

## Содержание

<b>I</b>	<b>Модуль 1</b>	<b>5</b>
<b>1</b>	<b>Лекция 1</b>	<b>5</b>
1.1	Как устроен курс . . . . .	5
1.2	Практика . . . . .	5
1.3	SymPy . . . . .	5
1.4	Рисуем графики . . . . .	6
1.4.1	Подключаем библиотеку . . . . .	6
1.4.2	Рисуем график функции $f(x)$ . . . . .	6
1.4.3	Изменяем размер картинки и граничные значения по осям . . . . .	6
1.4.4	Добавим названия осей и подпись к графику . . . . .	6
1.4.5	Добавим оси и подпишем под картинкой информацию об экстремумах и точках перегиба . . . . .	6
1.5	Метод градиентного спуска . . . . .	7
1.5.1	Теорема о градиенте . . . . .	7
1.5.2	Применение в машинном обучении . . . . .	7
1.5.3	Идея применения градиентного спуска . . . . .	7
1.5.4	Метод градиентного спуска на пальцах . . . . .	7
1.5.5	Метод градиентного спуска (одномерный случай) . . . . .	7
1.5.6	Метод градиентного спуска (общий случай) . . . . .	7
1.5.7	Параметр learning rate . . . . .	8
1.5.8	Теорема о поиске в выпуклой гладкой функции . . . . .	8
1.6	Реализация градиентного спуска на python . . . . .	8
<b>2</b>	<b>Лекция 2</b>	<b>9</b>
2.1	Линейная регрессия . . . . .	9
2.2	Функция ошибок MSE . . . . .	9
2.3	Матричное представление MSE . . . . .	9
2.4	Как решать задачу минимизации MSE . . . . .	9
2.4.1	Аналитическое решение . . . . .	9
2.4.2	Приближённое решение . . . . .	9
2.5	Производная по вектору . . . . .	9
2.5.1	Пример 1. Подсчёт градиента скалярного произведения . . . . .	9
2.5.2	Пример 2. Подсчёт градиента от матрицы . . . . .	10
2.6	Минимизация MSE с помощью градиента . . . . .	10
2.7	Градиент функции потерь . . . . .	10
2.8	Реализация на питоне . . . . .	10
2.8.1	Генерирование данных для задачи регрессии . . . . .	10
2.8.2	Функция подсчёта ошибки . . . . .	10
2.8.3	Реализация градиентного спуска . . . . .	11
2.8.4	Функция предсказания модели . . . . .	11
2.8.5	Применение градиентного спуска . . . . .	11
<b>3</b>	<b>Лекция 3</b>	<b>12</b>
3.0.1	Один из недостатков градиентного спуска . . . . .	12
3.1	Стохастический градиентный спуск . . . . .	12
3.2	Mini-Batch Gradient Descent . . . . .	12
3.3	Ещё один недостаток градиентного спуска . . . . .	12
3.4	Метод моментов (Momentum) . . . . .	12
3.5	Преимущества метода моментов . . . . .	12
3.6	Градиентный шаг . . . . .	13
3.7	AdaGrad (Adaptive Gradient) . . . . .	13
3.8	RMSPROP (Root Mean Square Propagation) . . . . .	13

3.9	Практикум на питоне . . . . .	13
3.9.1	Генерируем данные . . . . .	13
3.9.2	Библиотечное решение . . . . .	13
<b>4</b>	<b>Лекция 5. Линейная алгебра</b>	<b>15</b>
4.0.1	Упражнение . . . . .	15
4.0.2	Ответы на упражнение . . . . .	15
4.1	Выражение базиса через другой базис . . . . .	15
4.2	Программирование . . . . .	16
4.2.1	Нетривиальные вещи . . . . .	16
4.2.2	Отрисовка векторов . . . . .	16
<b>5</b>	<b>Лекция 6. Метод главных компонент (PCA)</b>	<b>17</b>
5.1	Постановка задачи выражения новых признаков . . . . .	17
5.1.1	Интерпретация линейного выражения признаков . . . . .	17
5.1.2	Пояснение: Проекция . . . . .	18
5.1.3	Сохранение информации . . . . .	18
5.2	Постановка задачи . . . . .	18
5.2.1	Ковариация . . . . .	19
5.3	Алгоритм поиска новых компонент . . . . .	19
5.3.1	Первый шаг . . . . .	19
5.4	Применение метода . . . . .	19
5.5	Доля объяснённой дисперсии . . . . .	19
5.6	Выбор числа главных компонент . . . . .	20
5.7	Упражнения . . . . .	20
5.7.1	Нахождение главных компонент . . . . .	20
5.7.2	То же самое в питоне . . . . .	20
5.7.2.1	Генерация данных . . . . .	20
5.7.2.2	Реализация метода главных компонент . . . . .	20
5.7.2.3	Вычисляем матрицу ковариаций . . . . .	20
5.7.2.4	Находим собственные векторы и проецируем все точки . . . . .	20
5.7.2.5	Восстановление данных . . . . .	21
5.7.2.6	Доля объяснённой дисперсии . . . . .	21
5.7.2.7	Что выдаёт PCA из sklearn . . . . .	21
5.8	Сингулярное разложение (SVD) . . . . .	21
5.9	Что лучше: PCA или SVD? . . . . .	21
5.10	Упражнение . . . . .	21
5.10.1	Находим ответы . . . . .	22
5.11	Связь SVD и PCA . . . . .	22
5.12	Сингулярное разложение в питоне . . . . .	22
<b>II</b>	<b>Модуль 2</b>	<b>23</b>
<b>6</b>	<b>Лекция 1. Множества</b>	<b>23</b>
6.1	Особые множества . . . . .	23
6.2	Упражнение . . . . .	23
6.3	Упражнение . . . . .	23
6.3.1	Определение равномошности . . . . .	23
6.4	Упражнение . . . . .	24
6.5	Ответы на упражнения . . . . .	24
6.6	Все ли бесконечные множества равномошны? . . . . .	24
6.7	Континуальность . . . . .	24
6.8	Теорема о удалении бесконечного множества . . . . .	24
6.9	Упражнение . . . . .	24
6.10	Упражнение . . . . .	24
6.11	Упражнение . . . . .	25
6.12	Упражнение . . . . .	25

<b>7</b>	<b>Лекция 2. Сигма-алгебра</b>	<b>26</b>
7.1	Пространство элементарных исходов . . . . .	26
7.2	Упражнение (классическая вероятность) . . . . .	26
7.3	Важная мысль . . . . .	26
7.4	Упражнение (геометрическая вероятность) . . . . .	26
7.5	Сигма-алгебра . . . . .	27
7.6	Упражнение . . . . .	27
7.7	Упражнение . . . . .	27
7.8	Зачем нужна $\sigma$ -алгебра . . . . .	27
7.9	Определение вероятности . . . . .	27
7.10	Упражнение . . . . .	28
7.11	Подытог . . . . .	28
7.12	Свойства вероятности . . . . .	28
7.12.1	Упражнение . . . . .	28
7.12.2	Упражнение . . . . .	28
7.13	Условная вероятность . . . . .	28
7.14	Формула Байеса . . . . .	28
7.15	Формула полной вероятности . . . . .	28
7.15.1	Упражнение . . . . .	28
7.16	Априорное и апостериорное распределение . . . . .	29
7.17	Байесовская вероятность в виде графа . . . . .	29
7.17.1	Аналитическая задача на формулу Байеса . . . . .	29
<b>8</b>	<b>Факультативная часть лекции 2</b>	<b>29</b>
8.1	Случайные величины и борелевская $\sigma$ -алгебра . . . . .	29
8.1.1	Пример . . . . .	29
8.2	Подготовка к Борелевской $\sigma$ -алгебре . . . . .	29
8.2.1	Упражнение . . . . .	30
8.3	Борелевская $\sigma$ -алгебра . . . . .	30
8.3.1	Открытое множество . . . . .	30
8.3.2	Внутренние точки . . . . .	30
8.3.3	Замкнутое множество . . . . .	30
8.4	Борелевское множество . . . . .	30
8.4.1	Упражнение . . . . .	30
8.5	Измеримость . . . . .	30
<b>9</b>	<b>Лекция 3. Комбинаторика</b>	<b>31</b>
9.1	Правило умножения и сложения . . . . .	31
9.1.1	Упражнение . . . . .	31
9.1.2	Сколько существует пятизначных чисел? . . . . .	31
9.1.3	Упражнение . . . . .	31
9.2	Принцип Дирихле . . . . .	31
9.2.1	Упражнение . . . . .	31
9.2.2	Упражнение . . . . .	31
9.3	Размещения . . . . .	32
9.3.1	Упражнение . . . . .	32
9.3.2	Упражнение . . . . .	32
9.4	Перестановки . . . . .	32
9.4.1	Упражнение . . . . .	32
9.4.2	Упражнение . . . . .	32
9.5	Сочетания . . . . .	32
9.5.1	Упражнение . . . . .	32
9.5.2	Упражнение . . . . .	33
9.5.3	Упражнение . . . . .	33
9.5.4	Упражнение . . . . .	33

<b>10 Лекция 4. Разлагай и властвуй, бином, беспорядки</b>	<b>34</b>
10.1 Дискретные случайные величины . . . . .	34
10.1.1 Упражнение . . . . .	34
10.2 Манифест . . . . .	34
10.3 Операции над случайными величинами . . . . .	34
10.3.1 Умножение случайных величин . . . . .	34
10.3.2 Сложение случайных величин . . . . .	35
10.3.3 Совместное распределение случайных величин . . . . .	35
10.4 Бином Ньютона . . . . .	35
10.5 Распределение Бернулли . . . . .	35
10.5.1 Упражнение . . . . .	35
10.5.2 Дисперсия . . . . .	35
10.6 Биномиальное распределение . . . . .	36
10.6.1 Ковариация (Cov) . . . . .	36
10.6.2 Корреляция (Corr) . . . . .	36
10.7 Формула флуктуаций исключений . . . . .	36
10.8 Беспорядки (субфакториал) ( $!n$ или $D_n$ ) . . . . .	36
10.8.1 Свойства беспорядков . . . . .	36
10.8.2 Доказать . . . . .	36
<b>11 Цепи Маркова и метод первого шага</b>	<b>37</b>
11.1 Упражнение . . . . .	37
11.2 Упражнение (Амёба) . . . . .	37
11.3 Геометрическое распределение . . . . .	37
11.3.1 Упражнение . . . . .	37
11.3.2 Коллекционер . . . . .	38
11.4 Марковская цепь . . . . .	38
11.4.1 Теорема об Эргодичности конечной цепи Маркова . . . . .	38
11.4.2 Упражнение . . . . .	38
<b>12 Непрерывные случайные величины, плотность распределения</b>	<b>40</b>
12.1 Функция распределения . . . . .	40
12.1.1 Свойства функции распределения . . . . .	40
12.2 Непрерывная случайная величина. Плотность . . . . .	40
12.2.1 Свойства плотности . . . . .	40
12.3 Упражнение . . . . .	41
12.4 Преобразования случайных величин . . . . .	42
12.4.1 Упражнение . . . . .	42
12.4.2 Упражнение на преобразование экспоненциального распределения . . . . .	42
12.4.3 Упражнение на преобразование равномерного распределения . . . . .	43
12.5 Теорема. Квантильное преобразование . . . . .	43
12.6 Формула для суммы случайных величин (формула свёртки) . . . . .	43
12.6.1 Упражнение . . . . .	44
12.6.2 Интеграл Гаусса-Пуассона . . . . .	44
12.6.3 Упражнение . . . . .	44
<b>13 Многомерные и условные распределения</b>	<b>45</b>

# Часть I

## Модуль 1

### 1. Лекция 1

#### 1.1. Как устроен курс

- 1 модуль - матан и линал
- 2 модуль - дискра, тервер
- 3/4 модуль - статистика

#### 1.2. Практика

Исследовать функцию

$$f(x) = x^3 - 3x^2 + 4$$

Найдём

#### 1.3. SymPy

В Питоне есть библиотека SymPy, которая предоставляет интерфейс для вычисления производных

```
!pip install sympy
```

Далее в питоне зададим переменную и производную:

```
import sympy as sp
x = sp.Symbol('x')
sp.diff(x**6)
```

Теперь будем анализировать функцию из практики:

```
def f(x):
    return x**3 - 3*x**2 + 4
```

Чтобы найти нули функции, надо решить уравнение  $f(x) = 0$ . В SymPy для этого есть функция solve:

```
sp.solve(f(x), x)
[-1, 2]
```

Теперь найдём производную функции  $f(x)$  и затем её нули, чтобы найти экстремумы

```
df_x = sp.diff(f(x))
#df_x == 3x^2
```

```
sp.solve(df_x, x)
# [0, 2]
```

```
f(0), f(2)
# (4, 0)
```

Точно также очень просто можем находить вторую производную и находить точки перегиба функции

```
d2f_x = sp.diff(df_x)
d2f_x
# 6x - 6
```

```
sp.solve(d2f_x, x)
#[1]
```

```
f(1)
#2
```

## 1.4. Рисуем графики

Что нам нужно будет сделать?

- Нарисовать график  $f(x)$ , подписать оси
- Напечатать под графиком при помощи *Markdown* экстремумы, точки перегиба и значения функции  $f(x)$  в этих точках

### 1.4.1. Подключаем библиотеку

```
import matplotlib.pyplot as plt
%matplotlib inline # -
```

### 1.4.2. Рисуем график функции $f(x)$

```
import numpy as np

x_values = [x for x in np.arange(-5, 5, 0.1)]
# or = np.linspace(-5, 5, 100)
f_values = [f(x)] for x in x_values]
```

```
plt.plot(x_values, f_values)
```

**Вставить график**

### 1.4.3. Изменяем размер картинки и граничные значения по осям

```
plt.figure(figsize=(10, 10))

plt.plot(x_values, y_values)

plt.xlim([-3, 5])
plt.ylim([-5, 7])
```

**Вставить график**

### 1.4.4. Добавим названия осей и подпись к графику

```
plt.title('Graph_of_fucntion_f(x)_with_extremum_and_dots_of...')

plt.xlabel('x')
plt.ylabel('f(x)')
```

**Вставить график**

### 1.4.5. Добавим оси и подпишем под картинкой информацию об экстремумах и точках перегиба

```
import numpy as np

x_values = [x for x in np.arange(-5, 5, 0.1)]
f_values = [f(x) for x in x_values]

plt.figure(figsize=(10,10))

plt.axvline(x=0, c = 'black')
plt.axhline(y=0, c = 'black')

plt.plot(x_values, f_values)

plt.xlim([-3, 5])
```

```
plt.ylim([-5, 7])

plt.title('Graph_of_fucntion_f(x)_with_extremum_and_dots_of_...')

plt.xlabel('x')
plt.ylabel('f(x)')

plt.show()
```

**Вставить график**

## 1.5. Метод градиентного спуска

### 1.5.1. Теорема о градиенте

**Градиент** - это вектор, в направлении которого функция растёт быстрее всего.

**Антиградиент** (вектор противоположный градиенту) - вектор, в направлении которого функция быстрее всего убывает.

### 1.5.2. Применение в машинном обучении

Для чего нам это нужно? В машинном обучении мы минимизируем значение функции, которая показывает ошибку модели. Иными словами: наша задача при обучении модели - найти такие веса  $\mathbf{w}$ , на которых достигается **минимум функции ошибок**.

В простейшем случае, если ошибка среднеквадратическая, то её график - парабола.

### 1.5.3. Идея применения градиентного спуска

На каждом шаге (на каждой итерации метода) движемся в сторону антиградиента функции потерь!

