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. 6307 _____ Трофимов Н.И.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель

Ознакомиться с методами кластеризации модуля Sklearn

Выполнение

Был загружен датасет, убраны столбцы с метками, откинуты наблюдения с пропущенными значениями и стандартизированы.

DBSCAN

На основе стандартизированных данных был проведен метод кластеризации BDSCAN с параметрами по умолчанию. Выведены метки кластеров, количество кластеров и процент наблюдений, не попавших ни в один кластер. Результаты представлены ниже.

```
Метки кластеров - {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1
6, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
35, -1}
Количество кластеров - 36
Доля не попавших в кластер - 0.7512737378415933
```

Затем были построены графики количества кластеров и процента не кластеризованных наблюдений в зависимости от максимальной рассматриваемой дистанции между наблюдениями и от минимального значения количества точек, образующих кластер. Результаты представлены на рисунках 1 и 2.

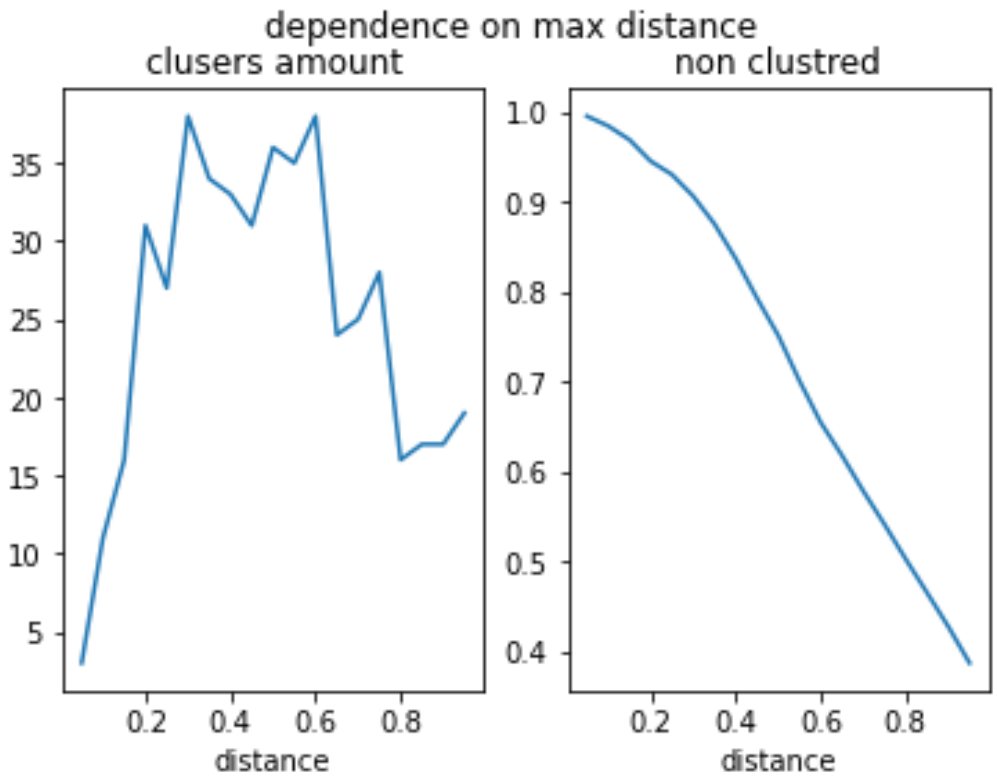

Рисунок 1. Зависимость от максимальной дистанции

Из графиков видно, что при малых значениях (<0.2) максимального расстояния кластера почти не образуются и большая часть точек становится некластеризованной, при значениях от 0.2 до 0.8 количество кластеров увеличивается, а некластеризованных данных становится меньше.

