Проект по случайным графам

Чегодаева Таисия и Купряков Дмитрий, ПАДИИ, 2 курс

25 мая 2025 г.

Часть I

Исследование свойств характеристики.

Глава 1

Исследовать, как ведет себя числовая характеристика τ в зависимости от параметров распределений θ и ν, зафиксировав размер выборки и параметр процедуры построения графа.

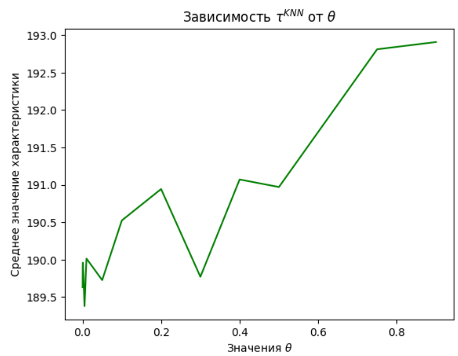
1.1 Характеристика τ^{KNN}.

1.1.1 Распределение LogNormal с $\mu = 0$ и параметром θ.

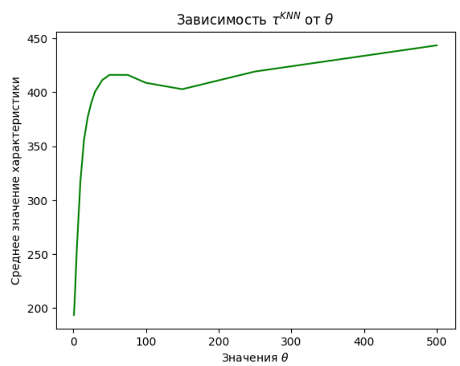
Зафиксируем размер выборки $n = 50$ и количество соседей $k = 5$. Число итераций для метода Монте-Карло равно 1000.

Сначала посмотрим на $\theta \in (0, 1)$.

При небольших θ среднее значение характеристики ≈ 190 и начинает расти при $\theta \longrightarrow 1$.



Теперь посмотрим на $\theta \in [1, 500]$.



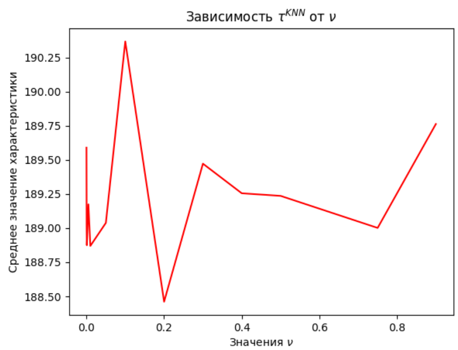
При $\theta \in [1, 100]$ наблюдается резкий рост среднего значения характеристики, после этого

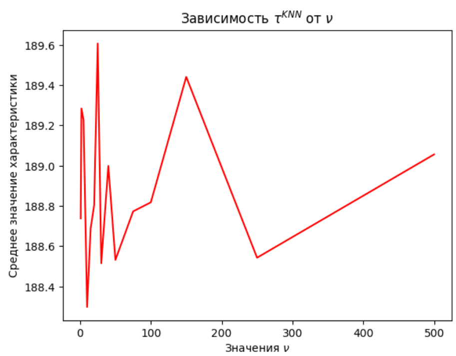
1.1.2 Распределение Exp с параметром λ.

Зафиксируем размер выборки $n = 50$ и количество соседей $k = 5$. Число итераций для метода Монте-Карло равно 1000.

Точно также будем перебирать $\nu \in (0, 1)$ и $\nu \in [1, 500]$.

Усредненная характеристика τ^{KNN} принимает значения в окрестности числа 189 независимо от параметра ν.





1.1.3 Распределение SkewNormal с параметром α.

Зафиксируем размер выборки $n = 100$ и количество соседей $k = 5$. Число итераций для метода Монте-Карло равно 1000.

