Математика для Data Science. Математический анализ. Шпаргалка

Содержание

Первая неделя. Введение, множества и доказательства

Объекты и целевая функция

Целевая функция — то, что мы хотим научиться вычислять для объектов из некоторого множества.

Признаки (или *фичи*) — набор того, что описывает объекты.

Регрессия — тип задач, где значение целевой функции может быть произвольным числом из некоторого промежутка.

Классификация — тип задач, в которых нужно отнести объект к одному из классов. *Бинарная классификация* — частный случай классификации, в которой возможных классов всего два.

Объект в машинном обучении часто представляют в виде *вектора* или *матрицы*. Неформально говоря, *вектор* — это набор значений признаков, а *матрица* — это табличка, в которой стоят значения признаков.

Функция потерь и данные в машинном обучении

Функция потерь показывает, насколько предсказанный ответ далёк от реального.

Пусть a — предсказанный ответ, y — реальный ответ. Приведём **примеры функций потерь** $L(y, a)$ в рассмотренных нами типах задач.

1. Задача регрессии.
   - Модуль отклонения: $L(y, a) = |y - a|$.
   - Квадрат отклонения: $L(y, a) = (y - a)^2$.
2. Задача бинарной классификации для классов 0 и 1.
   - *Индикаторная функция потерь*: $L(0, 0) = L(1, 1) = 1$ и для всех остальных аргументов $L(y, a) = 0$. Обозначается $L(y, a)$ как $\mathbf{1}\{y = a\}$.
   - Функция потерь, предсказывающая не класс объекта, а *вероятность* принадлежности объекта к одному из классов.

Обучающая выборка — набор размеченных данных, то есть набор объектов, для которых известно значение целевой функции.

Пусть объекты пронумерованы числами от 1 до n, и для этих объектов значения целевой функции – $y_1, y_2, \dots, y_n$ соответственно, а предсказание нашего алгоритма – $a_1, a_2, \dots, a_n$ соответственно. Приведём **примеры функций потерь для нашей выборки:**

1. Задача регрессии.
   - *Mean absolute error (MAE)* или *среднее отклонение по модулю* – это среднее арифметическое модулей отклонений:
     $MAE(y_1, y_2, \dots, y_n, a_1, a_2, \dots, a_n) = \frac{1}{n}(|y_1 - a_1| + |y_2 - a_2| + \dots + |y_n - a_n|)$.
   - *Mean squared error (MSE)* или *среднеквадратичная ошибка* — это среднее арифметическое квадратов отклонений:
     $MSE(y_1, y_2, \dots, y_n, a_1, a_2, \dots, a_n) = \frac{1}{n}((y_1 - a_1)^2 + (y_2 - a_2)^2 + \dots + (y_n - a_n)^2)$.
2. Задача бинарной классификации.
   - *Точность (accuracy)* — доля правильных ответов: $Acc(y_1, y_2, \dots, y_n, a_1, a_2, \dots, a_n) = \frac{1}{n}(\mathbf{1}\{y_1 = a_1\} + \mathbf{1}\{y_2 = a_2\} + \dots + \mathbf{1}\{y_n = a_n\})$.

Модель машинного обучения

Как правило, при решении задачи машинного обучения выбирается некоторый класс алгоритмов, где каждый конкретный алгоритм из класса задаётся *параметрами* или, иначе говоря, *весами*.

Алгоритм с фиксированными весами называется *моделью*.

Примеры классов алгоритмов:

1. Задача бинарной классификации.

   - Класс *константных функций*. Ему принадлежат алгоритмы, которые всегда выдают один и тот же ответ — постоянную величину c.
   - Класс *пороговых функций*. Ему принадлежат классификаторы вида $\mathbf{1}\{s \le t\}$. Здесь $\mathbf{1}$ — индикаторная функция, которая возвращает 1, если условие внутри фигурных скобок выполнено, и 0 иначе. За s обозначено значение признака, а t — фиксированое число, называемое *порогом* (от слова *threshold*).

2. Задача регрессии.

   - Класс *линейных функций*. Ему принадлежат функции вида $\hat{f}(r, d, p) = w_r r + w_d d + w_p p + w_0$, где r, d, p — значения признаков (в общем случае их n, где n — количество признаков), а w_r, w_d, w_p, w_0 — коэффициенты (в общем случае их $n+1$). w_0 называется *свободным коэффициентом*, его ещё называют *сдвигом* (или *bias*).

