Лекция 7:
о задаче кластеризации

Евгений Борисов

четверг, 1 ноября 2018 г.

кластеризация

метрический подход - использование расстояний между объектами

метрика - функция расстояния

$$\rho: X \times X \to [0, \infty)$$

аксиома тождества : $\rho(x, y) = 0 \Leftrightarrow x = y$

симметрия: $\rho(x, y) = \rho(y, x)$

неравенство треугольника: $\rho(x, z) \leqslant \rho(x, y) + \rho(y, z)$

кластеризация

метрика - функция расстояния

Евклидова метрика: $\rho(x,y)=\sqrt{\sum_i (x_i-y_i)^2}$

метрика Минковского: $\rho(x,y)=\sqrt[n]{\sum_i w_i|x_i-y_i|^n}$

метрика Чебышева: $\rho(x,y)=\max_i|x_i-y_i|$

косинусная метрика: $\rho(x,y)=\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\cdot\sqrt{\sum_i y_i^2}}$

кластеризация

о задаче: обучение «без учителя» (unsupervised learning)

дано:

X - объекты

$\rho: X \times X \to [0, \infty)$ - функция расстояния (метрика)

найти:

Y - кластеры (метки)

$a: X \to Y$ - кластеризатор

- кластер состоит из близких объектов

- объекты разных кластеров существенно разные

кластеризация

о некорректности (размытости) задачи кластеризации

- недостаточно точная постановка задачи
- много разных критериев качества
- число кластеров обычно заранее не известно
- результат сильно зависит от метрики
- нормировка данных может существенно изменять результат

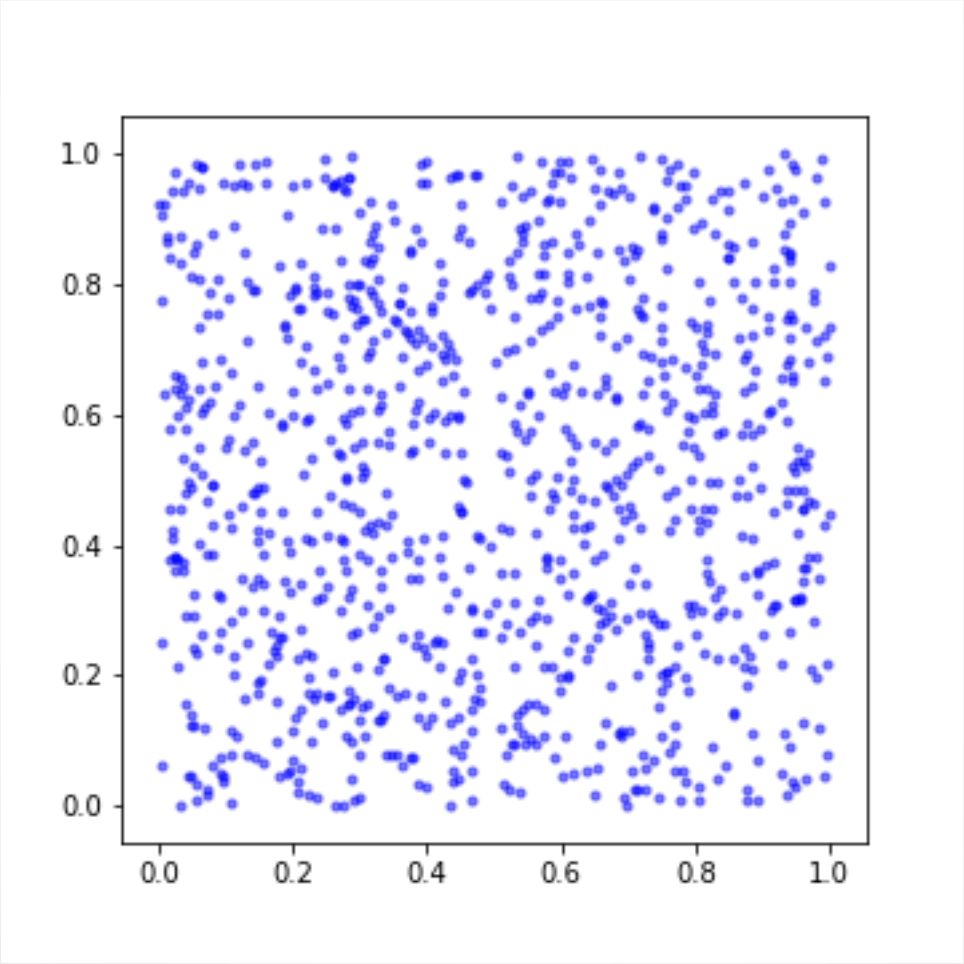
кластеризация

цели кластеризации

- предварительная обработка данных для упрощения основной задачи
- сжатие данных (оставляем один или несколько объектов от кластера)
- выделить нетипичные объекты
- построение иерархии объектов

