Математика для Machine Learning

Линдеманн Никита

5 мая 2019 г.

Содержание

1 Элементы математического анализа

1.1 Производная функции одной переменной

Определение 1.1.1. Производная функции одной переменной $y = f(x)$ в точке x_0 – это предел:

$$\frac{d}{dx}f(x_0) = \frac{d}{dx}y(x_0) = f'(x_0) = y'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

Если обозначить $\Delta x = x - x_0$, то производную функции в точке можно записать в виде:

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}.$$

Пример 1.1.1. Найдем по определению производную функции $y = x^2$ в точке $x_0 = 3$:

$$y'(3) = \lim_{x \to 3} \frac{x^2 - 3^2}{x - 3} = \lim_{x \to 3} \frac{(x-3)(x+3)}{x - 3} = \lim_{x \to 3}(x + 3) = 3 + 3 = 6.$$

Можно было действовать по-другому:

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

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:

$$(f(g(x)))' = f'_g(g) \cdot g'(x).$$

Пример 1.1.2. Найдем производную функции $y(x) = \dfrac{\cos(x^2 + 1)}{\ln(1 - 2x) + 3}$:

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

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:

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

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 1.1.1. *Функция примнимает свое локально максимальное или минимальное (экстремальное) значение только в тех точках, где ее производная равна нулю:*

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

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

Теорема 1.1.2. *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:*

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

Теорема 1.1.3. *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:*

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

1.2 Частные производные и градиент

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

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

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

Пример 1.2.1. 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:

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

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

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

Сформулируем главную теорему этого раздела.

Теорема 1.2.1. *Функция $y = f(x_1, x_2, \ldots x_n)$ принимает свое локально экстремальное значение только в тех точках, где ее градиент обращается в ноль:*

$$f(x_1, x_2, \ldots x_n) \to \min \;\Rightarrow\; \operatorname{grad} f(x_1, x_2, \ldots x_n) = \vec{0},$$

$$f(x_1, x_2, \ldots x_n) \to \max \;\Rightarrow\; \operatorname{grad} f(x_1, x_2, \ldots x_n) = \vec{0}.$$

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 $j = \overline{1, n}$.

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 $j = \overline{1, n}$.

Заметим, что при таком определении мы можем складывать только матрицы одного размера.

Пример 2.2.1.

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

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 $j = \overline{1, n}$.

Пример 2.2.2.

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

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 $j = \overline{1,n}$.

Из определения следует, что мы можем перемножать матрицы только тогда, когда количество столбцов у первой матрицы равно количеству строк у второй.

Пример 2.3.1. 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:

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

$$c_{12} = \sum_{k=1}^{2} a_{1k} b_{k2} = a_{11}b_{12} + a_{12}b_{22} = 0 \cdot 1 + (-1) \cdot 0 = 0,$$

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

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

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

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

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» и складываем.

Свойства умножения матриц:
1) Ассоциотивность: $A(BC) = (AB)C$.
2) Некоммутативность: $AB \neq BA$.
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 2.4.1.

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

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. Скалярным произведением векторов

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

называется

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

Пример 2.4.2. 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:

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

Пример 2.4.3. Найдем длину вектора $\mathbf{a}^T = (-2 \;\; 6 \;\; 0 \;\; 5)$. По определению:

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

Определение 2.4.5. Скалярное произведение векторов $\mathbf{a}$ и $\mathbf{b}$ – это произведение модулей 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) удобно при практическом вычислении, если известны координаты перемножаемых векторов.

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 2.5.1. 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 отождествляем 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 неотрицательное 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 (детерминант) квадратной матрицы $\underset{n\times n}{A} = (a_{ij})$ порядка n – это сумма по всем перестановкам из n элементов:

$$|\underset{n\times n}{A}| = \det \underset{n\times n}{A} = \sum_{\alpha_1 \dots \alpha_n} (-1)^{N(\alpha_1 \dots \alpha_n)} a_{1\alpha_1} \dots a_{n\alpha_n},$$

где $N(\alpha_1 \dots \alpha_n)$ – число инверсий в перестановке $(\alpha_1 \dots \alpha_n)$.

