Лекция 8. Кластеризация

Основы интеллектуального анализа данных

Полузёров Т. Д.

БГУ ФПМИ

Постановка задачи

Дано:
$\mathbb{X}$ — пространтсов объектов
$X = \{x_1, \ldots, x_\ell\}$ — обучающая выборка
$\rho : \mathbb{X} \times \mathbb{X} \to [0, +\infty]$ — функция расстояния между объектами

Необходимо:
Определить множество кластеров $\mathbb{Y}$ и построить алгоритм $a : \mathbb{X} \to \mathbb{Y}$ так, чтобы:

- каждый кластер состоял из близких объектов
- объекты разных кластеров были сущетсвенно различны

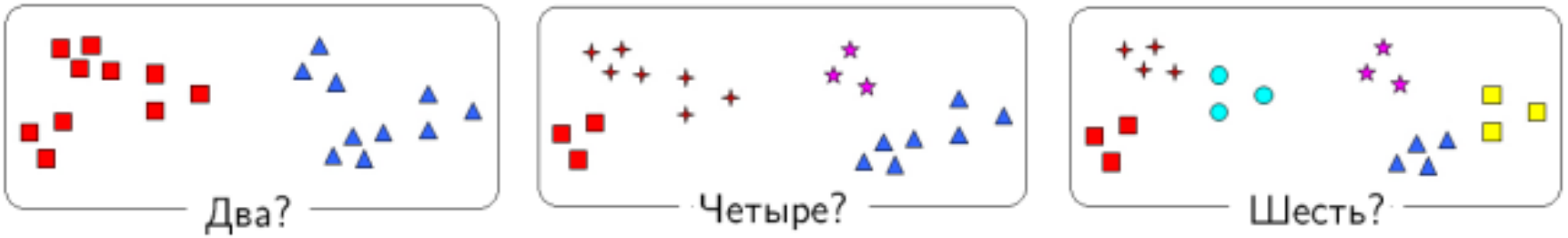
Это задача **кластеризации** — частный случай задач обучения без учителя.

Некорректность задачи кластеризации

Особенности задачи:

- точной постановки задачи нет
- непонятно как выбрать меру качества кластеризации
- априорно неизвестно число кластеров
- результат сильно зависит от меры расстояния ρ

Пример: сколько здесь кластеров?



Цели кластеризации

- **Упростить обработку данных**,
  разбить все множество X на группы схожих объектов, чтобы работать с каждой группой отдельно
- **Сократить объем данных**,
  оставить по одному представителю от группы
- **Выделить нетипичные объекты**,
  которые не относятся ни к одному из кластеров
- **Построить иерархию множества объектов**

Типы кластерных структур

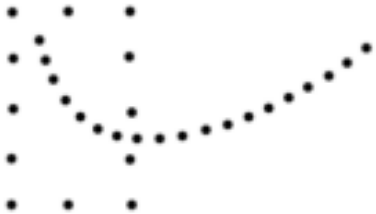
перемычки между кластерами

разреженный фон
из нетипичных объектов

перекрывающиеся кластеры

Типы кластерных структур

кластеры могут вообще отсутствовать



а это вообще не кластеры

- кластеры определяются субъективно
- каждый метод кластеризации имеет свои ограничения и способен работать только на некоторых типах кластеров

Метрическое пространство

Пусть известны только расстояния между оъъектами.
$a_i = a(x_i)$ — метка кластера объекта x_i

- Среднее внутрикластерное расстояния

$$F_0 = \frac{\sum_{i<j}[a_i = a_j]\rho(x_i, x_j)}{\sum_{i<j}[a_i = a_j]} \to \min$$

- Среднее межкластерное расстояние

$$F_1 = \frac{\sum_{i<j}[a_i \neq a_j]\rho(x_i, x_j)}{\sum_{i<j}[a_i \neq a_j]} \to \max$$

