

# Лабораторная работа 3

## Кластеризация

(<https://habr.com/ru/company/ods/blog/325654/>)

### ЗАДАНИЕ

**(!!!) При построении графиков обязательно подписывайте оси и делайте легенду, чтобы было понятно, что изображено**

**(!!!) Можно взять любую базу данных**

По ссылке (<https://scikit-learn.org/stable/modules/clustering.html>) находится обзор моделей кластеризации в sklearn.

Задачи:

#### **Часть 1 – работа с данными:**

Опишите используемую базу данных (или несколько, если хотите). Постройте диаграммы примерно как в лабе по классификации. Опишите выводы – замеченные закономерности и преобразования с базой, которые вы сделаете для дальнейшей работы.

#### **Часть 2 – классификация:**

Цель задания – добиться как можно большей точности по каждому из алгоритмов кластеризации. В отчете приведите описание того, как вы в итоге пришли к полученной точности, что вы делали для того, чтобы её повысить, и почему вы считаете, что сделали всё, что могли. Работать необходимо как с базой (признаками), так и с параметрами алгоритма.

**(!!!) Для определения точности используйте кросс-валидацию. Подсказка: признаки можно использовать не все, и можно даже их менять.**

**(!!!) Внимательно подойдите к выбору метрик и их анализу**

**Итак, в части 2 вам необходимо:**

1. Реализовать решение задачи кластеризации (просто как-то) каждым из рассмотренных алгоритмов.
2. Для каждого алгоритма составить список/таблицу настроечных параметров, описать их смысл, в каких случаях что используется и как это влияет (пусть предположительно) на результат. То есть, параметры, где всё очевидно из названия, расписывать не нужно, а для неочевидных и важных описать, что они означают, зачем они нужны, и как необходимо их менять в зависимости от данных. То есть, как вы будете их использовать для подстройки алгоритма.
3. Добиться максимально точного результата для каждого алгоритма, при этом сравнивать ещё время обучения и время работы моделей. Описать логику поиска лучших моделей, привести рассуждения, таблицы сравнения моделей, по которым вы определили лучший вариант, и т.д.
4. Сформулировать рекомендации по решению задачи кластеризации рассмотренными алгоритмами – в каких случаях какие алгоритмы и настроечные параметры нужно использовать.

**Вам также могут пригодиться:**

стандартизация/нормализация:

(<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>)

и метод главных компонент:

(<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>)

### **Заметка по базам данных**

Если у вас есть собственные данные, по которым нужно что-то классифицировать, либо вы нашли интересную базу данных – напишите мне на почту ([rtc.machinelearning@gmail.com](mailto:rtc.machinelearning@gmail.com)) письмо с темой “Подтверждение своей базы” и описанием базы. Я посмотрю и сообщу, подойдет она или нет.

Есть такие базы:

◆ New York City Airbnb Open Data

<https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>

На гугл-диске курса в папке *Datasets* → *AB\_NYC\_2019.csv*

◆ Титаник

Если возьмете эту базу, я ожидаю очень хорошего и подробного отчета, потому что по ней много материала.

<https://www.kaggle.com/c/titanic>

Данные можете скачать там же, либо взять с диска курса

◆ Продажи и рейтинги видеоигр

<https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/data>

◆ База с анонимизированными признаками автомобилей, где целевая переменная — время, необходимое, чтобы автомобиль прошёл тест (datasets → Mercedes).

<https://www.kaggle.com/c/mercedes-benz-greener-manufacturing/data>

◆ База для определения вида движения человека по данным со смартфона (выглядит сложно, там довольно много признаков и лежит это всё не очень удобно и понятно, но данных много. Не уверена, что с ней что-нибудь получится, но попробовать, наверно, можно)

<http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

◆ Можете поискать ещё сами в таком вот сборнике:

<https://archive.ics.uci.edu/ml/>

[datasets.htmlformat=&task=clu&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table](https://archive.ics.uci.edu/ml/datasets.htmlformat=&task=clu&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table)

Там сразу выбрана кластеризация (слева в модификаторах поиска).