Так как количество перестановок (биекций множества в себя) на множестве из n элементов ровно $n!$, то количество слагаемых в сумме при подсчете определителя через формулу полного разложения будет тоже равно факториалу от n.

Определение 2.5.4. Определитель – это полилинейная кососимметричная функция столбцов $\det : \underset{n\times n}{\mathbf{M}} \to \mathbb{R}$, такая, что детерминант еденичной матрицы равен еденице: $\det(E) = 1$.

Как и в случае со скалярным произведением, мы ввели два эквивалентных определения детерминанта. Заметим, что детерминант определен только для квадратных матриц, то есть только для матриц размеров $n \times n$.

Пример 2.5.2. Найдем определитель матрицы: $A = \begin{pmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix} = \begin{pmatrix}1 & 1\\ 3 & 4\end{pmatrix}$.

1. По первому определению:
   $|A| = \sum_{\alpha_1,\alpha_2} (-1)^{N(\alpha_1,\alpha_2)} a_{1\alpha_1} \cdot a_{n\alpha_2} = (-1)^{N(1,2)} a_{11} \cdot a_{22} + (-1)^{N(2,1)} a_{12} \cdot a_{21} = 1 \cdot 4 - 2 \cdot 3 = -2.$

2. По второму определению:
$\det(A) = \det\left[\begin{pmatrix}1\\3\end{pmatrix}, \begin{pmatrix}2\\4\end{pmatrix}\right] = \det\left[\begin{pmatrix}1\\0\end{pmatrix} + 3\begin{pmatrix}0\\1\end{pmatrix},\ 2\begin{pmatrix}1\\0\end{pmatrix} + 4\begin{pmatrix}0\\1\end{pmatrix}\right] = \det\left[\begin{pmatrix}1\\0\end{pmatrix},\ 2\begin{pmatrix}1\\0\end{pmatrix}\right] +$
$\det\left[\begin{pmatrix}1\\0\end{pmatrix},\ 4\begin{pmatrix}0\\1\end{pmatrix}\right] + \det\left[3\begin{pmatrix}0\\1\end{pmatrix},\ 2\begin{pmatrix}1\\0\end{pmatrix}\right] + \det\left[3\begin{pmatrix}0\\1\end{pmatrix},\ 4\begin{pmatrix}0\\1\end{pmatrix}\right] = 4\det\left[\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right] -$
$6\det\left[\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right] = 4\det(E) - 6\det(E) = -2.$

На практике, для вычисления определителей, обычно используют следствие из определений, а именно формулу разложения по i-ой строке (чаще всего по первой):

$$|\underset{n\times n}{A}| = \sum_{j=1}^{n}(-1)^{i+j}a_{ij}M_{ij},$$

где M_{ij} – дополнительный минор к элементу матрицы a_{ij}, то есть определитель матрицы, которая получается из исходной при вычеркивании i-ой строки и j-того столбца.

Для матриц второго и третьего порядка определитель выражается следующим образом (разложение по первой строке):

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

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

Пример 2.5.3. 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 2.5.5. Матрица называется вырожденой, если ее определитель равен нулю.

Свойства детерминанта:
1) $|A^T| = |A|$.
2) $|AB| = |A| \cdot |B|$.
3) Умножение всех элементов одной 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 хотя бы 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).

2.6 Обратная матрица

Определение 2.6.1. Матрица A^{-1} называется обратной к квадратной невырожденной матрице A, если $AA^{-1} = A^{-1}A = E$.

Пример 2.6.1. Найдем обратную матрицу к

$$A = \begin{pmatrix} 2 & 5 \\ 1 & 3 \end{pmatrix}.$$

Решим задачу несколькими способами.

