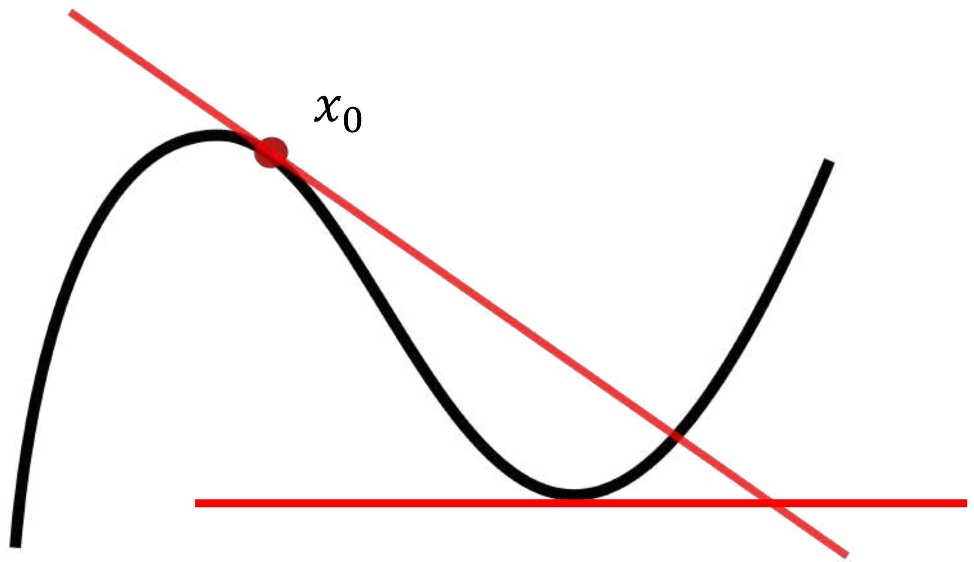
$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$



Производная

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

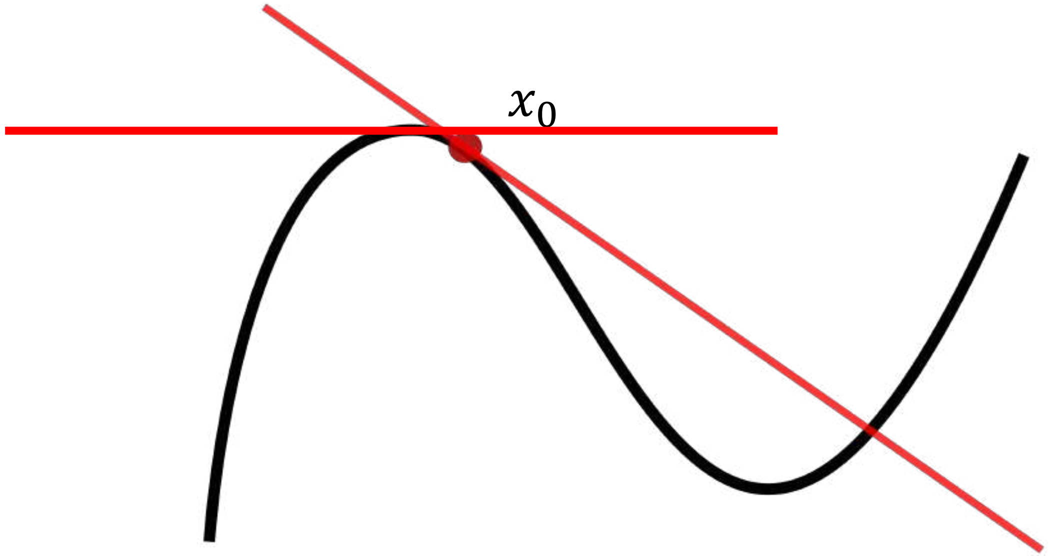
$$f'(x_0) = 0$$



Производная

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

$$f'(x_0) = 0$$



Градиент

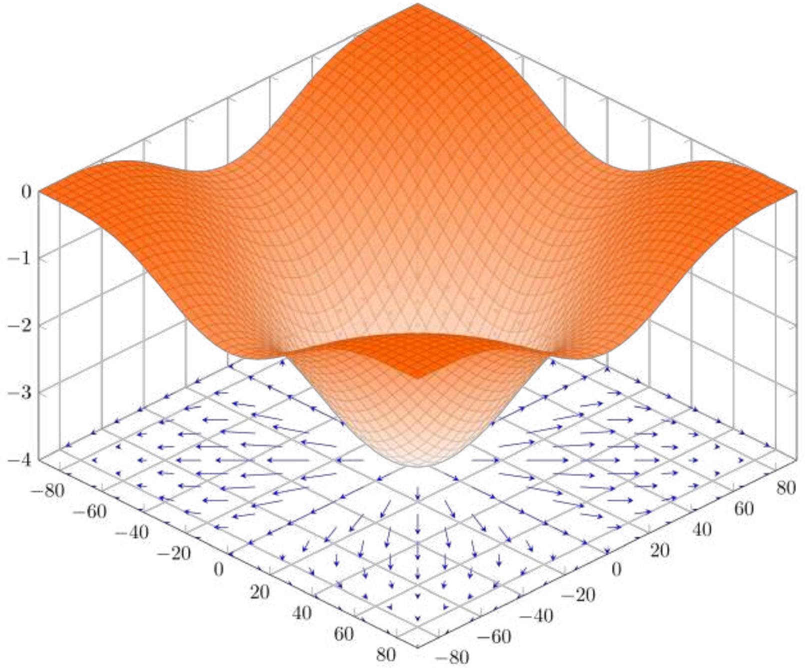
- Градиент — вектор частных производных

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

- 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!
- 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

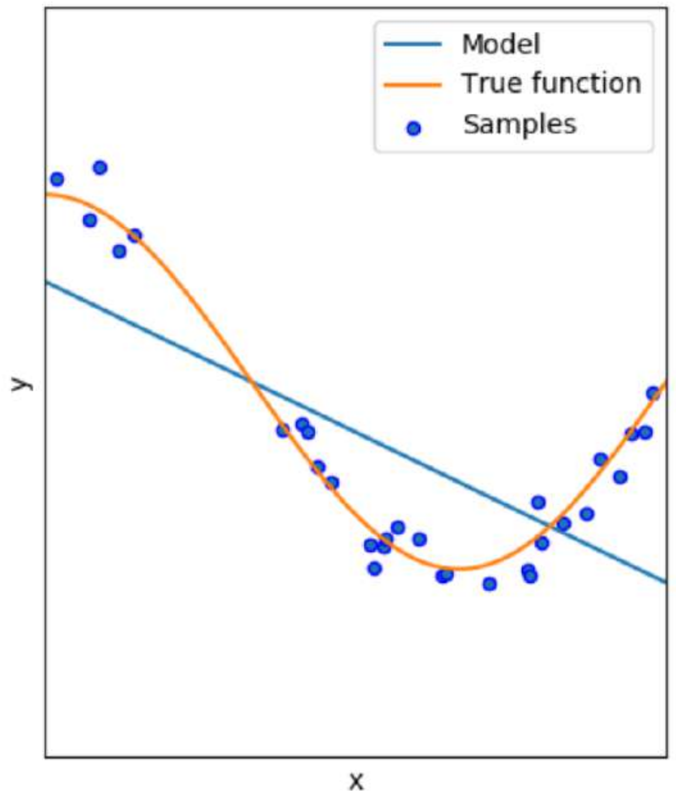
$$w = (X^T X)^{-1} X^T y$$

- 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

Переобучение и регуляризация линейных моделей

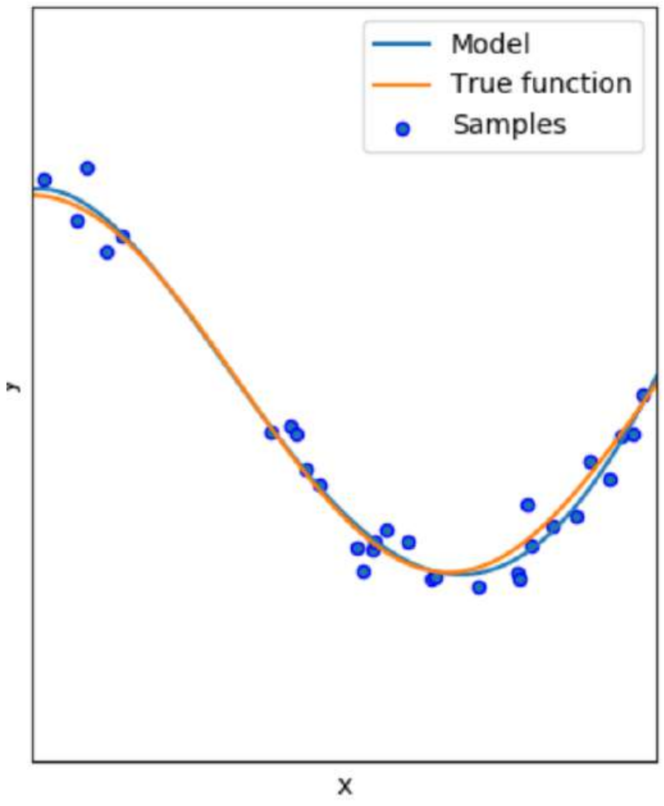
Нелинейная задача

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



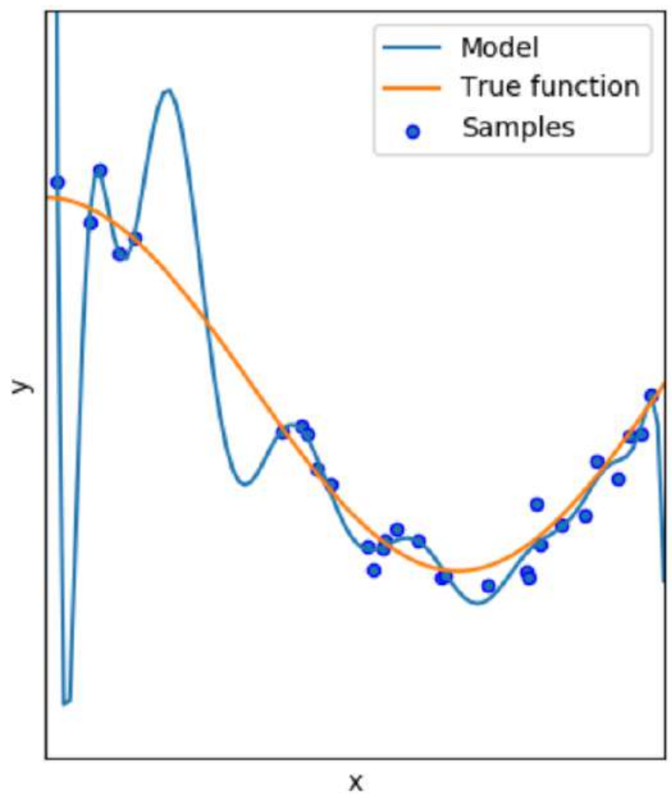
Нелинейная задача

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



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

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

- 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 не 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

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

- λ — коэффициент регуляризации

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:

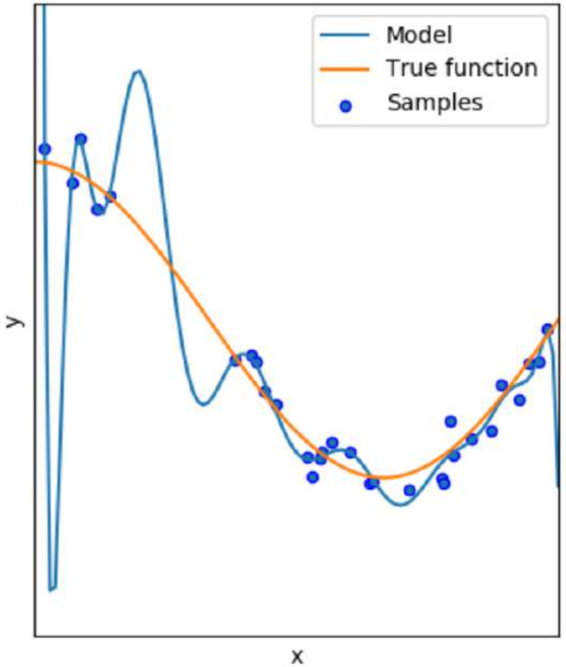
$$w = (X^T X + \lambda I)^{-1} X^T y$$

- Гребневая регрессия (Ridge regression)

Эффект регуляризации

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

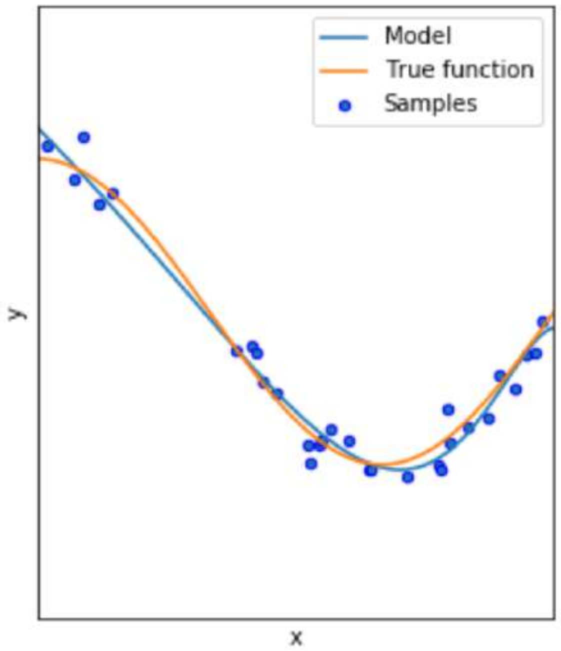
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 \to \min_w$$



Эффект регуляризации

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

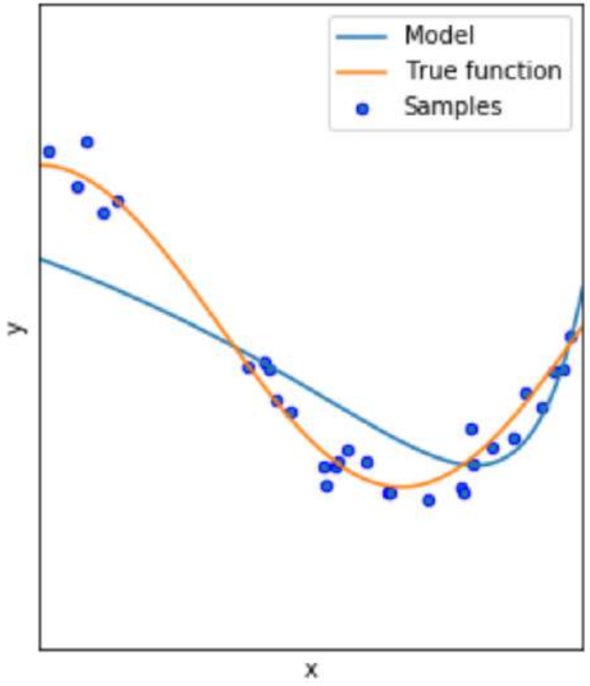
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + 0.01 \, \|w\|^2 \to \min_w$$



Эффект регуляризации

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

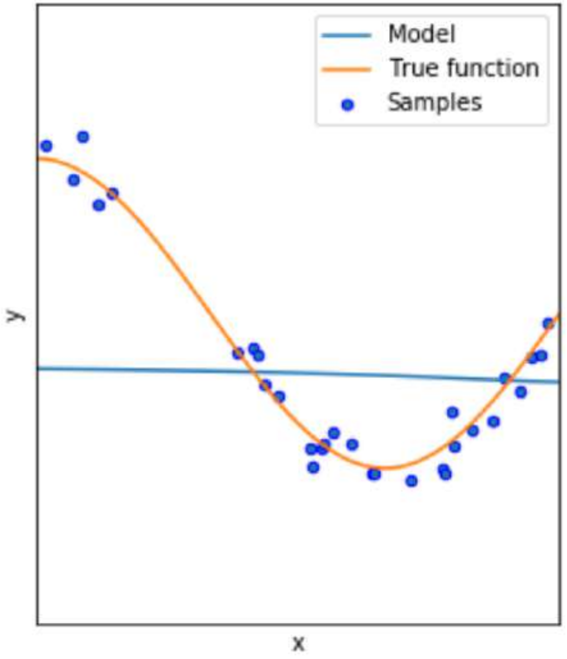
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + 1 \, \|w\|^2 \to \min_w$$



Эффект регуляризации

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

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



Лассо

- Регуляризованный функционал

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

- 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

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

- 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

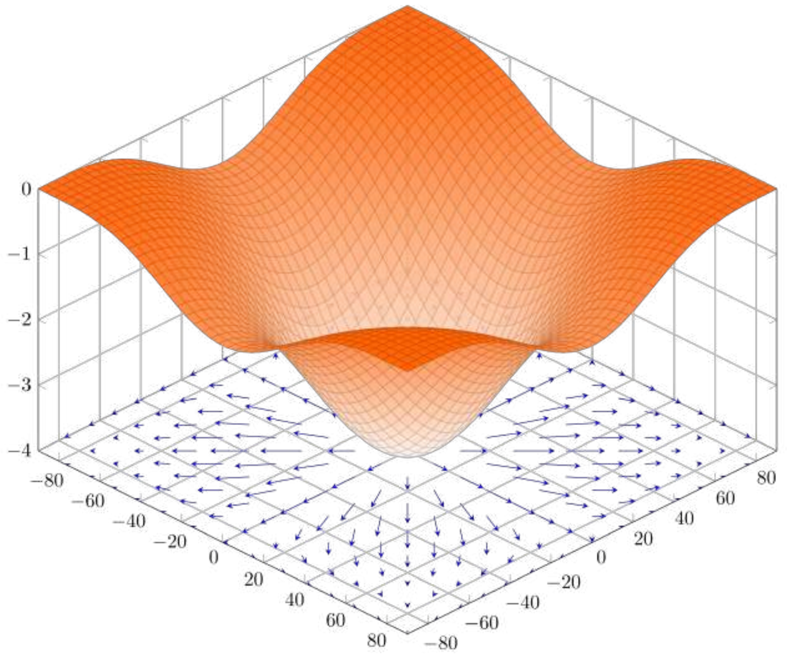
- Градиент — вектор частных производных

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

- У градиента есть важное свойство!

Важное свойство

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

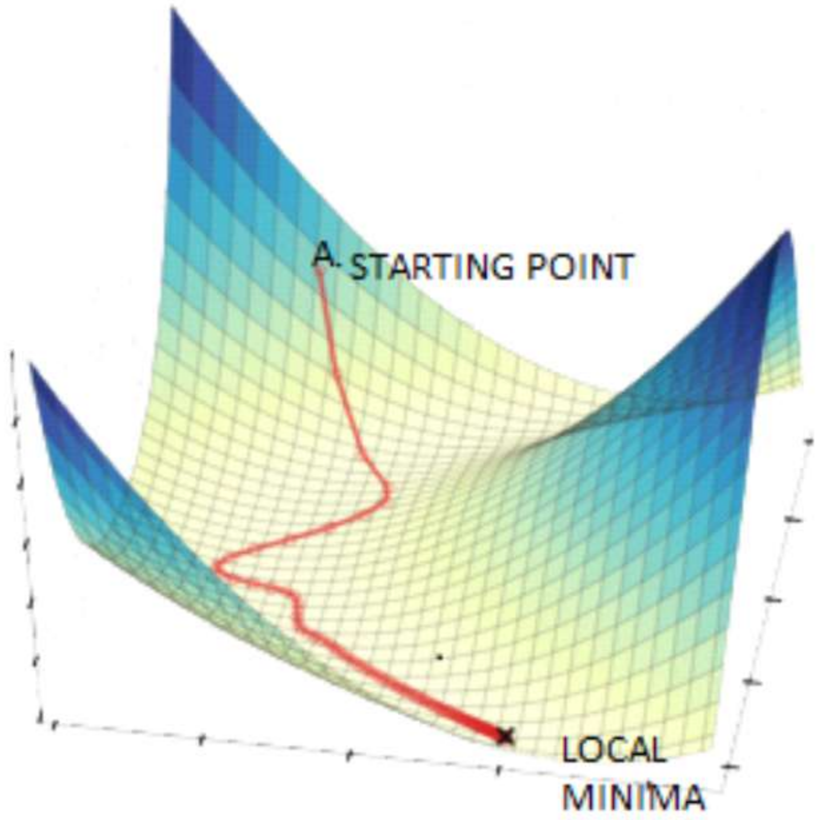


Важное свойство

- 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

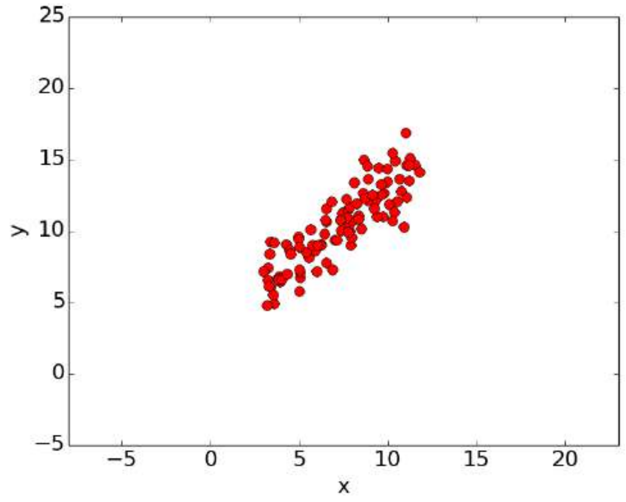
- Стартуем из случайной точки
- Сдвигаемся по антиградиенту
- Повторяем, пока не окажемся в точке минимума

