Проект по случайным графам

Ильин Павел, Кулешов Илья, ПАДИИ, 2 курс

30 мая 2025 г.

Оглавление

Введение **3**

I Исследование свойств характеристик графов **4**

1 Исследование поведения числовой характеристики τ в зависимости от параметров распределений **5**
1.1 Методология исследования 5
1.1.1 Описание числовых характеристик 5
1.1.2 Исследуемые распределения 5
1.1.3 Применяемые ML алгоритмы 5
1.1.4 Разное техническое 6
1.2 Характеристики графов при различных параметрах распределений . . . 6
1.2.1 Эксперимент 1: Normal vs Student-t 6
1.2.2 Эксперимент 2: Pareto vs Gamma 6

2 Исследование поведения характеристики τ в зависимости от параметров построения графа **8**
2.1 Эксперимент 1: Normal vs Student-t 8
2.2 Эксперимент 2: Pareto vs Gamma 8

3 Построение статистических критериев **10**
3.1 Критерии для Normal и Student-t 10
3.1.1 Построение множества A при $\alpha = 0.05$ 10
3.1.2 Оценка мощности критерия для Normal и Student-t 10
3.1.3 Оценка мощности критерия для Pareto и Gamma 10

II Применение машинного обучения для классификации распределений **12**

4 Классификация с использованием нескольких характеристик **13**
4.1 Подготовка данных 13
4.1.1 Формирование датасетов 13
4.2 Эксперимент 1: Normal vs Student-t 13
4.2.1 Результаты для различных размеров графов 13
4.2.2 Оценка мощности критерия по ML 14
4.3 Эксперимент 2: Pareto vs Gamma 14
4.3.1 Результаты для различных размеров графов 14
4.3.2 Оценка мощности критерия по ML 15

Заключение **16**
4.3.3 KNN 16
4.3.4 Dist 16

Введение

Данный отчёт описывает исследование свойств случайных графов, построенных на основе различных вероятностных распределений. В работе рассматриваются два типа графов: графы k-ближайших соседей (KNN) и дистанционные графы.

Цель работы: Исследовать поведение числовых характеристик случайных графов в зависимости от параметров распределений и параметров построения графов.

Задачи:

1. Исследовать поведение числа треугольников, минимального кликового числа, числа компонент и хроматического числа в зависимости от параметров распределений
2. Исследовать влияние параметров процедуры построения графа и размера выборки
3. Построить статистические и ML критерии и оценить их мощность

Часть I

Исследование свойств характеристик графов

Глава 1

Исследование поведения числовой характеристики τ в зависимости от параметров распределений

1.1 Методология исследования

1.1.1 Описание числовых характеристик

В работе исследуются 4 основные характеристики:

- τ^{KNN} - количество треугольников в графе k-ближайших соседей
- τ^{dist} - минимального кликовое число дистанционного графа
- τ^{KNN} - количество компонент связности в графе k-ближайших соседей
- τ^{dist} - хроматическое число дистанционного графа

1.1.2 Исследуемые распределения

Эксперимент 1: Normal vs Student-t

- Нормальное распределение $N(0, \alpha)$
- Распределение Стьюдента с ν степенями свободы

Эксперимент 2: Pareto vs Gamma

- Распределение Парето с параметром α
- Гамма-распределение с параметрами shape и scale

1.1.3 Применяемые ML алгоритмы

- Логистическая регрессия
- Решающее дерево
- Градиентный бустинг (CatBoost) с iterations = 50