1. Воспользуемся методом Гаусса. Он заключается в следующем: к матрице, к которой ищется обратная, справа приписывается через черту единичная матрица, после этого, считая исходную матрицу и приписанную единичную одной матрицей, элементарными преобразованиями (прямым и обратным ходом метода Гаусса) исходную матрицу приводим к диоганальному виду, а потом к единичному. Получившаяся справа матрица и будет обратной к исходной. В нашем случае:

$$\left(\begin{array}{cc|cc} 2 & 5 & 1 & 0 \\ 1 & 3 & 0 & 1 \end{array}\right) \xrightarrow{II=I-2II} \left(\begin{array}{cc|cc} 2 & 5 & 1 & 0 \\ 0 & -1 & 1 & -2 \end{array}\right) \xrightarrow{I=I+5II} \left(\begin{array}{cc|cc} 2 & 0 & 6 & -10 \\ 0 & -1 & 1 & -2 \end{array}\right) \to \left(\begin{array}{cc|cc} 1 & 0 & 3 & -5 \\ 0 & 1 & -1 & 2 \end{array}\right).$$

   Запись над стрелкой $II = I - 2II$ означает, что на следующем шаге вместо второй строки будет записана разность первой строки и удвоенной второй.

2. Воспользуемся методом неопределенных коэффициентов. Пусть искомая матрица имеет вид:

$$A^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

   тогда по определению должно выполнятся

$$AA^{-1} = \begin{pmatrix} 2 & 5 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = E.$$

   Перемножая матрицы, имеем систему:

$$\begin{cases} 2a + 5c = 1, \\ 2b + 5d = 0, \\ a + 3c = 0, \\ b + 3d = 1. \end{cases}$$

   Решая, находим, что $a = 3,\ b = -5,\ c = -1,\ d = 2$.

3. Воспользуемся явной формулой для обратной матрицы:

$$A^{-1} = \frac{1}{|A|} A_+^T,$$

где A_+ – матрица алгебраических дополнений соответствующих элементов матрицы A, это значит, что элементы матрицы A_+ равны $a_{+ij} = (-1)^{i+j} M_{ij}$, где M_{ij} – дополнительный минор, то есть определитель матрицы, получающийся из исходной вычеркиванием i-ой строки и j-того столбца. Тогда согласно этой формуле:

$$A^{-1} = \frac{1}{1} \begin{pmatrix} 3 & -5 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & -5 \\ -1 & 2 \end{pmatrix}.$$

Итак, мы нашли, что

$$\begin{pmatrix} 2 & 5 \\ 1 & 3 \end{pmatrix}^{-1} = \begin{pmatrix} 3 & -5 \\ -1 & 2 \end{pmatrix}.$$

Заметим, что из формулы для обратной матрицы можно получить простое мнемоническое правило отыскания обратных матриц для матриц размером 2×2: надо просто поменять местами элементы на главной диоганали и домножить на -1 элементы на побочной диоганали, разделив полученную матрицу на детерминант исходной, мы и получим обратную матрицу:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Свойства обратной матрицы:
1) $(AB)^{-1} = B^{-1}A^{-1}$.
2) $|A^{-1}| = |A|^{-1}$.
3) $(A^T)^{-1} = (A^{-1})^T$.
4) $(kA)^{-1} = k^{-1}A^{-1}$.
5) $E^{-1} = E$.

2.7 Псевдообратная матрица

Понятие обратной матрицы удобно, например, при решении совместной системы линейных уравнений:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n2}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n; \end{cases}$$

Вводя обозначения

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix},$$

можно переписать систему в виде

$$AX = b.$$

Тогда решение запишется так:

$$X = A^{-1}b.$$

Однако, в более общем случае система может быть несовместная или иметь более одного решения (такое возможно, когда количество линейно независимых уравнений меньше количества переменных). В таком случае не получиться так лаконично записать решение из-за того, что матрица A можеть быть не квадратной или вырожденной.

Тогда мы хотим найти такое решение системы, при котором величина $|AX - b|^2$ будет минимальной (почему минимизируется именно квадрат разности будет объяснено в разделе математической статистики). Для этого от исходной системы переходят к следующему уравнению:

$$A^T AX = A^T b,$$

решение которого формально записывается как

$$X = (A^T A)^{-1} A^T b.$$

Определение 2.7.1. A^+ называется псевдообратной матрицей для матрицы A, если она удовлетворяет следующим критериям:

1. $AA^+A = A$
2. $A^+AA^+ = A^+$ (A^+ является слабым обращением в мультипликативной полугруппе)
3. $(AA^+)^* = AA^+$ (это означает, что AA^+ – эрмитова матрица)
4. $(A^+A)^* = A^+A$ (A^+A – тоже эрмитова матрица)

Здесь $*$ обозначает операцию эрмитого сопряжения, то есть $A^* = \overline{A^T}$ (подчеркивание означает комплексное сопряжение).