Множества

Множество — математический объект, являющийся набором других объектов.

Объекты, из которых состоит множество, называют *элементами множества* или *точками множества*.

Множества обычно обозначают заглавными буквами латинского алфавита, а элементы множества — строчными.

Любой элемент содержится в множестве не больше одного раза. Множества, отличающиеся порядком элементов, считаются одинаковыми.

$x \in A$ читается как «x является элементом множества A» или «x принадлежит A».

$y \notin A$ читается как «y не принадлежит A».

Способы задания множества

- Перечислить его элементы внутри фигурных скобок.
- Задать описанием: «множество всех x, таких что для них выполнено условие P». Записывается это в форме $\{x \mid P(x)\}$.

Операции над множествами

- *Пересечение* множеств A и B — это множество $A \cap B := \{x \mid x \in A \text{ и } x \in B\}$. Здесь знак «:=» читается как «по определению равно».
- *Объединение* множеств A и B — это множество $A \cup B := \{x \mid x \in A \text{ или } x \in B\}$.
- *Разность* множеств A и B — это множество $A \setminus B := \{x \mid x \in A \text{ и } x \notin B\}$.

Множество A называется *подмножеством* множества B, если все элементы множества A также являются элементами множества B. Обозначение: $A \subset B$.

Пустое множество — это множество, в котором нет элементов.

Способы изобразить множества

1. На *диаграмме Эйлера* множества рисуются как круги, а внутри кругов располагаются элементы.
2. В общем случае, если про множества ничего не известно, рисуют диаграмму *Эйлера-Венна* — диаграмму Эйлера со всеми возможными пересечениями.

Некоторые часто встречающиеся множества

- $\mathbb{N} = \{1, 2, 3, \dots\}$ — множество *натуральных* чисел, то есть чисел, возникающих при счёте.
- $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ — множество *целых* чисел.
- $\mathbb{Q}$ — множество *рациональных* чисел, то есть чисел, которые можно записать в виде дроби $\frac{m}{n}$, где $m \in \mathbb{Z}$ и $n \in \mathbb{N}$. Числа $x \notin \mathbb{Q}$ называются *иррациональными*.

- $\mathbb{R}$ — множество *действительных* (или *вещественных*) чисел. Действительное число — это бесконечная десятичная дробь, то есть выражение вида $\pm a_0.a_1a_2a_3\ldots$, где $\pm$ — это знак $+$ или знак $-$, a_0 — целое неотрицательное число, и $a_i \in \{0, 1, 2, \ldots, 9\}$ для всех $i \geq 1$.
- $\mathbb{R}^n$ — множество всех наборов из n действительных чисел.

Некоторые часто встречающиеся подмножества $\mathbb{R}$

- При $a < b$ *отрезком* называется множество $[a, b] := \{x \mid a \leq x \leq b\}$. Точки a и b называются *граничными* точками отрезка.
- При $a < b$ *интервалом* называется множество $(a, b) := \{x \mid a < x < b\}$.
- *Замкнутыми лучами* называются множества $[a, +\infty) := \{x \mid a \leq x\}$ и $(-\infty, a] := \{x \mid x \leq a\}$. Точка a называется *граничной* точкой замкнутого луча.
- *Открытыми лучами* называются множества $(a, +\infty) := \{x \mid a < x\}$ и $(-\infty, a) := \{x \mid x < a\}$.
- Интервал $(x_0 - \varepsilon, x_0 + \varepsilon)$ называется *ε-окрестностью* точки x_0, где ε (читается как «эпсилон») — положительное действительное число.
- Проколотой *ε-окрестностью* точки x_0 называется ε-окрестность точки x_0, в которую не входит сама точка x_0.

How-to по доказательствам

Общепринятые сокращения:

- $\implies$ — *следствие.* $A \implies B$ означает следующее: если выполнено утверждение A, то выполнено утверждение B.
- $\iff$ — *равносильность* (читается «тогда и только тогда»). Выражение $A \iff B$ означает, что $A \implies B$ и $B \implies A$. То есть: если выполнено утверждение A, то выполнено утверждение B, и наоборот — если верно утверждение B, то верно утверждение A.
- *Квантор* всегда идёт вместе с переменной или набором переменных, после чего идёт утверждение, в котором этот x фигурирует. Есть два вида кванторов:
  1. $\exists$ — *квантор существования.* $\exists x$: $A(x)$ означает, что существует значение x, при подстановке которого утверждение $A(x)$ становится истинным.
  2. $\forall$ — *квантор всеобщности.* $\forall x$: $A(x)$ означает, что для любого значения x утверждение $A(x)$ истинно.
- $\neg$ — отрицание. Отрицание к утверждению A записывается как $\neg A$.

