1 Описание кода

1.1 Договор про хранение графов

Граф хранится в специальном классе со следующими полями:

- Вершины графа хранятся списком:

$$V_G = (x_1, x_2, \ldots, x_k)$$

  где k - количество вершин графа, x_i - координата i-ой вершины

- Рёбра графа хранятся множеством:

$$E_G = \{(a_1, b_1), (a_2, b_2), \ldots\}$$

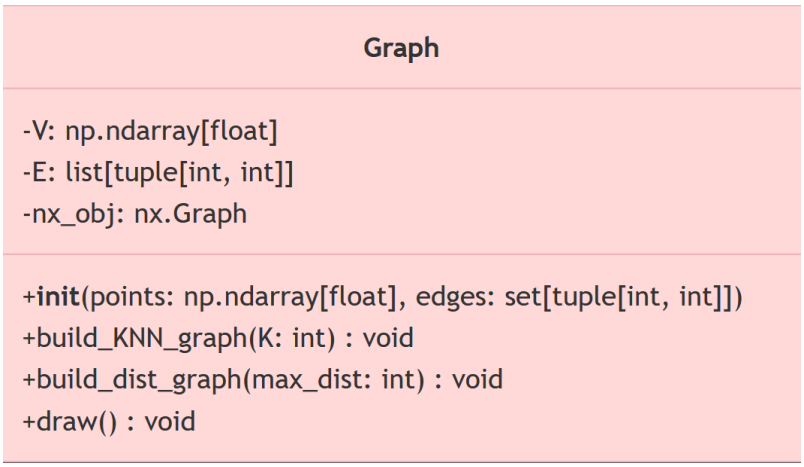
  где $a_i < b_i$ - номера вершин

- Также хранится объект `NetworkX.Graph` для удобного использования различных алгоритмов на графах

Класс имеет 3 метода:

1. Два метода для построения рёбер, соответствующих:
   - KNN-графу
   - Distance-графу
2. Один метод для визуализации графа

Ниже приведена UML-диграмма класса



1.2 Функции подсчета характеристик графа

В данном разделе представлены ключевые функции для анализа свойств графов. Все функции принимают на вход объект класса `Graph` и возвращают числовые характеристики.

```
def calculate_min_deg(G: Graph) -> int:
    """ Returns the minimum degree of a graph vertex """

def calculate_max_deg(G: Graph) -> int:
    """ Returns the maximum degree of a graph vertex """

def calculate_number_component(G: Graph) -> int:
    """ Returns the number of connected components of a graph """

def calculate_number_articul(G: Graph) -> int:
    """ Returns the number of articulation points of a graph """

def calculate_number_triangle(G: Graph) -> int:
    """ Returns the number of triangles in a graph """

def calculate_clique_number(G: Graph) -> int:
    """ Returns the click count of a graph """

def calculate_maxsize_independed_set(G: Graph) -> int:
    """ Returns the size of the maximum independent set """
```

Все функции реализованы с помощью библиотеки **NetworkX**. Почти все функции честным перебором дают точные значения характеристик.
Исключениями являются:

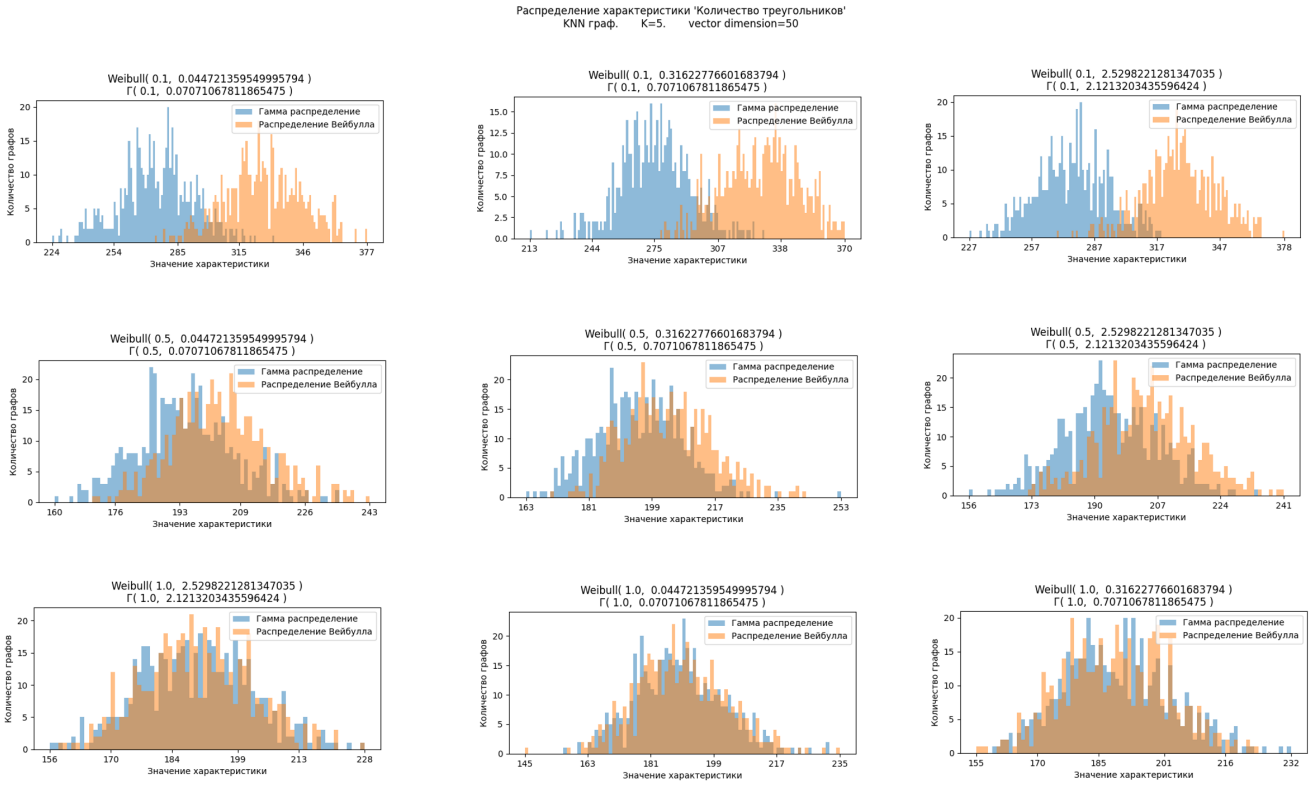
- Функция подсчета кликового числа. Реализована через жадный поиск хроматического числа графа, основано на том, что для дистанционного графа они совпадают почти наверное.
- Функция подсчета числа независимости графа. Реализована через подсчет кликового числа для дополнения.