2.8 Матричные разложения

Определение 2.8.1. Разложение матрицы A – представление матрицы A в виде произведения матриц, обладающих некоторыми определёнными свойствами (например, ортогональностью, симметричностью, диагональностью).

В теории матриц известно множество различных разложений (спектральное, полярное, сингулярное разложения, ЖНФ и ФНФ, LU, QR и QZ разложения или, например, разложения Холецкого и Шура). Для задач машинного обучения, а именно для понимания работы рекомендательных систем и анализа текста, нам понадобятся лишь некоторые из них.

3 Некоторые понятия теории вероятностей

3.1 Классическая вероятность

Машинное обучение как наука основано на теории вероятностей. Вообще говоря, это довольно сложная и содержательная наука, начала которой были положены еще в средние века при попытках анализа различного рода азартных игр. Многие ученные работали над этим разделом математики на протяжении многих веков, но свой современный вид теория вероятностей приобрела относительно недавно – в середине прошлого века. Аксиоматика, предложенная Андреем Николаевичем Колмогоровым, позволила формализовать теорию вероятностей, использовав уже хорошо разработанный на тот момент математический аппарат теории меры.

В основе теории вероятностей лежат такие понятия как борелевские подмножества, алгебры и сигма-алгебры, а сама вероятность понимается как мера на сигма-алгебре борелевских подмножеств. К счастью, чтобы понять суть машинного обучения, не обязательно знать эту науку так глубоко – для наших целей нам достаточно более простых определений.

Определение 3.1.1. Конечное дискретное вероятностное пространство – это тройка $(\Omega, \mathcal{F}, P)$, где:

1. $\Omega = \{\omega_1, \omega_2, \dots \omega_N\}$ – это произвольное непустое конечно множество, элементы которого называют элементарными исходами.
2. $\mathcal{F} = 2^\Omega$ – это множество всех подмножеств Ω, называемых (случайными) событиями.
3. $P : \mathcal{F} \to [0, 1]$ – функция, которая ставит в соответствие каждому событию какое-то вещественное число от 0 до 1, такая, что $P(\Omega) = 1$, то есть $\sum_{i=1}^{N} P(\omega_i) = 1$.

Определение 3.1.2. Число $P(\omega_i)$, сопоставленное элементарному исходу ω_i, будем называть вероятностью этого исхода и считать, что эта вероятность определена как предел отношения:

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

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

Определение 3.1.3. Вероятность события $A \in \mathcal{F}$ – это сумма вероятностей тех элементарных исходов, которые составляют событие A:

$$P(A) = \sum_{\omega_i \in A} P(\omega_i).$$

Обратим внимание, что мы рассматриваем модель эксперимента, при котором элементарные исходы являются взаимоисключающими и в результате опыта одно из них обязательно происходит. При этом наблюдается свойство устойчивости частоты случайного события: с увеличинием числа повторений опыта значение частоты появления случайного события стабилизируется возле некоторого неслучайного числа.

Пример 3.1.1. Рассмотрим опыт с бросанием игральной кости.
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:

$$P(\omega_i) = \frac{1}{6}.$$

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:

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

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 $\xi(\omega_i) = x_k$:

$$P_\xi(k) = \sum_{\xi(\omega_i) = x_k} P(\omega_i),$$

где за $P_\xi(k)$ обозначена вероятность того, что ξ принимает значение x_k, то есть 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 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 3.2.1. 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:

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

Необходимо понимать, что кроме дискретно распределенных случайных величин есть и другие, с непрерывным распределением. Непрерывные случайные величины в простейшем случае можно понимать как набор исходов, представляющий из себя вещественную прямую $\mathbb{R}$, тогда вероятности отдельных исходов превращаются в функцию распределения $F(a) = P(x < a)$, производная которой играет важную роль и называется плотностью распределения

$$f(x) = \frac{dF}{dx}.$$

Тогда условие нормировки превращается в условие равенства еденице интеграла от неотрицательной функции плотности:

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

3.3 Совместная вероятность и теорема Байеса

Определение 3.3.1. Совместная вероятность событий A и B – это вероятность одновременного наступления этих событий $P(A, B)$.