- Их отношение

$$\frac{F_0}{F_1} \to \min$$

Векторное пространство

Если объекты задаются вектором $x_i \in \mathbb{R}^n$

- Сумма средних внутрикластерных расстояний

$$\Phi_0 = \sum_{a \in \mathbb{Y}} \frac{1}{|X_a|} \sum_{i:a_i=a} \rho(x_i, \mu_a) \to \min$$

где $X_a = \{x_i \in X | a_i = a\}$ — кластер a, μ_a — центр кластера a

- Сумма межкластерных расстояний

$$\Phi_1 = \sum_{a,b \in \mathbb{Y}} \rho(\mu_a, \mu_b) \to \max$$

- отношение

$$\frac{\Phi_0}{\Phi_1} \to \min$$

Алгоритм: K-средник (K-means)

Algorithm 1: K-means

Input: X^ℓ, K

Output: центры кластеров $\mu_a, a = 1, \ldots, K$

случайно инициализировать $\mu_a, a = 1, \ldots, \mathrm{K}$

while *не перестанут изменяться* μ_a **do**

  отнести каждый x_i к ближайшему центру

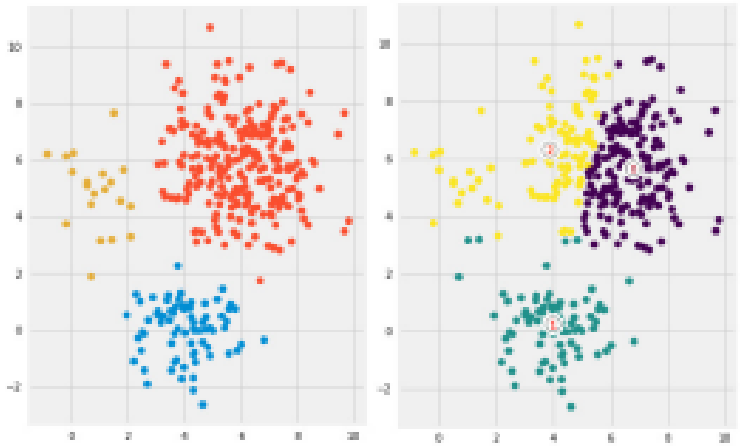
$$a_i := \arg\min_{a \in \mathbb{Y}} ||x_i - \mu_a||$$

