Математика для Machine Learning

Линдеманн Никита

Ноябрь 2018

Содержание

1 Элементы математического анализа

1.1 Производная функции одной переменной

Определение 1.1.1 *Производная* функции одной переменной $y = f(x)$ – это предел:

$$\frac{d}{dx}f(x) = \frac{dy}{dx} = f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}.$$

Пример

Найдем по определению производную функции $y = x^2$ в точке $x_0 = 3$:

$$y'(x) = \lim_{\Delta x \to 0} \frac{(x+\Delta x)^2 - x^2}{\Delta x} = \lim_{\Delta x \to 0} \frac{2x\Delta x + \Delta x^2}{\Delta x} = \lim_{\Delta x \to 0}(2x + \Delta x) = 2x.$$

Далее подставим значение $x_0 = 3$ в найденную функцию $y'(x) = 2x$:

$$y'(x_0) = y'(3) = 2 \cdot 3 = 6.$$

Конечно, на практике никто не считает производные по определению, для этого есть готовая таблица производных элементарных функций (которую полезно помнить наизусть) и правила дифференцирования.

Правила дифференцирования:

1) Линейность производной: $(\alpha f(x) + \beta g(x))' = \alpha f'(x) + \beta g'(x)$.
2) Производная произведения: $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$.
3) Производная частного:

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}.$$

4) Производная сложной функции:

$$(f(u))' = f'(u) \cdot u'.$$

Пример

Найдем производную функции $y(x) = \frac{\cos(x^2+1)}{\ln(1-2x)+3}$:

$$y'(x) = \frac{(\cos(x^2+1))' \cdot (\ln(1-2x)+3) - (\cos(x^2+1)) \cdot (\ln(1-2x)+3)'}{(\ln(1-2x)+3)^2}.$$

Отдельно найдем производные:

$$(\cos(x^2+1))' = 2x \cdot (-\sin(x^2+1)).$$

$$(\ln(1-2x)+3)' = -2 \cdot \frac{1}{1-2x}.$$

Окончательно имеем:

$$y'(x) = \frac{-2x\sin(x^2+1) \cdot (\ln(1-2x)+3) + (\cos(x^2+1)) \cdot \frac{2}{1-2x}}{(\ln(1-2x)+3)^2}.$$

Здесь стоит упомянуть про физический и геометрический смысл производной: ее можно интерпретировать как скорость изменения какой-либо величины (например, скорость материальной точки в механике) или как тангенс угла наклона касательной к графику функции. Именно из интерпретации производной как тангенса угла наклона касательной следуют важные прикладные теоремы, ради которых мы и ввели это понятие.

ТЕОРЕМА Функция примнимает свое локально максимальное или минимальное (экстремальное) значение только в тех точках, где ее производная равна нулю (обратное не всегда верно):

$$y(x_0) \to min \ \Rightarrow \ y'(x_0) = 0,$$
$$y(x_0) \to max \ \Rightarrow \ y'(x_0) = 0.$$

ТЕОРЕМА Функция $y = f(x)$ строго *возрастает* на промежутке (a, b), тогда и только тогда, когда в каждой точке этого промежутка производная $f'(x)$ строго положительна, или на математическом языке:

$$y = f(x) \nearrow, \ x \in (a, b) \Leftrightarrow \ \forall \ x \in (a, b) \hookrightarrow f'(x) > 0.$$

ТЕОРЕМА Функция $y = f(x)$ строго *убывает* на промежутке (a, b), тогда и только тогда, когда в каждой точке этого промежутка производная $f'(x)$ строго отрицательна, или на математическом языке:

$$y = f(x) \searrow, \ x \in (a, b) \Leftrightarrow \ \forall \ x \in (a, b) \hookrightarrow f'(x) < 0.$$

1.2 Частные производные и градиент

Определение 1.2.1 Рассмотрим функцию $f(x_1, x_2, \dots, x_n)$ от n переменных $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. *Частная производная* $f(x_1, x_2, \dots, x_n)$ по переменной x_k – это предел:

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

Нахождение частных производных почти ничем не отличается от нахождения производной функции одной переменной (а чаще даже бывает и проще) – просто надо найти производную функции $f(x_k)$, считая все остальные переменные $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$ константами.