Определение 3.3.2. Два события A и B называются независимыми, если $P(A, B) = P(A)P(B)$.

Обратите внимание, что независимость событий в теории вероятностей определяется сугубо формально. Ее не надо путать ни с отношением «причина - следствие», которого между зависимыми случайными величинами может и не быть, ни с корреляцией, которая отражает только линейную часть зависимости между случайными величинами.

Определение 3.3.3. Условная вероятность – вероятность наступления события A, если известно, что произошло событие B, которое имеет положительную вероятность:

$$P(A|B) = \frac{P(A, B)}{P(B)}.$$

Определение 3.3.4. События A и B условно независимы при условии C, если выполнено $P(A, B \mid C) = P(A|C)P(B|C)$.

Из этих определений следует важнейшая для машинного обучения теорема – теорема Байеса.

Теорема 3.3.1 (Байеса).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Доказательство.

$$P(A|B) = \frac{P(A, B)}{P(B)}, \quad P(B|A) = \frac{P(A, B)}{P(A)}.$$

Выразим и приравняем из первого и второго равенства $P(A, B)$:

$$P(A|B)P(B) = P(B|A)P(A).$$

Из последнего равенства следует утверждение теоремы. □

Чтобы из совместной вероятности получить вероятность того или иного исхода одной из случайных величин, можно воспользоваться формулой полной вероятности, то есть просуммировать одну случайную величину по другой (маргинализация):

$$P(A) = \sum_B P(A|B)P(B).$$

Пример 3.3.1. Рассмотрим классическую задачу на применение теоремы Байеса.
Предположим, что некий тест на какую-нибудь страшную болезнь с 95% точностью определяет, болен ли человек. Предположим так же, что болезнь достаточно распространена и имеется у 1% респондентов. Пусть некоторый человек получил позитивный результат теста, то есть, тест говорит, что страшная болезнь у человека присутствует. С какой вероятностью выбранный человек действительно болен?
Пусть событие A – результат теста положителен, событие B – человек действительно болен. Нам надо найти вероятность, что человек болен, при условии, что результат теста положителен, то есть $P(B|A)$. По теореме Байеса:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

$P(A|B) = 95\%$ – вероятность того, что результат теста положителен, при условии, что болезнь есть. $P(B) = 1\%$ – вероятность того, что человек болен. Чтобы найти $P(A)$ воспользуемся маргинализацией, учтя, что для B возможны всего два исхода: B и $\overline{B}$ – человек либо болен, либо здоров, тогда:

$$P(A) = \sum_B P(A|B)P(B) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B}) = 95 \cdot 1 + 5 \cdot 99 = 590.$$

Окончательно имеем:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{95 \cdot 1}{590} \approx 0{,}161 \approx 16\%.$$

4 Начала математической статистики

4.1 Математическое ожидание и дисперсия

Определение 4.1.1. Математическое ожидание случайной величины – это сумма произведений вероятности того, что случайная величиная принимает некоторое значение, на это значение:

$$E(\xi) = \sum_{i=1}^{|\Omega|} P(\xi_i)\xi(\omega_i)$$

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 4.1.1. 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, то 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:

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

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 4.1.2. 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 (поэтому этот же термин используется в оптике, но означает, конечно, совсем другое).

Утверждение 4.1.1. *Для дисперсии справедливо следующее соотношение:*

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

Доказательство. Используя линейность матожидания, имеем:

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

□

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 4.1.3. Среднеквадратичное отклонение (стандартное отклонение) случайной величины – это квадратный корень из дисперсии:

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

Среднеквадратичное отклонение характеризует отклонение случайной величины от ее математического ожидания.

Пример 4.1.2. 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:

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

4.2 Дискретные распределения

Дискретные распределения описывают события, исход которых представляет собой счетное множество: успех или неудача, целое число, орёл или решка и так далее. Описывается дискретное распределение вероятностью наступления каждого из возможных исходов события.