Правила построения отрицаний

1. $\neg(A \text{ или } B) = \neg A$ и $\neg B$. То есть отрицанием к утверждению вида «A или B» будет утверждение: «(A неверно) и (B неверно)».
2. $\neg(A \text{ и } B) = \neg A$ или $\neg B$. То есть отрицанием к утверждению вида «A и B» будет утверждение: «(A неверно) или (B неверно)».
3. $\neg(\forall x : A(x)) \iff \exists x : \neg A(x)$. То есть отрицанием к утверждению вида «для всех x верно $A(x)$» будет утверждение вида «существует x такой, что неверно $A(x)$».
4. $\neg(\exists x : A(x)) \iff \forall x : \neg A(x)$. То есть отрицанием к утверждению вида «существует x такой, что верно $A(x)$» будет утверждение вида «для всех x неверно $A(x)$».

Закон контрапозиции гласит, что для утверждений X и Y выполнено

$$(X \implies Y) \iff (\neg Y \implies \neg X).$$

То есть утвеждения «если X, то Y» и «если неверно Y, то X тоже неверно» эквивалентны.

Закон контрапозиции используется при *доказательстве от противного*: мы предполагаем, что доказываемое утверждение неверно, после чего выводим противоречие.

Функции

Функция — это соответствие между элементами двух множеств, такое что каждому элементу первого множества соответствует ровно один элемент второго множества. Пусть первое множество обозначено через X, второе через Y, а функция через f. Тогда мы будем говорить, что «функция f отображает X в Y», «функция f из X в Y» или $f : X \to Y$.

Элемент $x \in X$, к которому мы применяем функцию f, называется *аргументом* функции, а элемент $f(x) \in Y$ называется *значением* функции.

Если функция f отображает X в Y, то X называется *областью определения* функции f. Множество всех значений, которые принимает функция f, называется *областью значений* функции f.

Точкой минимума функции $f : X \to \mathbb{R}$ называется такой $x_{min} \in X$, что $f(x_{min}) \leq f(x)$ для всех $x \in X$.

Вторая неделя. Последовательности и пределы

Знакомство с последовательностями и пределом

Последовательность элементов множества X — это функция $f : \mathbb{N} \to X$. Действительно, обозначив $x_1 = f(1), x_2 = f(2), x_3 = f(3), \ldots$, мы получим последовательность.

Элемент x_i называют i-ым *членом* последовательности, а число i называют его *индексом*.

Последовательность $x_1, x_2, x_3, \ldots$ принято компактно записывать при помощи фигурных скобок: $\{x_n\}_{n=1}^{\infty}$ или просто $\{x_n\}$.

Числовая последовательность — это последовательность элементов множества $\mathbb{R}$. Далее слово «числовая» будет опускаться.

Виды ограниченных последовательностей

1. Последовательность *ограничена снизу*, если существует такое число $C_1 \in \mathbb{R}$, что для всех $n \in \mathbb{N}$ выполнено $x_n > C_1$.

2. Последовательность *ограничена сверху*, если существует такое число $C_2 \in \mathbb{R}$, что для всех $n \in \mathbb{N}$ выполнено $x_n < C_2$.

3. Последовательность *ограничена*, если она ограничена и сверху, и снизу. То есть, если существуют такие числа $C_1 \in \mathbb{R}$ и $C_2 \in \mathbb{R}$, что для всех $n \in \mathbb{N}$ выполнено $C_1 < x_n < C_2$.

Число a называется *пределом* последовательности $\{x_n\}$, если для любого $\varepsilon > 0$ найдётся натуральное число N, такое что $|x_n - a| < \varepsilon$ при всех $n \geq N$.

Другими словами, a — предел $\{x_n\} \iff \forall \varepsilon > 0 \, \exists N \in \mathbb{N} \, \forall n \geq N : |x_n - a| < \varepsilon$.

