Кластеризация

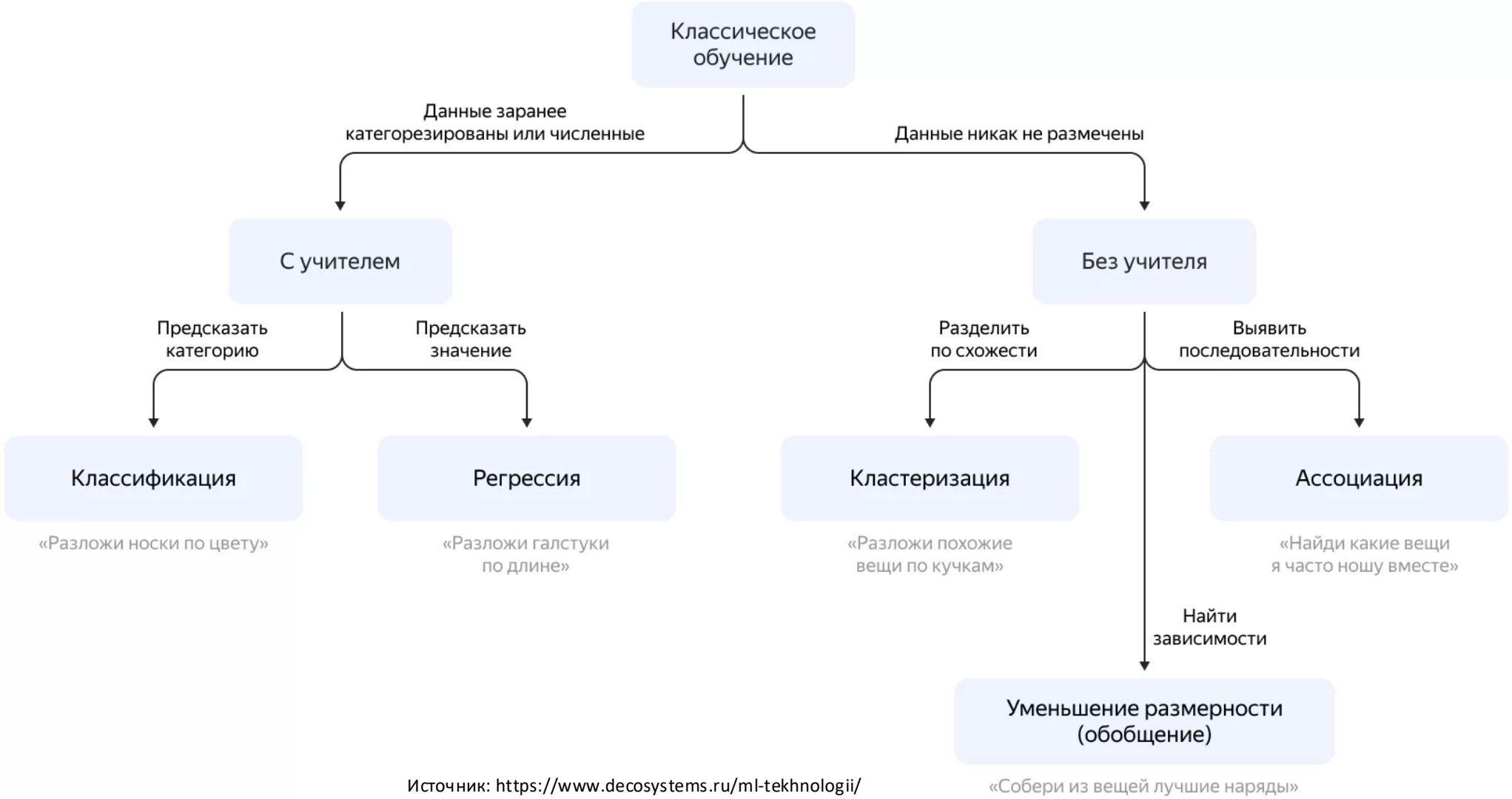
Паточенко Евгений

НИУ ВШЭ

План занятия

- Типы задач: обучение без учителя
- Кластеризация
- Метрики качества кластеризации
- Методы кластеризации

Типы задач в машинном обучении



Источник: https://www.decosystems.ru/ml-tekhnologii/

Обучение без учителя

- Обучение без учителя (unsupervised learning) — класс методов машинного обучения, в которых отсутствует целевая переменная и требуется восстановить структуру в данных

Обучение без учителя

- Обучение без учителя (unsupervised learning) — класс методов машинного обучения, в которых отсутствует целевая переменная и требуется восстановить структуру в данных
- Данные на входе не размечены

Обучение без учителя

- Обучение без учителя (unsupervised learning) — класс методов машинного обучения, в которых отсутствует целевая переменная и требуется восстановить структуру в данных
- Данные на входе не размечены
- Поскольку правильные ответы отсутствуют, возникают проблемы с измеримостью качества

Кластеризация

Кластеризация — задача обучения без учителя, цель которой за счет внутренней информации объектов выборки X найти «похожие» объекты и отнести их к одному классу

В англоязычной литературе для этого метода можно встретить название

unsupervised classification, т.е. классификация без учителя

Кластеризация

Даны объекты $x_1, \dots, x_l, x_j \in X$

Требуется выявить в данных K кластеров — таких областей, что объекты внутри одного кластера будут максимально похожи друг на друга, а объекты из разных кластеров — максимально друг на друга не похожи

Формализация задачи: необходимо построить алгоритм a: $X \to \{1, \dots, K\}$, сопоставляющий каждому объекту x номер кластера

