Лекция 2
Методы оптимизации

Градиентный спуск

Структура задач машинного обучения

$D = \{x, y\}_{i=1}^{N}$ — обучающая выборка.

$y^* = A(x; \mu)$ — метод предсказания

μ — параметры

$L(D, \mu) = E(D, \mu) + R(\mu)$ — функция потерь

$E(D, \mu)$ — функция ошибки

$R(\mu)$ — функция регуляризации

$L(D, \mu) \to \min\limits_{\mu}$ — процедура обучения

Ключевая задача в машинном обучении: **оптимизация**.

Напоминание: градиентный спуск

$$f(x) \to \min_x$$

η — величина шага (гиперпараметр)

1. Инициализация

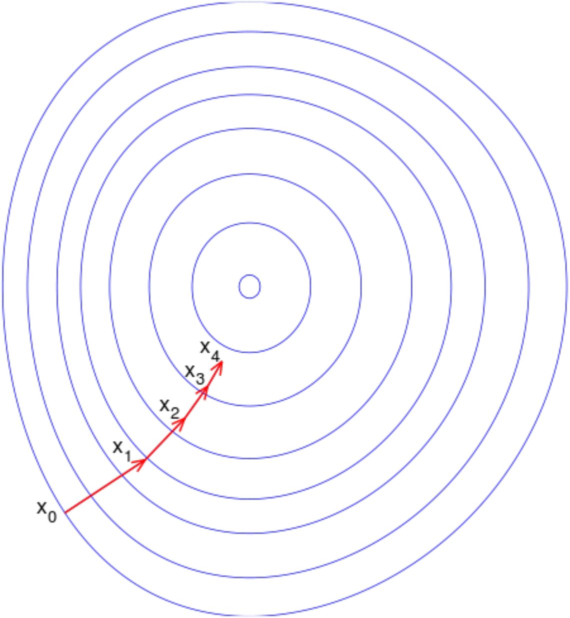
$$k = 0, \quad x_k = \text{начальное приближение}$$

2. Шаг в сторону сильнейшего убывания

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

3. Повторение до сходимости

$$k := k + 1, \quad \text{перейти к } 2$$


x0
x1
x2
x3
x4

Особенности градиентного спуска

Плюсы

- Достаточно универсальный метод
- Легко реализуем

Минусы

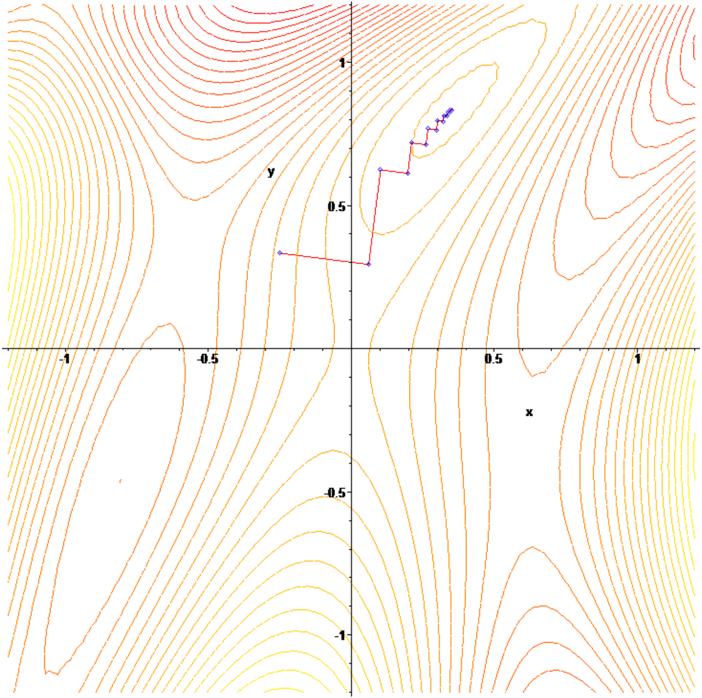
Особенности градиентного спуска

Плюсы

- Достаточно универсальный метод
- Легко реализуем

Минусы

- Попадает в локальные минимумы
- Шаги могут быть медленными
- Неэффективен для больших выборок
- Неприменим для недифференцируемых функций


y
1
0.5
-0.5
-1
-1
-0.5
0.5
1
x

Стохастические методы

Проблема эффективности

Пусть D содержит очень много элементов.