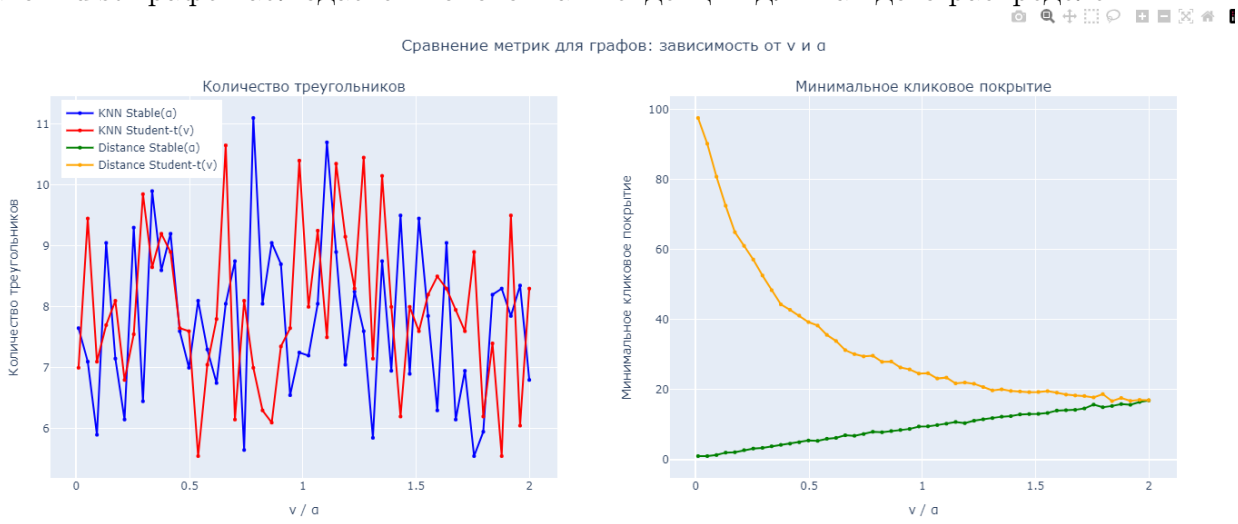
1.1.4 Разное техническое

- Все замеры проводятся методом Монте-Карло с 20 итерациями (кроме тех случаев, когда явно будет указано другое число)
- Базовое количество соседей для графа KNN - 5
- Базовое расстояние в дистанционном графе - 1.0
- Так как подсчет минимального кликового покрытия, подсчет хроматического числа - NP полные задачи, их вычисление происходит с помощью Python библиотеки networkx жадными методами.
- Для минимального кликового покрытия для удобства запускается подсчет хроматического числа дополнения графа (ввиду равенства данных величин)

1.2 Характеристики графов при различных параметрах распределений

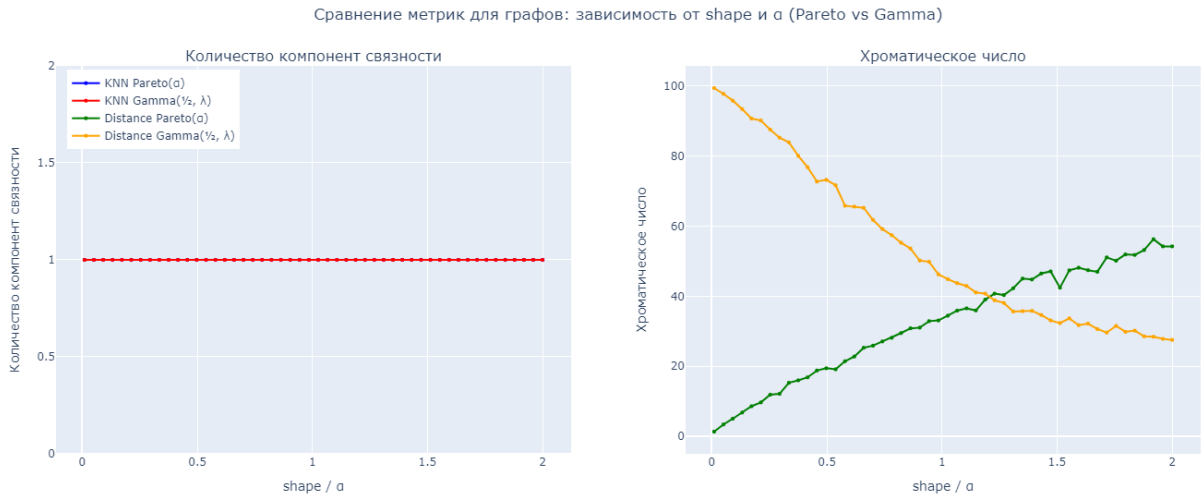
1.2.1 Эксперимент 1: Normal vs Student-t

На графике ниже видно, что параметры распределения не влияют никак на KNN граф, однако в dist графе наблюдается монотонная тенденция для каждого распределения



1.2.2 Эксперимент 2: Pareto vs Gamma

На графике ниже видно, что число компонент почти всегда - 1, однако в dist графе наблюдается монотонная тенденция для каждого распределения, но в различные стороны