Кластеризация vs классификация

Классификация

На основе обучающей выборки научиться восстанавливать зависимость

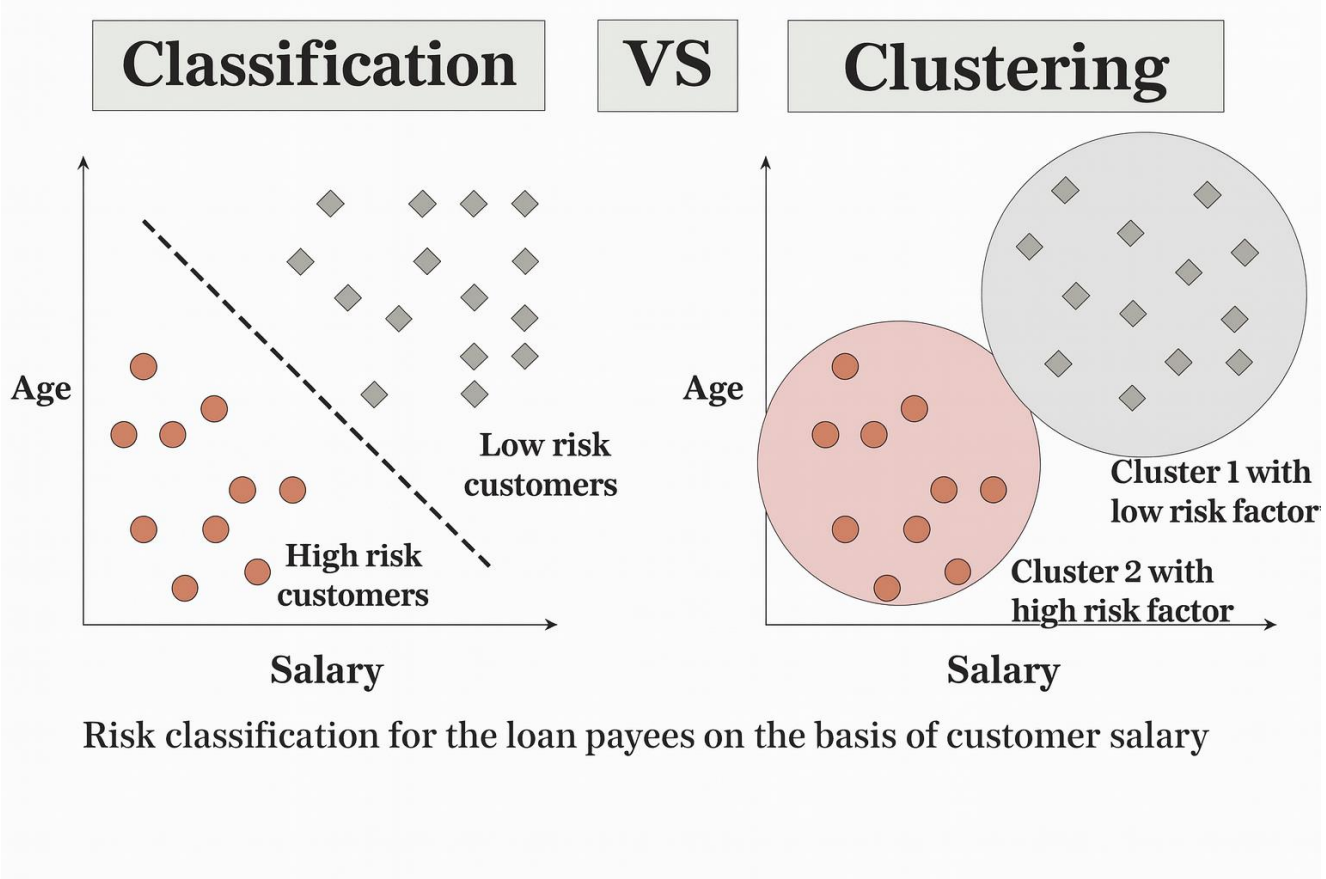
$a(X_{train}, Y_{train}): X_{test} \to X_{test}$

Кластеризация

Через имеющееся описание объектов X открыть их класс

$a: X \to Y$

Кластеризация vs классификация



Risk classification for the loan payees on the basis of customer salary

Источник: https://techdifferences.com/difference-between-classification-and-clustering.html

Кластеризация

Примеры задач:

- Поиск аномалий
- Выделение пользовательских сегментов
- Группировка документов / текстов
- Обработка геоданных

Метрики качества кластеризации

Внутренние метрики

Оценивают качество кластеризации, основываясь только на наборе данных:

— среднее внутрикластерное расстояние

— среднее межкластерное расстояние

— индекс Данна

— силуэт

Внешние метрики

Используют информацию об истинных метках объектов:

— RI, ARI

— гомогенность

— полнота

— V-мера

Метрики качества кластеризации

Среднее внутрикластерное расстояние

Один из самых простых способов посмотреть на качество кластеризации — замерить расстояние между объектами одного кластера:

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

Сумма расстояний между точками из одного и того же кластера делится на количество пар точек, принадлежащих к одному кластеру. Чем кучнее кластеры, тем меньшее значение будет принимать функция

Метрики качества кластеризации

Среднее межкластерное расстояние

Аналогично предыдущей метрике, но с одним изменением

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

Эту функцию максимизируем: чем более отдалены кластеры друг от друга, тем больше метрика

Метрики качества кластеризации

Индекс Данна (Dunn Index)

Стремимся одновременно минимизировать внутрикластерное расстояние и максимизировать межкластерное:

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

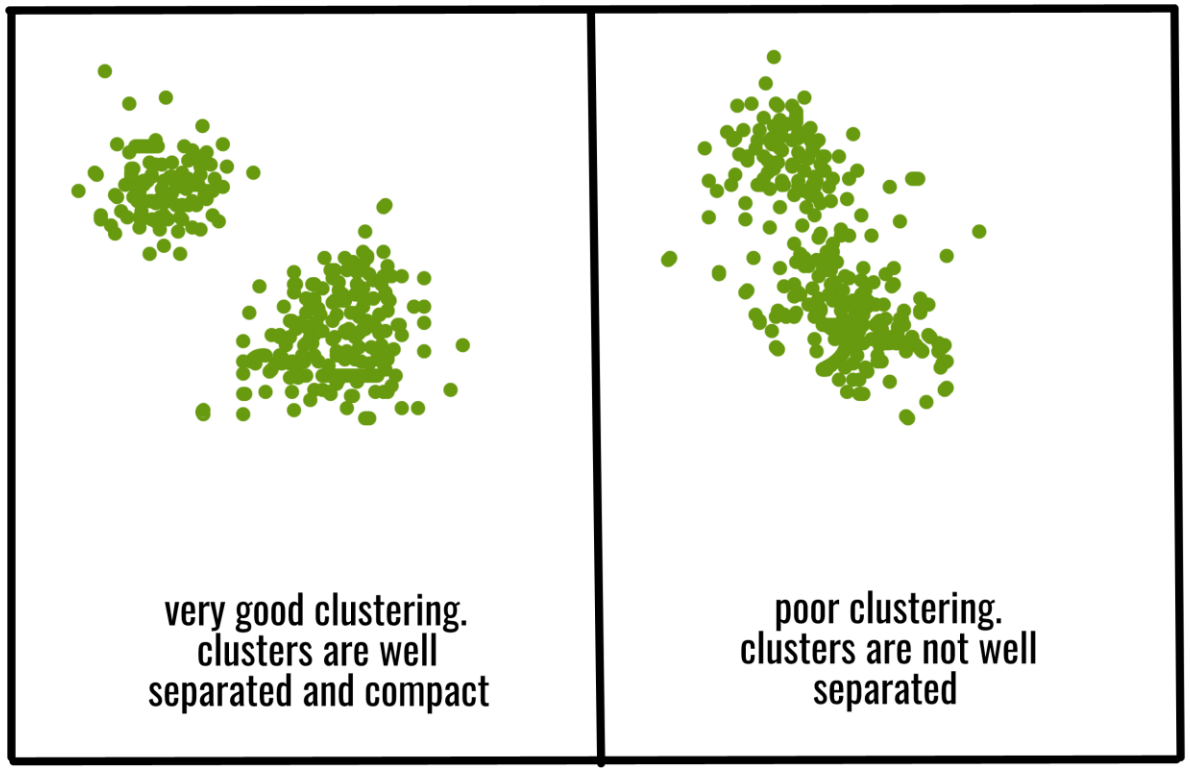
где

$F_1(i,j)$ — расстояние между кластерами i и j

$F_0(i)$ — внутрикластерное расстояние i-го кластера

Метрики качества кластеризации

Индекс Данна (Dunn Index)



Источник: https://progler.ru/blog/indeks-danna-i-indeks-bd-indeksy-klasternoy-validnosti-komplekt-1

Метрики качества кластеризации

Индекс Рэнда (Rand Index)

Предполагается, что известны истинные метки объектов Y.
RI — доля объектов, для которых исходное и полученное разбиения согласованы.
Выражает похожесть двух различных разбиений выборки

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

где

- 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

Метрики качества кластеризации

Индекс Рэнда (Rand Index)

Кластеризации с одинаковым значением RI

Источник: https://en.wikipedia.org/wiki/Rand_index

Метрики качества кластеризации

Скорректированный индекс Рэнда (Adjusted Rand Index)