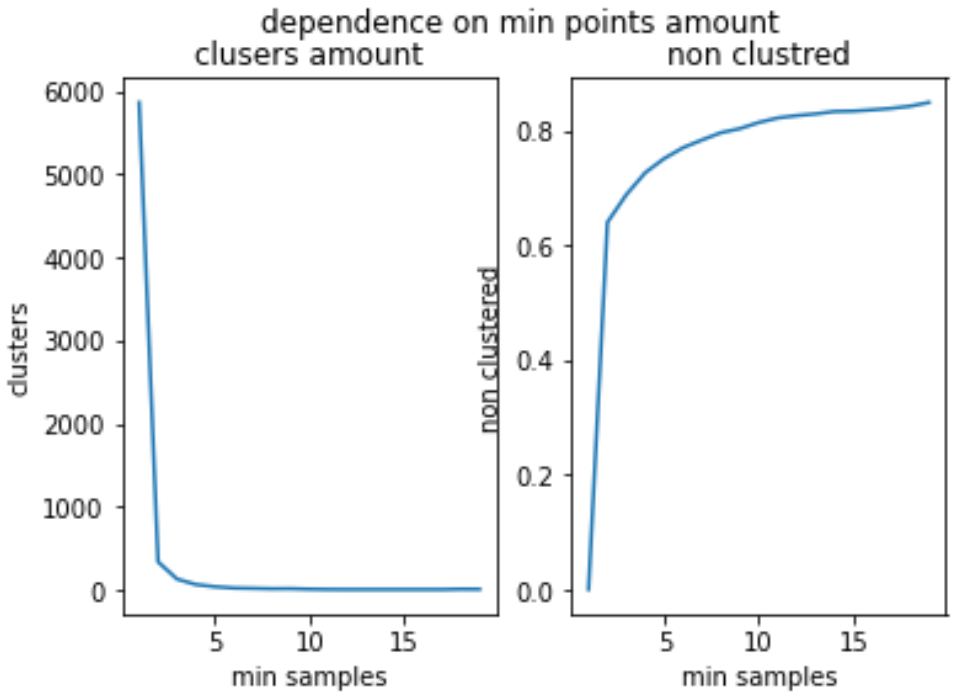

Рисунок 2. Зависимость от минимального количества точек

Из данных графиков видно, что при малых значениях (< 5) минимального количества точек для кластера – количество кластеров стремится к количеству точек, т.е. почти каждая точка представляет собой кластер, а некластеризованных данных (почти) нет. С увеличением данного параметра количество кластеров уменьшается, но растет количество некластеризованных данных.

Также были изучены все параметры, которые принимает алгоритм DBSCAN, они представлены в таблице 1.

Table 1. Параметры DBSCAN

Параметр	Описание
Eps	Максимальная дистанция между точками, чтобы считать их соседями

Min_samples	Минимальное количество точек вокруг необходимое для того, чтобы считать точку основной
Metric	Метрика, используемая для подсчета расстояний между точками
Metric_params	Дополнительные аргументы для метрики
Algorithm	Алгоритм, используемый методом NearestNeighbors для расчета расстояний и нахождений точек-соседей
Leaf_size	Размер листа, передаваемый в алгоритм
P	Степень метрики Минковского для расчета расстояний между точками
N_jobs	Количество параллельных вычислений

После построение графиков были определены значения параметров, при которых количество кластеров находится в диапазоне о 5 до 7 и процент некластеризованных данных не более 12%.

```
Min samples = 4::eps=1.75
```

При данных параметрах получается 5 кластеров и 9% некластеризованных данных.

После этого размерность была понижена до 2 методом главных компонент и данные визуализированы. Результат представлен на рисунке 3.