Каждая функция покрыта тестами.

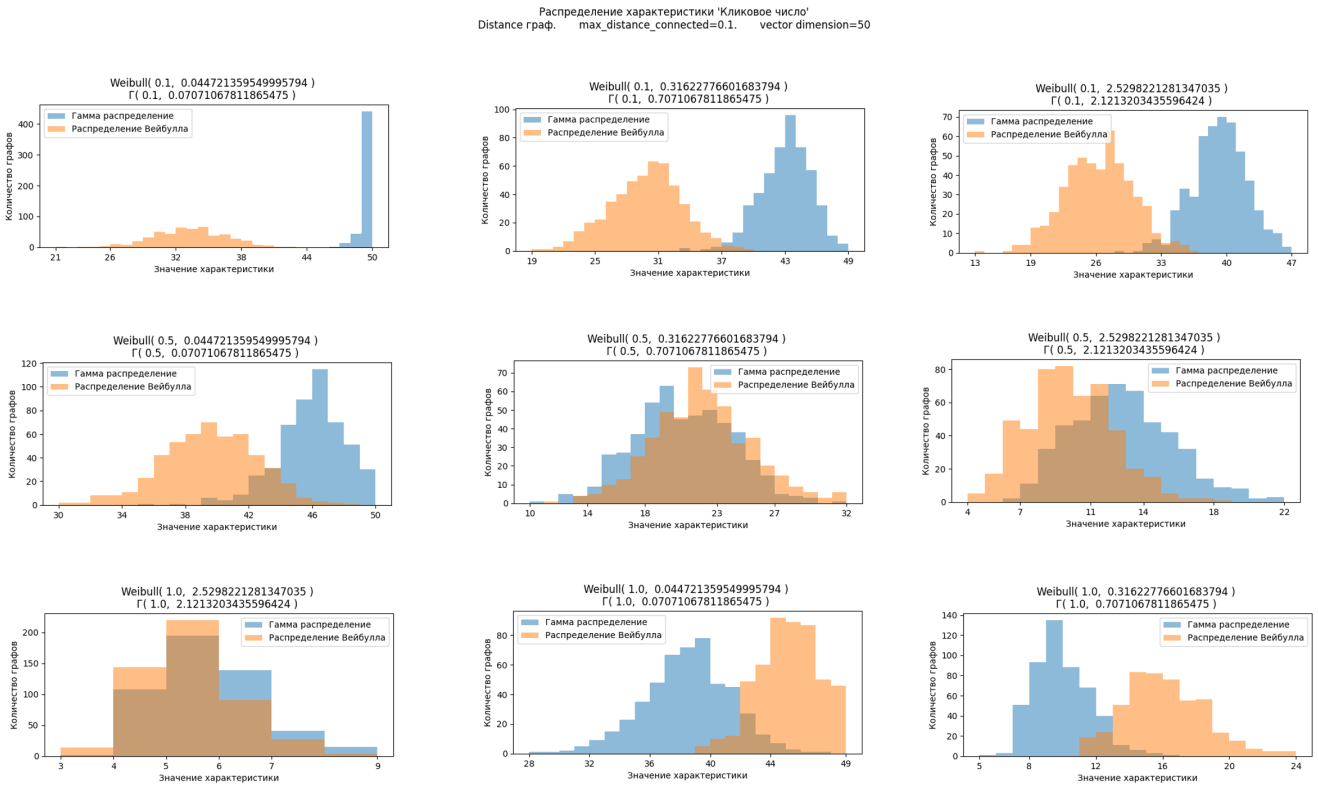
2. Part-I

2.1 Поведение характеристики в зависимости от параметров распределений (Минаков Д.Д)

Посмотрим на поведение характеристик при фиксированных параметрах построения графов, но с варьирующимися параметрами распределений.



Вывод: в зависимости от параметров распределений характеристика 'Количество треугольников' KNN графа может быть как хорошим признаком классификации (перекрытие гистограмм около нулевое), так и плохим (гистограммы практически идентичны).

