Unsupervised learning

Ушаков Роман

April, 2018

Сегодня в программе

1. Кластеризация
   - Что такое кластеризация?
   - С чем это едят?

2. Методы кластеризации
   - kMeans
   - Affinity Propagation
   - Агломеративная кластеризация

План

1. Кластеризация
   - Что такое кластеризация?
   - С чем это едят?
2. Методы кластеризации
   - kMeans
   - Affinity Propagation
   - Агломеративная кластеризация

Кластеризация

- **Входные данные:**
  - Признаковое описание: $X = \{x_1, \ldots, x_l\}$ — объекты из $\mathbb{R}^n$;
  - Матрица попарных расстояний: $D = \{d_1, \ldots, d_l\}$ — объекты из $\mathbb{R}^l$.

Кластеризация

- **Входные данные:**
  - Признаковое описание: $X = \{x_1, \ldots, x_l\}$ — объекты из $\mathbb{R}^n$;
  - Матрица попарных расстояний: $D = \{d_1, \ldots, d_l\}$ — объекты из $\mathbb{R}^l$.
- **Задача** разделить объекты на кластеры:
  (a) объекты **в одном** кластере **похожи** друг на друга
  (b) объекты **в разных** кластерах существенно **отличаются**

Кластеризация

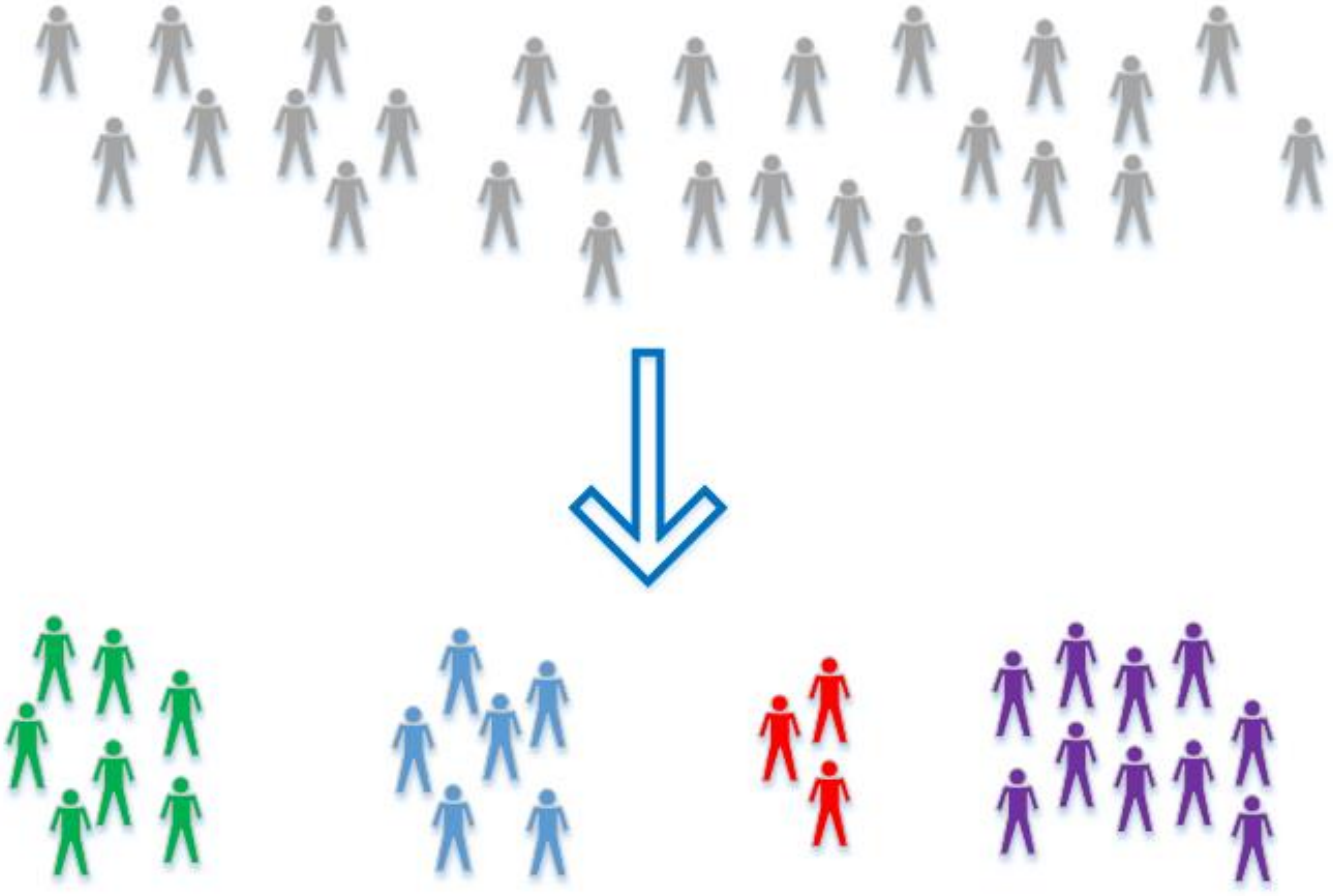
- **Входные данные:**
  - Признаковое описание: $X = \{x_1, \ldots, x_l\}$ — объекты из $\mathbb{R}^n$;
  - Матрица попарных расстояний: $D = \{d_1, \ldots, d_l\}$ — объекты из $\mathbb{R}^l$.
- **Задача** разделить объекты на кластеры:
  (a) объекты **в одном** кластере **похожи** друг на друга
  (b) объекты **в разных** кластерах существенно **отличаются**
- **Цели** кластеризации:
  - Понимание данных (разбиение на группы схожих объектов);
  - Сжатие данных (выбор представителей кластеров);
  - Обнаружение новизны

