

# Лабораторная работа №6

## Кластеризация (DBSCAN, OPTICS)

---

### Цель:

Ознакомиться с методами кластеризации модуля Sklearn

- [Контрольные вопросы](#)

### Выполнение:

#### 1. Загрузка данных:

---

1.1. Скачать датасет по ссылке <https://www.kaggle.com/datasets/muratkokludataset/pistachio-dataset>.

Данные представлены в виде xls(xlsx) файла.

#### Для информации

Для чтения файлов xls(xlsx) можно воспользоваться командой `pandas.read_excel` или преобразовать файл xls(xlsx) в формат csv.

Для преобразования xls(xlsx) в формат csv необходимо сохранить в файл (лист)

`Pistachio_Dataset` из файла `Pistachio_16_Features_Dataset.xls(xlsx)` в файл `Pistachio_16_Features_Dataset.csv`.

В файле `Pistachio_16_Features_Dataset.csv` в качестве разделителя значений может использоваться `;`.

В этом случае, при использовании `pandas.read_csv` необходимо указывать `sep=';'`

1.2. Загрузить данные из `Pistachio_16_Features_Dataset` в датафрейм. Подготовить данные для обработки (при необходимости - убрать столбец с метками и, если в наборе данных присутствуют пропуски, исключить наблюдения с пропущенными значениями и т.д.).

1.3. Понизить размерность пространства данных до 2, определить величину дисперсии, которую объясняют две компоненты. Выполнить кластеризацию методом k-средних для трех кластеров, подобрать оптимальное значение `n_init`. Вывести на экран получившееся разделение данных на кластеры.

## 2. DBSCAN

---

Далее, все операции осуществляются с данными из пункта 1.2.

2.1. Изучить теоретические сведения о кластеризации DBSCAN

- [Кластеризация пространственных данных – плотностные алгоритмы и DBSCAN](#)
- [Интересные алгоритмы кластеризации, часть вторая: DBSCAN / Хабр](#)
- [DBSCAN — Википедия](#)
- [DBSCAN — scikit-learn](#)
- [Кластеризация, DBSCAN — scikit-learn](#)
- [Clustering — scikit-learn](#)

2.2. Ответить на вопрос:

- Обосновать свой ответ.
- В случае ответа  - произвести стандартизацию данных ([StandardScaler](#)).

2.3. Провести кластеризацию методом DBSCAN с параметрами по умолчанию. Вывести метки кластеров, количество кластеров, а также процент наблюдений, которые кластеризовать не удалось.

- Пример: [Demo of DBSCAN clustering algorithm — scikit-learn 1.5.2 documentation](#)
- **Дать пояснения:**
  1. Что означает каждый из параметров (и его значение) в функции [DBSCAN](#)

2.4. Постройте график количества кластеров и процента не кластеризованных наблюдений в зависимости от максимальной рассматриваемой дистанции между наблюдениями. Минимальное значение количества точек образующих, кластер оставить по умолчанию.

2.5. Постройте график количества кластеров и процента не кластеризованных наблюдений в зависимости от минимального значения количества точек, образующих кластер. Максимальную рассматриваемую дистанцию между наблюдениями оставьте по умолчанию

2.6. Определите значения параметров, при котором количество кластеров получается от 5 до 7, и процент не кластеризованных наблюдений не превышает 12%.

2.7. Понижьте размерность данных до 2 . Визуализируйте результаты кластеризации, полученные в пункте 2.6 (метки должны быть получены на данных до уменьшения размерности).

## 3.OPTICS

---

3.1. Изучите теоретический материал

- [Кластеризация, OPTICS — scikit-learn](#)
- [Алгоритм кластеризации OPTICS — Википедия](#)
- [OPTICS — scikit-learn](#)
- [Demo of OPTICS clustering algorithm — scikit-learn](#)

3.2. Опишите параметры функции [OPTICS](#), а также какими атрибутами они обладают.

- Что означают значения параметров принимаемых по умолчанию?

3.3. Найдите такие параметры *max\_eps* и *min\_samples* функции OPTICS, при которых получаются результаты близкие к результатам DBSCAN из пункта 2.6. Вывести результаты на экран.