Предел последовательности x_n обозначается $\lim\limits_{n\to\infty} x_n$.

Способы записать, что $\lim\limits_{n\to\infty} x_n = a$

- x_n стремится к a,
- последовательность $\{x_n\}$ сходится к a,
- $x_n \xrightarrow[n\to\infty]{} a$ или, короче, $x_n \to a$.

Если у последовательности есть предел, то мы будем говорить, что последовательность является *сходящейся* или же просто *сходится*.

Арифметика пределов

Свойства предела последовательности

1. Если у последовательности есть предел, то он единственен.

2. Если последовательность $\{x_n\}$ сходится, то она ограничена.

3. Пусть $\lim\limits_{n\to\infty} x_n = a$ и $\lim\limits_{n\to\infty} y_n = b$, тогда

   - $\lim\limits_{n\to\infty} c \cdot x_n = c \cdot a$, где c — некоторое число ($c \in \mathbb{R}$);
   - $\lim\limits_{n\to\infty} (x_n + y_n) = a + b$;
   - $\lim\limits_{n\to\infty} (x_n \cdot y_n) = a \cdot b$;
   - если $b \neq 0$ и $y_n \neq 0$ для всех n, то $\lim\limits_{n\to\infty} \frac{x_n}{y_n} = \frac{a}{b}$.

Последовательности, стремящиеся к $+\infty$ и $-\infty$

1. Последовательность $\{x_n\}$ *стремится к плюс бесконечности*, если для любого $C \in \mathbb{R}$ начиная с некоторого номера $N \in \mathbb{N}$ все члены последовательности больше C.

   Обозначение: $x_n \to +\infty$ или $\lim\limits_{n\to\infty} x_n = +\infty$.

   Другими словами, $x_n \to +\infty \iff \forall C \in \mathbb{R}\ \exists N \in \mathbb{N}\ \forall n \geq N : x_n > C$.

2. Последовательность $\{x_n\}$ *стремится к минус бесконечности*, если для любого $C \in \mathbb{R}$ начиная с некоторого номера $N \in \mathbb{N}$ все члены последовательности меньше C.

   Обозначение: $x_n \to -\infty$ или $\lim\limits_{n\to\infty} x_n = -\infty$.

   Другими словами, $x_n \to -\infty \iff \forall C \in \mathbb{R}\ \exists N \in \mathbb{N}\ \forall n \geq N : x_n < C$.

Бесконечно малой последовательностью называют последовательность, которая сходится к нулю.

Свойства бесконечно малых последовательностей

1. Если a — предел последовательности $\{x_n\}$, то её можно представить в виде $\{a + \alpha_n\}$, где $\{\alpha_n\}$ — бесконечно малая последовательность.

2. Пусть $\{\alpha_n\}$, $\{\beta_n\}$ — бесконечно малые последовательности, $c \in \mathbb{R}$, тогда бесконечно малыми будут также последовательности $\{c \cdot \alpha_n\}$, $\{\alpha_n + \beta_n\}$ и $\{\alpha_n \cdot \beta_n\}$.

Третья неделя. Пределы, производные и исследование функций

Пределы функций и непрерывные функции

Пусть f — функция с областью определения $D \subset \mathbb{R}$ и значениями в $\mathbb{R}$. Число $a \in \mathbb{R}$ называется *пределом функции* f в точке x_0, если $\lim\limits_{n\to\infty} f(x_n) = a$ для любой последовательности $\{x_n\}$, такой что $\lim\limits_{n\to\infty}(x_n) = x_0$ и $x_n \in D \setminus \{x_0\}$ для всех n (мы предполагаем, что хотя бы одна такая последовательность существует). Предел функции f в точке x_0 обозначается $\lim\limits_{x\to x_0} f(x)$.

Свойства предела функции

1. Если у функции есть предел в точке, то он единственен.
2. Пусть даны две функции f и g с совпадающими областями определения, такие что $\lim\limits_{x\to x_0} f(x) = a$ и $\lim\limits_{x\to x_0} g(x) = b$, тогда
   - $\lim\limits_{x\to x_0} c \cdot f(x) = c \cdot a$, где $c \in \mathbb{R}$,
   - $\lim\limits_{x\to x_0} f(x) + g(x) = a + b$,
   - $\lim\limits_{x\to x_0} f(x) \cdot g(x) = a \cdot b$,
   - если $b \neq 0$, то $\lim\limits_{x\to x_0} \frac{f(x)}{g(x)} = \frac{a}{b}$.

