

Лекция 11

Обучение без учителя

Макаренко В.А., Габдуллин Р.А.

МГУ им. М.В. Ломоносова

17 марта 2023

Дано:

- Признаки объектов без ответов

Требуется:

- Найти зависимости между объектами

Типичные задачи:

- Кластеризация объектов
- Снижение размерности данных
- Поиск аномальных объектов
- Восстановление плотности распределения

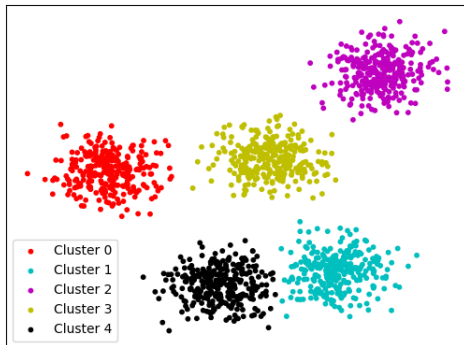


Рис.: Источник: medium.com

- Выборка: $X = \{x_i\}_{i=1}^{\ell}$, $x \in \mathbb{X}$.
- Цель: построить отображение $a : \mathbb{X} \rightarrow \{1, 2, \dots, K\}$.

Метрики качества в задаче кластеризации

Функция расстояния между объектами: $\rho : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$.

Центры кластеров: $\{c_k\}$, $1 \leq k \leq K$.

- Внутрикластерное расстояние (чем меньше, тем лучше):

$$\sum_{k=1}^K \sum_{i=1}^{\ell} [a(x_i) = k] \rho(x_i, c_k).$$

- Межкластерное расстояние (чем больше, тем лучше):

$$\sum_{i,j=1}^{\ell} [a(x_i) \neq a(x_j)] \rho(x_i, x_j)$$

- Индекс Данна (чем больше, тем лучше):

$$\frac{\min_{1 \leq k < k' \leq K} d(k, k')}{\max_{1 \leq k \leq K} d(k)},$$

где $d(k, k')$ – расстояние между кластерами k и k' , а $d(k)$ – внутрикластерное расстояние для k -го кластера.

K-средних (K-means)

Цель: определить центр $\{c_k\}_{k=1}^K$ каждого кластера и распределение объектов по кластерам.

Задача оптимизации – минимизация внутрикластерного расстояния:

$$\sum_{k=1}^K \sum_{i=1}^{\ell} [a(x_i) = k] \rho(x_i, c_k) \rightarrow \min_{c, a}.$$

Задаем начальное приближение для центров кластеров и запускаем итерационный процесс:

- 1 Относим каждый объект к ближайшему кластеру при фиксированных центрах:

$$a(x_i) = \operatorname{argmin}_{1 \leq k \leq K} \rho(x_i, c_k).$$

- 2 Обновляем центры кластеров при фиксированном распределении объектов по кластерам:

$$c_k = \operatorname{argmin}_c \sum_{i: a(x_i)=k} \rho(x_i, c).$$

К-средних (K-means)

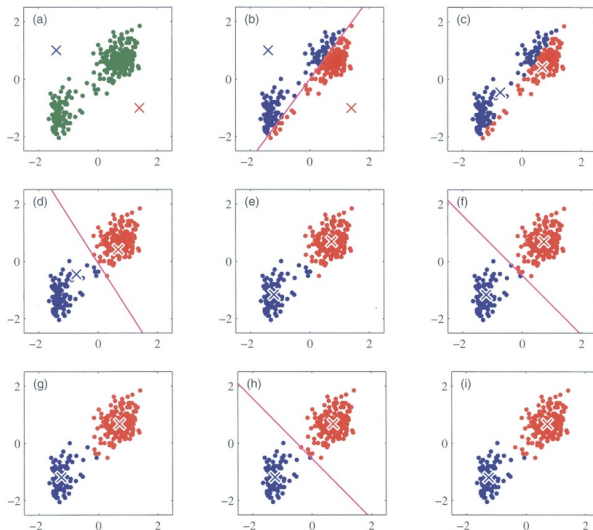


Рис.: Источник: neerc.ifmo.ru

Иерархическая кластеризация

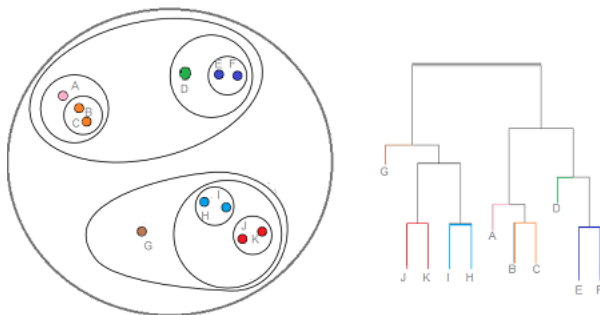


Рис.: Источник: laptrinhx.com

Виды:

- Агломеративные алгоритмы (слияние кластеров).
- Дивизионные алгоритмы (разъединение кластеров).