Будем перебирать $\theta = \{0.001, 0.01, 0.1, 0.5, 0.75, 1, 3, 5, 10, 15, 20, 50, 100, 500, 1000\}$.

Результаты

Усредненная характеристика τ^{KNN} при любых значениях параметра τ приближенно равна 9, но при больших значениях это приближение становится более заметным.

1.1.4 Распределение Normal с параметром-дисперсией σ и матожиданием 0.

Зафиксируем размер выборки $n = 100$ и количество соседей $k = 5$. Число итераций для метода Монте-Карло равно 1000.

Будем перебирать $\nu = \{0.001, 0.01, 0.1, 0.5, 0.75, 1, 3, 5, 10, 15, 20, 50, 100, 500, 1000\}$.

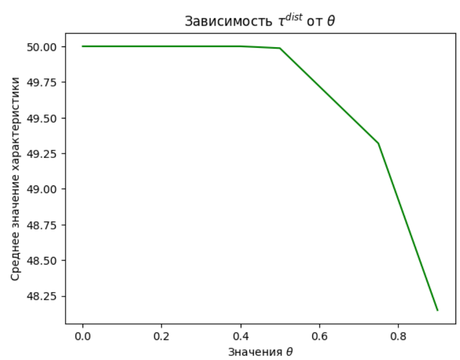
Результаты

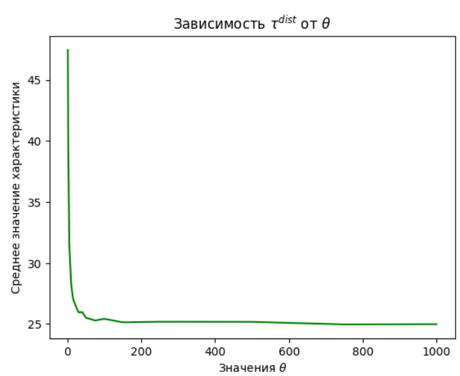
Усредненная характеристика τ^{KNN} принимает значения в окрестности числа 9 независимо от параметра ν. Но при больших значениях параметра можно заметить здесь, что характеристика τ^{KNN} начинает отклоняться от своего среднего значения.

1.2 Характеристика τ^{dist}.

1.2.1 Распределение LogNormal с $\mu = 0$ и параметром θ.

Зафиксируем размер выборки $n = 50$ и расстояние $dist = 5$. Число итераций для метода Монте-Карло равно 1000.

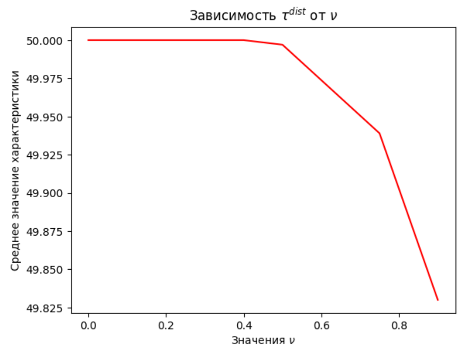


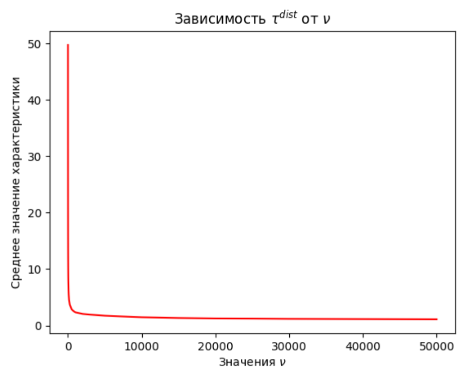


С увеличением θ среднее значение характеристики τ^{dist} уменьшается, и при $\theta \approx 100$ принимает значение 25. Затем на больших θ среднее значение немного увеличивается и колеблется около 27.

1.2.2 Распределение Exp с параметром λ.