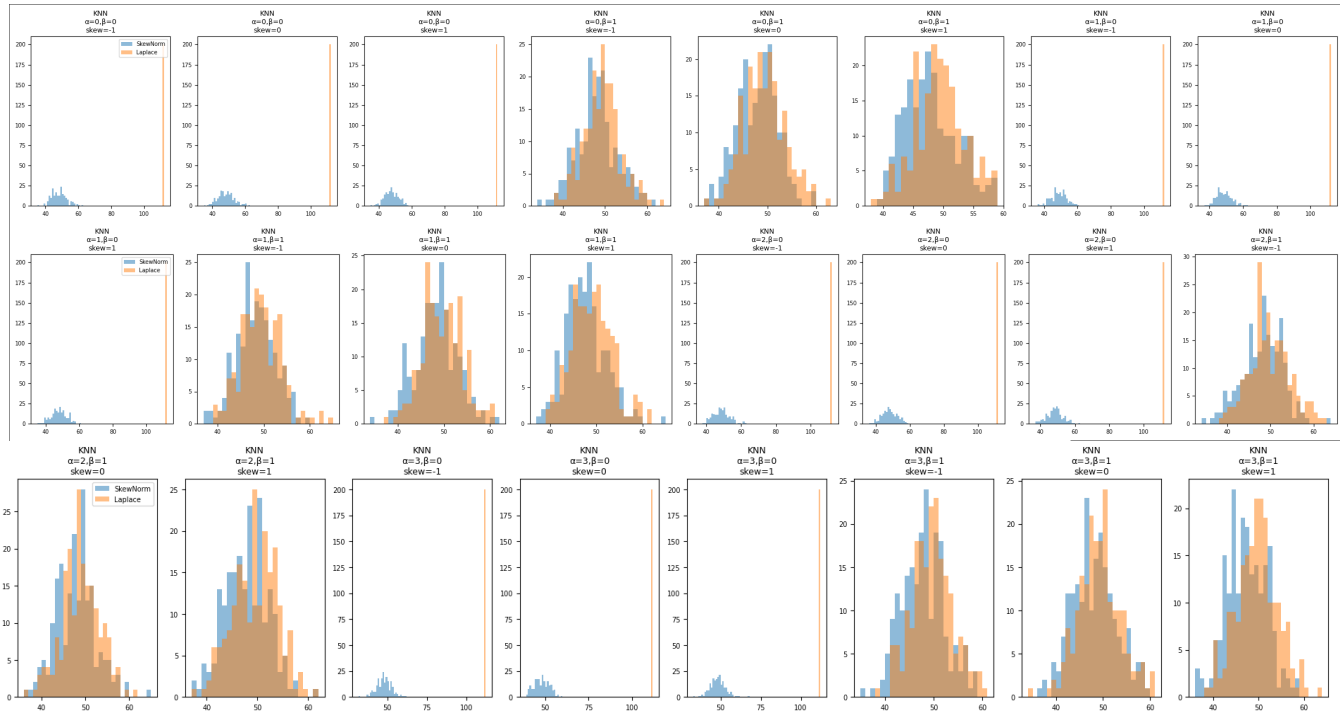


Вывод: в зависимости от параметров распределений характеристика ’Кликовое число’ Distance графа может быть как **очень хорошим** признаком классификации (перекрытие гистограмм нулевое), так и плохим (гистограммы практически идентичны).

2.1 Поведение характеристики в зависимости от параметров распределений (Иванова А.А)

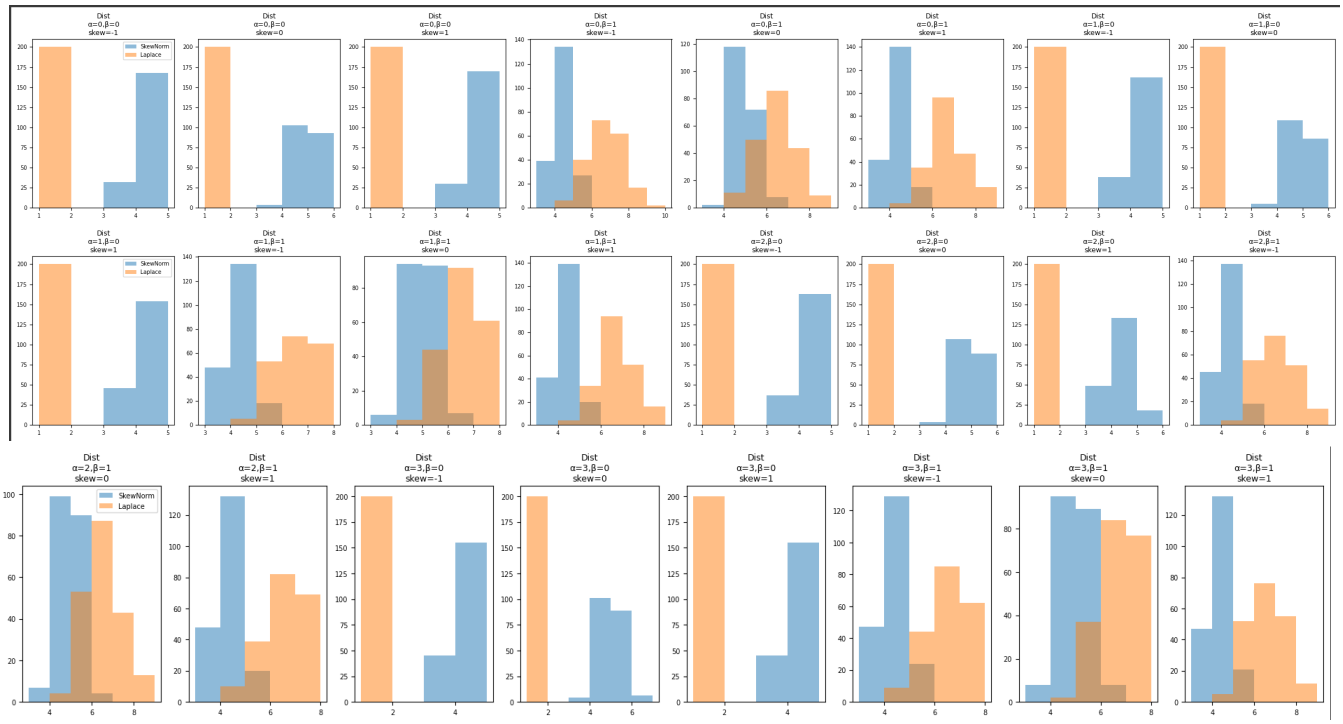
Посмотрим на поведение характеристик при фиксированных параметрах построения графов для распределения Лапласа и косого нормального, с варьирующимися параметрами распределений. Ниже графики, на которых перебираются различные параметры $\alpha_{laplace}, \beta_{laplace}, \alpha_{skew}$ при фиксированных параметрах графа:

- Размер графа = 40
- K в KNN = 3
- dist в Distance = 1
- характеристика для KNN графа - число треугольников (на этой странице)
- характеристика для Dist графа - максимальное независимое множество



Вывод: в зависимости от параметров распределений характеристика 'Количество треугольников' KNN графа может быть **очень хорошим** признаком классификации при хороших параметрах распределений, а при некоторых графики практически идентичны, распределения трудно отличимы.