Агломеративная кластеризация

- Сначала каждый объект содержится в собственном кластере.
- Итеративный процесс слияния двух ближайших кластеров, пока не выполнен критерий останова.
- На каждом шаге нужно уметь вычислять расстояние между кластерами и пересчитывать расстояние между новыми кластерами.
- Расстояние между одноэлементными кластерами определяется через расстояние между объектами:
$$R(\{x\}, \{y\}) = \rho(x, y).$$
- Для вычисления расстояния $R(U, V)$ между двумя кластерами U и V используют различные функции.

Функции расстояния между кластерами

- **Метод одиночной связи** (single linkage):

$$R_{\min}(U, V) = \min_{u \in U, v \in V} \rho(u, v).$$

- **Метод полной связи** (complete linkage):

$$R_{\max}(U, V) = \max_{u \in U, v \in V} \rho(u, v).$$

- **Метод средней связи** (Unweighted Pair Group Method with Arithmetic mean):

$$R_{\text{avg}}(U, V) = \frac{1}{|U| \cdot |V|} \sum_{u \in U} \sum_{v \in V} \rho(u, v).$$

- **Центроидный метод** (Unweighted Pair Group Method with Centroid average):

$$R_c(U, V) = \rho^2 \left(\sum_{u \in U} \frac{u}{|U|}, \sum_{v \in V} \frac{v}{|V|} \right).$$

- **Метод Уорда** (Ward's method):

$$R_{\text{ward}}(U, V) = \frac{|U| \cdot |V|}{|U| + |V|} \rho^2 \left(\sum_{u \in U} \frac{u}{|U|}, \sum_{v \in V} \frac{v}{|V|} \right).$$

Формула Ланса-Уильямса

Пересчет расстояния от нового кластера $W = U \cup V$ до кластера S :

$$R(W, S) = \alpha_U \cdot R(U, S) + \alpha_V \cdot R(V, S) + \beta \cdot R(U, V) + \gamma \cdot [R(U, S) - R(V, S)].$$

- **Метод одиночной связи** (single linkage):

$$\alpha_U = \frac{1}{2}, \quad \alpha_V = \frac{1}{2}, \quad \beta = 0, \quad \gamma = -\frac{1}{2}.$$

- **Метод полной связи** (complete linkage):

$$\alpha_U = \frac{1}{2}, \quad \alpha_V = \frac{1}{2}, \quad \beta = 0, \quad \gamma = \frac{1}{2}.$$

- **Метод средней связи** (Unweighted Pair Group Method with Arithmetic mean):

$$\alpha_U = \frac{|U|}{|V|}, \quad \alpha_V = \frac{|V|}{|W|}, \quad \beta = 0, \quad \gamma = 0.$$

- **Центроидный метод** (Unweighted Pair Group Method with Centroid average):

$$\alpha_U = \frac{|U|}{|V|}, \quad \alpha_V = \frac{|V|}{|W|}, \quad \beta = -\alpha_U \cdot \alpha_V, \quad \gamma = 0.$$

- **Метод Уорда** (Ward's method):

$$\alpha_U = \frac{|S| + |U|}{|S| + |W|}, \quad \alpha_V = \frac{|S| + |V|}{|S| + |W|}, \quad \beta = -\frac{|S|}{|S| + |W|}, \quad \gamma = 0.$$

Свойство монотонности расстояний

- R_t – расстояние между кластерами, выбранными для объединения на шаге t .
- Кластеризация называется монотонной, если:

$$R_2 \leq R_3 \leq \dots \leq R_m.$$

Теорема (Миллиган, 1979)

Если для коэффициентов в формуле Ланса-Уильямса выполняются следующие три условия, то кластеризация является монотонной:

- 1 $\alpha_U \geq 0, \quad \alpha_V \geq 0.$
- 2 $\alpha_U + \alpha_V + \beta \geq 1.$
- 3 $\min\{\alpha_U, \alpha_V\} + \gamma \geq 0.$

Из перечисленных выше расстояний теореме удовлетворяют все, кроме центроидного.

Дендрограмма

- По оси абсцисс – объекты.
- По оси ординат – расстояния между объединяемыми кластерами.
- Если кластеризация является монотонной, то дендрограмма не содержит самопересечений и будет наглядной.

