Санкт–Петербургский государственный университет

Математическое обеспечение и адмиистрирование информационных систем

Гусев Егор Игоревич

Вычислительный практикум

Отчет по заданию №12

Преподователь:
Т.О. Евдокимова

Санкт-Петербург
2021 г.

Содержание

1. Ссылка на код

Код доступен по ссылке на github.

2. Постановка задачи

1. Реализовать метод k- средних для кластеризации точек на плоскости.
2. Показать, что результат зависит от выбора изначальных центров и от способа определения расстояния между точками.

3. Теория

Метод заключается в выборе начальных центров кластеров и последующем итерационном распределении точек по кластерам и переопределении координат центров. На каждой итерации:

- Определяем кластер, к которому относится точка

$$l_j = \underset{i=1,\dots,k}{\operatorname{argmin}} \rho(x_j, c_i),$$

  где l_j — метка кластера, c_i — центр кластера, $\rho(x_j, c_i)$ — функция расстояния. В наших тестах будем использовать две функции расстояния: евклидово расстояние и расстояние городских кварталов.

- Пересчитываем координаты нового центра каждого из кластеров, используя среднее арифметическое.

Это происходит до тех пор, пока результаты на двух идущих подряд итерациях не окажутся одинаковыми.

4. Численный эксперимент

4.1 Описание

- Определим параметры k-means следующим образом: k = 5, N = 200, координатная область [0,100] x [0,100].

- Выберем 2 способа определения расстояния между точками: Евклидова и Манхэттенская нормы.
- Выберем 2 способа задания начальных координат центров: случайный в пределах области и в точках (0,0),(100,0),(100,100), (50, 0), (100,100).
- Реализуем 4 различных варианта функции k-средних.

4.2 Результаты

На рисунках видно, что на данной выборке точек более существенные различия заметны между реализациями с разными метриками расстояний между точками на плоскости. Выбор же начальных координат не так сильно повлиял на распределение кластеров в данном случае. Следовательно, результат работы метода k-средних зависит и от начальных параметров, и от способа определения расстояния.

На рисунках жёлтым, синим и зелёным цветами обозначены разные кластеры, а красным – центры кластеров на последней итерации.

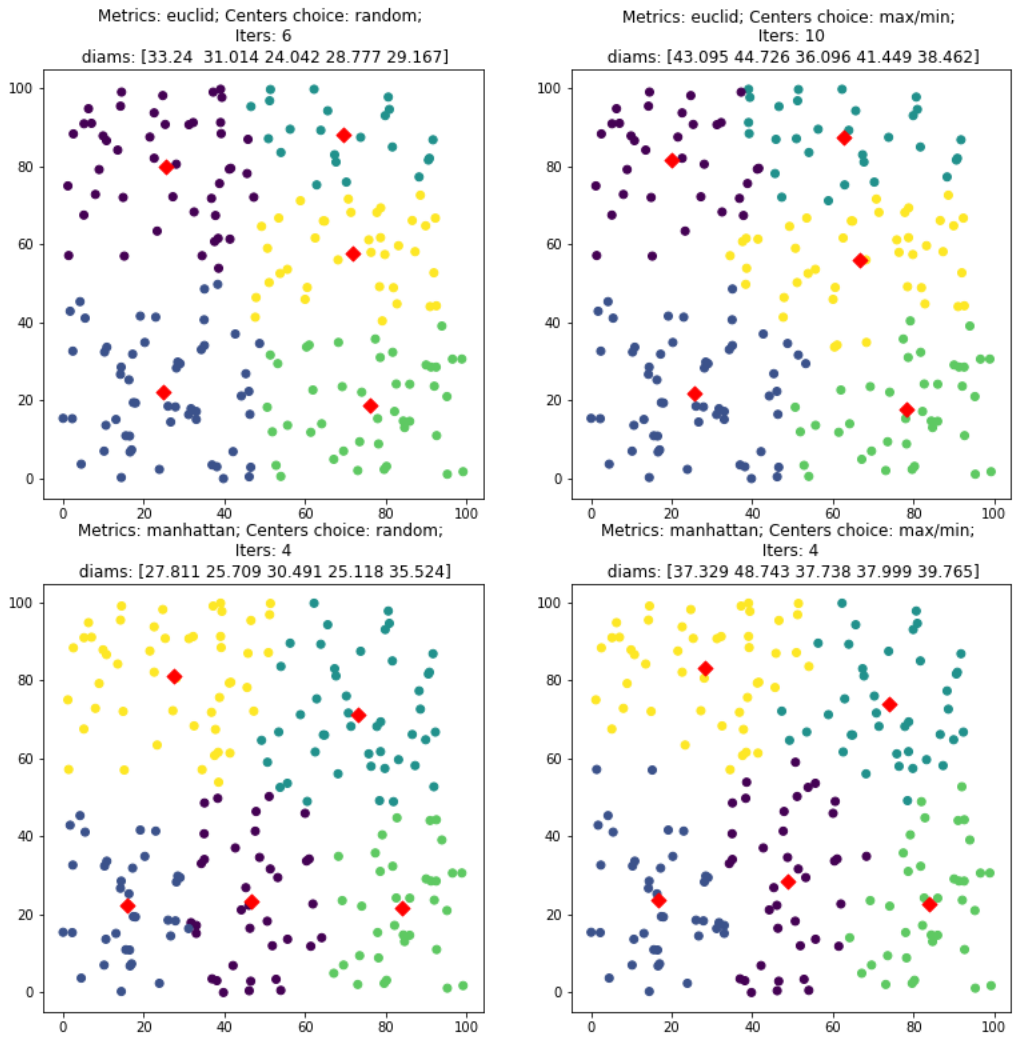

Рис. 1: Результат кластеризации N=200 равномерно распределенных точек на k=5 групп