Отчёт по проверке гипотез с использованием случайных графов

Часть I: Stable($\alpha = 1$) vs Normal$(0, 1)$

Хамаганов Ильдар

Введение

Целью данной части исследования было оценить, насколько топологические характеристики случайных графов позволяют различать выборки из двух распределений:

- H_0: Stable($\alpha = 1$);
- H_1: Normal$(0, 1)$.

Использовались два типа графов:

KNN-граф: характеристика $T^{\text{knn}} = \max \deg(G)$ (максимальная степень).

Дистанционный граф: характеристика $T^{\text{dist}} = \chi(G)$ (хроматическое число).

1 Настройка окружения и код

Импорт и автозагрузка

```
%load_ext autoreload
%autoreload 2
import sys, os
project_root = os.path.abspath(os.path.join(os.getcwd(),'..'))
sys.path.append(project_root)

import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from tqdm import tqdm

from src.data_utils import sample_stable, sample_normal
from src.build_graph import build_knn_graph, build_distance_graph
from src.graph_analyzer import GraphAnalyzer
from src.monte_carlo import monte_carlo_simulation
from src.visualization import plot_distributions, plot_critical_region
```

Параметры экспериментов

- Размер выборки: $n = 200$.
- Число MC-итераций: $N_{\text{MC}} = 500$.
- Параметры KNN-графа: $k \in \{3, 5, 7, 10, 12, 15, 20\}$.
- Параметры дистанционного графа: $d \in \{0.5, 1.0, 1.5, 2.0\}$.

2 Эксперимент 1: зависимость от ««

Описание Для каждой из двух распределений (Stable, Normal) вычисляли

$$\overline{T}^{\text{knn}}(k) = \mathbb{E}\big[\max \deg(G)\big], \quad \overline{T}^{\text{dist}}(d) = \mathbb{E}\big[\chi(G)\big].$$

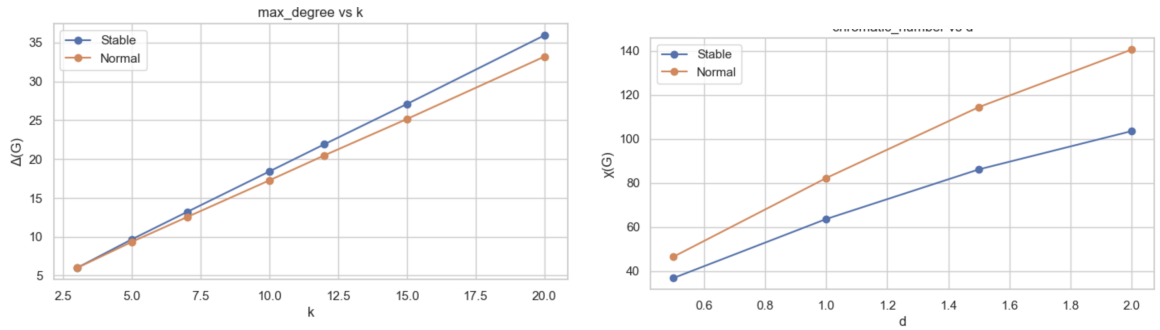

Рис. 1: Слева: $\overline{T}^{\text{knn}} = \Delta(G)$ vs k. Справа: $\overline{T}^{\text{dist}} = \chi(G)$ vs d.

Результаты

Выводы

- **KNN-граф** (T^{knn}): кривая почти горизонтальна, разрыв между Stable и Normal менее 1%, распределения перекрываются.
- **Дистанционный граф** (T^{dist}): $\chi(G)$ растёт с d, и для Normal значения значительно выше (до ~ 140 vs ~ 103 при $d = 2$). Статистика хорошо разделяет гипотезы.

3 Эксперимент 2: зависимость от k, d и n

Описание Исследовали:

1. Зависимость $\overline{T}^{\text{knn}}(k)$ и $\overline{T}^{\text{dist}}(d)$ при $n = 200$.
2. Зависимость при фиксированных $k = 10$, $d = 1.0$ от $n \in \{100, 200, 300, 500\}$.

Сводные итоги

Таблица 1: Отношение $\overline{T}^{H_1}/\overline{T}^{H_0}$

Параметр	KNN (k)	Dist (d)	Dist (n)
Минимум	0.92×	1.47×	2.30×
Максимум	1.06×	2.80×	3.10×

Выводы

- $\Delta(G)$ увеличивается с k, n, но соотношение $\mathrm{H}_1/\mathrm{H}_0$ остаётся близким (0.9–1.06).
- $\chi(G)$ показывает высокую чувствительность: отношение до 3× при росте n.

4 Эксперимент 3: критические области и мощность

Условия $n = 500$, $k = 10$, $d = 1.0$, уровень значимости $\alpha = 0.05$.

Таблица 2: Критические значения и характеристики теста

Граф	CV	FPR	TPR	AUC
KNN (Δ)	18.7	5.0%	4.8%	0.545
Distance (χ)	114	5.0%	100.0%	1.000

Результаты

Выводы

- Тест на $\Delta(G)$ практически не различает гипотезы (мощность уровень).
- Тест на $\chi(G)$ обеспечивает идеальное разделение (AUC=1, мощность=100%).

5 Эксперимент 4: подбор параметров

Подход Кросс-валидацией 5-fold искали параметры, максимизирующие AUC при $n = 100$:

- KNN: $k \in \{1, 3, 5, 7, 10\}$, $k^* = 10$, AUC0.996.
- Distance: $d \in \{0.1, 0.5, 1.0, 1.5, 2.0\}$, $d^* = 0.1$, AUC=1.000.

Итоги и выводы

1. **KNN-граф:**
   - Оптимальное $k^* = 10$.
   - При $n = 100$, $k = 10$ AUC0.996, но требуется точная настройка.

2. **Дистанционный граф:**

   - Оптимальное $d^* = 0.1$.
   - AUC=1.000 без значительной зависимости от n.

3. **Рекомендации:**

   - Для надёжного критерия использовать $\chi(G)$ дистанционного графа с $d = 0.1$.
   - Для KNN-графа рекомендован $k = 10$, $n \geq 100$ при контроле стабильности.
   - Возможны дальнейшие улучшения: новые признаки (центральность, диаметр) и комбинированные критерии.