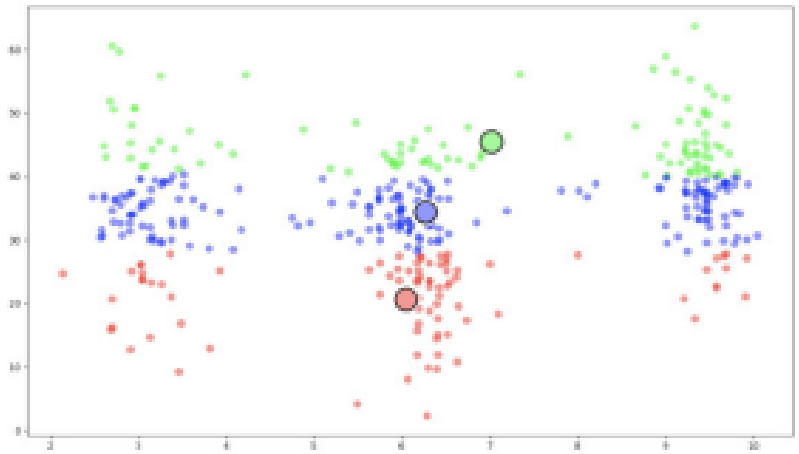
  вычислить новые положения центров

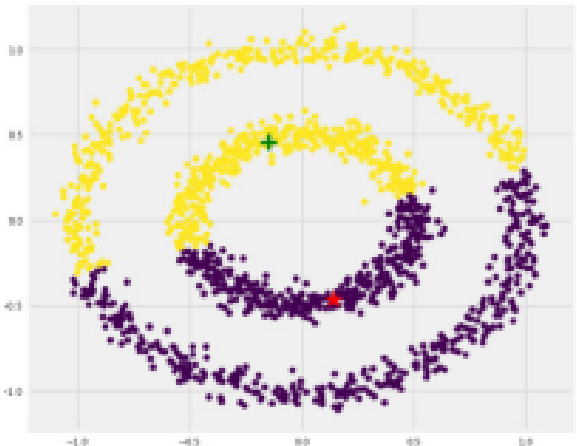
$$\mu_a := \frac{\sum_{i=1}^{\ell} [a_i = a] x_i}{\sum_{i=1}^{\ell} [a_i][a]}, a \in \mathbb{Y}$$

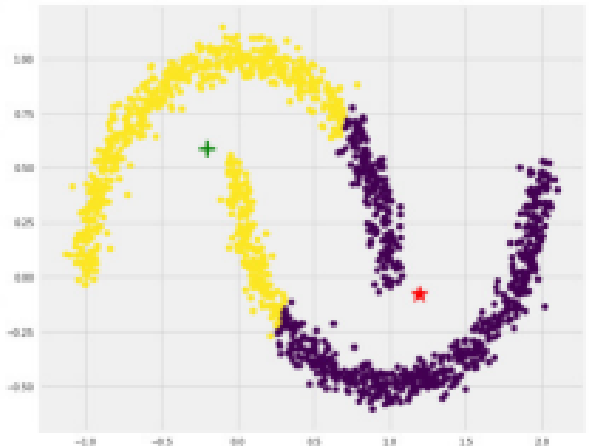
Неудачные примеры K-means

Причина — неудачное начальное приближение или существенная негауссовость кластеров









Алгоритм: DBSCAN
Density-Based Spatian Clustering of Applications with Noise

Зафиксируем 2 параметра:

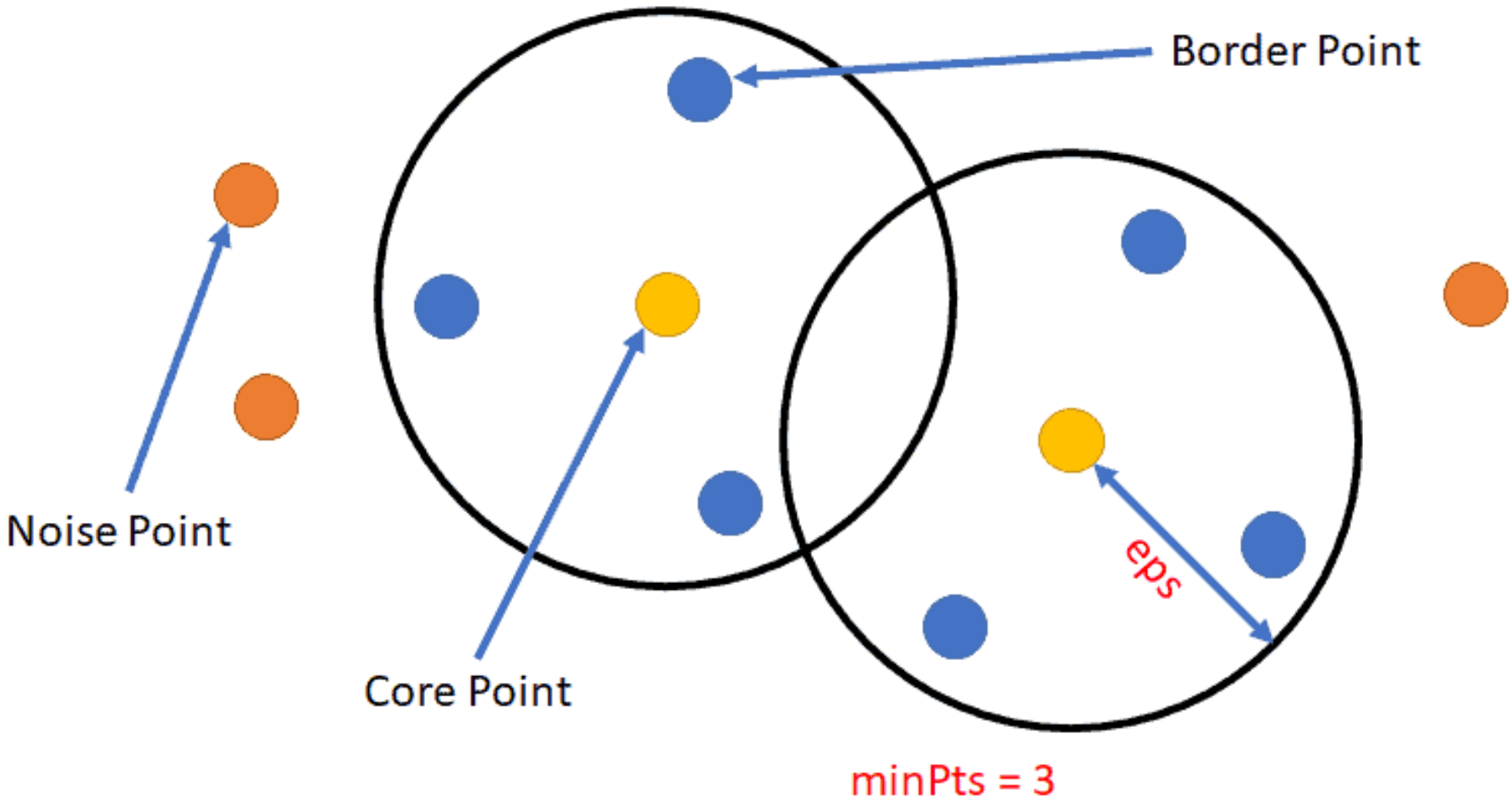
- ε — размер окрестности
- m — параметр плотности

ε-окресность точки $x \in U$ есть $U_\varepsilon(x) = \{u \in U : \rho(x, u) \leq \varepsilon\}$

3 типа объектов:

- **корневой** — имеет плотную окрестность, $|U_\varepsilon| \geq m$
- **граничный** — не корневой, но в окрестности корневого
- **шумовой** — все остальные

Типы точек



Algorithm 2: DBSCAN

Input: X^ℓ, ε, m

Output: разбиение на кластеры, определение шумовых объектов

$U := X$ — неразмеченные объектов $k := 0$ — номер кластера
while $U \neq \emptyset$ **do**
    взять случайный объект $x \in U$
    if $U_\varepsilon(x) < m$ **then**
        пометить шумовой
    else
        $k++$ — создать новый кластер, $K := U_\varepsilon(x)$
        while $K \neq \emptyset$ **do**
            $x' := K.pop()$ — не помеченый и не шумовой
            if $U_\varepsilon(x') \geq m$ **then**
                $K.add(U_\varepsilon(x'))$ и пометить x' как корневой
            else
                пометить x' как граничный

