

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

ОТЧЕТ
по лабораторной работе №6
по дисциплине «Машинное обучение»

Студенты гр. 6304

Преподаватель

Тимофеев А.А.

Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами кластеризации модуля Sklearn

Ход работы

Загрузка данных

1. Был создан датафрейм Pandas на основе загруженного датасета (<https://www.kaggle.com/arjunbhasin2013/ccdata>)

DBSCAN

1. Так как признаки в выборке соответствуют разным шкалам, была произведена стандартизация данных.
2. Была произведена кластеризация методом DBSCAN, выведены получившиеся метки кластеров, их количество, а также процент наблюдений, которые не удалось кластеризовать. Результаты представлены на рисунке 1.

```
Labels:
{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}
Num of clusters:
36
Non classified, %:
0.7512737378415933
```

Рисунок 1 – Результаты кластеризации

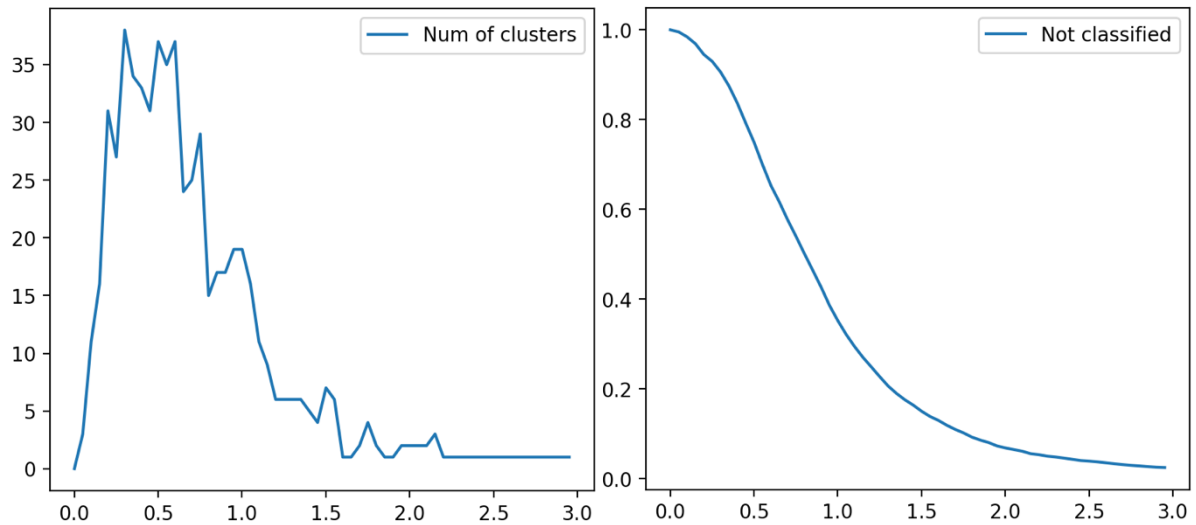
3. Описание параметров, которые DBSCAN принимает на вход, представлено в таблице 1.

Таблица 1 – Параметры DBSCAN

Название	Описание	Принимаемые значения	Значение по умолчанию
<i>eps</i>	Максимальное расстояние между двумя элементами, допускающее их соседство.	float	0.5
<i>min_samples</i>	Число элементов (или общий вес) в окрестности	int	5

	точки, чтобы рассматривать ее как основную. Сюда входит и сама точка.		
<i>metric</i>	Метрика, используемая при вычислении расстояния между элементами.	Строка или функция, принимаемая <code>sklearn.metrics.pairwise_distances</code>	<i>euclidean</i>
<i>metric_params</i>	Дополнительные ключевые аргументы для метрической функции.	dict	<i>None</i>
<i>algorithm</i>	Алгоритм, который будет использоваться модулем <code>NearestNeighbors</code> для вычисления точечных расстояний и поиска ближайших соседей.	auto, ball_tree, kd_tree, brute	<i>auto</i>
<i>leaf_size</i>	Размер листа, передаваемый в <code>BallTree</code> или <code>cKDTree</code> .	int	<i>30</i>
<i>p</i>	Степень метрики Минковского, которая будет использоваться для вычисления расстояния между точками.	float	<i>None</i>
<i>n_jobs</i>	Количество параллельных задач для запуска.	int	<i>None</i>

4. Был построен график зависимости количества кластеров и количества не кластеризованных данных от параметра *eps*. График представлен на



рисунке 2.

