Постановка задачи кластеризации

Дано:

X — пространство объектов;

$X^\ell = \{x_i\}_{i=1}^\ell$ — обучающая выборка;

$\rho\colon X \times X \to [0, \infty)$ — функция расстояния между объектами.

Найти:

$y_i \in Y$ — метки кластеров объектов:

— каждый кластер состоит из близких объектов;

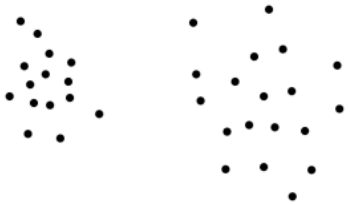
— объекты разных кластеров существенно различны.

Кластеризация — это *обучение без учителя*.

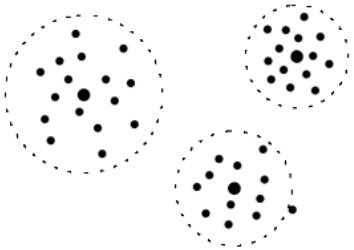
Решение задачи кластеризации принципиально неоднозначно:

- различные критерии качества кластеризации
- различные эвристические методы кластеризации
- различные варианты функции расстоянии ρ

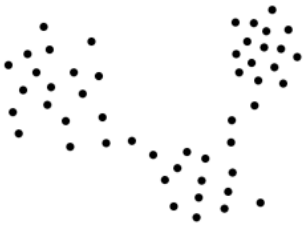
Типы кластерных структур



внутрикластерные расстояния, как правило, меньше межкластерных



кластеры с центром



кластеры могут соединяться перемычками

кластеры могут накладываться на разреженный фон из редко расположенных объектов

Типы кластерных структур

ленточные кластеры

перекрывающиеся кластеры

кластеры могут образовываться не по сходству, а по иным типам регулярностей

кластеры могут вообще отсутствовать

Метод k-средних (k-means)

Объекты x_i задаются векторами признаков $\big(f_1(x_i), \ldots, f_n(x_i)\big)$.

Вход: X^ℓ — обучающая выборка, параметр k;

Выход: центры кластеров μ_y, $y \in Y$;

1: начальное приближение центров μ_y, $y \in Y$;

2: **повторять**

3: отнести каждый x_i к ближайшему центру:
$y_i := \arg\min\limits_{y \in Y} \rho(x_i, \mu_y)$, $i = 1, \ldots, \ell$;

4: вычислить новые положения центров:

$$\mu_{yj} := \frac{\sum\limits_{i=1}^{\ell} [y_i = y] f_j(x_i)}{\sum\limits_{i=1}^{\ell} [y_i = y]}, \quad y \in Y, \quad j = 1, \ldots, n;$$

5: **пока** y_i не перестанут изменяться;

Мягкий вариант k-средних (EM-алгоритм)

Вход: X^ℓ — обучающая выборка, параметр k;

Выход: центры кластеров μ_y, $y \in Y$;

1: начальное приближение центров μ_y, $w_y := \frac{1}{|Y|}$, $y \in Y$;

2: **повторять**

3: оценить близость каждого x_i ко всем центрам:

$g_{iy} := w_y \exp\left(-\frac{1}{2}\rho^2(x_i, \mu_y)\right)$, $i = 1, \ldots, \ell$, $y \in Y$;

$g_{iy} := \dfrac{g_{iy}}{\sum\limits_{z \in Y} g_{iz}}$ — нормированные близости;

4: отнести каждый x_i к ближайшему центру:

$y_i := \arg\max\limits_{y \in Y} g_{iy}$, $i = 1, \ldots, \ell$;

5: новые положения центров и мощности кластеров:

$\mu_{yj} := \frac{1}{\ell w_y} \sum\limits_{i=1}^{\ell} g_{iy} f_j(x_i)$, $w_y := \frac{1}{\ell} \sum\limits_{i=1}^{\ell} g_{iy}$, $y \in Y$, $j = 1, \ldots, n$;

6: **пока** g_{iy} не перестанут изменяться;

Формирование начального приближения для k-средних

Вход: X^ℓ — обучающая выборка, параметры q, δ, k

Выход: $U \subset X^\ell$ — начальные приближения центров μ_y, $y \in Y$;

1: среднее расстояние до q ближайших соседей:

$R_i := \frac{1}{q} \sum_{j=1}^{q} \rho\left(x_i, x_i^{(j)}\right)$, для всех $i = 1, \ldots, \ell$,

где $x_i^{(j)}$ — j-й ближайший сосед объекта x_i;

2: отбросить шумовые объекты:

$X' := \left\{ x_i \in X^\ell \;\middle|\; R_i \leqslant \Delta \right\}$ при Δ: $|X'| = (1-\delta)\ell$;

3: выбрать пару самых удалённых объектов:

$U := \arg \max\limits_{x, x' \in X'} \rho(x, x')$;

далее последовательно присоединять к U по одному объекту, самому удалённому от уже выбранных:

4: **повторять** $k - 2$ **раз**

$U := U \cup \arg \max\limits_{x \in X'} \min\limits_{u \in U} \rho(x, u)$;

Недостатки k-means

- Чувствительность к выбору начального приближения
- Необходимость задавать k

Способы устранения этих недостатков:

- Эвристики для выбора начального приближения
- Мягкая кластеризация
- Мультистарт: несколько случайных инициализаций; выбор лучшей кластеризации по функционалу качества.
- Быстрые алгоритмы (k-means++, сэмплирование)
- Варьирование числа кластеров k в ходе итераций