Линейная регрессия

Лекция 2

План лекции

- Понятие линейных моделей
- Измерение ошибки в задачах регрессии
- Обучение линейной регрессии
- Градиентный спуск
- Стохастический градиентный спуск
- Модификации градиентного спуска
- Предобработка данных
- Переобучение
- Оценка качества моделей
- Регуляризация

Понятие линейных моделей

- Линейная регрессионная модель:

$$a(\vec{x}_i) = w_0 + \sum_{j=1}^{d} w_j x_{ij}\,,$$

где w_j – веса или весовые коэффициенты,
w_0 – свободный коэффициент или сдвиг (bias).

- В векторном виде:

$$a(\vec{x}_i) = w_0 + \langle \vec{w}, \vec{x}_i \rangle,$$

где $\vec{w} = (w_1, \dots, w_d)$, $\vec{x}_i = (x_{i1}, \dots, x_{id})$.

- В сокращенном векторном виде:

$$a(\vec{x}_i) = \langle \vec{w}, \vec{x}_i \rangle$$

Измерение ошибки в задачах регрессии

- Функция потерь:

$$L(y, y_{pred}) = L(y, a)$$

- Среднеквадратичная ошибка (mean squared error, MSE):

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

$$MSE(a, X) = \frac{1}{l}\sum_{i=1}^{l} L(y_i, a_i) = \frac{1}{l}\sum_{i=1}^{l} (a(\vec{x}_i) - y_i)^2$$

- Root mean squared error (RMSE):

$$RMSE(a, X) = \sqrt{\frac{1}{l}\sum_{i=1}^{l} (a(\vec{x}_i) - y_i)^2}$$

Измерение ошибки в задачах регрессии

- Коэффициент детерминации:

$$R^2(a, X) = 1 - \frac{\sum_{i=1}^{l} (a(\vec{x}_i) - y_i)^2}{\sum_{i=1}^{l} (y_i - \bar{y})^2} = 1 - \frac{\sigma^2}{\sigma_y^2}$$

где σ_y^2 – дисперсия y, σ^2 – дисперсия ошибки модели

- Среднее абсолютное отклонение (mean absolute error, MAE):

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

$$MAE(a, X) = \frac{1}{l} \sum_{i=1}^{l} |a(\vec{x}_i) - y_i|$$

Измерение ошибки в задачах регрессии

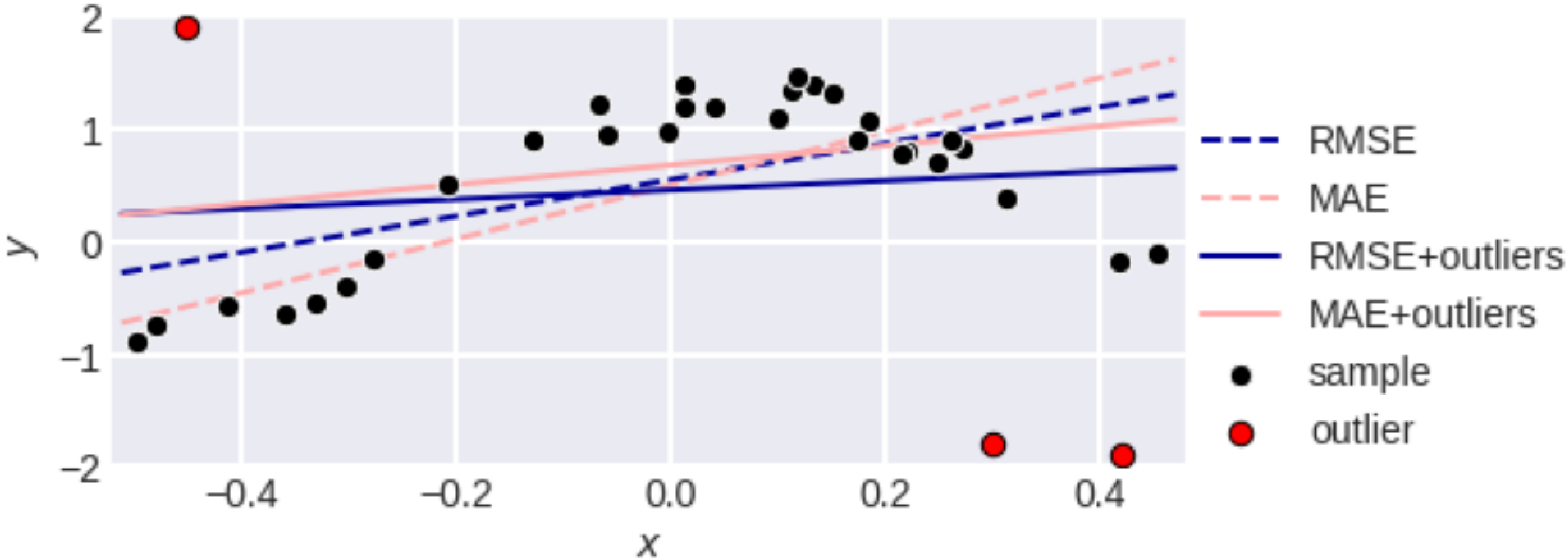
- Среднеквадратичная логарифмическая ошибка (mean squared logarithmic error, MSLE):

