Основная математика для ML

Шмидт Ян
Мой телеграм
Лендинг
RoadMap

18 мая 2025 г.

Оглавление

Дисклеймер

Обозначения. Красным текстом обозначены штуки, которые можно почитать по желанию, они не обязательны.

Глава 1

Алгебра

1.1 Нотация суммирования и произведения

$$\begin{aligned} a_1 + a_2 + \cdots + a_n &= \sum_{i=1}^{n} a_i \\ a_1 \cdot a_2 \cdot \ \ldots \ \cdot a_n &= \prod_{i=1}^{n} a_i \end{aligned} \tag{1.1}$$

1.2 Функции

Функцией f называется такое правило, по которому мы задаем соответствие рассматриваемому множеству X:

$$\begin{aligned} f(X) &= Y \\ X &\mapsto Y \end{aligned} \tag{1.2}$$

X называют *прообразом*, а Y — *образом*. Например, функция $f(x) = x^2$ каждому числу x сопоставляет x^2:

$$x \mapsto x^2 \tag{1.3}$$

Числу 2 будет соответствовать $2^2 = 4$, то есть

$$2 \mapsto 4, \tag{1.4}$$

пятерке — 25: $5^2 = 25$, то есть

$$5 \mapsto 25, \tag{1.5}$$

5 — прообраз, 25 — образ для функции $f(x) = x^2$, и так далее.

1.3 Линейная комбинация

Линейная комбинация — выражение, построенное на множестве элементов путём умножения каждого элемента на коэффициенты с последующим сложением ре-

зультатов (например, линейной комбинацией x и y будет выражение вида

$$ax + by, \tag{1.6}$$

где a и b — коэффициенты).

Если $v_1, \ldots, v_n$ — векторы, а $a_1, \ldots, a_n$ — скаляры, то линейная комбинация этих векторов со скалярами в качестве коэффициентов — это:

$$a_1v_1 + a_2v_2 + a_3v_3 + \cdots + a_nv_n = \sum_{i=1}^{n} a_iv_i \tag{1.7}$$

1.4 Линейная зависимость и независимость

При **линейной зависимости** существует нетривиальная линейная комбинация элементов этого множества, равная нулевому элементу: для векторов $v_1, \ldots, v_n$ найдутся такие числа $a_1, \ldots, a_n$, не равные одновременно нулю, что

$$\sum_{i=1}^{n} a_iv_i = 0 \tag{1.8}$$

При отсутствии такой комбинации, то есть, когда коэффициенты линейной комбинации равны нулю, векторы называются **линейно независимым**. Иными словами, если

$$\sum_{i=1}^{n} a_iv_i = 0 \tag{1.9}$$

выполняется только при $a_i = 0,\ i = \overline{1, n}$.

1.5 Скалярное произведение векторов

Скалярное произведение – полезная штука в МЛ, потому что может встречаться, например, в определении косинусного расстояния.

$$\begin{aligned} &\langle x, y\rangle = |x||y|\cos(\widehat{x, y}) = \sum_{i=1}^{n} x_iy_i, \\ &x, y \in \mathbb{R}^n, \end{aligned} \tag{1.10}$$

$\widehat{x, y}$ – угол между векторами x, y.

1.6 Матрицы

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix} \tag{1.11}$$

Матрицы можно складывать-вычитать, умножать, обращать (= делить), транспонировать. Пример транспонирования:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{pmatrix} \tag{1.12}$$

1.6.1 Ранг матрицы

Строки и столбцы матрицы являются элементами соответствующих векторных пространств:

- Столбцы матрицы A составляют элементы пространства размерности m;
- Строки матрицы A составляют элементы пространства размерности n.

Рангом матрицы называют количество линейно независимых столбцов матрицы (столбцовый ранг матрицы) или количество линейно независимых строк матрицы (строчный ранг матрицы). Этому определению эквивалентно определение ранга матрицы как порядка максимального отличного от нуля минора матрицы.

При элементарных преобразованиях ранг матрицы не меняется.

1.6.2 Элементарные преобразования над матрицами

Элементарными преобразованиями строк называют:

- Перестановку местами любых двух строк матрицы;
- Умножение любой строки матрицы на обратимую ненулевую константу k;
- Прибавление к любой строке матрицы другой строки, умноженной на некоторую константу.

Аналогично определяются элементарные преобразования столбцов. Элементарные преобразования обратимы.

