оМинистерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение

высшего образования

«Санкт-Петербургский политехнический университет Петра Великого»

(ФГАОУ ВО СПбПУ)

Институт промышленного менеджмента, экономики и торговли

Высшая инженерно-экономическая школа

**ЛАБОРАТОРНАЯ РАБОТА №4**

По дисциплине «Многомерный статистический анализ»

Кластеризация методом k-means

(семестр 2)

|  |  |  |  |
| --- | --- | --- | --- |
| Студент группы |  |  |  |
| 3740105/20101 |  | подпись, дата | К.С. Малышева. |

Оценка выполненной студентом работы:

|  |  |  |
| --- | --- | --- |
| Преподаватель,  Доцент, канд.физ-мат.наук |  |  |
|  | подпись, дата | Л.В. Павлова |

Санкт-Петербург – 2023

**Задача работы:**

Кластеризовать заранее подготовленные модельные данные с помощью метода k-means и kernel k-means.

**Ход работы:**

1. Подготовить модельные данные.

2. Реализовать k-means алгоритм.

3. Привести графическую иллюстрацию результатов нескольких кластеризации.

4. Реализовать ядерный алгоритм кластеризации (kernel k-means).

5. Сделать выводы.

**Постановка задачи**

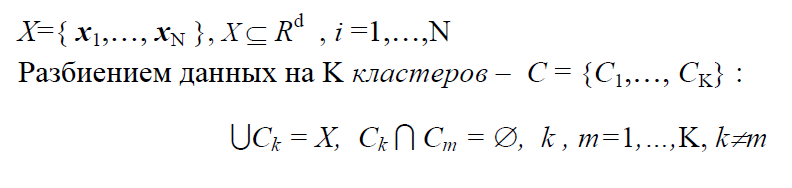
Автоматическая классификация объектов, или кластеризация, – разбиение совокупности объектов, статистически представленных матрицей «объект-свойство», на сравнительно небольшое число однородных групп, или кластеров. Кластерный анализ – это общее название множества вычислительных процедур, используемых при создании классификации. В результате работы с процедурами образуются кластеры, или группы очень похожих объектов.

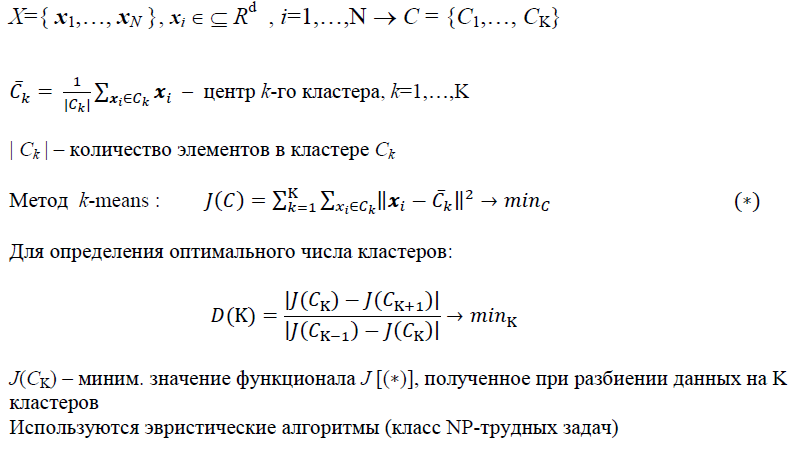
Для анализа множества объектов достаточно:

– сохранить описание типичного представителя каждого кластера;

– перечислить номера объектов, входящих в данный кластер;

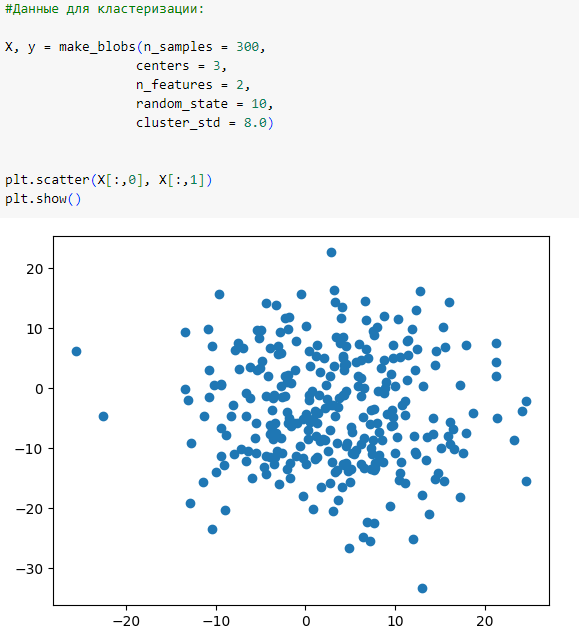
–указать максимальное отклонение каждого свойства от его среднего значения для данного кластера.