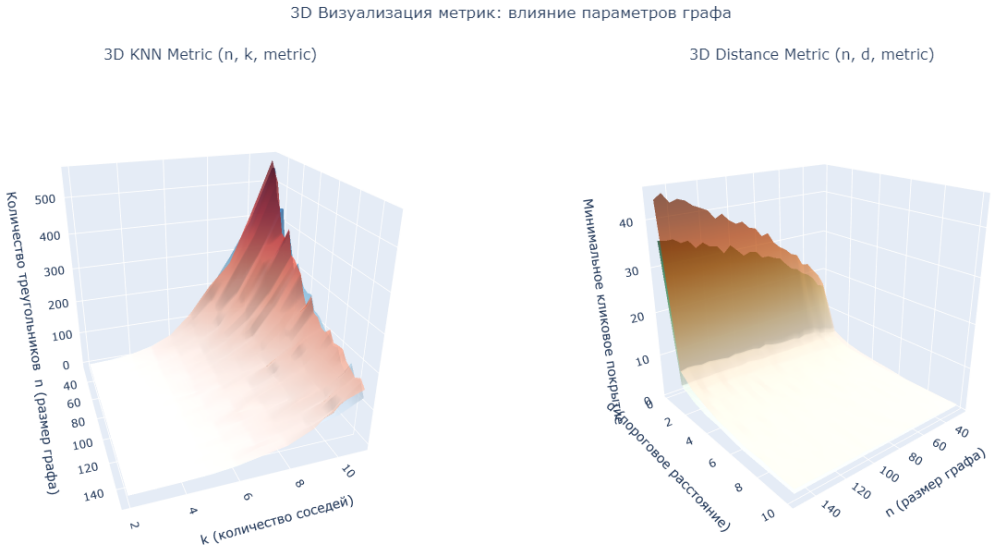

Сравнение метрик для графов: зависимость от shape и α (Pareto vs Gamma)
Количество компонент связности
KNN Pareto(α)
KNN Gamma(½, λ)
Distance Pareto(α)
Distance Gamma(½, λ)
Количество компонент связности
shape / α
Хроматическое число
Хроматическое число
shape / α

Глава 2

Исследование поведения характеристики τ в зависимости от параметров построения графа

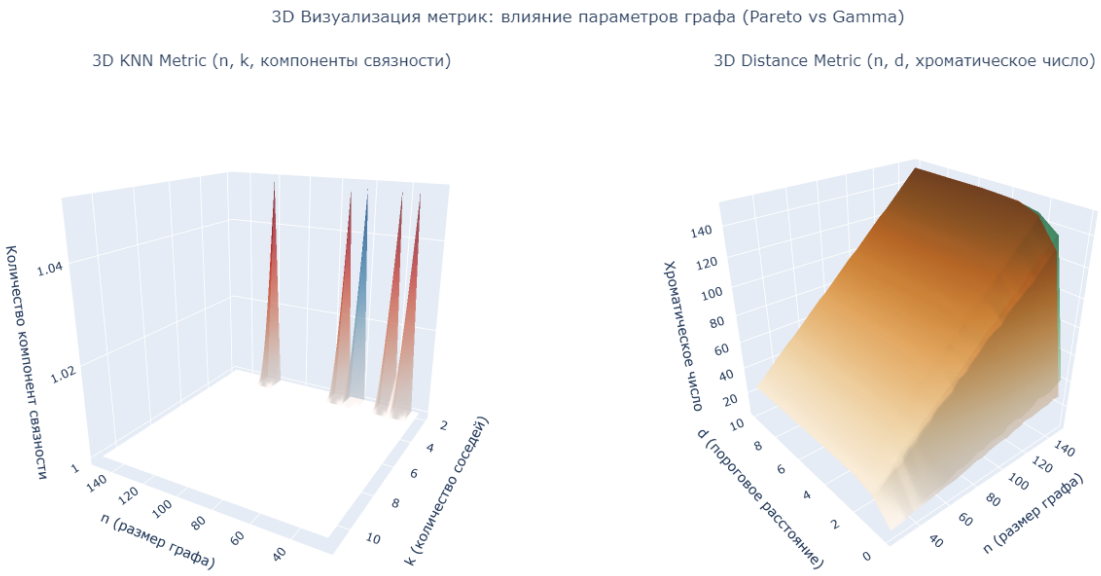
2.1 Эксперимент 1: Normal vs Student-t

Рассмотрим для начала число треугольников в KNN-графе: при увеличении n и k количество треугольников в каждом распределении увеличивается - в целом это логично (больше граф - больше ребер - больше треугольников). В случае дистанцинного графа, кликовое число растет вместе с n, но уменьшается с ростом d (тоже логично - больше ребер - меньше клик нужно).