Рассмторим последовательность из n независимых испытаний с двумя исходами: A и $\overline{A}$, которые назовем соответственно «успех» и «неудача», причем $P(A) = p \in (0,1)$, $P(\overline{A}) = 1 - p$. Такая схема испытаний называется схемой Бернулли, а сам опыт – опыт Бернулли.

Из комбинаторных соображений нетрудно получить, что при проведении опыта G по схеме Бернулли вероятность $P_n(k)$ события $A_n(k)$, состоящего в том, что при n повторениях опыта G событие A произойдет ровно k раз, равна:

$$P_n(k) = C_n^k p^k (1-p)^{n-k}.$$

Определение 4.2.1. Дискретная случайная величина ξ с реализациями $x_k = k$, $k = \overline{0,n}$ имеет биномиальное распределение с параметрами n – количество возможных исходов и $p \in (0,1)$, что записывается как $\xi \sim \mathrm{Bin}(n,p)$, если вероятность события $\xi = x_k$ определяется формулой Бернулли:

$$P(\xi = x_k) = P_k = C_n^k p^k (1-p)^{n-k} = C_n^k p^k q^{n-k}, \ q = 1-p.$$

Здесь $C_n^k = \dfrac{n!}{k!(n-k)!} = \dbinom{n}{k}$ – биномиальные коэффициенты, то есть коэффициенты в разложении бинома Ньютона:

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k} = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

Математическое ожидание и дисперсия биномиального распределния равны $E(\xi) = np$, $D(\xi) = npq = np(1-p)$.

Пример 4.2.1. Пусть монету подбрасывают три раза. Требуется найти ряд распределения (закон распределения, записанный в порядке возрастания случайной величины) числа ξ выпавших гербов.

Случайная величина ξ распределена по биномиальному закону с параметрами $n = 3$ и $p = 1/2$, поэтому ξ может принимать значения 0, 1, 2, 3 с вероятностями

$$P_0 = C_3^0 \left(\frac{1}{2}\right)^3 = \frac{1}{8}, \ P_1 = P_2 = C_3^1 \left(\frac{1}{2}\right)^3 = \frac{3}{8}, \ P_3 = 1 - (P_0 + P_1 + P_2) = \frac{1}{8}.$$

Таким образом, получаем следующий ряд распределения числа выпавших гербов:

ξ	0	1	2	3
P	1/8	3/8	3/8	1/8

Определение 4.2.2. Дискретная случайная величина ξ с реализациями $x_k = k$, $k = 0, 1, \ldots$ имеет распределение Пуассона с параметром $a > 0$, что записывается как $\xi \sim \mathrm{Pois}(a)$, если

$$P(\xi = x_k) = P_k = \frac{a^k}{k!} e^{-a}.$$

4.3 Непрерывные распределения

4.4 Оценки и статистики

4.5 Парная линейная регрессия

Определение 4.5.1. Ковариация двух случайных величин ξ и η, определенных в одном и том же вероятностном пространстве – это

$$\operatorname{cov}(\xi, \eta) = E[(\xi - E(\xi))(\eta - E(\eta))].$$

Ковариация – обобщение понятия дисперсии. Действительно, дисперсия есть ковариация случайной величины с самой собой: $\operatorname{cov}(\xi, \xi) = E[(\xi - E(\xi))(\xi - E(\xi))] = E[(\xi - E(\xi))^2] = D(\xi)$.

Утверждение 4.5.1. *Для ковариации справедливо следующее соотношение:*

$$\operatorname{cov}(\xi, \eta) = E(\xi\eta) - E(\xi)E(\eta).$$

Доказательство. Как и в доказательстве аналогичного свойства для дисперсии, используем линейность матожидания:

$$\operatorname{cov}(\xi, \eta) = E[(\xi - E(\xi))(\eta - E(\eta))] = E[\xi\eta - \xi E(\eta) - \eta E(\xi) + E(\xi)E(\eta)] = E(\xi\eta) - E(\xi)E(\eta) - E(\eta)E(\xi) + E(\xi)E(\eta) = E(\xi\eta) - E(\xi)E(\eta).$$

□

Определение 4.5.2. Линейный коэффициент корреляции случайных величин ξ и η – это

$$r(\xi, \eta) = \frac{\operatorname{cov}(\xi, \eta)}{\sigma(\xi)\sigma(\eta)}.$$

Линейный коэффициент корреляции изменяется от минус одного до одного и показывает, насколько сильна линейная зависимость величин: если $r(\xi, \eta)$ близок к еденице, то между случайными величинами ξ и η прямая линейная пропорциональность (чем больше ξ, тем больше η), а если $r(\xi, \eta)$ порядка минус одного, то связь между ξ и η линейная, но обратная: чем больше ξ, тем меньше η.