$$L(y, a) = \big(log(a + 1) - log(y + 1)\big)^2$$

- Средняя абсолютная процентная ошибка (mean absolute percentage error, MAPE):

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

Измерение ошибки в задачах регрессии



Обучение линейной регрессии

- В случае использования среднеквадратичной ошибки (MSE):

$$\frac{1}{l}\sum_{i=1}^{l}(\langle\vec{w},\vec{x}_i\rangle - y_i)^2 \to \min_{\vec{w}}$$

- В матричном виде:

$$\frac{1}{l}\|X\vec{w} - \vec{y}\|^2 \to \min_{\vec{w}},$$

где $X \in \mathbb{R}^{l\times d}, \vec{w} \in \mathbb{R}^d, \vec{y} \in \mathbb{R}^l$

Обучение линейной регрессии

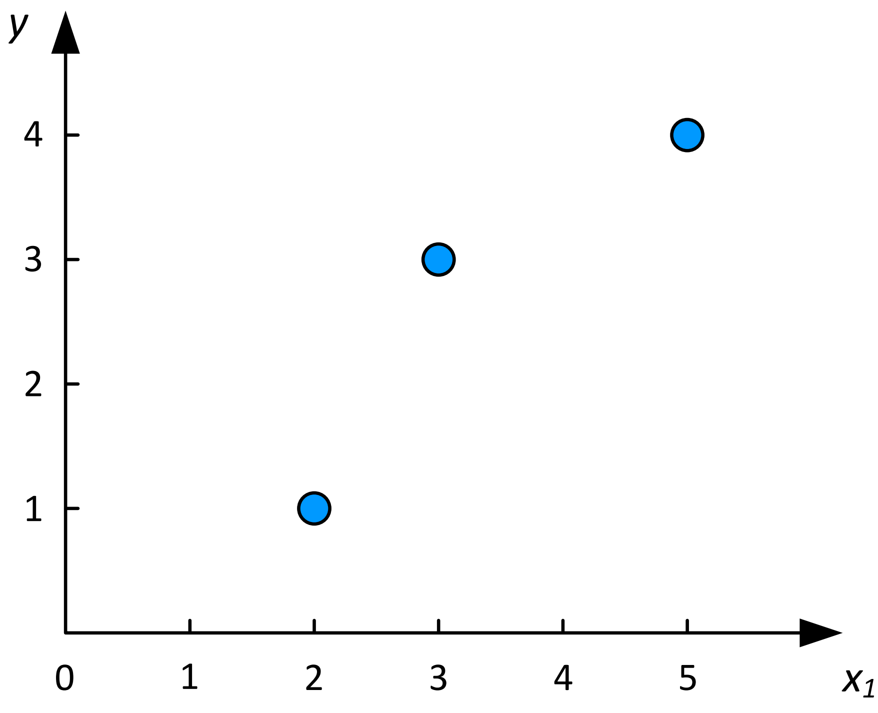
- После дифференцирования данного функционала по вектору $\vec{w}$, приравнивания к нулю и решения уравнения, получаем:

$$\frac{\partial}{\partial \vec{w}} \left(\frac{1}{l} \| X\vec{w} - \vec{y} \|^2 \right) = 0 \ \rightarrow \vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

