Занятие 11. Кластеризация

Колмагоров Евгений
ml.hse.dpo@yandex.ru

16 декабря 2024

План лекции

1. Введение в кластеризацию
2. Алгоритм K-means
3. Иерархическая кластеризация
4. DBSCAN
5. Метрики кластеризации

Unsupervised learning

Не всегда есть возможность или потребность собирать множество меток Y, чтобы обучить алгоритм машинного, зачастую есть необходимость за счёт машинного обучения получить такое множество Y.

Задача кластеризации

Кластеризация – задача обучения без учителя, цель которой за счёт внутренней информации объектов выборки X найти “похожие” объекты и отнести их к одному классу. (Зачастую кластеризацию в англоязычной литературе называют unsupervised classification, т.е. классификацией без учителя)

Отличие от классификации

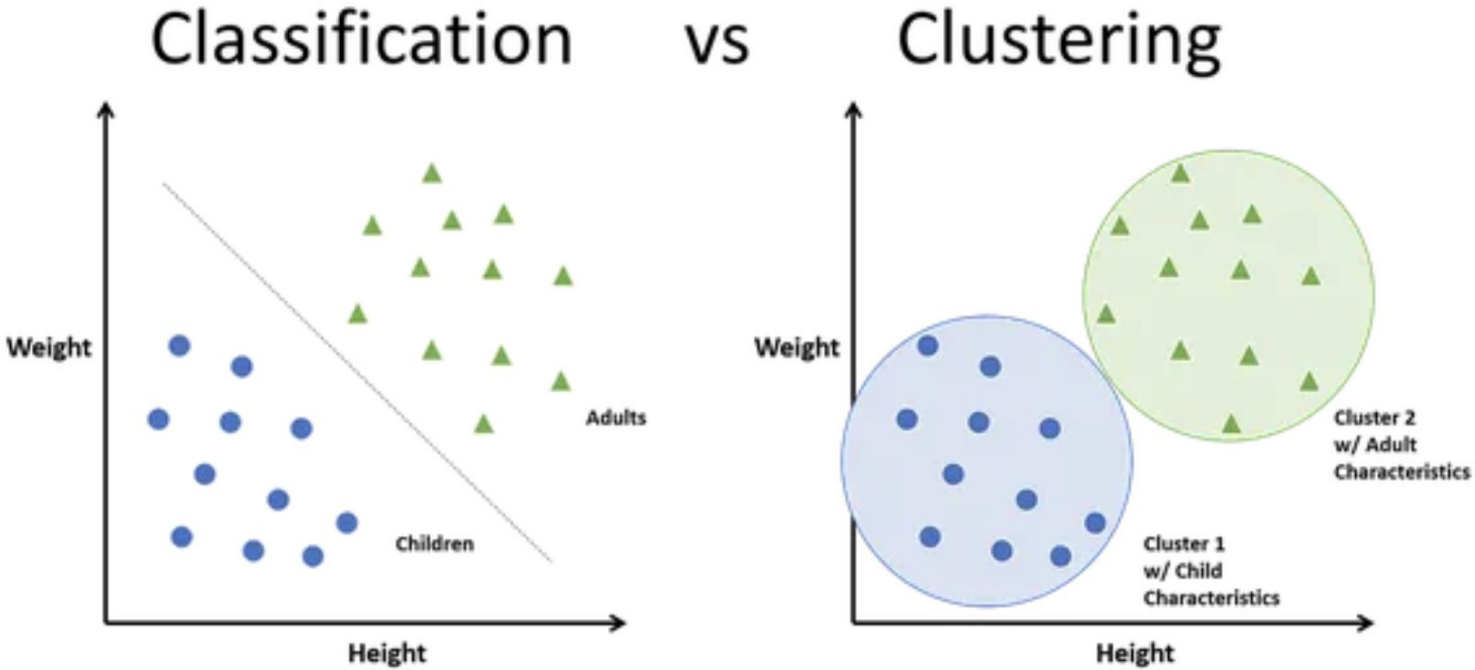
Принципиальное отличие между классификацией состоит в том, что в задаче классификации нужно на основе обучающей выборки (X_{train}, Y_{train}) научиться **восстанавливать зависимость**:

- $a(X_{train}, Y_{train}): X_{test} \rightarrow Y_{test}$

А в кластеризации через имеющееся описание объектов X **открыть их класс**:

- $a: X \rightarrow Y$

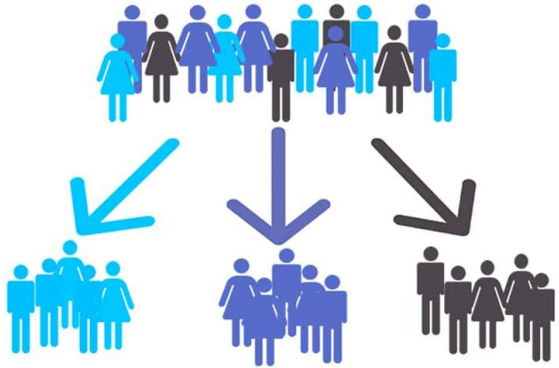
Отличие от классификации



Примеры задач

Задачи, которые могут быть решены с помощью кластеризации:

- Поиск аномалий
- Анализ социальных сетей
- Группировка документов
- Обработка геоданных
- Выделение пользовательских сегментов
- И много чего другого



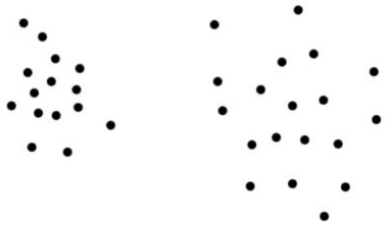
Неоднозначность решения задачи кластеризации

Результат работы алгоритма кластеризации неоднозначен из-за:

- Существует много критериев качества кластеризации
- Существует много эвристических критериев качества кластеризации
- Число кластеров $|Y|$ заранее, как правило, неизвестно
- Результат кластеризации существенно зависит от метрики d

Виды распределения данных

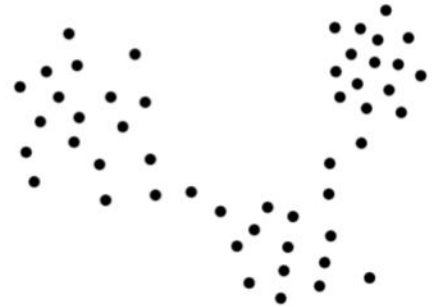
Также при решении задачи дополнительную сложность вносит разнообразие форм распределений данных