Обозначение: $A \sim B$ указывает на то, что матрица A может быть получена из B путём элементарных преобразований (или наоборот).

1.6.3 Виды матриц

Диагональные, квадратные, прямоугольные, треугольные, симметричные, ортогональные, блочные и многие другие. Список поможно посмотреть на англоязычной странице в вики.

1.6.4 Решение систем линейных уравнений (СЛУ) через матрицы

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \tag{1.13}$$

Решение:

$$Ax = b, \tag{1.14}$$

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}_{n \times n} \tag{1.15}$$

— матрица коэффициентов,

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}_{n \times 1} \tag{1.16}$$

— вектор свободных коэффициентов,

$$x = A^{-1}b \tag{1.17}$$

— решение системы через обратную матрицу

1.6.5 Определитель матрицы

$$A = \begin{vmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{vmatrix} = \det(A) \tag{1.18}$$

Определитель матрицы полезен тем, что от его значения зависит, с какой матрицей мы имеем дело. Если $\det A = 0$ (такая матрица называется **вырожденной**), то результат применения данной матрицы в качестве линейного оператора приведет к вырождению рассматримаемого пространства. Так, если применить вырожденную матрицу 2×2 к плоскости $\mathbb{R}^2$, то она выродится в одну-единственную прямую.

Интерпретация. Значение определителя показывает, по сколько раз растягивается (масштабируется) преобразуемое пространство. $\det A = 3$ будет означать, что все единичные кубы в $\mathbb{R}^n$ будут растянуты так, что их объем увеличится в 3 раза. Если $\det A = k < 0$, то пространство масштабируется в k раз, но поменяет свою ориентацию. Например, если рассмотреть снова пример с плоскостью, то при отрицательном определителе плоскость растянется или сожмется (если $|\det A| < 1$) и при этом зеркально отразится: оси x и y поменяются местами. В $\mathbb{R}^3$ чуть сложнее: в таком случае перестанет работать правило буравчика: пары осей поменяют свое положение.

1.7 Собственные числа, собственные векторы

Формально, пусть A, v – матрица линейного оператора (отображения) и произвольный ветктор соответственно. Если после применения матрицы к вектору (отображая как-то линейно вектор), выходит, что он всего лишь масштабировался на множитель λ, то говорят, что v – собственный вектор, а λ — собственное число.

$$Av = \lambda v \tag{1.19}$$

Как вычислять? Вычислять с. ч. можно так:

$$\begin{aligned} \det(A - \lambda E) = \begin{vmatrix} a_{11} - \lambda & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} - \lambda \end{vmatrix} = \\ = b_1\lambda^n + b_2\lambda^{n-1} + \dots b_{n-1}\lambda + b_n = 0, \end{aligned} \tag{1.20}$$

$b_1\lambda^n + b_2\lambda^{n-1} + \dots b_{n-1}\lambda + b_n = 0$ — характеристический полином, λ_i — собственные числа, это корни характеристиеского полинома, E — единичная диагональная матрица

$$E_{n\times n} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}, \tag{1.21}$$

С.ч. и векторы можно вычислять численными методами, например, степенным методом.

Глава 2

Математический анализ I

2.1 Кванторы

Квантор всеобщности — $\forall$, квантор существования — $\exists$; отрицание будет такое: $\nexists$. $\exists!$ – существует и единственно.

2.2 Интуитивное понятие предела $\lim$

Если коротко, то $\lim$ – это предельное состояние чего-то, как следует из названия. Можно рассматривать числовой предел последовательности точек, предел последовательности функций, предел значений функции и так далее.

Пример. Функция $f(x) = x^2$ своим пределом будет иметь бесконечность:

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} x^2 = \infty \tag{2.1}$$

Пример. Предел последовательности $\left\{\frac{1}{n^2}\right\}$ при $n \to \infty$ равен 0: $1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}, \frac{1}{25}, \ldots$ — в пределе получим нереально мелкое число, сколько угодно близкое к 0.

2.3 Производные функций (определение, свойства)

$$f'(x) = \lim_{\Delta\to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}, \tag{2.2}$$

Δx — маленькое приращение x, то есть $x + \Delta x$ и x очень мало отличаются друг от друга (на величину Δ).

2.4 Экстремум функции

Экстремум — максимальное или минимальное значение функции на заданном множестве. Точка, в которой достигается экстремум, называется *точкой экстремума*. Соответственно, если достигается минимум — точка экстремума называется *точкой минимума*, а если максимум — *точкой максимума*. В математическом анализе выделяют также понятие *локальный экстремум* (соответственно минимум или максимум).

