Санкт-Петербургский политехнический университет Петра Великого  
Институт компьютерных наук и технологий  
Высшая школа программной инженерии

Лабораторная работа №5

По дисциплине «Машинное обучение»

Выполнил студент гр 33534/5  Донцов А. Д.

Руководитель И. А. Селин

Санкт-Петербург  
2019 г.

# Постановка задачи

1) Разбейте множество объектов из набора данных pluton.csv на 3 кластера методом центров

тяжести (kmeans). Сравните качество разбиения в зависимости от максимального числа

итераций алгоритма (визуально и по подходящим метрикам из sklearn.metrics).

2) Сгенерируйте набор данных в двумерном пространстве, состоящий из 3 кластеров,

каждый из которых сильно “вытянут” вдоль одной из осей. Исследуйте качество кластеризации

методом k-медоидов или clara

(например,  https://github.com/salspaugh/machine\_learning/blob/master/clustering/kmedoids.py  ) в

зависимости от 1) использования стандартизации (sklearn.preprocessing.StandardScaler); 2) типа

метрики (достаточно евклидовой и манхэттенской). Объясните полученные результаты.

Альтернативные реализации методов k-медоидов и clara допустимы.

3) Постройте дендрограмму для набора данных votes.csv (число голосов, поданных за

республиканцев на выборах с 1856 по 1976 год). Строки представляют 50 штатов, а столбцы -

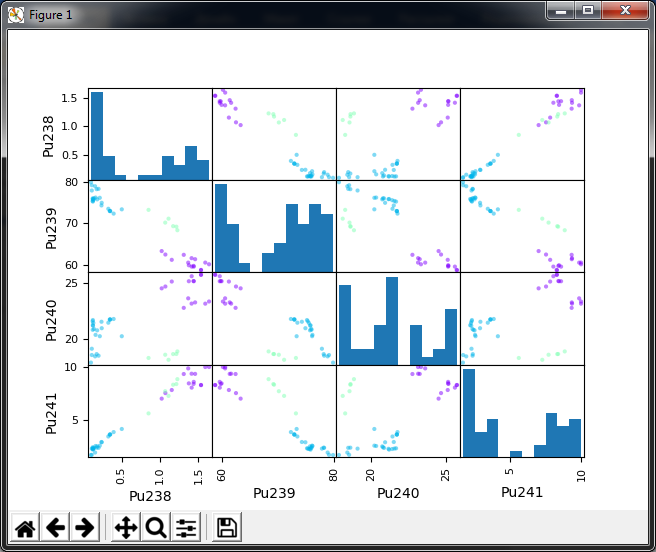
годы выборов (31). Проинтерпретируйте полученный результат.