Внутри кластерные расстояния больше, чем межкластерные

Ленточные кластеры

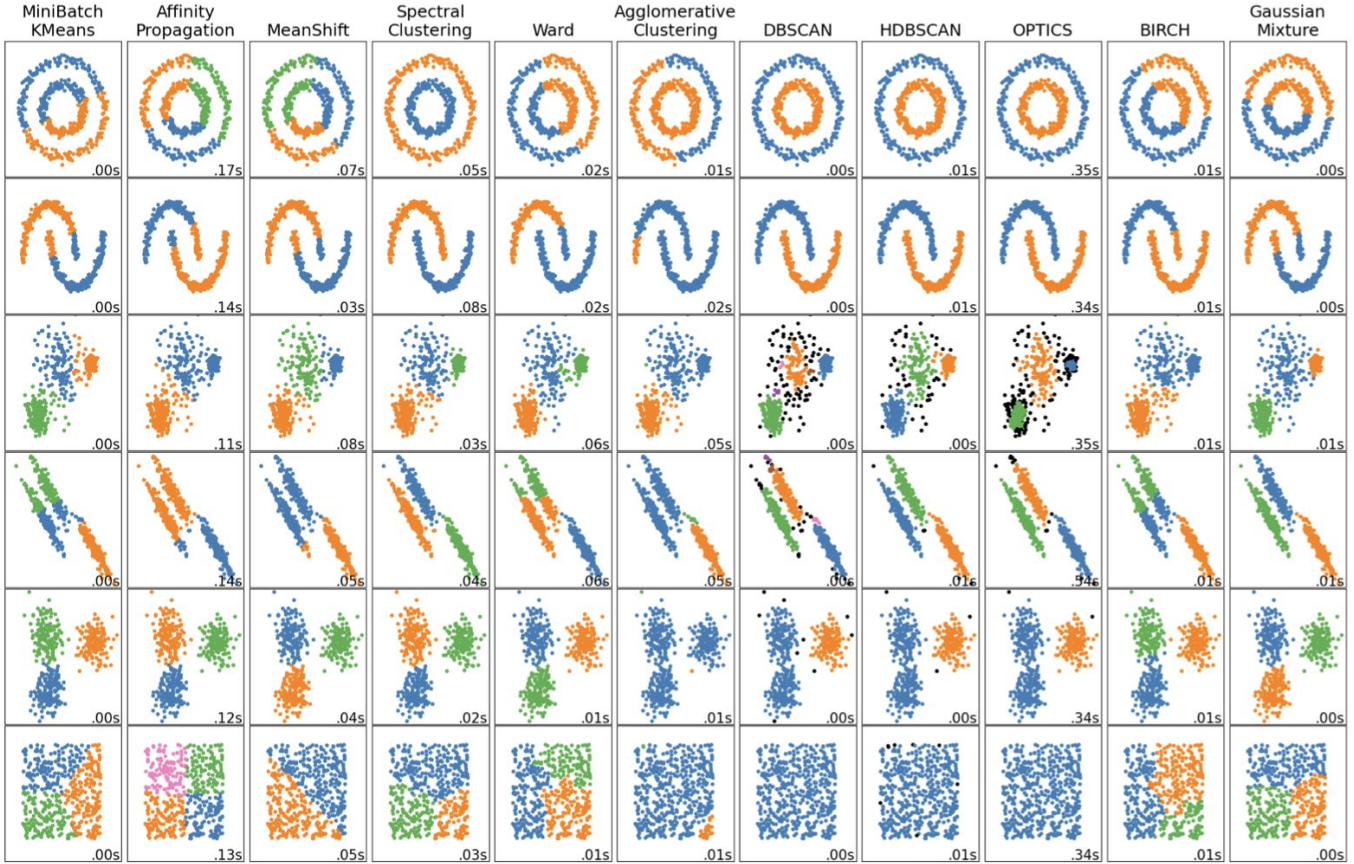
Виды распределения данных



Кластеры могут соединяться перемычками

Кластеры могут перекрываться

Какая из кластеризаций лучше?



Существующие алгоритмы

При построении алгоритма кластеризации можно руководствоваться различными эвристиками, в соответствии с которыми можно разделить выборку.

Рассмотрим наиболее известные для этого алгоритмы:

- Минимальное остовное дерево
- K-Means
- Иерархическая кластеризация
- DBSCAN

Кластеризация на графах

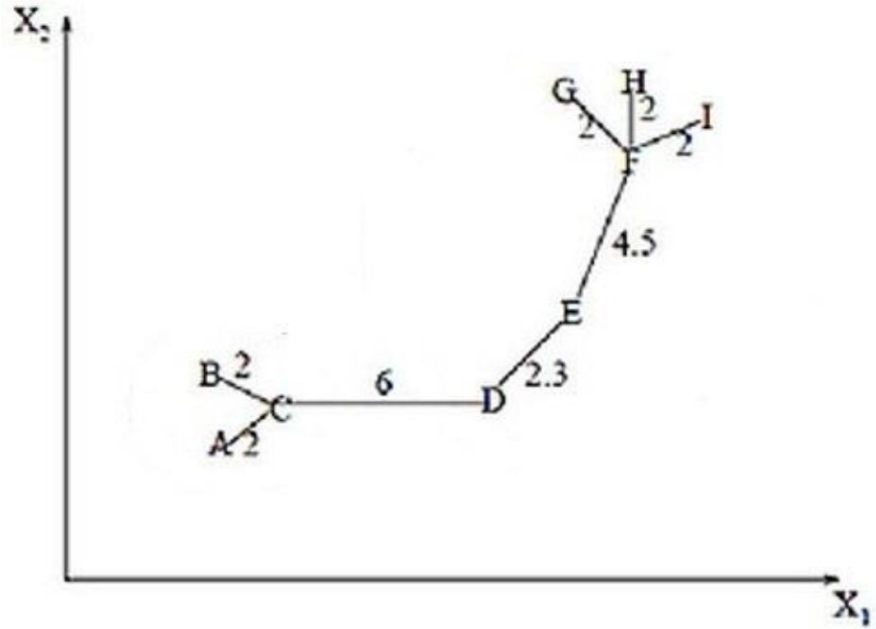
Вся выборка представляется как полный граф, где в вершинах стоят объекты из X, а на рёбрах указано расстояние между этими объектами.

Алгоритм состоит из следующих шагов:

- Построить минимальное остовное дерево по одному из этих алгоритмов
- На основе гиперпараметра K – число кластеров – удалить K-1 самых тяжёлых ребра, в результате чего получим K компонент связности

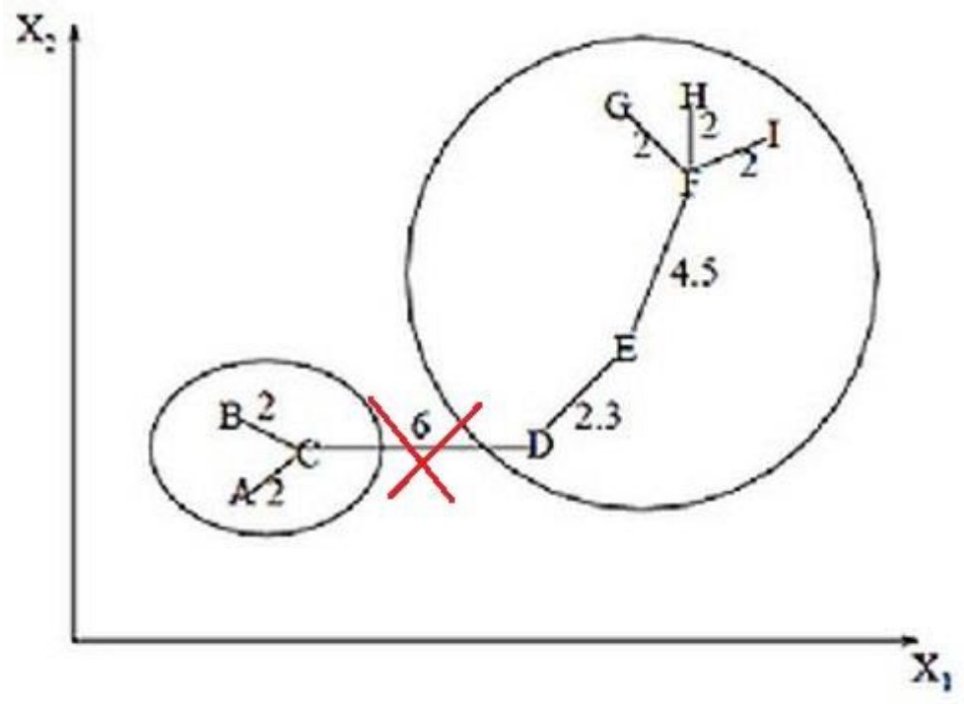
Минимальное остовное дерево

Шаг-1: Из полного графа строим минимальное остовное дерево по алгоритму Прима или Краскала



Разбиение дерева на связные компоненты

Шаг-2: Удаляем K-1 самых тяжёлых ребра. В итоге получим K компонент связности. Объекты, попадающие в одну компоненту связности, будем относить к одному кластеру

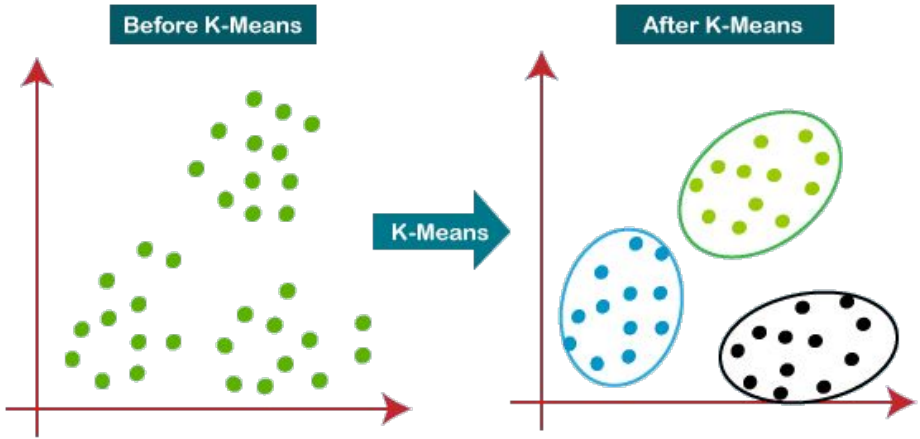


Метод K-средних (K-means)

Одним из самых популярных методов кластеризации является метод K-средних.