RI нормируется и принимает значения из отрезка [-1; 1] независимо от числа объектов N и числа кластеров

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

- $ARI = 1$ — разбиения совпадают
- $ARI > 0$ — разбиения похожи
- $ARI \approx 0$ — случайные разбиения
- $ARI < 0$ — разбиения непохожи

Метрики качества кластеризации

Гомогенность (однородность)

Предполагается, что известны истинные метки объектов Y

$$h = 1 - \frac{H(class|clust)}{H(class)}$$

где

- $H(class|clust)$ — условная энтропия классов с учетом назначений кластера
- $H(class)$ — энтропия классов

Метрики качества кластеризации

Гомогенность (однородность)

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

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

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

где:

- 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

Метрики качества кластеризации

Гомогенность (однородность)

$$h = 1 - \frac{H(class|clust)}{H(class)}$$

Худший случай — когда отношение энтропий оказалось равным единице (то есть гомогенность равна нулю): это значит, что энтропия от того, что выборка была поделена на кластеры никак не изменилась относительно исходной энтропии

В лучшем случае каждый кластер содержит элементы только одного класса (то есть гомогенность равна единице)

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

Метрики качества кластеризации

Полнота

Предполагается, что известны истинные метки объектов Y

$$c = 1 - \frac{H(clust|class)}{H(clust)}$$

где

- $H(clust|class)$ — условная энтропия кластеров с учетом меток классов
- $H(clust)$ — энтропия кластеров

Метрики качества кластеризации

Полнота

$$c = 1 - \frac{H(clust|class)}{H(clust)}$$

Худший случай — когда объекты из одного класса разбиты по разным кластерам

В лучшем случае объекты одного класса лежат в одном кластере

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

Метрики качества кластеризации

Гомогенность и полнота

Гомогенность и полнота принимают значения из отрезка $[0; 1]$, но метрики не нормализованы и зависят от числа кластеров!

При большом числе кластеров и малом числе объектов лучше использовать ARI

При числе кластеров меньше 10 и числе объектов больше 1000 проблема менее выражена, ее можно игнорировать и использовать метрики

Метрики качества кластеризации

V-мера

Среднее гармоническое гомогенности и полноты:

$$v = \frac{2hc}{h + c}$$

где

- h — гомогенность
- c — полнота

Гомогенность и полнота — аналоги точности и полноты в классификации. V-мера — аналог F-меры

Метрики качества кластеризации

Коэффициент силуэта (Silhouette coefficient)

Не требует знания истинных меток

$$s = \frac{b - a}{\max(a, b)}$$

где

- a — среднее расстояние от объекта до всех объектов из того же кластера
- c — среднее расстояние от объекта до объектов из ближайшего (не содержащего объект) кластера

Метрики качества кластеризации

Коэффициент силуэта (Silhouette coefficient)

Показывает, насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров.

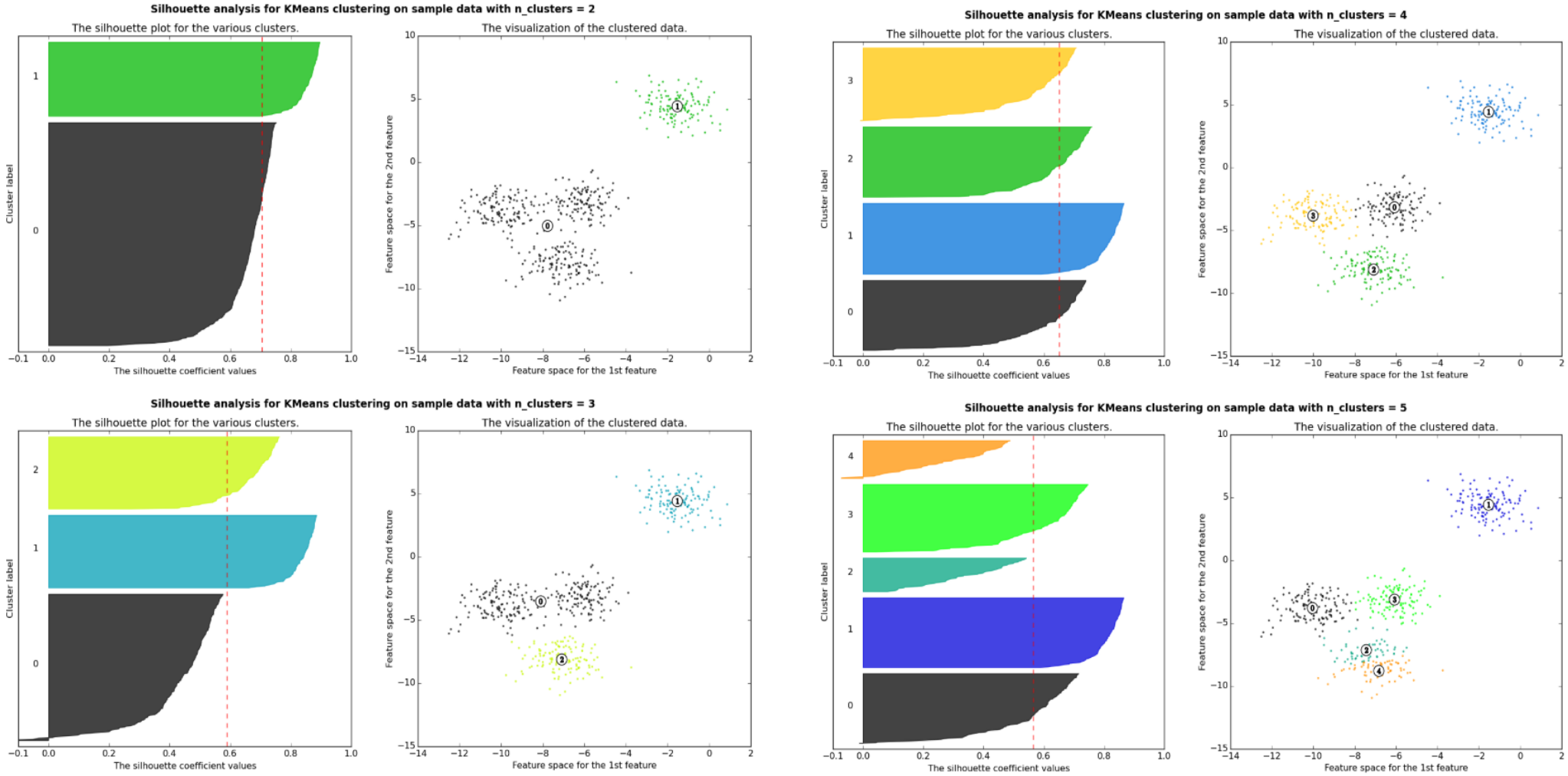
Силуэт выборки (S) — средняя величина силуэта по объектам. Принимает значения из отрезка $[-1; 1]$

- $S \approx -1$ — плохие разрозненные кластеризации
- $S \approx 0$ — кластеры накладываются друг на друга
- $S \approx 1$ — четко выраженные кластеры

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

Метрики качества кластеризации

Силуэт позволяет выбрать оптимальное количество классов k:



Источник: https://scikit-learn.sourceforge.net/dev/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Метрики качества кластеризации

- Если известны истинные метки, то лучше пользоваться V-мерой или ARI
- Если известно количество кластеров, то лучше выбрать коэффициент Дана
- Если совсем ничего неизвестно, хорошо работает силуэт