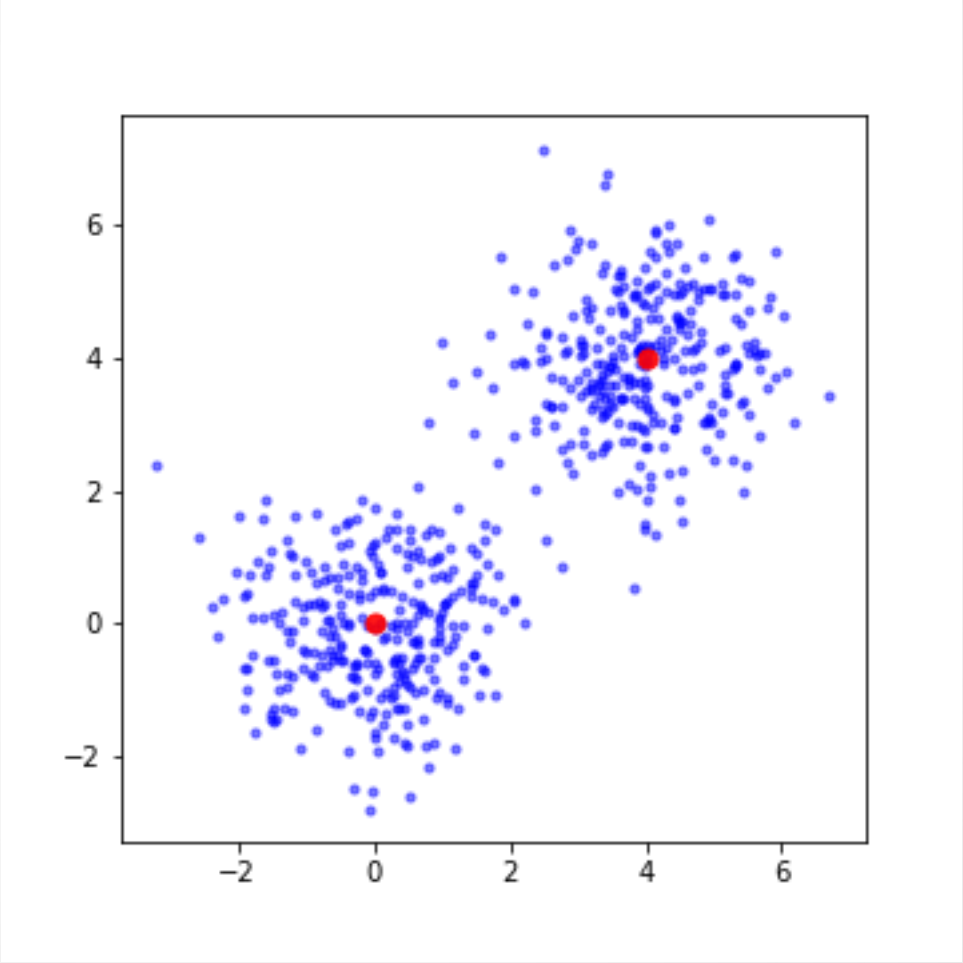
кластеризация

тип кластера: кластеры могут отсутствовать совсем



кластеризация

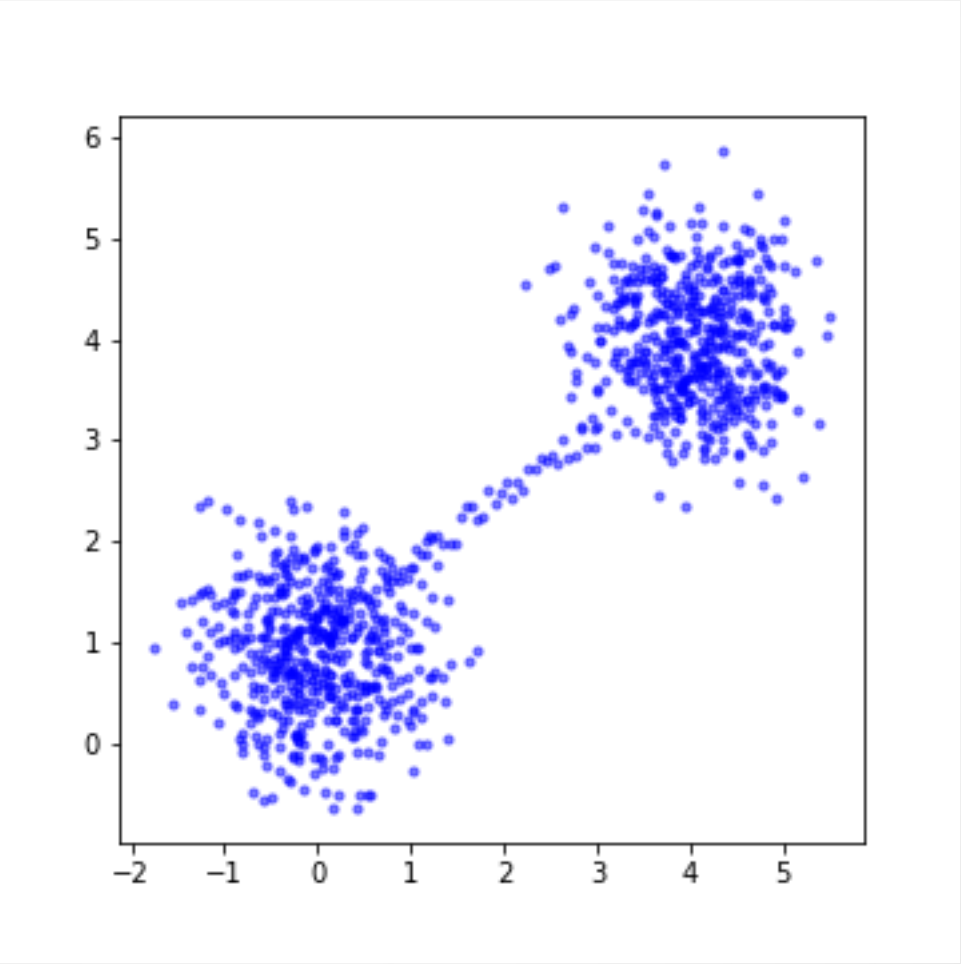
типы кластеров: кластеры с центром



другие типы кластеров:
• могут отсутствовать

кластеризация

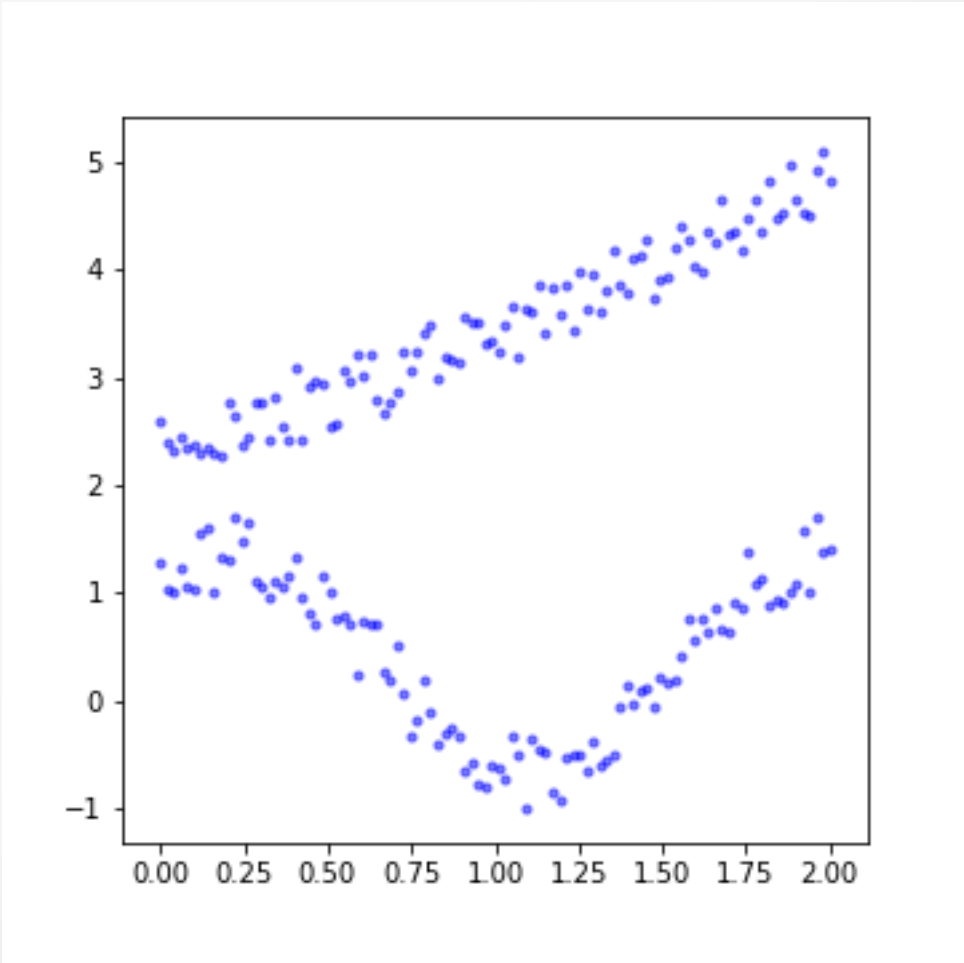
тип кластера: кластеры с перемычками



другие типы кластеров:
- могут отсутствовать
- кластеры с центром