2.2 Эксперимент 2: Pareto vs Gamma

Видно, что число компонент связности почти всегда 1 при различных n и парамерах построения KNN-графа, однако бывают и небольшие выбросы иногда. Для хроматического числа все поинтереснее - при минимальных d нам достаточно пары цветов (так как граф разряжен), но в среднем с ростом n и уменьшением d наблюдается тенденция на рост хроматического числа


3D Визуализация метрик: влияние параметров графа (Pareto vs Gamma)
3D KNN Metric (n, k, компоненты связности)
3D Distance Metric (n, d, хроматическое число)
Количество компонент связности
1.04
1.02
1
n (размер графа)
140
120
100
80
60
40
k (количество соседей)
2
4
6
8
10
Хроматическое число
140
120
100
80
60
40
20
10
d (пороговое расстояние)
8
6
4
2
0
n (размер графа)
40
60
80
100
120
140

Глава 3

Построение статистических критериев

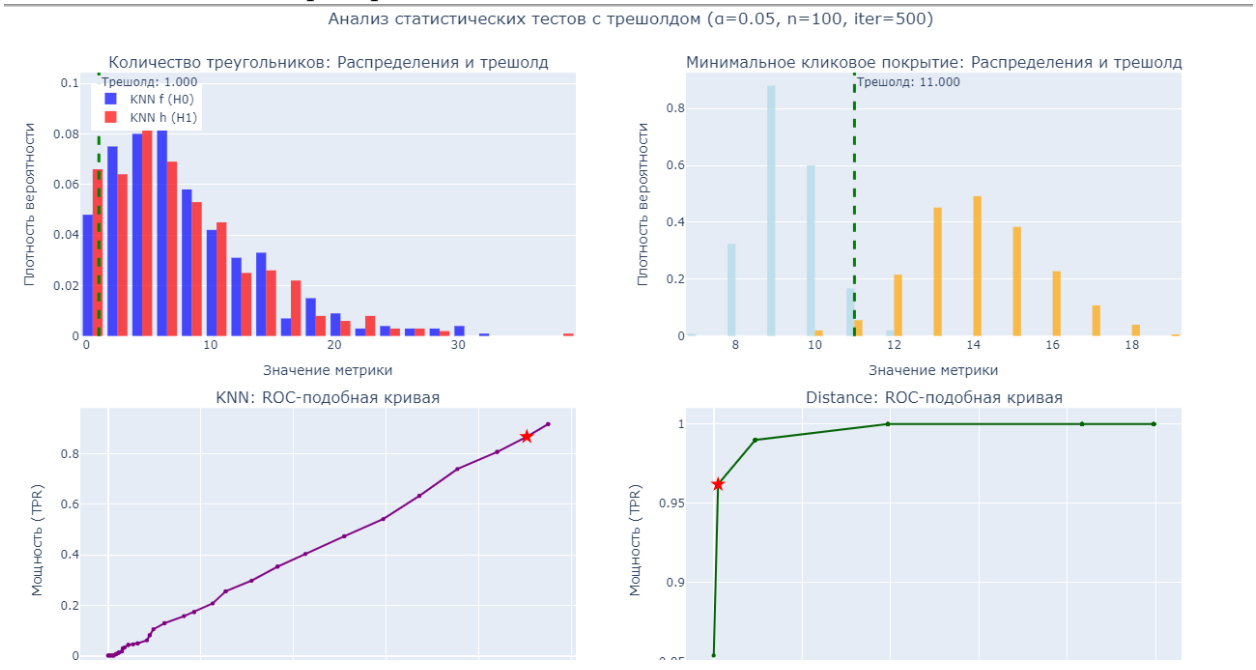
3.1 Критерии для Normal и Student-t

3.1.1 Построение множества A при $\alpha = 0.05$

Для подсчета статистического критерия разбиения используется 500 итераций метода Монте-Карло для формирования выборок, а затем взятие 95-го перцентиля плотности f (Normal и Pareto соответственно) для минимизации ошибки первого рода