То есть на каждом шаге движемся в направлении уменьшения ошибки.

Вектор градиента функции потерь обозначают **grad Q** или  $\nabla Q$

### 1.5.4. Метод градиентного спуска на пальцах

- Встаём в некоторую точку функции
- Вычисляем градиент
- Переходим в новую точку в направлении антиградиента
- Повторяем процесс из новой точки

### 1.5.5. Метод градиентного спуска (одномерный случай)

Пусть у нас только один вес -  $w$ .

Тогда при добавлении к весу  $w$  слагаемого  $-\frac{\partial Q}{\partial w}$  функция  $Q(w)$  убывает.

Тогда алгоритм выглядит следующим образом:

- Инициализируем вес  $w^{(0)}$
- На каждом следующем шаге обновляем вес, добавляя  $-\frac{\partial Q}{\partial w}(w^{(k-1)})$ :

$$w^{(k)} = w^{(k-1)} - \frac{\partial Q}{\partial w}(w^{(k-1)})$$

### 1.5.6. Метод градиентного спуска (общий случай)

Пусть  $w_0, w_1, \dots, w_n$  - веса, которые мы ищем.

Тогда  $\nabla Q(w) = \left\{ \frac{\partial Q}{\partial w_0}, \frac{\partial Q}{\partial w_1}, \dots, \frac{\partial Q}{\partial w_n} \right\}$

Тогда алгоритм выглядит так:

- Инициализируем веса  $w^{(0)}$  (заметим, что это вектор весов)

- На каждом шаге обновляем веса по формуле:

$$w^{(k)} = w^{(k-1)} - \nabla Q(w^{(k-1)})$$

### 1.5.7. Параметр learning rate

В формулу обычно добавляют параметр  $\eta$  - величина градиентного спуска (**learning rate**). Он отвечает за скорость движения в сторону антиградиента:

- Инициализируем веса  $w^{(0)}$  (заметим, что это вектор весов)
- На каждом шаге обновляем веса по формуле:

$$w^{(k)} = w^{(k-1)} - \eta \nabla Q(w^{(k-1)})$$

### 1.5.8. Теорема о поиске в выпуклой гладкой функции

Если функция  $Q(w)$  выпуклая и гладкая, а также имеет минимум в точке  $w^*$ , то метод градиентного спуска при аккуратно подобранному  $\eta$  через некоторое число шагов гарантированно попадает в малую окрестность точки  $w^*$ .

## 1.6. Реализация градиентного спуска на python

```
def gradient_descent(x_start, learning_rate, epsilon, num_iterations):
    x_curr = x_start
    df_x = sp.diff(f(x))

    trace = []
    trace.append(x_curr)

    for i in range(num_iterations):
        x_new = c_curr + df_x.subs(x, x_curr)
        trace.append(x_new)

        if abs(x_new - x_curr) < epsilon:
            return x_curr, trace

    return x_curr, trace
```



## 2. Лекция 2

### 2.1. Линейная регрессия

**Линейная регрессия** - функция  $a(x) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \dots + \omega_l x_l$ , где  $x$  - **вектор признаков**

Также есть **целевая переменная**, которую мы предсказываем -  $y$

$w$  - **веса** линейной регрессии

Запишем в другой форме:  $a(x) = \omega_0 + \sum_{i=1}^l \omega_i x_i$

Также мы это можем записать в другой форме: давайте добавим ещё один признак у всех объектов, который будет равен единице:  $x = (1, x_1, x_2, \dots, x_n)$ , и тогда всё записывается ещё красивее:

$$a(x) = \sum_{i=0}^l \omega_i x_i = (\vec{\omega}, \vec{x})$$

Где  $a(x) = (\vec{\omega}, \vec{x})$  - предсказание модели на объекте  $x$ . Но это предсказание для одного объекта.

Мы можем записать предсказания в матричном виде для нескольких объектов. Возьмём  $X$  - матрицу объект-признак. В каждой строке описан один объект, а кол-во строк - это кол-во объектов.

В матричном виде предсказание выглядит как  $a(X) = X \cdot w$

### 2.2. Функция ошибок MSE

**MSE** (Mean Squared Error) =  $\frac{1}{d} \sum_{i=1}^d (a(x_i) - y_i)^2$ , где  $d$  - количество данных. При этом в задаче обучения мы хотим  $MSE \rightarrow \min_{\vec{\omega}}$ . Мы уже научились и можем делать минимизацию функции с помощью Градиентного спуска.

### 2.3. Матричное представление MSE

$$MSE = \frac{1}{d} \|X\omega - y\|^2$$

Где  $\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + \dots + a_n^2}$

Соответственно  $\|\vec{a}\|^2 = a_1^2 + a_2^2 + \dots + a_n^2$

### 2.4. Как решать задачу минимизации MSE

#### 2.4.1. Аналитическое решение

Это решение, которое даёт точное решение

Решаем уравнение  $\nabla_{\omega} Q(\omega) = 0$

#### 2.4.2. Приближённое решение

С помощью GD шагаем  $\omega = \omega - \eta \nabla_{\omega} Q(\omega)$

### 2.5. Производная по вектору

Пусть у нас есть  $\vec{x} = (x_1, \dots, x_n)$

Градиент функции  $f(x)$  рассчитывается как  $\nabla_x f(x) = (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n})$

Если мы хотим взять градиент по матрице  $A$  от функции  $f$ , то это выглядит как  $\nabla_A f(A) = \begin{pmatrix} \frac{\partial f}{\partial A_{11}} & \dots & \frac{\partial f}{\partial A_{1n}} \\ \dots & \dots & \dots \\ \frac{\partial f}{\partial A_{n1}} & \dots & \frac{\partial f}{\partial A_{nn}} \end{pmatrix}$

#### 2.5.1. Пример 1. Подсчёт градиента скалярного произведения

Пусть у нас есть вектор весов  $\vec{\omega}$  и вектор  $\vec{x}$ . Есть скалярное произведение  $(\vec{\omega}, \vec{x})$ . Мы хотим посчитать  $\nabla_x (\vec{\omega}, \vec{x})$ .

Мы знаем, что  $\frac{\partial}{\partial x_i} (\vec{\omega}, \vec{x}) = \frac{\partial}{\partial x_i} (\omega_1 x_1 + \dots + \omega_n x_n) = \omega_i$

Тогда  $\nabla_x (\vec{\omega}, \vec{x}) = (\frac{\partial}{\partial x_1} (\vec{\omega}, \vec{x}), \frac{\partial}{\partial x_2} (\vec{\omega}, \vec{x}), \dots, \frac{\partial}{\partial x_n} (\vec{\omega}, \vec{x})) = (\omega_1, \omega_2, \dots, \omega_n) = \vec{\omega}$

### 2.5.2. Пример 2. Подсчёт градиента от матрицы

Пусть есть матрица  $A_{n \times n}$  и вектор  $\vec{x} \in \mathbb{R}^n$ .

Функция  $x^T A x$  - это число (давайте посмотрим на размерности)

$$(1 \times n)(n \times n)(n \times 1) = (1 \times n)(n \times 1) = (1 \times 1)$$

Теперь мы хотим от этой функции находить градиент:

$$\begin{aligned} \text{Давайте попробуем посчитать } \frac{\partial}{\partial x_i} x^T A x &= \frac{\partial}{\partial x} \sum_{j=1}^n x_j (A x)_j = \frac{\partial}{\partial x_i} \sum_{j=1}^n x_j \left( \sum_{k=1}^n a_{jk} x_k \right) = \frac{\partial}{\partial x_i} \sum_{j=1}^n \sum_{k=1}^n a_{jk} x_j x_k = \\ &= \sum_{j=1, j \neq i}^n a_{ji} x_j + \sum_{k=1, k \neq i}^n a_{ik} x_i + 2a_{ii} x_i = \sum_{i=1}^n \sum_{j=1}^n (a_{ij} + a_{ji}) x_j - i\text{-я производная.} \end{aligned}$$

$$\text{Тогда } \nabla_x (x^T A x) = (A + A^T) x$$

## 2.6. Минимизация MSE с помощью градиента

Вспоминаем, что MSE выглядит как  $\|y - X\omega\|^2$ . Но это можно переписать в явном виде без квадрата:

$$\|y - X\omega\|^2 = (y - X\omega)^T (y - X\omega) \rightarrow \min_{\omega}$$

Раскрываем скобки для поиска градиента:  $\nabla_{\omega} ((y^T y)^0 - \omega^T X^T y - y^T X \omega + \omega^T X^T X \omega) = 0$

$$\nabla_{\omega} (-\omega^T X^T y - y^T X \omega + \omega^T X^T X \omega) = -X^T y - X^T y + 2X^T X \omega = 0 \quad (\text{для последнего слагаемого смотрим вывод пункта 2.5.2})$$

Перекинем слагаемые в разные стороны:  $2X^T X \omega = 2X^T y$ . Сократим на двойку. Мы бы могли сократить матрицы, но обратной может не быть. Зато мы можем с каждой из сторон умножить на обратную матрицу к  $X^T X$ :

$$X^T X \omega = X^T y \rightarrow (X^T X)^{-1} (X^T X) \omega = (X^T X)^{-1} X^T y \rightarrow \omega = (X^T X)^{-1} X^T y$$

## 2.7. Градиент функции потерь

Из подсчитанного можем сказать, что градиент функции потерь для MSE будет выглядеть как

$$\nabla Q(\omega) = 2X^T (X\omega - y)$$

Но давайте будем находить  $\omega$  с помощью GD:

- На шаге обновления точки, у нас  $\omega_{next} = \omega_{prev} - \eta \nabla Q(\omega)$
- Запишем зная, чему равно  $\nabla Q(\omega)$ :  $\omega_{next} = \omega_{prev} - 2\eta X^T (X\omega_{prev} - y)$

## 2.8. Реализация на питоне

### 2.8.1. Генерирование данных для задачи регрессии

Давайте сгенерируем данные и визуализируем их:

```
import random
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline

X = np.linspace(-10, 10, 100)

print(X.shape)

y = X * (np.random.random_sample(len(X)) + 0.5)
X = X.reshape(len(X), 1)

print(X.shape)

plt.scatter(X, y)
```

### 2.8.2. Функция подсчёта ошибки

Также давайте напишем свой MSE:

```
def MSE(X, y, theta):
    m = len(y)

    error = (1./m) * (np.linalg.norm(X @ theta - y) ** 2)
    return error
```

### 2.8.3. Реализация градиентного спуска

Теперь у нас есть всё, чтобы реализовать свой градиентный спуск:

```
def gradient_descent(X, y, learning_rate, iterations):

    X = np.hstack((np.ones((X.shape[0], 1)), X)) # add column of ones
    params = np.random.rand(X.shape[1])

    m = X.shape[0]

    cost_track = np.zeros((iterations, 1))

    for i in range(iterations):
        params = params - 2./m * learning_rate * (X.T @ ((X @ params) - y))
        cost_track[i] = MSE(X, y, params)

    return cost_track, params
```

### 2.8.4. Функция предсказания модели

Записать предсказание модели можно очень просто:

### 2.8.5. Применение градиентного спуска

Применяем градиентный спуск:

```
track, weights = stochastic_gradient_descent(X, y, 0.01, 100)
plt.plot(track) # visualize errors
```

Теперь сделаем предсказание и посмотрим на визуализацию этого предсказания:

```
pred = predict(X, weights)
plt.scatter(X, y)
plt.plot(X, pred, '-', c = 'r')
```

### 3. Лекция 3

#### 3.0.1. Один из недостатков градиентного спуска

С точки зрения реализации есть следующий недостаток:

На каждом шаге вычисления  $\nabla Q(w)$  мы вычисляем производную по каждому весу от каждого объекта. То есть мы вычисляем целую матрицу производных - это затратно и по времени, и по памяти

#### 3.1. Стохастический градиентный спуск

##### Stochastic Gradient Descent

На каждом шаге мы выбираем **один случайный объект** и сдвигаемся в сторону антиградиента по этому объекту:

$$\omega^{(k)} = \omega^{(k-1)} - \eta_k \cdot \nabla q_{i_k}(\omega^{(k-1)})$$

где  $\nabla q_{i_k}(\omega)$  - градиент функции, вычисленный только по объекту с номером  $i_k$  (а не по всей обучающей выборке)

Если функция  $q(\omega)$  выпуклая и гладкая, а также имеет минимум в точке  $\omega^*$ , то метод стохастического градиентного спуска при аккуратно подобраном  $\eta$  (LR) через некоторое число шагов гарантированно попадает в малую окрестность точки  $\omega^*$ . Однако сходится метод медленнее, чем обычный градиентный спуск.

#### 3.2. Mini-Batch Gradient Descent

Промежуточное решение между классическим градиентным спуском и стохастическим вариантом

- Выбираем batch size (например, 32, 64, и т.д.). Разбиваем все пары объект-ответ на группы размера batch size
- На  $i$ -й итерации градиентного спуска вычисляем  $\nabla Q(\omega)$  только по объектам  $i$ -го батча:

$$\omega^{(k)} = \omega^{(k-1)} - \eta_k \nabla Q_i(\omega^{(k-1)})$$

где  $\nabla Q_i(\omega^{(k-1)})$  - градиент функции потерь, вычисленный по объектам из  $i$ -го батча

#### 3.3. Ещё один недостаток градиентного спуска

Мы можем застрять в локальном минимуме и не дойти до глобального минимума

#### 3.4. Метод моментов (Momentum)

Будем добавлять какие-то предыдущие значения шагов, которые будут аналогом инерции из физики:

Вектор инерции (усреднение градиента по предыдущим шагам):

$$h_0 = 0$$

$$h_k = \alpha h_{k-1} + \eta \nabla Q(w^{k-1})$$

Формула метода моментов:

$$w^{(k)} = w^{(k-1)} - h_k$$

Подробнее:

$$w^{(k)} = w^{(k-1)} - \eta \nabla Q(w^{(k-1)}) - \alpha h_{k-1}$$

#### 3.5. Преимущества метода моментов

Проще подбирать параметр  $\alpha$ , чем делать несколько запусков обычного градиентного спуска, потому что в многомерном случае сложнее генерировать эти данные и сложнее так находить глобальный минимум

Также иногда, когда у нас данных много, GD может работать неделю, и нам непозволительно делать много запусков

### 3.6. Градиентный шаг

В общем случае градиентный шаг может зависеть от номера итерации, тогда мы будем писать не  $\eta$ , а  $\eta_k$

- $\eta_k = c$  - это то что было у нас раньше - постоянный LR
- $\eta_k = \frac{1}{k}$  - здесь проблема в том, что мы можем на первом шаге стоять очень близко от глобального минимума и пролететь его
- $\eta_k = \lambda \left( \frac{s}{s_0 + k} \right)^p$ ,  $\lambda, s_0, p$  - параметры

### 3.7. AdaGrad (Adaptive Gradient)

Сумма квадратов обновлений

$$g_{k-1,j} = (\nabla Q(\omega^{(k-1)}))^2$$

Формула метода AdaGrad:

$$G_{k,j} = G_{k-1,j} + g_{k-1,j} = G_{k-1,j} + (\nabla Q(\omega^{(k-1)}))^2$$
$$\omega_j^{(k)} = \omega_j^{(k-1)} - \frac{\eta}{\sqrt{G_{k,j}} + \varepsilon} \cdot (\nabla Q(\omega^{(k-1)}))_j$$

Этот метод использует адаптивный шаг обучения по каждой из координат веса - тем самым мы регулируем скорость сходимости метода.