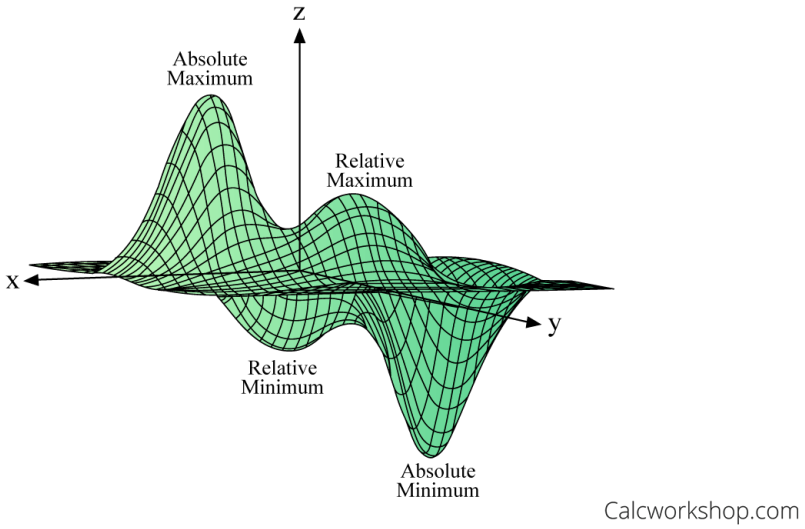


Иллюстрация экстремума функции двух переменных

2.5 Частные производные

В математическом анализе **частная производная** — одно из обобщений понятия производной на случай функции нескольких переменных. **Частная производная** — это предел отношения приращения функции **по выбранной переменной** к приращению этой переменной, при стремлении этого приращения к нулю:

$$\frac{\partial f(a_1, \cdots, a_n)}{\partial x_k} = \lim_{\Delta x_k \to 0} \frac{f(a_1, \ldots, a_k + \Delta x_k, \ldots, a_n) - f(a_1, \ldots, a_k, \ldots, a_n)}{\Delta x_k} \tag{2.3}$$

Пример. Если $f(x, y, z) = (x^5, y^3, z)$, то

$$\begin{cases} \dfrac{\partial f(x,y,z)}{\partial x} = (5x^4, y^3, z) \\ \dfrac{\partial f(x,y,z)}{\partial y} = (x^5, 3y^2, z) \\ \dfrac{\partial f(x,y,z)}{\partial z} = (x^5, y^3, 1) \\ \dfrac{\partial^2 f(x,y,z)}{\partial x^2} = (20x^3, y^3, z) \\ \dfrac{\partial f(x,y,z)}{\partial y \partial x} = (5x^4, 3y^2, z) \end{cases} \tag{2.4}$$

2.6 Градиент

Градиент функции $f(\mathbf{x}) = f(x_1, \ldots, x_n),\ x \in \mathbb{R}^n$ – вектор *первых* частных производных:

$$\operatorname{grad} f(\mathbf{x}) = \nabla f(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(\mathbf{x})}{\partial x_{n-1}} \\ \dfrac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix}_{n \times 1}. \tag{2.5}$$

Показывает направление наискорейшего роста функции.

2.7 Лапласиан

Вектор *вторых* частных производных:

$$\Delta f(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1^2} \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1^2} \\ \vdots \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_{n-1}^2} \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{pmatrix}_{n \times 1}. \tag{2.6}$$

2.8 Матрица Якоби

Матрица Якоби – матрица *первых* частных производных для вектор-функции $\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_{n-1}(\mathbf{x}) \\ f_n(\mathbf{x}) \end{pmatrix}_{n\times 1}$, $\mathbf{x} \in \mathbb{R}^n$:

$$\mathbf{J_f} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}_{m\times n} \tag{2.7}$$

Определитель матрицы Якоби – **якобиан** – может возникать, например, при переходе к новым координатам (от декартовых – к сферическим и т. д.).

2.9 Гессиан

Гессиан, или матрица Гессе – матрица *вторых* частных производных для функции $f(\mathbf{x})$ (не векторной, как для случая с Якобианом!):

$$\mathbf{H}_f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}. \tag{2.8}$$

Используются в методах второго порядка при численных методах оптимизации функций. Считать и хранить такую матрицу очень затратно при высоких размерностях, поэтому существуют разные приближенные методы (квази-ньютоновские).