кластеризация

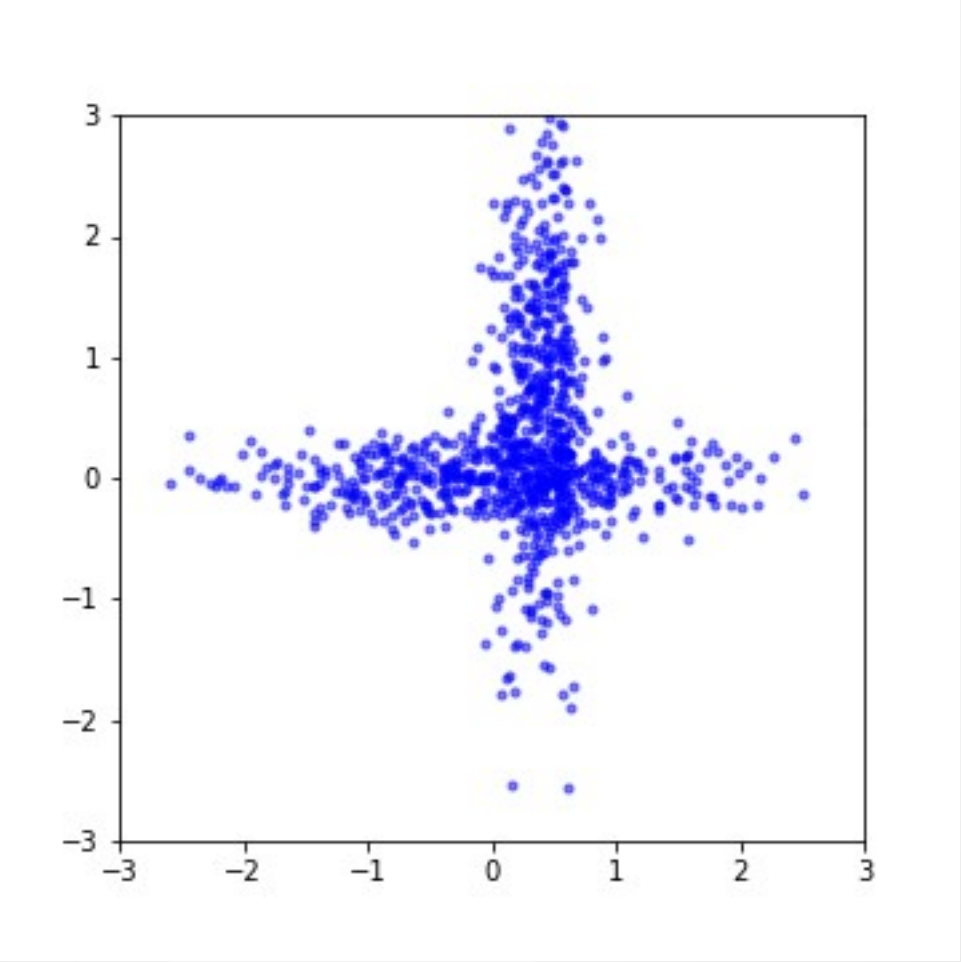
тип кластера: кластеры ленточные



другие типы кластеров:
- могут отсутствовать
- кластеры с центром
- кластеры с перемычками

кластеризация

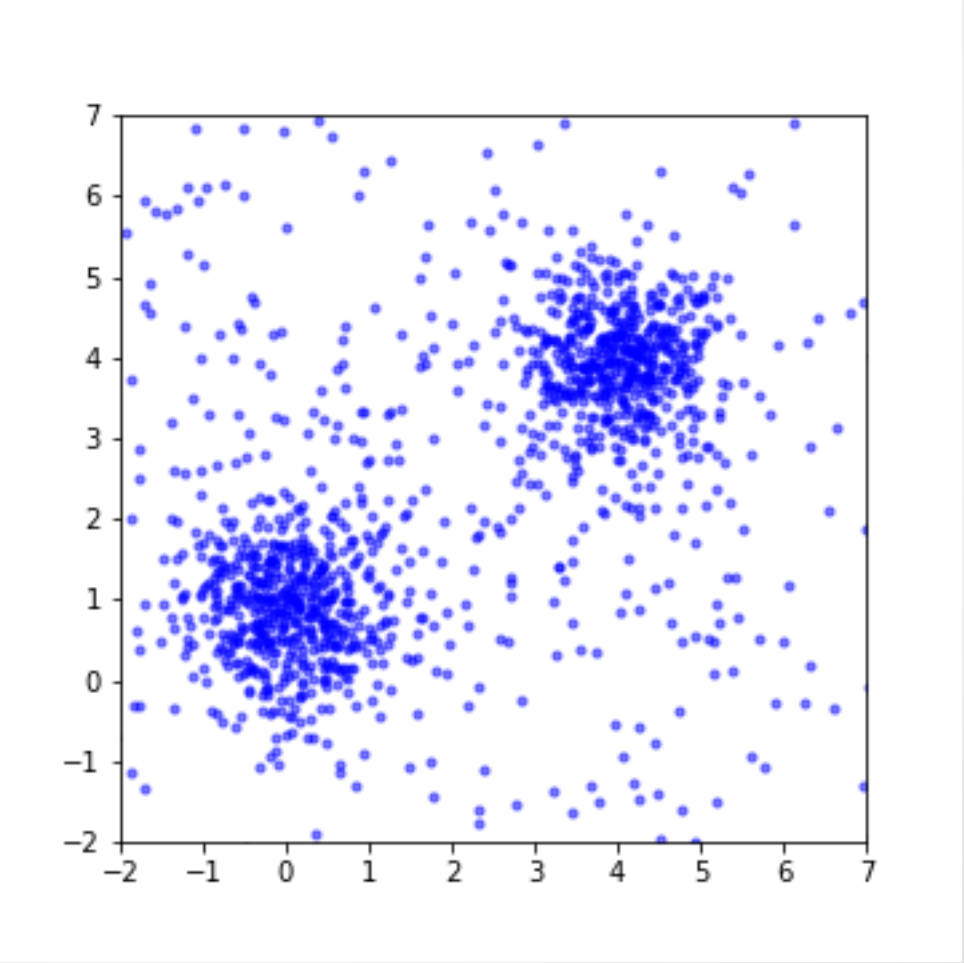
тип кластера: кластеры с наложением



другие типы кластеров:
- могут отсутствовать
- кластеры с центром
- кластеры с перемычками
- кластеры ленточные