Плюсы метода: происходит затухание величины шага

Минусы метода:  $G_{k,j}$  монотонно возрастает, поэтому шаги укорачиваются и мы можем не успеть дойти до минимума

### 3.8. RMSPROP (Root Mean Square Propagation)

Метод реализует экспоненциальное затухание градиентов

Формулы метода RMSprop (усреднённый по истории квадрат градиента):

$$G_{k,j} = \alpha \cdot G_{k-1,j} + (1 - \alpha) \cdot g_{k-1,j}$$
$$w_j^{(k)} = \omega_j^{k-1} - \frac{\eta}{\sqrt{G_{k,j}} + \varepsilon} \cdot (\nabla Q(\omega^{(k-1)}))_j$$

Если мы быстро сдвигались на последних шагах - то следующие будут маленькие

Если мы сдвигались медленно - то шаги будут большие

### 3.9. Практикум на питоне

#### 3.9.1. Генерируем данные

```
import numpy as np
from matplotlib import pylab as plt
%pylab inline

from sklearn.datasets import make_regression
```

```
X, y = make_regression(n_samples=10000)
print(X.shape, y.shape)
```

#### 3.9.2. Библиотечное решение

У нас 10000 объектов и 100 признаков. Для начала решим задачу аналитически "из коробки".

Решим сначала аналитически с помощью LinearRegression

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
```

```
lr = LinearRegression()
```

```
lr.fit(X, y)
```

```
print(mean_squared_error(y, lr.predict(X)))
```

Посмотрим на свободный член модели (intercept\_) и веса (coef\_)

```
print(lr.intercept_, lr.coef_[:5])
```

Смотрим всего 5 весов, так как в реале их там 100.

Теперь решим с помощью градиентного спуска:

```
from sklearn.linear_model import SGDRegressor
```

```
sgd = SGDRegressor(alpha=0.00000001)
```

```
sgd.fit(X, y)
```

```
print(mean_squared_error(y, sgd.predict(X)))
```

И теперь точно также посмотрим свободный член и веса

```
print(sgd.intercept_, sgd.coef_[:5])
```

## 4. Лекция 5. Линейная алгебра

Что такое векторное пространство? Это набор элементов (векторов), которые можно складывать и умножать на скаляр.

Что означает пространство  $\mathbb{R}^2$ ? Что базис состоит из двух векторов. В каждом по две координаты.

Что такое линейный оператор? Линейное отображение векторного пространства в себя. Задаётся матрицей

Матрица поворота на угол  $\alpha$

$$\begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix}$$

### 4.0.1. Упражнение

- Что делает оператор

$$\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$$

- Что делает оператор

$$\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$$

- Что делает оператор

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

### 4.0.2. Ответы на упражнение

- Растягивает всё пространство в 3 раза (гомотетия)
- Матрица поворота на 45 градусов
- Проекция на первую координату (ось)

## 4.1. Выражение базиса через другой базис

Если есть два базиса  $e$  и  $g$  и есть линейный оператор  $A : e \rightarrow e'$ , то можно найти линейный оператор  $A'$  для базиса  $g$  как

$$A' = M^{-1}AM$$

где  $M$  - базис  $g$  выраженный через базис  $e$ .

Мы можем получить одно и то же преобразование разными матрицами, потому что для л.о. могли взять разные базисы. Но у этих л.о. есть базовые характеристики/свойства, которые не меняются.

Геометрический смысл определителя матрицы л.о. показывает как изменилась площадь, натянутая на векторы, которые мы отображали. Если  $A$  - матрица л.о.,  $x$  - векторы, которые образуют фигуру,  $x' = Ax$  - отображённые векторы, то

$$\frac{S'}{S} = \det A$$

где  $S$  - площадь  $x$ , а  $S'$  - площадь  $x'$

Вспомним о собственном векторе - собственный вектор не меняет направление после применения л.о. Длину - он может изменить. То есть, если  $v$  - собственный вектор, тогда  $Av = \lambda v$ , где  $\lambda$  - скаляр.

Вспомним, что если у нас многочлен нечётной степени, то хотя бы один единственный вещественный корень у него есть. Это нам говорит о том, что при матрице с нечётной длиной - всегда есть минимум один собственный вектор.

А теперь вспомним, что есть такая штука, как след -  $\text{tr } A$ , который равен сумме чисел на главной диагонали. След равен сумме собственных чисел, а значит - он тоже не меняется.

Подытожим: необходимое, но не достаточное условие для того, чтобы сказать, что какие-то два оператора задают одно и то же преобразование:

- Определитель - одинаковый
- Собственные значения - одинаковые
- След - одинаковый

## 4.2. Программирование

Для задач линейной алгебры есть две библиотеки - `numpy.linalg` и `scipy.linalg`. Обе эти библиотеки имеют одинаковые библиотеки, но `scipy.linalg` - поддерживает больше методов.

### 4.2.1. Нетривиальные вещи

функция `np.linalg.eig` - возвращает собственные значения и собственные векторы, в виде двух возвращаемых переменных. При чём если мы приняли в  $v$  собственные векторы, то первый собственный вектор будет не  $v[0]$ , а  $v[[0][0], v[1][0], \dots, v[n-1][0]]$ .

### 4.2.2. Отрисовка векторов

```
fig, ax = plt.subplots(nrows=1, ncols=2)

origin = np.array([[0, 0], [0, 0]])
U =      np.array([[1, 0], [0, 1]])

ax[0].quiver(*origin, U[:,0], U[:,1], color='g', scale=7)
```

Данный код отрисует два вектор -  $(1, 0)$  и  $(0, 1)$



## 5. Лекция 6. Метод главных компонент (РСА)

РСА - Principal Component Analysis.

Что мы хотим - снизить размерность у данных.

Зачем это нужно?

- Ускорить вычисления
- Сократить память
- Упростить интерпретируемость модели - проще смотреть на 10 признаков, чем на 100
- Уменьшить шанс переобучения
- Хотим лучше учиться - могут быть шумовые и неинформативные признаки

Как мы можем это реализовать?

- Когда мы убираем какие-то признаки, это называется отбором признаков
- Также мы можем выразить новые признаки через старые, при чём новых будет меньше - придумать признаки

Цель метода главных компонент: мы хотим придумать новые признаки, каким-то образом выражающиеся через старые, причём новых признаков хочется получить меньше, чем старых. И метод главных компонент **линейно** выражает новые признаки через старые.

Допустим мы хотим из 100 признаков сделать 10. Какой критерий должен быть для этих признаков? Какие из них являются хорошими

### 5.1. Постановка задачи выражения новых признаков

- $x_1, \dots, x_n$  - исходные числовые признаки
- $z_1, \dots, z_d$  - новые числовые признаки,  $d \leq n$

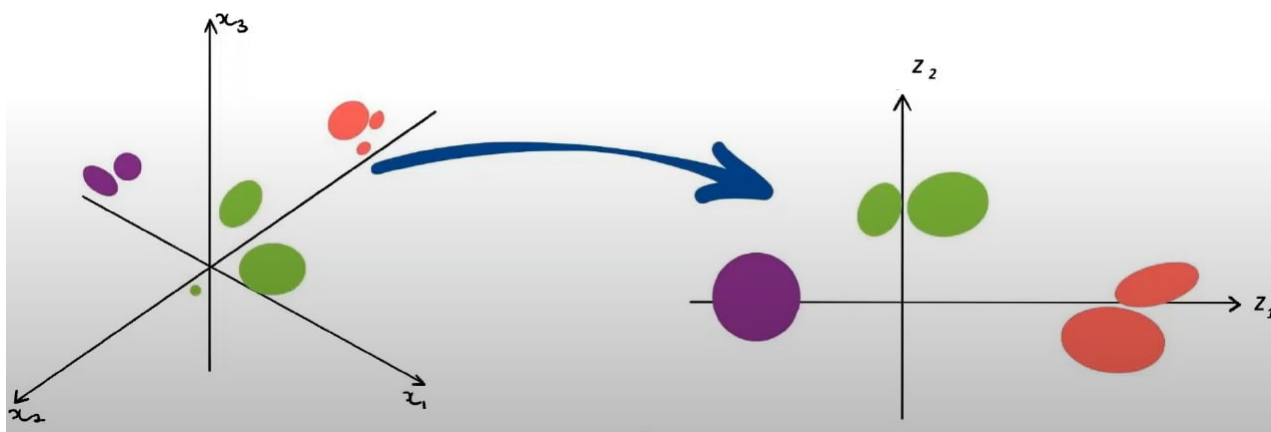
Хотим:

- Чтобы новые признаки  $z_j$  линейно выражались через исходные признаки  $x_i$
- Чтобы при переходе к новым признакам было потеряно наименьшее кол-во информации

#### 5.1.1. Интерпретация линейного выражения признаков

$$\begin{cases} z_1 = u_{11}x_1 + \dots + u_{1n}x_n \\ z_2 = u_{21}x_1 + \dots + u_{2n}x_n \\ \dots \\ z_d = u_{d1}x_1 + \dots + u_{dn}x_n \end{cases}$$

Геометрическая интерпретация: новые признаки  $z_j$  - это проекции исходных признаков  $x_i$  на некоторые векторы (компоненты)  $u$ . Мы проецируем пространство признаков размерности  $n$  на некоторое линейное подпространство размерности  $d$ :



### 5.1.2. Пояснение: Проекция

- Проекция вектора  $x$  на вектор (компоненту)  $u_i$ :

$(x, u_i)$  - скалярное произведение

- Проекция выборки  $X$  на компоненту  $u_i$ :

$$Xu_i$$

### 5.1.3. Сохранение информации

Также мы хотим, чтобы было потеряно наименьшее кол-во исходной информации

Дисперсия выборки, посчитанная в новых признаках, показывает, как много информации нам удалось сохранить после понижения размерности, поэтому дисперсия в новых признаках должна быть максимальной.

Пример: Хотим спроецировать двумерные данные  $X$  на одномерный вектор  $u$  так, чтобы дисперсия проекции  $Xu$  была максимальной:

Рис. 1: Плохое проецирование

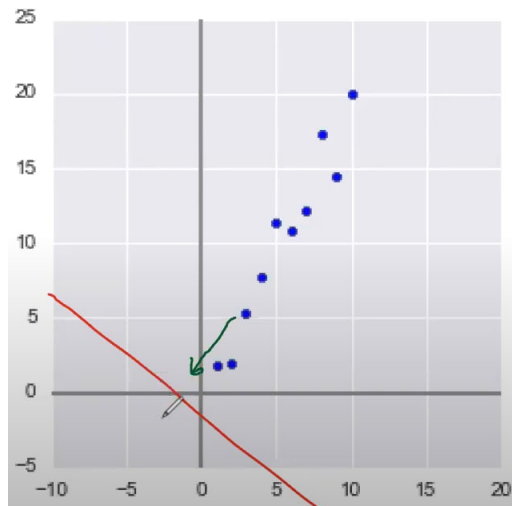
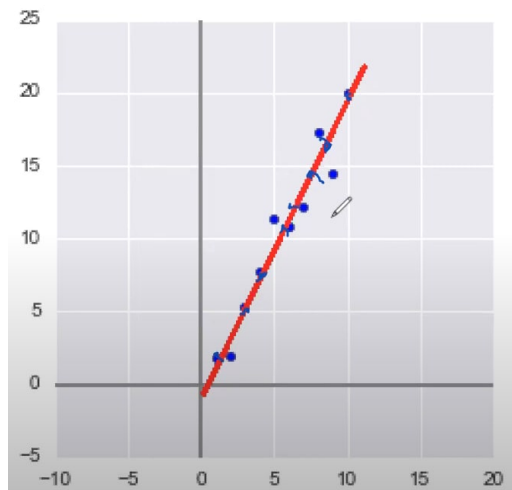


Рис. 2: Хорошее проецирование



## 5.2. Постановка задачи

Будем искать такие компоненты  $u_1, u_2, \dots, u_d$ , что:

- Они ортогональны, то есть  $\forall i, j : (u_i, u_j) = 0$
- Они нормированы, т.е.  $\|u_i\| = 1$

- Дисперсия проекции выборки на них максимальна:  $\mathbb{D}(Xu_i) \rightarrow \max_{u_i}, i = 1, \dots, d$
- Норма разности изначальных объектов и восстановленных - минимальна:

$$X \rightarrow Z = Xu \rightarrow u^T Z \iff \|X - u^T Z\|^2 \rightarrow \min_u$$

Последние два высказывания эквивалентны

- Мы уже выяснили, что проекция выборки  $X$  на компоненту  $u_i$ :  $Xu_i$
- Тогда проекция выборки на первые  $d$  компонент, задаваемых столбцами матрицы  $U_d = \langle u_1, u_2, \dots, u_d \rangle$ :

$$XU_d$$

- Тогда дисперсия проекции - это след ковариационной матрицы:

$$\text{tr}((XU_d)^T(XU_d)) = \sum_{i=1}^d \|Xu_i\|^2 \rightarrow \max_u$$

### 5.2.1. Ковариация

Напомним, что матрица ковариации  $\Sigma = \text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)^T]$

Заметим тогда, что дисперсия проекции на вектор  $u_i$  есть ни что иное, как  $\mathbb{D}(Xu_i) = \|Xu_i\|^2$

## 5.3. Алгоритм поиска новых компонент

### 5.3.1. Первый шаг

Будем искать первую компоненту,  $u_1$ :

$$\begin{cases} \|Xu_1\|^2 \rightarrow \max_{u_1} \\ \|u_1\|^2 = 1 \end{cases}$$

Как такое делать?

Запишем лагранжиан

$$L(u_1, \lambda) = \|Xu_1\|^2 + \lambda(\|u_1\|^2 - 1)$$

Найдём  $\frac{\partial L}{\partial u_1} = 2X^T Xu_1 + 2\lambda u_1 = 0$

Получаем, что

$$X^T Xu_1 = -\lambda u_1$$

И это будет являться собственным вектором. Для подстановки будем использовать

$$\|Xu_1\|^2 = u_1^T X^T Xu_1 = \lambda u_1^T u_1 = \lambda \rightarrow \max_{u_1}$$

и найдём максимальное собственное значение, т.к.  $u_1^T u_1 = \|u_1\|^2 = 1$

## 5.4. Применение метода

Когда главные компоненты найдены, можно проецировать на них и новые данные

$$Z' = X'U_d$$

## 5.5. Доля объяснённой дисперсии

- Упорядочим собственные значения матрицы  $X^T X$  по убыванию  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$
- Доля дисперсии, объяснённой  $j$ -й компонентой (explained variance ratio), высчитывается как

$$\delta_j = \frac{\lambda_j}{\sum_{i=1}^n \lambda_i}$$

- Доля дисперсии, объясняемой первыми  $k$  компонентами:

$$\delta_j = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i}$$

## 5.6. Выбор числа главных компонент

Эффективная размерность выборки - это наименьшее целое  $m$ , при котором доля необъяснённой дисперсии

$$E_m = \frac{\|ZU^T - X\|^2}{\|X\|^2} = \frac{\lambda_{m+1} + \dots + \lambda_n}{\sum \lambda_i} \leq \varepsilon$$

Пользуемся этим равенством как фактом.

Также есть критерий крутого склона, когда первые  $m$  компонент дают много объяснения дисперсии, а начиная с какого-то  $m + 1$  очень мало. Тогда мы возьмём первые  $m$  компонент

## 5.7. Упражнения

### 5.7.1. Нахождение главных компонент

Берём матрицу  $X = \begin{pmatrix} 1 & 3 \\ 0 & 2 \\ 0 & 0 \\ 3 & 3 \end{pmatrix}$ , вычитаем среднее в столбцах  $\rightarrow \begin{pmatrix} 0 & 1 \\ -1 & 0 \\ -1 & -2 \\ 2 & 1 \end{pmatrix}$ , находим матрицу ковариации

$$\text{cov}(\tilde{X}) = \tilde{X}^T \tilde{X} = \begin{pmatrix} 0 & -1 & -1 & 2 \\ 1 & 0 & -2 & 1 \end{pmatrix} \times \begin{pmatrix} 0 & 1 \\ -1 & 0 \\ -1 & -2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 4 \\ 4 & 6 \end{pmatrix}$$

$$\left| \begin{pmatrix} 6-\lambda & 4 \\ 4 & 6-\lambda \end{pmatrix} \right| = 0 \iff \lambda_1 = 10, \lambda_2 = 2$$

$$\text{Для значения } \lambda_1 = 10 \text{ находим собственные значения } \begin{pmatrix} -4 & 4 \\ 4 & -4 \end{pmatrix} \times \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} \iff v_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

Для значения  $\lambda_2 = 2$  - аналогично

### 5.7.2. То же самое в питоне

#### 5.7.2.1. Генерация данных .