Рассмотрим следующую задачу (задачу линейной парной регрессии). Пусть у объекта есть пара признаков (x, y). Пусть так же нам удалось набрать статистику: зависимость одного признака от другого, вид которой изображен на рисунке 1.

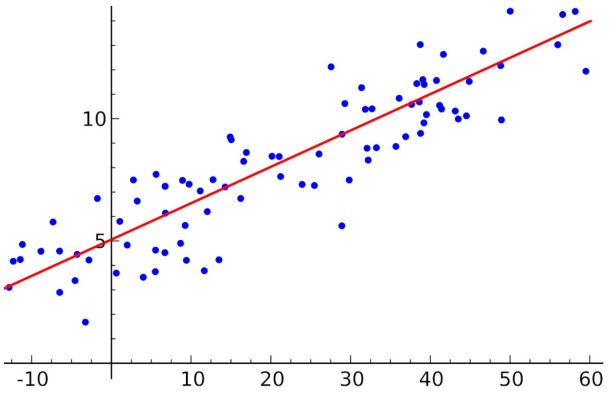

Рис. 1

Из графика видно, что, хотя точки и не ложатся на прямую, все же можно провести прямую, которая будет отражать зависимость $y(x)$. Встает вопрос – как провести эту прямую? Давайте из всех прямых выделим такую, которая лучше всего будет аппроксимировать нашу зависимость. Чтобы это сделать, надо определиться с так называемой функцией потерь, которая будет количественно отражать, насколько хорошо проведена прямая. Первое, что приходит на ум – просто взять сумму модулей расстояний от каждой точки до прямой:

$$S = \sum_{i=1}^{n} |(y(x_i) - \tilde{y}(x_i))|,$$

где $\tilde{y}(x_i) = kx_i + b$ – искомая прямая, а $y(x_i) = kx_i + b + \varepsilon_i$ – то, что нам удалось померить (данные, на основе которых мы будем строить уравнение регрессии). Такая функция ошибок тоже используются, но чаще в качестве штрафа за неправильное предсказание берут сумму квадратов расстояний от каждой точки до прямой. Этот выбор обусловлен теоремой Гаусса-Маркова.

Теорема 4.5.1 (**Гаусса-Маркова**). *Пусть в модели парной регрессии наблюдения y связаны с x зависимостью: $y_i = kx_i + b + \varepsilon_i$, и при этом выполнены следующие условия:*

1. *Модель данных правильно специфицирована*
2. *Все x_i детерминированы и не все равны между собой*
3. *Ошибки не носят систематического характера, то есть $E(\varepsilon_i) = 0 \ \forall\, i$*
4. *Дисперсия всех ошибок одинакова*
5. *Ошибки некоррелированы, то есть $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall\, i, j$*

Тогда в этих условиях оценки метода наименьших квадратов (МНК) оптимальны в классе линейных несмещённых оценок для поиска уравнения регресси $y = kx + b$.

Таким образом, наша функция ошибок имеет вид:

$$S(k, b) = \sum_{i=1}^{n} (y(x_i) - \tilde{y}(x_i))^2 = \sum_{i=1}^{n} (y(x_i) - kx_i - b)^2.$$

Заметим, что функция S не имеет максимума – как бы «плохо» мы не провели прямую (какое бы большое значение не принимала бы S), можно провести ее еще «хуже» (сделать значение S еще больше). А значит, если мы найдем экстремум функции S, то это будет минимум, а это как раз то, что нам нужно – в этом случае сумма растояний от точек до прямой будет минимальна.