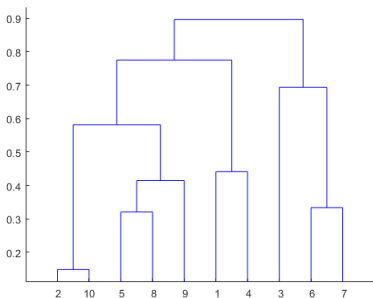


Рис.: Источник: [mathworks.com](https://www.mathworks.com)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

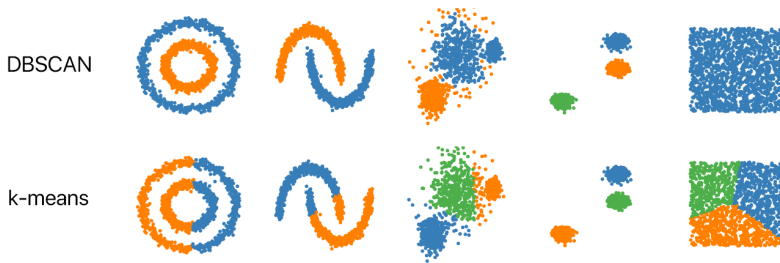


Рис.: Источник: swiftpack.co

Входные данные и параметры алгоритма:

- Обучающая выборка (набор точек) x_1, x_2, \dots, x_ℓ .
- Радиус окружности ε .
- Минимальное число точек m в окрестности.

Типы точек:

- 1 **Основные точка.** В круге радиуса ε с центром в такой точке содержится как минимум m точек выборки.
- 2 **Прямо достижимая из основной точки b .** Такая точка находится на расстоянии, не большем ε от основной точки b .
- 3 **Достижимая из основной точки b .** Точка a называется достижимой из основной точки b , если существует путь $p_1 = b, p_2, p_3, \dots, p_n = a$ такой, что p_{i+1} прямо достижима из p_i (точки p_1, \dots, p_{n-1} являются основными).
- 4 **Шумовая точка.** Точка, которая не является основной и не является достижимой ни для любой основной точки (в ε -окрестности этой точки содержится меньше m точек и нет ни одной основной).

DBSCAN. Типы точек

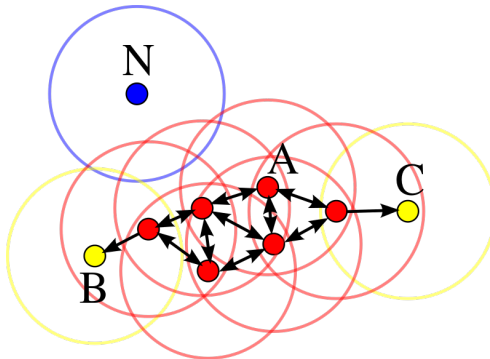


Рис.: Источник: Википедия

- $m = 4$.
- Красные точки – основные, желтые – достижимые, синие – шумовые.

DBSCAN. Кластеризация

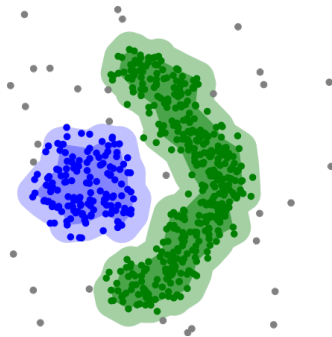


Рис.: Источник: [Википедия](#)

- Основная точка вместе со всеми достижимыми из нее (как основными, так и нет) формирует кластер.
- Каждый кластер содержит хотя бы одну основную точку.
- Неосновные точки формируют «границу» кластера, так как не могут быть использованы для достижения других точек.
- Все точки попарно связаны по плотности (точки p и q связаны по плотности, если существует точка o , из которой они достижимы).

Автокодировщик

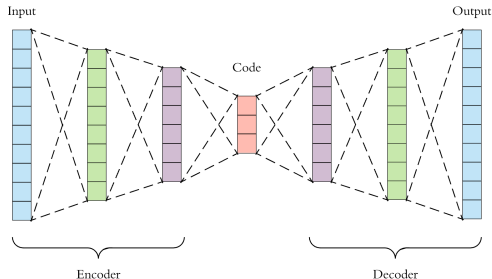


Рис.: Источник: medium.com

- Нелинейное снижение размерности.
- Модель состоит из двух частей: кодировщик g и декодировщик f .
- Модель учит тождественную функцию $x \approx f(g(x))$, минимизируя функционал $L(x, f(g(x)))$.