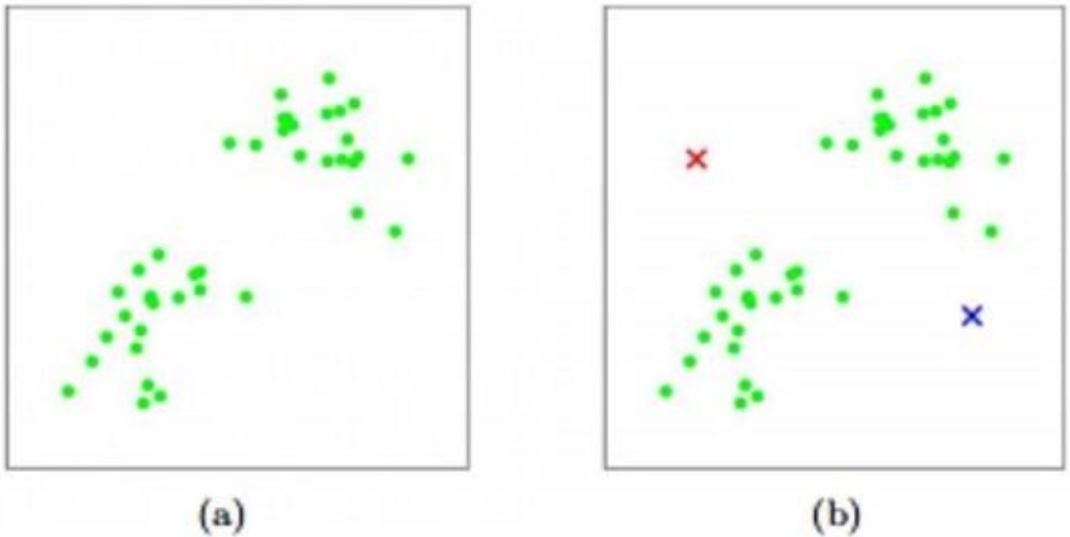
Алгоритм состоит из двух повторяющихся шагов:

- Отнести каждый объект к ближайшему к нему центру кластера
- Пересчитать центры полученных кластеров



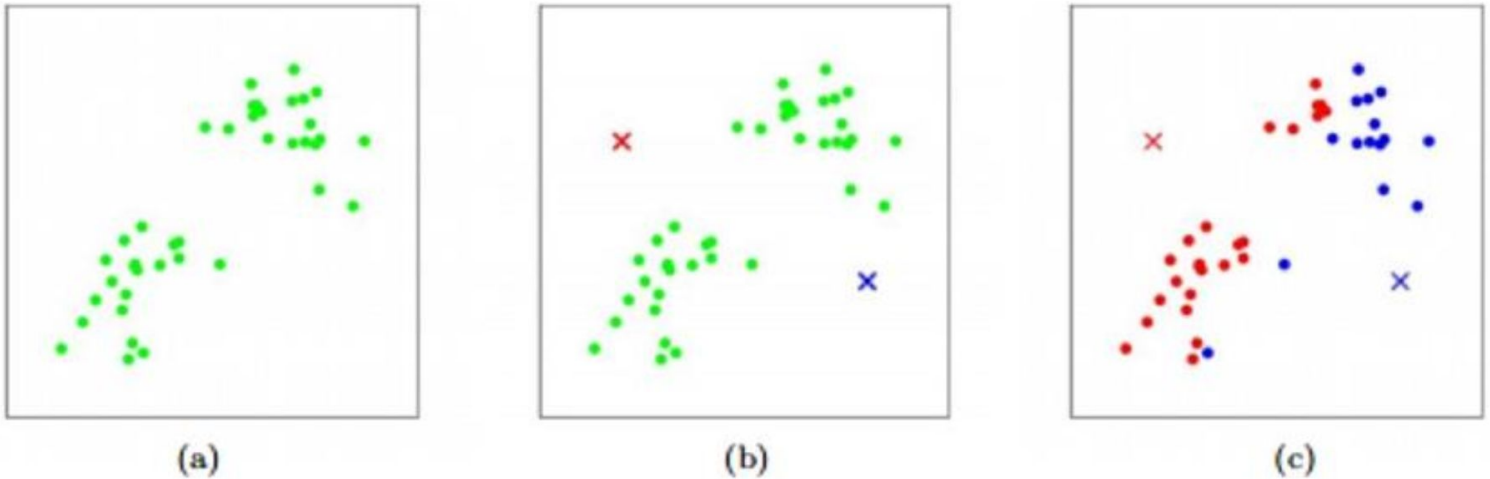
Метод K-средних (K-means)

- ***Дано***: выборка $x_1, \ldots, x_N$
- ***Гиперпараметры***: K – число кластеров
- ***Инициализация***: случайно выбрать центры кластеров $\boldsymbol{c_1, c_2, \ldots, c_K}$



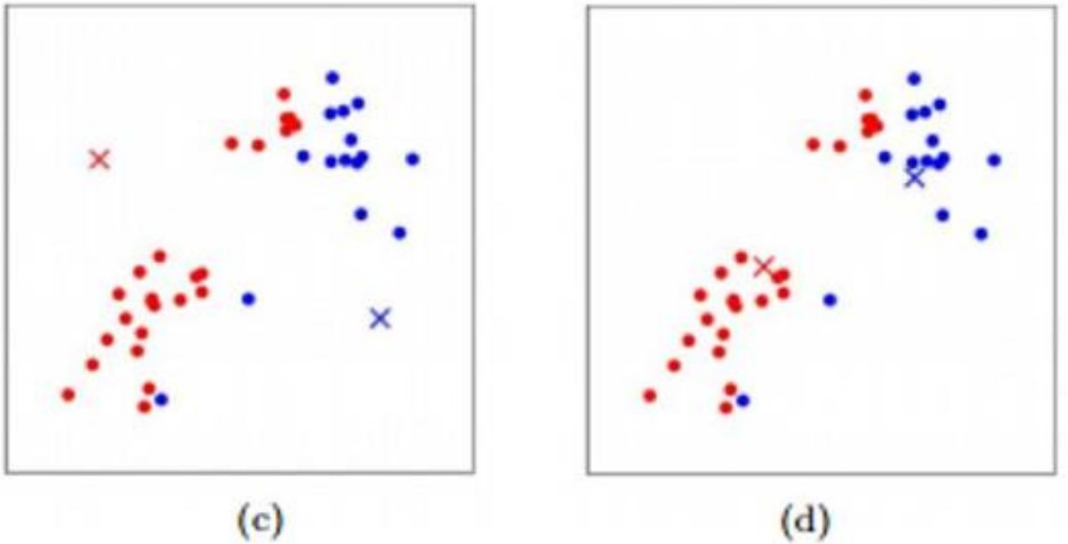
Метод K-средних (K-means)

- ***Дано***: выборка $x_1, \ldots, x_N$
- ***Гиперпараметры***: K – число кластеров
- ***Шаг-1:*** каждый объект отнести к ближайшему к нему центру кластера



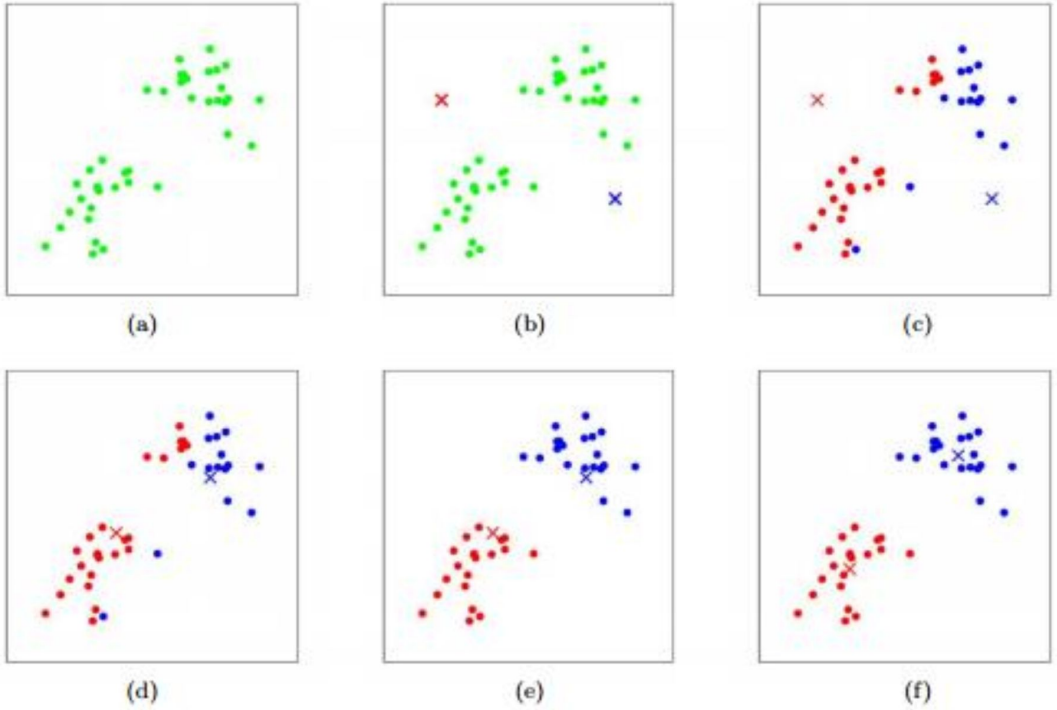
Метод K-средних (K-means)

- ***Дано***: выборка $x_1, \ldots, x_N$
- ***Гиперпараметры***: K – число кластеров
- ***Шаг-2:*** пересчитать центры полученных кластеров