2.10 Метрика (расстояние)

Функция $\rho(x, y)$ – *метрика* и определяет расстояние между точками x, y в $\mathbb{R}^n$, если

- $\rho(x, y) = 0 \Leftrightarrow x = y$ – тождество
- $\rho(x, y) \geq 0$ – неотрицательность
- $\rho(x, y) = \rho(y, x)$ – симметричность

- $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ – неравенство треугольника

Пусть $x, y \in \mathbb{R}^n$ – точки в n-мерном пространстве (например, евклидовом).

Часто используются:

- **Евклидова метрика**:

$$\rho(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.9}$$

- **Манхеттенская метрика (метрика городских кварталов)**:

$$\rho(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.10}$$

- **Расстояние Левенштейна** [1]: метрика сходства между двумя строковыми последовательностями. Чем больше расстояние, тем более различны строки. Для двух одинаковых последовательностей расстояние равно нулю. По сути, это минимальное число односимвольных преобразований (удаления, вставки или замены), необходимых, чтобы превратить одну последовательность в другую.

  Например, если сравнить слова «лупа», «пупа», то они отличаются в 1 позиции, поэтому расстояние Левенштейна = 1 (нужно сделать 1 замену, чтобы слова стали равными).

- **Расстояние Вассерштейна**: определяет расстояние между вероятностными распределениями. Чем-то похоже на KL-дивергенцию и дивергенцию Йенсена-Шеннона, но имеет некоторые теоретические преимущества. Почитать можно в вики и тут, тут.

2.11 Норма

Норма вектора задает его «длину» в рассматриваемом пространстве. Обозначается через $\|\cdot\|$. Сначала общий случай: пусть $\mathbf{x} \in \mathbb{R}^n, x = (x_1, \ldots, x_n)$. Тогда L_p-норма этого вектора:

$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} \tag{2.11}$$

При $p = 1$ получаем **манхеттенскую норму** (L_1-норму):

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \tag{2.12}$$

при $p = 2$ получаем **евклидову норму** (L_2-норму):

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}, \tag{2.13}$$

[1] Можно почитать на хабре

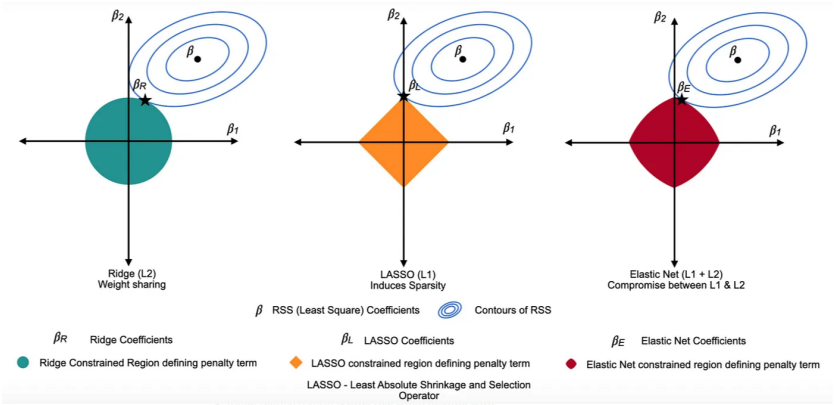
Пример. L_1-норма используется в **регуляризаторе LASSO**:

$$L = \underbrace{\mathcal{L}}_{\text{loss}} + \underbrace{\alpha\|w\|_1}_{\text{penalty}} = \mathcal{L} + \alpha \sum_w |w|, \tag{2.14}$$

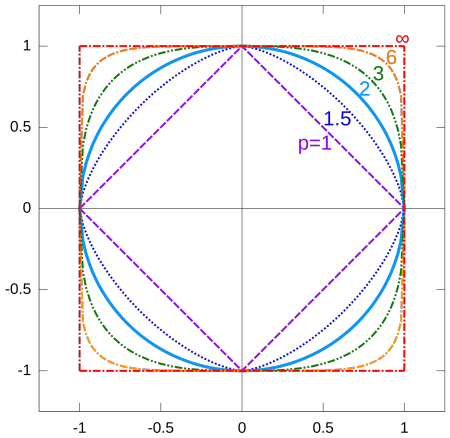
а L_2-норма – в **Ridge регуляризаторе**:

$$L = \underbrace{\mathcal{L}}_{\text{loss}} + \underbrace{\alpha\|w\|_2}_{\text{penalty}} = \mathcal{L} + \alpha \sum_w w^2, \tag{2.15}$$

а α называется штрафным коэффициентом, weigh decay или **коэффициентом регуляризации**.