кластеризация

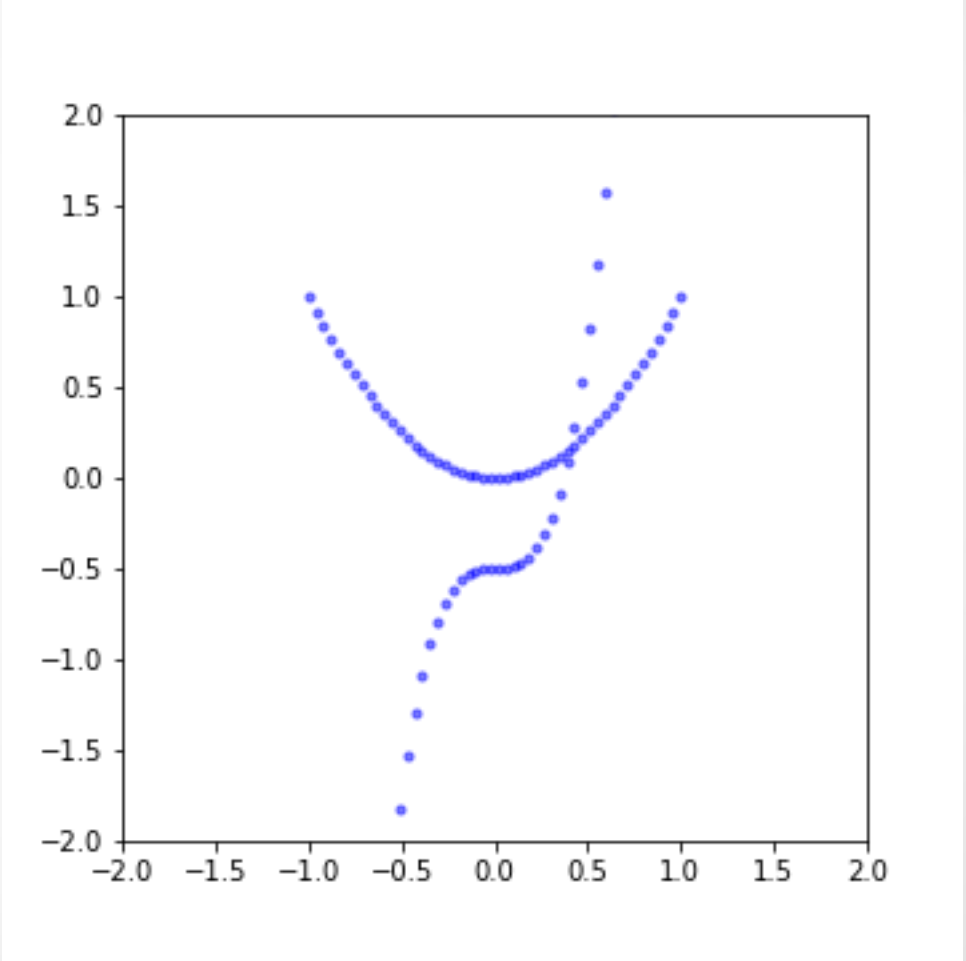
тип кластера: кластеры с шумом



другие типы кластеров:
- могут отсутствовать
- кластеры с центром
- кластеры с перемычками
- кластеры ленточные
- кластеры с наложением

кластеризация

тип кластера: кластеры по типу регулярности



другие типы кластеров:
- могут отсутствовать
- кластеры с центром
- кластеры с перемычками
- кластеры ленточные
- кластеры с наложением
- кластеры с шумом

кластеризация

оценки кластеризации $a : X \to Y$

$$ri = \frac{\sum_{i<j} [a_i = a_j] \rho(x_i, x_j)}{\sum_{i<j} [a_i = a_j]} \to min$$

среднее внутрикластерное расстояние

$$ro = \frac{\sum_{i<j} [a_i \neq a_j] \rho(x_i, x_j)}{\sum_{i<j} [a_i = a_j]} \to max$$