Из математического анализа известно, что функция нескольких переменных принимает свое экстремальное значение только в тех точках, где ее градиент равен нулю. Из этого условия получаем систему линейных уравнений:

$$\begin{cases} \dfrac{\partial S}{\partial k} = 0, \\ \dfrac{\partial S}{\partial b} = 0; \end{cases} \Rightarrow \begin{cases} -2 \displaystyle\sum_{i=1}^{n} (y_i - kx_i - b)x_i = 0, \\ -2 \displaystyle\sum_{i=1}^{n} (y_i - kx_i - b) = 0; \end{cases}$$

По свойству линейности суммы, имеем:

$$\begin{cases} k\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i y_i, \\ k\sum\limits_{i=1}^{n} x_i + b = \sum\limits_{i=1}^{n} y_i; \end{cases}$$

Вводя обозначение среднего арифметического $\overline{\alpha} = \dfrac{1}{n}\sum\limits_{i=1}^{n} \alpha_i$ и деля каждое уравнение системы на n, получим:

$$\begin{cases} k \cdot \overline{x^2} + b \cdot \overline{x} = \overline{xy}, \\ k \cdot \overline{x} + b = \overline{y}; \end{cases}$$

Выражая из второго уравнения b и подставляя его в первое, находим k:

$$k = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}.$$

Заметим, что если $E(x) = \overline{x}$, то можно записать: $k = \dfrac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \dfrac{\operatorname{cov}(x, y)}{D(x)}$.

Зная k, находим b:

$$b = \overline{y} - k \cdot \overline{x} = \overline{y} - \overline{x} \cdot \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}.$$

Искомое уравнение парной регресси имеет вид:

$$\tilde{y}(x) = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \cdot x + \overline{y} - \overline{x} \cdot \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}.$$

Пример 4.5.1. Рост-вес

4.6 Множественная линейная регрессия

На практике редко используется парная регрессия, описанная в предыдущем параграфе – обычно один из параметров, который мы хотим предсказать (он называется целевой переменной) зависит не от одного фактора, а от многих. Давайте попробуем обобщить решение предыдущей задачи на этот случай.

Пусть у нас есть целевая переменная y, которую мы хотим научиться предсказывать, и k параметров $(x_1, x_2, \dots x_k)$, от которых y зависит линейно: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$. Пусть так же нам удалось собрать n штук наблюдений – как ведет себя y в зависимости от параметров. Аналогично задаче парной регресси мы хотим предсказать $\tilde{y}$ так, чтобы сумма квадратов расстояний от y до $\tilde{y}$ была минимальной:

$$S(\beta_0, \beta_1, \dots \beta_k) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik})^2 \to \min.$$

Введем обозначения:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \; Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; W = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Тогда задача минимизации функции S превращается в задачу нахождения минимума квадрата длины вектора $Y - XW$:

$$S = |Y - XW|^2 = (Y - XW)^T(Y - XW) \to \min.$$

Посчитаем градиент функции S:

$$\operatorname{grad} S = \begin{pmatrix} -2\sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) \\ -2\sum\limits_{i=1}^{n} x_{i1}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) \\ \vdots \\ -2\sum\limits_{i=1}^{n} x_{ik}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) \end{pmatrix} = -2X^T(Y - XW).$$

Приравнивая градиент к нулю, получим аналитическое решение:

$$-2X^T(Y - XW) = 0,$$

$$X^T XW = X^T Y,$$

$$W = (X^T X)^{-1} X^T Y.$$

Пример 4.6.1. ?????????????

4.7 Метод максимального правдоподобия

5 Основные методы машинного обучения

5.1 Градиентный спуск

Рассмотрим еще раз задачу линейной регресси.

6 Список литературы

1. С. М. Никольский – Курс математического анализа
2. Д. В. Беклемишев – Курс аналитической геометрии и линейной алгебры
3. Д. В. Беклемишев – Дополнительные главы линейной алгебры
4. А. Н. Ширяев – Вероятность
5. А.И. Кибзун, Е.Р. Горяинова, А.В. Наумов, А.Н. Сиротин – Теория вероятностей и математическая статистика
6. С. Николенко, А. Кадурин, Е. Архангельская – Глубокое обучение. Погружение в мир нейронных сетей
7. Джоэл Грас – Data Science. Наука о данных с нуля