```
x = np.arange(1, 11)
y = 2 * x + np.random.randn(10) * 2
X = np.vstack((x, y))

plt.scatter(X[0], X[1])
```

#### 5.7.2.2. Реализация метода главных компонент .

```
Xcentered = X - X.mean(axis=1, keepdims=True)
m = np.mean(X, axis=1)
```

```
print("Mean_vector:", m)
print(Xcentered)
```

```
plt.scatter(Xcentered[0], Xcentered[1])
```

#### 5.7.2.3. Вычисляем матрицу ковариаций .

```
covmat = np.cov(Xcentered)
print(covmat, '\n')
print('Variance_of_X:', np.cov(Xcentered)[0, 0])
print('Variance_of_Y:', np.cov(Xcentered)[1, 1])
print('Covariance_of_X_and_Y:', np.cov(Xcentered)[0, 1])
```

#### 5.7.2.4. Находим собственные векторы и проецируем все точки .

```
eignums, vecs = np.linalg.eig(covmat)
v = vecs[:, np.argmax(eignums)]
Xnew = np.dot(v, Xcentered)
print(Xnew)
```

#### 5.7.2.5. Восстановление данных .

```
n = 3
Xrestored = np.dot(Xnew[n], v) + m
print('Restored:\n', Xrestored)
print('Original:\n', X[:, n])
```

#### 5.7.2.6. Доля объяснённой дисперсии .

```
100 * max(eignums) / sum(eignums)
```

#### 5.7.2.7. Что выдаёт PCA из sklearn .

```
from sklearn.decomposition import PCA
```

```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(Xcentered.T)
```

### 5.8. Сингулярное разложение (SVD)

SVD - Singular Value Decomposition

**Теорема.** Матрицу  $A \in \mathbb{R}^{m \times n}$  можно представить в виде

$$A = U \Sigma V^T$$

где

- $U \in \mathbb{R}^{m \times m}$ ,  $V \in \mathbb{R}^{n \times n}$  - ортогональные матрицы
- $\Sigma \in \mathbb{R}^{m \times n}$  - диагональная матрица с ненулевыми элементами  $\sigma_i = \sqrt{\lambda_i}$ , где  $\lambda_i$  - собственные значения матрицы  $A^T A$

При этом

- Столбцы матрицы  $U$  являются собственными векторами матрицы  $AA^T$
- Столбцы матрицы  $V$  являются собственными векторами матрицы  $A^T A$

По аналогии с PCA можно показать, что в столбцах  $V$  и стоят главные компоненты - столбцы  $V$  являются её собственными векторами, и мы на них проецируем, и по сути, её столбцы и являются компонентами.

### 5.9. Что лучше: PCA или SVD?

- Существуют вычислительные трудности с нахождением собственных значений, в этом недостаток PCA
- Существует итерационный алгоритм для нахождения SVD (без нахождения собственных значений)

Поэтому вычислительно эффективнее использовать SVD при прочих равных.

### 5.10. Упражнение

- Разложить матрицу  $X \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$  в SVD

### 5.10.1. Находим ответы

Найдём собственные значения матрицы  $X^T X = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$

Найдём собственные значения  $\left| \begin{pmatrix} 5 - \lambda & 4 \\ 4 & 5 - \lambda \end{pmatrix} \right| = 0 \iff \lambda_1 = 9, \lambda_2 = 1$

Тогда получаем, что  $\Sigma = \begin{pmatrix} \sqrt{9} & 0 \\ 0 & \sqrt{1} \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$

Собственные векторы  $v_1 = \begin{pmatrix} \frac{-1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix}, v_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix} \rightarrow V = \begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}$

Можем посчитать  $U$  как  $U = X V \Sigma^{-1} = \begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$

### 5.11. Связь SVD и PCA

- Столбцы матрицы  $V$  - это собственные векторы матрицы  $X^T X$ , т.е. векторы  $v_1, \dots, v_n$  - главные компоненты
- Столбцы матрицы  $U \Sigma$  - это новые признаки  $z = Xv$  ( $X = U \Sigma V^T \iff U \Sigma = X V$ )
- Сингулярные числа матрицы  $\Sigma$  - это корни из собственных чисел матрицы  $X^T X$

Для снижения размерности - берём первые  $k$  столбцов матрицы  $U$  и верхний  $k \times k$  квадрат матрицы  $\Sigma$ , тогда матрица  $U_K \Sigma_k$  содержит  $k$  новых признаков, соответствующих первым  $k$  компонентам.

### 5.12. Сингулярное разложение в питоне

```
import numpy as np

A = np.array([[3, 4, 3], [1, 2, 3], [4, 2, 1]])

U, D, VT = np.linalg.svd(A)