– *нормальное уравнение* (normal equation)

Обучение линейной регрессии

- Пример: пусть даны три точки $(2, 1)$, $(3, 3)$, $(5, 4)$
- Требуется построить линейную регрессионную модель на основе нормального уравнения



Обучение линейной регрессии

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \end{bmatrix}, \qquad \vec{y} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$$

$$x_0 \quad x_1$$

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

Обучение линейной регрессии

$$X = \begin{array}{c} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \end{bmatrix} \\ \begin{array}{cc} x_0 & x_1 \end{array} \end{array}, \qquad \vec{y} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$$

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \end{bmatrix} =$$

Обучение линейной регрессии

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \end{bmatrix}, \qquad \vec{y} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$$

$$x_0 \quad x_1$$

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 3 & 10 \\ 10 & 38 \end{bmatrix}$$

$$2 \times 3 \qquad 3 \times 2 \qquad 2 \times 2$$

Обучение линейной регрессии

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

Обучение линейной регрессии

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

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

Обучение линейной регрессии

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

$$(X^T X)^{-1} = \begin{bmatrix} 2.7 & -0.7 \\ -0.7 & 0.2 \end{bmatrix}$$

Обучение линейной регрессии

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

$$(X^T X)^{-1} = \begin{bmatrix} 2.7 & -0.7 \\ -0.7 & 0.2 \end{bmatrix}$$

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

Обучение линейной регрессии

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

$$(X^T X)^{-1} = \begin{bmatrix} 2.7 & -0.7 \\ -0.7 & 0.2 \end{bmatrix}$$

$$(X^T X)^{-1} X^T = \begin{bmatrix} 1.29 & 0.57 & -0.85 \\ -0.28 & -0.07 & 0.36 \end{bmatrix}$$

Обучение линейной регрессии

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

$$(X^T X)^{-1} = \begin{bmatrix} 2.7 & -0.7 \\ -0.7 & 0.2 \end{bmatrix}$$

$$(X^T X)^{-1} X^T = \begin{bmatrix} 1.29 & 0.57 & -0.85 \\ -0.28 & -0.07 & 0.36 \end{bmatrix}$$

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y} =$$

Обучение линейной регрессии

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

$$(X^T X)^{-1} = \begin{bmatrix} 2.7 & -0.7 \\ -0.7 & 0.2 \end{bmatrix}$$

$$(X^T X)^{-1} X^T = \begin{bmatrix} 1.29 & 0.57 & -0.85 \\ -0.28 & -0.07 & 0.36 \end{bmatrix}$$

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y} = \begin{bmatrix} -0.43 \\ 0.93 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

