Математика для Data Science. Математический анализ. Шпаргалка

Содержание

Четвёртая неделя. Градиентный спуск

Одномерный градиентный спуск

Для поиска минимума дифференцируемой функции $f : [a, b] \to \mathbb{R}$ мы можем использовать следующий **Алгоритм *градиентного спуска* в одномерном случае**:

1. Выберем какую-нибудь точку $r_1 \in [a, b]$.
2. Обозначим за i номер шага градиентного спуска. Сейчас $i = 1$.
3. Вычислим $f'(r_i)$.
4. Если $f'(r_i) = 0$, то алгоритм останавливается.

   Если $f'(r_i) > 0$, то мы сдвигаемся влево — выбираем $\delta > 0$ и назначаем $r_{i+1} = r_i - \delta$.

   Если $f'(r_i) < 0$, то мы сдвигаемся вправо — выбираем $\delta > 0$ и назначаем $r_{i+1} = r_i + \delta$.
5. Заменяем i на $i + 1$ и повторяем шаги 3, 4, 5.

Если в градиентом спуске мы делаем шаг на $-\lambda f'(r_i)$ для некоторого положительного числа $\lambda > 0$, то такое λ называется *learning rate* или *скоростью обучения*. В таком случае в 4 пункте алгоритма $r_{i+1} = r_i - \lambda f'(r_i)$.

$\mathbb{R}^n$: расстояния и векторы

$\mathbb{R}^n$ — это множество упорядоченных наборов вида $(x_1, x_2, \ldots, x_n)$, таких что $\forall i : x_i \in \mathbb{R}$. Каждый такой набор называется *точкой* $\mathbb{R}^n$.

Мы называем f *функцией многих переменных*, если f отображает D в $\mathbb{R}$, где $D \subset \mathbb{R}^n$ для какого-то n. Другими словами, область определения f должна быть подмножеством $\mathbb{R}^n$, а область значений f — подмножеством $\mathbb{R}$.

Евклидово расстояние между точками $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ и $b = (b_1, \ldots, b_n) \in \mathbb{R}^n$ определяется как

$$d(a, b) := \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_1 - b_1)^n}.$$

Точка $a \in \mathbb{R}^n$ называется *пределом последовательности* $\{x_i\}$, где $x_i \in \mathbb{R}^n$, если для любого $\varepsilon > 0$ найдётся натуральное число N, такое что $d(x_i, a) < \varepsilon$ при всех $i \geq N$ (т.е все x_i лежат в ε-окрестности точки a при $i \geq N$).

Неформальное определение *векторного пространства*:

- Все элементы $\mathbb{R}^n$ называются *векторами*, а само множество $\mathbb{R}^n$ называется *векторным пространством.*
- Векторы можно складывать друг с другом. Результатом сложения также будет вектор из этого же векторного пространства.

  В общем случае сумма векторов $(a_1, a_2, \ldots, a_n), (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$ определяется так:

  $$(a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n) \in \mathbb{R}^n.$$
- Также векторы можно умножать на числа.

  В общем случае умножение вектора $(a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ на число $c \in \mathbb{R}$ (это число называется *скаляром*) определяется так:

  $$c(a_1, a_2, \ldots, a_n) = (ca_1, ca_2, \ldots, ca_n) \in \mathbb{R}^n.$$

Мы иногда будем называть элементы $\mathbb{R}^n$ точками, а иногда векторами.

Длина вектора $x = (x_1, x_2, \ldots, x_n)$ определяется так:

$$||x|| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

.

Дифференциал

- Функции вида $a_1\Delta x_1 + \dots + a_n\ \Delta x_n$ называются *линейными функциями* от $(\Delta x_1, \dots,\ \Delta x_n)$.
- Выражение $a_1\Delta x_1 + \dots + a_n\ \Delta x_n$ называют линейным приращением функции f.
- А функцию $g(x_1 + \Delta x_1, \dots, x_n + \Delta x_n) = f(x_1, \dots,\ x_n) + a_1\Delta x_1 + \dots + a_n\Delta x_n$ называют *линейным приближением* функции f в точке x.
- **Неформальное определение дифференциала**

$$f(x_1 + \Delta x_1, \dots, x_n + \Delta x_n) - f(x_1, \dots, x_n) \approx d_x f(\Delta x_1, \dots, \Delta x_n) := a_1\Delta x_1 + \dots + a_n\Delta x_n.$$

  В общем случае коэффициенты $a_1, \dots, a_n$ зависят от выбранной точки $x = (x_1, \dots, x_n)$.