<u>Пример</u>

Найдем дифференциал функции $f(x, y, z) = x \cos(2y - z) + xy$ по формуле

$$df(x, y, z) = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz.$$

Для этого найдем частные производные:

$$\frac{\partial f}{\partial x} = f'_x = \cos(2y - z) + y,$$
$$\frac{\partial f}{\partial y} = f'_y = 2x \cdot (-\sin(2y - z)) + x,$$
$$\frac{\partial f}{\partial z} = f'_z = -x \cdot (-\sin(2y - z)).$$

Окончательно имеем:

$$df(x, y, z) = (\cos(2y - z) + y)dx + (-2x \cdot \sin(2y - z) + x)dy + (x \cdot \sin(2y - z))dz.$$

Определение 1.2.2 *Градиентом* функции $f(x_1, x_2, \ldots, x_n)$ от n переменных называется n-мерный вектор, составленный из частных прозводных функции f:

$$\mathbf{grad}\, f(x_1, x_2, \ldots, x_n) = \begin{pmatrix} f'_{x_1} \\ f'_{x_2} \\ \vdots \\ f'_{x_n} \end{pmatrix}.$$

Смысл градиента таков: это вектор, указывающий в направлении наибольшего возрастания функции $f(x_1, x_2, \ldots, x_n)$, значение которой меняется от одной точки пространства к другой, а по величине (модулю) равный скорости роста этой функции в этом направлении.

Пример

Найдем градиент функции $f(x_1, x_2, x_3, x_4) = x_2 \sin(x_1 x_3) + 2x_4$ в точке $(\frac{3\pi}{2}, 0, 1, -2)$. Частые производные:

$$f'_{x_1} = x_2 x_3 \cos(x_1 x_3),$$
$$f'_{x_2} = \sin(x_1 x_3),$$
$$f'_{x_3} = x_2 x_1 \cos(x_1 x_3),$$
$$f'_{x_4} = 2.$$

Подставляя в частные производные соответствующие значения переменных, находим:

$$\mathbf{grad}\, f = \begin{pmatrix} 0 \\ -1 \\ 0 \\ 2 \end{pmatrix}.$$

2 Основы линейной алгебры

2.1 Векторы и матрицы

Определение 2.1.1 *Вектор* – это упорядоченный одномерный массив чисел.

Мы будем обозначать векторы строчными латинскими буквами полужирным шрифтом и записывать их в виде вектор-столбца:

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}.$$

Определение 2.1.2 *Матрица* – это упорядоченный двумерный массив чисел.

Множество матриц с m строками и n столбцами, то есть матрицы размеров $m \times n$, будем обозначать $\mathbf{M}_{m\times n}$. Произвольную матрицу $A \in \mathbf{M}_{m\times n}$ будем обозначать заглавными латинскими буквами и записывать в следующем виде:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix},$$

где на пересечении i-той строки с j-тым стобцом будет находится элемент матрицы a_{ij}. То есть первый индекс элемента матрицы – номер строки, второй – номер столбца, в котором находится элемент.

Заметим, что вектор – частный случай матрицы: действительно, матрица размерами $m\times 1$ – в точности и есть вектор-столбец. Значит, если мы научимся работать с матрицами, то сразу же получим и правила работы с векторами. Поэтому далее пока что не будем отдельно рассматривать векторы, а будем работать с матрицами произвольных размеров $m \times n$, считая, что возможны случаи $n = 1$.

2.2 Сложение и вычитание матриц, умножение матриц на число

Определение 2.2.1 Пусть даны две матрицы A, $B \in \mathbf{M}_{m\times n}$: $A = (a_{ij})$, $B = (b_{ij})$. Матрица $C = (c_{ij})$ называется *суммой матриц A и B*, если $c_{ij} = a_{ij} + b_{ij}$ для любых $i = \overline{1, m}$ и $j = \overline{1, n}$. Будем обозначать $A + B = C$.

Определение 2.2.2 Пусть даны две матрицы A, $B \in \mathbf{M}_{m\times n}$: $A = (a_{ij})$, $B = (b_{ij})$. Матрица $C = (c_{ij})$ называется *разностью матриц A и B*, если $c_{ij} = a_{ij} - b_{ij}$ для любых $i = \overline{1, m}$ и $j = \overline{1, n}$. Будем обозначать $A - B = C$.

Очевидно, что при таком определении мы можем складывать только матрицы одного размера, и это достаточно важное замечание!