Рисунок 2 – Зависимость количества кластеров и количества не кластеризованных данных от параметра *eps*

Из графика видно следующее:

- При очень маленьком *eps* кластеры не создаются, и следовательно процент не кластеризованных данных близок 1
- С ростом *eps* количество кластеров увеличивается и достигает своих максимальных значений в интервале *eps* [0.4, 0.7]
- С продолжением роста *eps* количество кластеров уменьшается, так как в них попадает все больше и больше точек, а количество не кластеризованных данных стремится к 0.

5. Был построен график зависимости количества кластеров и количества не кластеризованных данных от параметра *min_samples*. График представлен на рисунке 3.

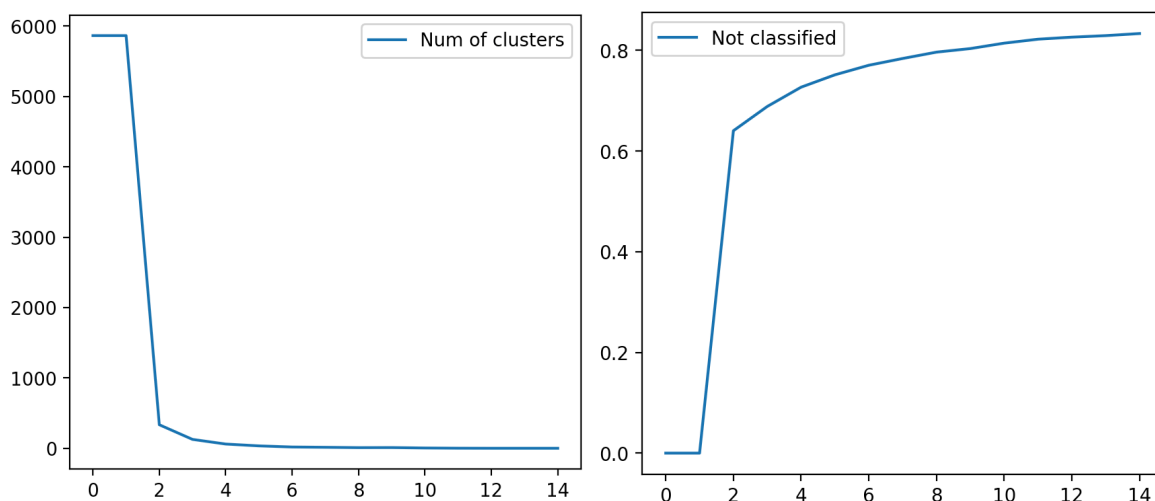


Рисунок 3 – Зависимость количества кластеров и количества некластеризованных данных от параметра *min_samples*

Из графика видно следующее:

- Количество кластеров изначально очень высоко, так как по сути каждая точка представляет собой кластер, при это количество не кластеризованных данных равняется 0
 - С ростом минимального числа точек в окрестности уменьшается количество кластеров, и соответственно повышается количество не кластеризованных данных. Это следует из того, что становится меньше основных точек, способных образовать вокруг себя кластер, и, следовательно становится больше выбросов.
6. Были определены значения параметров, при которых количество кластеров получается от 5 до 7, и процент не кластеризованных наблюдений не превышает 12%.
- $eps = 2, min_samples = 3$
7. Были визуализированы результаты кластеризации, полученные со значениями параметров из пункта 6, на данных с размерностью 2. Результат показан на рисунке 4.