Формальное определение дифференциала. Пусть f это функция от n переменных. Функция $d_x f(\Delta x_1, \dots, \Delta x_n) := a_1\Delta x_1 + \dots + a_n\ \Delta x_n$ называется дифференциалом функции f в точке $x = (x_1, \dots, x_n)$, если следующий предел существует и равен нулю:

$$\lim_{(\Delta x_1, \dots, \Delta x_n) \to (0, \dots, 0)} \frac{f(x_1 + \Delta x_1, \dots,\ x_n + \Delta x_n) - (f(x) + a_1\Delta x_1 + \dots + a_n\ \Delta x_n)}{||(\Delta x_1, \dots, \Delta x_n)||} :=$$
$$:= \lim_{(\Delta x_1, \dots, \Delta x_n) \to (0, \dots, 0)} \frac{f(x_1 + \Delta x_1, \dots, x_n + \Delta x_n) - (f(x) + d_x f(\Delta x_1, \dots, \Delta x_n))}{||(\Delta x_1, \dots, \Delta x_n)||} = 0$$

Обозначив вектор $(\Delta x_1, \dots, \Delta x_n)$ за Δx, получим, что формула из предыдущего определения эквивалентна такой:

$$\lim_{\Delta x\ \to 0} \frac{f(x + \Delta x) - (f(x) + d_x f(\Delta x))}{||\Delta x||} = 0.$$

Здесь $x, \Delta x$ и $(x + \Delta x)$ – векторы из n переменных. Ноль в выражении $\lim\limits_{\Delta x\ \to 0}$ это сокращённая запись вектора $(0, \dots, 0)$. Ноль в правой части равенства это просто число $0 \in \mathbb{R}$ (не вектор).

Если у функции f существует дифференциал в точке x, то функция f называется *дифференцируемой в точке* x.

Функция f называется *дифференцируемой*, если она дифференцируема во всех точках своей области определения.

Свойства дифференциала

1. **Единственность дифференциала.** Пусть f – функция от n переменных. Если у функции f существует дифференциал в точке x, то этот дифференциал единственен.
2. **Дифференциал произведения на константу.** Пусть f дифференцируема в точке x. Тогда для любого числа $c \in \mathbb{R}$ функция cf дифференцируема в точке x, и
$$d_x(cf) = c \cdot d_x f$$
3. **Дифференциал суммы.** Пусть f и g дифференцируемы в точке x. Тогда функция $f + g$ дифференцируема в точке x, и
$$d_x(f + g) = d_x f + d_x g$$
4. **Дифференциал произведения.** Пусть f и g дифференцируемы в точке x. Тогда функция $f \cdot g$ дифференцируема в точке x, и
$$d_x(f \cdot g) = f(x) \cdot d_x g + g(x) \cdot d_x f.$$
Заметьте, что в этом выражении $f(x)$ и $g(x)$ это просто числа, потому что точка x зафиксирована.
5. **Дифференциал частного.** Пусть f и g дифференцируемы в точке x. Пусть g определена и не равна нулю в некоторой окрестности точки x. Тогда функция $\frac{f}{g}$ дифференцируема в точке x, и
$$d_x\left(\frac{f}{g}\right) = \frac{g(x) \cdot d_x f\ - f(x) \cdot d_x g}{g(x)^2}.$$
Заметьте, что в этом выражении $f(x)$ и $g(x)$ это просто числа, потому что точка x зафиксирована.

6. **Дифференциал сложной функции.** Пусть f – функция от одной переменной, а g – функция от n переменных. Тогда $f(g(x))$ это функция от n переменных (эта функция называется композицией функций f и g). Пусть g дифференцируема в точке x, а f имеет производную в точке $g(x)$. Тогда функция $f(g(x))$ тоже дифференцируема в точке x и её дифференциал равен

$$f'(g(x)) \cdot d_x g.$$

Заметьте, что в этом выражении $f(g(x))$ это просто число.

Частная производная

Пусть дана функция $f : \mathbb{R}^n \to \mathbb{R}$ и точка $x = (x_1, \dots, x_n) \ \in \mathbb{R}^n$. Тогда *частной производной* по k-ой координате называется предел

$$\frac{\partial f}{\partial x_k} := \lim_{t \to 0} \frac{f(x_1, \dots, x_k + \Delta x_k, \dots, x_n) - f(x_1, \dots, x_k, \dots, x_n)}{\Delta x_k}.$$

При вычислении частной производной по x_k можно считать все остальные переменные в формуле константами. Или можно воспользоваться таким алгоритмом:

1. В формуле для f подставить конкретные значения для всех координат, кроме k-ой. То есть мы подставляем следующие $(n-1)$ чисел: первую координату точки x, вторую координату точки x, и т.д. – все кроме k-ой координаты точки x. Получится функция от одной переменной – от переменной x_k.