Далее графики для дистанционного графа :

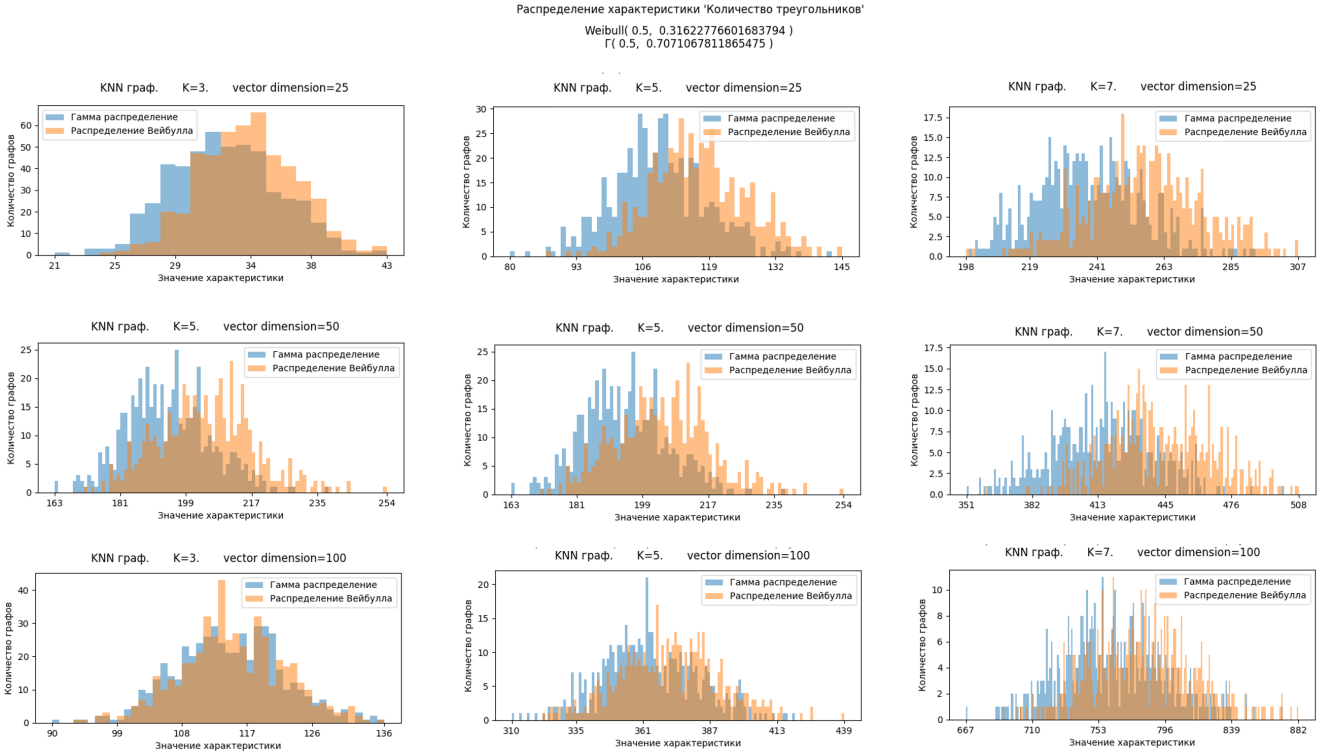


Вывод: в целом характеристика 'Размер максимального независимого множества' неплохая характеристика, при некоторых параметрах она лучше разделяет распределения, в некоторых хуже, но в целом всегда неплохо.

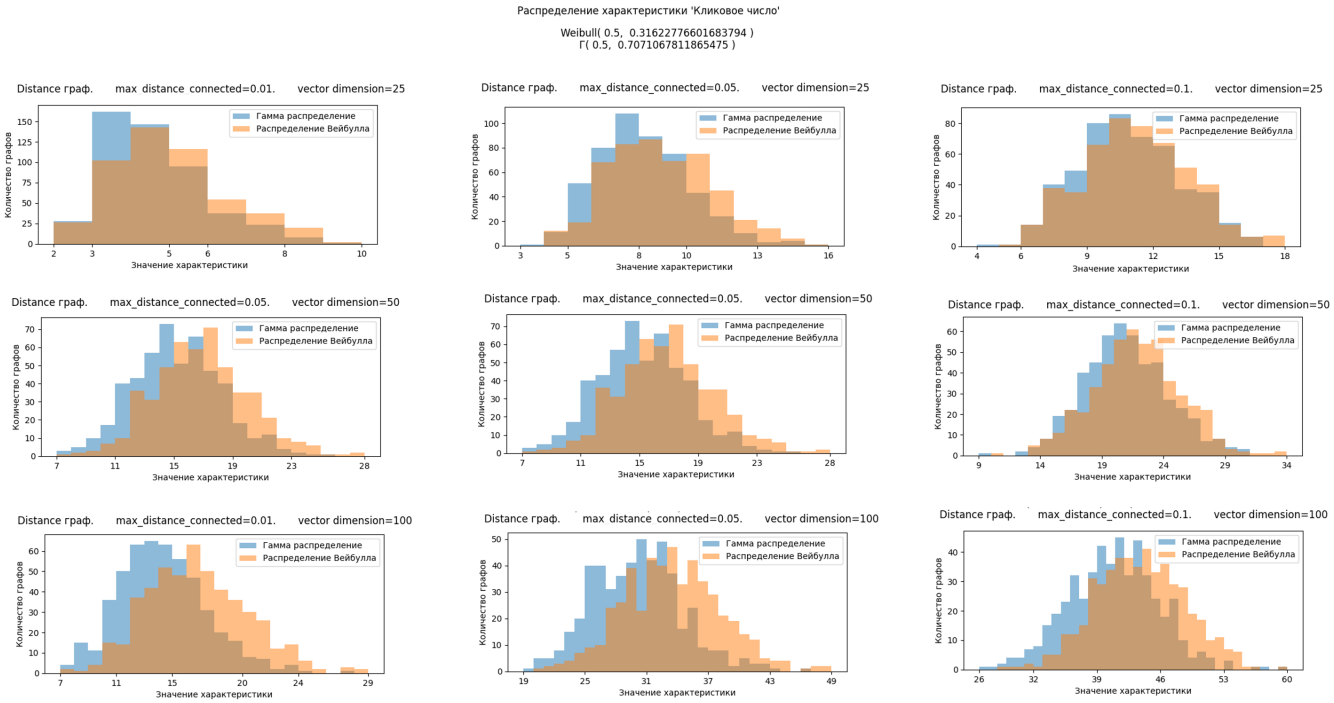
2.2 Поведение характеристики в зависимости от построения (Минаков Д.Д.)