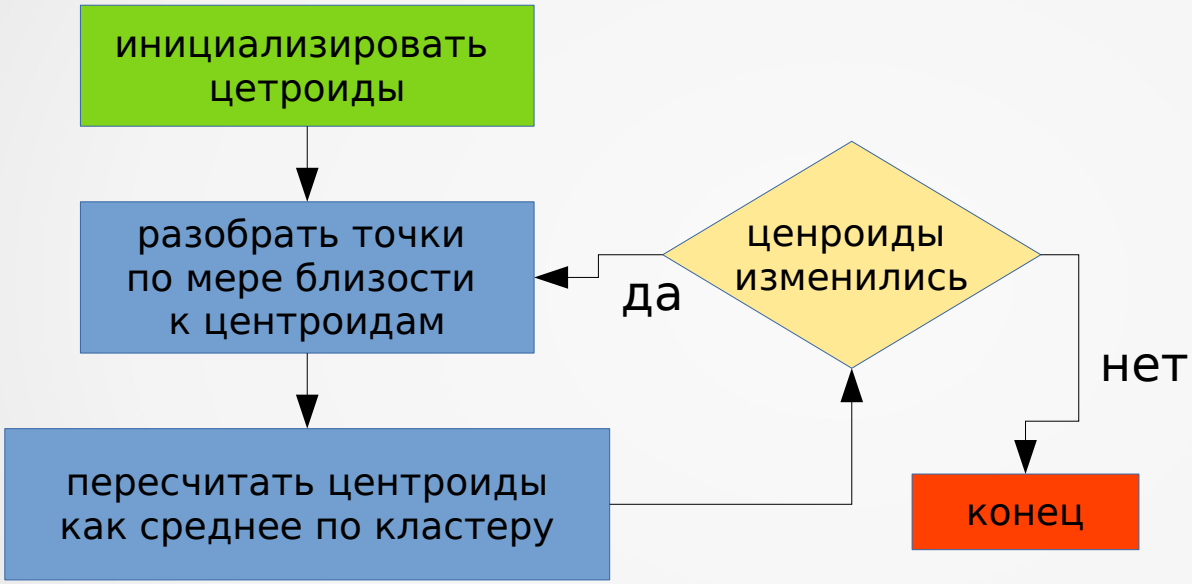
среднее межкластерное расстояние

отношение внутрикластерного и межкластерного расстояний

$$\frac{ri}{ro} \to min$$

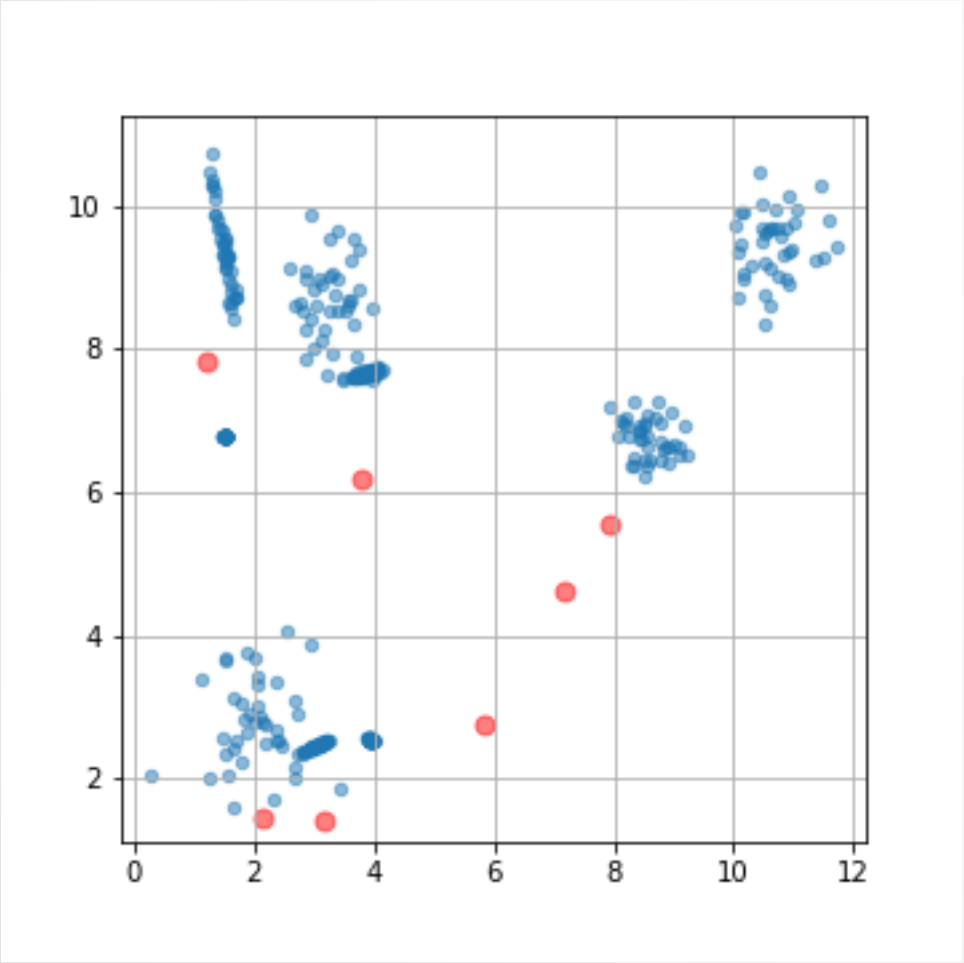
кластеризация

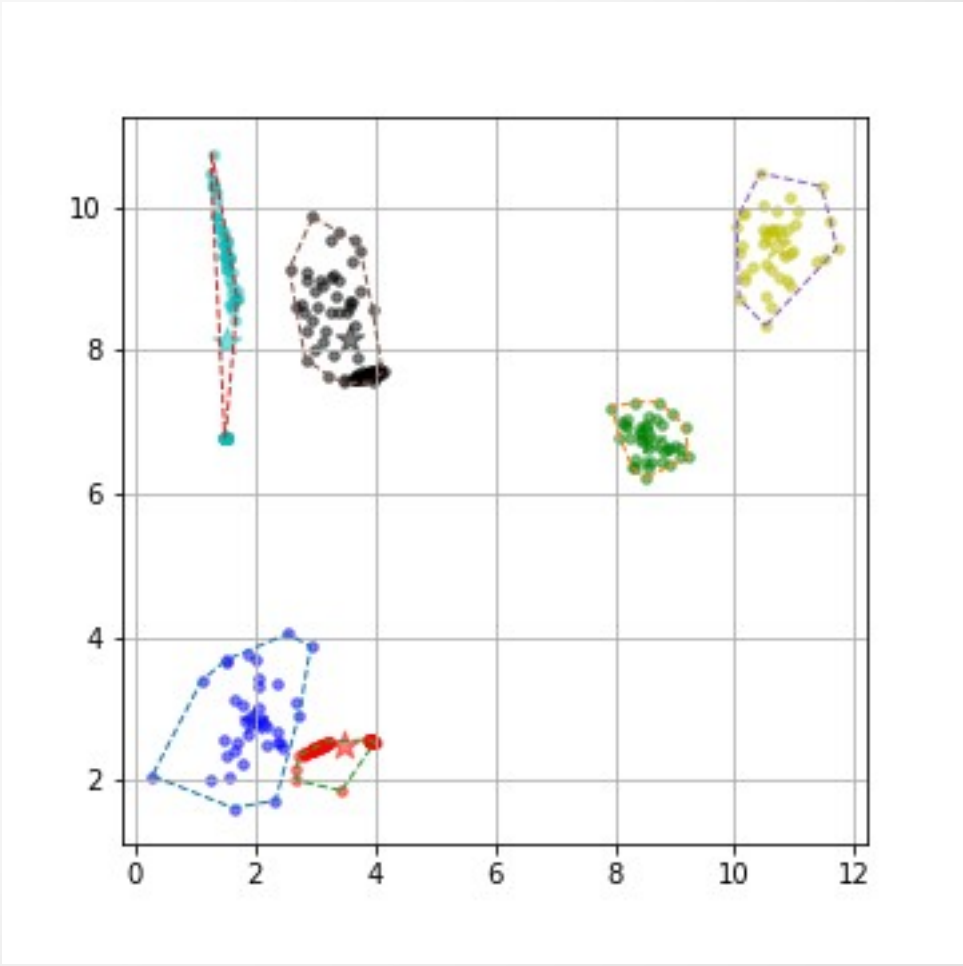
k-means
количество кластеров как параметр,
цель - найти точки-центроиды



кластеризация

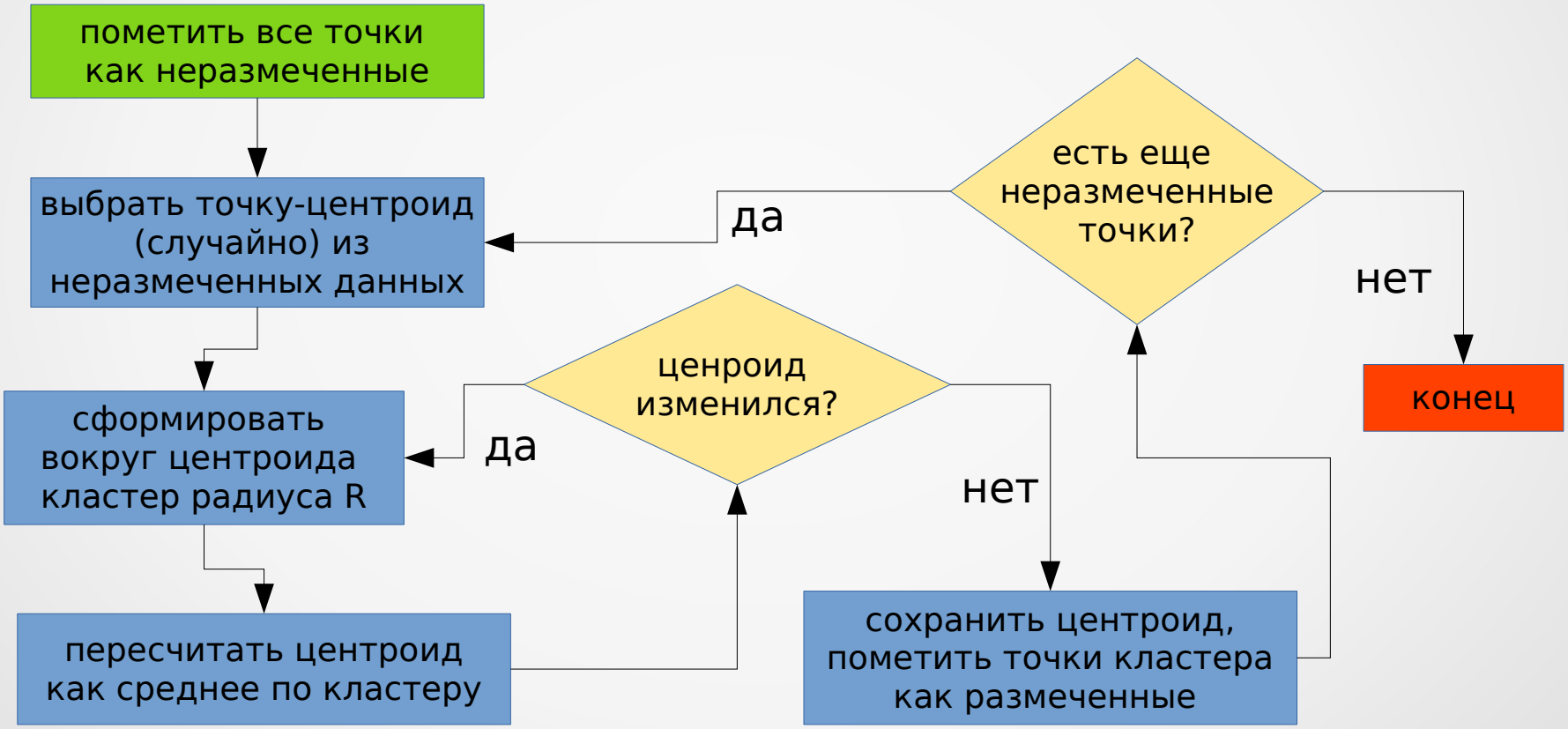
k-means: начальное состояние и результат