2. У полученной функции от одной переменной вычислить производную.

3. Найти эту производную в конкретной точке – подставляем k-ую координату точки x.

Функция, полученная в Пункте 1 описывает, как ведёт себя f на прямой, проходящей через точку x и параллельной k-ой координатной оси. То есть мы фиксируем все координаты, кроме k-ой, и разрешаем изменять только k-ую координату. Выражение, полученное в пункте 1 называют *ограничением* функции f на эту прямую. Найденная частная производная описывает скорость роста функции f вдоль этой прямой в точке x.

Теорема. Дана функция f от n переменных. Пусть у f в точке x существует дифференциал $d_x f(\Delta x_1, \dots, \Delta x_n) = a_1 \Delta x_1 + \dots + a_n \ \Delta x_n$ и частные производные $\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}$. Тогда

$$a_1 = \frac{\partial f}{\partial x_1}, \dots, a_n \ = \frac{\partial f}{\partial x_n}.$$

То есть для любого $j = 1, \dots, n$ число a_j равно частной производной функции f по j-ой координате, вычисленной в точке x. Другими словами:

$$d_x f(\Delta x_1, \dots, \Delta x_n) = \frac{\partial f}{\partial x_1} \Delta x_1 + \frac{\partial f}{\partial x_2} \Delta x_2 \ + \dots + \frac{\partial f}{\partial x_n} \Delta x_n,$$

где все частные производные вычислены в точке x.

Теорема. Дана функция f от n переменных. Пусть f определена в некоторой окрестности точки x, и в точке x у f существуют частные производные по всем координатам. Тогда x может быть точкой локального минимума или максимума только если все частные производные равны нулю.

Следствие. Пусть в точке x также существует дифференциал $d_x f$. Точка x может быть точкой локального минимума или максимума, только если $d_x f = 0$ (то есть $d_x f(\Delta x_1, \dots, \Delta x_n) = 0$ для любых $\Delta x_1, \dots, \Delta x_n$).

Мы можем интерпретировать $\frac{\partial f}{\partial x_k}$ как функцию, которая отображает каждую точку $x \in \mathbb{R}^n$ в частную производную $\frac{\partial f}{\partial x_k}$ вычисленную в этой точке (для тех $x \in \mathbb{R}^n$, в которых $\frac{\partial f}{\partial x_k}$ определена).

Свойства частной производной как функции

Пусть у функций f и g определены частные производные по x_k. Тогда для частной производной выполнены следующие утверждения, аналогичные утверждениям для обычной производной:

1. у функции $f + g$ определена частная производная по x_k и $\frac{\partial (f+g)}{\partial x_k} = \frac{\partial f}{\partial x_k} + \frac{\partial g}{\partial x_k}$,

2. у функции cf определена частная производная по x_k и $\frac{\partial(cf)}{\partial x_k} = c\frac{\partial f}{\partial x_k}$, где $c \in \mathbb{R}$,

3. у функции fg определена частная производная по x_k и $\frac{\partial(fg)}{\partial x_k} = \frac{\partial f}{\partial x_k}g + f\frac{\partial g}{\partial x_k}$,

4. у постоянной функции c частная производная по x_k равна нулю.

Направление и градиент

Вектор длины 1 называется *направлением.*

Введём обозначение для вектора $a := (a_1, \dots, a_n)$. Соответственно, длина этого вектора равна $||a|| = \sqrt{a_1^2 + \dots + a_n^2}$.

Теорема. Среди всех направлений $(\Delta x_1, \dots, \Delta x_n)$ функция $d_x f(\Delta x_1, \dots, \Delta x_n) = a_1 \Delta x_1 + \dots + a_n\, \Delta x_n$ достигает минимального значения на направлении $(\Delta x_1, \dots, \Delta x_n) = \left(\frac{-a_1}{||a||},\ \frac{-a_2}{||a||}, \dots\ , \frac{-a_n}{||a||}\right) = -\frac{a}{||a||}$. При этом по теореме из предыдущего урока $a_k = \frac{\partial f}{\partial x_k}$.

Направлением ненулевого вектора a называется вектор $\frac{a}{||a||}$.

Для нулевого вектора (вектора, состоящего из одних нулей) направление не определено. Два вектора с совпадающими направлениями называются *сонаправленными*, а с противоположными направлениями — *противонаправленными.*

Вектор

$$\nabla f(x) := \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n}\right)$$

называется *градиентом* функции f в точке x.

Тем самым, теорема из этого урока говорит, что направление противоположное направлению градиента — это направление наискорейшего убывания функции. Другими словами, шаг градиентного спуска нужно делать против направления градиента. То есть в направлении вектора $(-\nabla f(x))$.