Функция f называется *непрерывной* в точке x_0, если f определена в точке x_0 и $\lim\limits_{x\to x_0} f(x)$ существует и равен $f(x_0)$.

Функция f называется *непрерывной*, если она непрерывна во всех точках своей области определения.

Точки области определения f, в которых f не является непрерывной, называются *точками разрыва*.

Есть и эквивалентное первому определение непрерывности: функция f называется *непрерывной* в точке x_0, если $x_0 \in D$, и для любого $\varepsilon > 0$ найдётся $\delta > 0$, такое что выполнено $|f(x) - f(x_0)| < \varepsilon$ для всех $x \in D$, удовлетворяющих $|x - x_0| < \delta$.

Свойства непрерывных функций Пусть f и g — непрерывные функции с совпадающими областями определения, тогда

- функция $c \cdot f$, где $c \in \mathbb{R}$, непрерывна,
- функция $f + g$ непрерывна,
- функция $f \cdot g$ непрерывна.

Производные: интуиция без доказательств

Мгновенной скоростью в момент времени t_0 называется предел $\lim\limits_{t\to t_0} \frac{S(t)-S(t_0)}{t-t_0}$, где $S(t)$ — расстояние, пройденное к моменту времени t.

Обозначив в этом определении $t - t_0$ за Δt, получим, что $\lim\limits_{t\to t_0} \frac{S(t)-S(t_0)}{t-t_0} = \lim\limits_{\Delta t\to 0} \frac{S(t_0+\Delta t)-S(t_0)}{\Delta t}$.

Производные: формально с доказательствами

Пусть функция f определена на некотором интервале, и точка x_0 принадлежит этому интервалу. Тогда *производной* функции f в точке x_0 называется число $\lim\limits_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}$ или, эквивалентно, число $\lim\limits_{t\to 0} \frac{f(x_0+t)-f(x_0)}{t}$. Производная обозначается $f'(x_0)$.

Если производная в точке x_0 существует, то функция f называется *дифференцируемой* в точке x_0.

Функция f называется *дифференцируемой* на некотором интервале, если она дифференцируема в каждой точке этого интервала. Тогда производной функции f называется функция f', которая отображает x в $f'(x)$.

Производная: вычисления без доказательств.

Производные часто встречающихся функций

- $c' = 0$, где $c \in \mathbb{R}$
- $(x^n)' = nx^{n-1}$, где n — произвольное действительное число, например: $x^{\frac{1}{2}} = (\sqrt{x})' = \frac{1}{2\sqrt{x}}$
- $(a^x)' = a^x \cdot \ln a$, в частности $(e^x)' = e^x$
- $(\log_a x) = \frac{1}{x \ln a}$, в частности $(\ln x) = \frac{1}{x}$
- $(\sin x)' = \cos x$
- $(\cos x)' = -\sin x$
- $(\operatorname{tg} x)' = \frac{1}{\cos^2 x}$
- $(\operatorname{ctg} x)' = -\frac{1}{\sin^2 x}$

Свойства производных

Пусть функции f и g определены и дифференцируемы на интервале (a, b), тогда

1. функция cf тоже дифференцируема на (a, b) и $(cf)' = cf'$, где c — произвольное действительное число,
2. функция $f + g$ тоже дифференцируема на (a, b) и $(f + g)' = f' + g'$,
3. функция fg тоже дифференцируема на (a, b) и $(fg)' = f'g + fg'$,
4. если g не обращается в ноль на этом интервале, то функция $\frac{f}{g}$ тоже дифференцируема на (a, b) и $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$.

Пусть теперь функция f непрерывна и дифференцируема в точке x_0, а g непрерывна и дифференцируема в точке $y_0 = f(x_0)$. Тогда

5. функция $g(f(x))$ непрерывна и дифференцируема в точке x_0 и её производная в точке x_0 равна $g'(y_0)f'(x_0) = g'(f(x_0))f'(x_0)$.

   Функция $g(f(x))$ называется *композицией* функций f и g и обозначается $g \circ f$.

Исследование функций при помощи производных.