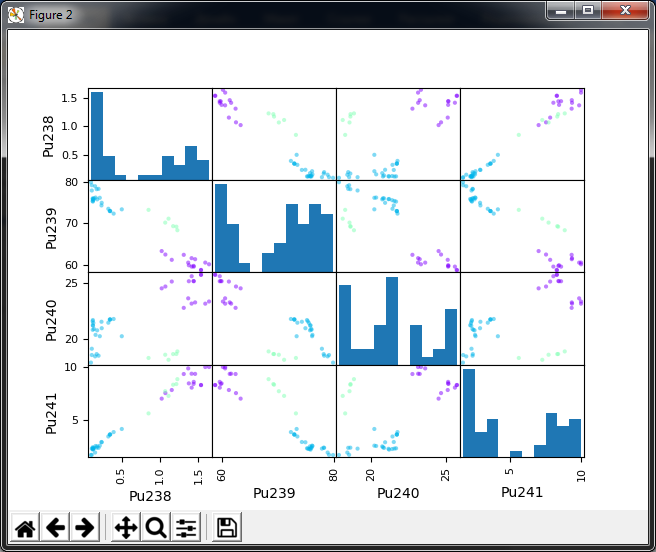
# Ход работы

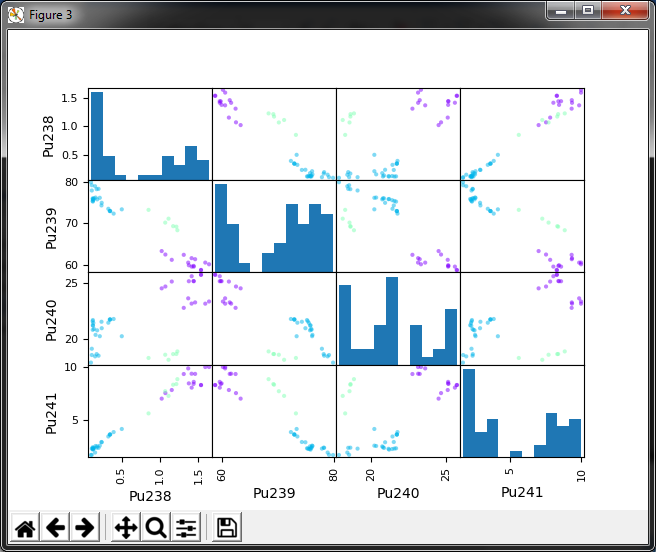
1. Для теста классификатора были выбраны значения 10, 20, 30 для параметра max\_iter. Качество разделения оценивалась с помощью метрик calinski\_harabaz\_score, davies\_bouldin\_score.
2. Для тестов использовался датасет с параметрами:   
   centers = [[5, 4], [10, 4.5], [2, 6]]  
   X, y = make\_blobs(n\_samples=300, n\_features=2, centers=centers, cluster\_std=0.2, random\_state=5)  
   В качестве метрик были выбраны homogeneity\_score, completeness\_score.
3. Для датасета votes.csv была построена дендрограмма.