Метрики качества кластеризации

- Если известны истинные метки, то лучше пользоваться V-мерой или ARI
- Если известно количество кластеров, то лучше выбрать коэффициент Дана
- Если совсем ничего неизвестно, хорошо работает силуэт

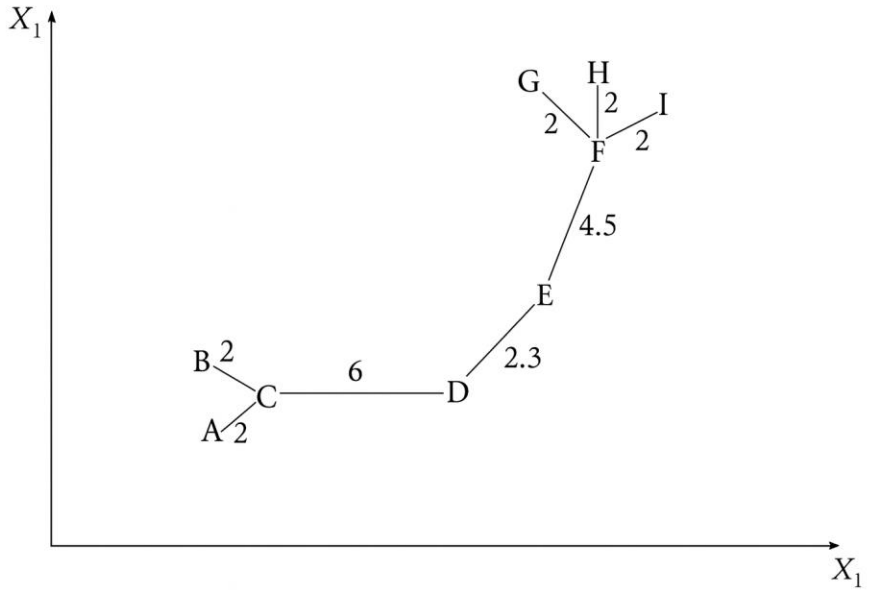
- Если есть возможность собрать разметку в достаточном объеме, то лучше решать задачу классификации

Методы кластеризации: графы

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

Алгоритм состоит из следующих шагов:

1. Построить минимальное остовное дерево по алгоритму по алгоритму Прима или Краскала

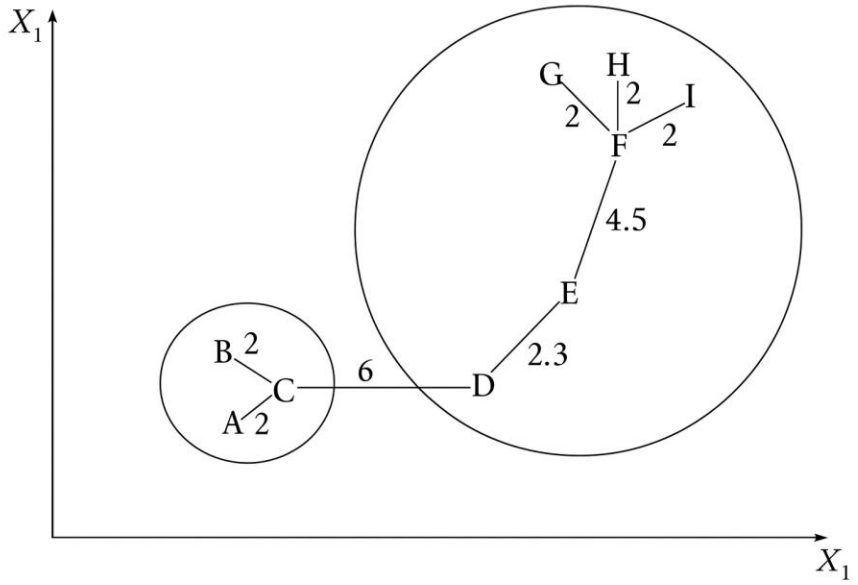


Методы кластеризации: графы

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

Алгоритм состоит из следующих шагов:

1. Построить минимальное остовное дерево по алгоритму по алгоритму Прима или Краскала
2. На основе гиперпараметра K — число кластеров — удалить $K - 1$ самых тяжелых ребра, в результате чего получим K компонент связности



Методы кластеризации: K-средних (K-means)

Один из самых популярных методов.

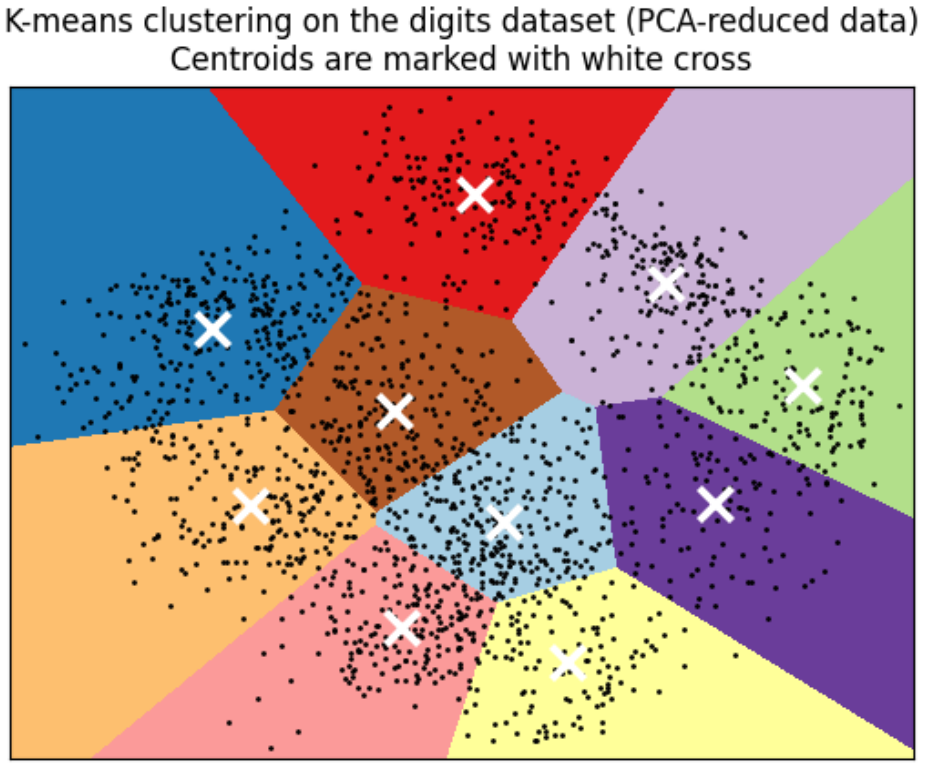
Направлен на выбор центроидов, которые минимизируют инерцию или критерий суммы квадратов внутри кластера:

$$\sum_{i=1}^{n} \min_{\mu_i \in C} (\|x_i - \mu_i\|^2)$$

Методы кластеризации: K-средних (K-means)

Алгоритм состоит из следующих шагов:

1. Отнести каждый объект к ближайшему к нему центру кластера (ему присваивается центроид)