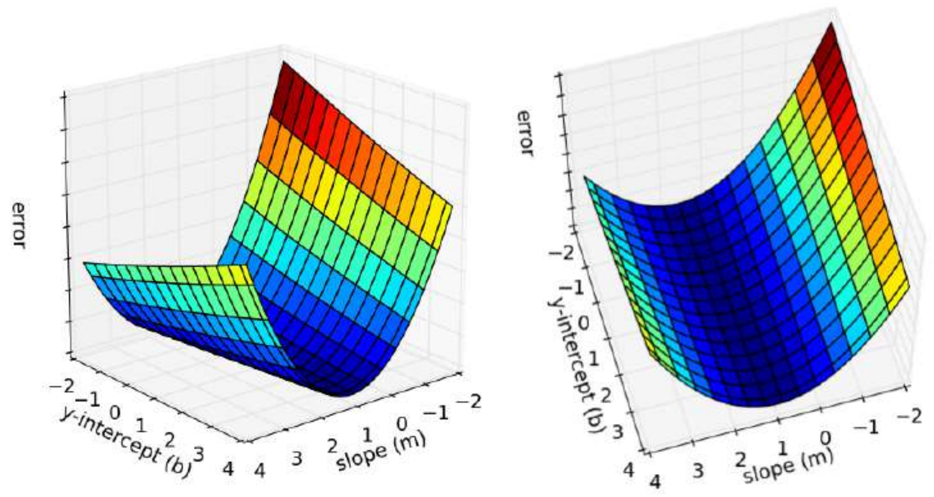
Парная регрессия

- 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

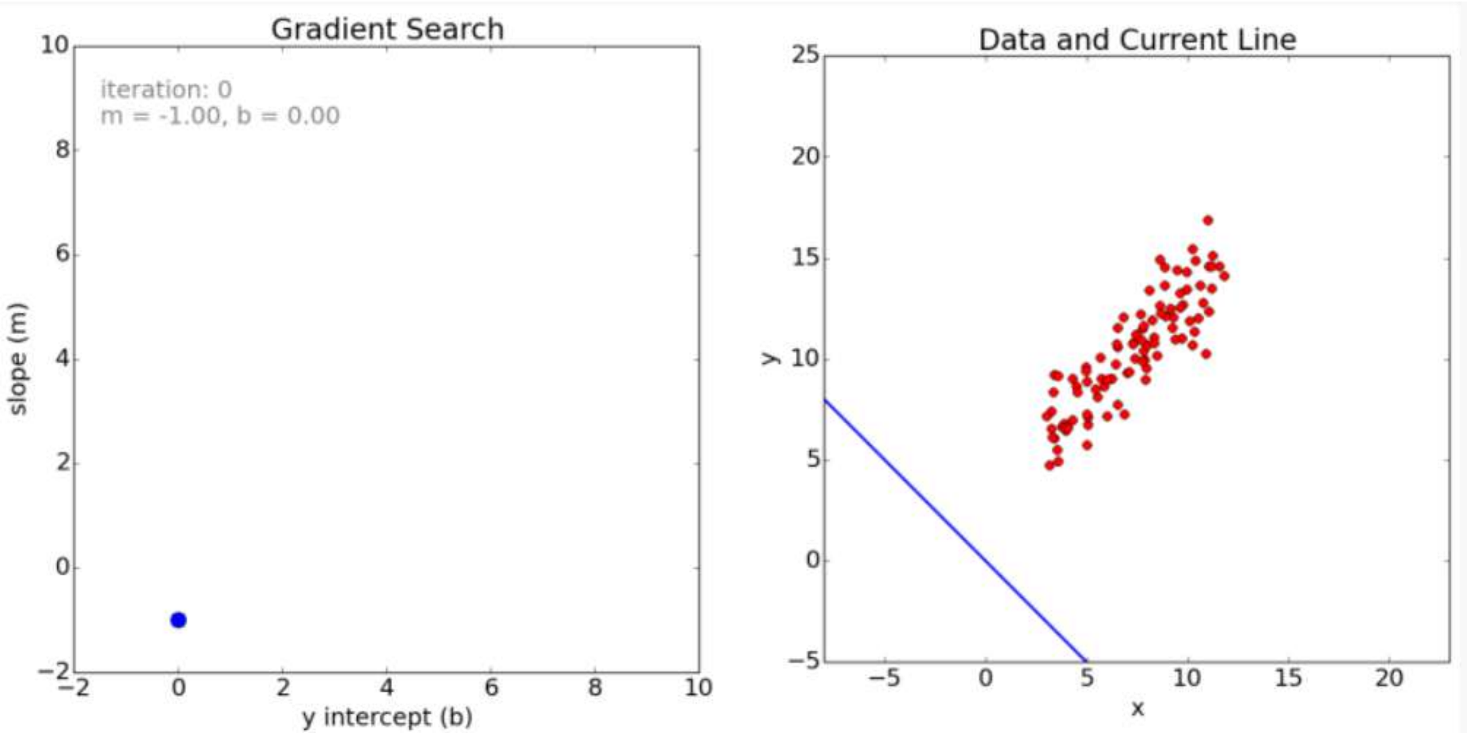
- Останавливаем процесс, если

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

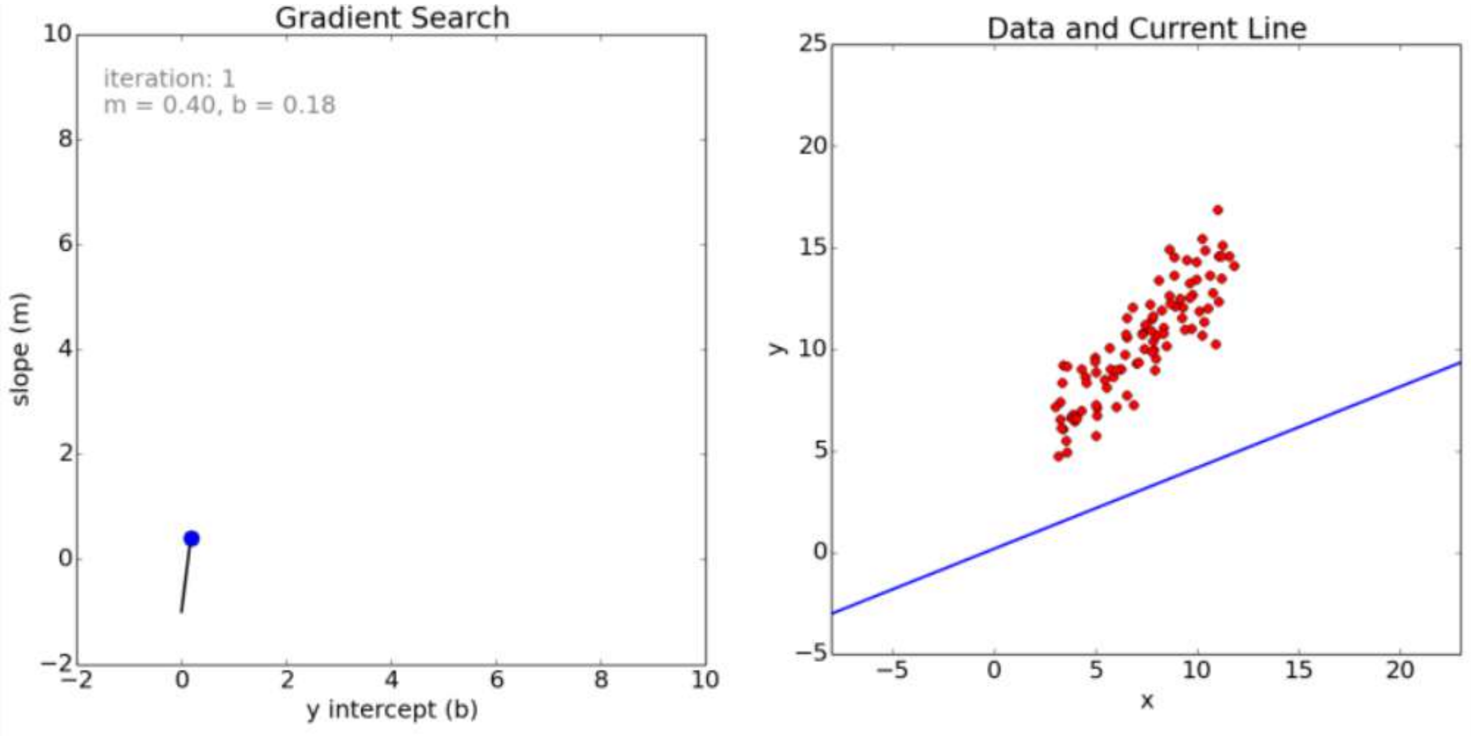
- Другой вариант:

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

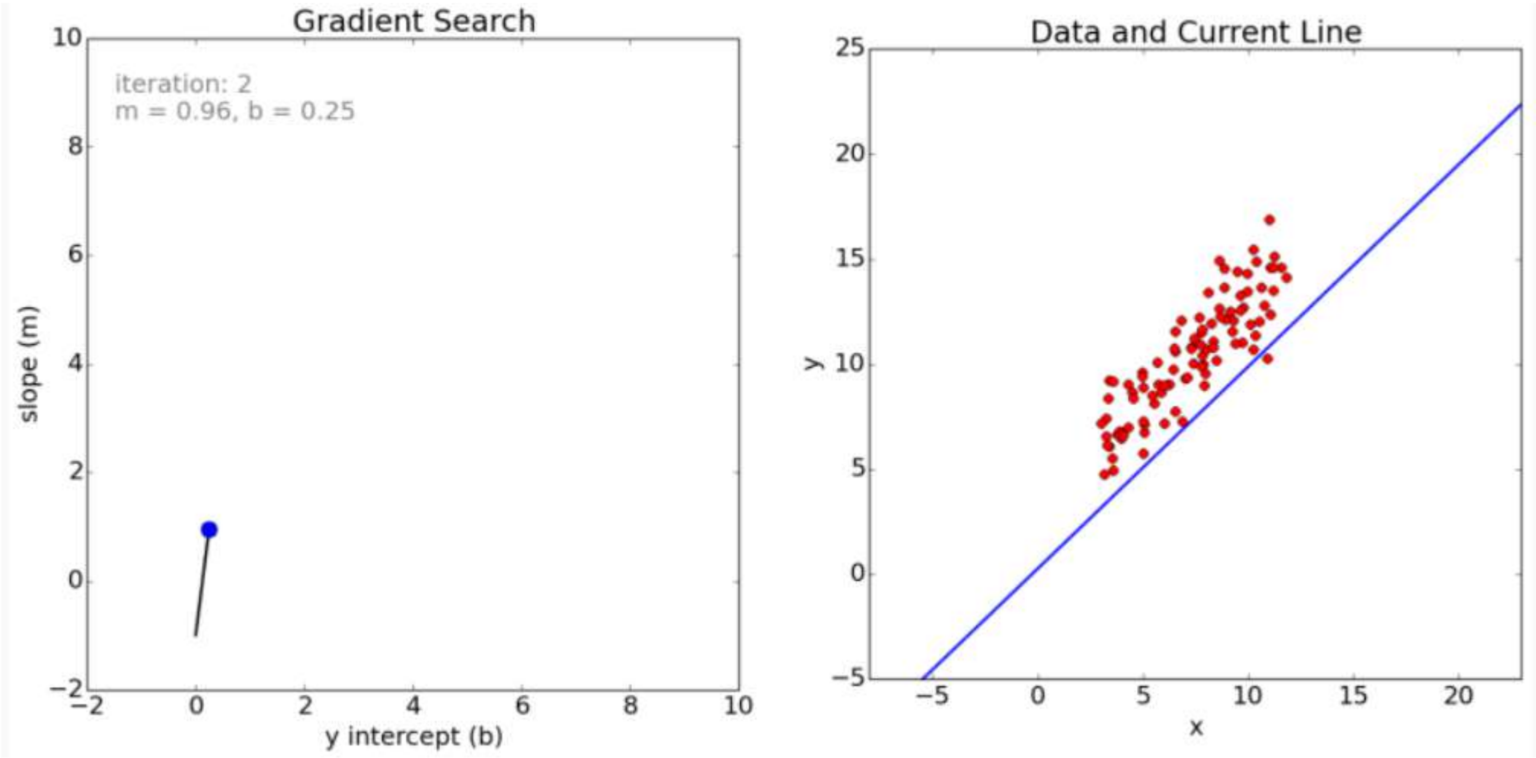
Парная регрессия



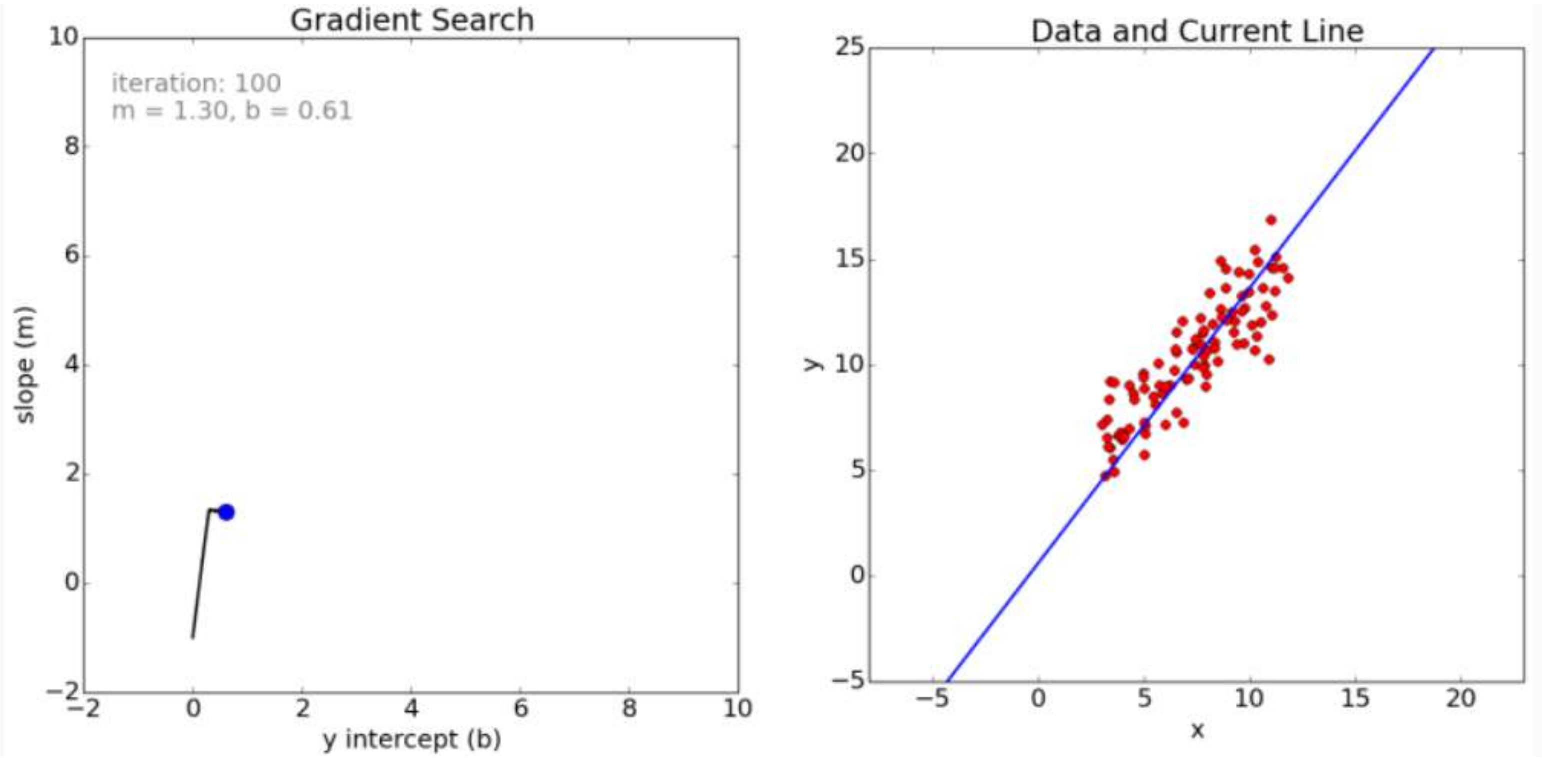
Парная регрессия



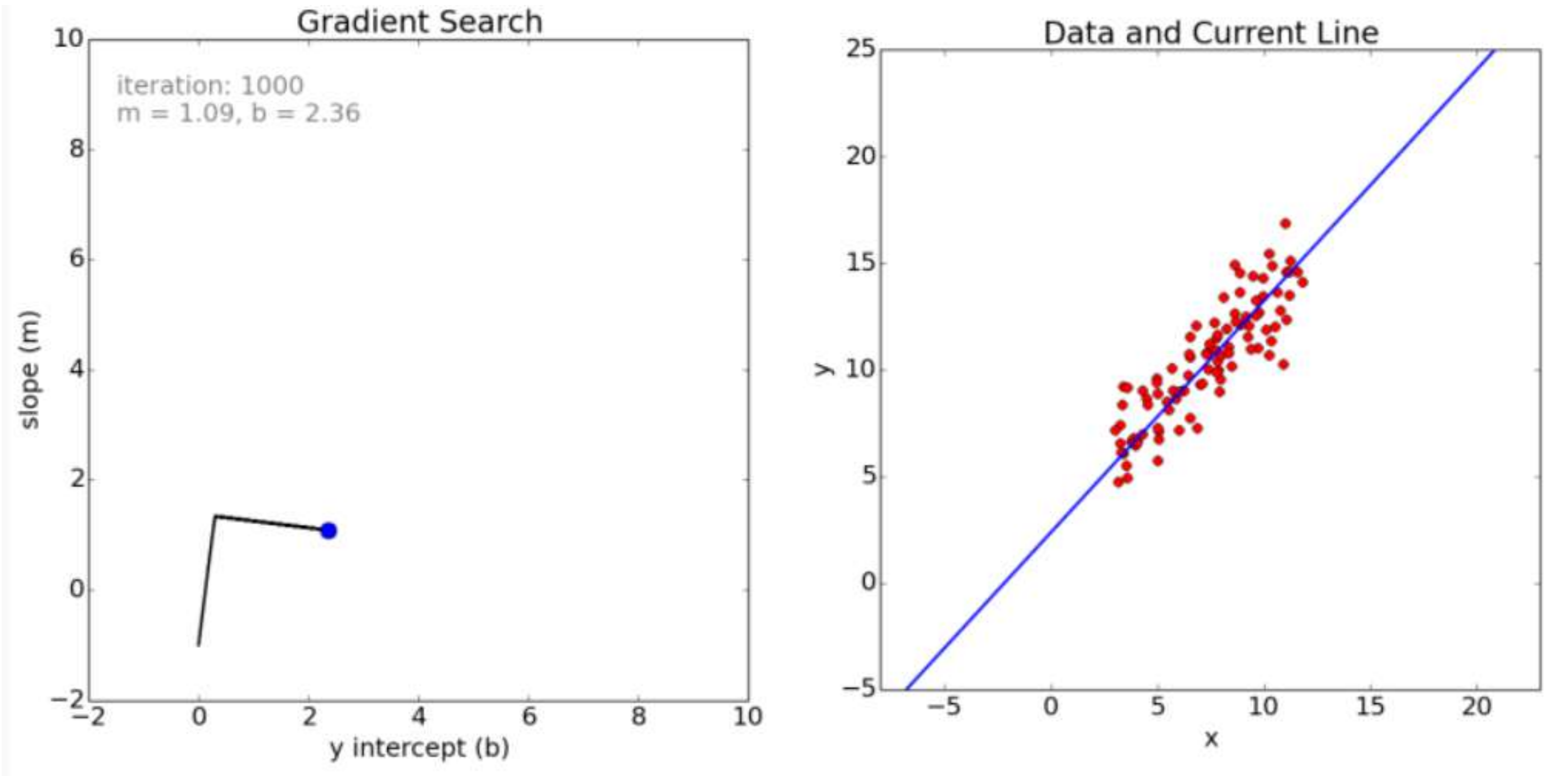
Парная регрессия



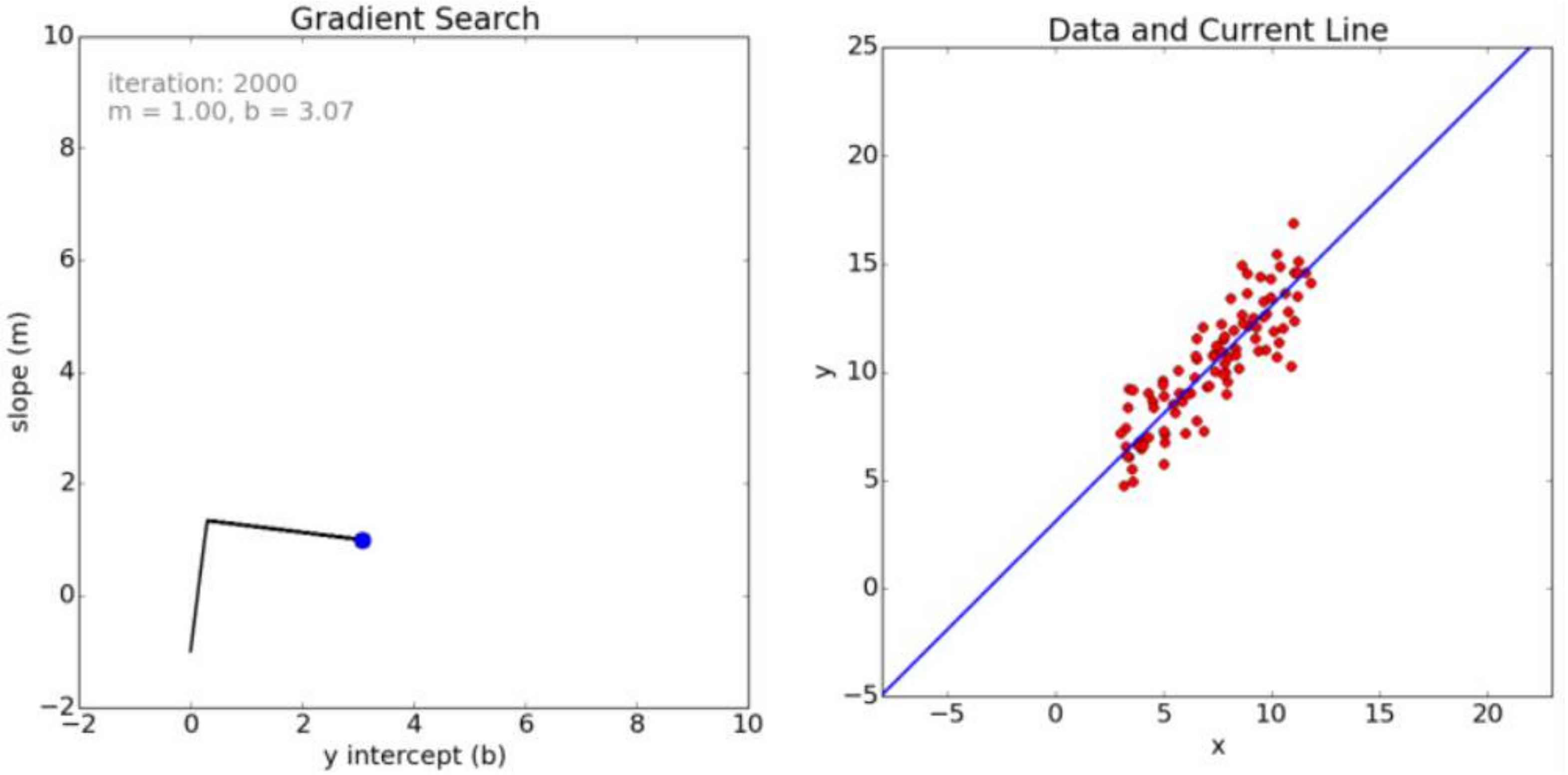
Парная регрессия

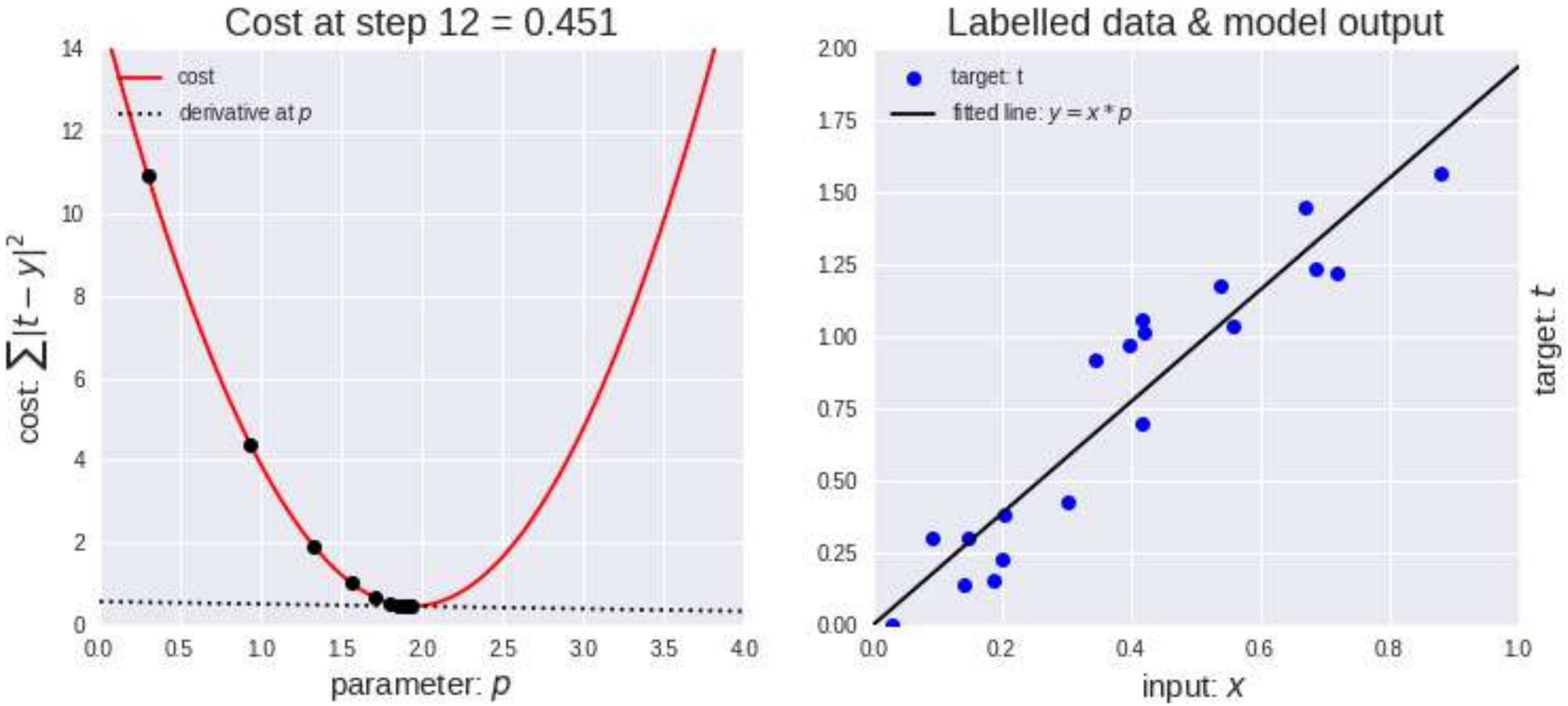