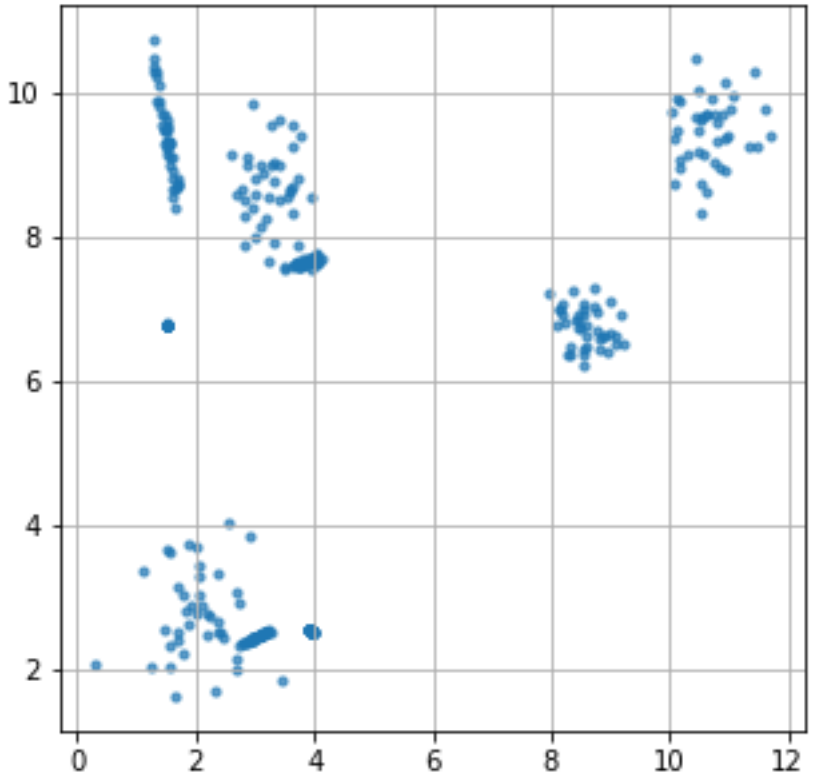
кластеризация

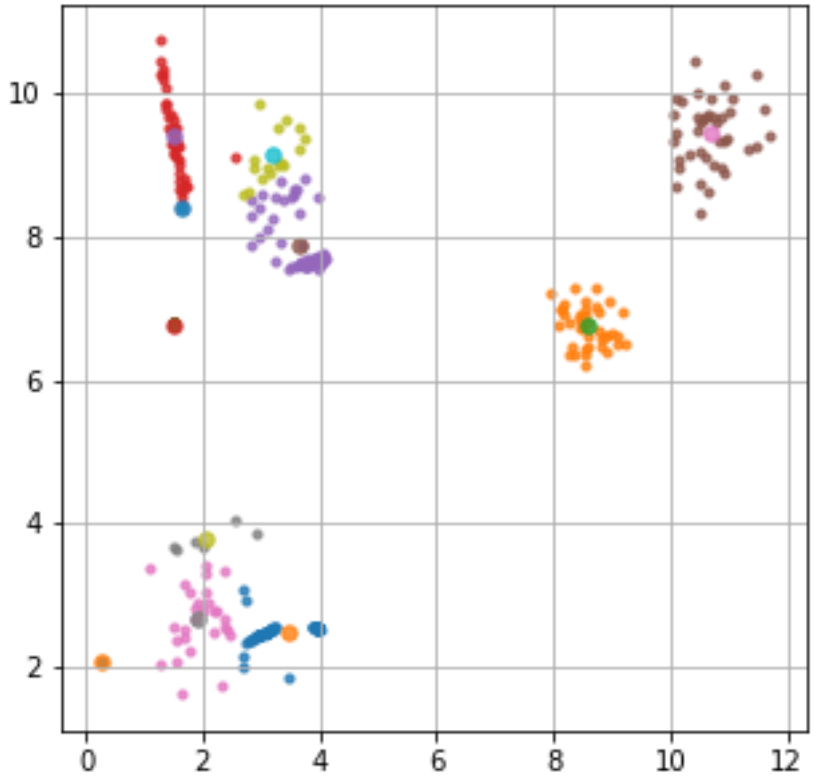
метод ФОРЭЛ
фиксируем радиус R кластеров,
цель - найти точки-центроиды