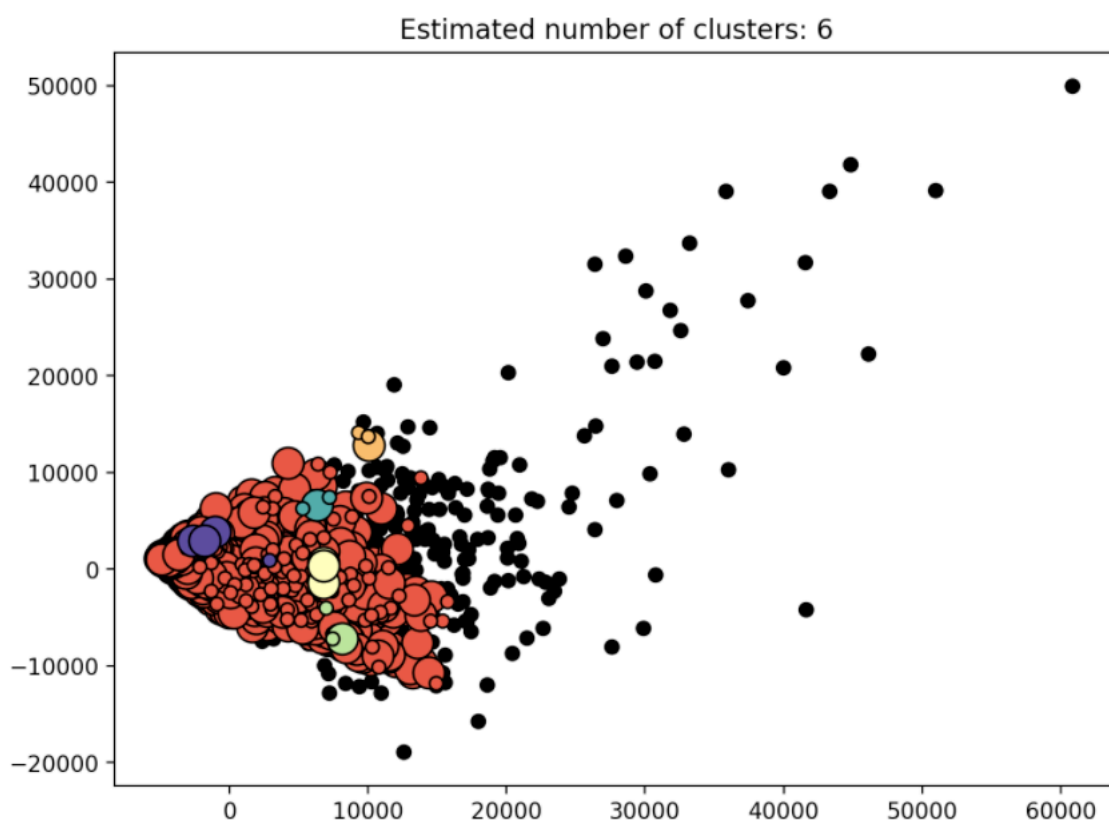


Рисунок 4 – Результаты кластеризации

Из графика следует, что большинство элементов выборки были определены к одному кластеру.

OPTICS

1. Параметры метода OPTICS представлены в таблице 2, атрибуты в таблице 3.

Таблица 2 – Параметры метода OPTICS

Название	Описание	Принимаемые значения	Значение по умолчанию
min_samples	Число элементов в окрестности точки, чтобы рассматривать ее как основную.	int > 0 0 < float < 1	5

<i>max_eps</i>	Максимальное расстояние между двумя элементами, допускающее их соседство.	float	np.inf
<i>metric</i>	Метрика, используемая при вычислении расстояния между элементами.	Строка или функция	<i>minkowski</i>
p	Параметр для метрики Минковского	int	2
<i>metric_params</i>	Дополнительные ключевые аргументы для метрической функции.	dict	<i>None</i>
cluster_method	Метод извлечения, используемый для извлечения кластеров с использованием вычисляемой достижимостью и упорядочения.	{ xi, dbscan }	xi
<i>eps</i>	Максимальное расстояние между двумя элементами, допускающее их соседство. По умолчанию значение соответствует <i>max_eps</i> , используется только при <i>cluster_method=dbscan</i>	float	None

<code>xi</code>	Определяет минимальную крутизну на графике достижимости, который составляет границу кластера. <i>Используется только при <code>cluster_method=xi</code>.</i>	$0 < \text{float} < 1$	0.05
<code>predecessor_correction</code>	Коррекция кластеров в соответствии с предшественниками. <i>Используется только при <code>cluster_method=xi</code>.</i>	bool	True
<code>min_cluster_size</code>	Минимальное количество элементов в кластере OPTICS	int > 0 $0 < \text{float} < 1$	None
<i>algorithm</i>	Алгоритм, который будет использоваться модулем NearestNeighbors.	auto, ball_tree, kd_tree, brute	<i>auto</i>
<i>leaf_size</i>	Размер листа, передаваемый в BallTree или cKDTree.	int	30
<i>n_jobs</i>	Количество параллельных задач для запуска.	int	None