Парная регрессия



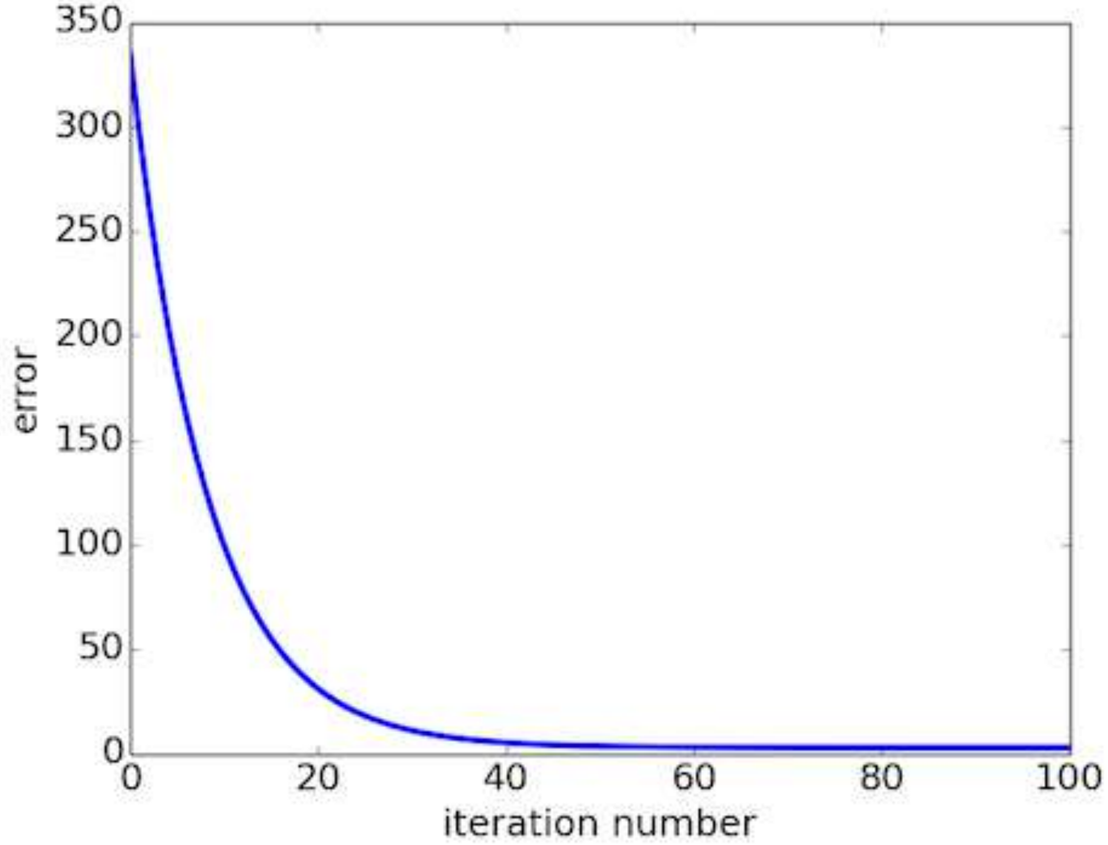
Парная регрессия





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: $y = x * p$
target: t
input: x
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:

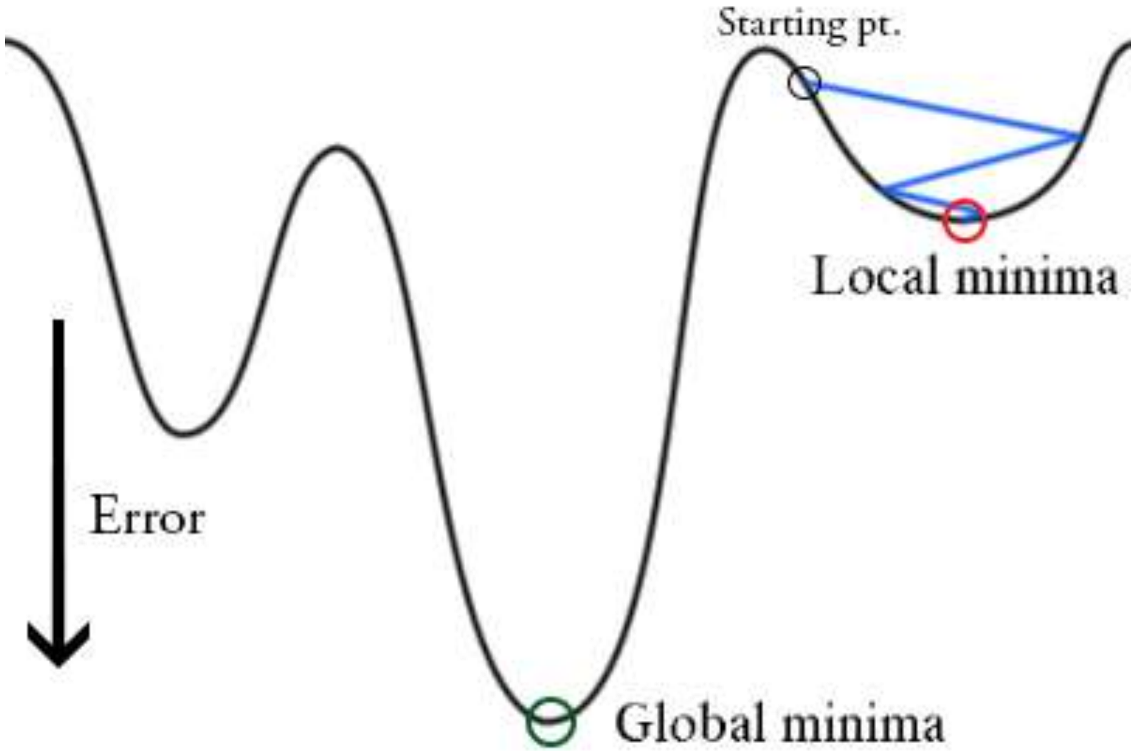
$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