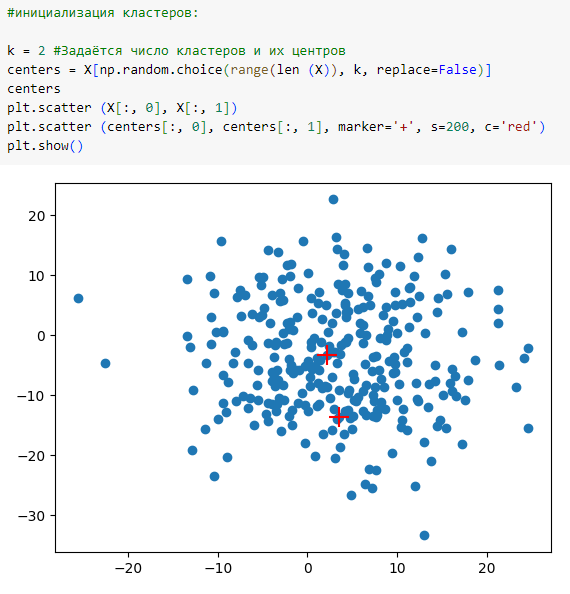
**1. Подготовить модельные данные.**

Генерируем данные для кластеризации.



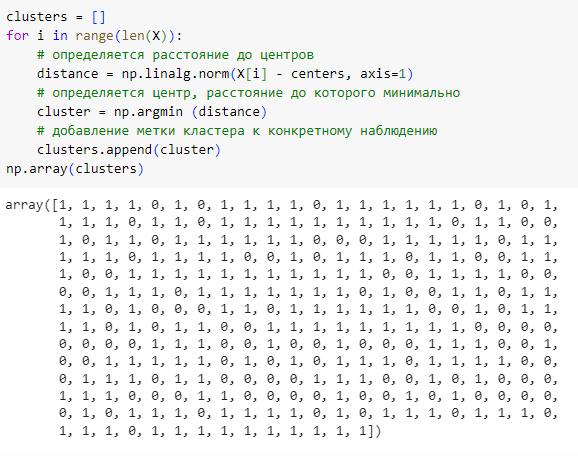
**2. Реализовать k-means алгоритм.**

Задаём начальное число кластеров и инициализируем случайным образом начальные центры кластеров.

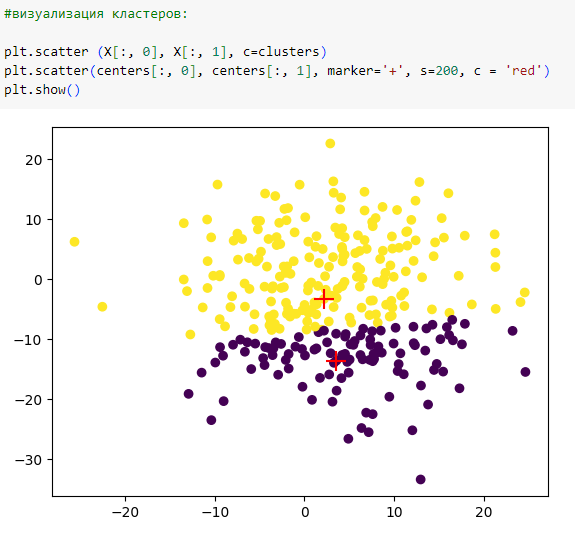


Определяем принадлежность наблюдений к конкретному кластеру на основе минимального расстояния.