# Результат работы программы

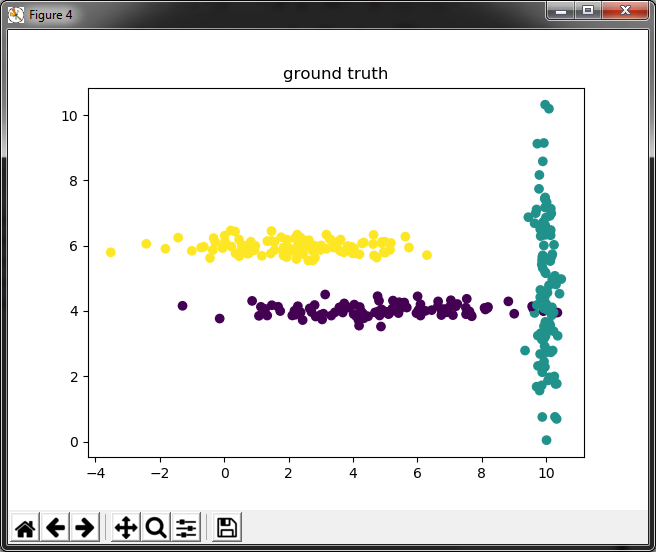
## Задание 1

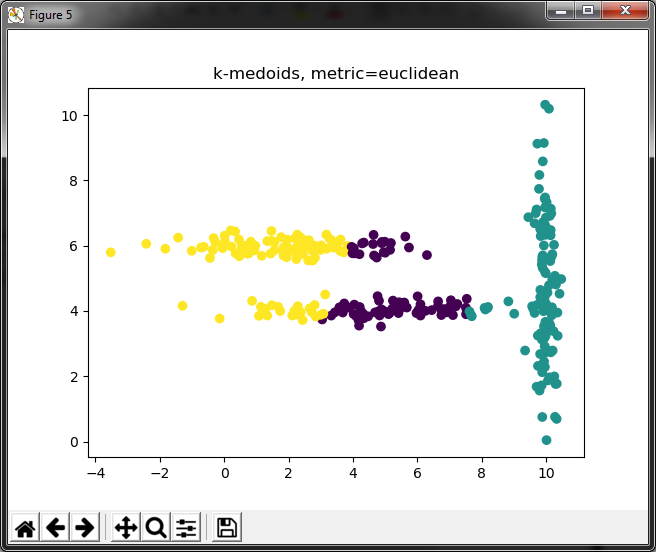


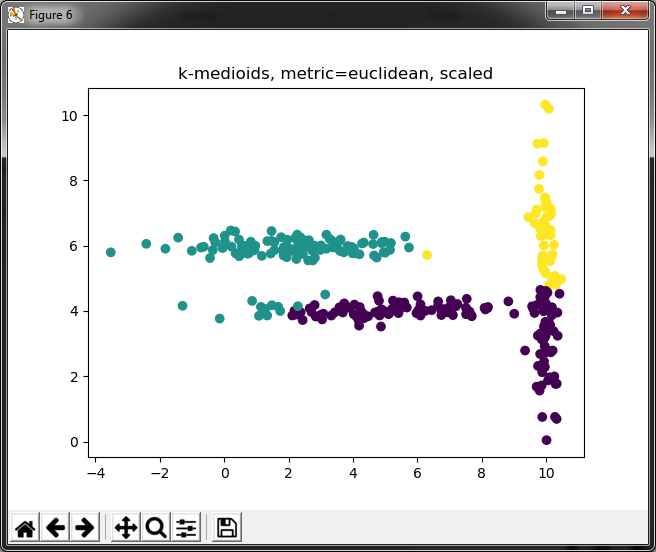


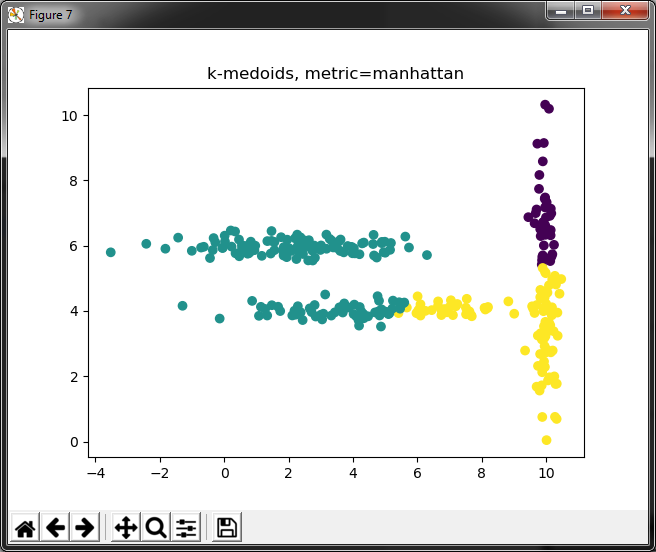


## Задание 2

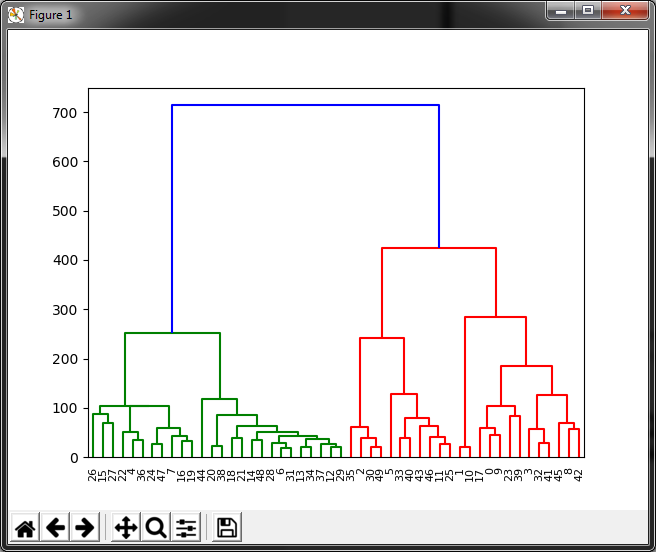








## Задание 3



# Вывод

1. Для метрики calinski-harabaz был получен результат 243.31015661304917, для метрики davies-bouldin 0.4183126928050051 на каждой итерации алгоритма. Данные метрики по-разному оценивают качество кластеризации. Можно сделать вывод о высоком качестве, т.к. значение метрики davies-bouldin близко к 0, а расстояние согласно метрике calinski-harabaz невелико.
2. Тесты на заданных датасетах дали следующие результаты:

|  |  |  |
| --- | --- | --- |
| Metric | homogeneity-score | completeness-score |
| euclidean | 0.5682437663870256 | 0.5717416095279674 |
| euclidean, scaled | 0.5496790897337038 | 0.6098018746625112 |
| manhattan | 0.6063029853671632 | 0.6107646573369031 |
| manhattan, scaled | 0.6320284206430118 | 0.6702740938217145 |

Данные метрики оценивают однородность и полноту кластеризации. Таким образом, можно сказать о приемлемом качестве разбиения несмотря на вытянутость вдоль оси.

1. Построенная дендрограмма позволяет сделать вывод о распределении голосов по группам штатов. Зеленая группа – штаты, где побеждают демократы, красные – республиканцы. Расстояние между группами по вертикальной оси свидетельствует о разнице в голосах. Чем меньше расстояние – тем меньше разница.