Обучение линейной регрессии

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y}$$

$$(X^T X)^{-1} = \begin{bmatrix} 2.7 & -0.7 \\ -0.7 & 0.2 \end{bmatrix}$$

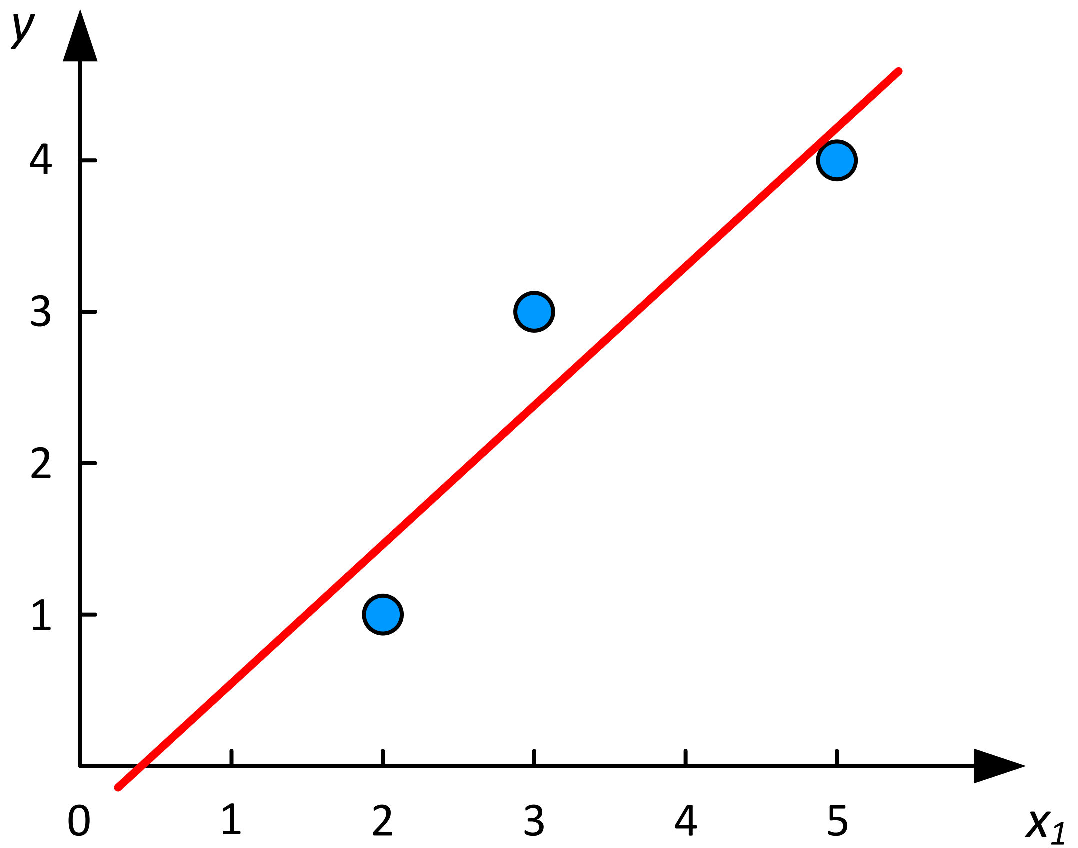
$$(X^T X)^{-1} X^T = \begin{bmatrix} 1.29 & 0.57 & -0.85 \\ -0.28 & -0.07 & 0.36 \end{bmatrix}$$

$$\vec{w}_{opt} = (X^T X)^{-1} X^T \vec{y} = \begin{bmatrix} -0.43 \\ 0.93 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$a(\vec{x}) = w_0 + w_1 x_1 = -0.43 + 0.93 x_1$$