Пусть дана функция f с областью определения D. Точка $x_0 \in D$ называется *точкой локального минимума*, если существует такой $\varepsilon > 0$, что $f(x) \geq f(x_0)$ для всех $x \in D \cap (x_0 - \varepsilon, x_0 + \varepsilon)$. При этом $f(x_0)$ называется *локальным минимумом*.

Точка x_0 называется *точкой минимума* (иногда говорят *точкой глобального минимума*) функции f, если $f(x_0) \leq f(x)$ для всех x из области определения f. Число $f(x_0)$ называют *минимумом функции* или *глобальным минимумом функции* или *минимальным значением функции*.

x_0 называется *точкой перегиба* функции $f(x)$, если производная в этой точке равна нулю: $f'(x_0) = 0$, но функция не достигает в x_0 ни локального минимума, ни локального максимума.

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 $(x_1, x_2, \dots, x_n)$, 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 $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ и $b = (b_1, \dots, b_n) \in \mathbb{R}^n$ определяется как

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

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 $(a_1, a_2, \dots, a_n), (b_1, b_2, \dots, b_n) \in \mathbb{R}^n$ определяется так:

  $$(a_1, a_2, \dots, a_n) + (b_1, b_2, \dots, b_n) = (a_1 + b_1, a_2 + b_2, \dots, a_n + b_n) \in \mathbb{R}^n.$$

- Также векторы можно умножать на числа.

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

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

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

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

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

.

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 $(-\nabla f(x))$.

Пятая неделя. Модификации градиентного спуска

Градиентный спуск

Для поиска минимума функции $f : \mathbb{R}^n \to \mathbb{R}$ мы можем использовать следующий
Алгоритм градиентного спуска:

1. Выберем какую-нибудь начальную точку $r_0 \in \mathbb{R}^n$.
2. Обозначим за i номер шага градиентного спуска. Сейчас $i = 1$.
3. Вычислим градиент $\nabla f(r_i)$.
4. Если $\nabla f(r_i) = 0$, то алгоритм останавливается. Иначе выбираем $\delta > 0$ и сдвигаемся на δ в направлении $(-\nabla f(r_i))$. Называем точку, в которую мы попадаем, r_{i+1}.
5. Заменяем i на $i+1$ и повторяем шаги 3, 4, 5.

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

Линейная регрессия и градиентный спуск

В задаче линейной регрессии мы представляем объекты в виде набора признаков, каждый из которых является некоторым числом. А затем пробуем найти для каждого признака коэффициент такой, чтобы при сложении признаков с данными коэффициентами мы получали что-то близкое к нашей целевой функции.

Обозначения в задаче линейной регрессии

1. Объект это x, объект номер i из обучающей выборки это $x^{(i)}$.
2. Каждый объект задаётся n признаками: $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}) \in \mathbb{R}^n$. Итого, $x_j^{(i)}$ это j-ый признак i-ого объекта.
3. Значение целевой функции на объекте $x^{(i)}$ обозначается за $y^{(i)} \in \mathbb{R}$.
4. Обучающая выборка это m пар: $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$.
5. Предсказание нашей модели на объекте $x^{(i)}$ обозначается за $\hat{y}^{(i)}$.
6. Значение квадратичной функции потерь на i-ом объекте это $L(y^{(i)}, \hat{y}^{(i)}) := (\hat{y}^{(i)} - y^{(i)})^2$.
7. Значение функции потерь для всей выборки это $\frac{1}{2} \sum\limits_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$.
8. Модель задаётся набором параметров $\theta = (\theta_0, \theta_1, \ldots, \theta_n) \in \mathbb{R}^{n+1}$.
9. Для каждого объекта вводим фиктивный признак x_0, который всегда равен 1. Тогда $\hat{y} = h_\theta(x) := \sum\limits_{j=0}^{n} \theta_j x_j$.
10. Функция потерь $J(\theta) := \frac{1}{2} \sum\limits_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 = \frac{1}{2} \sum\limits_{i=1}^{m} \left(\sum\limits_{j=0}^{n} \theta_j x_j^{(i)} - y^{(i)} \right)^2$. Её-то мы и будем минимизировать.