<u>Пример</u>

$$\begin{pmatrix} 1 & -4 & 0 \\ 5 & 3 & -1 \end{pmatrix} + \begin{pmatrix} 4 & 5 & -7 \\ 0 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 1 & -7 \\ 5 & 5 & 0 \end{pmatrix}.$$

Определение 2.2.3 Пусть дана матрица $A \in \mathbf{M}_{m\times n}$: $A = (a_{ij})$ и число λ. Матрица $C = (c_{ij})$ называется *произведением матрицы A на число λ*, если $c_{ij} = \lambda a_{ij}$ для любых $i = \overline{1, m}$ и $j = \overline{1, n}$. Будем обозначать $\lambda A = C$.

Пример:

$$8 \cdot \begin{pmatrix} 1 & 7 \\ 2 & -3 \end{pmatrix} = \begin{pmatrix} 8 & 56 \\ 16 & -24 \end{pmatrix}.$$

Так как мы ввели сложение и вычитание матриц и умножение матриц на число покомпонентно, то все свойства чисел, связанные с этими операциями, остаются верными и для матриц.

2.3 Умножение матриц

Определение 2.3.1 Пусть даны две матрицы $A \in \mathbf{M}_{m\times n}$ и $B \in \mathbf{M}_{n\times p}$: $A = (a_{ij})$, $B = (b_{ij})$. Матрица $C = (c_{ij}) \in \mathbf{M}_{m\times p}$ называется *произведением матриц A и B*, если

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

для любых $i = \overline{1, m}$ и $j = \overline{1, n}$. Будем обозначать $A \cdot B = C$.

Из определения следует, что мы можем перемножать матрицы, только если количество столбцов у первой равно количеству строк у второй. Это важное замечание!

Пример
Найдем $C = AB$, где

$$A = \begin{pmatrix} 0 & -1 \\ 4 & 5 \\ -8 & 6 \end{pmatrix}, \; B = \begin{pmatrix} 3 & 1 \\ -2 & 0 \end{pmatrix}.$$

1. Найдем размер результирующей матрицы. Для этого запишем размеры перемножаемых матриц, в данном случае это 3×2 и 2×2. Убеждаемся, что количество столбцов у первой матрицы равно количеству строк у второй, значит, мы можем перемножить матрицы. Размер результирующей матрицы будет равен «количество строк первой матрицы $\times$ количество столбцов второй матрицы», то есть 3×2.

2. Последовательно найдем элементы результирующей матрицы $C = c_{ij}$. Для этого обозначим матрицы $A = (a_{ij})$, $B = (b_{ij})$:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}, \; B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

Тогда по определению:

$$c_{11} = \sum_{k=1}^{2} a_{1k} b_{k1} = a_{11} b_{11} + a_{12} b_{21} = 0 \cdot 3 + (-1) \cdot (-2) = 2,$$

$$c_{21} = \sum_{k=1}^{2} a_{2k} b_{k1} = a_{21} b_{11} + a_{22} b_{21} = 0 \cdot 1 + (-1) \cdot 0 = 0,$$

$$c_{21} = \sum_{k=1}^{2} a_{2k} b_{k1} = a_{21} b_{11} + a_{22} b_{21} = 4 \cdot 3 + 5 \cdot (-2) = 2,$$

$$c_{22} = \sum_{k=1}^{2} a_{2k} b_{k2} = a_{21} b_{12} + a_{22} b_{22} = 4 \cdot 1 + 5 \cdot 0 = 4,$$

$$c_{31} = \sum_{k=1}^{2} a_{3k} b_{k1} = a_{31} b_{11} + a_{32} b_{21} = (-8) \cdot 3 + 6 \cdot (-2) = -36,$$

$$c_{32} = \sum_{k=1}^{2} a_{3k} b_{k2} = a_{31} b_{12} + a_{32} b_{22} = (-8) \cdot 1 + 6 \cdot 0 = -8.$$

3. Запишем ответ:

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 2 & 4 \\ -36 & -8 \end{pmatrix} = 2 \cdot \begin{pmatrix} 1 & 0 \\ 1 & 2 \\ -18 & -4 \end{pmatrix}.$$

Таким образом, можно вывести мнемоническое правило для произведения матриц: чтобы получить элемент c_{ij} результирующей матрицы, умножаем i-ую строку первой матрицы на j-ый столбец второй матрицы «почленно».