Таблица 3 – Атрибуты метода OPTICS

Название	Описание	Принимаемые значения
labels_	Метки кластера для каждой точки	array (n_samples)
reachability_	Расстояния достижимости для элементов, индексированные по порядку элементов.	array (n_samples)
ordering_	Упорядоченный список элементов для кластеров.	array (n_samples)
core_distances_	Расстояние, на котором каждый элемент становится основной точкой, индексируется по порядку элементов.	array (n_samples)
predecessor_	Точки, откуда был достигнут элемент, проиндексированные по порядку элементов.	array (n_samples)
cluster_hierarchy_	Список кластеров в виде [начало, конец] в каждой строке, включая все индексы. Кластеры упорядочены в соответствии с (конец, -начало) (по	array (n_samples, 2)

	<p>возрастанию), так что более крупные кластеры, включающие более мелкие кластеры, идут после более мелких. <i>Используется только при $cluster_method=xi$.</i></p>	
--	---	--

2. Были определены параметры max_eps и $min_samples$, при которых результаты кластеризации приблизительно схожи с результатами DBSCAN из пункта 6: $max_eps = 2$, $min_samples = 3$, при этом число кластеров равнялось 6, а процент не кластеризованных данных составлял 6%.

Процесс определения основных точек в OPTICS идентичен соответствующему процессу в DBSCAN, однако в OPTICS для точек вычисляются и сохраняются расстояния достижимости, на основе которых точки выстраиваются в кластере, сохраняя при этом иерархическую структуру.

3. Полученные результаты были визуализированы. График представлен на рисунке 5.