План

1. Кластеризация
   - Что такое кластеризация?
   - С чем это едят?
2. Методы кластеризации
   - kMeans
   - Affinity Propagation
   - Агломеративная кластеризация

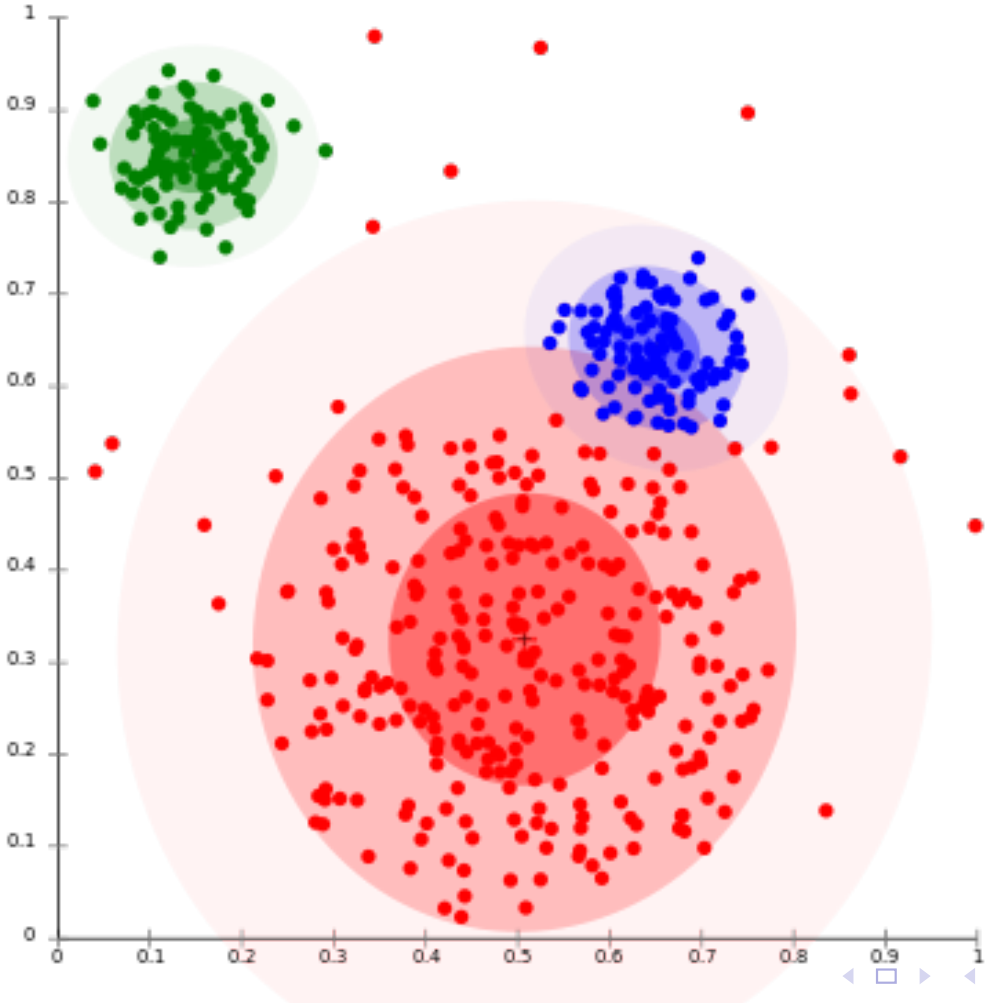
Примеры использования

Сегментация пользователей



Примеры использования

Поиск аномалий

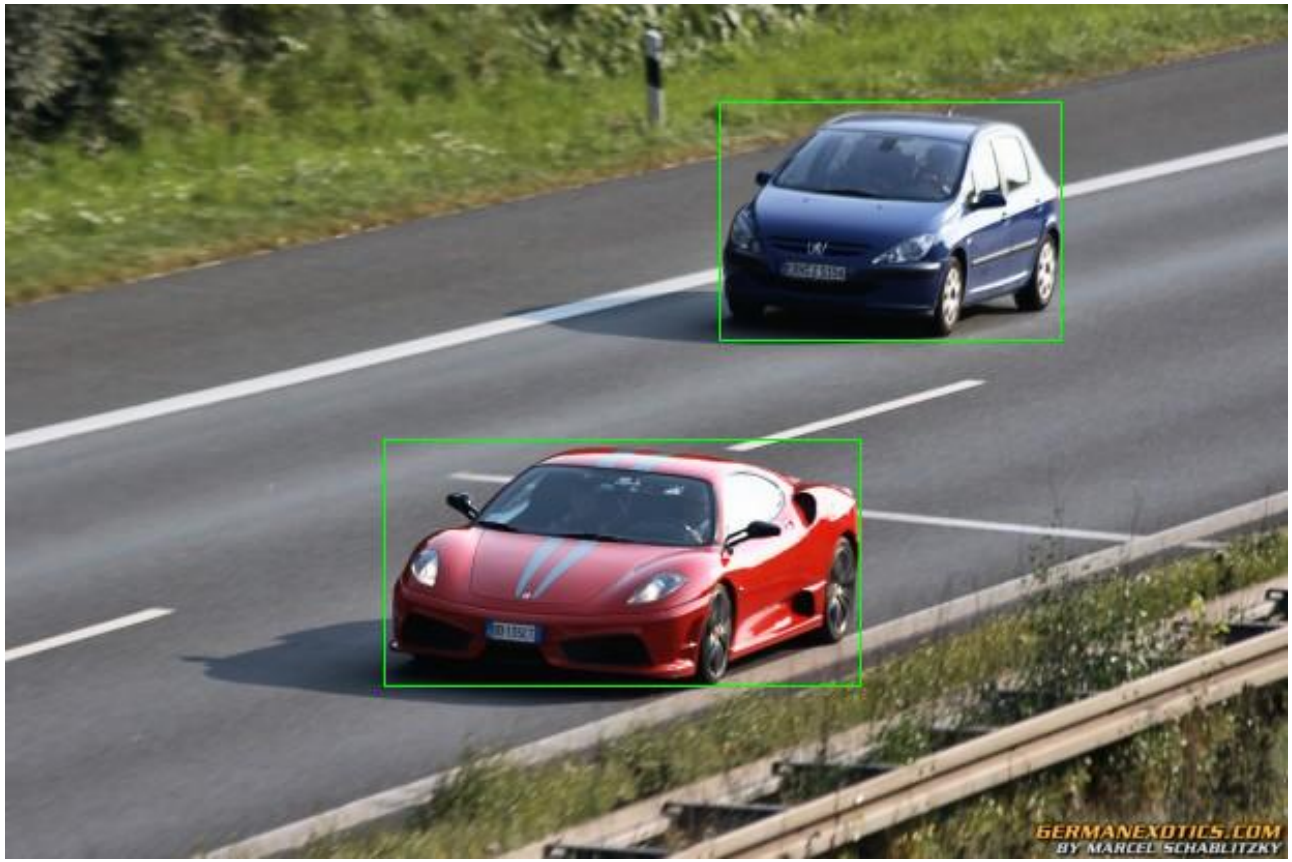


Примеры использования

Определение тематики текста

Примеры использования

Сегментация изображений



План

1. Кластеризация
   - Что такое кластеризация?
   - С чем это едят?

2. Методы кластеризации
   - kMeans
   - Affinity Propagation
   - Агломеративная кластеризация

kMeans

5 простых шагов:

- Выбраем количество кластеров k, которое нам кажется оптимальным для наших данных;

kMeans

5 простых шагов:

- Выбраем количество кластеров k, которое нам кажется оптимальным для наших данных;
- Раскидываем случайным образом в пространство наших данных k точек (центроидов);

kMeans

5 простых шагов:

- Выбраем количество кластеров k, которое нам кажется оптимальным для наших данных;
- Раскидываем случайным образом в пространство наших данных k точек (центроидов);
- Для каждой точки нашего набора данных посчитать, к какому центроиду она ближе;

kMeans

5 простых шагов:

- Выбраем количество кластеров k, которое нам кажется оптимальным для наших данных;
- Раскидываем случайным образом в пространство наших данных k точек (центроидов);
- Для каждой точки нашего набора данных посчитать, к какому центроиду она ближе;
- Переместить каждый центроид в центр выборки, которую мы отнесли к этому центроиду;