- Долго считать градиент
- Можно сделать мало итераций за разумное время
- Плохое решение

Идея: приближенно оценивать градиент.

Стохастический градиентный спуск (SGD)

Разложим функцию потерь:

$$L(D, \mu) = \frac{1}{N} \sum_{i=1}^{N} l(x_i, \mu).$$

Стохастический градиентный спуск (SGD)

Разложим функцию потерь:

$$L(D, \mu) = \frac{1}{N} \sum_{i=1}^{N} l(x_i, \mu).$$

Разложим градиент:

$$\nabla_\mu L(D, \mu) = \frac{1}{N} \sum_{i=1}^{N} \nabla_\mu l(x_i, \mu).$$

Стохастический градиентный спуск (SGD)

Разложим функцию потерь:

$$L(D, \mu) = \frac{1}{N} \sum_{i=1}^{N} l(x_i, \mu).$$

Разложим градиент:

$$\nabla_\mu L(D, \mu) = \frac{1}{N} \sum_{i=1}^{N} \nabla_\mu l(x_i, \mu).$$

Пусть $I = (i_1, i_2, \ldots, i_m)$ — небольшая подвыборка D

Приблизим градиент:

$$\nabla_\mu L(D, \mu) \approx \frac{1}{m} \sum_{I} \nabla_\mu l(x_i, \mu).$$

Стохастический градиентный спуск (SGD)

$$\sum_{i=1}^{N} f_i(x) \to \min_x$$

η — величина шага, m — размер подвыборки

1. Инициализация
$$k = 0, \quad x_k = \text{начальное приближение}$$
2. Шаг *примерно* в сторону сильнейшего убывания
$$I := \text{случайная подвыборка размера } m$$
$$x_{k+1} = x_k - \eta \sum_I \nabla f_i(x_k)$$
3. Повторение до сходимости
$$k := k + 1, \quad \text{перейти к } 2$$

SGD + Mini-batches

На каждой итерации:

- Перемешать выборку
- Разбить выборку на равные части размера m (мини-батчи)

$$I_1 + I_2 + \ldots + I_{N/m} = \{1, \ldots, N\}$$

- Для $j = 1, \ldots, N/m$

$$x := x - \eta \sum_{I_j} \nabla f_i(x)$$

Возможный вариант: $m = 1$.

Методы 2 порядка

Ряд Тейлора

$f(x)$ — одномерная бесконечно дифференцируемая функция

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots =$$

$$= \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n.$$

(не всегда сходится)

Ряд Тейлора и градиентный спуск

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до первой производной

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

Ряд Тейлора и градиентный спуск

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до первой производной

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

Минимизируем

$$f(x_0) + f'(x_0)(x - x_0) \to \min_x$$

Ряд Тейлора и градиентный спуск

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до первой производной

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

Минимизируем

$$f(x_0) + f'(x_0)(x - x_0) \to \min_x$$

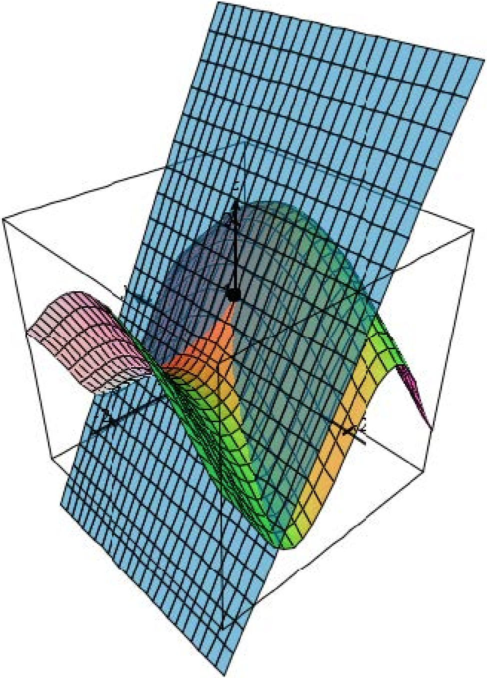
$f'(x_0) > 0$ — минимум в направлении $x \to -\infty$
$f'(x_0) < 0$ — минимум в направлении $x \to +\infty$
Оптимальный шаг: в направлении $-f'(x_0)$

Многомерный градиентный спуск

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до первой производной

$$f(x) \approx f(x_0) + (x - x_0)^T \nabla f(x_0)$$



Многомерный градиентный спуск

Задача:

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до первой производной

$$f(x) \approx f(x_0) + (x - x_0)^T \nabla f(x_0)$$

Когда достигается максимум $\langle x - x_0, \nabla f(x_0) \rangle$?