Обучение линейной регрессии

$$a(\vec{x}) = -0.43 + 0.93x_1, \qquad a(1) = 0.5, \qquad a(5) = 4.2$$



Градиентный спуск

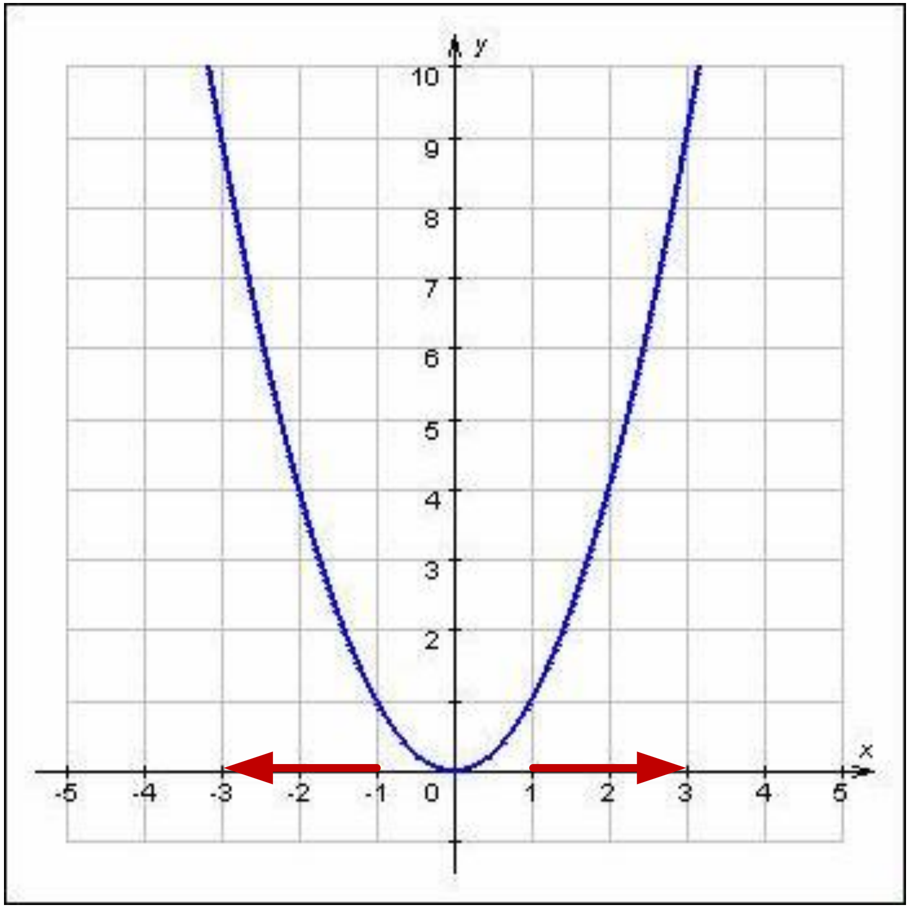
- *Градиентом* функции $f: \mathbb{R}^d \to \mathbb{R}^d$ называется вектор её частных производных (∇ – оператор набла, оператор Гамильтона):

$$\nabla f(x_1, \dots, x_d) = \left(\frac{\partial f}{\partial x_j}\right)_{j=1}^{d} = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_d}\right)$$

- Градиент является направлением наискорейшего роста функции, а *антиградиент* ($-\nabla f$) – направлением наискорейшего убывания