3. Останавливаемся, если

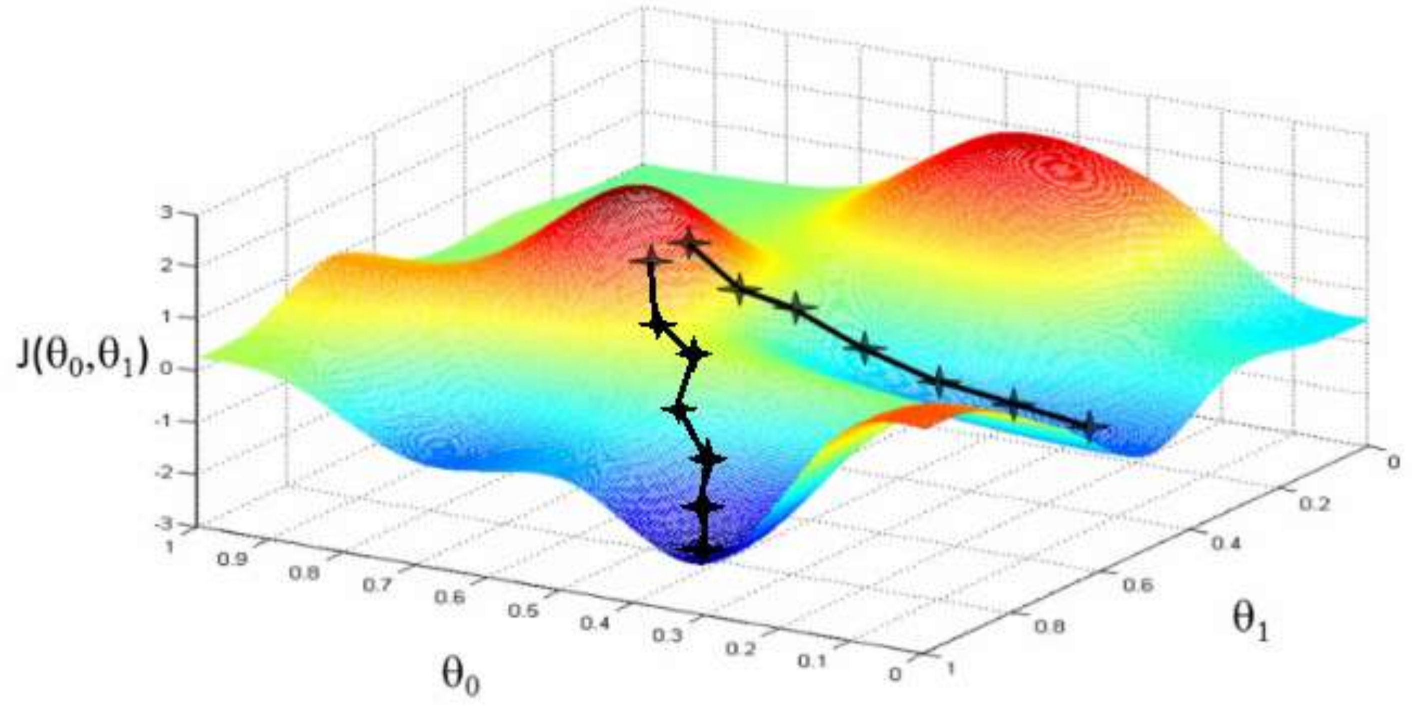
$$\|w^t - w^{t-1}\| < \varepsilon$$

Локальные минимумы

- Градиентный спуск находит только локальные минимумы



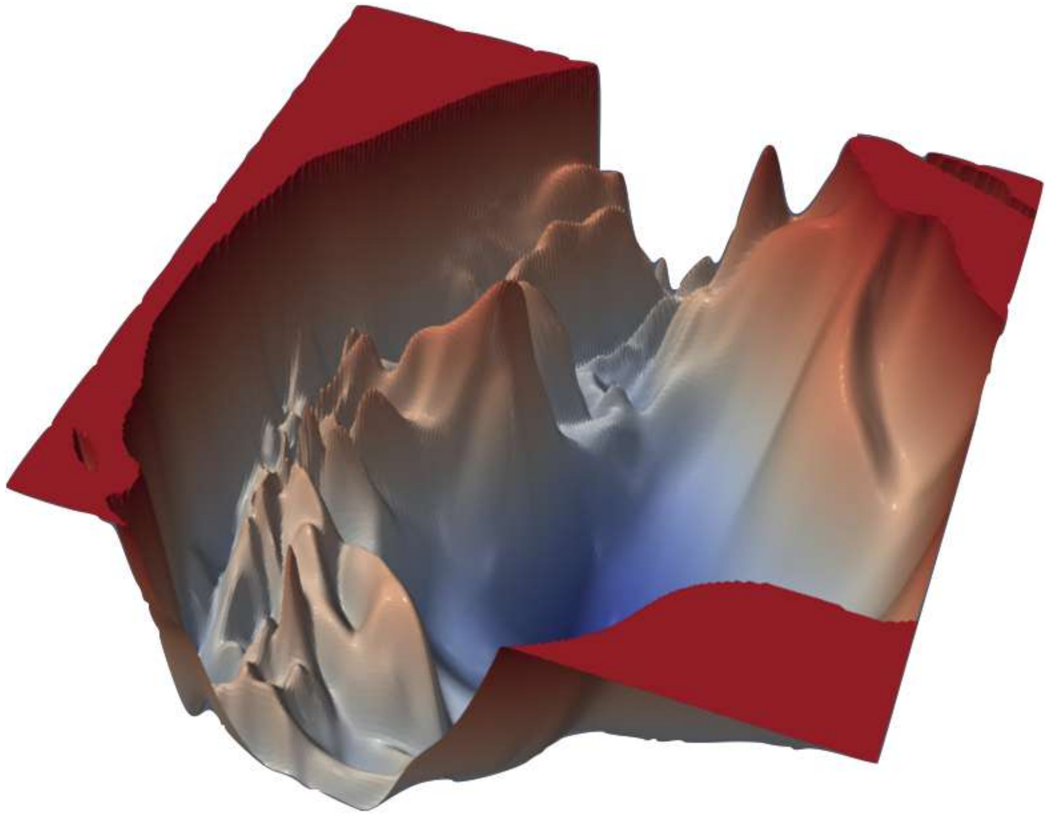
Локальные минимумы



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

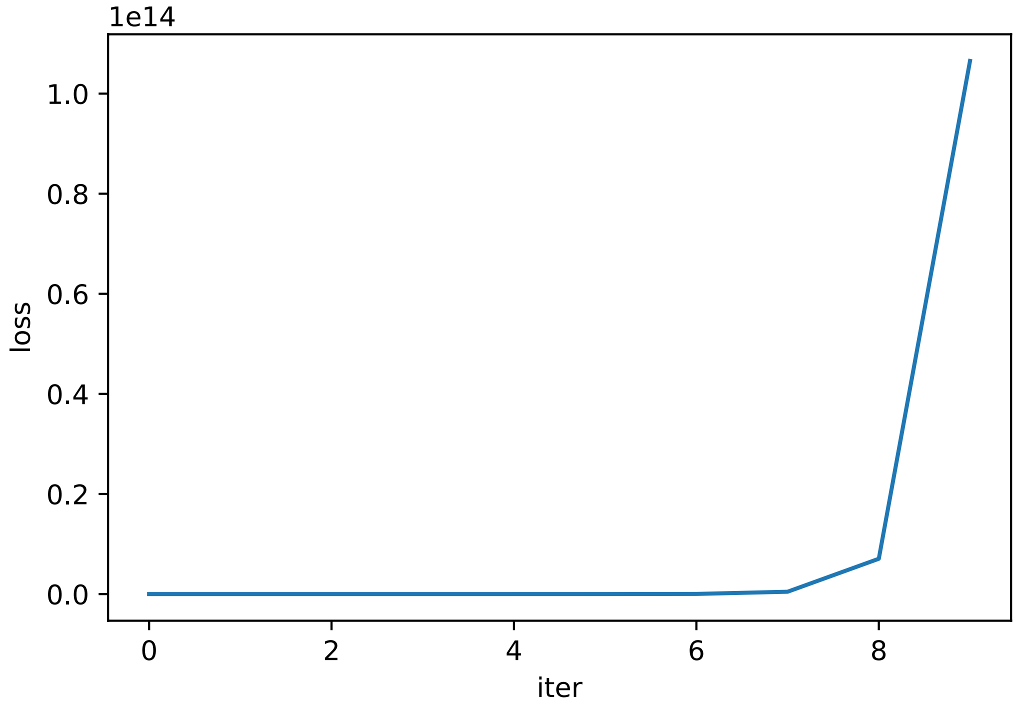
Длина шага

```
[[  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:

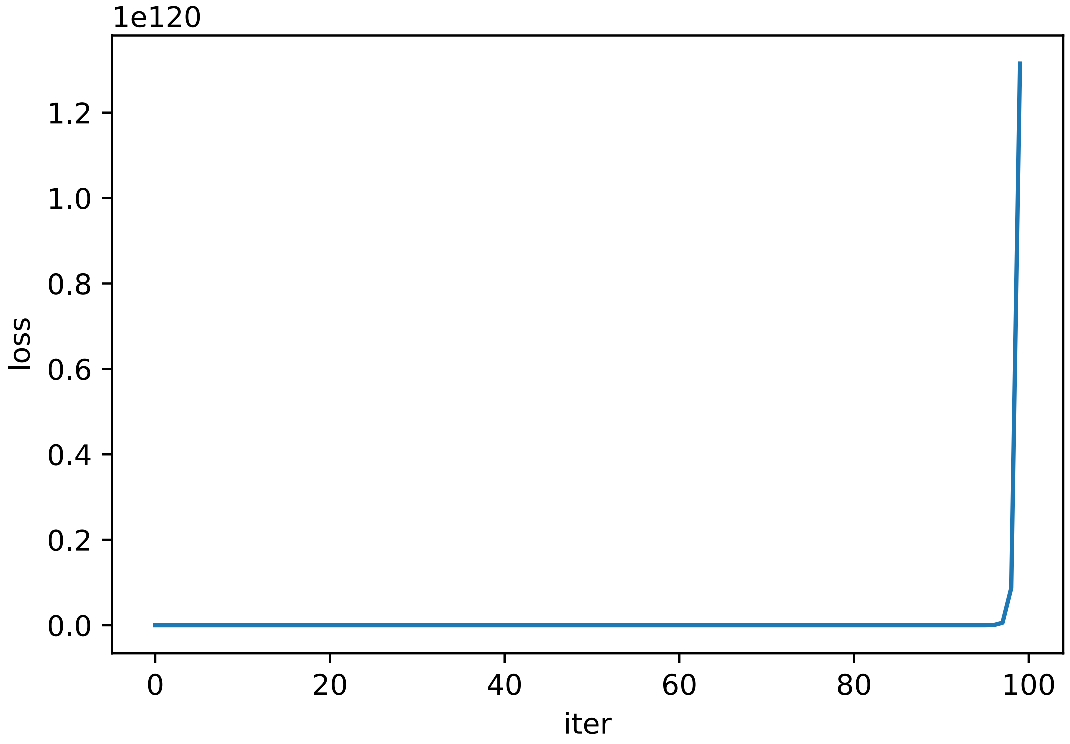
[26.52, 564.80, 682.90, 5097.71, 12110.87]



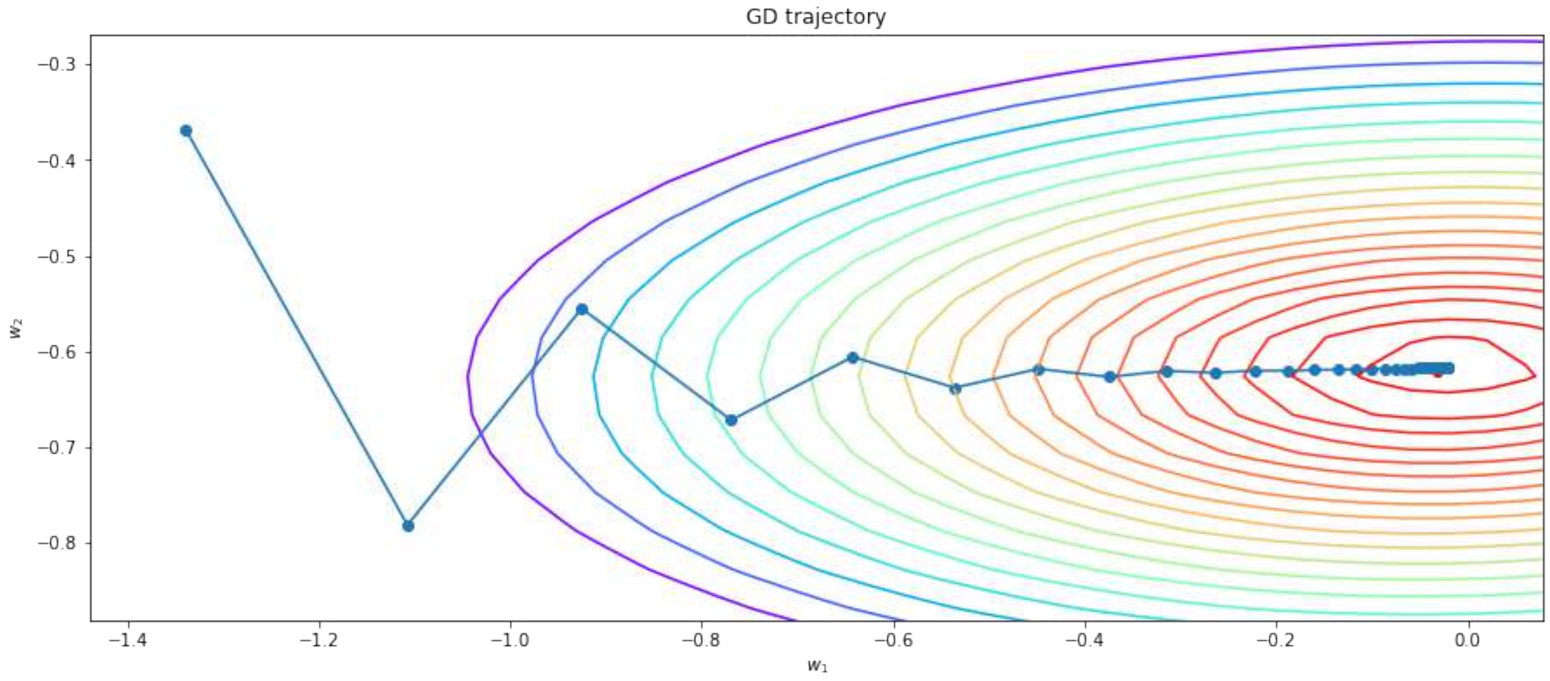
Длина шага

Градиент на первом шаге:

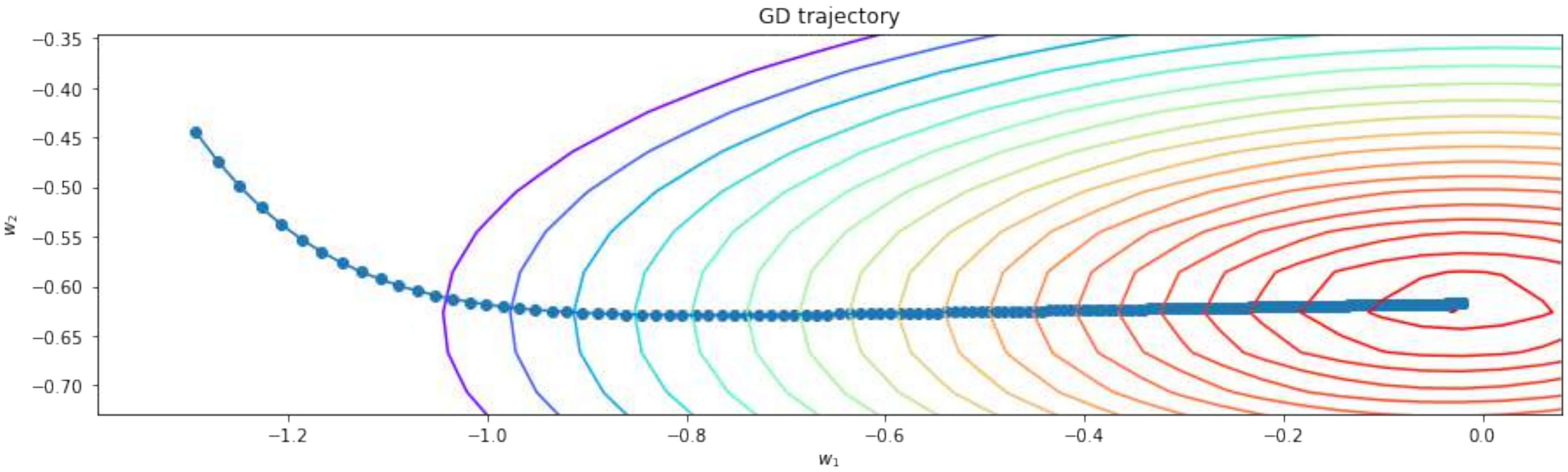
[26.52, 564.80, 682.90, 5097.71, 12110.87]



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

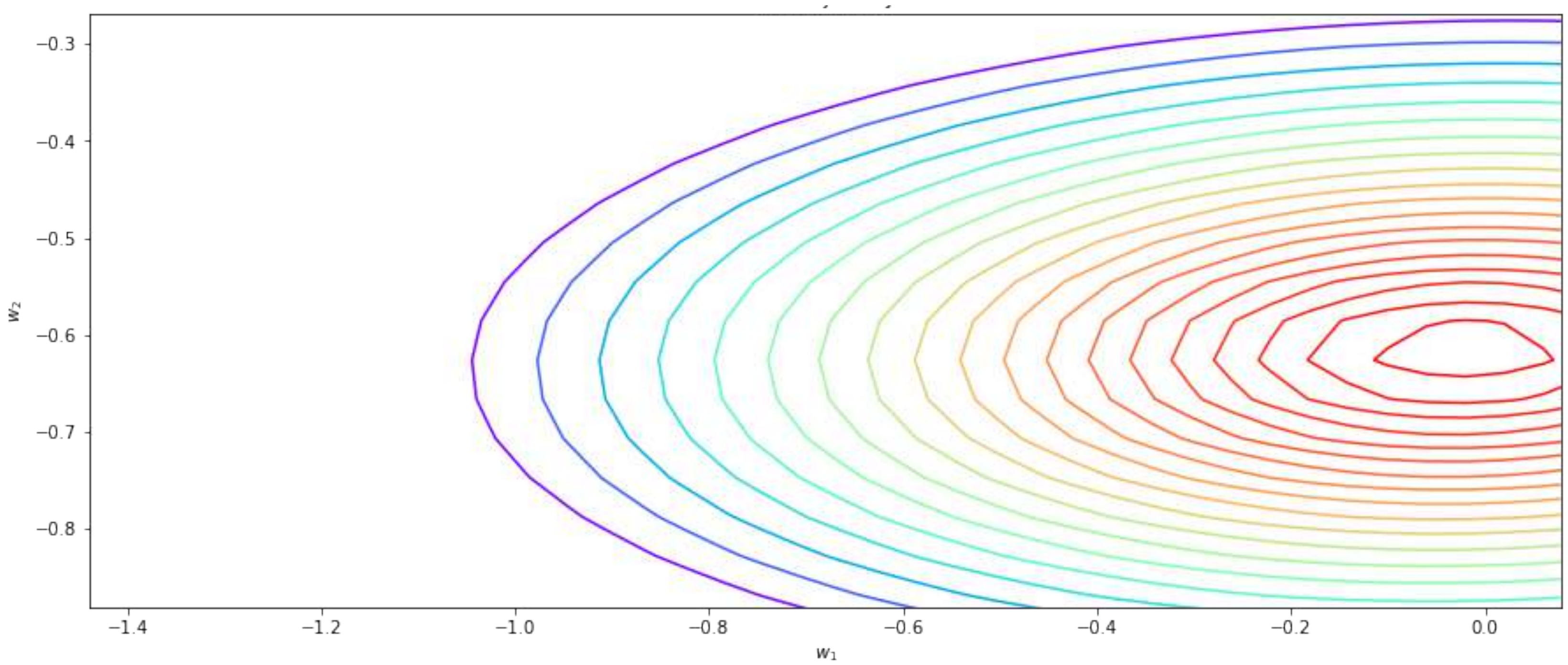
- Например: $\eta_t = \frac{1}{t}$

- Ещё вариант: $\eta_t = \lambda \left(\frac{s}{s+t}\right)^p$

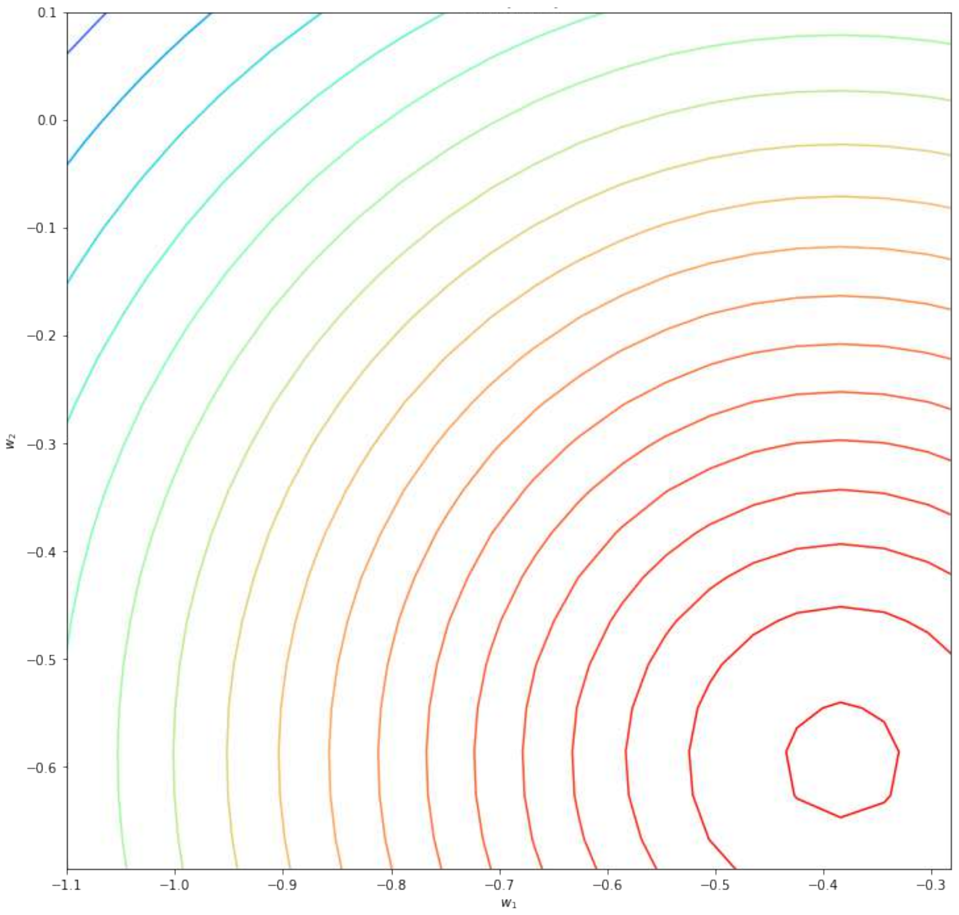
Масштабирование признаков

```
[[  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:

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

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

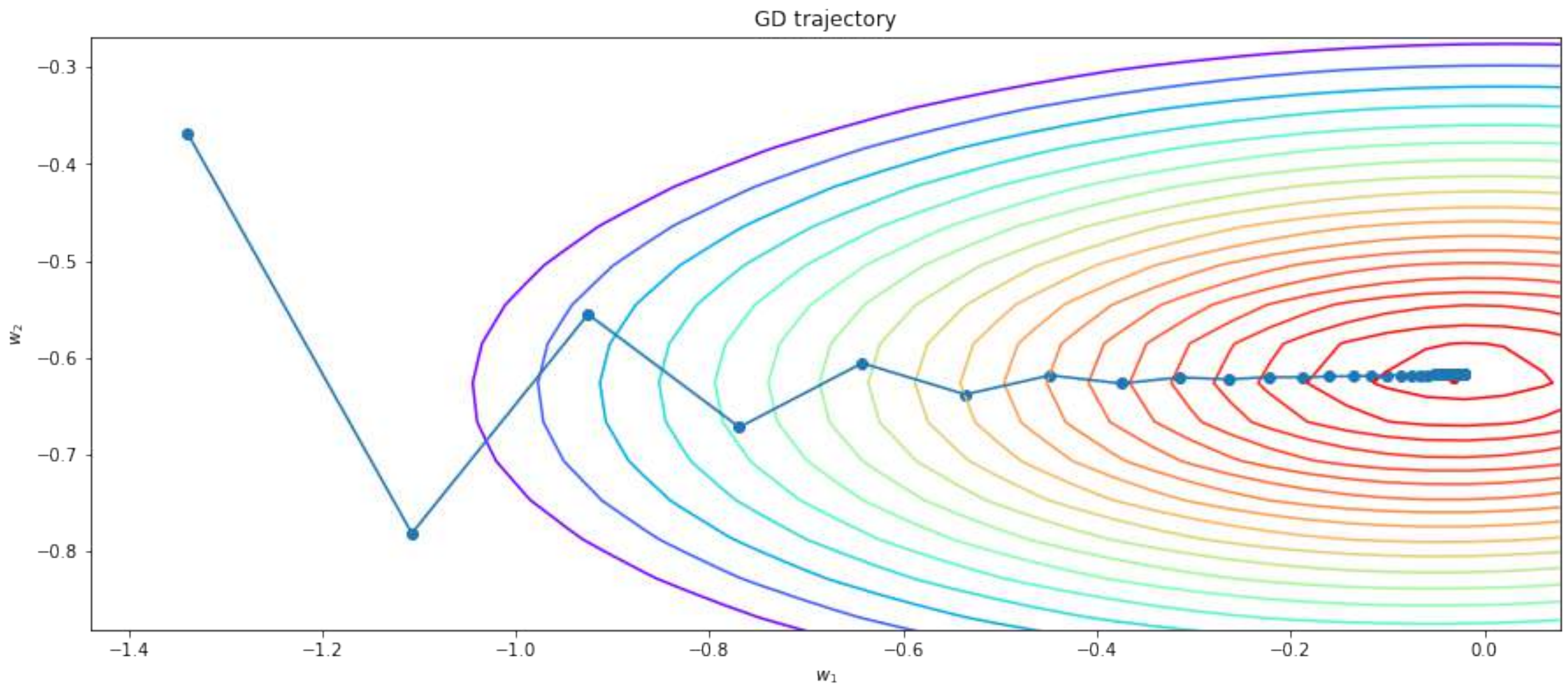
2. Повторять, каждый раз выбирая случайный объект i_t:

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

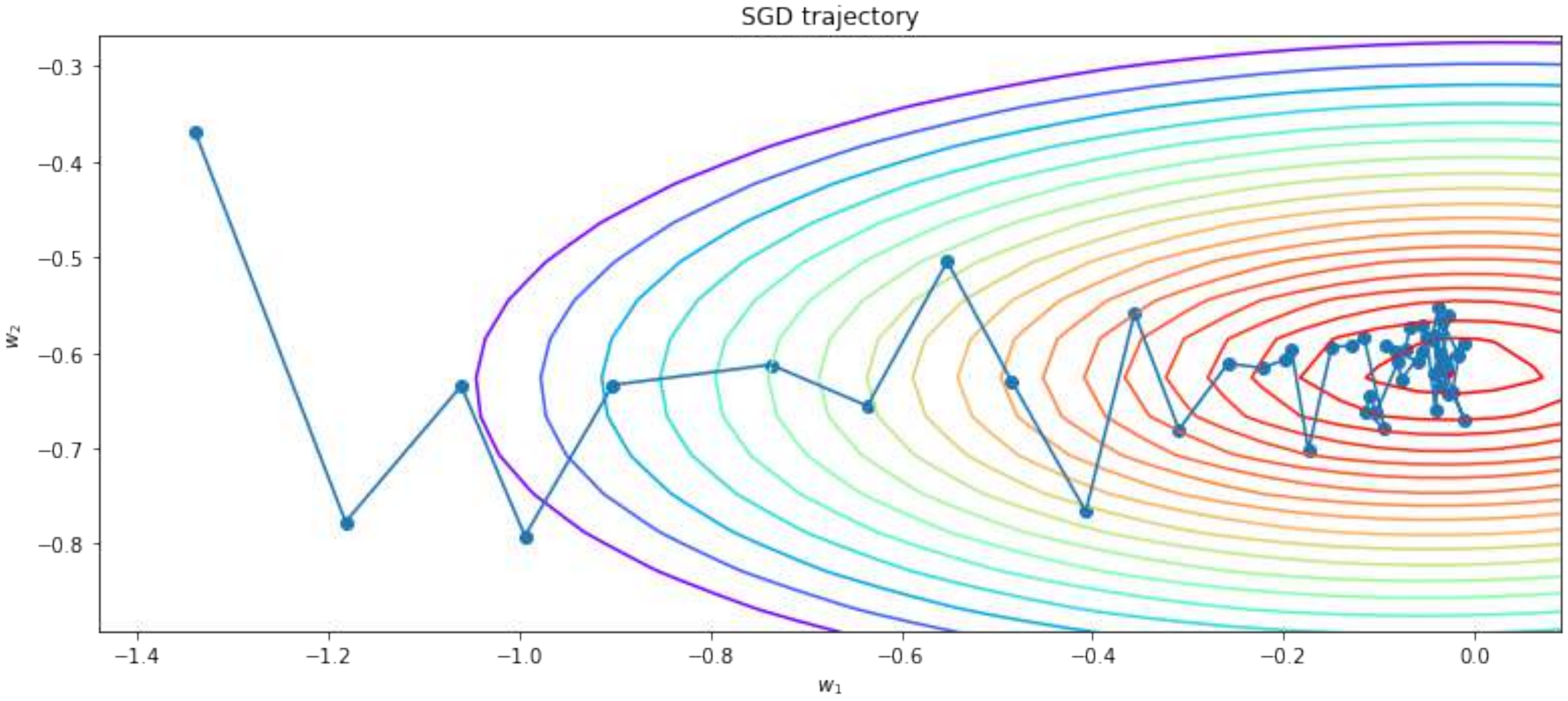
3. Останавливаемся, если

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

Градиентный спуск



Стохастический градиентный спуск



Стохастический градиентный спуск

1. Начальное приближение: w^0

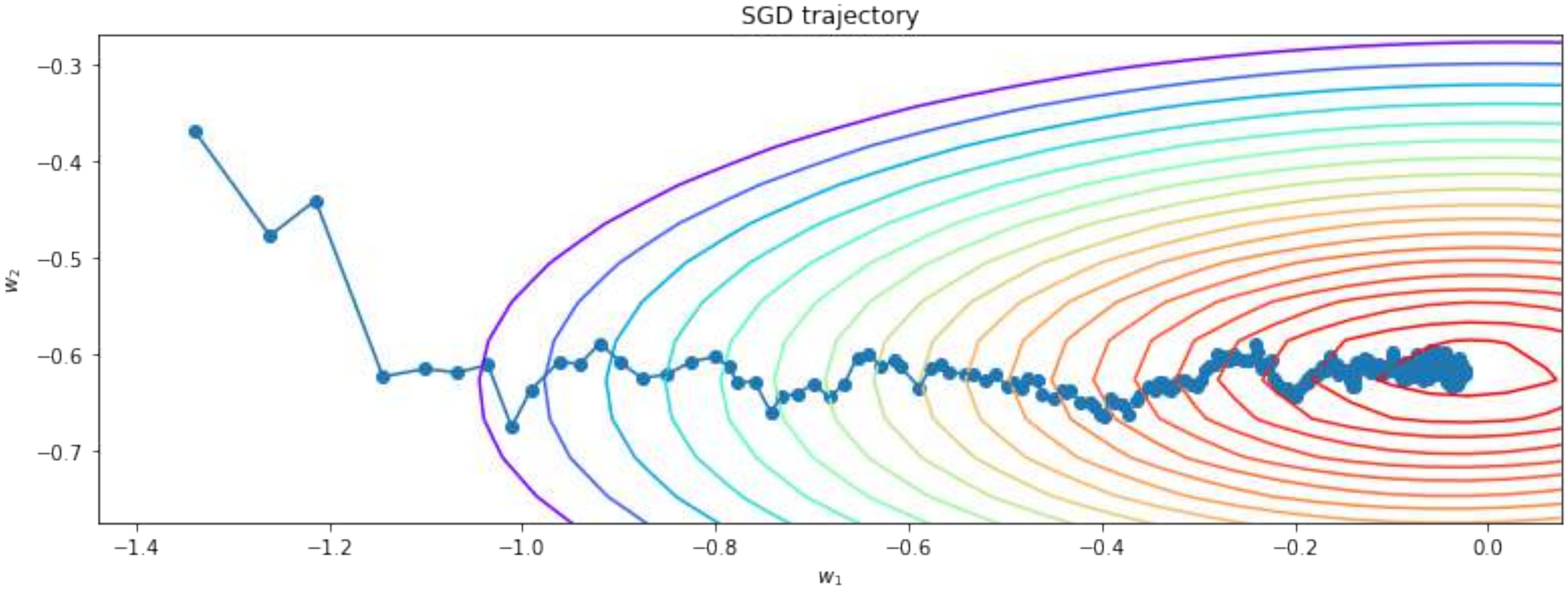
2. Повторять, каждый раз выбирая случайный объект i_t:

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

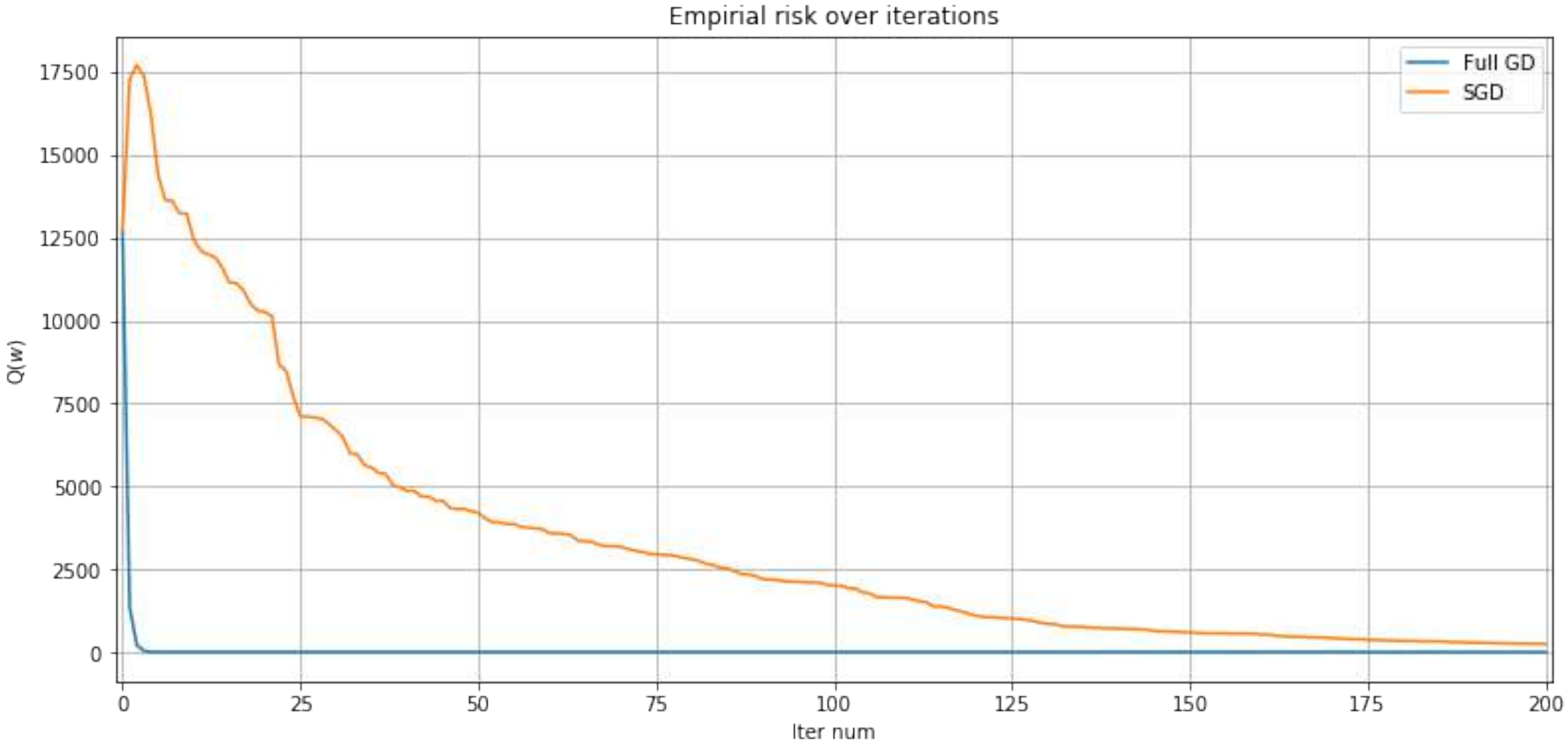
3. Останавливаемся, если ошибка на валидационной выборке перестала падать

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

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