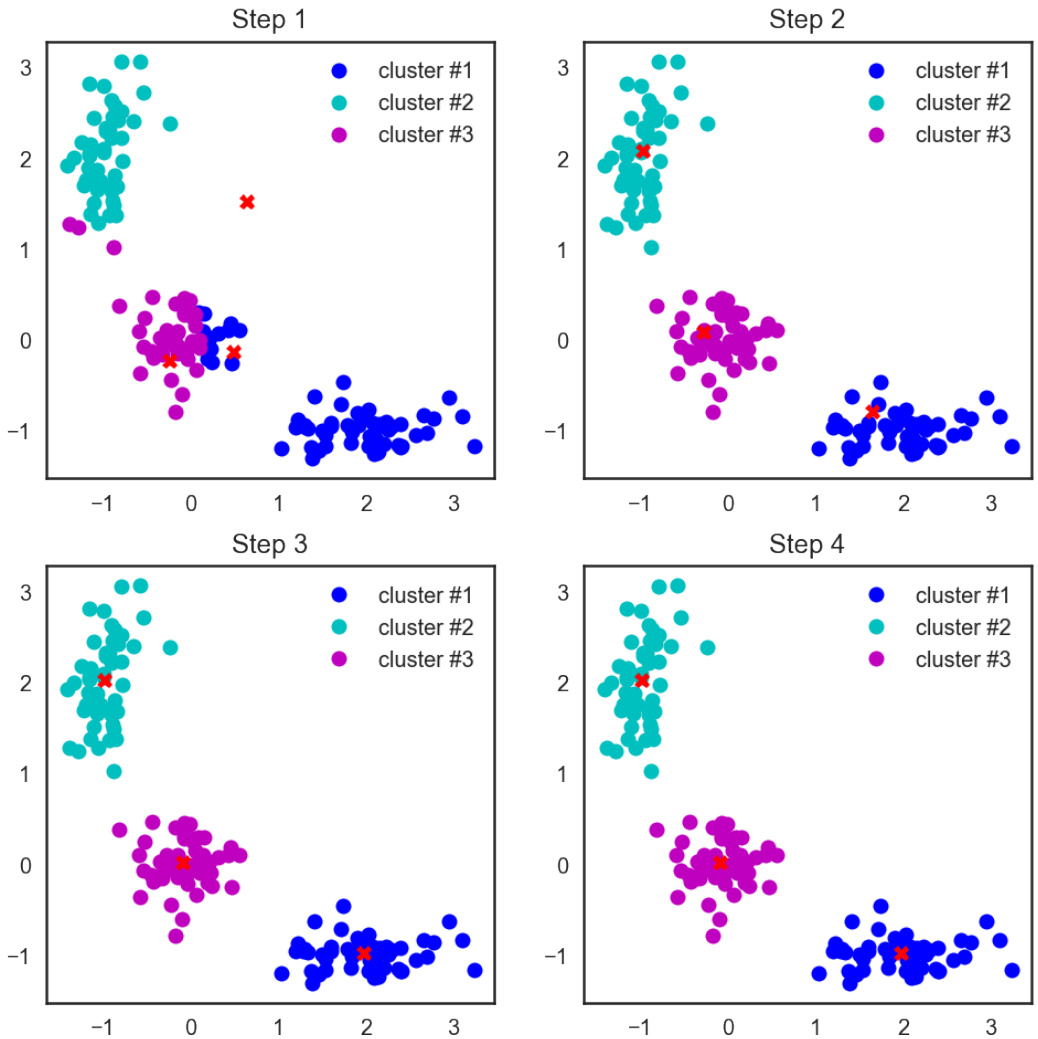
kMeans

5 простых шагов:

- Выбраем количество кластеров k, которое нам кажется оптимальным для наших данных;
- Раскидываем случайным образом в пространство наших данных k точек (центроидов);
- Для каждой точки нашего набора данных посчитать, к какому центроиду она ближе;
- Переместить каждый центроид в центр выборки, которую мы отнесли к этому центроиду;
- Повторять последние два шага фиксированное число раз, либо до тех пор пока центроиды не "сойдутся".