Регуляризаторы LASSO (L_1), Ridge (L_2) и их комбинация – ElasticNet



Разные нормы, иллюстрация единичных окружностей

Глава 3

Теория вероятностей и математическая статистика

3.1 Основные понятия

Случайная величина — это результат случайного эксперимента. Обозначают часто как ξ (кси). Например, случайная величина — подбрасывание монетки. **Вероятность** изменяется на промежутке $p \in [0, 1]$, где $p = 0$ — событие точно не произойдет, $p = 1$ — сто процентов произойдет. Вероятность «50/50» — $p = 0.5$.

3.2 Типы случайных величин: дискретные и непрерывные

- **Дискретные**: подбрасывание монеты, количество звонков в день, ...
- **Непрерывные**: те, которые задаются функцией распределения и соответствующей функцией плотности распределения

3.3 Плотность распределения случайной величины

Плотность распределения — любая функция, которая удовлетворяет специальным условиям: это нормированная функция, то есть если проинтегрировать ее от $-\infty$ до $+\infty$ (то есть по всему вещественному промежутку в одномерном $\mathbb{R}^1$ случае), то получим единицу:

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \tag{3.1}$$

3.4 Некоторые распределения

3.4.1 Нормальное (гауссовское)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.2}$$

– плотность распределения, причем параметр μ — **математическое ожидание** (среднее значение), медиана и мода распределения, а параметр σ — **среднеквадратическое отклонение**, σ^2 — **дисперсия** распределения.

Пример. Если $\mu = 0, \sigma = 1$, то это **стандартное нормальное распределение**. Рассмотрим чуть подробнее его плотность:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3.3}$$

Поскольку

$$\int_{-\infty}^{+\infty} e^{-x^2/2} dx = \sqrt{2\pi}, \tag{3.4}$$

то

$$\int_{-\infty}^{+\infty} f(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} = 1 \tag{3.5}$$

Вот почему плотность нормального распредления выглядит именно так, откуда там взялся коэффициент с π. И это еще раз проливает свет на определение плотности вероятностного распределения.

3.4.2 Экспоненциальное

Случайная величина X имеет экспоненциальное распределение с параметром $\lambda > 0$, если её плотность вероятности имеет вид:

$$f_X(x) = \begin{cases} \lambda\, e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{3.6}$$

3.4.3 Бернулли

Случайная величина X имеет распределение Бернулли, если она принимает всего два значения: 1 и 0 с вероятностями p и $q \equiv 1 - p$ соответственно. Таким образом: $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = q$. Принято говорить, что событие $\{X = 1\}$ соответствует «успеху», а событие $\{X = 0\}$ — «неудаче». Эти названия условные, и в зависимости от конкретной задачи могут быть заменены на противоположные.

3.4.4 Пуассона

Выберем фиксированное число $\lambda > 0$ и определим дискретное распределение, задаваемое следующей функцией вероятности:

$$p(k) \equiv \mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{3.7}$$

3.5 Математическое ожидание

Идейно, **мат ожидание случайной величины** — это что ты с точки зрения математики ожидаешь получить, случайно сэмплируя величины из распределения при стремлении количества сэмплирований к бесконечности. Обозначают символом $E, \mathbb{E}$.

Например, бросая монетку бексонечное число раз, в пределе получим, что орел/решка выпадают в среднем одинаковое кол-во раз. То есть мат ожидание такой случайной величины (она, кстати, Бернуллевская, дискретная) равно 0.5.

Для непрерывных распределений мат ожидание — это интеграл

$$\mathbb{E}\xi = \int_{-\infty}^{+\infty} x f(x) dx, \tag{3.8}$$

для дискретных — сумма

$$\mathbb{E}\xi = \sum_{i=1}^{n} a_i p_i, \tag{3.9}$$

a_i, p_i — значение случайной величины и ее вероятность соответственно, n — количество экспериментов.