3.1.2 Оценка мощности критерия для Normal и Student-t

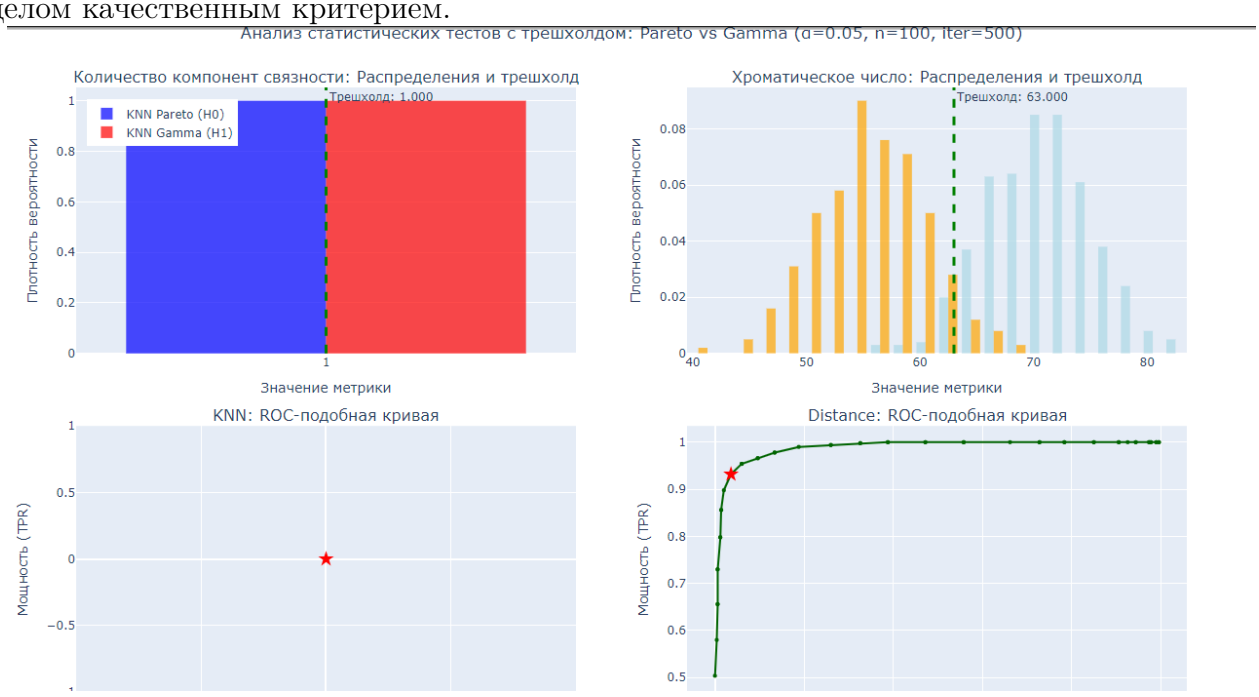
Видно, что для KNN графа и его характеристики сложно выбрать хороший критерий в целом. Однако для дистанционного графа Student-t и Normal сильно различаются по своим средним, что позволяет критерию для минимизации ошибки первого рода быть в целом качественным критерием.



3.1.3 Оценка мощности критерия для Pareto и Gamma

Видно, что для KNN графа и его характеристики сложно выбрать хороший критерий в целом. Однако для дистанционного графа Student-t и Normal сильно различаются по

своим средним, что позволяет критерию для минимизации ошибки первого рода быть в целом качественным критерием.



Часть II

Применение машинного обучения для классификации распределений

Глава 4

Классификация с использованием нескольких характеристик

4.1 Подготовка данных

4.1.1 Формирование датасетов

Для генерации датасетов были выбраны размеры графа : n = [25, 100, 500], затем с помощью 50 итераций метода Монте-Карло формировались наблюдения вида [размер графа, параметр построения, характеристика, тип распределения].

Для анализа был взят дистанционный граф.