Теперь посмотрим на поведение характеристик в зависимости от параметров построения графов, при фиксированных распределениях

$$\text{Weibull}(\frac{1}{2}, \frac{1}{\sqrt{10}}) \qquad \Gamma(\frac{1}{2}, \frac{1}{\sqrt{2}})$$



Вывод: при большом размере выборки, количество треугольников выглядит как не самая удачная характеристика для классификации, однако, при относительно небольшой выборке (≤ 50), эта характеристика может оказаться неплохим второстепенным признаком.



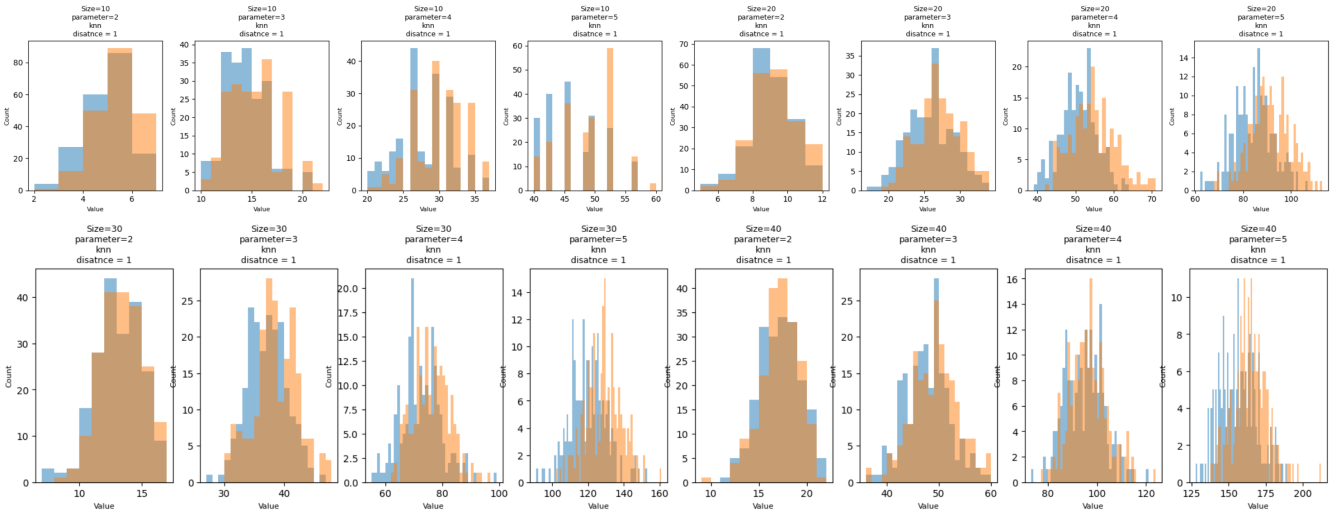
Вывод: кликовое число, с точки зрения задачи классификации, для данных распределений является посредственным признаком, независимо от размера выборки и расстояния связи. Убедиться в этом еще раз мы сможем в **Part-II**.

2.2 Поведение характеристики в зависимости от построения (Иванова А.А.)

Теперь посмотрим на поведение характеристик в зависимости от параметров построения графов, при фиксированных распределениях

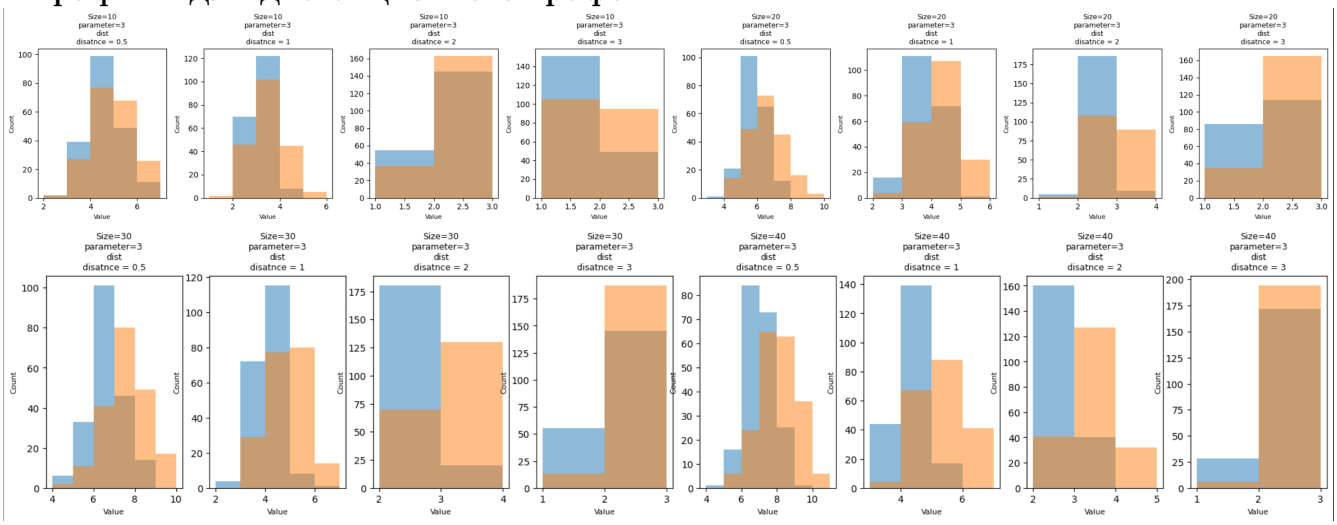
$$\text{Laplace}(\ 0, \frac{1}{\sqrt{2}}) \qquad\qquad Skewnormal(1)$$