Зафиксируем размер выборки $n = 50$ и расстояние $dist = 5$. Число итераций для метода Монте-Карло равно 1000.





При больших ν среднее значение τ^{dist} стремится к 1.

Замечание: для экспоненциального распределения видно более резкое уменьшение значения характеристики по сравнению с lognormal распределением.

1.2.3 Распределение SkewNormal с параметром α.

Зафиксируем размер выборки $n = 100$ и расстояние $dist = 1$. Число итераций для метода Монте-Карло равно 1000.

Будем перебирать
$\theta = \{0.001, 0.01, 0.1, 0.5, 0.75, 1, 3, 5, 10, 15, 20, 50, 100, 500, 1000, 10000, 150000, 300000, 500000, 15000000\}$.

Результаты.

Характеристика τ^{dist} при $\theta \in (0, 1)$ принимает в среднем значение 5, а при больших θ принимает значения, близкие к 3. Это хорошо видно на графике.

1.2.4 Распределение Normal с параметром-дисперсией σ и матожиданием 0.

Зафиксируем размер выборки $n = 100$ и расстояние $dist = 5$. Число итераций для метода Монте-Карло равно 1000.

Будем перебирать
$\nu = \{0.001, 0.01, 0.1, 0.5, 0.75, 1, 3, 5, 10, 15, 20, 50, 100, 500, 1000, 10000, 150000, 300000, 500000, 15000000\}$

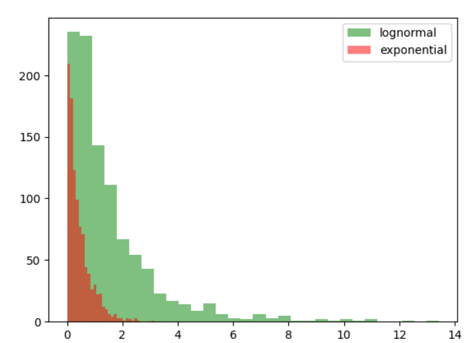
Результаты

Характеристика τ^{dist} при $\nu \in (0, 0.5)$ принимает значение 1 (т.е. при таких ν граф – полный). С увеличением параметра растет среднее значение характеристики (можно посмотреть здесь).

Глава 2

Исследовать, как ведет себя числовая характеристика τ в зависимости от параметров процедуры построения графа и размера выборки при фиксированных значениях $\theta = \theta_0$ и $\nu = \nu_0$.

[Тасина вставка] Сначала посмотрим на LogNormal и Exp распределения при данных θ_0 и ν_0:



Видно, что для построения дистанционного графа брать `dist` > 5 бессмысленно, т.к. при больших значениях `dist` число рёбер в графе стремится к $\binom{n}{2}$, где n – число вершин, соответственно, хроматическое число становится равным n для обоих распределений.

Для нас же важно понимать, как различить между собой эти распределения, поэтому гораздо интереснее смотреть на графы с меньшим числом рёбер и смотреть на `dist` ≤ 5.

2.1 Характеристика τ^{KNN}.

2.1.1 Распределение LogNormal с $\mu = 0$ и $\theta = \theta_0 = 1$ и распределение Exp с параметром $\nu = \nu_0 = \frac{1}{\sqrt{e^2 - e}}$.

Картинку смотрите тут: `11.png`.

Замечания:

- τ^{KNN} для Exp распределения растет медленнее, чем для LogNormal распределения.

- Стоит посмотреть на значения характеристики при бОльших размерах выборки (еще не смотрела), т.к. при увеличении выборки разница между значениями растет.

Возможно, еще добавлю что-то сюда.

2.1.2 Распределение SkewNormal с параметром $\alpha_0 = 1$.

Будем перебирать параметры с 1000 итерациями метода Монтэ-Карло:
1. $n\_samples = \{[1, 5, 10, 25, 50, 100, 300]\}$
2. $k\_neighbours = \{1, 3, 5, 7, 9, 15, 20\}$

Результаты

Можно заметить, что средняя величина характеристики τ^{KNN} увеличивается, по мере роста перебираемых параметров. Но также часто встречаются ситуация, когда среднее значение совпадает с реальным.