kMeans

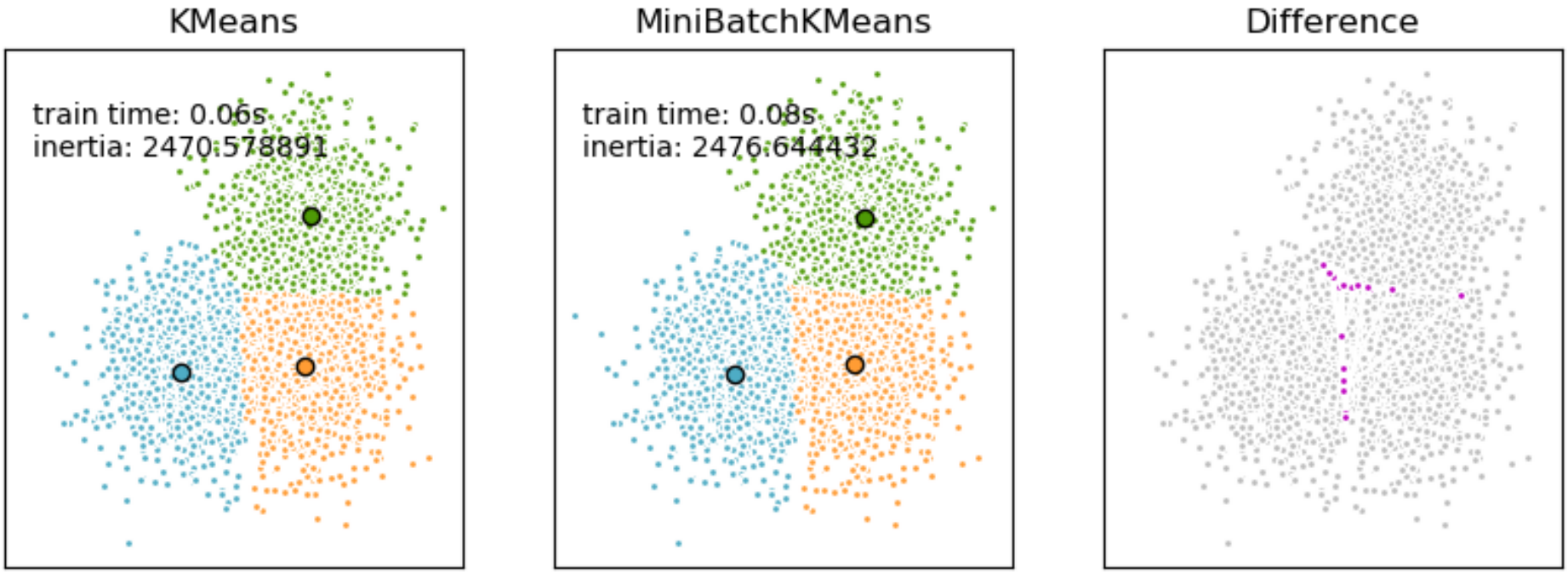
Пошаговая визуализация алгоритма



kMeans

Сложность алгоритма: $O(l^{nk+1})$, где n – размерность пространств, k – количество кластеров и l – количество объектов.

MiniBatch kMeans:



kMeans


Когда сделал Кминс кластеризацию
знания
это
сила
memesmix.net

План

1. Кластеризация
   - Что такое кластеризация?
   - С чем это едят?
2. Методы кластеризации
   - kMeans
   - Affinity Propagation
   - Агломеративная кластеризация

Affinity Propagation

- $s(x_i, x_j)$ – правило «похожести», $S = (s(x_i, x_j))$ – матрица «схожести».

Affinity Propagation

- $s(x_i, x_j)$ – правило «похожести», $S = (s(x_i, x_j))$ – матрица «схожести».
- $r(x_i, x_k)$ – насколько x_i хочет видеть x_k своим представителем, $R = (r(x_i, x_j))$ – матрица «ответственности» (responsibility).

Affinity Propagation

- $s(x_i, x_j)$ – правило «похожести», $S = (s(x_i, x_j))$ – матрица «схожести».
- $r(x_i, x_k)$ – насколько x_i хочет видеть x_k своим представителем, $R = (r(x_i, x_j))$ – матрица «ответственности» (responsibility).
- $a(x_i, x_k)$ – насколько хорошо x_k готова представлять интересы x_i, $A = (a(x_i, x_j))$ – матрица «доступности» (availability).