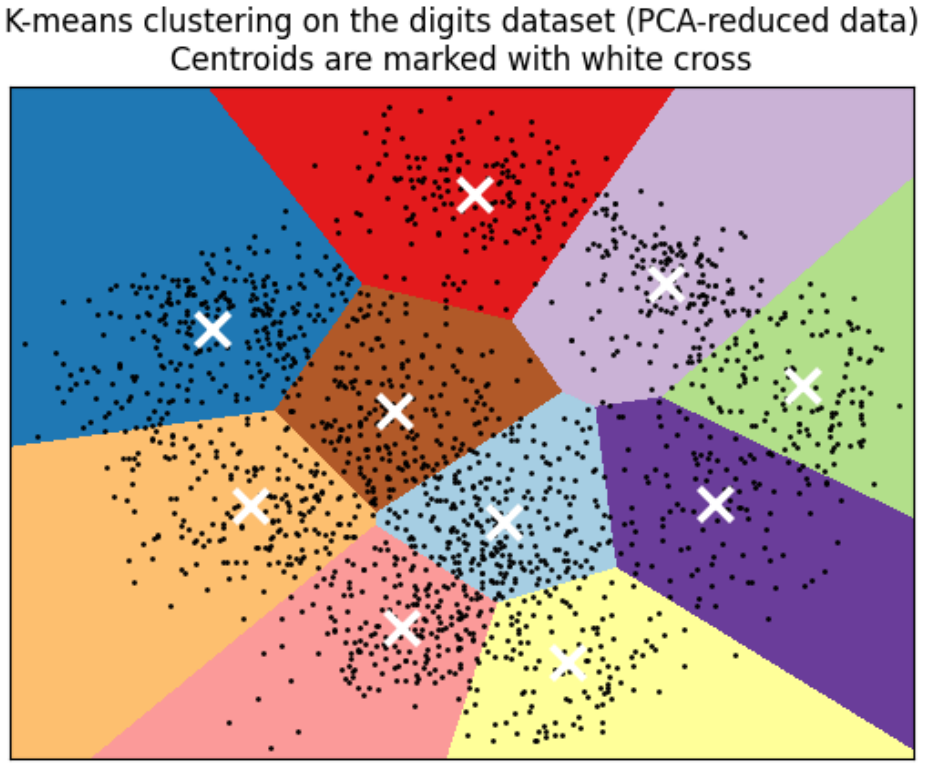


Источник: https://scikit-learn.ru/stable/auto_examples/cluster/plot_kmeans_digits.html

Методы кластеризации: K-средних (K-means)

Алгоритм состоит из следующих шагов:

1. Отнести каждый объект к ближайшему к нему центру кластера (ему присваивается центроид)
2. Пересчитать центры кластеров (создаются новые центроиды путем взятия среднего значения всех выборок, присвоенных каждому предыдущему центроиду)

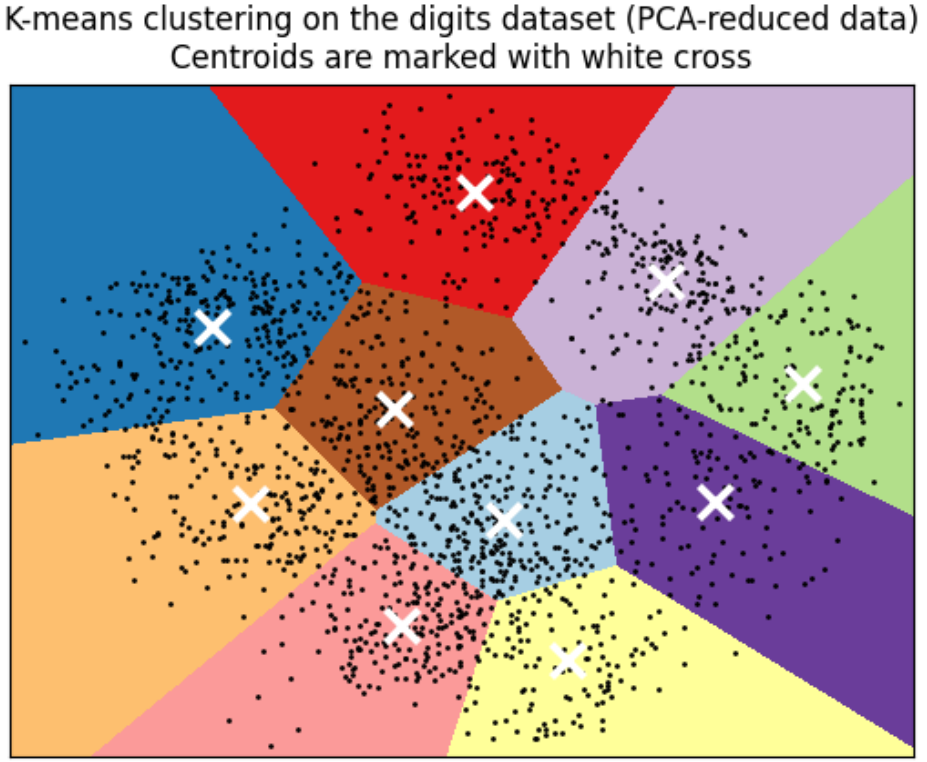


Источник: https://scikit-learn.ru/stable/auto_examples/cluster/plot_kmeans_digits.html

Методы кластеризации: K-средних (K-means)

Алгоритм состоит из следующих шагов:

1. Отнести каждый объект к ближайшему к нему центру кластера (ему присваивается центроид)
2. Пересчитать центры кластеров (создаются новые центроиды путем взятия среднего значения всех выборок, присвоенных каждому предыдущему центроиду)
3. Повторять 1 и 2 до сходимости (вычисляется разница между старым и новым центроидами, и алгоритм повторяет эти последние два шага до тех пор, пока это значение не станет меньше порогового значения. Другими словами, это повторяется до тех пор, пока центроиды не перестанут значительно перемещаться)

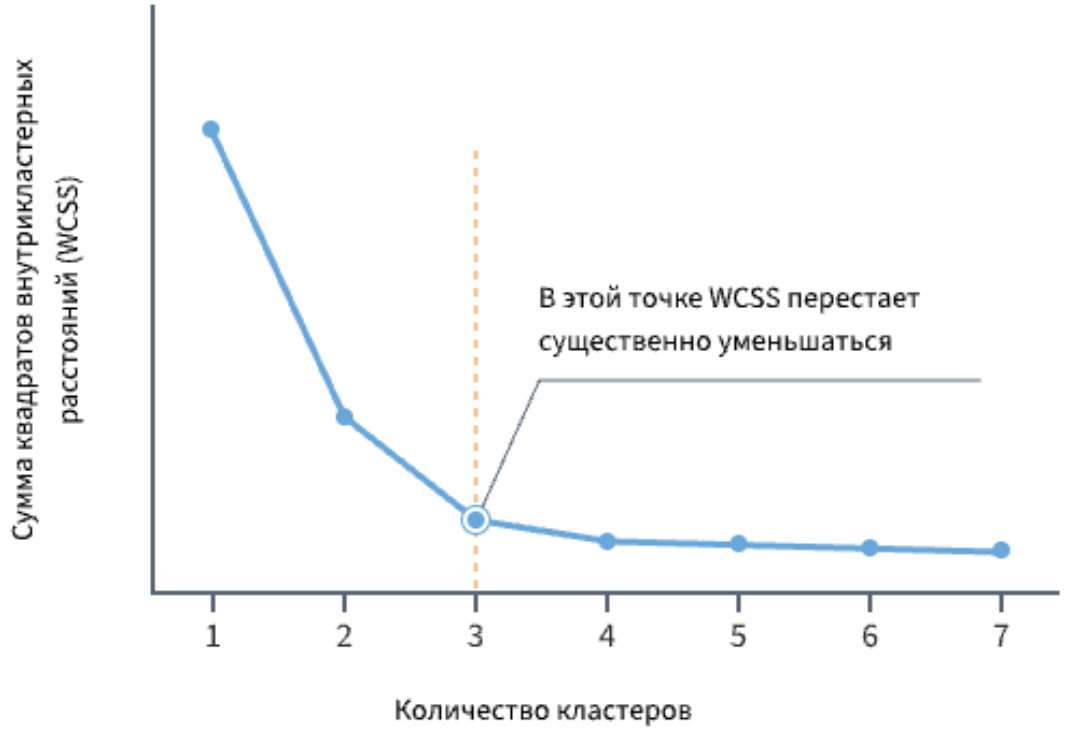


Источник: https://scikit-learn.ru/stable/auto_examples/cluster/plot_kmeans_digits.html