Метод K-средних (K-means)

- ***Дано***: выборка $x_1, \ldots, x_N$
- ***Гиперпараметры***: K – число кластеров
- ***Остановка:*** повторять шаг-1 и шаг-2 до сходимости



Оптимизируемый функционал

Метод K-средних с евклидовым расстояние производит оптимизацию следующего функционала:

$$Q = \frac{1}{N \cdot K} \sum_{k=1}^{K} \sum_{i=1}^{N} (\mu_k - x_i)^2 \mathbb{I}[a(x_i) = k] \rightarrow min_{\mu_1,\ldots,\mu_K}$$

Оптимизируемый функционал

На шаге пересчёта центра k-го кластера оптимизируется внутренняя сумма:

$$\sum_{i=1}^{N}(\mu_k - x_i)^2 \mathbb{I}[a(x_i) = k] \to min_{\mu_k}$$

Если взять производную данной суммы по μ_k и приравнять её к 0, то получим, что

$$\mu_k = \frac{1}{I_k} \sum_{i=1}^{N} x_i \cdot \mathbb{I}[a(x_i) = k]$$

Оптимизируемый функционал

На шаге пересчёта объектов кластера оптимизируется весь функционал Q

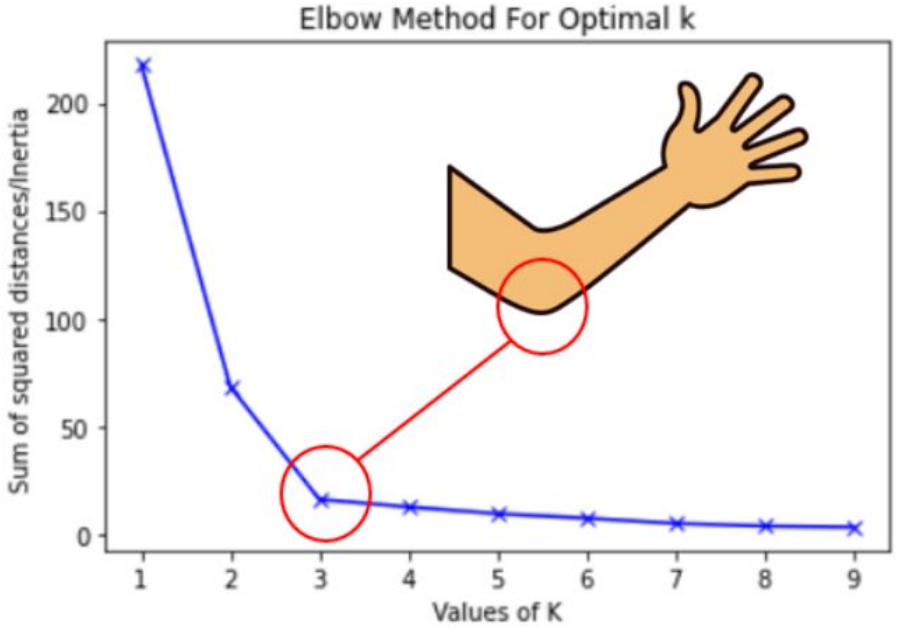
$$Q = \frac{1}{N \cdot K} \sum_{k=1}^{K} \sum_{i=1}^{N} (\mu_k - x_i)^2 \mathbb{I}[a(x_i) = k]$$

Таким образом, чтобы индикаторная функция I была равна 1 только на ближайшем к данному объекту центре

Правило локтя для выбора K (elbow method)

Для выбора оптимального значения K использую эвристическое правило под названием правило “локтя”.

Выбираем такое значение K, когда происходит значительное уменьшение внутрикластерного расстояния

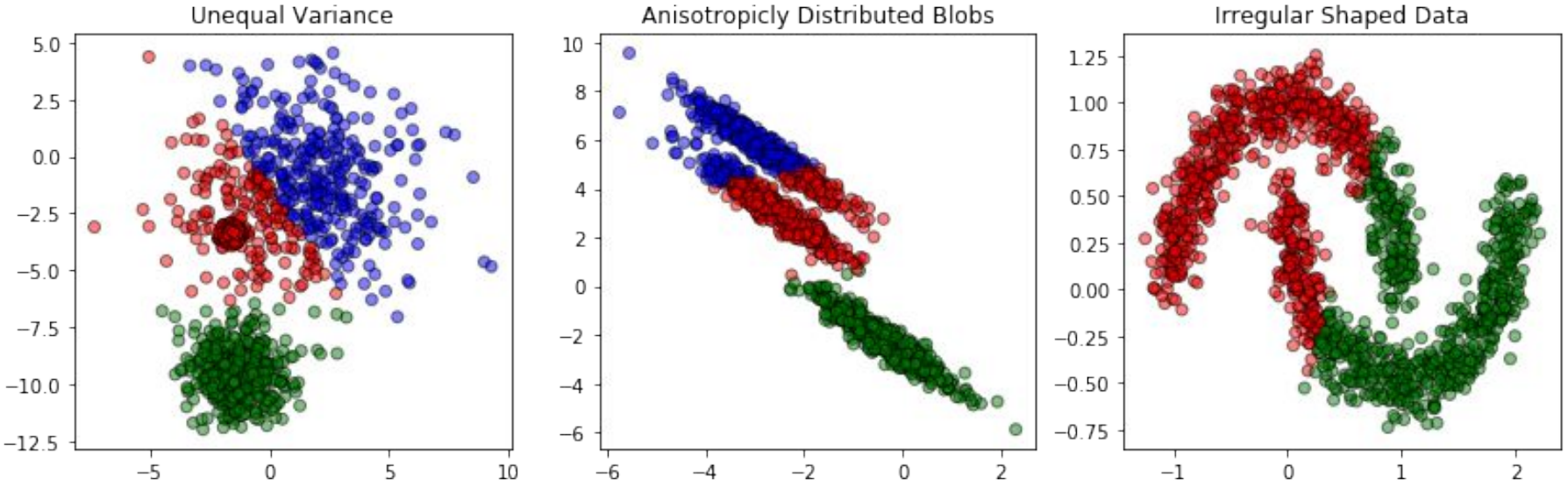


Line plot between K and inertia

K-Means в сегментации изображений

Зона применимости алгоритма K-средних

Алгоритм K-средних хорошо себя показывает там, где в распределении данных есть ярко выраженные центры кластеров, и все объекты сконцентрированы вокруг них, если это не так алгоритм работает плохо



Иерархическая кластеризация

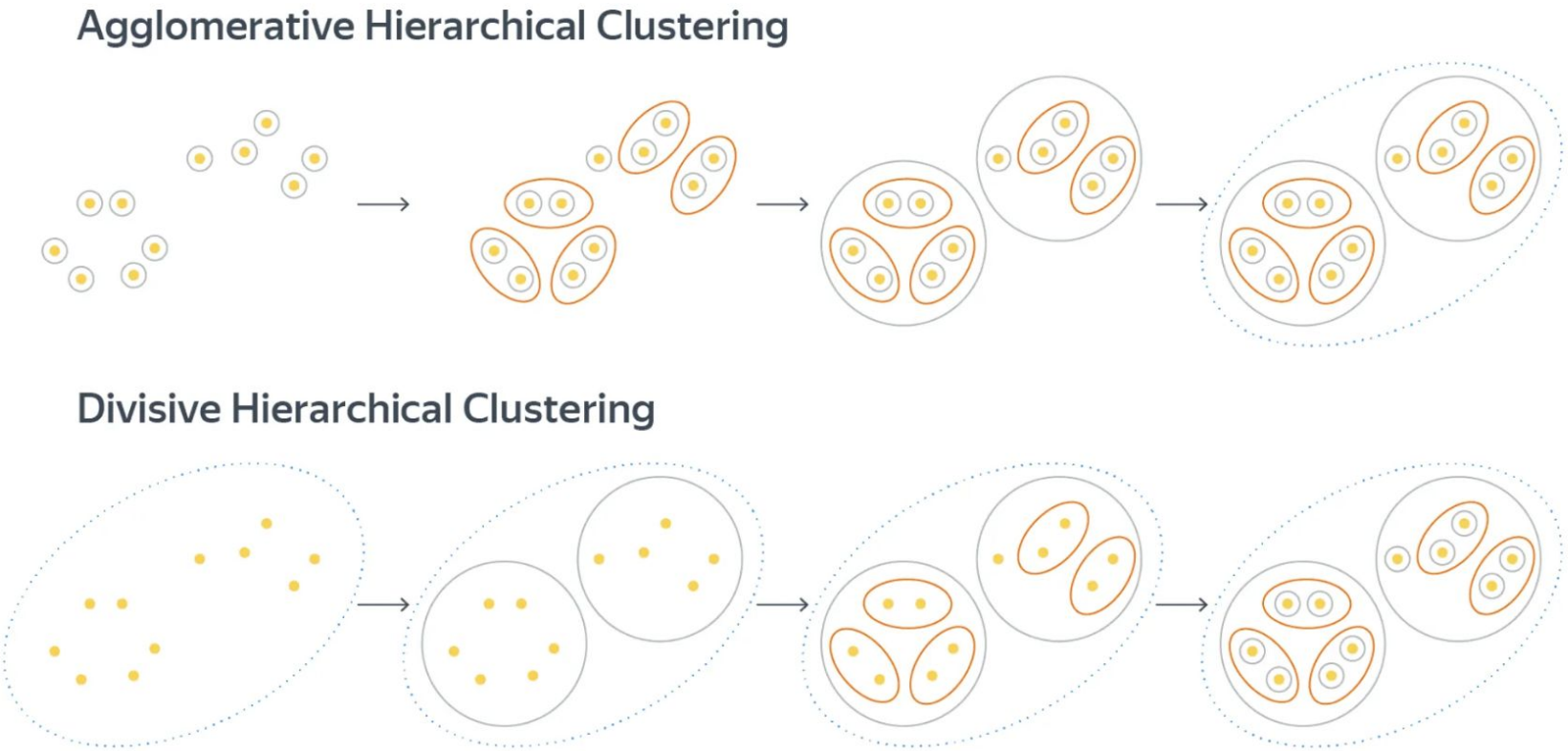
Другой класс алгоритмов кластеризации использует идею иерархическом разбиения всех данных X на подмножества вложенных кластеров $\{X_1, X_2, \ldots, X_k\}$:

$$X_k \subset X_{k-1} \subset .. \subset X_1 \subset X$$

Существует два вида иерархических моделей с разными видами направленности построения подмножеств

- **Агломеративные** – строят каждый последующий кластер через объединение предыдущих
- **Дивизионные** – строят каждый последующий кластер разбивая предыдущие на два

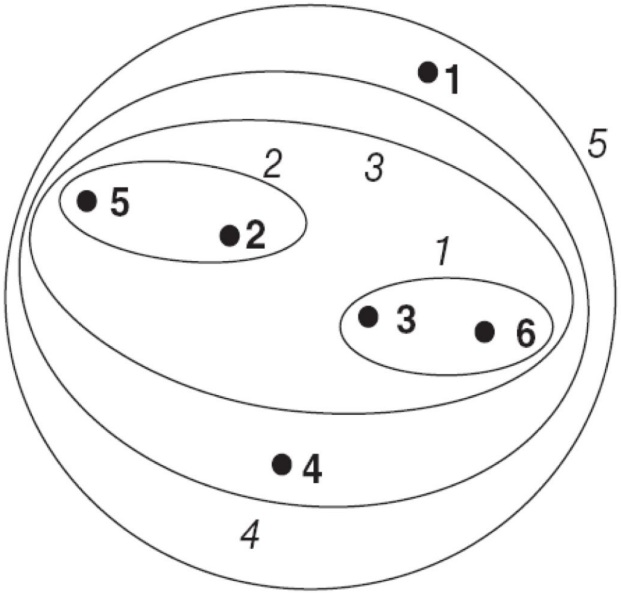
Виды иерархических кластеризаций



Агломеративная кластеризация

Идея агломеративной кластеризации состоит из двух шагов:

- Изначально создать столько кластеров сколько есть объектов в выборке
- Повторять итеративно слияние двух ближайших кластеров, пока не выполнится критерий останова



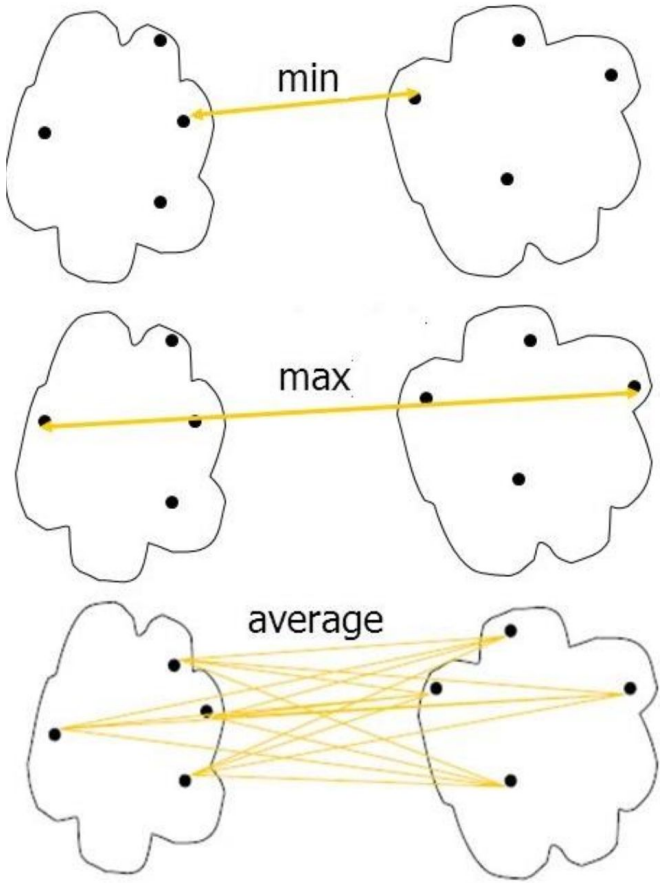
Как вычислять расстояние между кластерами

Существует 3 различных способа считать расстояние между кластерами U и V:

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

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

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



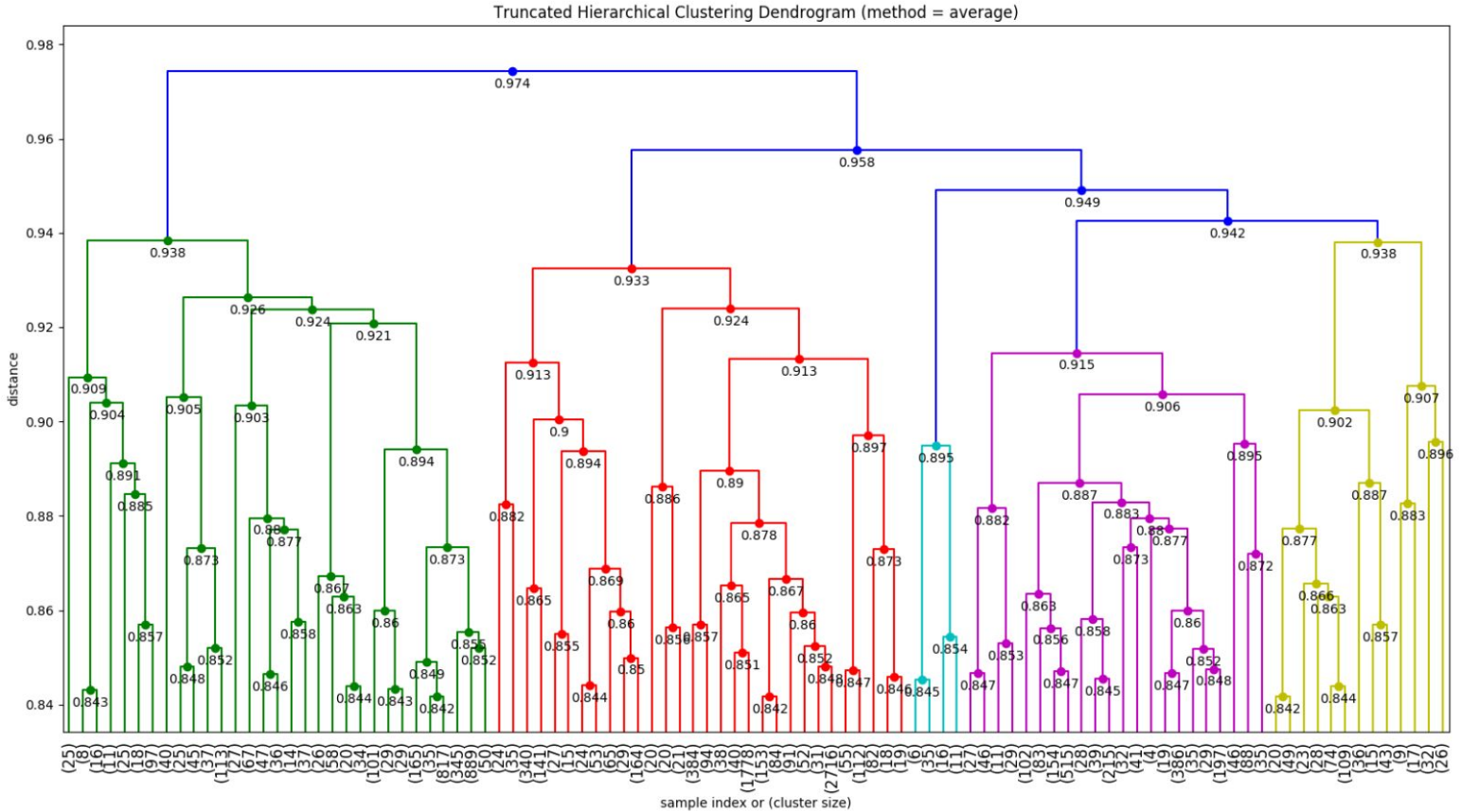
Формулы пересчёта расстояний Ланса-Вильямса

Так как последующие кластера строятся исходя из объединения предыдущих, то справедлива следующая формула вычисления расстояний:

$$\begin{aligned} R(U \cup V, S) = \alpha_U \cdot R(U, S) + \\ + \alpha_V \cdot R(V, S) + \\ + \beta \cdot R(U, V) + \\ + \gamma \cdot \left| R(U, S) - R(V, S) \right|, \end{aligned}$$

где α_U, α_V, β, γ — числовые параметры.

