

Проект по случайным графам

Чегодаева Таисия и Купряков Дмитрий, ПАДПИ, 2 курс

29 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 $\theta \rightarrow 1$.

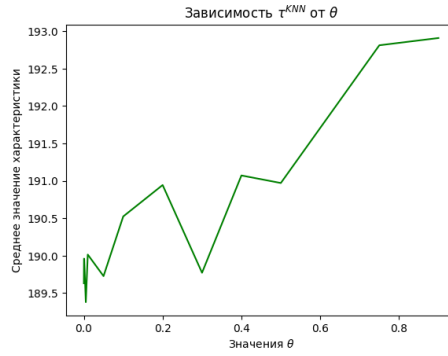


Рис. 1.1: LogNormal распределение с варьирующимся параметром θ

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

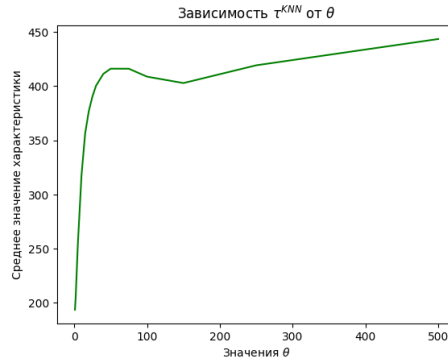


Рис. 1.2: LogNormal распределение с варьирующимся параметром θ

При $\theta \in [1, 100]$ наблюдается резкий рост среднего значения характеристики, после этого кривая выравнивается и выходит на плато.

1.1.2 Распределение Ехр 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 .

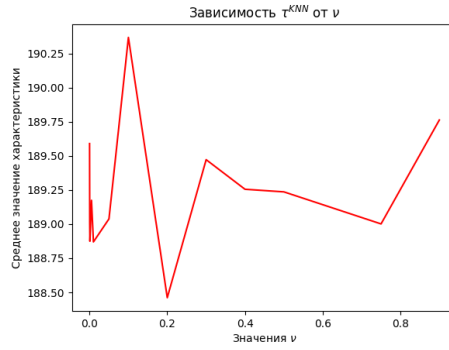


Рис. 1.3: Экспоненциальное распределение с варьирующимся параметром ν

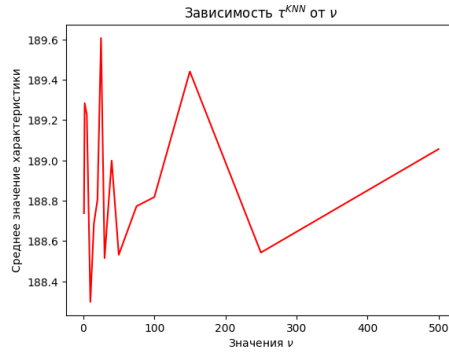


Рис. 1.4: Экспоненциальное распределение с варьирующимся параметром ν

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.

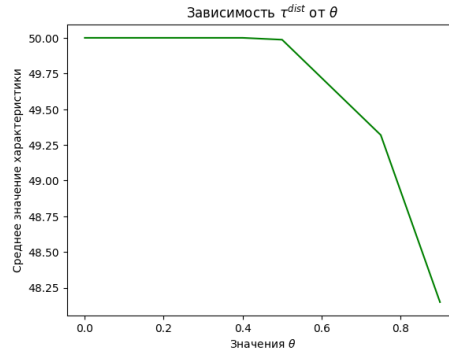


Рис. 1.5: LogNormal распределение с варьирующимся параметром θ

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 Exr с параметром λ .

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

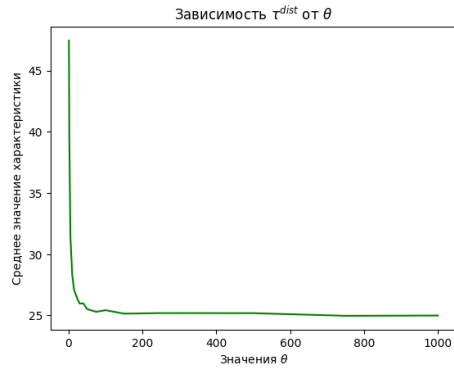


Рис. 1.6: LogNormal распределение с варьирующимся параметром θ

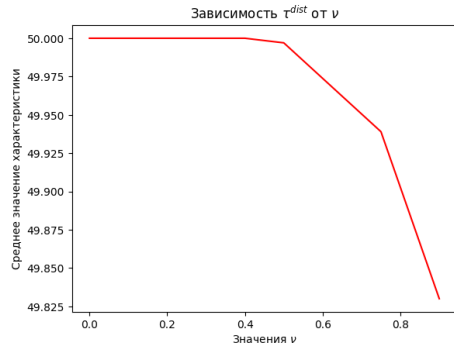


Рис. 1.7: Экспоненциальное распределение с варьирующимся параметром ν

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

$\theta = \{0.001, 0.01, 0.1, 0.5, 0.75, 1, 3, 5, 10, 15, 20, 50, 100, 500, 1000, 10000, 150000, 300000, 500000, 1500000\}$.

Результаты.