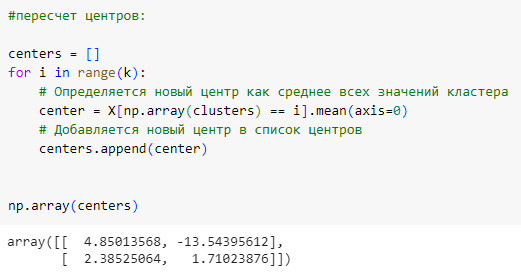


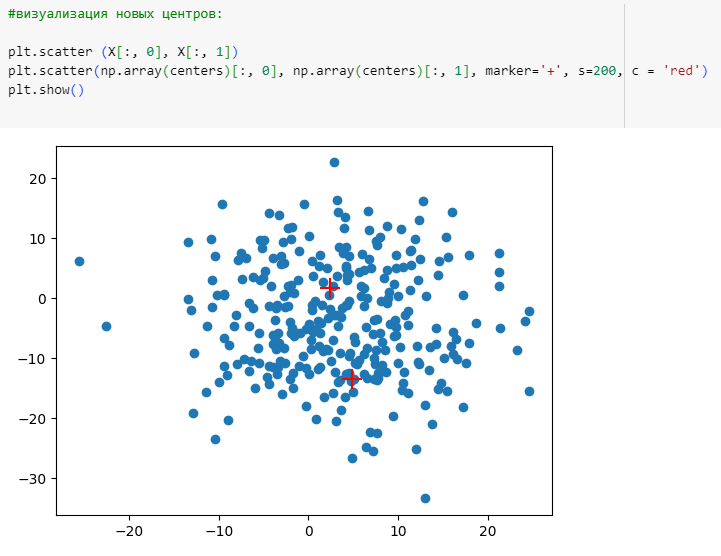


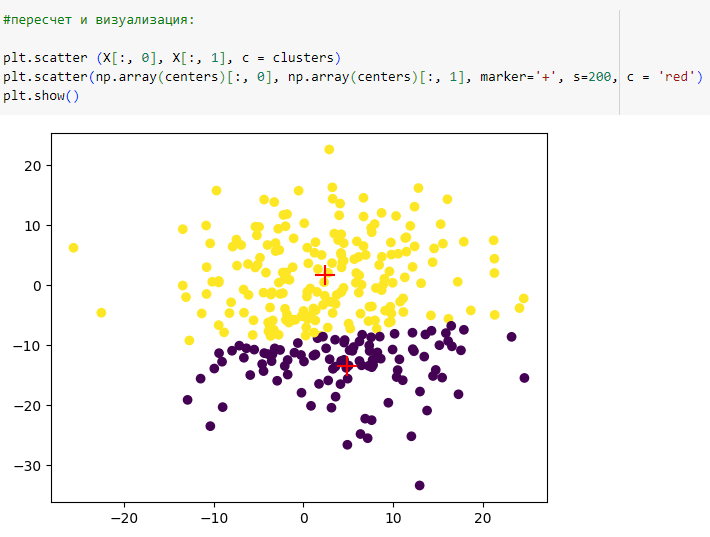
Визуализируем полученный результат кластеризации.



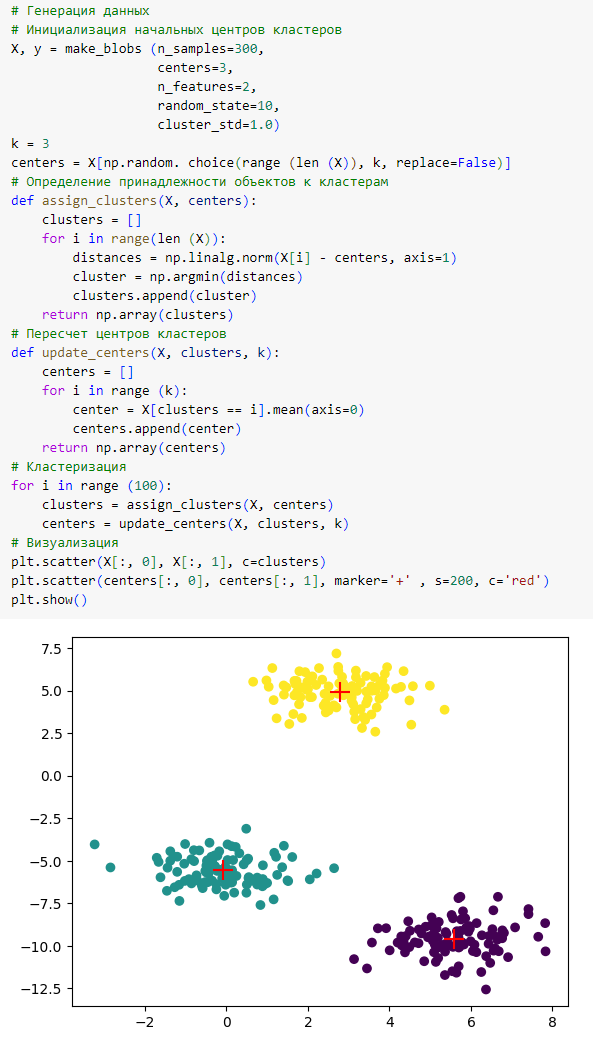
Пересчитываем новые центры кластеров и визуализируем результат.







Повторим кластеризацию для трех кластеров для хорошо разделенных данных.