Свойства умножения матриц:
1) Ассоциотивность: $A(BC) = (AB)C$.
2) Антикоммутативность: $AB \neq BA$.
3) Дистрибутивность: $A(B + C) = AB + AC$, $(A + B)C = AC + BC$.
4) Ассоциативность и коммутативность относительно умножения на число: $(\lambda A)B = \lambda(AB) = A(\lambda B)$.

2.4 Транспонирование, скалярное произведение, длина вектора

Определение 2.4.1 Пусть дана матрица $A \in \mathbf{M}_{m\times n}$: $A = (a_{ij})$. Матрица $C = (c_{ij}) \in \mathbf{M}_{n\times m}$ называется *транспонированной к* A, если $c_{ij} = a_{ji}$ для любых $i = \overline{1, m}$ и $j = \overline{1, n}$. Будем обозначать транспонированную к A матрицу как A^T.

Пример

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

То есть при транспонировании матрицы i-ая строка становится i-ым столбцом, а j-ый столбец стоновится j-ой строкой.

Свойства транспонирования:
1) $(A + B)^T = A^T + B^T$.
2) $(A - B)^T = A^T - B^T$.
3) $(\lambda A)^T = \lambda A^T$.
4) $(AB)^T = B^T A^T$.

Определение 2.4.2 *Скалярное произведение векторов*

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

это

$$\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T \cdot \mathbf{b} = \begin{pmatrix} a_1 & a_2 & \dots & a_n \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = \sum_{k=1}^{m} a_k b_k.$$

Пример

Найдем скалярное произведение векторов $\mathbf{a}^T = (-2 \ \ 6 \ \ 0 \ \ 5)$ и $\mathbf{b}^T = (4 \ \ 3 \ \ 8 \ \ -2)$:
По определению имеем:

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{k=1}^{4} a_k b_k = (-2) \cdot 4 + 6 \cdot 3 + 0 \cdot 8 + 5 \cdot (-2) = 0.$$

Мы получили, что векторы $\mathbf{a}$ и $\mathbf{b}$ ортогональны.

Определение 2.4.3 Два ненулевых вектора называются *ортогональными*, если их скалярное произведение равно нулю.

Определение 2.4.4 *Длина (модуль) вектора* $\mathbf{a}^T = (a_1 \ \ a_2 \ \ \dots \ \ a_m)$ – квадратный корень из скалярного произведения этого вектора на себя, то есть:

$$a = |\mathbf{a}| = \sqrt{\sum_{k=1}^{m} (a_k)^2}\,.$$

Пример

Найдем длину вектора $\mathbf{a}^T = (-2 \ \ 6 \ \ 0 \ \ 5)$. По определению:

$$a = |\mathbf{a}| = \sqrt{\sum_{k=1}^{4} (a_k)^2} = \sqrt{(-2)^2 + 6^2 + 0^2 + 5^2} = \sqrt{65}.$$

Определение 2.4.5 *Скалярное произведение векторов* $\mathbf{a}$ и $\mathbf{b}$ – это произведение модулей (длин) этих векторов на косинус угла α между ними:

$$\langle \mathbf{a}, \mathbf{b} \rangle = ab \cdot \cos(\alpha).$$

Мы привели два определения скалярного произведения векторов, что может вызывать некоторую путаницу – ведь определения очень разные. На сомом деле эти определения эквивалентны (результат не будет зависить от способа вычисления), просто определение (2.4.5) более общее, а (2.4.2) более удобно при практическом вычислении.

2.5 Ранг и определитель матрицы

Определение 2.5.1 Система столбцов $A_1, A_2, \dots, A_n$ из $\mathbf{M}_{m\times 1}$ называется *линейно зависимой*, если некоторая их нетривиальная линейная комбинация равна нулевому столбцу, то есть если существуют числа $\lambda_1, \lambda_2, \dots, \lambda_n$ такие, что хотя бы одно из этих чисел не нулевое (нетривиальность) и выполнено:

$$\sum_{k=1}^{n} A_k\lambda_k = O = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Пример

Докажем, что следующие столбцы линейно зависимы:

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} -4 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 7 \\ 9 \end{pmatrix}.$$

Действительно, существует нетривиальная линейная комбинация этих столбцов, равная нулевому вектору (напомним, что мы отожествляем понятие матрицы размера $m \times 1$ с вектором):

$$3 \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} -4 \\ 1 \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ 7 \\ 9 \end{pmatrix} = O.$$

Определение 2.5.2 Целое неотрицительное число r называется *рангом матрицы* A и обозначается $r = \operatorname{rg} A$, если из столбцов этой матрицы можно выбрать r линейно независимых столбцов, но нельзя выбрать $r+1$ линейно независимых столбцов.