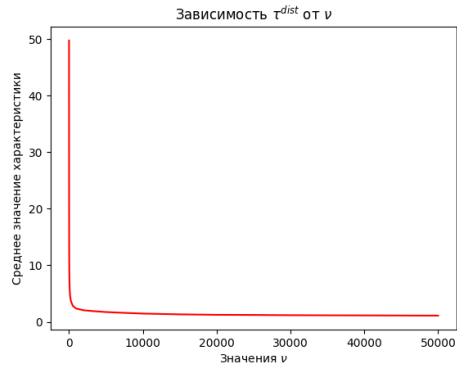


Рис. 1.8: Экспоненциальное распределение с варьирующимся параметром ν

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

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

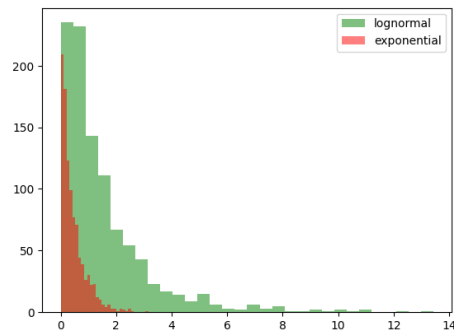
Результаты

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 растёт медленнее, чем для LogNormal распределения.

- При увеличении выборки разница между значениями характеристики для двух разных распределений растёт. Это, конечно, хорошо, но computationally может быть неприятно. Хочется смотреть и на небольшую выборку, а в нашем случае разница между распределениями на небольших размерах почти не различима.

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 $\theta = \theta_0 = 1 + \frac{1}{\sqrt{e^2 - e}}$ 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 трудно определить какую-либо границу между ними. Выходит, что работать с KNN-графом довольно неприятно. Посмотрим на дистанционный граф.

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.

Часть II

Несколько характеристик проверки гипотезы.

Глава 4

Тасина часть

Сгенерировала три набора данных для дистанционных графов:

- 10000 значений характеристик для графов на 25 вершинах (по 5000 значений для каждого распределения).
- 10000 значений характеристик для графов на 100 вершинах (по 5000 значений для каждого распределения).
- 2000 значений характеристик для графов на 500 вершинах (по 1000 значений для каждого распределения). Размер меньше, чем у предыдущих двух наборов, т.к. для графов на 500 вершинах классификация довольно простая, и это видно еще из 1 части задачи по построению множества A .

Для дистанционных графов зафиксировала $dist = 1$, т.к. для больших значений решать задачу не очень интересно – тогда граф становится все больше похож на полный. А при маленьких значениях многие характеристики для обоих распределений совпадают. Ну и судя по графикам из 1 части, именно выбранное значение $dist$ рассмотреть будет интересно, и в то же время не очень сложно.

Я использую следующие характеристики:

- хроматическое число графа.
- число треугольников.
- максимальная степень вершины.

Размер максимального независимого множества не использую, т.к. он выражается через хроматическое число (при условии, что хроматическое число равно кликовому числу – что верно в дистанционных графах):

$$\alpha(G) = n - \chi(G),$$

где n – число вершин в графе. Соответственно, никакой новой информации эта характеристика не несет.

В качестве классификаторов я использую логистическую регрессию, дерево решений и случайный лес.

4.1 Число вершин в графе = 25.

Для графов на 25 вершинах лучше всего отработала логистическая регрессия. Возможно, это произошло, т.к. между характеристиками двух распределений граница не такая уж и заметная, поэтому эвристические алгоритмы (типа дерева решений или случайного леса) работали хуже.

Если смотреть на ошибку первого рода, то она не превосходит $\alpha = 0.05$, т.е. все три классификатора подходят как статистические критерии.

При этом лучшим алгоритмом по мощности и по точности является логистическая регрессия.

4.2 Число вершин в графе = 100.

В этом случае результаты вообще прекрасные – граница между характеристиками разных распределений стала более видимой, поэтому и точность стала выше.

И конечно, все три классификатора дают ошибку первого рода меньше, чем $\alpha = 0.05$.

Тут лучше всего отработали дерево решений и случайный лес.

4.3 Число вершин в графе = 100.

Тут, как и предполагалось, высокая точность при использовании всех алгоритмов, т.к. разделение между распределениями довольно четкое, и это было видно еще в 1 части задания.

Если смотреть на важность признаков, то довольно понятен алгоритм работы дерева решений: в основе его работы лежит отбор признаков через предикаты, и при взгляде на картинку выше становится понятно, что если использовать, например, такой предикат: [Хроматическое число ≤ 350], то классификация распределений уже будет оптимальной. Поэтому признак "хроматическое число" является самым важным для дерева решений (и, судя по значениям, единственным используемым).

В этом случае все алгоритмы показали ошибку первого рода $= 0$ и мощность $= 1$.

4.4 Общий вывод по трем экспериментам.

- Для логистической регрессии наиболее важными признаками являются хроматическое число и максимальная степень вершины.
- Для каждого набора данных дерево решений считало важным лишь 1 признак, и он был разным для разных наборов. Кажется, это обусловлено лишь особенностью задачи и данных (конкретнее: по большей части все зависит от числа вершин, на которых строим граф), другого разумного объяснения не придумала.
- Случайный лес – ансамбль нескольких деревьев решений, поэтому все признаки относительно одинаковые по важности.

Но в общем и целом все получилось, точность высокая у всех алгоритмов!