A_remake = (U @ np.diag(D) @ VT)
print(A_remake)
```

# Часть II

## Модуль 2

Преподавать будет Филипп Ульянов

### 6. Лекция 1. Множества

Что такое **множества**? Это перечисление каких-то элементов. Обычно мы множество записываем в фигурные скобки:  $A = \{5, 42, 7\}$  - множество, которое состоит из числа 5, 42 и 7.

Также мы можем задать из каких элементов состоит наше множество, если их долго перечислять:  $B = \{2^{k-1} | k \in \mathbb{N}\} = \{2^0, 2^1, 2^2, 2^3, \dots\}$ . То есть мы указываем свойство.

Третий способ задать множество - написать алгоритм получения элементов:  $1 \in C$  или  $a \in C \rightarrow 2 \cdot a \in C$ .

#### 6.1. Особые множества

- $\mathbb{N}$  - множества натуральных чисел
- $\mathbb{Z}$  - целые числа
- $\mathbb{Q}$  - рациональные числа
- $\mathbb{R}$  - вещественные числа
- $\mathbb{C}$  - комплексные числа
- $\emptyset$  - пустое множество
- $\mathbb{U}$  - универсальное множество - множество из всех существующих элементов
- $2^A$  - булеан - множество всех подмножеств

#### 6.2. Упражнение

$$A = \{1, 4, 5\}, B = \{4, 3\}, C = \{2\}$$

$$A \cap B = \{4\}$$

$$A \cup B = \{1, 3, 4, 5\}$$

$$A \cap C = \emptyset$$

$$A \setminus B = \{1, 5\}$$

$$A \Delta B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B) = \{1, 3, 5\}.$$

$$A \times B = \{(1, 4), (1, 3), (4, 4), (4, 3), (5, 4), (5, 3)\}$$

$$2^A = \{\{1, 4, 5\}, \emptyset, \{1\}, \{4\}, \{5\}, \{1, 4\}, \{1, 5\}, \{4, 5\}\}$$

$$|A| = 3 - \text{мощность множества}$$

$$|2^A| = 2^{|A|} = 2^3 = 8$$

Существует очень много свойств, для примера:

- $A \cup B \cap C = (A \cap C) \cup (B \cap C)$
- $(A \cup B) \times C = (A \times C) \cup (B \times C)$
- $\overline{\overline{A \cup B}} = \overline{A} \cap \overline{B}$  - закон де Моргана

#### 6.3. Упражнение

Отель имеет  $\infty$  номеров. Все номера заняты, мы не можем заселиться.

Давайте переселим всех людей на номер  $+1$  и тогда освобождается первый номер и мы можем заселиться.

А что если приехало  $\infty$  новых людей? Сможем ли мы их заселить?

Давайте каждого человека селить не в номер  $+1$ , а в номер  $\cdot 2$ . Тогда освобождаются все нечётные номера и мы их можем заселить.

##### 6.3.1. Определение равномощности

Множества  $A$  и  $B$  называются равномощными, если между их элементами существует взаимно-однозначное соответствие (биекция).

## 6.4. Упражнение

Сравнить множества

- $|\mathbb{Z}|$  и  $|\mathbb{N}|$
- $|\mathbb{N} \times \mathbb{N}|$  и  $|\mathbb{N}|$
- $|\mathbb{Q}|$  и  $|\mathbb{N}|$

## 6.5. Ответы на упражнения

- Очень легко. Просто нумеруем целые числа в порядке 0, -1, 1, -2, 2 итд, и получаем биекцию
- Аналогично с пунктом ниже
- На самом деле их мощности совпадают, потому что мы можем построить биекцию. А эту биекцию мы можем сделать в виде алгоритма. Мы можем нарисовать координатную сетку, и обходить её "улиткой" начиная из (0,0). Если мы в очередной шаг наступаем на рациональное число, то увеличим каунтер встреченных чисел и говорим какое по номеру это число. Рано или поздно мы встретим любое из рациональных чисел. При этом встречать рациональные числа мы не перестанем, потому что любое число вида  $\frac{1}{n}$  или  $\frac{n-1}{n}$  является рациональным. Доказано

Определение: говорят, что множество  $A$  счётное, и оно равномощно множеству  $|\mathbb{N}|$ .

## 6.6. Все ли бесконечные множества равномощны?

Ответ: нет. Доказательство:

Пусть у нас есть множество бесконечных бинарных чисел. Мы их начинаем нумеровать (строить биекцию из  $\mathbb{N}$ ). Предположим такая биекция есть. Но давайте выпишем каждое число в строку. Поменяем бит на главной диагонали. Тогда мы получим на диагонали новое число, которое не равно ни одному из уже имеющихся, ведь с  $i$ -м числом оно различается в  $i$ -м бите.

## 6.7. Континуальность

Определение: Множество  $B$  для которого мощность совпадает с  $\mathbb{R}$  называется континуальным.

## 6.8. Теорема о удалении бесконечного множества

Если из бесконечного множества  $M$  вырезать счётное подмножество  $A$  и осталось бесконечное кол-во элементов, то  $|M \setminus A| = |M|$

## 6.9. Упражнение

За  $S$  сейчас и в дальнейшем обозначаем континуальное множество. Множество бесконечных бинарных последовательностей.

Является ли  $|[0; 1]| = |S|$ ?

$\forall X \in [0; 1]$  можно записать в виде бесконечной последовательности из 0 и 1. Можно бинарным поиском делать эту последовательность: если лежит  $\leq m$ , то пишем 0, если  $> m$ , то 1.

## 6.10. Упражнение

Последовательности похожи, если отличаются на конечное число элементов.

Есть элемент из бесконечного числа нулей. Какова мощность множества элементов  $|A|$  всех похожих на этот элемент?

Как минимум счётно, так как мы можем нумеровать последовательность слева в виде двоичной системы счисления, а справа вставить бесконечное кол-во нулей.



### 6.11. Упражнение

Отношение похожести разбиывает все последовательности на классы эквивалентности. Какова мощность таких множеств? Допустим элемент множества  $A_i$

У нас есть всего два варианта: либо это континуальное множество, либо счётное.

У  $S$  мощность континуум.  $S$  мы можем получить, как  $S = \bigcup_{i=1}^{\infty} A_i$ .

Если мы объединим бесконечное множество счётных множеств, то мы получаем континуальное. Но мы знаем, что при объединении счётных множеств  $(\mathbb{N} \times \mathbb{N})$  мы получаем счётное. Тогда мы приходим к противоречию и получаем, что мощность  $A_i$  - континуально.

### 6.12. Упражнение

Дракон захватил гномов и поставил их в шеренгу. На каждого гнома надет колпак. Каждый гном видит все колпаки впереди стоящих перед ним гномов, но свой и гномов позади колпаки - не видит.

Дракон начинает задавать гномам вопрос: "какой колпак на тебе надет?". Если гном отвечает правильно - дракон его отпускает. Если гном отвечает неправильно - дракон его съедает.

Вопрос: существует ли здесь какая-то стратегия такая, что погибнет конечное число гномов?

## 7. Лекция 2. Сигма-алгебра

### 7.1. Пространство элементарных исходов

Все варианты, которые произойти или не произойти для какого-то случайного действия - называются пространством элементарных исходов и обозначаются  $\Omega$ . Для события подкидывания игральной кости, у нас элементарные исходы могут быть - чем выпала кость, и тогда  $\Omega = \{1, 2, 3, 4, 5, 6\}$ . Сами исходы обычно изображаются  $\omega_i$ .

Все события, с которыми мы обычно имеем дело - это какие-то множества  $A \subset \Omega$ , которые лежат внутри пространства элементарных исходов

### 7.2. Упражнение (классическая вероятность)

$\Omega = \{1, 2, 3, 4, 5, 6\}$   $A$  - выпало чётное число на кубике, тогда  $A = \{2, 4, 6\} \subset \Omega$ .

Будем считать, что кубик правильный, и  $\mathbb{P}(\omega_i) = \frac{1}{6}$ . Но часто события не являются равновероятными.

Но в нашем случае  $\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = \frac{1}{2}$

### 7.3. Важная мысль

Нам надо уметь считать мощность множеств там, где события равновероятны  $\Rightarrow$  нам нужно обладать знаниями комбинаторики.

Может быть также, что вероятность определяется не для конечных множеств, а для бесконечных множеств, например для площадей

### 7.4. Упражнение (геометрическая вероятность)

Допустим есть промежуток времени в 1 час с 9 до 10 утра. В этот час Саша и Катя хотят встретиться. Но они приходят в случайное время, ждут 15 минут и уходят. Какова вероятность, что Саша и Катя встретятся?

Как устроено пространство элементарных исходов и как можно параметризовать наше событие?

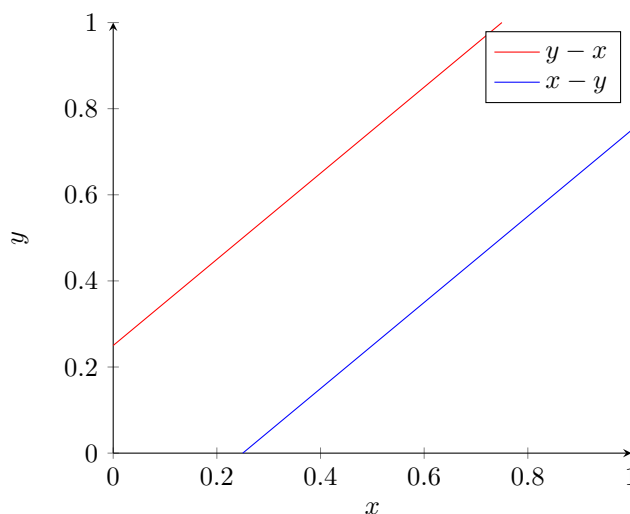
Можно сказать, что час - это отрезок  $[0; 1]$ , а ждут они 0.25 ( $\frac{1}{4}$  часа - 15 минут).

Тогда  $\Omega = [0; 1] \times [0; 1]$  - времена, когда пришли девочки.  $x$  - когда пришла Саша,  $y$  - когда пришла Катя. Тогда мы можем нарисовать квадрат. Тогда для нас множество, которое удовлетворяет событию -  $A = \{(x, y) \in \Omega \mid |x - y| \leq 0.25\}$

Что означает  $|x - y| \leq \frac{1}{4}$ ?

$$A = \begin{cases} x > y & x - y \leq \frac{1}{4} \\ x < y & y - x \leq \frac{1}{4} \end{cases}$$

Давайте нарисуем это:



Площадь между этими прямыми нам и подходит для ответа.

Но конкретно в этом случае нам проще найти площади, которые отвечают за то, что они не встретятся.

$$\mathbb{P}(\text{Встреча есть}) = 1 - \mathbb{P}(\text{Встречи нет}) = 1 - \left(\frac{3}{4}\right)^2 = \frac{7}{16}$$

Следующий вопрос: можно ли  $\forall A \in \Omega$  считать событием?

Ответ: иногда да, иногда нет, зависит от сигма-алгебры  $\mathbb{F}$   
А как правильно определять вероятность  $\mathbb{P}$

## 7.5. Сигма-алгебра

Вероятностная тройка -  $(\Omega, \mathbb{F}, \mathbb{P})$ .

$\sigma$ -алгебры: список множеств, которым мы считаем событием

Удобное определение: набор множеств  $\mathbb{F}$ , которые являются подмножествами множества  $\Omega$ , называется  $\sigma$ -алгеброй, если:

- $\emptyset \in \mathbb{F}, \Omega \in \mathbb{F}$
- Если событие  $A_1, A_2, \dots, A_n \in \mathbb{F}$ , тогда любая их комбинация  $(\cup, \cap, \setminus, \Delta, \bar{A})$  тоже лежит в  $\mathbb{F}$

Формальное определение:

- $\Omega \in \mathbb{F}$
- $A \in \mathbb{F} \rightarrow A^c = \Omega \setminus A \in \mathbb{F}$
- $A_1, A_2, \dots, A_n \in \mathbb{F} \rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathbb{F}$

и это является минимальным набором требований для  $\sigma$ -алгебры

## 7.6. Упражнение

$$A \cap B = \overline{\overline{A} \cup \overline{B}}$$

## 7.7. Упражнение

Все мы помним  $\Sigma = \{1, 2, 3, 4, 5, 6\}$  для бросания кубика. Но есть Вася, который знает у чисел только остаток от деления на 3. Тогда он не различает между собой события выпало 1/4, 2/5, 3/6. Для васи набор событий  $\mathbb{F}$  - те события, которые он может понять. Если более формально, то  $\mathbb{F} = \{\{1, 4\}, \{2, 5\}, \{3, 6\}, \emptyset, \Omega, \{1, 4, 2, 5\}, \{1, 4, 3, 6\}, \{2, 5, 3, 6\}, \Omega\}$ . Это все события, которые воспринимает Вася. Где-то объединение простейших событий, например  $\{1, 4, 2, 5\}$  - он, допустим не разглядел, что выпало, но знает, что остаток от деления у этого числа - не ноль.

- Самая богатая  $\sigma$ -алгебра:  $2^{\Omega}$
- Самая бедная  $\sigma$ -алгебра:  $\{\emptyset, \Omega\}$

## 7.8. Зачем нужна $\sigma$ -алгебра

Зачем?

- моделирование наделённостью информации
- технические цели в теоремах

Всегда существует самый маленький набор множеств, который порождает  $\sigma$ -алгебру, и поэтому очень часто записывают  $\mathbb{F} = \sigma(\text{набор событий})$

## 7.9. Определение вероятности

Вероятность  $\mathbb{P}$  просто задаётся функцией  $\mathbb{P} : \mathbb{F} \rightarrow [0; 1]$

И перед тем, как идти дальше, скажем, что  $A \sqcup B$  обозначает объединение, которое ещё говорит о том, что  $A \cup B = \emptyset$ , то есть  $A$  и  $B$  точно не пересекаются

Аксиомы Колмогорова

- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(A) = 1 - \mathbb{P}(\bar{A})$
- $\mathbb{P}(A \sqcup B) = \mathbb{P}(A) + \mathbb{P}(B)$
- $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

## 7.10. Упражнение

Подкидываем монетку:  $\Sigma = \{0, 1\}$  - выпал орёл/решка.

$$\mathbb{F} = \sigma(0, 1) = 2^\Sigma$$

$$\mathbb{P} : \mathbb{F} \rightarrow [0; 1]$$

Если монетка правильная, то  $\mathbb{P}(0) = \mathbb{P}(1) = \frac{1}{2}$ . Если монетка смещённая, то, например, может быть  $\mathbb{P}(0) = \frac{1}{4}$  и  $\mathbb{P}(1) = \frac{3}{4}$

## 7.11. Подытог

Тройка  $(\Omega, \mathbb{F}, \mathbb{P})$  задаёт модель.

## 7.12. Свойства вероятности

Правило сложения:  $\mathbb{P} = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

### 7.12.1. Упражнение

Есть кубик. Событие  $A$  - выпало чётное число, событие  $B$  - выпало число, которое делится на 3. Тогда

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(2 \cup 4 \cup 6) + \mathbb{P}(3 \cup 6) - \mathbb{P}(6) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$$

Вытекает, что  $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

### 7.12.2. Упражнение

Если три множества, то

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$$

## 7.13. Условная вероятность

Допустим мы знаем, что  $B$  точно произошло.

Чему тогда равно  $\mathbb{P}(A|B)$ ?

Как пример:  $A$  - кубик выпал на 2,  $B$  - выпало чётное

$$\text{Такая вероятность } \mathbb{P}(A|B) = \frac{|A \cap B|}{|B|} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

И в другом виде это записывается ещё как  $\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B) = \mathbb{P}(B|A) \cdot \mathbb{P}(A)$

## 7.14. Формула Байеса

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

## 7.15. Формула полной вероятности

Допустим мы "нарезали"  $\Omega = \bigcup_{i=1}^k H_k$ , при этом  $\forall i, j : H_i \cap H_j = \emptyset$ , то есть  $\Omega = \bigsqcup_{i=1}^k H_i$ , тогда