Графики для KNN графа :



Вывод: При любом размере выборки количество треугольников не выглядит хорошей характеристикой, потому что графики практически идентичны. Лучшее, что можно получить, при количестве вершин 40 и количестве соседей в графе - 5 (правый нижний график)

Графики для дистанционного графа



Вывод: Максимальное независимое множество, с точки зрения задачи классификации, для данных распределений является признаком получше, например для 40 вершин и дистанции 2, распределения уже неплохо различимы, на основе этой характеристики и будем строить критическое множество.

2.3 Построение критического множества и оценка мощности критерия

Построение критического множества $\mathcal{A}$ происходит следующим образом:
1. Генерируем большое количество графов, с фиксированными параметрами и считаем для каждого из них характеристику.
2. Считаем 95% перцентиль = A_{crit} - это будет крайнее значение множества $\mathcal{A}$
3. Теперь, если значение характеристики графа $\leq A_{crit}$, то принимаем гипотезу H_0, иначе отвергаем

Минаков Д.Д.

Для распределений Weibull($\frac{1}{2}, \frac{1}{\sqrt{10}}$) Γ($\frac{1}{2}, \frac{1}{\sqrt{2}}$)

и **Distance** графа, по характеристике **кликовое число** построим критическое множество A.

H_0- гамма распределение, H_1- распределение Вейбулла.
Получим:
Критическое значение $A_{crit} = 46$.
Мощность критерия = 0.31120000
Ошибка 1 рода : 0.02190000

Иванова А.А.

Для распределений Laplace($0, \frac{1}{\sqrt{2}}$) $Skewnormal(1)$

и **Distance** графа, по характеристике **размер максимального независимого множества** построим критическое множество A.

H_0- косое нормальное распределение, H_1-распределение Лапласа.
Получим:
Критическое значение $A_{crit} = 5$
Мощность критерия = 0.30330000
Ошибка 1 рода = 0.00270000

Итого:
В обоих случаях получили неплохой критерий, с мощностями в 30% и 31% и ошибками первого рода 0.27% и 2.19%

3. Part-II

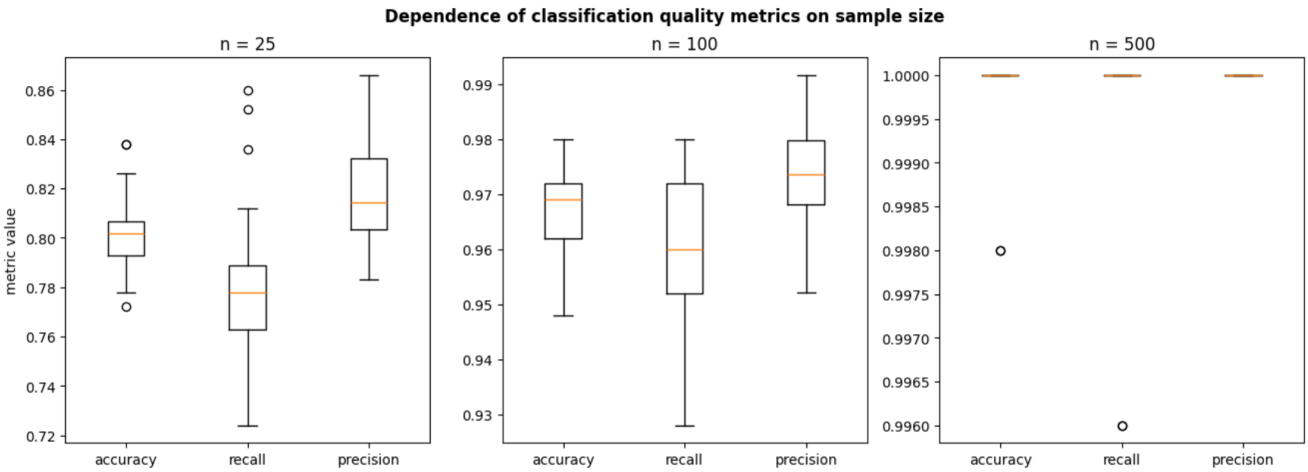
3.1 Применяем классификационные алгоритмы (Минаков Д.Д.)