- **Дать пояснения:**

1. В чем отличия от алгоритма метода OPTICS от метода DBSCAN

2. Как было определено, что результаты получились близкими? (Картинки похожи - отвечать нельзя!)

3.4. Постройте график достижимости (reachable plot) (см. [Demo of OPTICS clustering algorithm — scikit-learn](#))

- **Дать пояснения:**

1. "График `reachable plot` показывает, ..."
2. "Из графиков видно, что ..."

3.5. Исследуйте работу метода OPTICS с использованием различных метрик (`metric`). Исследуйте результаты кластеризации с использованием не менее 8 различных метрик.

- **Дать пояснения:**

1. Что означает и как влияет на результат выбранная метрика?
  2. Использование какой метрики дало наилучший результат и почему?
-

# Контрольные вопросы

---

## DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

---

Несколько контрольных вопросов по теме DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

### Базовые понятия:

1. Что такое кластер в контексте DBSCAN?
2. Что такое *шум* в контексте DBSCAN?
3. Какие два ключевых параметра используются в DBSCAN и как они влияют на результат кластеризации? ( $\epsilon$  (эпсилон) и MinPts)
4. Объясните понятия *основной точки*, *граничной точки* и *точки шума*.
5. Как DBSCAN определяет плотность точек в наборе данных?

### Применение алгоритма:

6. Опишите основные шаги алгоритма DBSCAN.
7. Как DBSCAN обрабатывает кластеры различной формы и размера?
8. В чем преимущество DBSCAN перед k-means?
9. В чем недостатки DBSCAN?
10. Как выбрать оптимальные значения для параметров  $\epsilon$  и MinPts?
11. Как влияет размерность данных на производительность DBSCAN? ("проклятие размерности")
12. Какие методы можно использовать для определения оптимального значения  $\epsilon$ ?

### Сравнение с другими методами:

13. Сравните DBSCAN с k-means. В каких случаях предпочтительнее использовать один алгоритм вместо другого?
14. Сравните DBSCAN с иерархической кластеризацией.

### Продвинутые вопросы:

15. Как DBSCAN может быть использован для обнаружения аномалий?
16. Как можно ускорить работу DBSCAN для больших наборов данных?
17. Что такое HDBSCAN и чем он отличается от DBSCAN?

## OPTICS (Ordering Points To Identify the Clustering Structure)

---

Несколько контрольных вопросов по теме OPTICS (Ordering Points To Identify the Clustering Structure):

**Базовые понятия:**

1. Какова основная цель алгоритма OPTICS?
2. Как OPTICS связан с DBSCAN?
3. Что такое *core-distance* и *reachability-distance*? Объясните их значения и как они вычисляются.
4. Что представляет собой *упорядочивание точек* в OPTICS и почему оно важно?
5. Как визуализируются результаты OPTICS? Что такое *reachability-distance plot*?

**Применение алгоритма:**

6. Опишите основные шаги алгоритма OPTICS.
7. Как интерпретировать *reachability-distance plot* для определения кластеров?
8. Какие параметры используются в OPTICS и как они влияют на результат? ( $\epsilon$  (эпсилон) и MinPts)
9. В чем преимущества OPTICS по сравнению с DBSCAN?
10. В чем недостатки OPTICS?

**Сравнение с другими методами:**

11. Сравните OPTICS с DBSCAN. В каких случаях предпочтительнее использовать OPTICS?
12. Сравните OPTICS с иерархической кластеризацией.

**Продвинутые вопросы:**

13. Как можно автоматически извлекать кластеры из *reachability-distance plot*? Какие методы существуют для этого? (например, *steep down areas*, *Xi-steep down areas*)
14. Как параметр  $\epsilon$  влияет на результаты OPTICS, и почему его часто устанавливают в "бесконечность"?
15. Как OPTICS справляется с кластерами различной плотности?
16. Как можно ускорить работу OPTICS для больших наборов данных?
17. Какие модификации OPTICS существуют, и какие проблемы они решают? (например, DeLi-Clu, HiSC)