Свойства ранга:

1) $\operatorname{rg} A^T = \operatorname{rg} A$.

2) Если A – квадратная матрица и $|A| = 0$, то строки и столбцы матрицы линейно зависимы.

3) Линейная (не)зависимость столбцов матрицы эквивалентна линейной (не)зависимости строк.

4) Если все элементы строки или столбца матрицы умножить на отличное от нуля число, то ранг матрицы не изменится.

5) Ранг не изменится, если к элементам любой строки или столбца матрицы прибавить соответствующие элементы другой строки или столбца, умноженные на одно и то же число.

6) Ранг не изменится, если переставить два любых столбца или две строки матрицы.

Определение 2.5.3 *Определитель (или детерминант)* матрицы размера 2×2 – это

$$|A| = \det A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Детерминант матрицы размера 3×3 – это

$$|A| = \det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

Заметим, что детерминант определен только для квадратных матриц, то есть только для матриц размеров $n \times n$!

Пример

Найдем определитель:

$$|A| = \det A = \begin{vmatrix} 0 & 1 & 3 \\ 2 & 3 & 5 \\ 3 & 5 & 7 \end{vmatrix} = 0 \cdot \begin{vmatrix} 3 & 5 \\ 5 & 7 \end{vmatrix} - 1 \cdot \begin{vmatrix} 2 & 5 \\ 3 & 7 \end{vmatrix} + 3 \cdot \begin{vmatrix} 2 & 3 \\ 3 & 5 \end{vmatrix} = 0 - (14 - 15) + 3 \cdot (10 - 9) = 4.$$

Мы получили, что наша матрица не вырождена.

Определение 2.5.4 Матрица называется *вырожденой*, если ее определитель равен нулю.

Свойства детерминанта:

1) $|A^T| = |A|$.

2) $|AB| = |A| \cdot |B|$.

3) Умножение всех элементов строки или столбца определителя на некоторое число равносильно умножению определителя на это число.

4) Если матрица содержит нулевую строку или столбец, то определитель этой матрицы равен нулю.

5) Если две строки или два столбца матрицы линейно зависимы, то определитель этой матрицы равен нулю.

6) При перестановке двух любых строк или столбцов определитель матрицы меняет знак.

7) Определитель не изменится, если к элементам любой строки или столбца матрицы прибавить соответствующие элементы другой строки или столбца, умноженные на одно и то же число.

8) Определитель матрицы треугольного вида равен произведению элементов, стоящих на главной диагонали:

$$|A| = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{vmatrix} = \prod_{k=1}^{n} a_{kk} = a_{11} \cdot a_{22} \cdot \ldots \cdot a_{nn}.$$

Заметим, что последнее свойство позволяет легко находить определитель матриц большого порядка, если перед этим привести матрицу к диоганальному виду (что всегда можно сделать, опираясь на предпоследнее свойство и используя алгоритм Гаусса).

3 Некоторые понятия теории вероятностей

3.1 Классическая вероятность

Определение 3.1.1 *Дискретное вероятностное пространство* – это любой конечный набор элементов (*элементарных исходов*) $\Omega = \{\omega_1, \omega_2, \dots \omega_N\}$, каждому из которых сопоставлено некоторое положительное число $\omega_k \longmapsto P_k$ $(k = \overline{1, N})$, причем для этих чисел выполнено условие нормировки:

$$\sum_{k=1}^{N} P_k = 1.$$

Определение 3.1.2 Число, сопоставленное элементарному исходу ω_i, будем называть *вероятностью* этого исхода и считать, что эта вероятность определена как предел отношения:

$$P_i = \lim_{N \to \infty} \frac{N(\omega_i)}{N},$$

где $N(\omega_i)$ – это количество испытаний, при котором исход ω_i наступил, а N – общее число испытаний.

Определение 3.1.3 *Событие* – это любое подмножество A множества элементарных исходов Ω: $A \subseteq \Omega$.