Дендрограмма

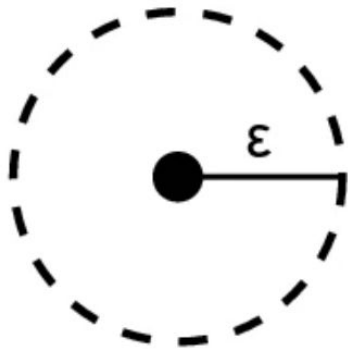


Алгоритм DBSCAN

Алгоритм DBSCAN (Density-based spatial clustering of applications with noise) развивает идею кластеризации через выделение связных компонент.

Он использует идею поиска кластеров на основе выделения участков в данных с заданной плотностью.

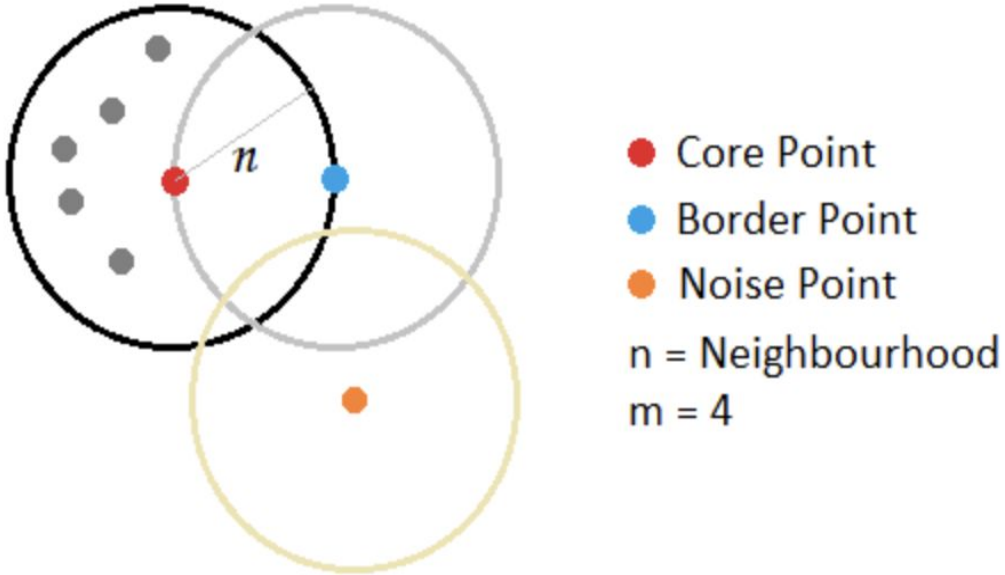
- Плотность объекта x_i определяется как количество других точек выборки в шаре $B(x_i, \varepsilon)$ радиуса ε



Виды точек

Существует три типа точек:

- ***Основные (core-points)*** – точки, в окрестности которых более чем N_0 объектов выборки, где N_0 гиперпараметр алгоритма
- ***Граничные (border points)*** – точки, в окрестности которых есть основные, но их число меньше чем N_0
- ***Шумовые (noise points)*** – точки, в окрестности которых нет основных точек и содержится менее N_0 объектов



Шаги алгоритма

Дано: выборка $x_1, \ldots, x_N$

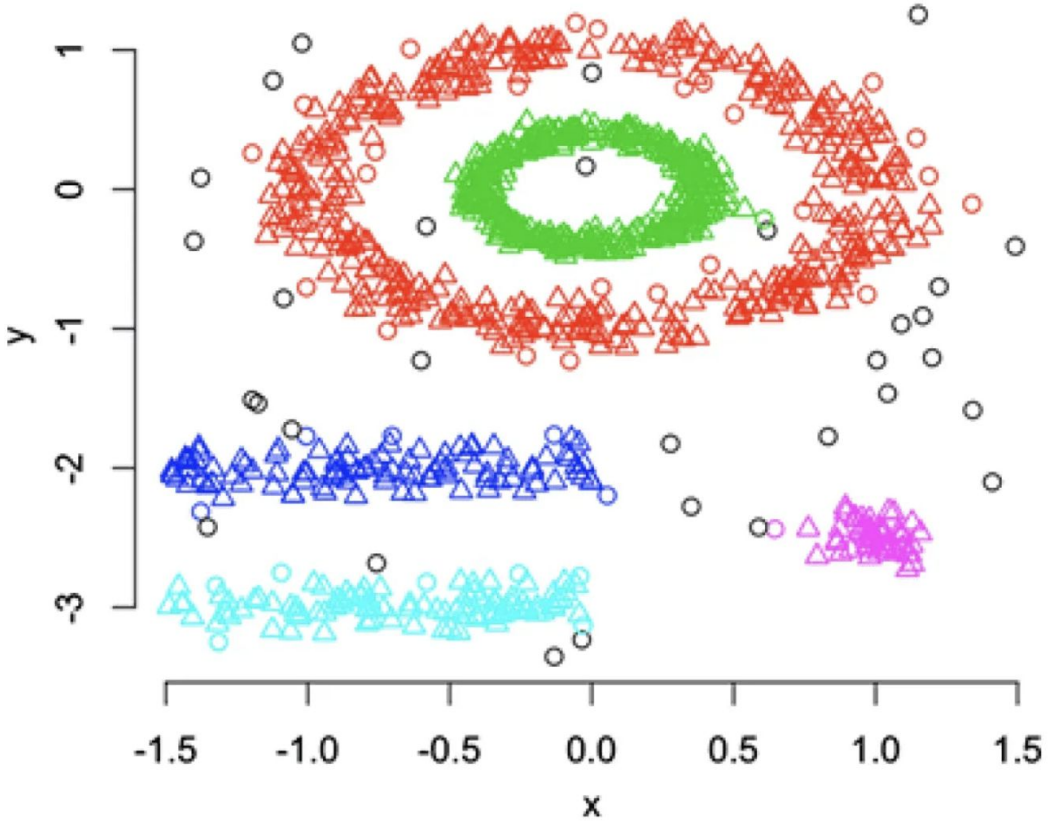
Гиперпараметры: ε – окрестность рассматриваемого шара, N_0 – минимальное количество точек, чтобы считать точку основной

Шаги:

1. Удалить все шумовые точки
2. Найти все основные точки, и если имеется пересечение между двумя окрестностями основных точек, то соединить их ребром
3. В полученном графе выделить компоненты связности – они и будут кластерами
4. Граничные точки относятся к тому кластеру, куда попала ближайшая основная к ним точка

Геометрия выделяемых кластеров

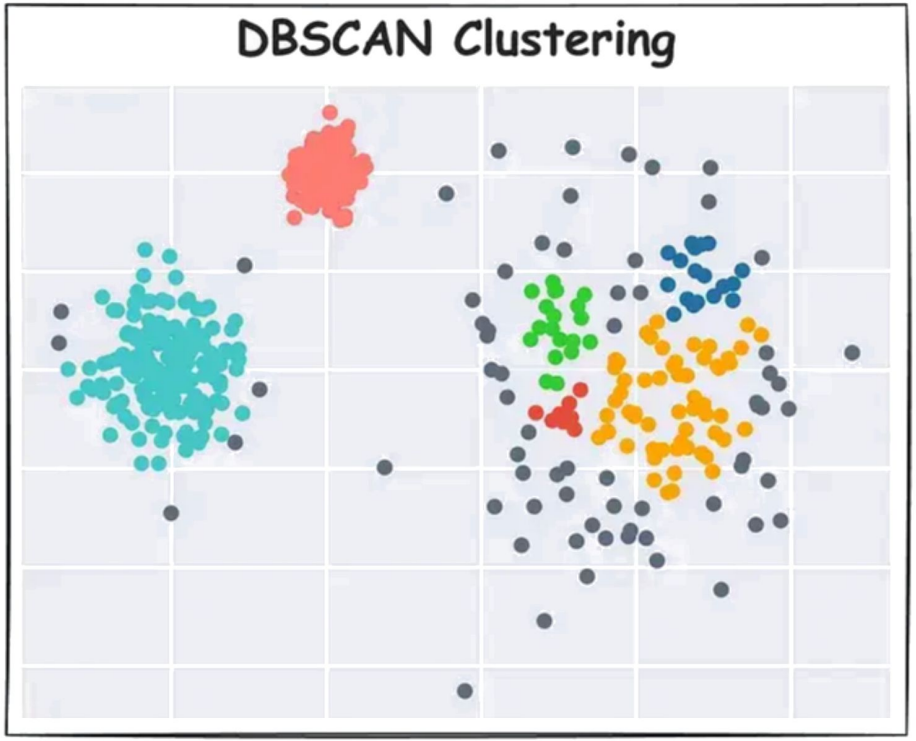
Так как алгоритм основан на поиске плотных участков, то форма выделяемых компонент может быть произвольной