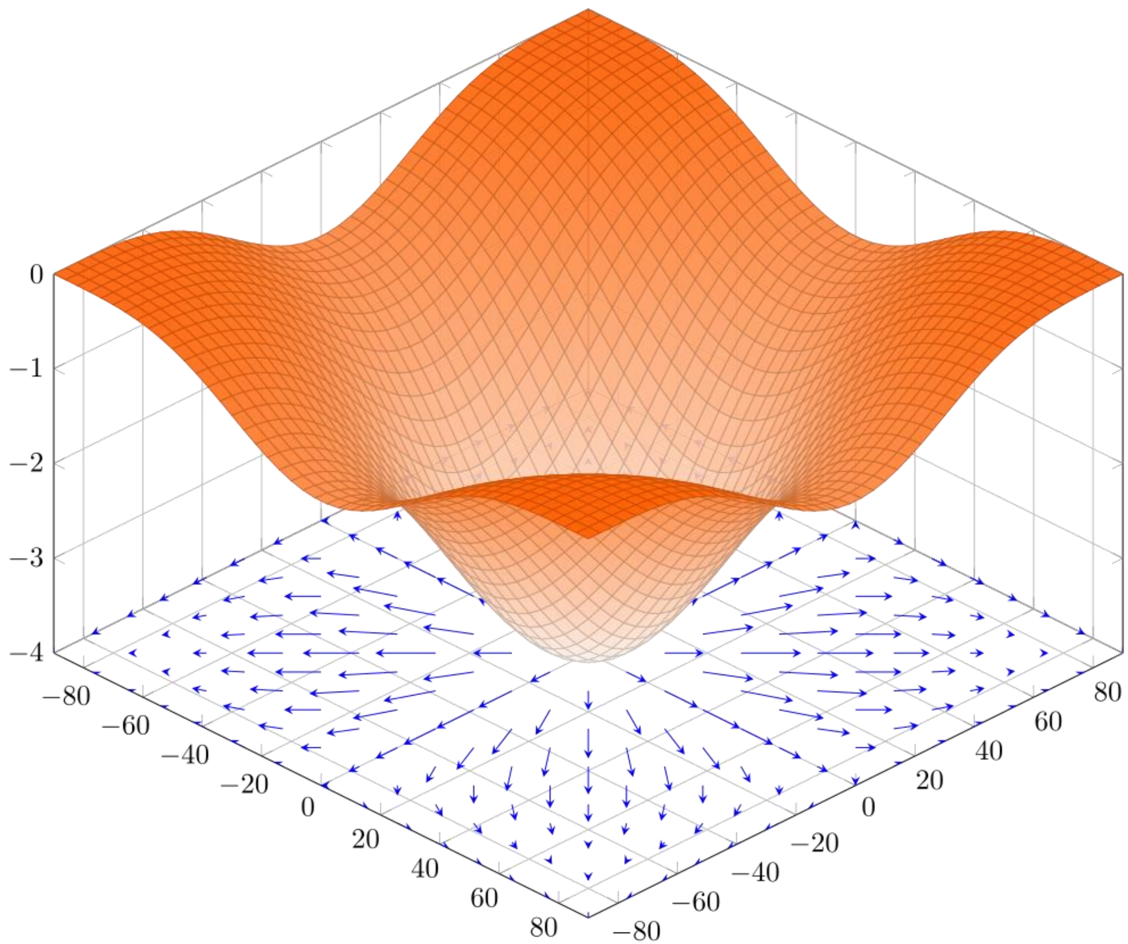
Градиентный спуск

- *Пример*: функция $y = x^2$, производная $\frac{dy}{dx} = 2x$



Градиентный спуск

- *Пример*: функция $y = -(\cos^2 x_1 + \cos^2 x_2)^2$



Градиентный спуск

Алгоритм градиентного спуска:

1. Выбрать начальную точку $\vec{w}^{(0)}$
2. Повторять до сходимости:

$$\vec{w}^{(k)} = \vec{w}^{(k-1)} - \eta_k \nabla Q\left(\vec{w}^{(k-1)}\right),$$

где k – номер шага,

$Q(\vec{w})$ – функция ошибки для набора параметров $\vec{w}$,

η_k – скорость спуска (длина k-го шага).

- Условия останова:
  - ошибка не уменьшается в течение нескольких итераций
  - вектор весов *почти* перестает изменяться
  - достигнуто максимальное число итераций

