МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №6

по дисциплине «Машинное обучение»

Тема: Кластеризация (DBSCAN, OPTICS)

Студент гр. 6304 _____ Иванов Д.В.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами кластеризации модуля Sklearn.

Загрузка данных

1. Датасет скачан и загружен в датафрейм.

```
import pandas as pd
import numpy as np
data = pd.read_csv('CC GENERAL.csv').iloc[:,1:].dropna()
```

DBSCAN

1. Проведена кластеризация методом k-средних.

```
k_means = KMeans(init='k-means++', n_clusters=3, n_init=15)
k_means.fit(no_labeled_data)
```

2. Исходные данные стандартизированы.

```
data = np.array(data, dtype='float')
min_max_scaler = preprocessing.StandardScaler()
scaled_data = min_max_scaler.fit_transform(data)
```

3. Проведена кластеризацию методом DBSCAN при параметрах по умолчанию. Метки кластеров, количество кластеров и процент наблюдений, которые кластеризовать не удалось:

```
0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, -1
36
0.7512737378415933
```

4. Построен график количества кластеров и процента не кластеризованных наблюдений в зависимости от максимальной рассматриваемой дистанции между наблюдениями (рис. 1).

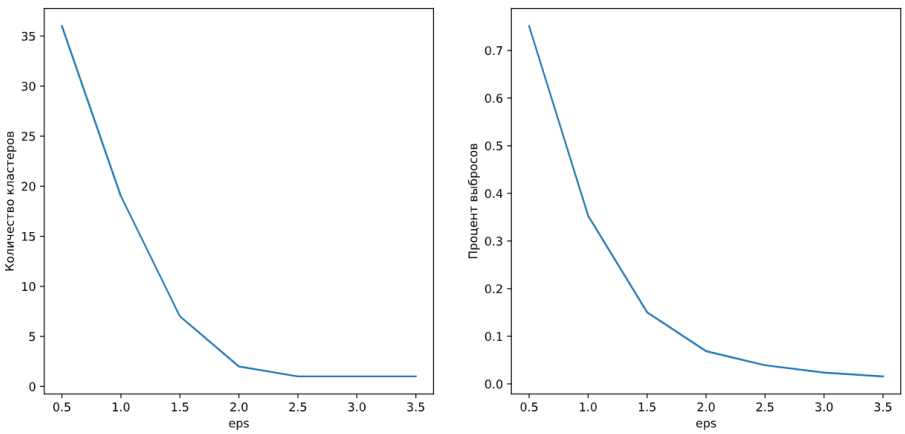

Рис. 1 — График количества кластеров и процента не кластеризованных наблюдений в зависимости от eps

5. Построен график количества кластеров и процента не кластеризованных наблюдений в зависимости от минимального значения количества точек, образующих кластер (рис. 2).