3.6 Дисперсия

Дисперсия — это мера разброса данных. Чем больше дисперсия, тем больше разброс значений, и наоборот. Для расчета дисперсии в математике используется формула, которая учитывает разницу каждого значения от среднего значения и возводит эту разницу в квадрат:

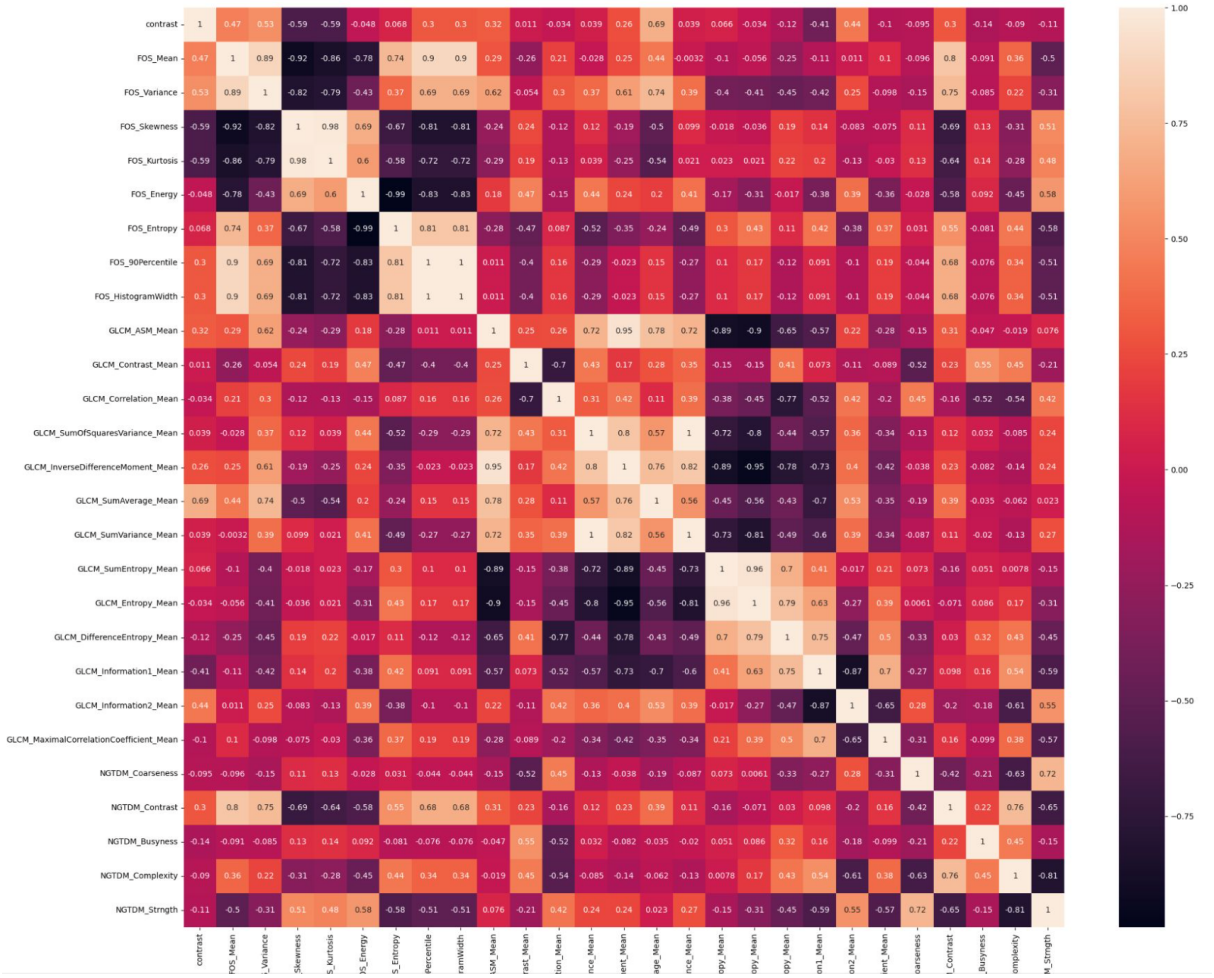
$$\mathbb{D}\xi = \mathbb{E}(\xi - \mathbb{E}\xi)^2 = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2 = \int_{-\infty}^{+\infty} x^2 f(x) dx - \left(\int_{-\infty}^{+\infty} x f(x) dx \right)^2 \tag{3.10}$$

В дискретном случае:

$$\mathbb{D}\xi = \mathbb{E}(\xi - \mathbb{E}\xi)^2 = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2 = \sum_{i=1}^{n} a_i^2 p_i - \left(\sum_{i=1}^{n} a_i p_i \right)^2 \tag{3.11}$$

3.7 Корреляция

Бывает разной: линейной, ранговой, нелинейной. Корреляции Пирсона (линейная), Спирмена (ранговая), Кенделла (ранговая) и другие (φ_k — нелинейная). По-разному показывают, как переменные зависят друг от друга.