Чтобы сделать шаг градиентного спуска, нам нужно найти градиент, то есть

$$\nabla J(\theta) := \left(\frac{\partial J}{\partial \theta_0}(\theta), \frac{\partial J}{\partial \theta_1}(\theta), \ldots, \frac{\partial J}{\partial \theta_n}(\theta) \right).$$

Тогда из точки $(\theta_0, \theta_1, \ldots, \theta_n)$ после одного шага градиентного спуска с learning rate α мы попадём в точку

$$\left(\theta_0 - \alpha \frac{\partial J}{\partial \theta_0}(\theta), \theta_1 - \alpha \frac{\partial J}{\partial \theta_1}(\theta), \ldots, \theta_n - \alpha \frac{\partial J}{\partial \theta_n}(\theta) \right).$$

Стохастический градиентный спуск

Один шаг градиентного спуска задаётся следующим преобразованием координат вектора $(\theta_0, \theta_1, \ldots, \theta_n)$:

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial J}{\partial \theta_j}(\theta) = \theta_j - \alpha \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) \cdot x_j^{(i)} \text{ для всех } j \in \{0, 1, \ldots, n\}$$

Итак, в алгортиме градиентного спуска мы на каждом шаге должны считать значения функции потерь для всех m объектов выборки.

А алгоритм *стохастического градиентного спуска (или SGD)* же выглядит следующим образом.

Мы перемешиваем наши m объектов из обучающей выборки. Делаем m шагов — по одному для каждого объекта $x^{(i)}$. На одном шаге координаты вектора θ меняются так:

$$\theta_j \leftarrow \theta_j - \alpha \left(\hat{y}^{(i)} - y^{(i)}\right) \cdot x_j^{(i)} \text{ для всех } j \in \{0, 1, \ldots, n\}$$

Потом повторяем эту процедуру k раз. Каждая такая итерация из m шагов называется одной *эпохой*. А число k соответственно называется *количеством эпох*.

Разберём теперь, как работает *mini-batch gradient descent*.

Мы выбираем число, которое будет размером *mini-batch*, например 16. Перемешиваем нашу обучающую выборку из m элементов. Разбиваем её на $\frac{m}{16}$ частей по 16 объектов. Каждая такая часть называется mini-batch. Делаем $\frac{m}{16}$ шагов — по одному шагу для каждого mini-batch. Одна такая процедура называется одной эпохой. Так же, как и в SGD мы можем сделать несколько эпох.

Градиентный спуск с моментом

Опишем алгоритм *градиентного спуска с моментом*.

Пусть $\alpha > 0$ — learning rate, а $0 < \beta < 1$ — коэффициент момента. Пусть также $V_0 = 0$.

Пусть после $(t-1)$-ого шага градиентного спуска с моментом мы находимся в какой-то точке. Обозначим за S_t градиент функции потерь в этой точке. Положим

$$V_t := \beta V_{t-1} + (1-\beta) S_t$$

или, эквивалентно,

$$V_t = (1-\beta)(\beta^{t-1} S_1 + \beta^{t-2} S_2 + \cdots + \beta^2 S_{t-2} + \beta^1 S_{t-1} + S_t)$$

и сделаем шаг на $(-\alpha) V_t$.

Ещё можно делать шаг не на $(-\alpha V_t)$, а на $\frac{-\alpha}{1-\beta^t} V_t$. Величина $\frac{1}{1-\beta^t} V_t$ при этом называется *экспоненциально взвешенным средним* от $S_1, S_2, \ldots, S_t$.

RMSprop

Как и раньше, обозначим через S_t градиент функции потерь на t-ом шаге. Кроме того, пусть $0 < \beta_2 < 1$ и ε — очень маленькая константа, обычно её берут равной 10^{-8}.

Пусть $s_{t,j}$ это j-ая координата вектора S_t. То есть $S_t = (s_{t,1}, s_{t,2}, s_{t,3}, \ldots)$.

Обозначим $u_{t,j} := (1-\beta_2)(\beta_2^{t-1} s_{1,j}^2 + \beta_2^{t-2} s_{2,j}^2 + \cdots + \beta_2^2 s_{t-2,j}^2 + \beta_2^1 s_{t-1,j}^2 + s_{t,j}^2)$.

Разделим j-ую координату S_t на $\sqrt{\frac{1}{1-\beta_2^t} u_{t,j}} + \varepsilon$. Мы делаем такую операцию с каждой координатой вектора S_t. Умножаем полученный вектор на $(-\alpha)$ и на такой вектор делаем шаг градиентного спуска.