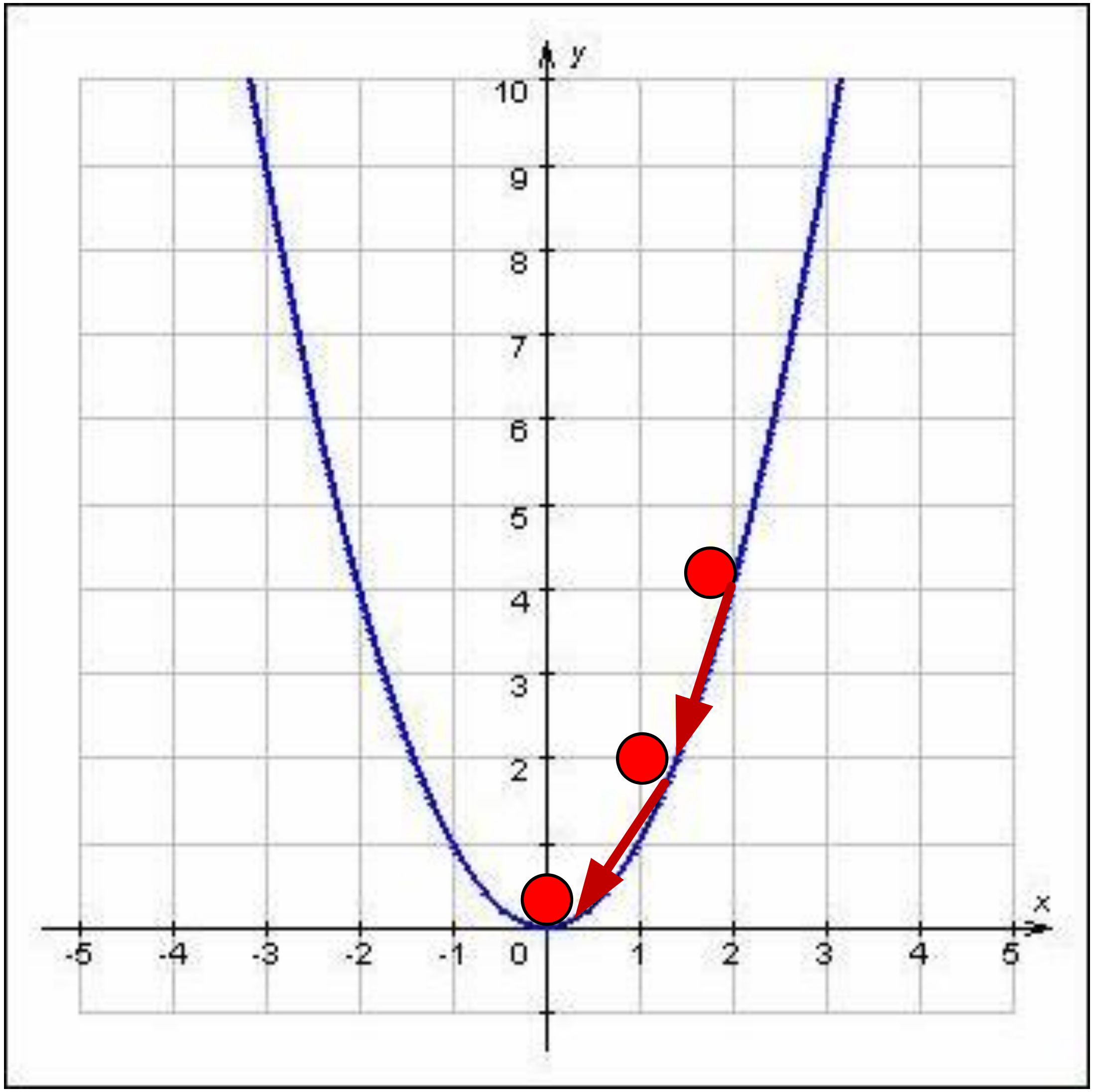
Градиентный спуск

- Скорость спуска:
  - слишком высокая $\rightarrow$ переход через минимум
  - слишком низкая $\rightarrow$ медленная сходимость