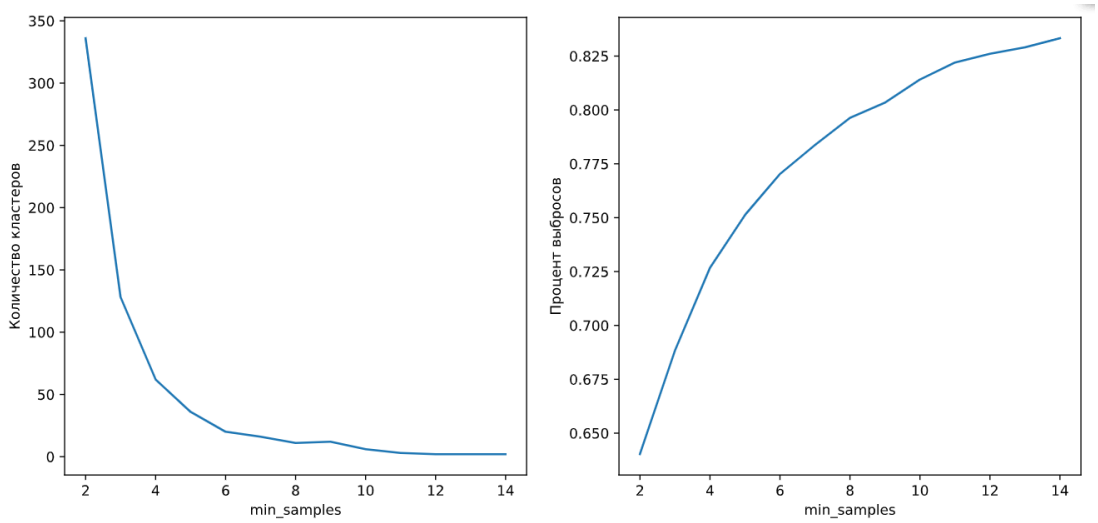

Рис. 2 — График количества кластеров и процента не кластеризованных наблюдений в зависимости от min_samples

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

```
samples, eps -> count_of_clusters, percent
3, 2.0 -> 6, 0.0629
```

7. Размерность данных понижена до 2 с использованием метода главных компонент, визуализированы результаты кластеризации.

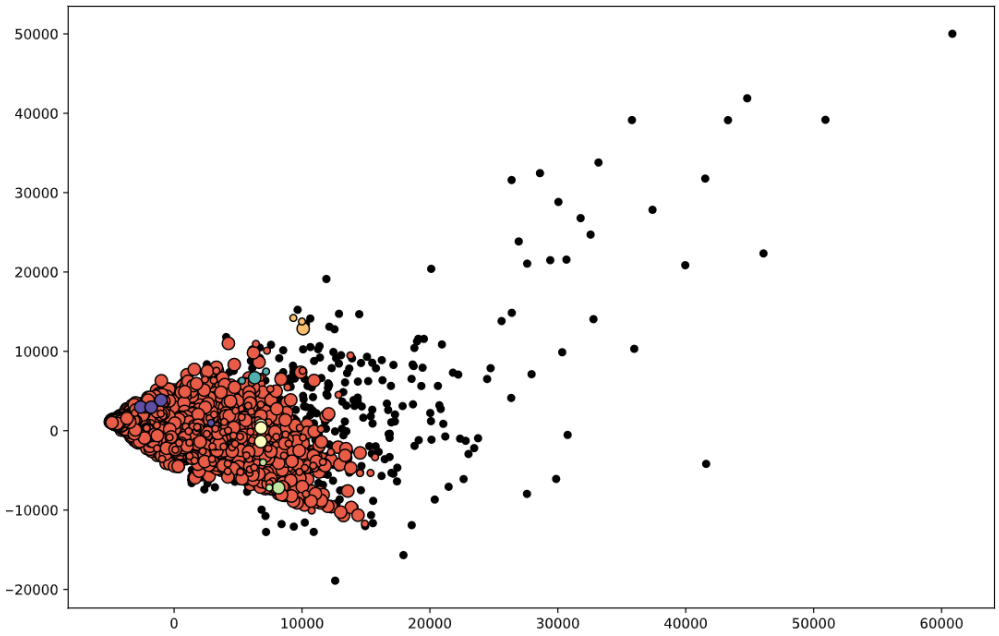

Рис. 3 — Результат кластеризации

OPTICS

1. Результаты кластеризации методом OPTICS похожи на результаты DBSCAN (количество полученных кластеров = 6, процент не кластеризованных данных = 6%.) при использовании метода кластеризации “dbscan”, соответственно *max_eps* = 2, *max_samples* = 3. Основные точки в OPTICS выбираются так же, как и в DBSCAN, но для каждой точки определяется дистанция достижимости.

