

Московский физико-технический институт  
Физтех-школа прикладной математики и информатики

В В Е Д Е Н И Е В А Н А Л И З Д А Н НЫХ  
IV СЕМЕСТР

Лектор:

h\nu

Автор: *Киселев Николай*  
*Репозиторий на Github*

весна 2025

## Содержание

<b>1 Задача линейной регрессии</b>	<b>2</b>
1.1 Прямой подход . . . . .	2
1.2 Градиентный спуск (GD) . . . . .	2
1.3 Стохастический градиентный спуск (SGD) . . . . .	3

# 1 Задача линейной регрессии

## 1.1 Прямой подход

**Определение 1.1.**  $\mathcal{M} = \{y : \mathbb{R}^d \rightarrow \mathbb{R} | y(x) = x^T \theta, \theta \in \mathbb{R}^d\}$  — множество рассматриваемых нами моделей.

Наша цель — получить наилучшую модель, то есть оценить  $\theta$ .

Пусть  $\hat{\theta}$  — оценка  $\theta$ . Тогда  $\hat{y}(x) = x^T \hat{\theta}$  — предсказание для  $x$ .

Пусть  $x_1, \dots, x_n$  — объекты,  $Y_1, \dots, Y_n$  — таргеты. Пусть  $\hat{Y}_i = \hat{y}(x_i)$

Введем функционал ошибки:

$$\mathcal{L}(y, z) = (y - z)^2$$

И тогда получаем, что

$$F(\theta) = \sum_{i=1}^n \mathcal{L}(Y_i, \hat{Y}_i) = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - x_i^T \hat{\theta})^2 = \|Y - X \hat{\theta}\|^2$$

Где

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}$$

Мы хотим минимизировать  $F(\theta)$

**Утверждение 1.1.** Если  $XX^T$  не вырождена, то  $\hat{\theta} = (X^T X)^{-1} X^T Y$

*Доказательство.*

$$F(\theta) = \|Y - X\theta\|^2 = (Y - X\theta)^T (Y - X\theta) = Y^T Y - 2Y^T X\theta + \theta^T X^T X\theta$$

Посчитаем градиент данной функции по  $\theta$ :

$$\nabla F(\theta) = -2X^T Y + 2X^T X\theta = 0$$

Домножим на  $(X^T X)^{-1}$  слева:

$$\hat{\theta} = \underbrace{(X^T X)^{-1} X^T}_{\text{псевдообратная матрица}} Y$$

Т.к. функция  $\|Y - X\theta\|^2$  квадратична, точка с нулевым градиентом является точкой минимума.  $\square$

## 1.2 Градиентный спуск (GD)

Пусть у нас есть задача  $f(x) \rightarrow \min_x$ .

**Замечание.**  $\nabla f$  — направление наибольшего роста  $f(x)$  в точке  $x$ .

**Идея:** будем идти в противоположную сторону. Пусть  $x_0$  — начальное приближение. Будем действовать по следующему алгоритму: будем постепенно делать шаги, каждый новый шаг определяется формулой  $x_{t+1} = x_t - \eta \nabla f$ , где  $\eta$  — шаг метода.

**Пример.** Рассмотрим  $f(x) = x^2 \Rightarrow \nabla f = 2x$ . GD даст нам  $x_t = x_t - 2\eta x_t$ . При  $\eta = 1$  мы получим  $x_{t+1} = -x_t$ .

**Пример.** Рассмотрим  $f(x) = x^4 \Rightarrow \nabla f = 4x^3$ . GD даст нам  $x_t = x_t - 2\eta x_t$ . При  $\eta = 1$  мы получим  $x_{t+1} = -4x_t$ , т.е. наша последовательность не сходится.

Применим GD к задаче линейной регрессии:  $F(\theta) = \|Y - X\theta\|^2 \rightarrow \min_\theta$

$$\nabla F(\theta) = -2X^T Y + 2X^T X\theta = 2X^T(X\theta - Y)$$

Шаг GD (занесем константу 2 в  $\eta$ ):

$$\theta_{t+1} = \theta_t - \eta X^T(X\theta_t - Y) = \theta_t - \eta \sum_{i=1}^n x_i(x_i^T \theta_t - Y_i)$$

Что мы можем сказать про градиентный спуск?

- + Не надо обращать матрицу.
- Если  $n$  велико, то каждый шаг выполняется долго

Возникает еще одна идея: а что если считать градиент не для каждого из  $n$  элементов, а для некоторого количества из них. Таким образом, мы приходим к идее стохастического градиентного спуска.

### 1.3 Стохастический градиентный спуск (SGD)

Возьмем индексы  $I = \underbrace{\{i_1, \dots, i_k\}}_{\text{батч}} \sim U\{1, 2, \dots, n\}$  (отвечающие равномерному распределению), где  $k$  — размер батча. Тогда шаг стохастического градиентного спуска будет определяться по формуле:

$$\theta_{t+1} = \theta_t - \eta \frac{n}{k} \sum_{i \in I} x_i(x_i^T \theta_t - Y_i)$$

Здесь множитель  $\frac{n}{k}$  добавлен для нормировки: т.к. мы взяли  $k$  объектов из  $n$ , то полученный градиент будет приблизительно в  $\frac{n}{k}$  меньше исходного. Рассмотрим данные операции в матричном виде:

Пусть  $X_I$  — матрица из строк матрицы  $X$  с индексами  $i_1, \dots, i_k$ ,  $Y_I$  — вектор из элементов вектора  $Y$  с индексами  $i_1, \dots, i_k$ . Таким образом, шаг SGD будет иметь следующий вид:

$$\theta_{t+1} = \theta_t - \eta \frac{n}{k} X_I^T (X_I \theta_t - Y_I)$$

Итого, процедура имеет следующий вид:

1. Сгенерировать набор  $I$
2. Вычислить  $\theta_{t+1} = \dots$