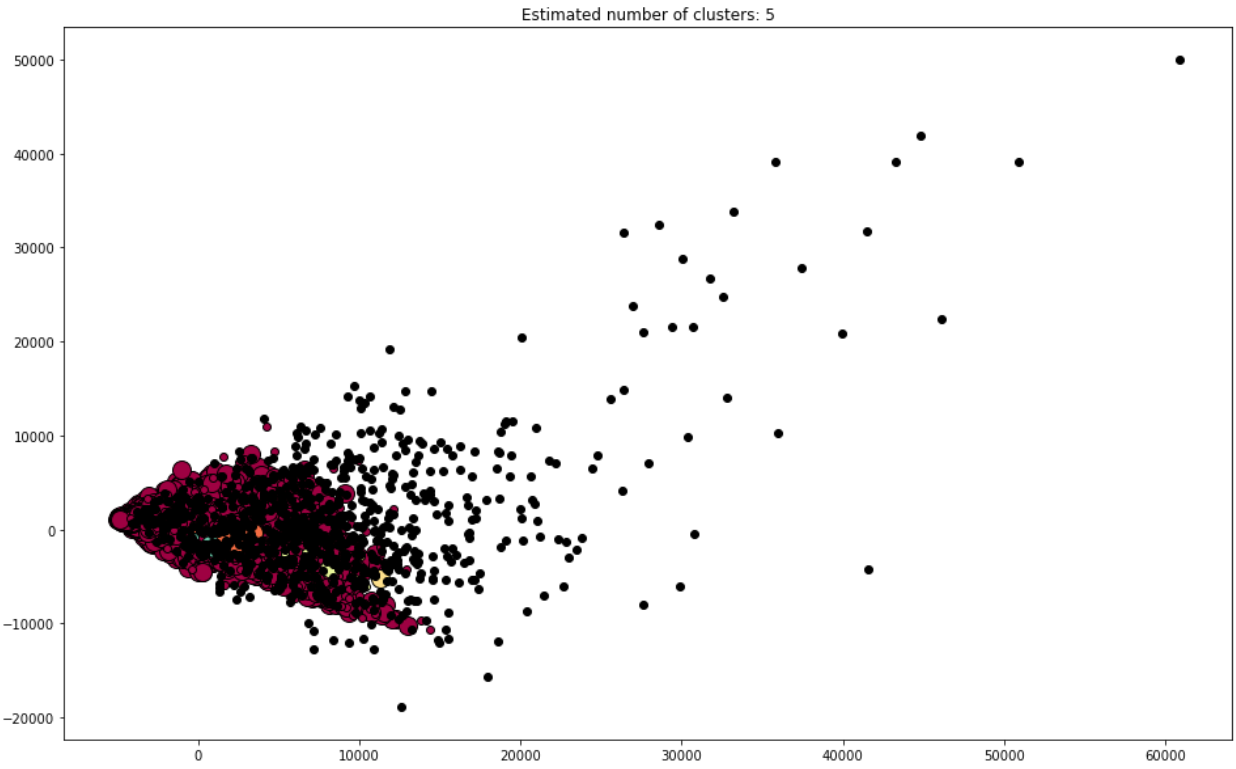

Рисунок 3. Результаты при найденных параметрах

Из данного графика видно, что большинство точек были определены к одному кластеру.

OPTICS

Перед применением алгоритма OPTICS были изучены его параметры и атрибуты, их описание представлено в таблице 2.

Параметр	Описание
Min_samples	Количество точек вокруг, чтобы рассматривать данную как основную
Max_eps	Максимальное расстояние между двумя точками
Metric	Метрика, используемая для вычисления расстояния между элементами

P	Параметр для метрики Минковского
Metric_params	Дополнительные параметры для метрик
Cluster_method	Метод извлечения используемый для извлечения кластеров
Eps	Максимальная дистанция для двух точек для метода dbscan
xi	Определяет минимальную крутизну на графике достижимости, который составляет границу кластера
Predecessor_correction	Коррекция кластеров в соответствии с предшественниками
Min_cluster_size	Минимальное количество элементов в кластере
Algorithm	Алгоритм для расчета ближайших соседей
Leaf_size	Размер листа
Атрибут	Описание
Labels_	Метки кластеров для каждой точки
Reachability_	Расстояния достижимости для элементов
Ordering_	Упорядоченный список элементов для кластера
Core_distances_	Расстояние, на котором каждый элемент становится основной точкой

Predecessor_	Точка, из которой была поулчена выборка
Cluster_hierarchy_	Список кластеров

Далее были определены параметры метода OPTICS (max_eps, min_samples) при которых результаты получаются примерно равными результатам алгоритма DBSCAN

```
Min samples = 3::eps=2.0
```

При данных параметрах получается 6 кластеров и 6% некластеризованных данных.