кластеризация

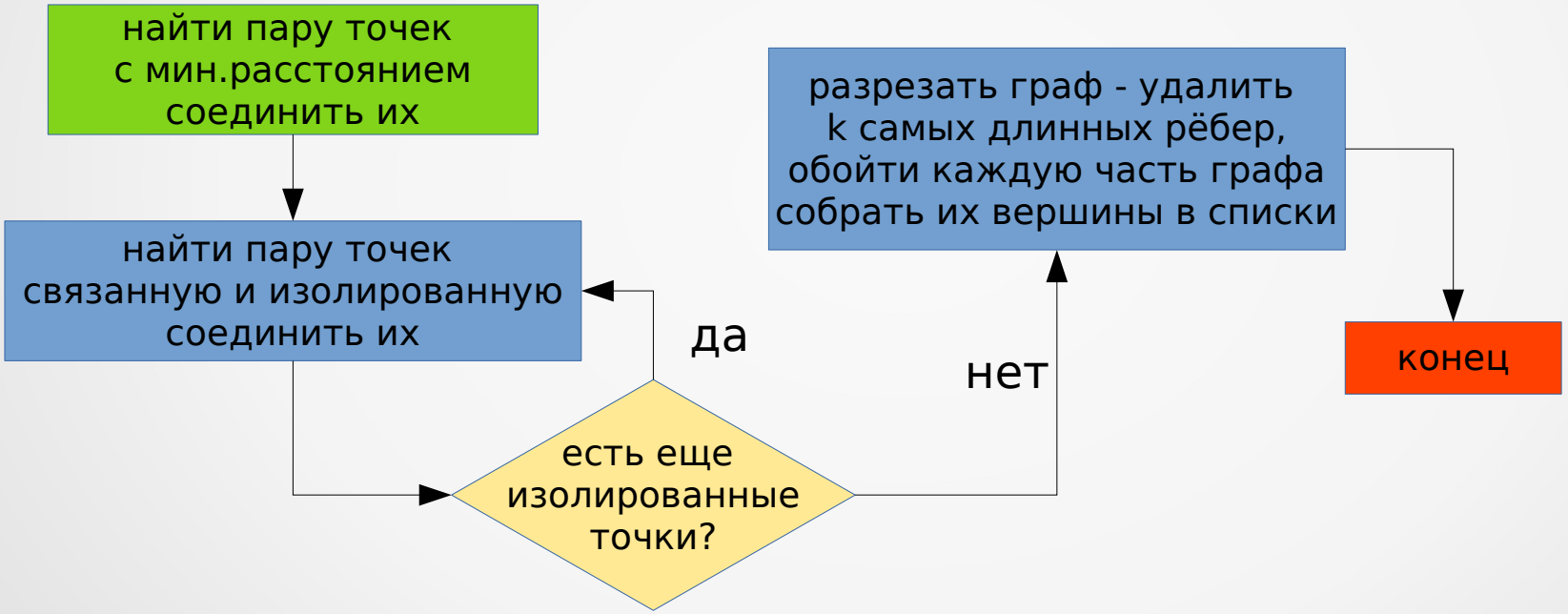
ФОРЭЛ: начальное состояние и результат





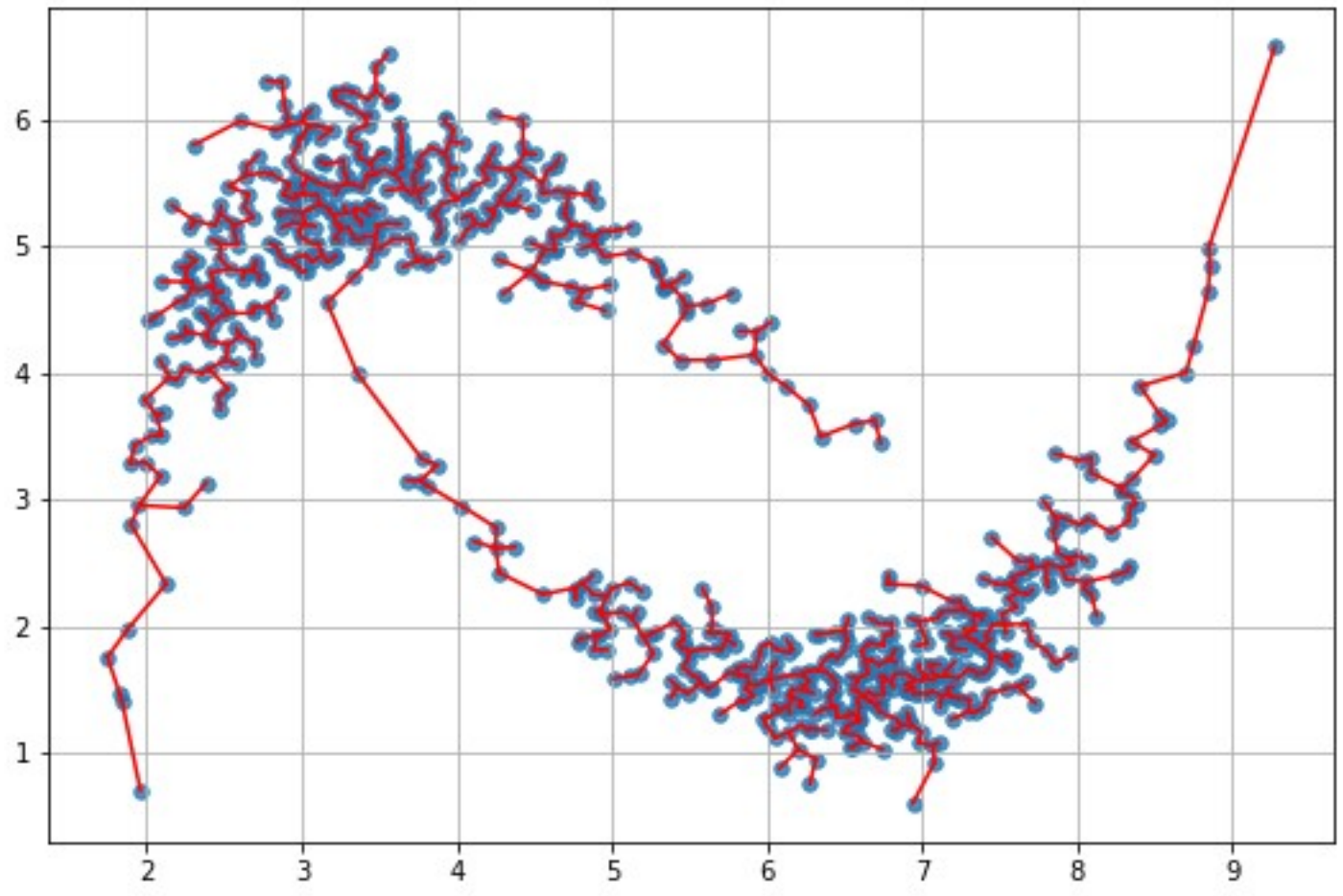
кластеризация

метод КНП (Кратчайший Незамкнутый Путь)
параметр - количество кластеров k
цель - построить ациклический граф на точках