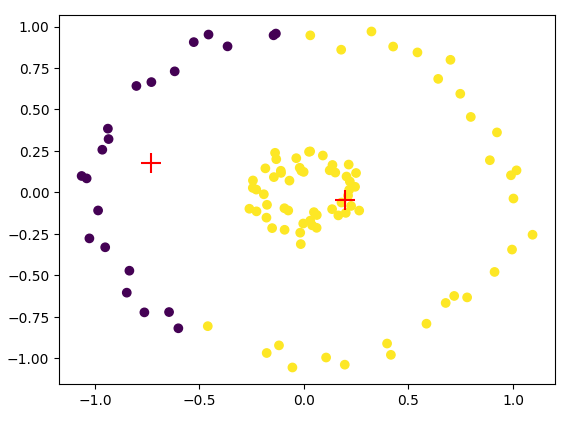


Далее попробуем реализовать алгоритм k-means для нелинейно разделенных данных.

Сгенерируем следующие данные.



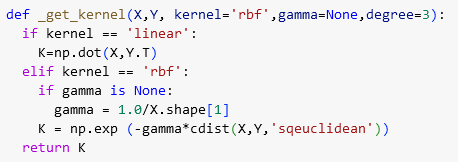
И повторим разработанный алгоритм для них. Результат кластеризации приведем на графике.

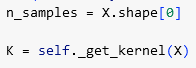


Таким образом, мы видим, что обычный алгоритм k-means не справляется с кластеризацией такого типа данных.

**4. Реализовать ядерный алгоритм кластеризации (kernel k-means).**

Создадим ядерные функции: линейное и гаусcово ядро (rbf). И вычислим ядерную функцию для наших данных (в данном примере используется гауссово ядро с параметрами gamma=5 и degree=3).