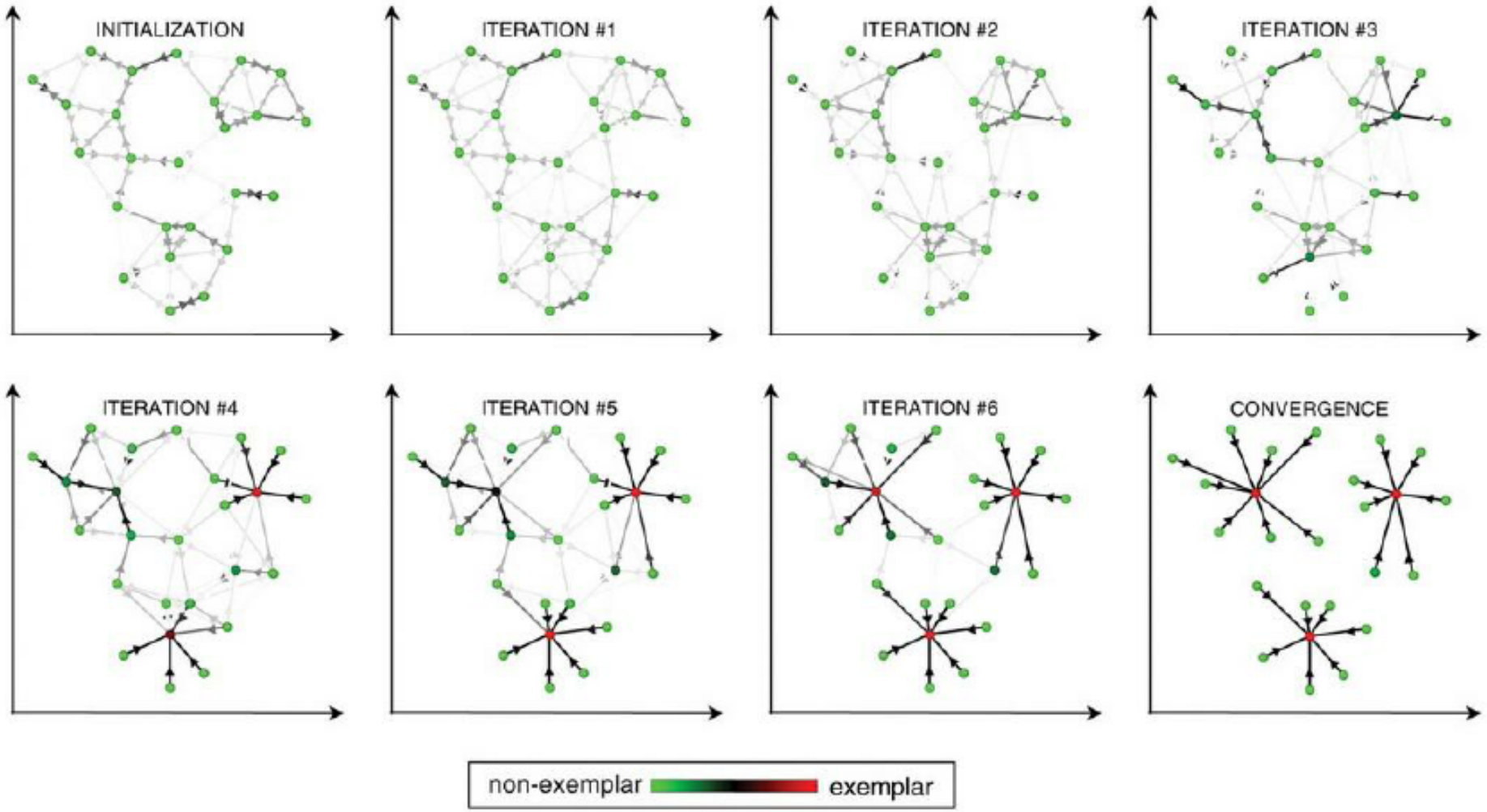
Affinity Propagation

Матрицы R и A обновляются по очереди:

- $r_{i,k} \leftarrow s_{i,k} - max_{k' \neq k}(a_{i,k'} + s_{i,k'})$
- $a_{i,k} \leftarrow min(0, r_{k,k} + \sum_{i' \notin i,k} max(0, r_{i',k}))$
- $a_{k,k} \leftarrow \sum_{i' \neq k} max(0, r_{i',k})$

$c_i \leftarrow argmax_k(r_{i,k} + a_{i,k})$

Affinity Propagation



План

1. Кластеризация
   - Что такое кластеризация?
   - С чем это едят?
2. Методы кластеризации
   - kMeans
   - Affinity Propagation
   - Агломеративная кластеризация

Агломеративная кластеризация

4 простых шага

- Присваиваем каждой точке свой кластер;

Агломеративная кластеризация

4 простых шага

- Присваиваем каждой точке свой кластер;
- Сортируем попарные расстояния между кластерами по возрастанию;

Агломеративная кластеризация

4 простых шага

- Присваиваем каждой точке свой кластер;
- Сортируем попарные расстояния между кластерами по возрастанию;
- Берём пару ближайших кластеров, склеиваем их в один и пересчитываем центр кластера;

Агломеративная кластеризация

4 простых шага

- Присваиваем каждой точке свой кластер;
- Сортируем попарные расстояния между кластерами по возрастанию;
- Берём пару ближайших кластеров, склеиваем их в один и пересчитываем центр кластера;
- Повторяем два последних пункта до тех пор, пока все данные не склеятся в один кластер.