$$\sum_{i=1}^k \mathbb{P}(H_i) = 1$$

$$\text{При этом } \mathbb{P}(A) = \sum_{i=1}^k \mathbb{P}(A \cap H_i) = \sum_{i=1}^k \mathbb{P}(A|H_i) \cdot \mathbb{P}(H_i)$$

### 7.15.1. Упражнение

Есть школа, в которой три класса - А, Б и В. В них одинаково человек, но

- В классе А 30% учеников любят географию
- В классе Б 40% учеников любят географию
- В классе В 70% учеников любят географию

Мы взяли случайного ученика и спросили, любит ли он географию. Какова вероятность, что этот ученик любит географию?

$\mathbb{P}(\text{Петя} \heartsuit \text{географию}) = \mathbb{P}(A) \cdot \mathbb{P}(\heartsuit|A) + \mathbb{P}(B) \cdot \mathbb{P}(\heartsuit|B) + \mathbb{P}(V) \cdot \mathbb{P}(\heartsuit|V) = \frac{1}{3} \cdot 0.3 + \frac{1}{3} \cdot 0.4 + \frac{1}{3} \cdot 0.7 = \frac{14}{30} = \frac{7}{15}$ .  
Давайте посчитаем, какова вероятность, что Петя из класса А, Б или В если он любит географию:

- Из класса А:  $\mathbb{P}(A|\heartsuit) = \frac{\mathbb{P}(\heartsuit|A) \cdot \mathbb{P}(A)}{\mathbb{P}(\heartsuit)} = \frac{0.3 \cdot \frac{1}{3}}{\frac{7}{15}} = \frac{1}{10} \cdot \frac{15}{7} = \frac{3}{14}$
- Из класса Б:  $\mathbb{P}(B|\heartsuit) = \frac{\mathbb{P}(\heartsuit|B) \cdot \mathbb{P}(B)}{\mathbb{P}(\heartsuit)} = \frac{0.4 \cdot \frac{1}{3}}{\frac{7}{15}} = \frac{2}{15} \cdot \frac{15}{7} = \frac{2}{7}$
- Из класса В:  $\mathbb{P}(V|\heartsuit) = \frac{\mathbb{P}(\heartsuit|V) \cdot \mathbb{P}(V)}{\mathbb{P}(\heartsuit)} = \frac{0.7 \cdot \frac{1}{3}}{\frac{7}{15}} = \frac{7}{30} \cdot \frac{15}{7} = \frac{1}{2}$

## 7.16. Априорное и апостериорное распределение

Априорное распределение - это распределение вероятностей, которое мы получаем до эксперимента

Апостериорное распределение - это распределение вероятностей, которое мы получили после эксперимента

В нашем случае экспериментом было то, что мы спрашивали у Пети, любит ли он географию.

## 7.17. Байесовская вероятность в виде графа

Затехать граф

### 7.17.1. Аналитическая задача на формулу Байеса

Оценить долю спама в комментариях. Нужно было следить за тем, как после выкатки алгоритма изменится процент спама в комментах

Как измерялся спам?

- Толokersы - ошибаются много, но размечают тоже много
- Штатские модераторы - не ошибаются никогда

Мы можем взять выборку, чтобы её оценили толokersы, а потом после толokersов передать её модерам, чтобы они оценили качество толokersов.

Допустим доля спама в комментариях  $q$ , а доля спама, которую дали толokersы -  $\nu$

**Написать Сандомирской, как нормально затехать это**

Вероятность, когда модеры отметили как спам, а толokersы как не спам обозначим за  $P_1$ , а наоборот за  $P_2$ .

Тогда мы знаем, что  $\nu = (1 - q)p_2 + q(1 - p_1)$  (там когда граф затехаю будет видно).

И получается, что  $\mathbb{P}(\text{spam} | \text{Толока отметила spam}) = \frac{q(1-p_1)}{\nu} = \frac{q(1-p_1)}{q(1-p_1)+p_2(1-q)}$

## 8. Факультативная часть лекции 2

### 8.1. Случайные величины и борелевская $\sigma$ -алгебра

Случайная величина - функция, которая отображает пространство элементарных исходов  $\Omega$  на множество  $\mathbb{R}$   $X : \Omega \rightarrow \mathbb{R}$ , но на ней нет требований, какие есть в обычных вероятностях.

#### 8.1.1. Пример

При бросании орла и решки у нас уже не обязательны исходы {орёл, решка}, мы можем записать это как  $\{1, 0\}, \{-1, 1\}, \{42, 15\}$ .

После подбрасывания монетки какая мысль возникает? Если я знаю значение случайной величины  $X$ , то значит я понимаю какие события произошли.

### 8.2. Подготовка к Борелевской $\sigma$ -алгебре

Когда берут  $\sigma$ -алгебру, то часто говорят о том, что эта  $\sigma$ -алгебра порождается случайной величиной  $X$ :  $\sigma(X) = \sigma(\{X \leq t\} | t \in \mathbb{R})$  -  $\sigma$ -алгебра, порождённая случайной величиной  $X$

### 8.2.1. Упражнение

Почему мы можем все события так написать?

$$\{X^2 > 7\} \in \sigma(X)$$

Мы можем переписать это в виде  $X \in (-\infty; -\sqrt{7}) \cup (\sqrt{7}; +\infty)$

Левое событие выглядит как  $X < \sqrt{7}$ , правое как  $X > \sqrt{7}$ . Обозначим их как  $B$  и  $A$ .

Тогда  $\bar{A} = X \leq \sqrt{7} \in \sigma(X)$  по определению  $\Rightarrow A \in \sigma(X)$

$$B = \bigcup_{i=1}^{\infty} \frac{\{x \leq \sqrt{7} - \frac{1}{i}\}}{B_i}$$

Следует, что  $\forall B_i \in \sigma(X) \Rightarrow B \in \sigma(X)$

Почему в определении именно такие множества?

Ответ: Потому что такими событиями можно описать все события из борелевского  $\sigma$ -алгебры.

## 8.3. Борелевская $\sigma$ -алгебра

Борелевская  $\sigma$ -алгебра -  $\sigma$ -алгебра, порождённая открытыми подмножествами множества  $\Sigma$

### 8.3.1. Открытое множество

Открытое множество - все точки которого - внутренние.

### 8.3.2. Внутренние точки

Рассмотрим интервал  $(2; 4)$  и точку  $7 - (2; 4) \cup \{7\}$ . Это не открытое множество.

виды точек:

- Внутренняя точка  $A$  -  $\delta$  окрестность точки лежит в множестве:  $\exists U_\varepsilon(a) \subset A$
- Граничная точка  $A$  -  $\forall U_\varepsilon(a)$  верно, что  $U_\varepsilon \cap A \neq \emptyset$  и  $U_\varepsilon(a) \cap \bar{A} \neq \emptyset$
- Изолированная точка  $A$  -  $\exists U_\varepsilon(a) : U_\varepsilon \cap A = \{a\}$

### 8.3.3. Замкнутое множество

Замкнутое множество - дополнение к открытому.

## 8.4. Борелевское множество

Борелевское множество - любое множество из Борелевской  $\sigma$ -алгебры

### 8.4.1. Упражнение

Если у нас есть  $\mathbb{B}(\mathbb{R})$ , входит ли  $(2; 100) \in \mathbb{B}(\mathbb{R})$ ? Да, потому что оно открытое.

Будет ли  $[2; 100] \in \mathbb{B}(\mathbb{R})$ ? Да, потому что  $[2; 100] = \mathbb{R} \setminus ((-\infty; 2) \cup (100; +\infty))$ .

Будет ли  $\{7\} \in \mathbb{B}(\mathbb{R})$ ? Да, потому что  $\{7\} \in \mathbb{R} \setminus ((-\infty; 7) \cup (7; +\infty))$

Будет ли  $[2; 100) \in \mathbb{B}(\mathbb{R})$ ? Да, потому что их можно разбить на отрезки и интервалы и свести к предыдущим случаям.

Все привычное нам множества - Борелевские.

Не Борелевские множества существуют, но нам не привычны.

## 8.5. Измеримость

Говорят, что случайная величина  $X$  измерима относительно  $\sigma$ -алгебры  $\mathbb{F}$ , если  $\sigma(X) \subseteq \mathbb{F}$ .

На самом деле, при задании  $\Omega, (\Omega, \mathbb{F})$  - измеримое пространство.

Обычно любое множество из  $\mathbb{F}$  - измеримо.

Ещё говорят, что случайная величина  $Y$  измерима относительно случайной величины  $X$  - это значит, что  $\sigma$ -алгебры  $Y$  является подмножеством  $\sigma$ -алгебры  $X$ :  $\sigma(Y) \subseteq \sigma(X)$

Теорема:  $Y$  измеримо относительно  $X \iff Y = f(X)$

## 9. Лекция 3. Комбинаторика

Мы сказали на прошлой лекции, что  $\mathbb{P}(A) = \frac{|A|}{|\Omega|} \rightarrow$  надо научиться считать число элементов.

### 9.1. Правило умножения и сложения

$$A = \{a_1, \dots, a_n\}$$
$$B = \{b_1, \dots, b_m\}$$

Если мы хотим взять элемент из  $A$  или из  $B$ , то для этого есть  $n + m$  способов.

Если мы хотим взять 2 объекта, сначала из  $A$ , а потом из  $B$ , то для этого есть  $n \cdot m$  способов

#### 9.1.1. Упражнение

Допустим номер машины имеет вид CDDDC777, где C - буква, а D - цифра. Раз мы набираем по очереди, то мы получаем, что работает правило умножения. Допустим мы хотим получить номер, где из букв складывается АУЕ. Тогда таких вариантов будет  $1 \cdot 10 \cdot 10 \cdot 10 \cdot 1 \cdot 1 - 12^3 = 10^3 - 1$  (единичку вычли, потому номера с тремя нулями не бывает). Всех вариантов  $12^3 \cdot 10^3 - 12^3$ . Тогда  $\mathbb{P}(АУЕ) = \frac{10^3 - 1}{10^3 - 12^3 - 12^3}$

#### 9.1.2. Сколько существует пятизначных чисел?

На первую позицию может поставить только одну из 9 цифр (0 не можем), на оставшиеся любую из 10, тогда ответ:  $9 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 9 \cdot 10^4$

#### 9.1.3. Упражнение

Есть:

- $A$  - русский алфавит из 33 букв
- $B$  - 10 цифр
- $C$  - английский алфавит из 26 букв

Сколько строк длины 8 мы можем сгенерировать, используя все эти символы?

Алфавит, из которого мы можем набирать символы, имеет мощность  $33 + 10 + 26 = 69$ , и соответственно ответ:  $69^8$

### 9.2. Принцип Дирихле

Если у нас есть  $n$  ящиков и  $n + 1$  кролик, тогда хотя бы в одном ящике будет сидеть хотя бы 2 кролика.

#### 9.2.1. Упражнение

В ящике есть 6 чёрных и 6 белых носков. Сколько минимально нужно достать носков, чтобы получилась хотя бы одна пара носков? Если мы достанем 3 носка, то как минимум 2 из них будут одного цвета и будет пара.

#### 9.2.2. Упражнение

В ящике есть 101 пуговица, каждая из которых 1 из 11 цветов.

Доказать: Найдётся либо 11 пуговиц одного цвета, либо 11 пуговиц разного цвета.

Предположим, что нет 11 пуговиц одного цвета и нет 11 различных цветов.

Тогда одного цвета существует не более 10 пуговиц и существует не более 10 цветов. Максимально таких пуговиц будет  $10 \cdot 10 = 100$ . Тогда нам нужно положить ещё одну пуговицу. Но если мы положим пуговицу существующего, то будет 11 пуговиц одного цвета. Если положим пуговицу цвета, которого ещё не было - то будет 11 различных цветов.

### 9.3. Размещения

Существуют двух видов:

- Без повторений  $A_n^m$
- С повторениями  $\overline{A}_n^m$

Допустим есть  $A = \{a_1, \dots, a_n\}$ , мы хотим отобрать  $m$  штук, и при этом порядок важен. На первое место мы можем поставить один объект из  $n$ . На второе место один объект из  $n-1$  объектов. И так далее, получим, что кол-во способов это сделать -  $n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-m+1) = \prod_{i=n-m+1}^n i$  и это размещения без повторений

Если мы говорим, что элементы могут повторяться, то на любое место мы можем поставить любой из  $n$  элементов и по правилу умножения кол-во таких вариантов -  $n^m$

#### 9.3.1. Упражнение

Есть 16 команд. Каким количеством способов могут распределиться среди них первые три места? Мы можем поставить какую-то команду на первое место, из оставшихся на второе, и из оставшихся на третье. Это Размещения с повторениями. Ответ:  $16 \cdot 15 \cdot 14$

#### 9.3.2. Упражнение

Есть  $B = \{1, 2, 3, 4\}$ . Сколько существует трёхзначных чисел, составленных из этих цифр? Это размещения с повторениями и ответ  $4^3$ .

### 9.4. Перестановки

Перестановки также бывают

- Без повторений  $A_n^n = P(n) = n!$
- С повторениями  $P(n_1, \dots, n_k) = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}$ , где  $n = \sum_{i=1}^k n_i$

#### 9.4.1. Упражнение

Допустим у вас есть 40 книг и вас спрашивают сколькими способами вы можете расставить их на полке. Это есть ни что иное. как  $A_{40}^{40} = P(40) = 40!$

#### 9.4.2. Упражнение

Есть

- 3 кебаба
- 4 лукума
- 2 пахлав

Сколько есть способов ими пообедать, если обеды между собой отличаются только порядком поглощения блюд? Это  $\frac{(3+4+2)!}{3! \cdot 4! \cdot 2!}$

### 9.5. Сочетания

Тоже бывают

- Без повторений  $C_n^m = \frac{n!}{m!(n-m)!}$
- С повторениями  $\overline{C}_n^m = C_{n+m-1}^m$

#### 9.5.1. Упражнение

Допустим есть  $A = \{a_1, \dots, a_n\}$  и нужно взять оттуда  $m$  объектов, но порядок не важен. Если мы возьмём  $A_n^m$ , то получим набор из  $m$  объектов, но у них порядок важен. Чтобы избавиться от порядка - поделим на  $m!$  - то, как внутри могут располагаться объекты. Тогда это будет  $\frac{n!}{m!(n-m)!}$



### 9.5.2. Упражнение

Есть два друга. У одного друга 9 уникальных монет, а у второго 7 уникальных марок. Им захотелось обменяться. 2 монеты можно обменять в 2 марки. Сколько способов таких обменов существует?  $C_9^2 \cdot C_7^2$

### 9.5.3. Упражнение

Есть магазин с 4 видами пироженных. Хочется купить 7 штук.

Закодируем покупки 1101110011 - мы купили  $1+1 = 2$  пирожных первого вида, потом поставили 0 (сказали, что их мы больше не покупаем). Дальше купили  $1+1+1 = 3$  пирожных второго вида и поставили 0. Дальше стоит 0, потому что пирожных третьего вида мы вообще не купили, и пирожных четвертого вида мы купили 2 штуки. Тогда это сводится к задаче: "как среди 10 символов выбрать 3, которые будут нулями".  $\frac{10!}{7!3!} = C_{10}^7 = C_{4+7-1}^7 = \frac{10!}{7!3!}$

### 9.5.4. Упражнение

Есть остановка, на которой стоит 10 человек.

Приезжает поезд, в котором есть 10 вагонов.

Все вагоны и люди - разные (различимы между собой).

$\mathbb{P}(1 \text{ вагон пустой}) = \frac{m}{n}$

$$n = 10!0 = \overline{A}_{10}^{10}$$

$m = 10! \cdot 9 \cdot 10 = \text{пассажиры идут в вагоны} \cdot \text{в какой вагон идут} \cdot \text{выбрать пустой вагон}$

**Может быть не до конца затежал**

## 10. Лекция 4. Разлагай и влавствуй, бином, беспорядки

### 10.1. Дискретные случайные величины

Случайная величина - функция, которая делает отображение  $X : \Omega \rightarrow \mathbb{R}$