Определение 3.1.4 *Вероятность события* A – это сумма вероятностей тех элементарных исходов, которые составляют событие A:

$$P(A) = \sum_{i:\,\omega_i \in A} P_i.$$

Пример

Рассмотрим опыт с бросанием игральной кости. В нашей терминологии:

1) Вероятностное пространство – это множество всевозможных элементарных исходов: $\Omega = \{\omega_i \mid \omega_i = \{\text{выпала } i\text{-ая грань}\},\ i = \overline{1, 6}\}$.

2) Так как в данном эксперименте естественно предположить, что выпадение любой грани равновероятно, то с учетом условия нормировки получается, что i-ая грань выпадет с вероятностью:

$$P_i = \frac{1}{6}.$$

3) Пример события – выпала четная грань. Действительно, выпадение четной грани есть подмножество $A = \{\omega_2, \omega_4, \omega_6\}$ всех элементарных исходов Ω. Тогда вероятность такого события – сумма:

$$P(A) = \sum_{i:\,\omega_i \in A} P_i = \sum_{i=1}^{3} P_i = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

3.2 Случайные величины

Определение 3.2.1 *Случайная величина* ξ – это любая функция, сопоставляющая элементарному исходу из множества Ω некоторое вещественное число, то есть:

$$\xi : \Omega \longrightarrow \mathbb{R}.$$

Очевидно, что по определению функции, случайная величина может принимать только одно из значений $\xi \in \{x_1, x_2, ..., x_m\}$, где $m \leqslant N(\Omega)$ (условимся для удобства записывать значения случайной величины в порядке возрастания: $x_i < x_{i+1}$).

Определение 3.2.2 *Вероятность того, что ξ принимает некоторое значение x_k* равна сумме вероятностей событий ω_i для которых $\xi(\omega_i) = x_k$ или

$$P_\xi(k) = \sum_{i:\xi(\omega_i)=x_k} P_i \ ,$$

где за $P_\xi(k)$ обозначена величина $P(\xi = x_k)$.

Таким образом, случайную величину удобно интерпритировать как результат какого-либо случайного события, то есть $\omega_i \longrightarrow \xi(\omega_i)$.

Обратим внимание на то, что под ξ в зависимости от контекста мы можем понимать немного разные вещи. Запись $\xi(\omega_i) = x_j$ означает, что случайная величина ξ ставит в соответствие элементарному исходу ω_i вещественное число x_j, и здесь ξ – функция. Так же под ξ мы можем понимать само значение, которое ставится в соответствие какому-то элементарному исходу.

Пример

Рассмотрим опыт с бросанием уже двух разных игральных костей (что на самом деле эквивалентно двум последовательным бросанием одной игральной кости). Пусть случайная величина – это сумма очков, выпавшая на костях. Тогда очевидно, что $\xi \in \{2, 3, \ldots, 12\}$.

Найдем вероятность того, что $\xi = 7$: это реализуется при элементарных исходах (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1), где на первом месте записано количество очков, выпавших на первой кости, а на втором – количество очков, выпавших на второй кости. Вероятность любого исхода (i, j) равна 1/36 при $i, j = \overline{1, 6}$. Тогда по определению:

$$P(\xi) = P(7) = 6 \cdot \frac{1}{36} = \frac{1}{6}.$$

3.3 Математическое ожидание и дисперсия

Определение 3.3.1 *Математическое ожидание* случайной величины – это сумма произведений вероятности того, что случайная величиная принимает некоторое значение, на это значение:

$$E(\xi) = \sum_{i=1}^{N(\Omega)} P_i \xi(\omega_i)$$

По сути матожидание – это некое среднее значение случайной величины при достаточно большом количестве испытаний.

Свойства математического ожидания:

1)$E(c\xi) = cE(\xi)$.

2)$E(\xi + \eta) = E(\xi) + E(\eta)$.

3)$\xi \geqslant 0 \Rightarrow E(\xi) \geqslant 0$.

4)$\forall\, i, j \hookrightarrow \xi_i \geqslant \eta_j \Rightarrow E(\xi) \geqslant E(\eta)$.