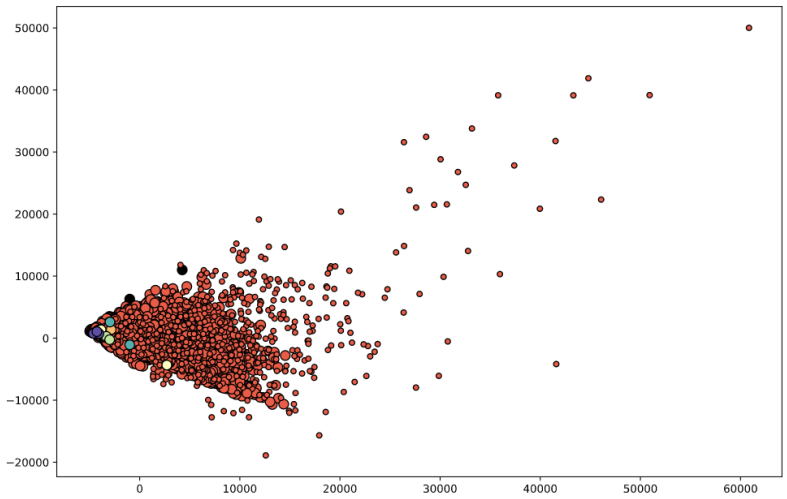

Рис. 4 — Результат кластеризации.

2. Построен график достижимости.

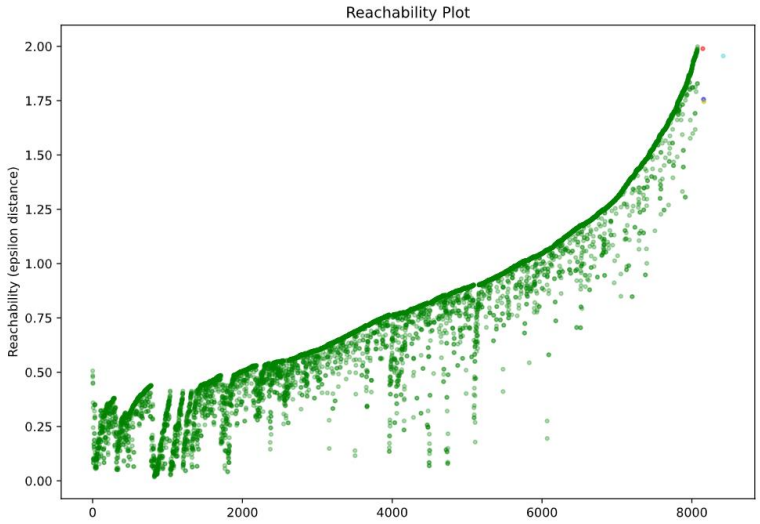

Рис. 17 — График достижимости

3. Исследована работа метода OPTICS с использованием метрик: *manhattan*, *euclidean*, *canberra*, *braycurtis*, *chebyshev*.

metric	min_samples	max_eps	cluster_num	non_clust
manhattan	10	1	6	0.81
		2	2	0.49
		10	1	0.01
	5	1	30	0.74
		2	12	0.44
		10	1	0.01
euclidean	10	1	3	0.42
		2	1	0.08
		10	1	0.00
	5	1	19	0.36
		2	2	0.07
		10	1	0.00
canberra	10	1	19	0.89
		2	6	0.55
		10	1	0.00
	5	1	18	0.50
		2	1	0.00
		10	1	0.00
braycurtis	10	1	1	0.00
		2	1	0.00
		10	1	0.00

	5	1	1	0.00
		2	1	0.00
		10	1	0.00
chebyshev	10	1	1	0.12
		2	1	0.02
		10	1	0.00
	5	1	3	0.10
		2	1	0.01
		10	1	0.0

Вывод

В ходе выполнения лабораторной работы произведено знакомство с кластеризацией методами DBSCAN и OPTICS из модуля Sklearn. Для исходного набора данных оба метода производят разбиение либо на большое количество кластеров, либо на один единственный, так же часто не классифицируется большая часть данных.