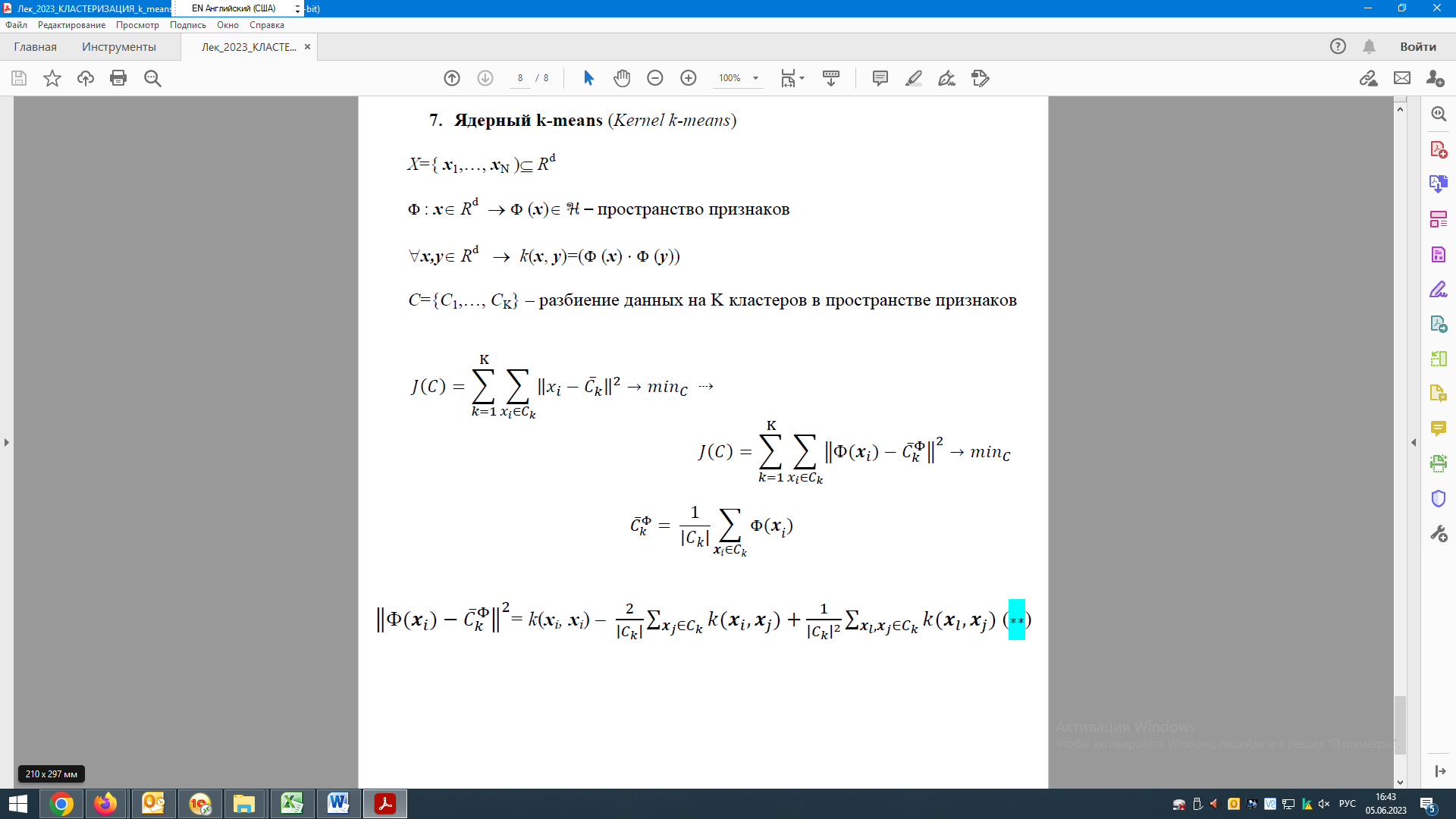




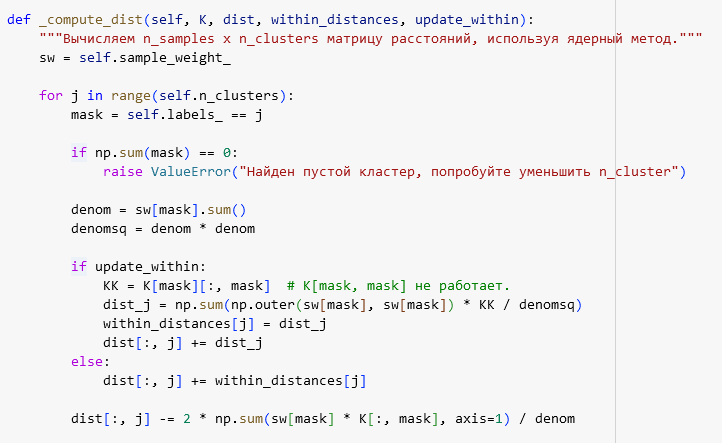
Далее инициализируем центроиды случайными точками из данных.



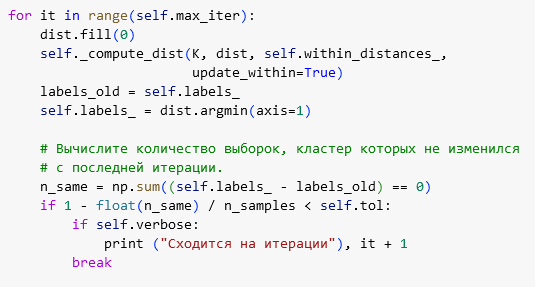
Создаем и вычисляем n\_samples x n\_clusters матрицу расстояний, используя ядерный метод.



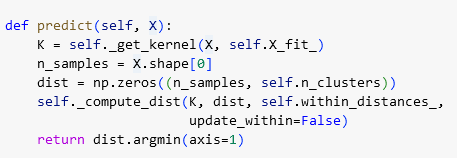




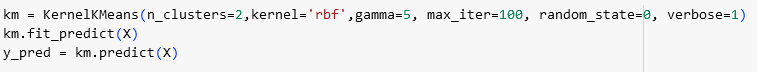
Повторяем данный процесс до достижения сходимости (в данном случае максимальное количество итераций = 100).

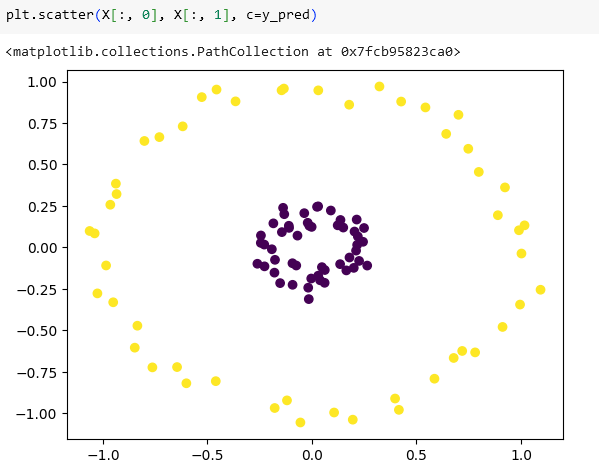


Далее обновляем центроиды и пересчитываем значения еще раз.



Для наших данных получается следующий результат.





Таким образом, данные разделились как было необходимо.

Поэтому можно сделать вывод о том, что простой алгоритм k-means целесообразно использовать в случае линейной разделимости данных. Для случая нелинейных данных наиболее подходящим методом является kernel k-means.