Полученные результаты изображены на рисунке 4 вместе с графиком достижимости.

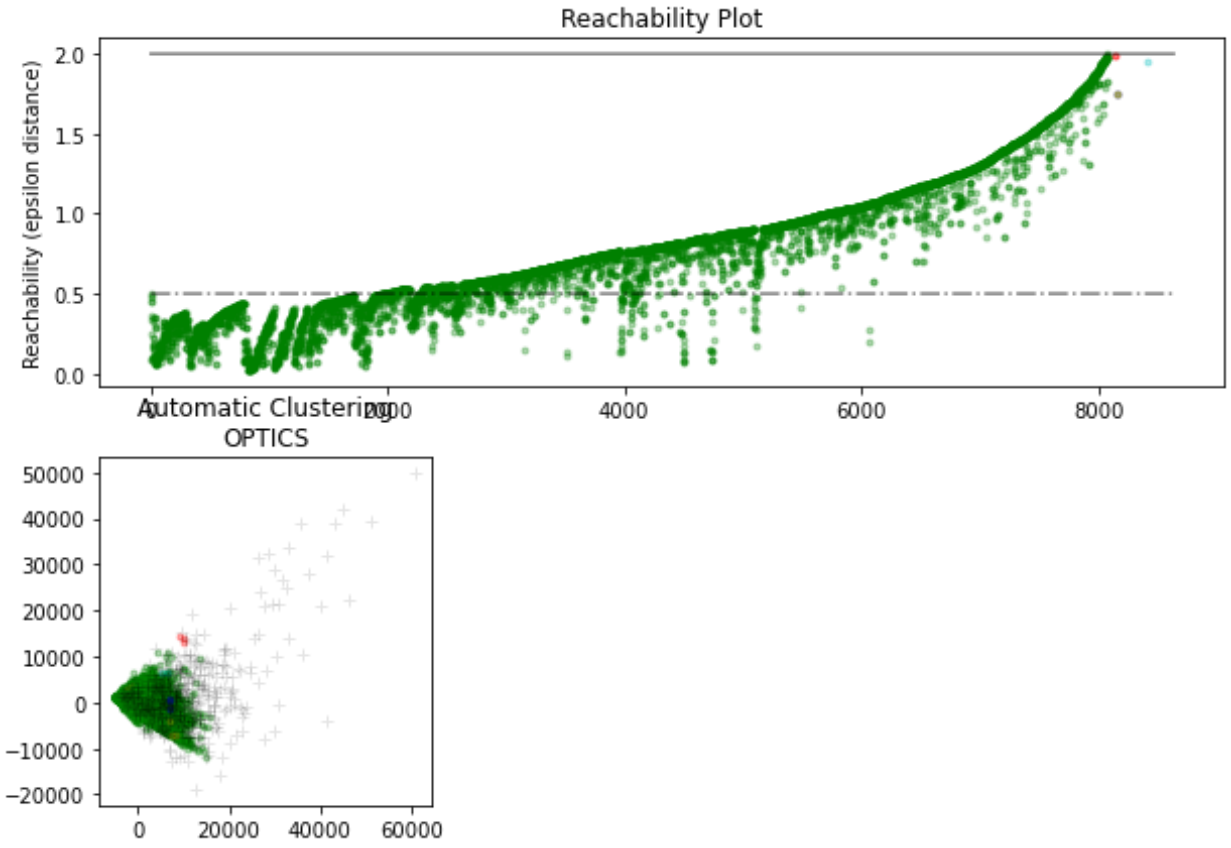

Рисунок 4. Результаты алгоритма OPTICS

Основные точки в алгоритмах DBSCAN и OPTICS определяются одинаково, однако во втором случае для точек сохраняются расстояния достижимости, на основе которых точки формируются в кластере.

