Санкт-Петербургский государственный университет

Математико-механический факультет

Литвинов Степан Сергеевич

Задача кластеризации (k-means)

Практическая работа

Санкт-Петербург
2022

Оглавление

1. Постановка задачи **3**

2. Теорминимум **4**

3. Тесты **5**

4. Доп **6**

4.1. Теорминимум 6

4.2. Тесты 7

5. Код **9**

1. Постановка задачи

Построить кластеризации (для одинаковых N и k), используя два разных“расстояния” и разные начальные центры (рандомные и крайние (⊠ max/min покоординатам), используя метод k-means.

2. Теорминимум

Выбираем начальные центры кластеров. В наших тестах будем использовать два способа выбора начальных центров: случайный выбор и выбор центров, равных максимуму/минимуму по координатам.

На каждой итерации:

- Определяем кластер, к которому относится точка

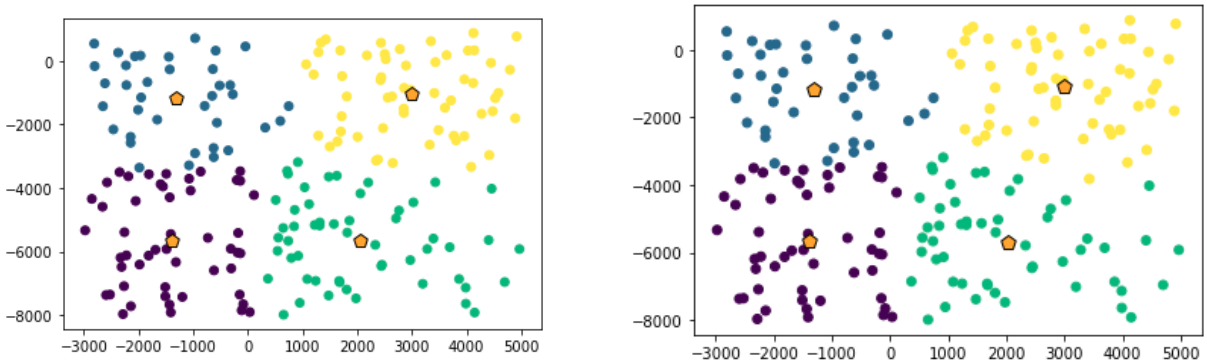
$$l_j = \arg\min_{i=1,\dots,k} \rho(x_j, c_i),$$

  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.

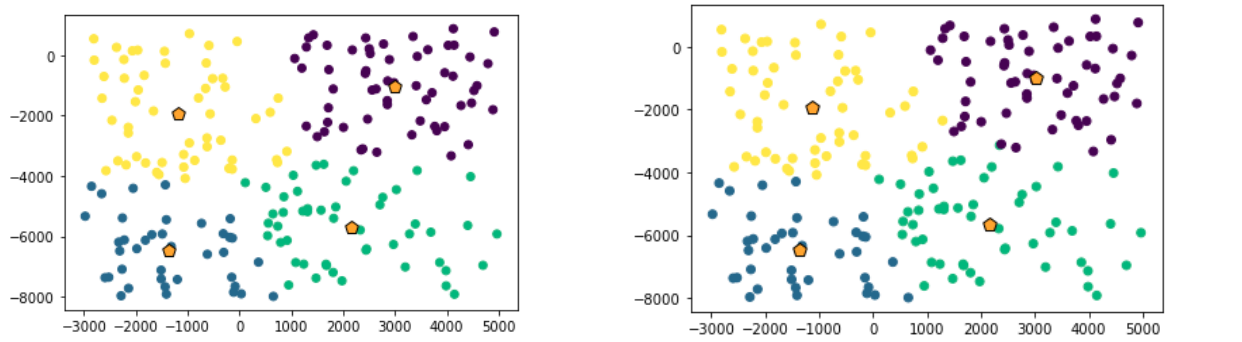
Продолжаем процесс до тех пор, пока составы кластеров не перестанут меняться.

3. Тесты



a) b)

Рис. 1: Результаты кластеризации при случайном выборе начальных центров. Функция расстояния: a) евклидово расстояние, b) расстояние городских кварталов.



a) b)

Рис. 2: Результаты кластеризации при выборе начальных центров, равных максимуму/минимуму по координатам. Функция расстояния: a) евклидово расстояние, b) расстояние городских кварталов.