Агломеративная кластеризация

Расстояния между кластерами

- Single linkage — минимум попарных расстояний между точками из двух кластеров
  $d(C_i, C_j) = min_{x_i \in C_i, x_j \in C_j} ||x_i - x_j||$

Агломеративная кластеризация

Расстояния между кластерами

- Single linkage — минимум попарных расстояний между точками из двух кластеров
  $d(C_i, C_j) = min_{x_i \in C_i, x_j \in C_j} ||x_i - x_j||$
- Complete linkage — максимум попарных расстояний между точками из двух кластеров
  $d(C_i, C_j) = max_{x_i \in C_i, x_j \in C_j} ||x_i - x_j||$

Агломеративная кластеризация

Расстояния между кластерами

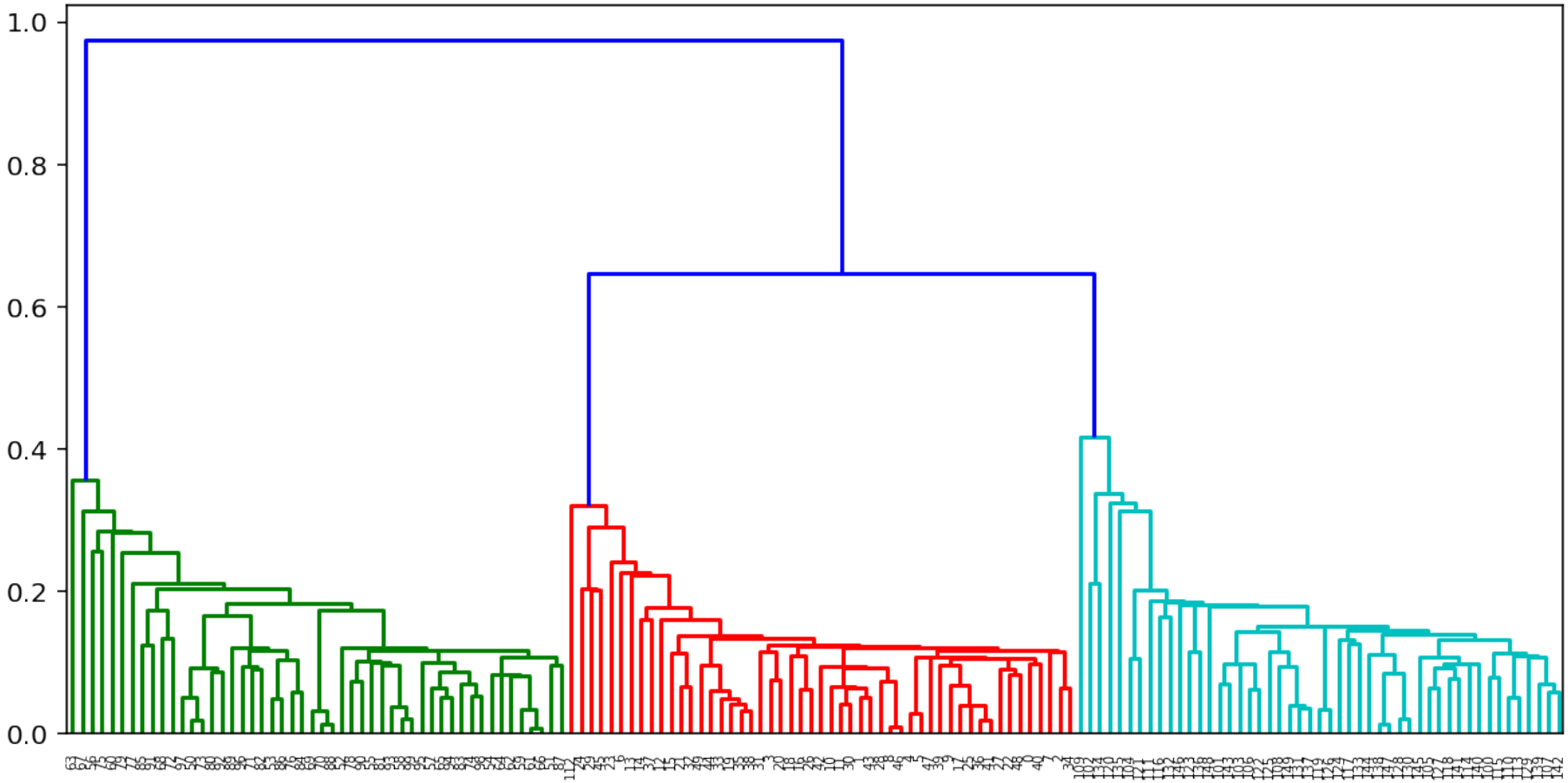
- Single linkage — минимум попарных расстояний между точками из двух кластеров
  $d(C_i, C_j) = min_{x_i \in C_i, x_j \in C_j} ||x_i - x_j||$
- Complete linkage — максимум попарных расстояний между точками из двух кластеров
  $d(C_i, C_j) = max_{x_i \in C_i, x_j \in C_j} ||x_i - x_j||$
- Average linkage — среднее попарных расстояний между точками из двух кластеров
  $d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x_i \in C_i} \sum_{x_j \in C_j} ||x_i - x_j||$