Пример

Рассмотрим следующую игру. В коробке лежит 10 фишек: 5 штук со зачением -5, 4 штуки со значением 2,5 и 1 фишка со значением 10. Предлагается вытащить одну из фишек и, если на фишке написано положительное число, то ваш выйгрыш – очки на фишке, умноженные на 10, но если вы вытащили фишку с отрицательным значением – вы должны отдать 50 рублей.

Посчитаем матожидание выйгрыша. Для этого составим закон распределения случайной величины ξ – выйгрыша:

ξ_i	-5	2,5	10
$P(\xi_i)$	0,5	0,4	0,1

По определению математического ожидания:

$$E(\xi) = \sum_{i=1}^{N(\Omega)} P(\xi_i)\xi_i = \sum_{i=1}^{3} P(\xi_i)\xi_i = 0,5 \cdot 5 + 0,4 \cdot 2,5 + 0,1 \cdot 10 = -0,5.$$

Таким образом, математическое ожидание данной игры проигрышно, то есть, сыграв много раз в эту игру, вы уйдете, оказавшись в минусе в среднем на 5 рублей.

Определение 3.3.2 *Дисперсией* случайной величины называется величина

$$D(\xi) = E((\xi - E(\xi))^2).$$

Дисперсия позволяет количественно оценить, насколько сильно конкретное значение случайной величины отличается от среднего значения, то есть насколько значения рассеяны, а рассеяние с латыни переводится не иначе, как дисперсия.

Свойства дисперсии:

1) $D(\xi) = E(\xi^2) - E(\xi)^2$.
2) $D(c\xi) = c^2 D(\xi)$.
3) $D(\xi) \geqslant 0$.
4) $D(\xi) = 0 \Leftrightarrow \forall\, i \hookrightarrow \xi(\omega_i) = E(\xi)$.
5) $D(\xi + \eta) \geqslant D(\xi) + D(\eta)$.

Первое свойство имеет важное прикладное значение. Именно по этой формуле чаще всего и считают дисперсию, так как это проще, чем по определению (в этом мы сможем убедиться на последующем примере).

Определение 3.3.3 *Среднеквадратичное отклонение (стандартное отклонение)* случайной величины – это квадратный корень из дисперсии:

$$\sigma(\xi) = \sqrt{D(\xi)}.$$

Среднеквадратичное отклонение так называется, поскольку характеризует отклонение случайной величины от ее математического ожидания.

Пример

Вернемся к предыдущему примеру. Там мы нашли неутешительное математическое ожидание $E(\xi) = -0,5$ нашей игры, и сейчас нам предстоит вычислить её дисперсию двумя способами: по определению и по первому свойству «дисперсия – это матожидание квадрата минус квадрат матожидания». Для поиска дисперсии первым способом составми таблицу:

ξ_i	-5	2,5	10
$P(\xi_i)$	0,5	0,4	0,1
$\xi_i - E(\xi)$	-4,5	3	10,5
$(\xi_i - E(\xi))^2$	20,25	9	110,25
$(\xi_i - E(\xi))^2 \cdot P(\xi_i)$	10,125	3,6	11,025

Суммируя значения последней строки, получаем дисперсию:

$$D(\xi) = 10,125 + 3,6 + 11,025 = 24,75.$$

Чтобы найти дисперсию по первому свойству, необходимо вычислить величины:

$$E(\xi^2) = \sum_{i=1}^{3} P(\xi_i)\xi_i^2 = 0,5 \cdot 25 + 0,4 \cdot 6,25 + 0,1 \cdot 100 = 25.$$

$$E(\xi)^2 = \left(\sum_{i=1}^{3} P(\xi_i)\xi_i\right)^2 = (0,5 \cdot (-5) + 0,4 \cdot 2,5 + 0,1 \cdot 10)^2 = (-0,5)^2 = 0,25.$$

Тогда, по формуле $D(\xi) = E(\xi^2) - E(\xi)^2$ получим такой же результат, как и в первом случае:

$$D(\xi) = 25 - 0,25 = 24,75.$$

Из этого примера видно, что последний способ вычисления дисперсии действительно проще, чем подсчет дисперсии по определению.

Размерность дисперсии – рубли в квадрате, поэтому сложно по дисперсии оценить разброс выйгрыша или проигрыша в игре. А вот по стандартному отклонению уже становится яснее, насколько сильно разбросаны эти значения относительно вычесленного матожидания:

$$\sigma(\xi) = \sqrt{D(\xi)} = \sqrt{24,75} \approx 5.$$