Недостатки DBSCAN

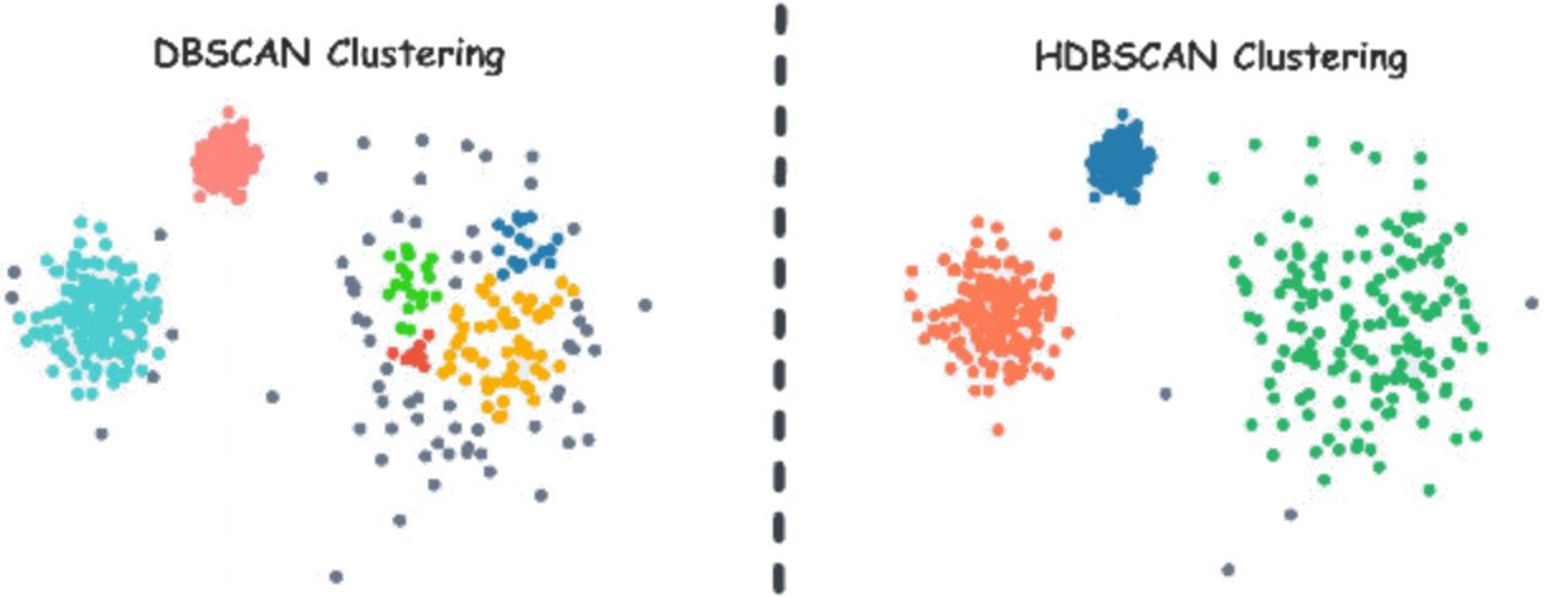
Несмотря на все свои плюсы алгоритм

- Чувствителен к настройке своих гиперпараметров
- Плохо справляется с данными, у которых кластера переменной плотности



Hierarchical DBSCAN

Улучшенной версией DBSCAN является вариация алгоритма, которая использует иерархический подход в объединении кластеров найденных алгоритмом DBSCAN



Проблема выбора алгоритма кластеризации

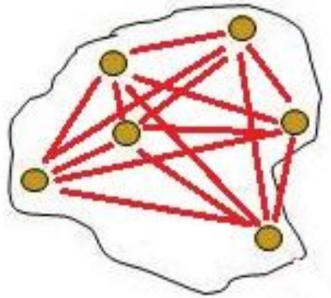
Мы рассмотрели несколько примеров алгоритмов решающих одну и ту же задачу, но как среди описанного множества выбрать самый лучший алгоритм.

- И если в обучении с учителем было валидационное множество, по которому какой из алгоритмов работает лучше по выбранной метрике, то в задаче кластеризации не всё так очевидно.
- Существует множество различных метрик, которые оценивают тот или иной аспект кластеризации

Внутреннее межкластерное расстояние

Одним из самых простых способов посмотреть на качество кластеризации можно замерив расстояние между объектами одного кластера:

$$F_0 = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}\rho(x_i, x_j)\mathbb{I}[a(x_i) = a(x_j)]}{\sum_{i=1}^{n}\sum_{j=i}^{n}\mathbb{I}[a(x_i) = a(x_j)]}$$



- Смысл метрики в том, что чем более кучные будут получаться кластера тем меньшее значение будет она принимать

Среднее межкластерное расстояние

Аналогично среднему межкластероному расстоянию вводится среднее межкластерное:

$$F_1 = \frac{\sum_{i=1}^{n} \sum_{j=i}^{n} \rho(x_i, x_j) \mathbb{I}[a(x_i) \neq a(x_j)]}{\sum_{i=1}^{n} \sum_{j=i}^{n} \mathbb{I}[a(x_i) \neq a(x_j)]}$$

- Чем более будут отдалены кластеры друг от друга, тем больше будет данная метрика

Индекс Данна (Dunn index)

Хорошая кластеризация та, у которой минимально внутрикластерное расстояние и максимально межкластерное.

Для оценки этих двух составляющих применяется индекс Дана:

$$DI = \frac{\min\limits_{1 \le i < j \le K} F_1(i,j)}{\max\limits_{1 \le i \le K} F_0(i)}$$

- $F_1(i, j)$ – расстояние между кластерами i и j
- $F_0(i)$ – внутрикластерное расстояние i-го кластера

Индекс Данна (Dunn index)

$$DI = \frac{\min\limits_{1 \le i < j \le K} F_1(i,j)}{\max\limits_{1 \le i \le K} F_0(i)}$$

Rand Index (RI)

Предположим, что известны истинные метки объектов Y.

Тогда можно вычислить рассчитать следующую метрику:

$$RI = \frac{a+b}{C_N^2} = \frac{2(a+b)}{N(N-1)}$$

- a – число пар объектов, попавших в один кластер
- b – число пар объектов с разными метками и попавшими в разные кластера
- C_N^2 – число всевозможных пар

Rand Index (RI)

$$RI = \frac{a+b}{C_N^2} = \frac{2(a+b)}{N(N-1)}$$

Rand Index показывает какая доля объектов, для которых исходное и полученное разбиение согласованы

Кластеризации с одинаковым значением RI

Гомогенность (Homogenity)

Предположим, что известны истинные метки объектов Y.

Введём следующие обозначения

- n – общее число объектов в выборке
- n_k – число объектов в кластере под номером k
- m_c – число объектов в классе номер c
- n_{ck} – количество объектов из класса c в кластере k

Гомогенность (Homogenity)

Введём следующие энтропии для мультиномиальных распределений m_c / n, n_k/n, n_{kc}/n

$$H_{class} = -\sum_{c=1}^{C} \frac{m_c}{n} \log \frac{m_c}{n}$$

$$H_{clust} = -\sum_{k=1}^{K} \frac{n_k}{n} \log \frac{n_k}{n}$$

$$H_{class|clust} = -\sum_{c=1}^{C} \sum_{k=1}^{K} \frac{n_{ck}}{n} \log \frac{n_{ck}}{n_k}$$

Гомогенность (Homogenity)

Определим гомогенность кластеризации, как

$$Homogeneity = 1 - \frac{H_{class|clust}}{H_{class}}$$

- Худший случай, когда отношение энтропий оказалось 1, то есть энтропия от того, что выборка была была поделена на кластеры никак не изменилась относительно исходной энтропии
- Наилучший случай, когда каждый кластер содержит элементы только одного класса

Тривиальный случай получить наилучшую гомогенность – выделить каждый объект в отдельный кластер