Агломеративная кластеризация

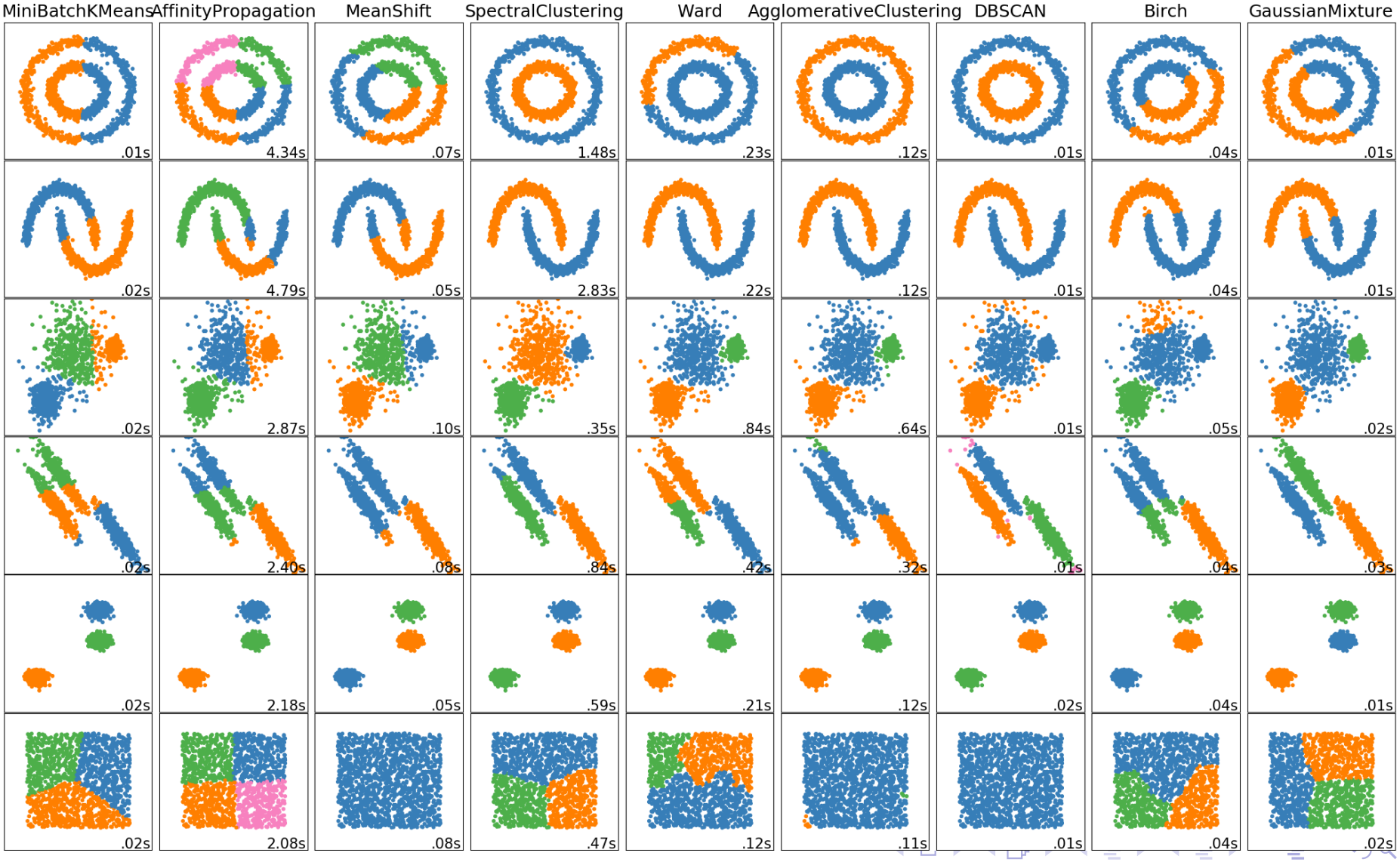
Расстояния между кластерами

- Single linkage — минимум попарных расстояний между точками из двух кластеров
  $d(C_i, C_j) = min_{x_i \in C_i, x_j \in C_j} ||x_i - x_j||$
- Complete linkage — максимум попарных расстояний между точками из двух кластеров
  $d(C_i, C_j) = max_{x_i \in C_i, x_j \in C_j} ||x_i - x_j||$
- Average linkage — среднее попарных расстояний между точками из двух кластеров
  $d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x_i \in C_i} \sum_{x_j \in C_j} ||x_i - x_j||$
- Centroid linkage — расстояние между центроидами двух кластеров
  $d(C_i, C_j) = ||\mu_i - \mu_j||$

Агломеративная кластеризация



Сравнение алгоритмов



Итог

1. Кластеризация — разбиение множества объектов по “сообществам“
2. kMeans подходит для большого числа задач, начинать проводить анализ нужно с него
3. Чаще всего, количество кластеров — гиперпараметр