4.2 Эксперимент 1: Normal vs Student-t

4.2.1 Результаты для различных размеров графов

В целом лучше всего себя показал градиентный бустинг CatBoost (подписан как ctb). Линейная модель не смогла грамотно оценить зависимость по такому малому количеству данных, а решающее дерево просто средне обучилось.

	model	n	f1_score
0	logreg	25	0.308±0.0
1	decision_tree	25	0.824±0.0
2	ctb	25	0.933±0.0
3	logreg	100	0.4±0.0
4	decision_tree	100	0.778±0.0
5	ctb	100	0.857±0.0
6	logreg	500	0.667±0.0
7	decision_tree	500	0.824±0.0
8	ctb	500	0.933±0.0

4.2.2 Оценка мощности критерия по ML

Несмотря на неплохие показатели f1-score, который является средним гармоническим для ошибок первого и второго рода, сам критерий ошибки первого рода не получился достаточно качественным

	model	n	pow
0	logreg	n=25	0.285714
1	decision_tree	n=25	0.000000
2	ctb	n=25	0.857143
3	logreg	n=100	0.285714
4	decision_tree	n=100	0.000000
5	ctb	n=100	0.714286
6	logreg	n=500	0.571429
7	decision_tree	n=500	0.000000
8	ctb	n=500	0.857143

4.3 Эксперимент 2: Pareto vs Gamma

4.3.1 Результаты для различных размеров графов

В целом лучше всего себя показал градиентный бустинг CatBoost (подписан как ctb), хотя в среднем модели получили чуть худшее качество, чем в предыдущем эскперимете

	model	n	f1_score
0	logreg	25	0.182±0.0
1	decision_tree	25	0.632±0.0
2	ctb	25	0.667±0.0
3	logreg	100	0.182±0.0
4	decision_tree	100	0.671±0.034
5	ctb	100	0.778±0.0
6	logreg	500	0.182±0.0
7	decision_tree	500	0.68±0.031
8	ctb	500	0.737±0.0

4.3.2 Оценка мощности критерия по ML

Несмотря на ухудшение качества общего, мощность критерия в лучшем случае осталась неизменной относительно предыдущего эксперимента

	model	n	pow
0	logreg	n=25	0.142857
1	decision_tree	n=25	0.000000
2	ctb	n=25	0.428571
3	logreg	n=100	0.142857
4	decision_tree	n=100	0.000000
5	ctb	n=100	0.857143
6	logreg	n=500	0.142857
7	decision_tree	n=500	0.000000
8	ctb	n=500	0.857143

Заключение

Основные результаты

Мы исследовали поведение 4 различных характеристик для двух типов графов в различных распределениях и определили различные критерии для оценки принадлежности графа к какому либо типу (по исходному распределению).

4.3.3 KNN

Для KNN графов распределения были не так важны - и это понятно, ведь для каждой вершины мы берем всегда k ближайших соседей и неважно как они далеко, поэтому есть гипотеза, что действительно критичным для KNN графа является не просто само распределение, а локальные различия в изменениях плотности - так например при подсчете треугольников для KNN графа равномерного распределения будут в среднем получатся последовательные клики размера k, а в распределениях с явными пиками, в этих пиках вполне могут получаться клики размера 2k (если достаточно точек было изначально). Если же распределения локально имеют различные виды участков, (Student-t и Normal), то граф будет слишком случайным по количеству треугольников (что было видно выше на картинке). Аналогично для подсчета числа компонент (только тут все еще менее интересно)

4.3.4 Dist

Для дистанционных графов все интереснее, чем для KNN, ведь здесь на различные характеристики графа влияет уже не только изменение плотности, но и среднее и дисперсия распределения. Ввиду этого можно наблюдать различное поведение. Также графическим меодом было доказано, что минимальное кликовое число графа можно считать через хроматическое число дополнения графа. (3D график хроматического числа инвертировать по n и d и получим график минимального кликового покрытия). В конечном итоге, именно большая динамичность и большая случайность таких графов, позволяют качественно разделять различные исходные распределения между собой даже в одномерном случае, в отличие от KNN

Литература

[1] NetworkX developers. NetworkX: Network Analysis in Python. https://networkx.org/

[2] Scikit-learn developers. Scikit-learn: Machine Learning in Python. https://scikit-learn.org/

[3] CatBoost developers. CatBoost: Gradient Boosting on Decision Trees. https://catboost.ai/

[4] NumPy developers. NumPy: The fundamental package for scientific computing with Python. https://numpy.org/