Полнота

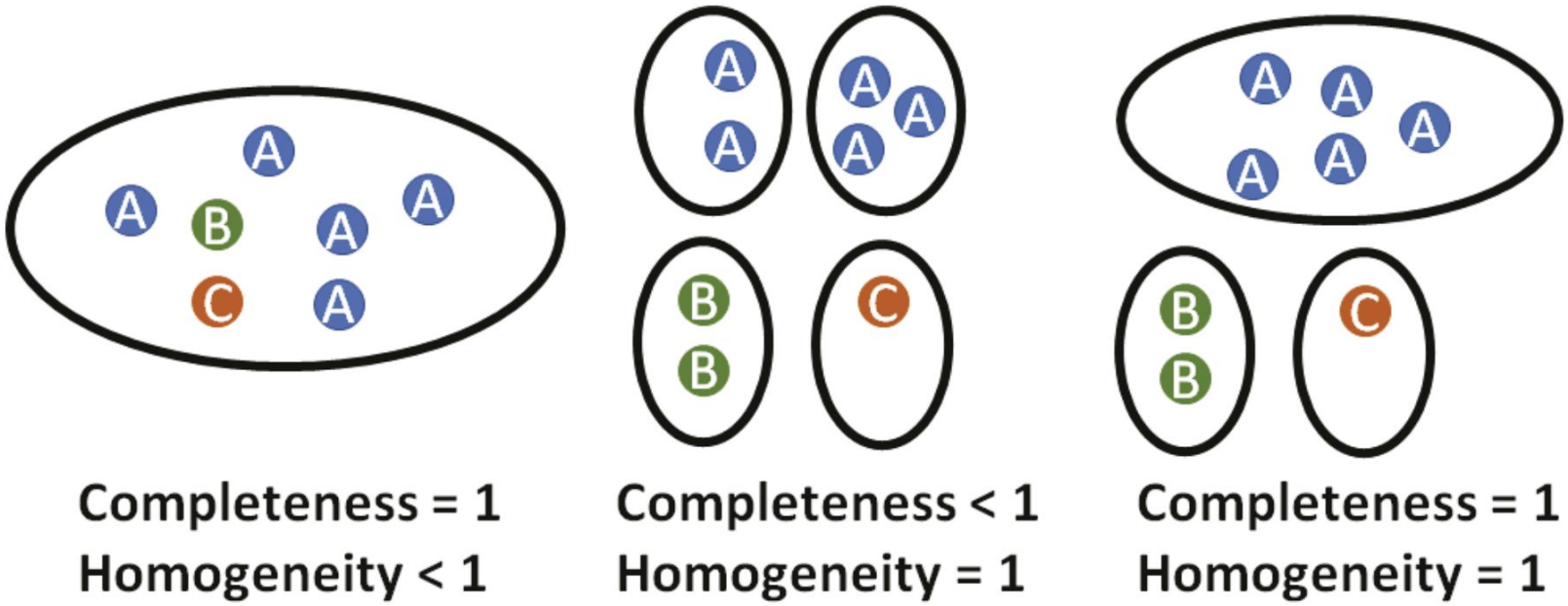
Полнота напоминает по формуле гомогенность:

$$Completeness = 1 - \frac{H_{clust|class}}{H_{clust}}$$

- Худший случай – объекты одного из одного класса разбиты по разным кластерам
- Лучший случай – объекта одного класса лежат в одном кластере

Тривиальный случай получить наилучшую полноту – положить все объекты в один кластер

Связь гомогенности и полноты



Связь гомогенности и полноты

Гомогенность и полнота кластеризации – это в некотором смысле аналоги точности и полноты классификации. Аналог F-меры для задачи кластеризации тоже есть, он называется V-мерой и связан с гомогенностью и полнотой той же формулой, что и F-мера с точностью и полнотой:

$$V_1 = \frac{2 \cdot Homogeneity \cdot Completeness}{Homogeneity + Completeness}$$

Коэффициент силуэта (Silhouette coefficient)

Введём коэффициент силуэта для объекта x_i как

$$S(x_i) = \frac{B(x_i) - A(x_i)}{\max(B(x_i), A(x_i))}$$

Где $B(x_i)$ – среднее расстояние до объектов ближайшего кластера, $A(x_i)$ – среднее расстояние до объектов своего кластера

Коэффициент силуэта для всей выборки вычисляется как среднее значение по всем объектам

$$S = \frac{1}{N} \sum_{i=1}^{N} S(x_i)$$

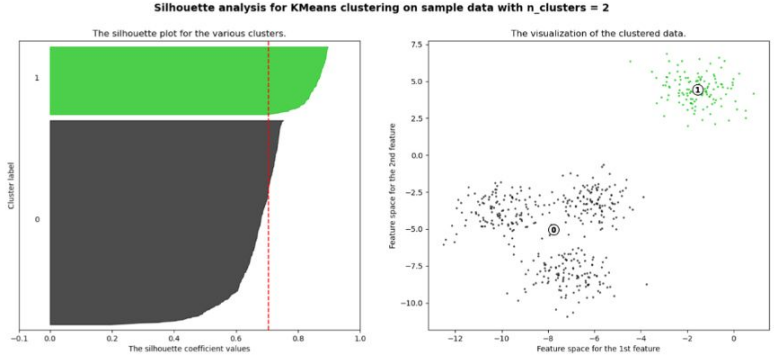
Коэффициент силуэта (Silhouette coefficient)

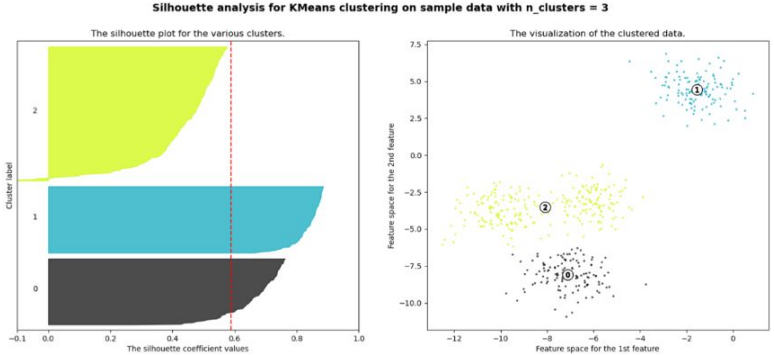
Коэффициент силуэта S принимает значения из отрезка [-1, 1]

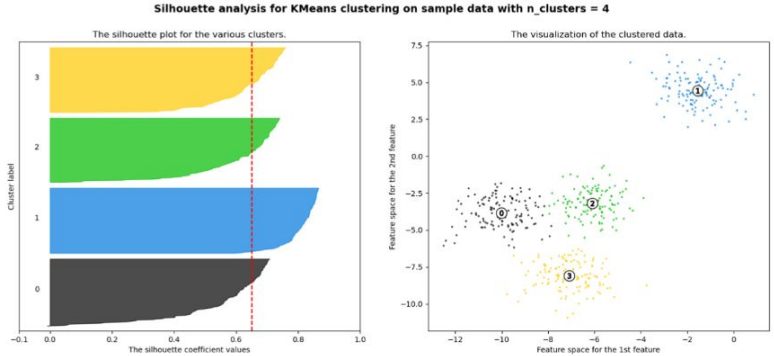
- S ≈ -1 – плохие разрозненные кластеризации
- S ≈ 0 - кластеры накладываются друг на друга
- S ≈ 1 - чётко выраженные кластеры

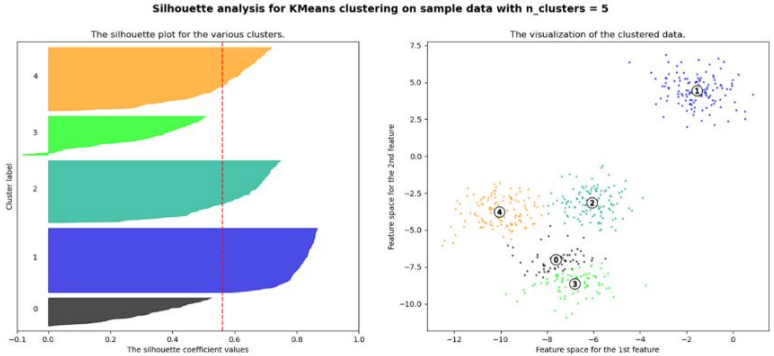
Силуэт зависит от формы кластеров и достигает наибольших значений на более выпуклых кластерах

Коэффициент силуэта для поиска оптимального K









Выводы

- Если есть возможность собрать разметку, то лучше решать задачу классификации
- Задача кластеризации некорректно поставленная задача, так как на одной и той же выборке данных могут устраивать различные варианты кластеризации
- Необходимо выбирать метод кластеризации в зависимости от распределения данных
- Есть неоднозначность в выборе метрик. Так, например, если известны истинные метки, то лучше пользоваться V-мерой, если известно количество кластеров то коэффициентом Дана, а если совсем ничего неизвестно, то коэффициентом силуэта