Выберем 3 модели - линейная, логистическая и ridge регрессии, обучим на них классификатор и сравним качество

N	Модель	Precision	Accuracy	Recall
25	Logistic	0.8071	0.8037	0.7980
	Linear	0.8426	0.8077	0.7567
	Ridge	0.8426	0.8077	0.7567
100	Logistic	0.9700	0.9693	0.9687
	Linear	0.9907	0.9597	0.9280
	Ridge	0.9907	0.9597	0.9280
500	Logistic	1.0000	0.9993	0.9987
	Linear	1.0000	0.9993	0.9987
	Ridge	1.0000	0.9993	0.9987

Все линейные модели дают очень близкий результат, поэтому зафиксируем в качестве модели обычную **логистическую регрессию**

Проведем кроссвалидацию для оценки дисперсии метрик



Занесем данные в таблицу

Выводы:

- Увеличение выборки снижает дисперсию экспоненциально
- **Стабильность метрик:**
  - Precision демонстрирует самую низкую дисперсию на всех выборках
  - Recall наиболее чувствителен к размеру выборки

Размер выборки	**Метрика**	**Среднее**	**Дисперсия**
N=25	Accuracy	0.80	0.000271
	Recall	0.78	0.001231
	Precision	0.81	0.000460
N=100	Accuracy	0.97	6.384×10^{-5}
	Recall	0.96	0.000166
	Precision	0.97	9.624×10^{-5}
N=500	Accuracy	0.999	3.6×10^{-7}
	Recall	0.999	1.44×10^{-6}
	Precision	0.99	0.0

- **Оптимальный размер:** Для данной задачи выборка размером n=100 уже обеспечивает отличные результаты, а дальнейшее увеличение (n=500) лишь незначительно улучшает метрики и их стабильность.

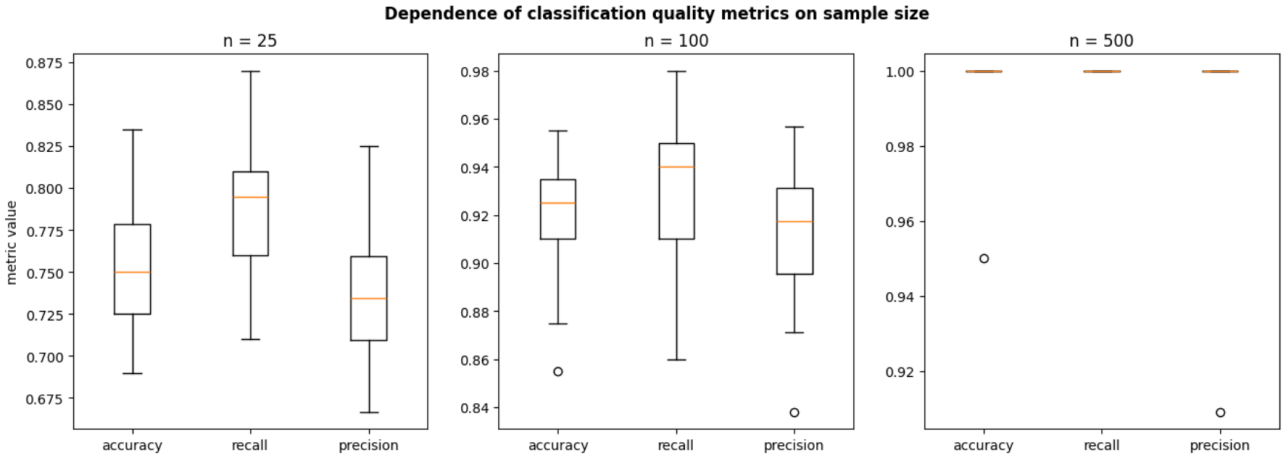
3.1 Применяем классификационные алгоритмы (Иванова А.А.)

Выберем 3 модели - линейная, логистическая и ridge регрессии, обучим на них классификатор и сравним качество

Размер выборки	Модель	Precision	Accuracy	Recall
N=25	Logistic	0.730	0.749	0.789
	Linear	0.734	0.752	0.789
	Ridge	0.734	0.752	0.789
N=100	Logistic	0.916	0.919	0.923
	Linear	0.906	0.919	0.935
	Ridge	0.906	0.919	0.935
N=500	Logistic	1.000	1.000	1.000
	Linear	1.000	1.000	1.000
	Ridge	1.000	1.000	1.000

В качестве самой удобной и не проигрывающей по качеству модели выберем логистическую регрессию и дальше будем анализировать дисперсию и важность характеристик относительно нее.

Проведем кроссвалидацию для оценки дисперсии метрик



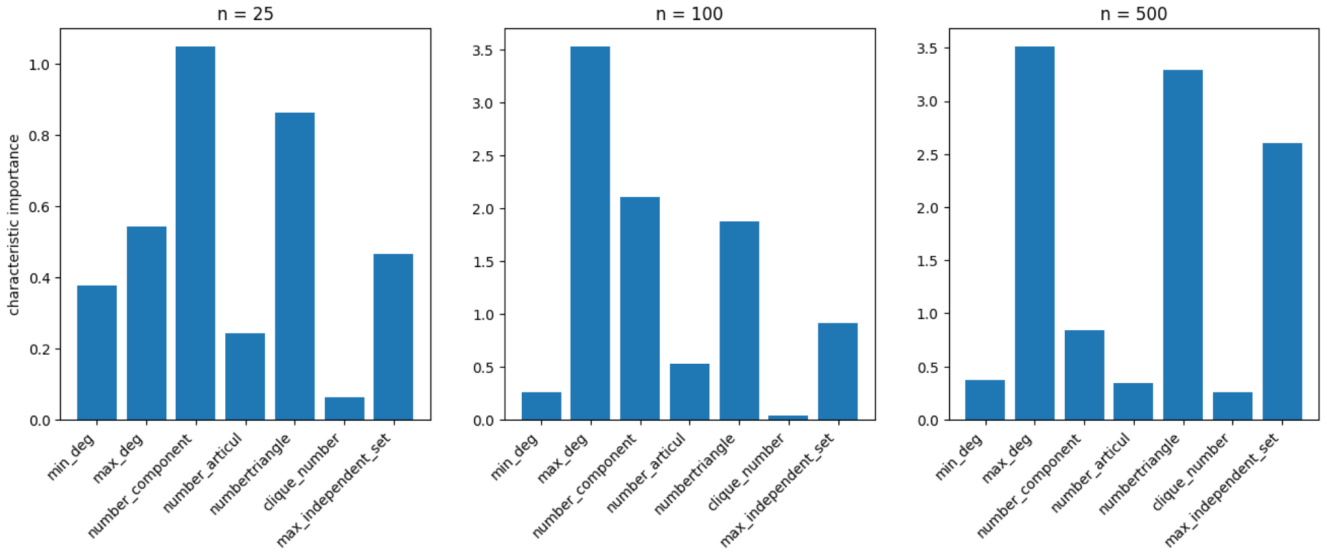
Размер выборки	Метрика	Среднее	Дисперсия
N=25	Accuracy	0.75	0.001028
	Recall	0.79	0.001366
	Precision	0.73	0.001303
N=100	Accuracy	0.92	0.000411
	Recall	0.94	0.000906
	Precision	0.91	0.000593
N=500	Accuracy	0.99	4×10^{-5}
	Recall	1.00	0.0
	Precision	1.00	0.00016