2.1.3 Распределение Normal с параметром-дисперсией $\sigma_0 = 1$ и матожиданием 0.

Будем перебирать параметры с 1000 итерациями метода Монтэ-Карло:
1. $n\_samples = \{[1, 5, 10, 25, 50, 100, 300]\}$
2. $k\_neighbours = \{1, 3, 5, 7, 9, 15, 20\}$

Результаты

Можем наблюдать такую же тенденцию – с ростом параметров растет среднее значение характеристики, даже значения принимаются такие же со сдвигом на небольшой ϵ.

2.2 Характеристика τ^{dist}.

2.2.1 Распределение LogNormal с $\mu = 0$ и $\theta = \theta_0 = 1$ + распределение Exp с параметром $\nu = \nu_0 = \frac{1}{\sqrt{e^2 - e}}$.

Картинку смотрите тут: `14.png`.

Пара замечаний:

- τ^{dist} для Exp распределения растет быстрее, чем для LogNormal распределения.

- пока что τ^{dist} рассматривать и изучать приятнее/проще, чем τ^{KNN}.

2.2.2 Распределение SkewNormal с параметром $\alpha_0 = 1$.

Будем перебирать параметры с 1000 итерациями метода Монтэ-Карло:
1. $n\_samples = \{[1, 5, 10, 25, 50, 100, 300]\}$
2. $dists = \{0.001, 0.01, 0.1, 0.5, 1, 3, 5\}$

Результаты

Можно заметить, что больше всего на значение характеристики τ^{dist} влияет параметр $n\_samples$, а с увеличением параметра $dist$ увеличивается количество ребер из-за этого уменьшается количество независимых вершин.

2.2.3 Распределение Normal с параметром-дисперсией $\sigma_0 = 1$ и матожиданием 0.

Будем перебирать параметры с 1000 итерациями метода Монтэ-Карло:
1. $n\_samples = \{[1, 5, 10, 25, 50, 100, 300]\}$
2. $k\_neighbours = \{1, 3, 5, 7, 9, 15, 20\}$

Результаты

Для каждого значения параметра $n\_samples$ можем заметить довольно плотное распределение среднего значения характеристики τ^{dist}, но с ростом этого параметра растет количество выбросов и колебания.

Глава 3

Построить множество A в предположении $\theta = \theta_0$ и $\nu = \nu_0$ при максимальной допустимой вероятности ошибки первого рода $\alpha = 0.05$. Оценить мощность полученного критерия.

3.1 Характеристика τ^{KNN}.

Для визуализаций смотрите картинку `15.png`.

Распределения смешаны между собой, и невозможно определить какую-либо границу между ними. Выходит, что работать с KNN-графом довольно трудно. Посмотрим на дистанционный граф.

Но я все-таки хотела бы посмотреть на KNN-граф на большом числе вершин и уже после этого определиться с ответом.

3.2 Характеристика τ^{dist}.

Для визуализаций смотрите картинку `16.png`.

А вот тут четко просматривается граница между двумя распределениями, особенно при бОльших размерах выборки. Построим множество A (синие пунктирные линии на графике).

Посмотреть картинку можно тут: `17.png`.

При увеличении $dist$ и размера выборки граница между двумя распределениями становится более явной. И даже есть примеры, когда мощность максимальна и равна 1. Однако при небольших размерах выборки и маленьких $dist$ распределения довольно трудно различимы. В таких случаях и ошибка первого рода большая.

Вывод: если дана выборка достаточного размера, то при выборе правильного $dist$ (кажется, что значения 2, 3, 5 подходят) можно построить дистанционный граф так, что по хроматическому числу этого графа будет возможно определить исходное распределение.