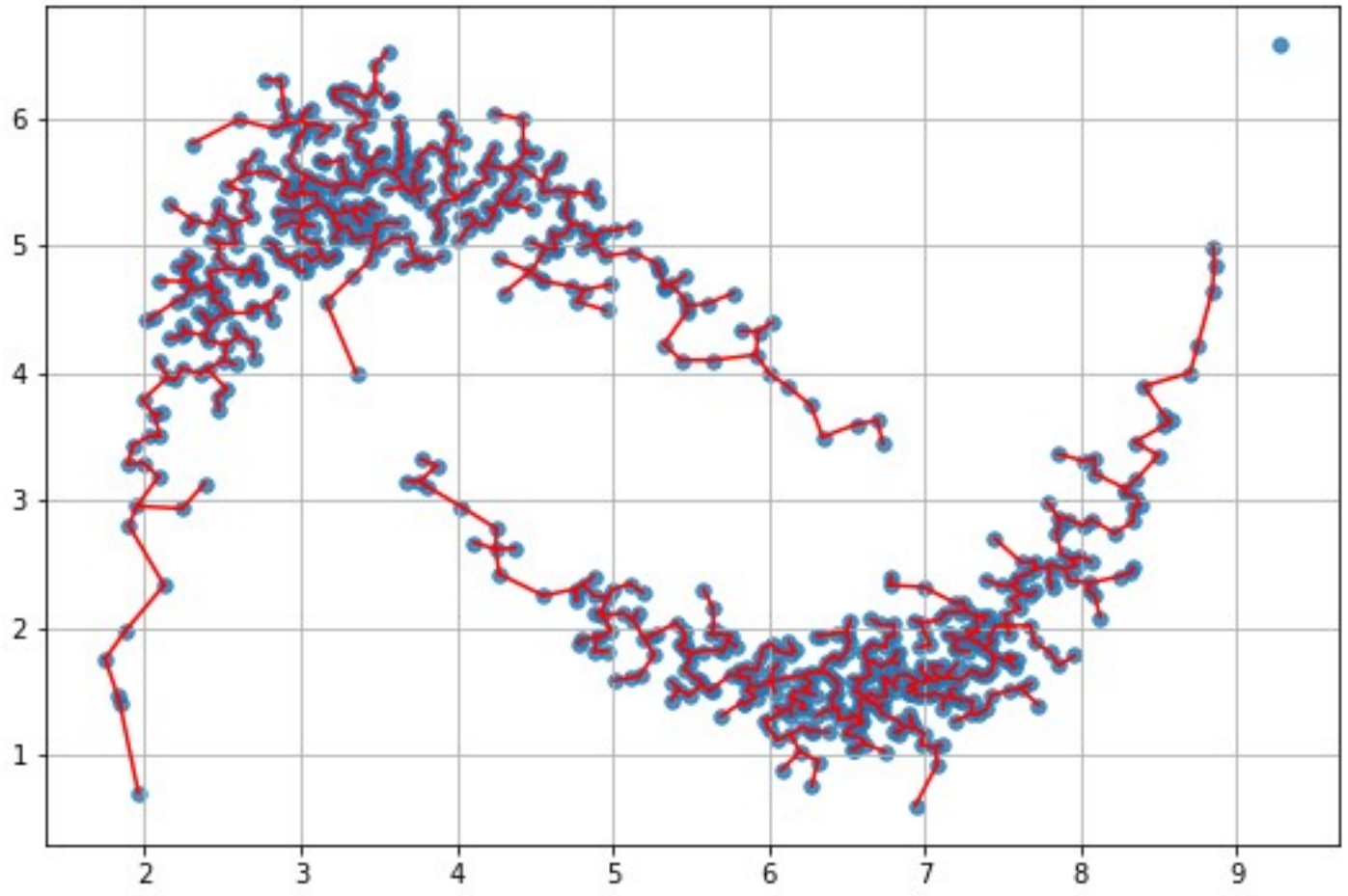
кластеризация

КНП: полный граф



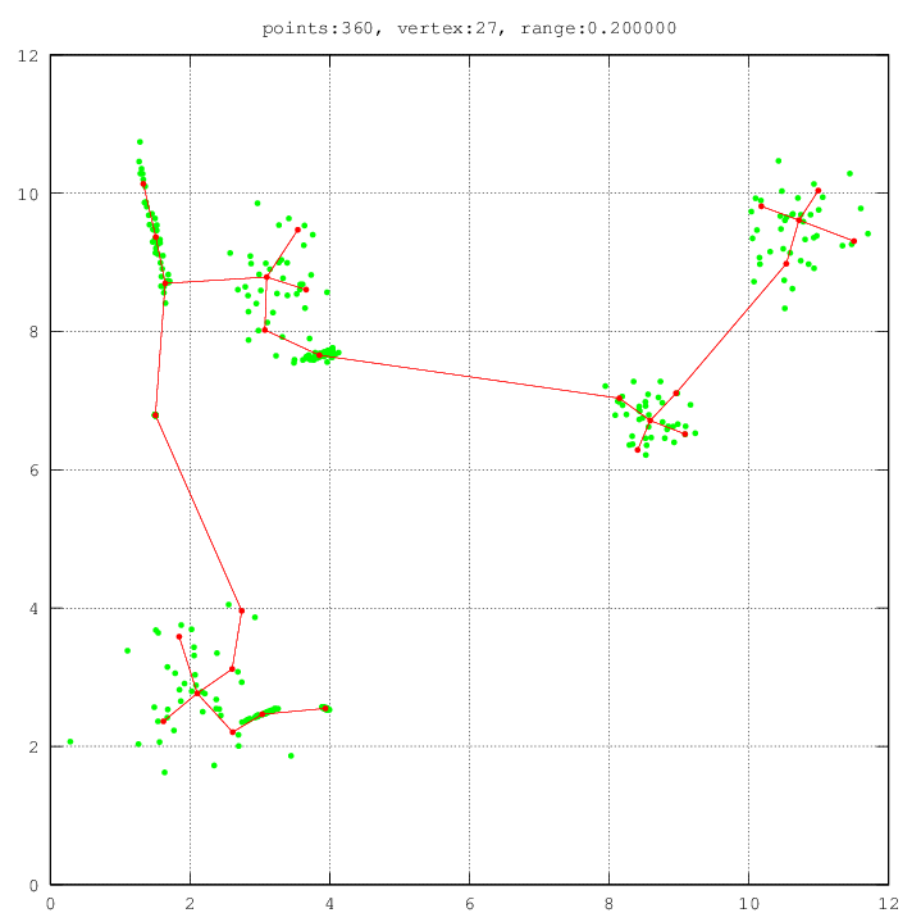
кластеризация

КНП: разрезанный граф



кластеризация

ФОРЭЛ + КНП



кластеризация

dbscan
иерархическая кластеризация
частичное обучение SSL

кластеризация: литература

git clone *https://github.com/mechanoid5/ml_lectorium.git*

- К.В. Воронцов Методы кластеризации. - курс "Машинное обучение" ШАД Яндекс 2014
- Е.С.Борисов Кластеризатор на основе алгоритма k-means. *http://mechanoid.kiev.ua/ml-k-means.html*
- Е.С.Борисов Метод кластеризации КНП. *http://mechanoid.kiev.ua/ml-knp.html*
- Е.С.Борисов Метод кластеризации ФОРЭЛ. *http://mechanoid.kiev.ua/ml-forel.html*
- Е.С.Борисов Метод иерархической кластеризации. *http://mechanoid.kiev.ua/ml-lnwl.html*
-

кластеризация

Вопросы ?

кластеризация: практика

источники данных для экспериментов

sklearn.datasets
UCI Repository
kaggle

задание

реализовать итоговый обход графа для КНП
реализовать комбинированный метод ФОРЭЛ+КНП
применить кластеризаторы для разных наборов данных
и посчитать оценку результатов кластеризации