4. Доп

4.1. Теорминимум

Когда значение k равно 1, сумма квадрата внутри кластера будет большой. По мере увеличения значения k сумма квадратов расстояний внутри кластера будет уменьшаться.

Наконец, мы построим график между значениями k и суммой квадрата внутри кластера, чтобы получить значение k. Мы внимательно рассмотрим график. В какой-то момент значение по оси x резко уменьшится. Эта точка будет считаться оптимальным значением k.

4.2. Тесты

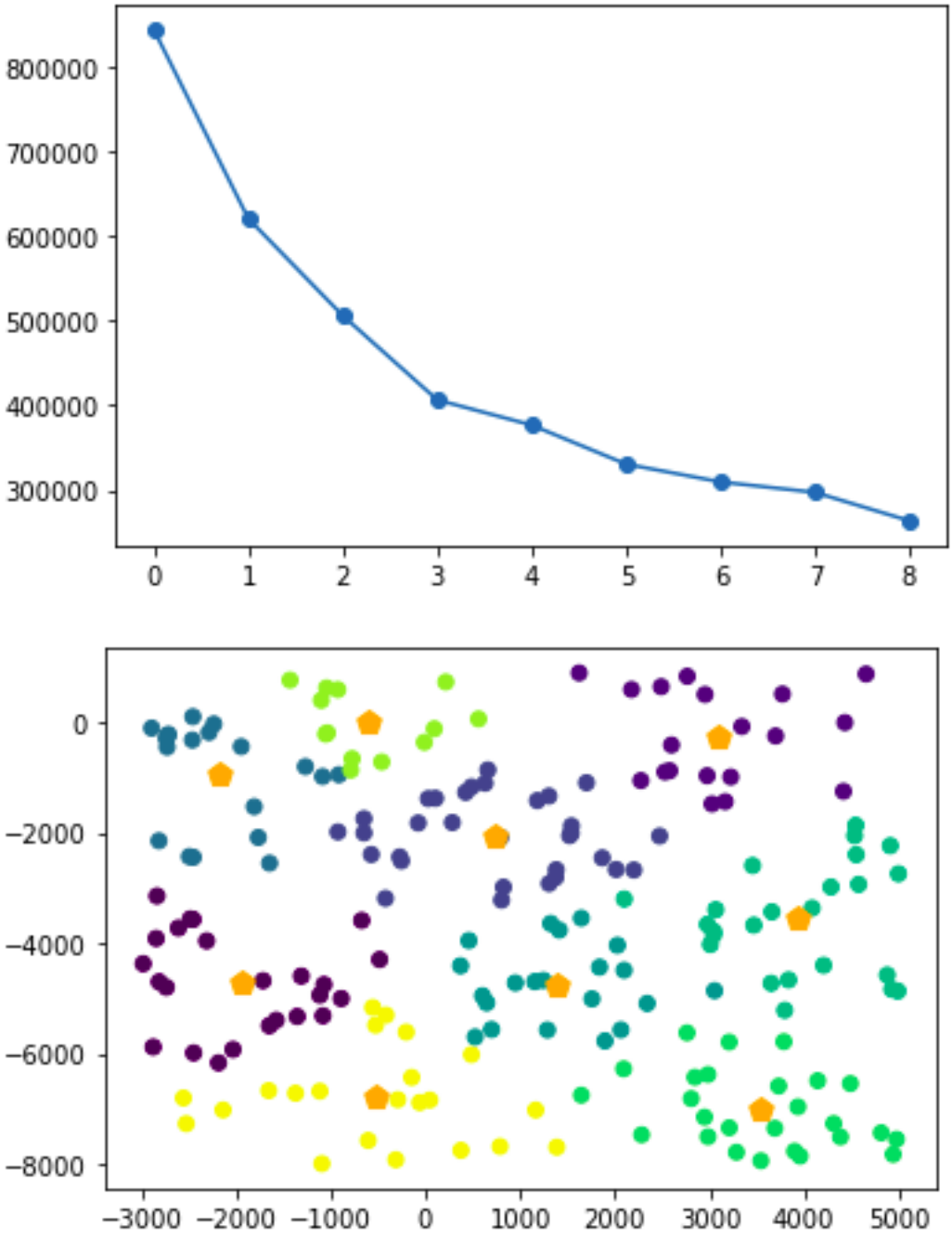


Рис. 3: Манхэттанское расстояние

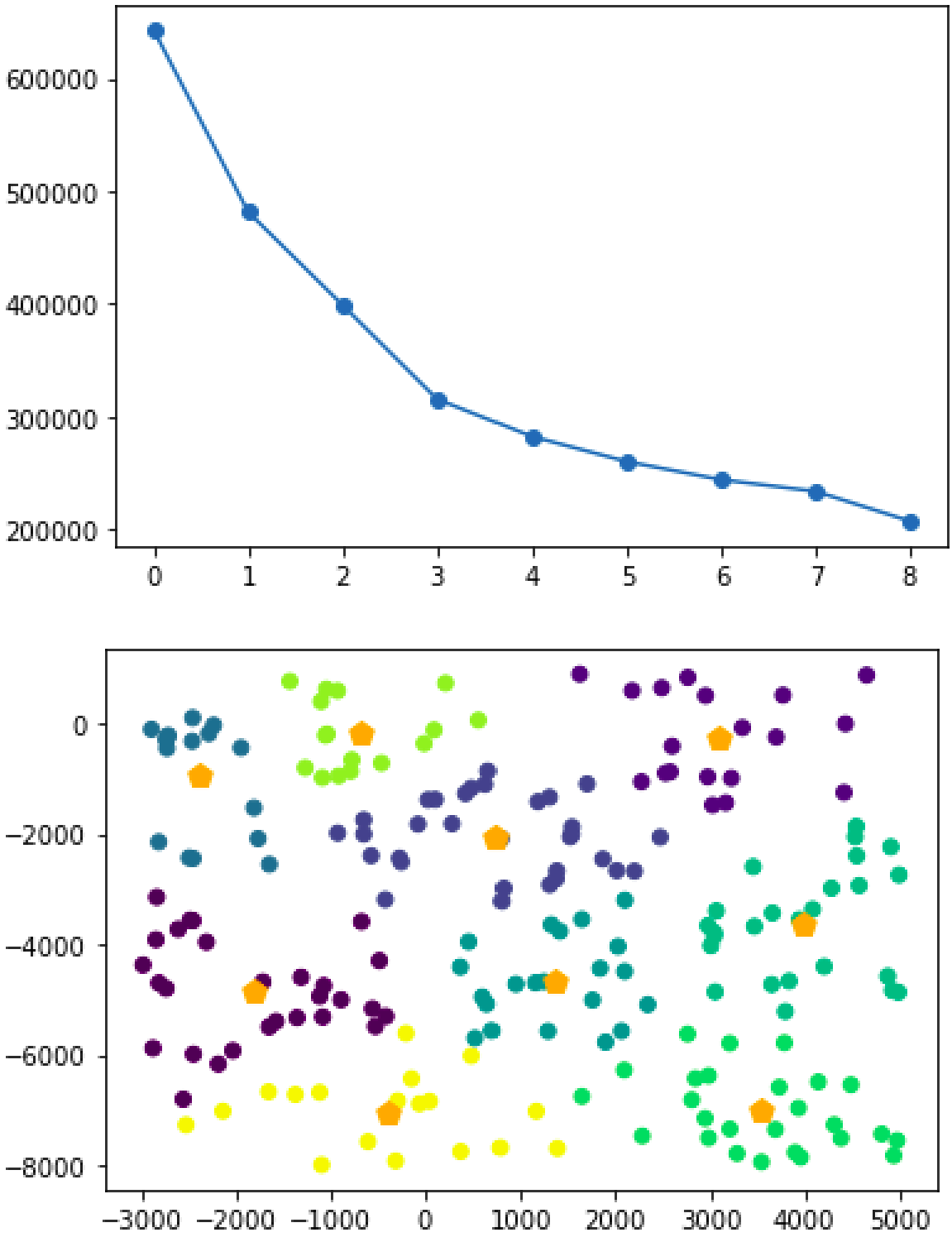


Рис. 4: Евклидово расстояние

5. Код

Можно посмотреть здесь