Пример матрицы корреляций Пирсона между разными признаками

3.8 Выборка, генеральная совокупность

Генеральная совокупность — это вообще **все возможные** существующие на свете объекты, которые нам интересно рассматривать. Например, можно рассматривать множество всех людей на планете как генеральную совокупность людей. Тогда выборкой будет называться (определение по самому слову) некоторое подмножество, в которое входят отобранные (например, случайно) люди. Можно выбрать конкретных людей. Выборка светловолосых и темноволосых людей, выборка очень низких и очень высоких, выборка из всех петербуржцев и так далее.

Обычно выборку обозначают за X.

Дополнительно Ковариации, ковариационная матрица, многомерные случайные величины, моменты, центральная предельная теорема, γ- и β-распределения, статистические тесты, алгебра множеств, счетная аддитивность, σ-алгебры, аксиоматика Колмогорова, расстояния между распределениями

Глава 4

Численные методы. Методы оптимизации

4.1 Что такое минимизация функционала

Пусть целевая функция имеет вид:

$$F(\vec{x}) : \mathbb{X} \to \mathbb{R} \tag{4.1}$$

И задача оптимизации задана следующим образом

$$F(\vec{x}) \to \min_{\vec{x} \in \mathbb{X}}, \tag{4.2}$$

то есть надо минимизировать $F(x)$ на множестве всех $x \in \mathbb{X}$.

$$\min_{\vec{x} \in \mathbb{X}} F(\vec{x}) \tag{4.3}$$

— функционал качества / задача оптимизации / **минимизация функционала** / минимизация целевой функции и проч.

С максимизацией все так же с точностью до знака.

4.2 Градиент. Градиентный спуск

Градиент — направление наискорейшего роста функции. **Антиградиент** — градиент с минусом, то есть направление наискорейшего уменьшения функции

Основная идея метода заключается в том, чтобы идти в направлении наискорейшего спуска, а это направление задаётся антиградиентом $-\nabla F$:

$$\vec{x}_{n+1} = \vec{x}_n - \lambda \nabla F\left(\vec{x}_n\right) \tag{4.4}$$

За счет этого из заданной точки мы движемся в направлении наискорейшего убывания функции, таким образом попадем в конечном итоге в желаемый минимум.

Проблема в том, что нужно правильно подобрать λ — **градиентный шаг**. Если сделать его слишком большим, то шаги будут чересчур большими и можем перепрыгнуть через точку минимума; если сделать слишком маленьким, то будет очень долго сходиться алгоритм. Обычно берут что-то вроде $\lambda = 10^{-3}$, но это тоже гиперпараметр, поэтому только практикой можно понять, какой λ подходит лучше всего.

Глава 5

Теоретическая информатика

5.1 Энтропия Шеннона

Информационная энтропия — мера неопределённости некоторой системы. Например, в последовательности букв, составляющих какое-либо предложение на русском языке, разные буквы появляются с разной частотностью, поэтому неопределённость появления для некоторых букв меньше, чем для других.

Информационная двоичная энтропия, при отсутствии информационных потерь, рассчитывается по **формуле Хартли**:

$$I = \log_2 N, \tag{5.1}$$

где N — мощность алфавита, I — количество информации в каждом символе сообщения. Для случайной величины x, принимающей n независимых случайных значений x_i с вероятностями p_i $(i = 1, ...n)$, формула Хартли переходит в **формулу Шеннона**:

$$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{5.2}$$

Здесь $-\log_2 p_i$ означает измеряемое **в битах** (отсюда $-\log_{\mathbf{2}} p_i$) количество информации, содержащейся в том событии, что случайная величина приняла значение x_i, $H(x)$ — количество информации, которое в среднем приходится на одно событие (для предложений на русском языке — количество информации в среднем на одну букву).

Если измерять информацию не **в битах**, то основание логарифма поменяется на любое другое.

Дополнительно Условная энтропия, перекрестная (совместная) энтропия, пропускная способность канала, основы кодирования, дерево Хаффмана, расстояние Левенштейна, дифференциальная энтропия