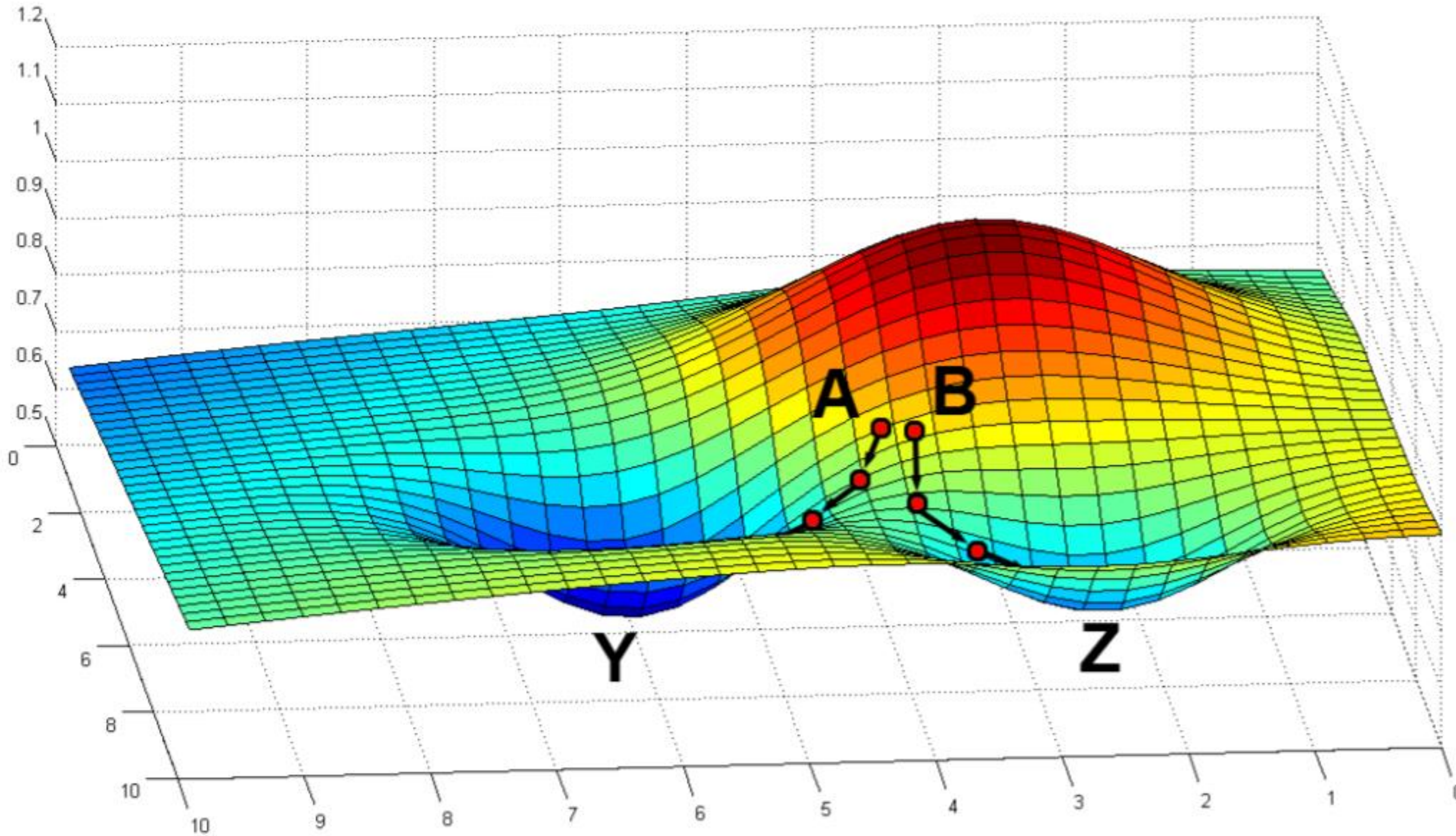
- Варианты вычисления скорости спуска:
  - константная: $\eta_k = const$
  - линейное уменьшение (linear decay): $\eta_k = \eta_0 \left(1 - \frac{k}{K}\right)$
  - экспоненциальное уменьшение (exponential decay):

$$\eta_k = \eta_0 e^{-\frac{k}{K}}$$

Градиентный спуск



Градиентный спуск



Градиентный спуск

- Для среднеквадратичной ошибки:

$$\nabla Q(\vec{w}) = \nabla_{\vec{w}}\left(\frac{1}{l}\|X\vec{w} - \vec{y}\|^2\right) = \frac{2}{l}X^T(X\vec{w} - \vec{y})$$

Стохастический градиентный спуск

- Функционал ошибки представим в виде суммы l функций ошибок:

$$Q(\vec{w}) = \frac{1}{l}\sum_{i=1}^{l} L_i(\vec{w})$$

- При градиентном спуске необходимо вычислять градиент всей суммы:

$$\nabla Q(\vec{w}) = \frac{1}{l}\sum_{i=1}^{l} \nabla L_i(\vec{w})$$

- Если выборка большая, вычисление градиента трудоемко

Стохастический градиентный спуск

- Оценить градиент суммы функций можно градиентом одного случайно взятого i_k слагаемого:

$$\nabla Q(\vec{w}) \approx \nabla L_{i_k}(\vec{w})$$

- Метод стохастического градиентного спуска (stochastic gradient descent, SGD):

$$\vec{w}^{(k)} = \vec{w}^{(k-1)} - \eta_k \nabla L_{i_k}\left(\vec{w}^{(k-1)}\right)$$

- Градиентный спуск по мини-батчам (mini-batch gradient descent):

$$\nabla Q(\vec{w}) \approx \frac{1}{n} \sum_{j=1}^{n} \nabla L_{i_{kj}}(\vec{w}),$$

где i_{kj} – случайно выбранные номера слагаемых