Случайные величины тоже делятся на

- Дискретные:  $\Omega \rightarrow A$ , где  $A$  - счётное
- Непрерывные:  $A$  - континуальное, и существует  $\exists f_x(t)$  - функция плотности распределения
- Другие (смешанные, говорить не будем - на практике бесполезные)

#### 10.1.1. Упражнение

Вася подкидывал монетку

- Если выпал орёл - записывал 0
- Если выпала решка - записывал 1

Монетка честная, то есть

$X$	0	1
$\mathbb{P}(x = k)$	$\frac{1}{2}$	$\frac{1}{2}$

Это называется закон распределения и это один из способов задания дискретной случайной величины. Потом после Васи пришёл ещё Игорь, но он уже записывал  $-1$  и  $1$  вместо  $0$  и  $1$

$X$	-1	1
$\mathbb{P}(y = k)$	$\frac{2}{3}$	$\frac{1}{3}$

Как можем видеть - у Игоря монетка нечестная

### 10.2. Манифест

- $X, Y, Z$  - случайные величины
- $x, y, z$  - значения
- $A, B, C, D$  - события  $\mathbb{E}$  - математическое ожидание  $Var, \mathbb{D}$  - дисперсия  $Cov, Corr$  - ковариация и корреляция

### 10.3. Операции над случайными величинами

Как будут выглядеть случайные величины:

- $Z = X \cdot Y$
- $W = X + Y$
- Совместное распределение  $(X, Y)$  - они могут быть зависимыми, независимыми

#### 10.3.1. Умножение случайных величин

Какие значения случайных величин  $Z$  могут быть?

$X$	$Y$	$X \cdot Y$
0	-1	0
0	1	0
1	-1	-1
1	1	1

$Z$	0	-1	1
$\mathbb{P}(z = k)$	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

### 10.3.2. Сложение случайных величин

Каковы могут быть результат  $W$ ?

$X$	$Y$	$X + Y$
0	-1	-1
0	1	1
1	-1	0
1	1	2

### 10.3.3. Совместное распределение случайных величин

$Y/X$	0	1
-1	$\frac{2}{6}$	$\frac{2}{6}$
1	$\frac{1}{6}$	$\frac{1}{6}$

## 10.4. Бином Ньютона

плз, очень тривиально, почитайте где-нибудь, я просто даже не знаю как кратко расписать

## 10.5. Распределение Бернулли

Распределением Бернулли называют дискретную случайную величину, у которой

$X$	0	1
$\mathbb{P}(x = k)$	$1 - p$	$p$

с параметром  $p$

Если какое-то  $X$  распределено по параметру Бернулли, то пишут  $X \sim \text{Bern}(p)$  или  $X \sim B(p)$

### 10.5.1. Упражнение

$X$	-1	0	1
$\mathbb{P}(x = k)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{2}{4}$

Нужно найти:

- $\mathbb{E}(X)$
- $\text{Var}(X)$
- $\sigma(X) = \sqrt{\text{Var}(X)}$

Для дискретной величины:

$$\mathbb{E} = \sum_{k=1}^n \mathbb{P}(x = k) \cdot k$$

И в нашем случае

$$\mathbb{E} = \frac{1}{4} \cdot (-1) + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 = \frac{1}{4}$$

### 10.5.2. Дисперсия

Дисперсия - это мат ожидание отклонения случайной величины от математического ожидания в квадрате:

$$\mathbb{E}[(x - \mathbb{E})^2] = \mathbb{E}(x^2) - \mathbb{E}^2(x)$$

$$\text{Var}(x) = \sum_{k=1}^n \mathbb{P}(x = k) \cdot (k - \mathbb{E}(x))^2$$

Для нашей задачи:  $\text{Var}(x) = \frac{1}{4}(-1 - \frac{1}{4})^2 + \frac{1}{4}(0 - \frac{1}{4})^2 + \frac{1}{2}(1 - \frac{1}{4})^2$

$$\mathbb{E}(x^2) = \frac{1}{4}(-1)^2 + \frac{1}{4} \cdot (0)^2 + \frac{1}{2} \cdot 1^2 = \frac{3}{4}$$

$$\mathbb{E}^2(x) = \left(\frac{1}{4}\right)^2 = \frac{1}{16}$$

$$\text{Var}(x) = \frac{12}{16} - \frac{1}{16} = \frac{11}{16}$$

$$\sigma(x) = \sqrt{\text{Var}(x)} = \frac{\sqrt{11}}{4}$$

## 10.6. Биномиальное распределение

$Y \sim \text{Bin}(n, p)$ , если  $n$  - число испытаний, а  $p$  - вероятность успеха.

Как в данном случае найти  $\mathbb{E}(Y)$ ?

В нашем случае  $y = x_1 + x_2 + \dots + x_n$ , где

$$\frac{x_i}{\mathbb{P}(x_i = k)} \mid \begin{array}{c|c} 0 & 1 \\ \hline 1-p & p \end{array}$$

Где  $x_1, x_2, \dots, x_n \sim iid$  - независимые случайные величины: independent identical distributed.

В случае  $iid$  -  $\mathbb{E}(\alpha \cdot X + b \cdot Y) = \alpha \cdot \mathbb{E}(x) + b \cdot \mathbb{E}(y)$

И  $\text{Var}(\alpha \cdot X + Y) = \alpha^2 \cdot \text{Var}(X) + \text{Var}(Y) + 2\alpha \cdot \text{Cov}(X, Y)$

### 10.6.1. Ковариация (Cov)

$\text{Cov}(X, Y) = \mathbb{E}[(Y - \mathbb{E}(y)) \cdot (X - \mathbb{E}(X))] = \mathbb{E}(Y \cdot X) - \mathbb{E}(Y) \cdot \mathbb{E}(X)$  - смотрят на сколько  $Y$  отклоняется от своего мат ожидания, а  $X$  от своего.

### 10.6.2. Корреляция (Corr)

$\text{Corr}(X, y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)} \in [-1; 1]$  - здесь уходят размерности.

Ну и  $\mathbb{E}(X \cdot Y) = \mathbb{X} \cdot \mathbb{Y} + \text{Cov}(X, Y)$

А  $\text{Var}(X \cdot Y)$  было дано в качестве упражнения и автор когда-нибудь выразит это

Очевидно, что  $\text{Var}(X) \geq 0$  и  $\text{Var}(X) = 0 \iff \mathbb{P}(x = k) = 1$

$\mathbb{E}(Y) = \mathbb{E}(x_1 + \dots + x_n) = n \cdot \mathbb{E}(x_i) = n \cdot p$

## 10.7. Формула фключений исключений

Тоже из школы - если кто хочет затехать - пишите @math\_hse\_boy

## 10.8. Беспорядки (субфакториал) (!n или $D_n$ )

Игра киллер: есть игра, где каждому в конверте написано имя, кого он должен встретиться в лагере  $1 \times 1$  и "убить". Каждый одновременно и охотник, и жертва.

Играем не до последнего выжившего, а на наибольшее кол-во баллов (убийств).

То есть если  $n$  участников, то в начале есть  $n$  конвертов. Проблема заключается в том, что кто-то вполне вероятно может получить себя в конверте после убийства своей цели ( $X$ ). Назовём таких людей суицидниками. Вопрос: каково  $\mathbb{X}$ ?. В самом начале это  $\mathbb{E}(X) = \mathbb{E}(Y_1 + \dots + Y_n) = n \cdot \frac{1}{n} = 1$ , то есть в среднем в начале игры один суицидник будет сразу.

Сколько существует раскладок без суицидников? Это кол-во и называется **беспорядками**.

Ну тут по формуле включений-исключений. Sorry.

Но, если  $e^x$  разложить в ряд Тейлора, то  $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots = \sum_{i=1}^{\infty} \frac{x^i}{i!}$ ,

тогда  $e^{-x} = 1 - \frac{x}{1!} + \frac{x^2}{2!} - \frac{x^3}{3!} + \dots = \sum_{i=0}^{\infty} \frac{(-1)^i \cdot x^i}{i!}$

Тогда итоговая вероятность, что суицидников не будет в самом начале, будет  $\mathbb{P}(\alpha'_1, \dots, \alpha'_n) = \frac{100! \cdot e^{-1}}{100!} = \frac{1}{e}$

А более формально  $D_n = n!(1 - \frac{1}{1!} + \frac{1}{2!} - \dots + \frac{(-1)^n}{n})$

И  $\lim_{n \rightarrow \infty} \frac{D_n}{n!} = e^{-1}$

### 10.8.1. Свойства беспорядков

- $D_{n,k}$  обозначает беспорядки из  $n$  элементов, где ровно  $k$  в порядке. Тогда  $D_{n,k} = C_n^k \cdot D_{n-k}$
- $(D_n - n \cdot D_{n-1}) = -(D_{n-1} - (n-1)D_{n-2})$
- $D_n = (n-1) \cdot (D_{n-1} + D_{n-2})$

### 10.8.2. Доказать

Чему равно  $\sum_{k=0}^n D_{n,k}$ , автор ещё не делал.

## 11. Цепи Маркова и метод первого шага

### 11.1. Упражнение

Есть игрок Маша и Саша. Они После тусы решают кто пойдёт мыть посуду. Если кто-то выкидывает шестёрку, то на этом игра прекращается, и этот игрок идёт мыть посуду. Если на кубике выпало другое число, то ребята продолжают подкидывать кубик, пока у кого-то не выпадет 6.

Пусть Саша подкидывает кубик первым. С какой вероятностью Саша пойдёт мыть посуду?

**Затехать состояния**

Получается система уравнений:

$$\begin{cases} P_{1 \rightarrow 3} = \frac{1}{6} \cdot 1 + \frac{5}{6} \cdot P_{2 \rightarrow 3} \\ P_{2 \rightarrow 3} = \frac{1}{6} \end{cases}$$

Тогда, решая системы, получаем, что у нас  $P_{1 \rightarrow 3} = \frac{1}{6} + \frac{25}{36} \cdot P_{1 \rightarrow 3}$

И это  $\frac{11}{36} P_{1 \rightarrow 3} \Rightarrow P_{1 \rightarrow 3} = \frac{6}{25}$

Очевидный вопрос - почему игра не получится бесконечной? Вероятность того, что игра не остановится на  $n$ -м шаге  $= \left(\frac{5}{6}\right)^n \xrightarrow{n \rightarrow \infty} 0$

### 11.2. Упражнение (Амёба)

Что происходит с Амёбой? **Затехать состояния**

Какова вероятность, что  $\mathbb{P}(\text{Популяция амёб умрёт} = A)$

$$\mathbb{P}(A) = \frac{1}{4} \cdot 1 + \frac{1}{4} \mathbb{P}(A) + \frac{1}{2} \cdot \mathbb{P}(A) \cdot \mathbb{P}(A)$$

$$x = \frac{1}{4} + \frac{1}{4}x + \frac{1}{2}x^2$$

$$\frac{1}{2}x^2 - \frac{3}{4}x + \frac{1}{4} = 0$$

$$2x^2 - 3x + 1 = 0$$

$$\text{Тогда } D = 9 - 8 = 1$$

$$\text{И } x_1 = \frac{3+1}{4} = 1$$

$$\text{И } x_2 = \frac{3-1}{4} = \frac{1}{2}$$

### 11.3. Геометрическое распределение

$X \sim \text{Geom}(p)$  называется, если эквивалентно задаче:

$X$  - число подкидывание монетки до первого орла (само подкидывание орла - не включается), с вероятностью выпадения какой-то стороной  $p$  и другой  $1 - p$ .

#### 11.3.1. Упражнение

А как тогда найти при  $X \sim \text{Geom}(p)$   $\mathbb{E}(X)$ ?

С вероятностью  $p$  мы заканчиваем подбрасывание. С вероятностью  $1 - p$  будем подбрасывать как минимум ещё один раз.

Строим уравнения:

$$\mathbb{E}(X) = p \cdot 1 + (1 - p) \cdot \mathbb{E}(X + 1)$$

$$\mathbb{E}(X) = p + (1 - p) \cdot (e + 1)$$

$$\mathbb{E}(X) = 1 + (1 - p) \cdot e$$

$$\mathbb{E}(X) - \mathbb{E}(X) + p \cdot e = 1$$

И приходим к выводу, что  $\mathbb{E}(X) = \frac{1}{p}$

И при  $p = \frac{1}{2}$  получаем  $\mathbb{E}(X) = 2$

Как найти  $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$ ?

$$\mathbb{E}(X^2) = p + \mathbb{E}(X^2) + 2\mathbb{E}(X) + 1 - p\mathbb{E}(X^2) - 2p\mathbb{E}(X) - p = \mathbb{E}(X^2) + 2\mathbb{E}(X) + 1 - p\mathbb{E}(X^2) - 2p\mathbb{E}(X)$$

Тогда вычтем с каждой стороны  $\mathbb{E}(X^2)$  и получим  $2\mathbb{E}(X) + 1 - p\mathbb{E}(X^2) = 0 = 2 \cdot \frac{1}{p} + 1 - p\mathbb{E}(X^2) - 2p \cdot \frac{1}{p} \rightarrow$

$$p\mathbb{E}(X^2) = \frac{2}{p} + 1 - 2 = \frac{2}{p} - 1$$

$$\text{И тогда } \mathbb{E}(X^2) = \frac{2}{p^2} - \frac{1}{p} = \frac{2-p}{p^2}$$

$$\text{И тогда } \text{Var}(X) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

### 11.3.2. Коллекционер

Есть киндер-сюрпризы. В одной коллекции 30 игрушек. С каждого яйца вероятность выпадения конкретной игрушки  $\frac{1}{30}$

$X$  - кол-во шоколадок для покупки, чтобы собрать всю коллекцию.

Тогда каково  $\mathbb{E}(X)$ ?

Пусть представим  $X = X_1 + X_2 + \dots + X_{30}$

Где  $X_i$  - число яиц до выпадения  $i$  уникальных игрушек.

Мы понимаем, что матожидание  $\mathbb{E}(X_1) = 1$ , так как при покупке первого яйца мы гарантировано получаем новую игрушку.

Тогда, так как мы уже что-то знаем про геометрическое распределение,  $\mathbb{E}(X_2) = \frac{29}{30} \cdot 1 + \frac{1}{30} \cdot \mathbb{E}(X_2 + 1)$

И получится  $\mathbb{E}(X_2) = \frac{30}{29}$

По аналогии  $\mathbb{E}(X_3) = \frac{30}{28}$

Выписываем ещё 27 уравнений и складываем всё.

$$X = 1 + \frac{30}{29} + \frac{30}{28} + \frac{30}{27} + \dots + \frac{30}{2} + \frac{30}{1} = 30 \left( \frac{1}{30} + \frac{1}{29} + \frac{1}{28} + \dots + \frac{1}{2} + 1 \right) = 30 \sum_{i=1}^{30} \frac{1}{i} \sim 30 \ln 30 \sim 102$$

Если яйцо стоит 30 рублей и в магазине только 1 коллекция, то надо примерно  $3k$  рублей.

### 11.4. Марковская цепь

Марковская цепь - это последовательность случайных величин  $X_0, X_1, X_2, \dots$

И второе условие -  $\mathbb{P}(X_{n+1} = k | \underbrace{X_n, X_{n-1}, X_{n-2}, \dots, X_1, X_0}_{\mathcal{F}_n}) = \mathbb{P}(X_{n+1} = k | X_n)$

И помимо есть ещё начальное состояние в котором мы находимся  $\mathbb{P}(X_0 = k) = \lambda_k$

#### 11.4.1. Теорема об Эргодичности конечной цепи Маркова

Для конечной цепи Маркова есть матрица состояний, которая описывает вероятность перехода из одного состояния в другое.

$$\begin{pmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,n-1} & p_{1,n} \\ p_{2,1} & p_{2,2} & \dots & p_{2,n-1} & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{n-1,1} & p_{n-1,2} & \dots & p_{n-1,n-1} & p_{n-1,n} \\ p_{n,1} & p_{n,2} & \dots & p_{n,n-1} & p_{n,n} \end{pmatrix}$$

$\mathbb{P}(X = j) = \pi_j$  - вероятность в бесконечном блуждании по состояниям оказаться в состоянии  $j$ . Называется это стационарным распределением.

Теорема говорит о том, что если цепь Маркова имеет конечное количество состояний и для какого-то  $n_0$  все элементы матрицы  $P^{n_0}$  отличны от нуля, тогда  $\forall i, j \exists \lim_{n \rightarrow \infty} p_{i,j}^n = \pi_j > 0$

Вероятности  $\pi_j$  являются наименьшим решением системы уравнений.

При этом:

- $\sum_j \pi_j = 1$
- $\pi_j = \sum_i \pi_i \cdot p_{i,j}$

#### 11.4.2. Упражнение

**Затехать состояния**

$$x_0 = \begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix}, P = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix}$$

$$\text{Тогда } x_1^T = x_0^T \cdot P = (0.3 \quad 0.7) \cdot \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} = (0.3 \cdot 0.3 + 0.7 \cdot 0.6 \quad 0.3 \cdot 0.7 + 0.7 \cdot 0.4) = (0.51 \quad 0.49)$$

$$\text{И в общем случае получаем } x_{t+1}^T = x_t^T \cdot P$$

$$\text{Это можно представить в виде } x_{t+1}^T = \overbrace{x_{t-1}^T \cdot P \cdot P}^{x_t}$$

И можем продолжить так редуцировать

$$x_{t+1}^T = \overbrace{\overbrace{x_{t-2}^T \cdot P \cdot P}^{x_{t-1}}}^{x_t} \cdot P \cdot P \cdot P$$