Также был запущен алгоритм OPTICS с использованием различных метрик, результаты представлены на рисунках 5 и 6.

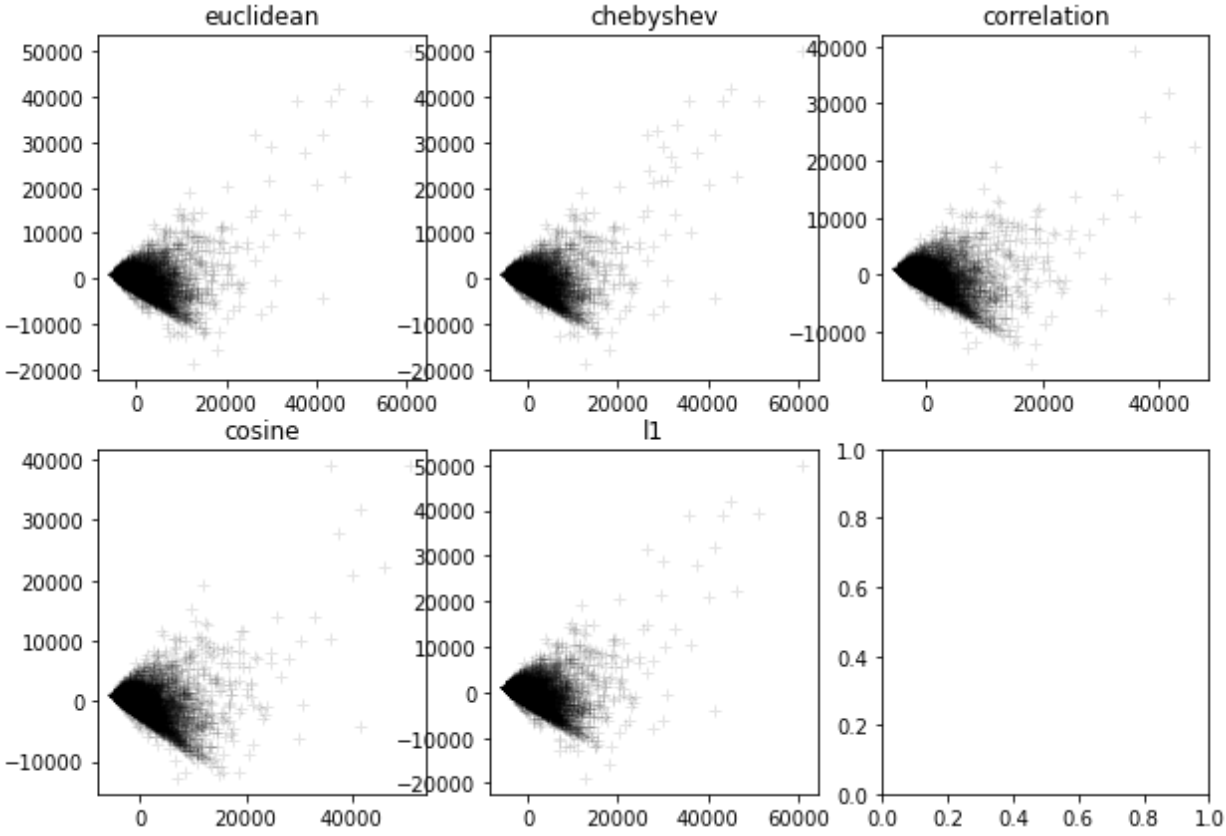

Рисунок 5. Алгоритм OPTICS с различными методами

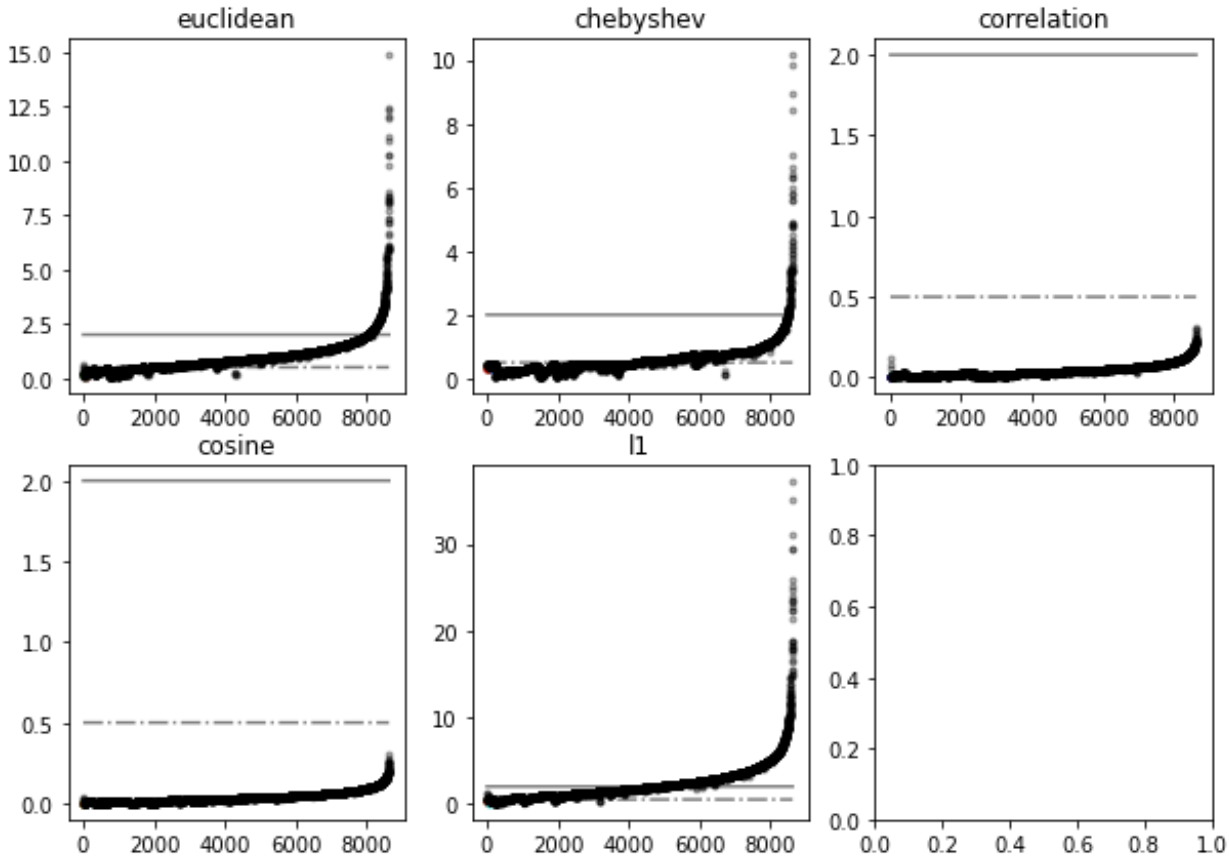

Рисунок 6 Алгоритм OPTICS с разными методами графики достижимости

Как видно из данных графиков, основные различия видны лишь на графиках достижимости, т.к. в зависимости от выбранного метода по-разному считаются расстояния между точками.

ВЫВОД

В данной лабораторной работе были рассмотрены методы кластеризации DBSCAN и OPTICS реализованные в пакете Sklearn. На наборе данных были запущены данные алгоритмы с различными параметрами.