Многомерный градиентный спуск

Задача:

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до первой производной

$$f(x) \approx f(x_0) + (x - x_0)^T \nabla f(x_0)$$

Когда достигается максимум $\langle x - x_0, \nabla f(x_0) \rangle$?

Ответ: когда $(x - x_0)$ и $\nabla f(x_0)$ параллельны

Многомерный градиентный спуск

Задача:

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до первой производной

$$f(x) \approx f(x_0) + (x - x_0)^T \nabla f(x_0)$$

Когда достигается максимум $\langle x - x_0, \nabla f(x_0) \rangle$?

Ответ: когда $(x - x_0)$ и $\nabla f(x_0)$ параллельны

Наибольшее возрастание: $x - x_0 = \eta \nabla f(x_0)$
Наибольшее убывание: $x - x_0 = -\eta \nabla f(x_0)$

Минимизация 2 порядка

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до второй производной

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2$$

Минимизация 2 порядка

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до второй производной

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

Минимизируем

$$f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 \to \min_x$$

Минимизация 2 порядка

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до второй производной

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

Минимизируем

$$f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 \to \min_x$$

Минимум при

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

Минимизация 2 порядка (многомерная)

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до второй производной

$$f(x) \approx f(x_0) + (x - x_0)^T \nabla f(x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0)$$

Минимизация 2 порядка (многомерная)

$$f(x) \to \min_x$$

Разложим в ряд Тейлора до второй производной

$$f(x) \approx f(x_0) + (x - x_0)^T \nabla f(x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0)$$

Минимизируем

$$f(x_0) + (x - x_0)^T \nabla f(x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0) \to \min_x$$

Минимум при

$$x = x_0 - \left(\nabla^2 f(x_0)\right)^{-1} \nabla f(x_0)$$

Метод Ньютона

$$f(x) \to \min_x$$

1. Инициализация

$$k = 0, \quad x_k = \text{начальное приближение}$$

2. Шаг в точку примерного минимума

$$x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

3. Повторение до сходимости

$$k := k + 1, \quad \text{перейти к } 2$$

Метод Ньютона (модификация)

$$f(x) \to \min_x$$

η — величина шага (гиперпараметр)

1. Инициализация

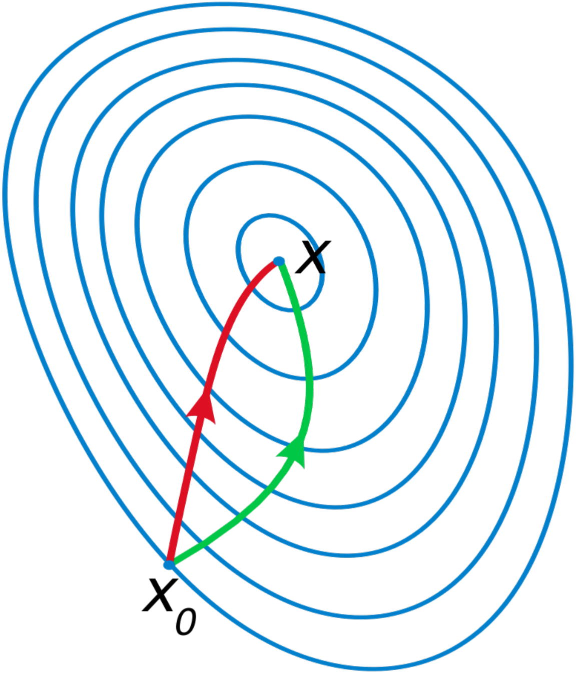
$$k = 0, \quad x_k = \text{начальное приближение}$$

2. Шаг в сторону примерного минимума

$$x_{k+1} = x_k - \eta \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

3. Повторение до сходимости

$$k := k + 1, \quad \text{перейти к } 2$$


x
x0

Почему «метод Ньютона»?