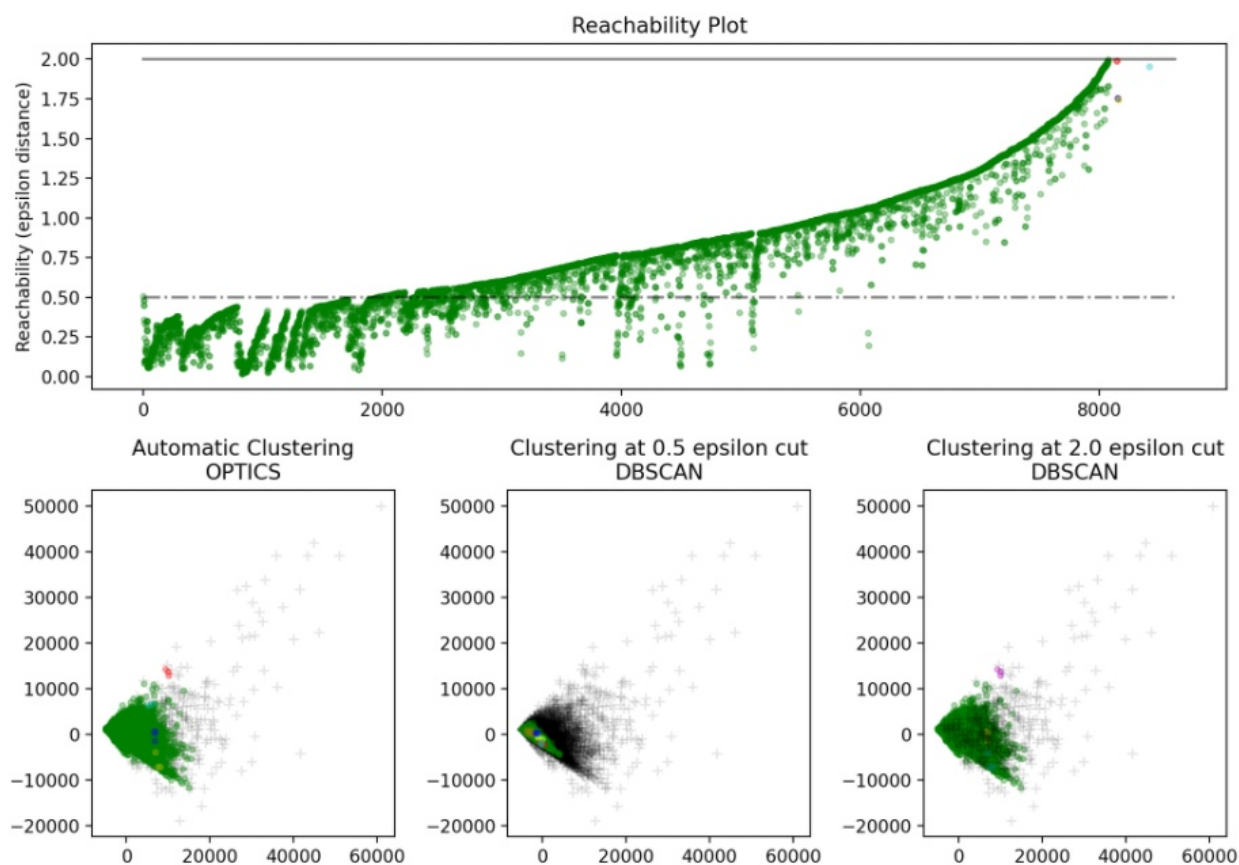


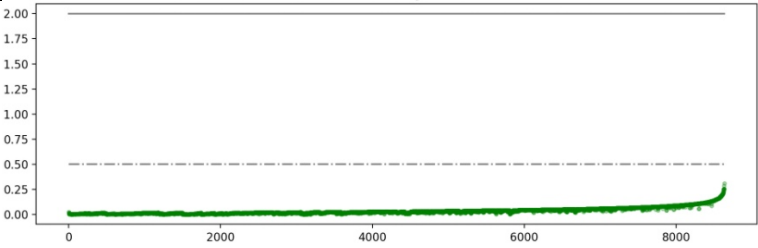
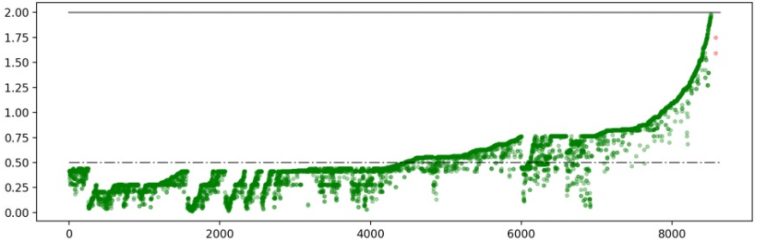
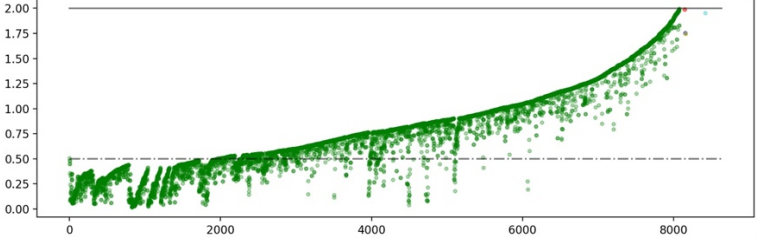
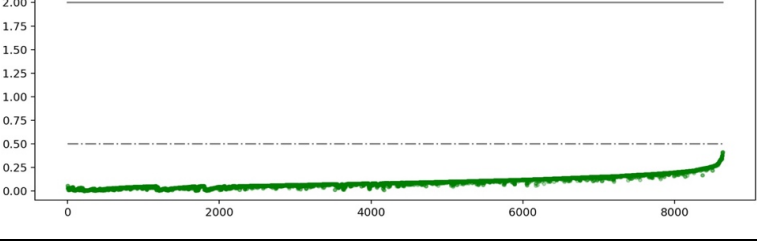
Рисунок 5 – Результаты кластеризации

Как и в случае с DBSCAN большинство данных принадлежат одному и тому же кластеру.

4. Было проведено исследование работы метода OPTICS с различными метриками. Результаты представлены в таблице 4.

Таблица 4 – Сравнение результатов OPTICS при различных метриках

Название	Кол-во кластеров	Выбросов, %	График достижимости
<i>cityblock</i>	6	6	

<i>cosine</i>	6	6	
<i>chebyshev</i>	6	6	
<i>l2</i>	6	6	
<i>braycurtis</i>	6	6	

Различия между результатами при различных метриках видны лишь на графике достижимости в виду того, что расстояние между точками измеряется разными способами.

Выводы

В ходе выполнения данной лабораторной работы было произведено знакомство с методами кластеризации модуля Sklearn. Кластеризация производилась с помощью методов DBSCAN и OPTICS.