Вывод :
При увеличении размера выборки, дисперсия снижается достаточно быстро.

Оптимальный размер:
В зависимости от задачи можно использовать:

- граф на 100 вершинах (если нужны быстрые вычисления и неплохое качество)
- граф на 500 вершинх, если нужна высокая точность, и есть время на вычисления (т.к. вычисления на графе для 500 вершинах кратно увеличиваются)

3.2 Важность характеристик, как признаков классификации (Минаков Д.Д.)



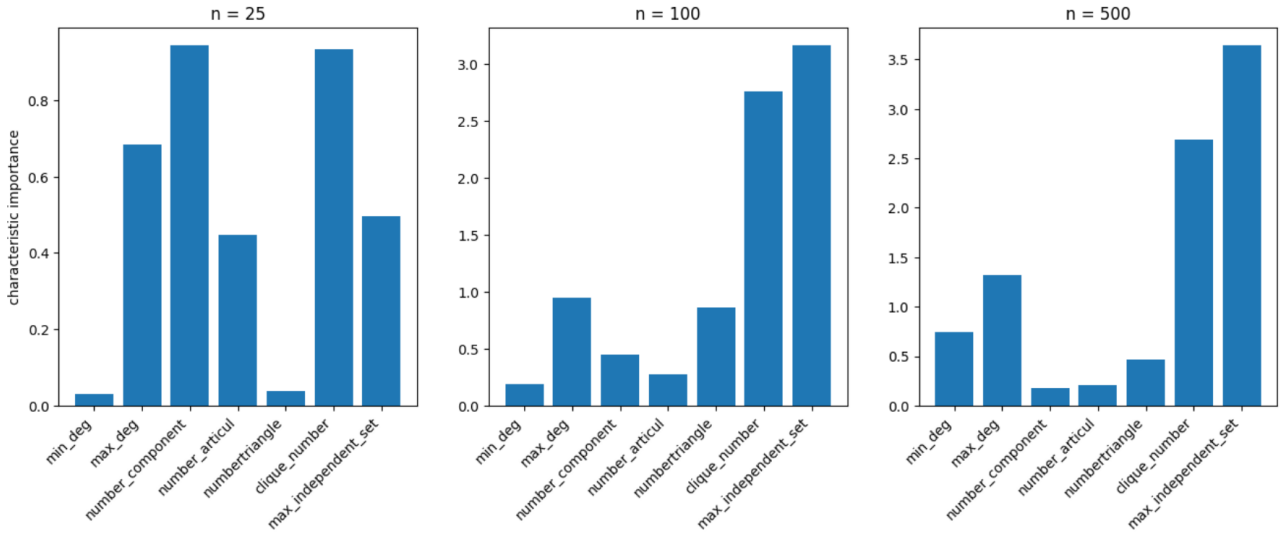
Вывод:

Характеристика графа	**n=25**	**n=100**	**n=500**
Минимальная степень			
Максимальная степень			
Количество компонент связности			
Количество точек сочленения			
Количество треугольников			
Кликовое число графа			
Число независимости			

Высокая важность
Средняя важность
Низкая важность

3.2 Важность характеристик, как признаков классификации (Иванова А.А.)

Аналогично, будем использовать **логистическую регрессию**



Вывод:

Характеристика графа	**n=25**	**n=100**	**n=500**
Минимальная степень			
Максимальная степень			
Количество компонент связности			
Количество точек сочленения			
Количество треугольников			
Кликовое число графа			
Число независимости			

Высокая важность
Средняя важность
Низкая важность

3.3 Выводы о вероятности ошибки первого рода и мощности модели как критерия

Поскольку удобнее было в обоих случаях гипотезу H_0 обозначить в датасете, как класс 1, то ошибка первого рода считается как $1 - recall$, а мощность критерия как $\frac{TN}{TN+FP}$

Итоги Минакова Д. Д

Размер выборки	**Ошибка I рода (α)**	**Мощность критерия**
N = 25	0.20	0.80
N = 100	0.03	0.97
N = 500	0.01	1.00

Итоги Ивановой А.А.

Размер выборки	**Ошибка I рода (α)**	**Мощность критерия**
N = 25	0.21	0.7086
N = 100	0.06	0.9153
N = 500	0.00	1.0000

Общий анализ:
На 500 вершинах можно точно отличить распределения, потому что у них сильно отличаются характеристики, но это достаточно долго, поэтому лучше выбрать 100 вершин, ошибка первого рода 0.06 и 0.03 соответственно, мощность 0.92 и 0.97 соответственно, но при этом обсчет 10к графов займет всего 10 минут

Общий вывод:
Классификатор имеет высокую мощность и маленькую ошибку первого рода на 3000 вычислениях, что делает его отличным статистическим критерием.