Задача:

$$g(x) = 0$$

Итеративное решение:

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$$

Почему «метод Ньютона»?

Задача:

$$g(x) = 0$$

Итеративное решение:

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$$

Задача:

$$f(x) \to \min_x \quad \Rightarrow \quad f'(x) = 0$$

Итеративное решение:

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

Методы 0 порядка

Покоординатный спуск

(Производную нельзя вычислить)

$$f(x) = f(x_1, x_2, \ldots, x_n) \to \min_x$$

Минимизируем вдоль одной координаты i

$$f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) \to \min_z$$

Минимизация линейным поиском
(перебор всех значений в окрестности).

Покоординатный спуск

$$f(x) = f(x_1, x_2, \ldots, x_n) \to \min_x$$

1. Инициализация

$$k = 0, \quad x^k = \text{начальное приближение}$$

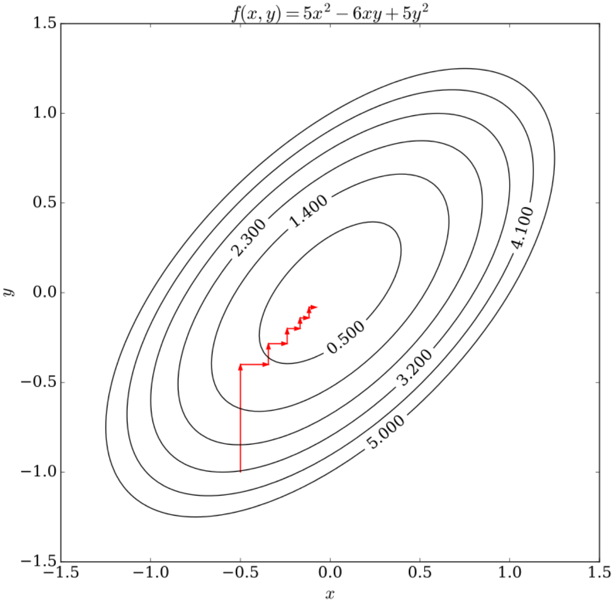
2. Минимизируем вдоль координаты i для $i = 1, \ldots, n$

$$z^* = \arg\min_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)$$

$$x^{k+1} = (x_1^k, \ldots, x_{i-1}^k, z^*, x_{i+1}^k, \ldots, x_n^k)$$

$$k := k + 1$$

3. Повторение до сходимости



$f(x,y) = 5x^2 - 6xy + 5y^2$
2.300
1.400
0.500
3.200
4.100
5.000
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
x
y

Вариации градиентного спуска

GD + линейный поиск

$$f(x) \to \min_x$$

1. Инициализация

$$k = 0, \quad x_k = \text{начальное приближение}$$

2. Минимизация в направлении сильнейшего убывания

$$\alpha^* = \arg\min_\alpha f(x_k - \alpha \nabla f(x_k))$$

$$x_{k+1} = x_k - \alpha^* \nabla f(x_k)$$

3. Повторение до сходимости

$$k := k + 1, \quad \text{перейти к } 2$$

GD + переменный шаг

$$f(x) \to \min_x$$

η — базовая величина шага (гиперпараметр)

1. Инициализация

$$k = 0, \quad x_k = \text{начальное приближение}$$

2. Шаг в сторону сильнейшего убывания

$$x_{k+1} = x_k - \frac{\eta}{k} \nabla f(x_k)$$

3. Повторение до сходимости

$$k := k + 1, \quad \text{перейти к 2}$$

GD + инерция

$$f(x) \to \min_x$$

η — величина шага, α — влияние инерции

1. Инициализация

$$k = 0, \quad x_k = \text{начальное приближение}, \quad m_k = 0$$

2. Шаг в сторону сильнейшего убывания

$$m_{k+1} := -\eta \nabla f(x_k) + \alpha m_k$$

$$x_{k+1} = x_k + m_{k+1}$$

3. Повторение до сходимости

$$k := k + 1, \quad \text{перейти к } 2$$

Резюме

Обзор методов

- Градиентный спуск
- Стохастический градиентный спуск
- Метод Ньютона
- Покоординатный спуск

Источники

- G. Venter. Review of Optimization Techniques.
- Википедия :)