K-means vs DBSCAN

DBSCAN

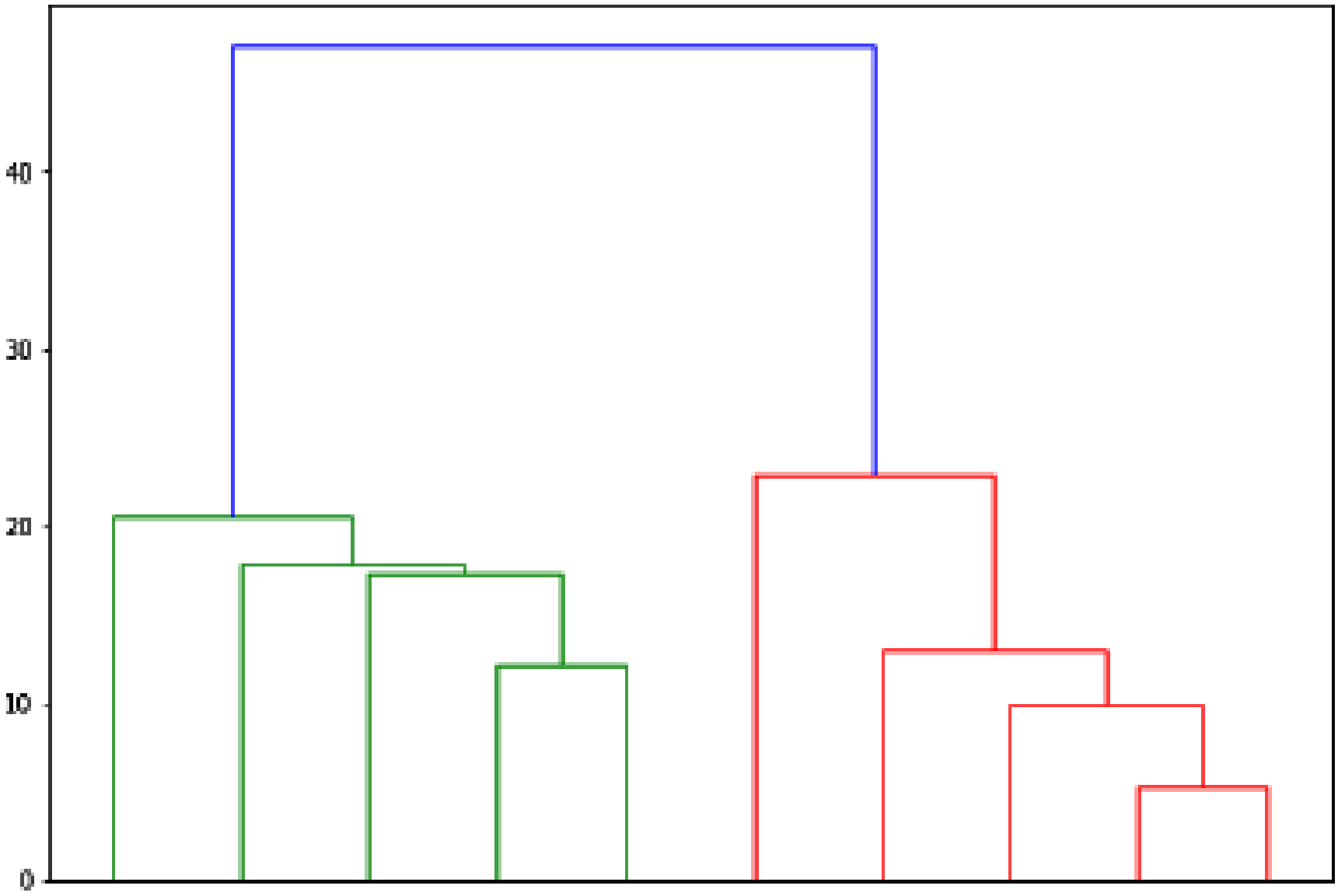
k-means

R_{UV} — мера расстояния между кластерами U и V

Algorithm 3: Иерархическая кластеризация

$C_1 := \{\{x_1\}, \ldots, \{x_\ell\}\}$
for $t = 2, \ldots, \ell$ **do**
    найти в C_{t-1} пару кластеров (U, V) с минимальным R_{UV}
    слить их в один кластер
    $W := U \cup V$
    $C_t := C_{t-1} \cup \{W\} \setminus \{U, V\}$

Дендрограмма



Итого

- Кластеризация — частный случай обучения без учителя
- Ключевая концепция — близость похожих объектов
- Изначально задача поставлена некорректно
- Каждый алгоритм подходит для определенного типа кластеровной структуры