И мы можем точно также продолжать редуцировать и получить

$$x_{t+1}^T = x_0^T \cdot P^{t+1}$$

Из линейной алгебры знаем, что матрицу можно представить в виде  $C^{-1} \cdot D \cdot C$ , где  $D$  - диагональная из собственных значений, а  $C$  состоит из собственных векторов.

Как быстро возводить  $P^n$ ?  $P^n = \underbrace{C \cdot D \cdot \overbrace{C^{-1} \cdot C}^{=E} \cdot D \cdot C^{-1} \dots C \cdot D \cdot C^{-1}}_{n \text{ раз}} = C \cdot D^n \cdot C^{-1}$ , где чтобы полу-

чить  $D^n$  нужно просто диагональные элементы возвести в степень  $n$ .

Собственные значения в нашем случае  $\lambda_1 = 0.3, \lambda_2 = 1$ .

А собственные векторы  $\begin{pmatrix} 6 \\ 7 \end{pmatrix}$  и  $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$

И тогда наша матрица  $P = \begin{pmatrix} 6 & 1 \\ 7 & -1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 & 0 \\ 0 & -0.3 \end{pmatrix} \cdot \begin{pmatrix} 6 & 1 \\ 7 & -1 \end{pmatrix}$

А  $P^n = \begin{pmatrix} 6 & 1 \\ 7 & -1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1^n & 0 \\ 0 & (-0.3)^n \end{pmatrix} \cdot \begin{pmatrix} 6 & 1 \\ 7 & -1 \end{pmatrix} = \begin{pmatrix} 6 & 1 \\ 7 & -1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 6 & 1 \\ 7 & -1 \end{pmatrix} =$  **Досчитаю позже**  
**и конец лекции перепишу**

И тогда у нас получается  $x_\infty^T = x_0^T \cdot P^\infty = \begin{pmatrix} 0.3 & 0.7 \end{pmatrix} \cdot \begin{pmatrix} \frac{6}{13} & \frac{6}{13} \\ \frac{7}{13} & \frac{7}{13} \end{pmatrix} = \begin{pmatrix} \frac{3}{10} \cdot \frac{6}{13} + \frac{7}{10} \cdot \frac{7}{13} & \frac{3}{10} \cdot \frac{6}{13} + \frac{7}{10} \cdot \frac{7}{13} \end{pmatrix}$

## 12. Непрерывные случайные величины, плотность распределения

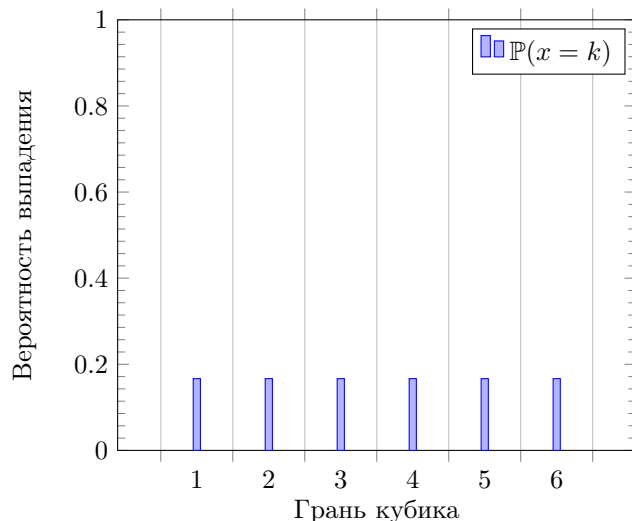
### 12.1. Функция распределения

Мы говорили, что клёво было бы найти такие функции, которые  $X : \Omega \rightarrow \mathbb{R}$  и дальше с ними работать.

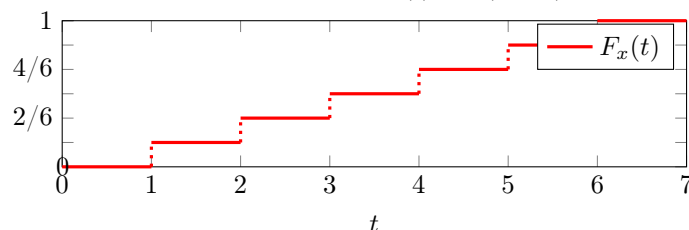
Что это за функции такие?

Допустим у нас есть кубик:

$X$	1	2	3	4	5	6
$\mathbb{P}(x = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$



**Функция распределения** -  $F_x(t) = \mathbb{P}(t \leq x)$  В нашем случае **Передать тех легенды**



#### 12.1.1. Свойства функции распределения

- $0 \leq F_x(x) \leq 1$
- $\forall x_1 \leq x_2 : F_x(x_1) \leq F_x(x_2)$
- $F_x(X)$  непрерывна справа
- Для непрерывных функций распределения:  $\lim_{x \rightarrow -\infty} F_x(x) = 0, \lim_{x \rightarrow +\infty} F_x(x) = 1$

### 12.2. Непрерывная случайная величина. Плотность

Случайная величина  $X$  называется непрерывной, если  $\exists f_x(t) \geq 0 \forall t : F_x(t) = \int_{-\infty}^t f_x(t)dt$  и эта функция называется плотностью

#### 12.2.1. Свойства плотности

- $f_x(t) \geq 0$
- $\int_{-\infty}^{+\infty} f_x(t)dt = 1$
- $\mathbb{P}(a \leq x \leq b) = \int_a^b f_x(t)dt = F_x(b) - F_x(a)$
- $\mathbb{P}(x = a) = \int_a^a f_x(t)dt = 0$



### 12.3. Упражнение

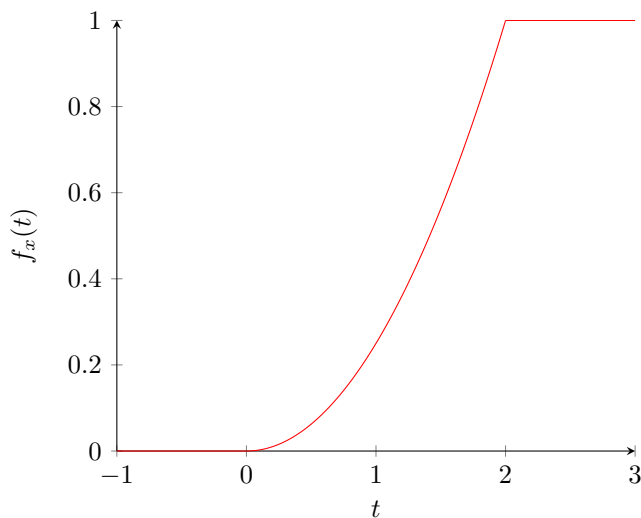
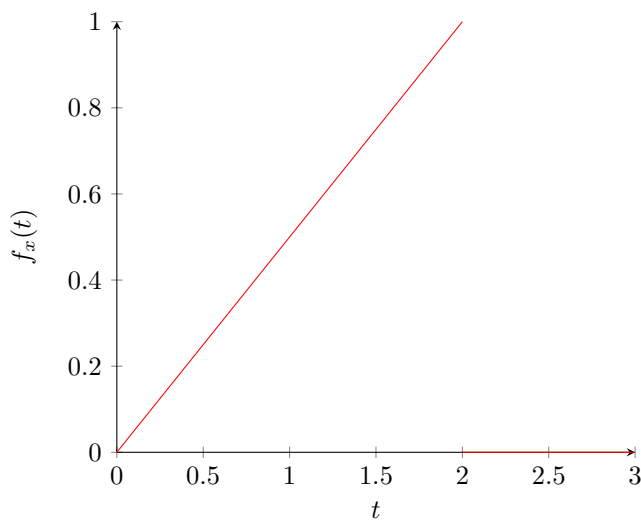
Случайная величина  $X$  распределена на отрезке  $[0; 2]$  с функцией распределения  $F_x(t) = c \cdot t^2$ .  
Что хотим найти?

1.  $c$
2.  $f_x(t)$
3.  $\mathbb{P}(x < 1)$
4.  $\mathbb{P}(x \in [1; 1.5])$
5.  $F_x(-5), F_x(10)$
6.  $\mathbb{E}(X), \text{Var}(X)$

Находим:

1. Знаем, что  $f_x(t) = F'_x(t) = 2ct$ . Давайте тогда решим наше уравнение, зная что  $\int_0^2 2ct \, dt = 2c \frac{t^2}{2} \Big|_0^2 = 1 = 2c2 - 0 = 4c = 1$ . Тогда получаем, что  $c = \frac{1}{4}$ .
2. Тогда получается, что

$$f_x(t) = \begin{cases} \frac{t}{2} & t \in [0; 2] \\ 0 & t \notin [0; 2] \end{cases}$$



$$3. \mathbb{P}(x < 1) = \mathbb{P}(x \leq 1) = F(1) = \int_0^1 f(t)dt = \frac{1}{4}(1 - 0) = \frac{1}{2}$$

$$4. \mathbb{P}(1 \leq x \leq 1.5) = F(1.5) - F(1) = \frac{1.5^2}{4} - \frac{1^2}{4} = \frac{2.25-1}{4} = \frac{1.25}{4}$$

$$5. F(-5) = 0, F(10) = 1$$

$$6. \mathbb{E}(X) = \int_{-\infty}^{+\infty} f_x(t) \cdot t \, dt = \int_0^2 \frac{t^2}{2} dt = \frac{1}{2} \cdot \frac{t^3}{3} \Big|_0^2 = \frac{1}{6}(2^3 - 0^3) = \frac{8}{6} = \frac{4}{3}$$

$$\text{А } \mathbb{E}(X^2) = \int_{-\infty}^{+\infty} f_x(t) \cdot t^2 \, dt = \int_0^2 \frac{t^3}{2} dt = \frac{1}{2} \cdot \frac{t^4}{4} = \frac{1}{8}(2^4 - 0^4) = \frac{16}{8} = 2$$

$$\text{И } \text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = 2 - \frac{16}{9} = \frac{2}{9}$$

## 12.4. Преобразования случайных величин

Есть экспоненциальная случайная величина  $X \sim \text{Exp}(\alpha)$ , если

$$F_x(t) = \begin{cases} 1 - e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

### 12.4.1. Упражнение

Найти:

1.  $f_x(t)$
2.  $\mathbb{E}(X)$
3.  $\text{Var}(X)$

Найдём:

1.

$$f_x(t) = F'_x(t) = \begin{cases} \alpha \cdot e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

$$2. \mathbb{E}(X) = \int_0^{+\infty} t \cdot \alpha \cdot e^{-\alpha t} dt = (\text{интегрирование по частям}) = \int_0^{+\infty} -t de^{-\alpha t} = -te^{-\alpha t} \Big|_0^{+\infty} - \int_0^{+\infty} -e^{-\alpha t} dt =$$

$$\lim_{t \rightarrow \infty} \frac{t}{e^{\alpha t}} - \lim_{t \rightarrow 0} \frac{t}{e^{\alpha t}} - \frac{1}{\alpha} e^{-\alpha t} \Big|_0^{+\infty} = 0 - 0 - \frac{1}{\alpha} (e^{-\infty} - e^0) = -\frac{1}{\alpha}$$

$$3. \mathbb{E}(X^2) = \int_0^{+\infty} t^2 \cdot \alpha \cdot e^{-\alpha t} dt = -\frac{2}{\alpha^2} \text{ Тут где-то бага, должно быть без минуса}$$

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \frac{2}{\alpha^2} - \frac{1}{\alpha^2} = \frac{1}{\alpha^2}$$

### 12.4.2. Упражнение на преобразование экспоненциального распределения

$X \sim \text{Exp}(1)$

Как будет распределена

1.  $Y = X^2 - f_y(t)?$
2.  $Z = \sqrt{X} - f_z(t)?$

Такие преобразования называются функциями случайного аргумента.

Сделать это очень просто - нужно воспользоваться функциями распределения и воспользоваться их свойствами:

$$1. F_y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(X^2 \leq t) = \mathbb{P}(-\sqrt{t} \leq X \leq \sqrt{t}) = F_x(\sqrt{t}) - \overbrace{F_x(-\sqrt{t})}^{=0 \text{ из-за определения}} = 1 - e^{-\sqrt{t}}$$

$$\text{А плотность будет } f_y(t) = F'_y(t) = [F_x(\sqrt{t})]' = \frac{1}{2\sqrt{t}} e^{-\sqrt{t}}$$

$$2. F_z(t) = \mathbb{P}(Z \leq t) = \mathbb{P}(\sqrt{X} \leq t) = \mathbb{P}(X \leq t^2) = F_x(t^2) = 1 - e^{-(t^2)}.$$

$$\text{А плотность будет } f_z(t) = F'_z(t) = [F_x(t^2)]' = 2t \cdot F'_x(t) = 2t f_x(t) = 2te^{-t}$$

Дальше можно сделать с  $Z$  что угодно

### 12.4.3. Упражнение на преобразование равномерного распределения

$X \sim U([a, b])$ , если

$$f_x(t) = \begin{cases} \frac{1}{b-a} & t \in [a; b] \\ 0 & t \notin [a; b] \end{cases}$$

и

$$F_x(b) = \int_{-\infty}^b f_x(t) dt = \begin{cases} \frac{t}{b-a} & t \in [a; b] \\ 0 & t \leq a \\ 1 & t > b \end{cases}$$

Найдём:

1.  $\mathbb{E}(X)$
2.  $Var(X)$

Рассчёт:

$$1. \mathbb{E}(X) = \int_a^b t \frac{1}{b-a} dt = \frac{1}{b-a} \int_a^b t dt = \frac{1}{2(b-a)} t^2 \Big|_a^b = \frac{1}{2(b-a)} (b^2 - a^2) = \frac{(b-a)(a+b)}{2(b-a)} = \frac{a+b}{2}$$

$$2. \mathbb{E}(X^2) = \int_a^b t^2 \frac{1}{b-a} dt = \frac{1}{b-a} \int_a^b t^2 dt = \frac{1}{3(b-a)} t^3 \Big|_a^b = \frac{1}{3(b-a)} (b^3 - a^3) = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

$$\text{И тогда } Var(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{4(b^2+ab+a^2)-3(b^2+2ab+b^2)}{12} = \frac{b^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}$$

### 12.5. Теорема. Квантильное преобразование

Допустим есть  $X \sim Exp(\alpha)$

Теперь возьмём такую  $Y = F_x(X) = 1 - e^{-\alpha X}$ . Чему равно распределение случайной величины  $Y$ ?

$$F_y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(1 - e^{-\alpha X} \leq t) = \mathbb{P}(e^{-\alpha X} \geq 1 - t) = \mathbb{P}(-\alpha X \geq \ln(1 - t)) = \mathbb{P}(X \leq -\frac{1}{\alpha} \ln(1 - t)) = F_x(-\frac{1}{\alpha} \ln(1 - t)).$$

Прежде чем переходить дальше, надо сказать, что у нас ломается область определения: для  $Dom(X) = [0; +\infty)$ ,  $Dom(Y) = [1 - e^{-0}; 1 - e^{-\infty}] = [0; 1]$

Теперь преобразовываем дальше:  $F_x(-\frac{1}{\alpha} \ln(1 - t)) = 1 - e^{\ln(1-t)} = 1 - 1 + t = t$ .

Тогда  $Y \sim U([0; 1])$

А что теперь, если взять  $X \sim F_\theta$ , где  $F_\theta$  - любая функция распределения. А  $Y$  всё так же  $Y \sim F_x(X)$ ? Это всегда будет  $\sim U[0; 1]$  и это называется квантильным преобразованием.

$$F_y(t) = \mathbb{P}(y \leq t) = \mathbb{P}(F_x(x) \leq t) = \mathbb{P}(x \leq F_x^{-1}(t)) = F_x(F_x^{-1}(t)) = t$$

Почему это хорошо? Компьютер хорошо генерирует равномерное распределение. Тогда можно получить произвольное распределение из равномерного:

1. Генерируем  $X \sim U([0; 1])$
2. Генерируем  $F_y^{-1}(x)$  (Для примера если хотим экспоненциальное распределение:  $y_i = -\frac{1}{\alpha} \ln(1 - x_i)$  - экспоненциальное распределение с параметром  $\alpha$ ).

Для нормального распределения приём не работает и для функции распределения, у которой нет обратной функции - тоже не работает.

### 12.6. Формула для суммы случайных величин (формула свёртки)

Пусть есть  $X, Y$  - независимые случайные величины и известные их  $f_x(t_1)$  и  $f_y(t_2)$

Как будет распределена случайная величина  $Z = X + Y$  -  $f_z(t_3)$ ?

$$\text{Попробуем снова сделать через функции распределения: } F_z(t) = \mathbb{P}(z \leq t) = \mathbb{P}(x+y \leq t) = \int \int f(t_1, t_2) dt_1 dt_2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{t-t_1} f_x(t_1) f_y(t_2) dt_1 dt_2 = \int_{-\infty}^{+\infty} f_x(t_1) dt_1 \int_{-\infty}^{t-t_1} f_y(t_2) dt_2$$

$$\text{А мы знаем, что } f_z(t) = F'_z(t) = \int_{-\infty}^{+\infty} f_x(t_1) dt_1 \left( \int_{-\infty}^{t-t_1} f_y(t_2) dt_2 \right)'_t$$

$$\text{Вспомним, что мы умеем } \frac{d}{dt} \int_{\alpha(t)}^{\beta(t)} f(x, t) dx = \int_{\alpha(t)}^{\beta(t)} \frac{df(x, t)}{dt} dx + f(\beta(t), t) \cdot \frac{d\beta(t)}{dt} - f(\alpha(t), t) \cdot \frac{d\alpha(t)}{dt}$$

Тогда продолжим  $\int_{-\infty}^{+\infty} f_x(t_1) dt_1 \left( \int_{-\infty}^{t-t_1} f_y(t_2) dt_2 \right)' = \int_{-\infty}^{+\infty} f_x(t_1) f_y(t-t_1) \cdot 1 dt_1$

Тогда  $f_z(z) = \int_{-\infty}^{+\infty} f_x(x) f_y(z-x) dx$  или же то же самое  $f_z(z) = \int_{-\infty}^{+\infty} f_y(y) f_x(z-y) dy$

### 12.6.1. Упражнение

Пусть есть  $X$  с нормальным распределением:  $X \sim N(\mu, \sigma^2)$ , где  $f_x(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$

Найдём:

1.  $\mathbb{E}(X)$
2.  $Var(X)$

Приступим (ой, крч там оказалось, что интеграл неберущийся, поэтому просто на слово поверьте):

1.  $\mu$
2.  $\sigma^2$

Пусть  $X \sim N\left(\mu_1, \sqrt{\frac{1}{2}}\right)$  и  $Y \sim N\left(\mu_2, \sqrt{\frac{1}{2}}\right)$ , и есть  $Z = X + Y$ , тогда  $Z \sim N\left(\mu_1 + \mu_2, \sqrt{\frac{1}{2}} + \sqrt{\frac{1}{2}}\right)$

А  $f_z(t) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi}} e^{(y-\mu_1)^2} \cdot \frac{1}{\sqrt{\pi}} e^{-(z-y+\mu_2)^2} dy = \text{Автор потом дорешает.}$

### 12.6.2. Интеграл Гаусса-Пуассона

Есть табличный интеграл  $\int_0^{+\infty} e^{-x^2} dx = \frac{\sqrt{\pi}}{2}$

### 12.6.3. Упражнение

Пусть  $X$  распределено на отрезке  $[0; 2]$  и  $F_x(t) = ct^3$

Найти:

1.  $c$
2.  $\mathbb{P}(x \leq 0.3 | x \leq 0.6)$
3.  $cov(x + 1, \frac{1}{x})$

Находим:

$$1. f_x(t) = F'_x(t) = 3ct^2, \text{ а } \int_0^2 f_x(t) dt = \int_0^2 3ct^2 dt = 3c \int_0^2 t^2 dt = \frac{3ct^3}{3} \Big|_0^2 = c(8 - 0) = 1 = 8c = 1 \rightarrow c = \frac{1}{8}$$

Можно ещё решать другим способом:  $F_x(2) - F_x(0) = 1 = c2^3 - c0^3 = 8c = 1 \rightarrow c = \frac{1}{8}$

$$2. \text{ По теореме Байеса знаем, что } \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \text{ тогда } \mathbb{P}(x < 0.3 | x < 0.6) = \frac{\mathbb{P}(x < 0.3)}{\mathbb{P}(x < 0.6)} = \frac{f_x(0.3)}{f_x(0.6)} = \frac{0.3^2}{0.6^2} = \left(\frac{3}{6}\right)^2 = \frac{1}{8}$$

$$3. cov\left(x + 1, \frac{1}{x}\right) = cov\left(x, \frac{1}{x}\right) + \overbrace{cov\left(1, \frac{1}{x}\right)}^{=0} = cov\left(1, \frac{1}{x}\right) = \mathbb{E}\left(x \cdot \frac{1}{x}\right) - \mathbb{E}(x) \cdot \mathbb{E}\left(\frac{1}{x}\right)$$

$$\mathbb{E}(x) = \int_0^2 x \cdot f_x(x) dx = \int_0^2 \frac{3x^3}{8} dx = \frac{3}{8 \cdot 4} (2^4 - 0^4) = \frac{3 \cdot 16}{8 \cdot 4} = \frac{3}{2}$$

$$\mathbb{E}\left(\frac{1}{x}\right) = \int_0^2 \frac{1}{x} f_x(x) dx = \int_0^2 \frac{3x}{8} dx = \frac{3}{8 \cdot 2} (2^2 - 0^2) = \frac{3}{4}$$

$$\text{И получается } cov\left(x + 1, \frac{1}{x}\right) = \mathbb{E}\left(x \cdot \frac{1}{x}\right) - \mathbb{E}(x) \cdot \mathbb{E}\left(\frac{1}{x}\right) = \mathbb{E}(1) - \mathbb{E}(x) \cdot \mathbb{E}\left(\frac{1}{x}\right) = 1 - \frac{3}{2} \cdot \frac{3}{4} = 1 - \frac{9}{8} = -\frac{1}{8}$$

## 13. Многомерные и условные распределения