Методы кластеризации: K-средних (K-means)

Правило локтя для выбора K

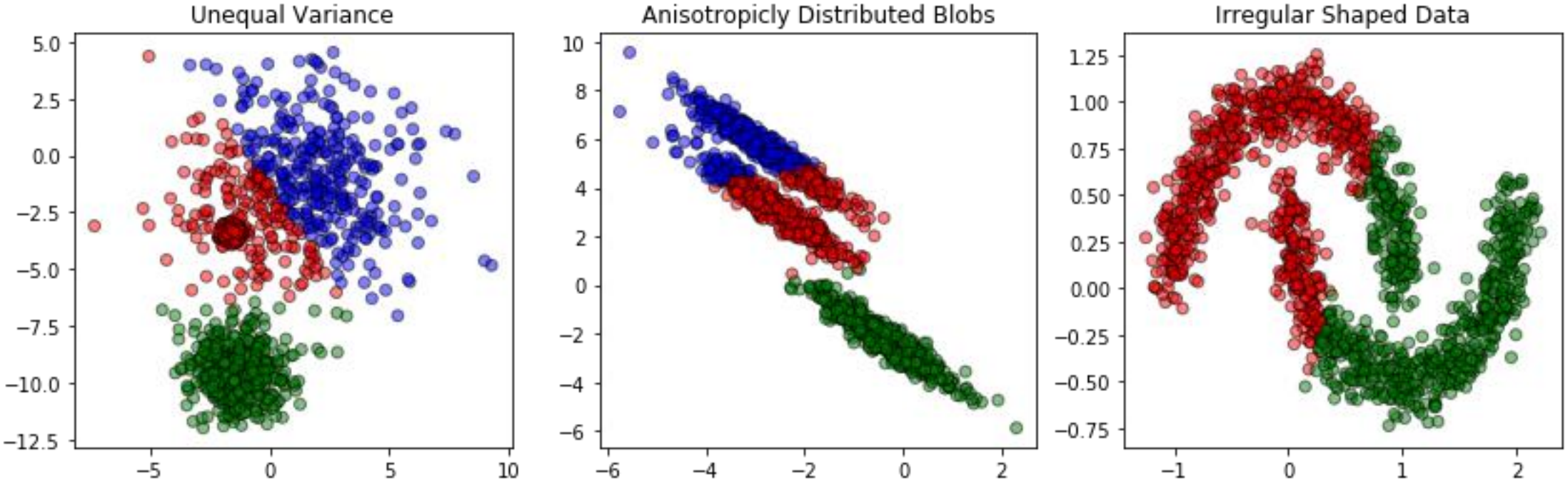
Выбираем такое значение K, когда происходит значительное уменьшение внутрикластерного расстояния



Источник: https://wiki.loginom.ru/articles/elbow-method.html

Методы кластеризации: K-средних (K-means)

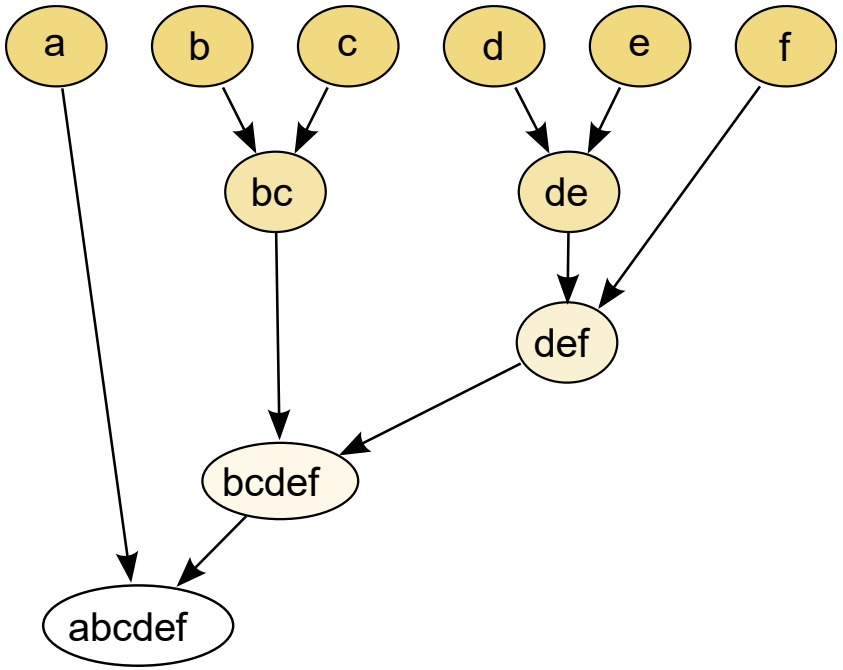
Алгоритм K-means хорошо показывает себя только в том случае, если в распределении данных есть ярко выраженные центры кластеров и все объекты сконцентрированы вокруг них



Источник: https://zerowithdot.com/mistakes-with-k-means-clustering

Методы кластеризации: иерархическая

Иерархия кластеров представлена в виде дерева (дендрограммы). Корень дерева — это уникальный кластер, в котором собраны все образцы, а листья — это кластеры только с одним образцом

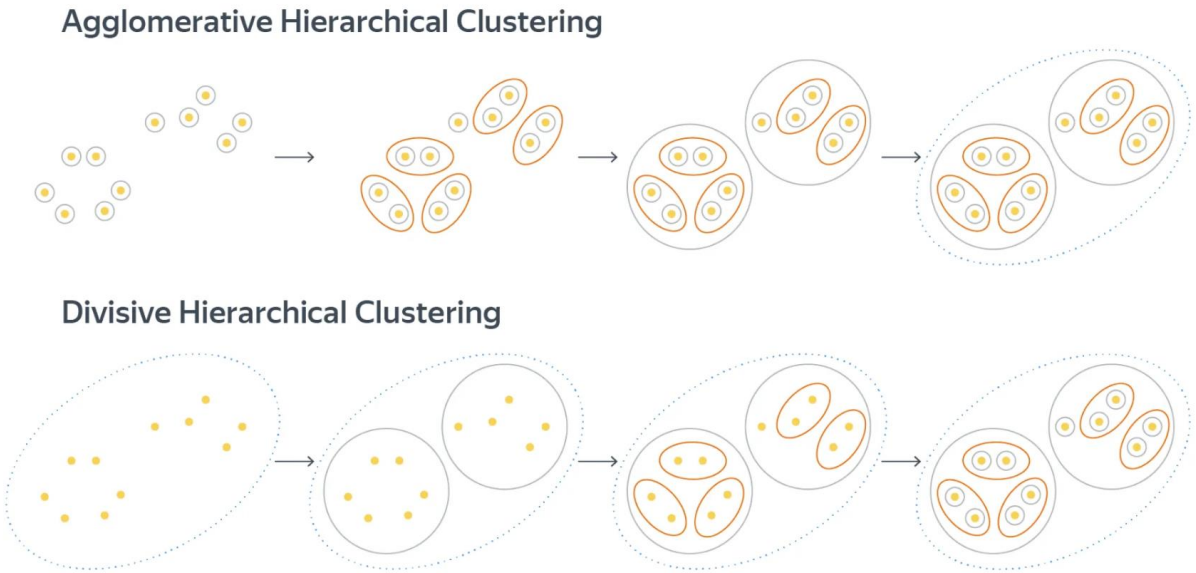


Источник: https://ru.wikipedia.org/wiki/Иерархическая_кластеризация

Методы кластеризации: иерархическая

Виды моделей в зависимости от алгоритма построения:

1. Агломеративные (строят каждый следующий кластер путем объединения предыдущих)
2. Дивизионные (строят каждый следующий кластер путем разбиения предыдущих)



Источник: https://education.yandex.ru/handbook/ml/article/klasterizaciya

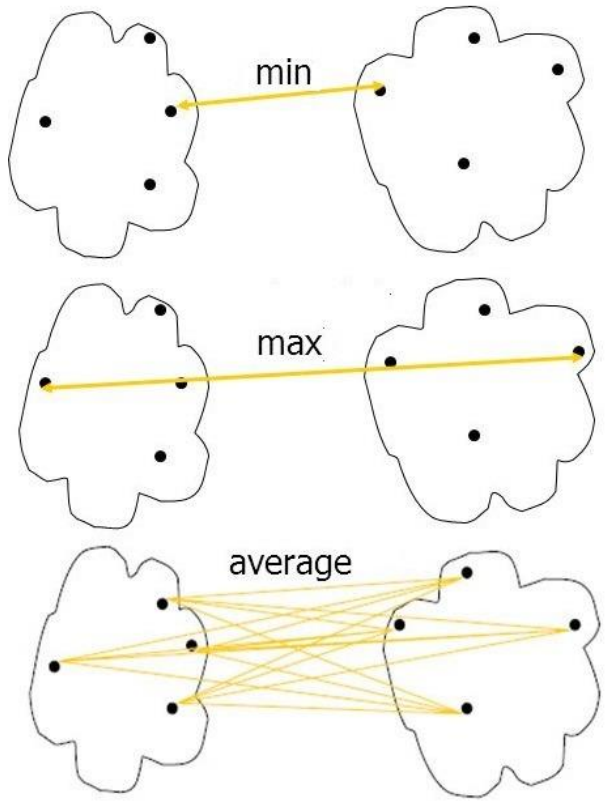
Методы кластеризации: иерархическая

Расстояние между кластерами U и V можно посчитать тремя способами:

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

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

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

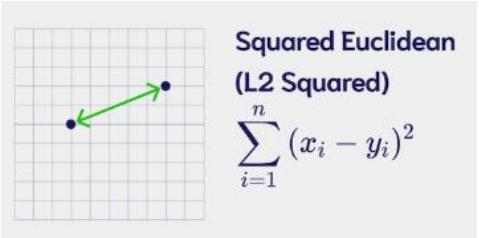


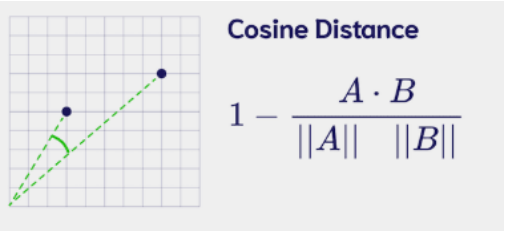
Источник: https://education.yandex.ru/handbook/ml/article/klasterizaciya

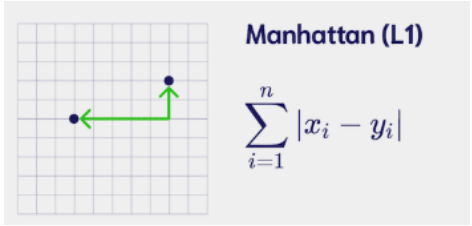
Методы кластеризации: иерархическая

Расстояние между объектами:

- Евклидова метрика
- Манхэтаннское расстояние
- Косинусная близость



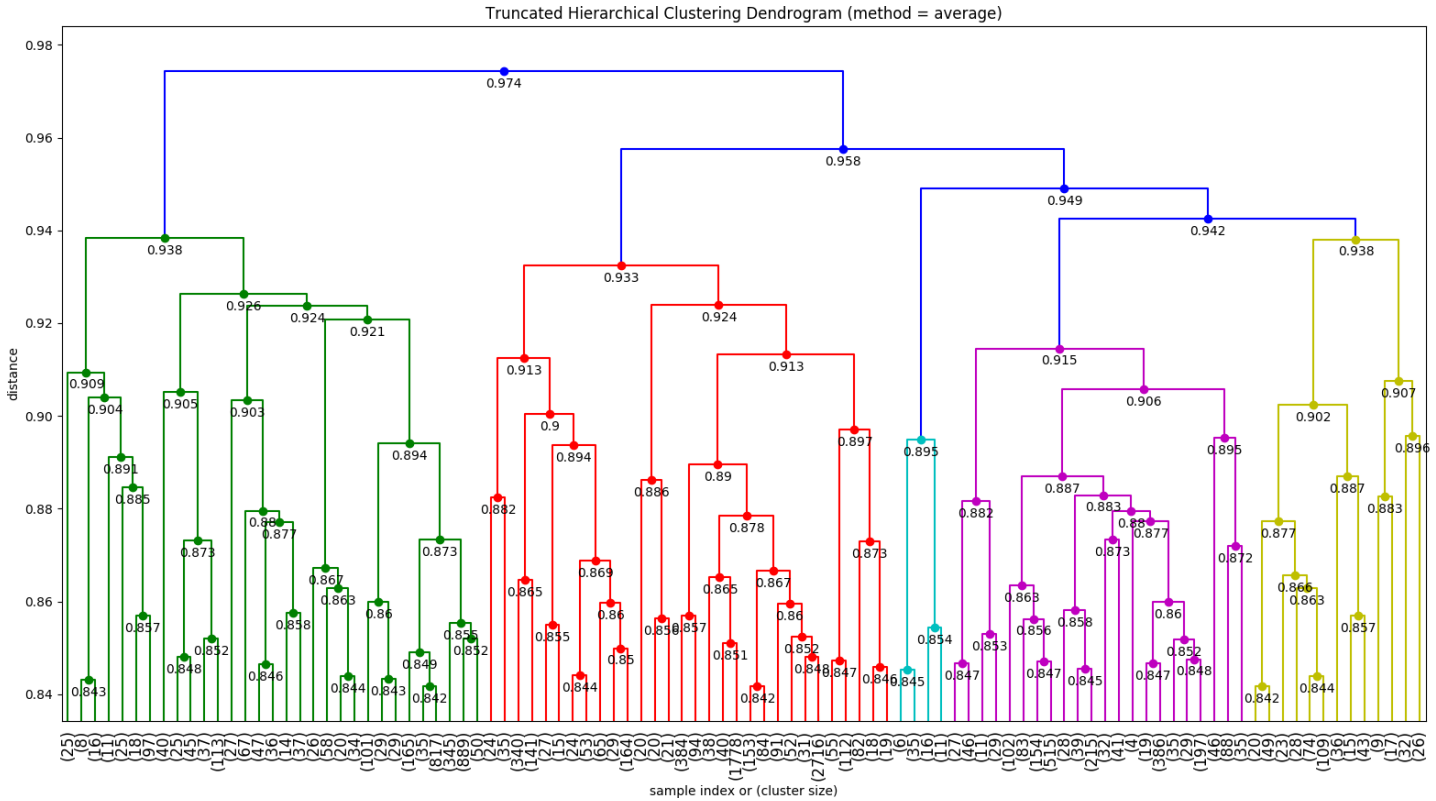




Источник: https://weaviate.io/blog/distance-metrics-in-vector-search

Методы кластеризации: иерархическая

Пример дендрограммы



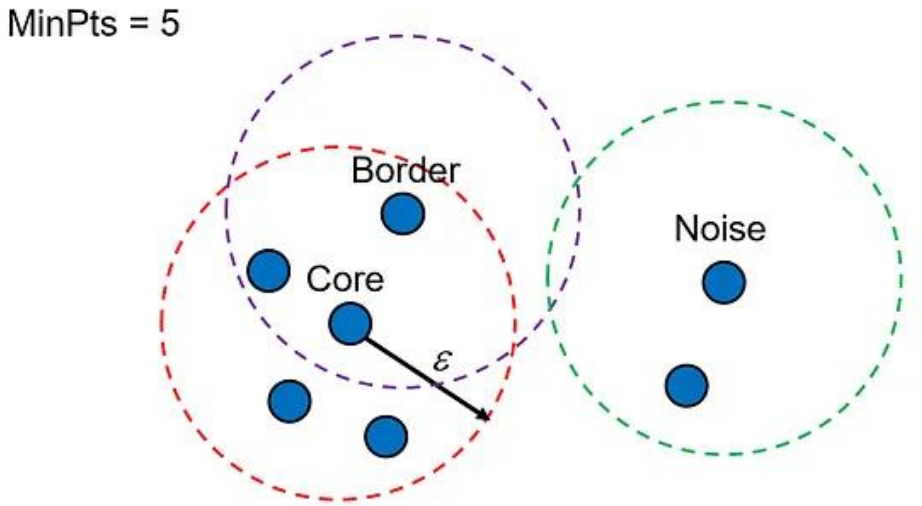
Источник: https://stackoverflow.com/questions/54629587/retrieve-clusters-with-n-members-from-hierarchical-clustering

Методы кластеризации: DBSCAN

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

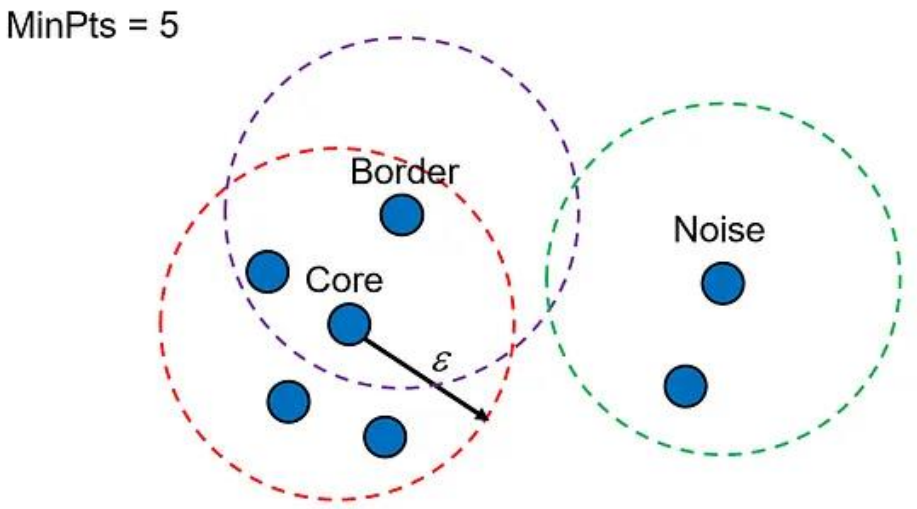
Использует идею поиска кластеров на основе выделения участков в данных с заданной плотностью

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



Источник: https://ai.plainenglish.io/dbscan-density-based-clustering-aaebd76e2c8c

Методы кластеризации: DBSCAN



Источник: https://ai.plainenglish.io/dbscan-density-based-clustering-aaebd76e2c8c

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

Методы кластеризации: DBSCAN

Алгоритм:

1. Выбрать точку без метки
2. Если в окрестности ε меньше, чем N_0 точек, пометить ее как шумовую
3. Найти все основные точки, и если имеется пересечение между двумя окрестностями основных точек, то соединить их ребром
4. В полученном графе выделить компоненты связности – они и будут кластерами
5. Граничные точки отнести к тому кластеру, куда попала ближайшая основная к ним точка

Гиперпараметры:

eps (ε) — размер окрестности

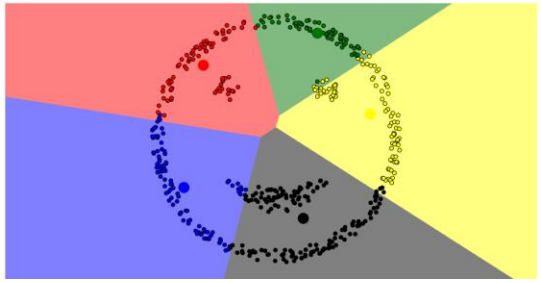
N_0 — минимальное количество точек в окрестности, чтобы считать точку основной

Интерактивная визуализация алгоритма:

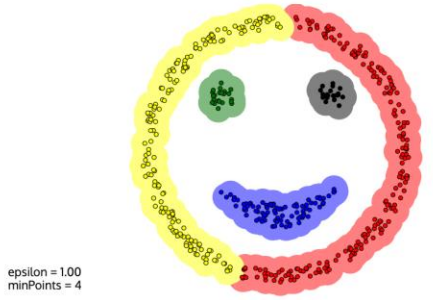
https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Методы кластеризации: DBSCAN

Плюсы:
Сам определяет количество кластеров (по модулю задания ε и N_0)
Успешно справляется со сложными формами кластеров

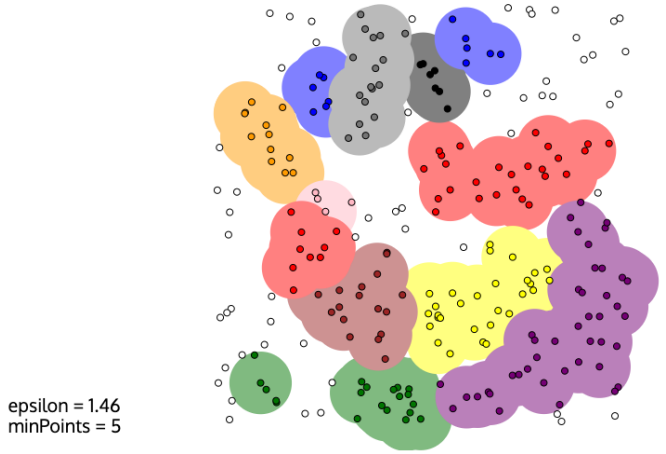


K-means (K=5)



DBSCAN ($\varepsilon = 1, N_0 = 4$)

Минусы:
Долгое время работы
Чувствителен к настройке гиперпараметров
Плохо справляется с данными, у которых кластеры переменной плотности



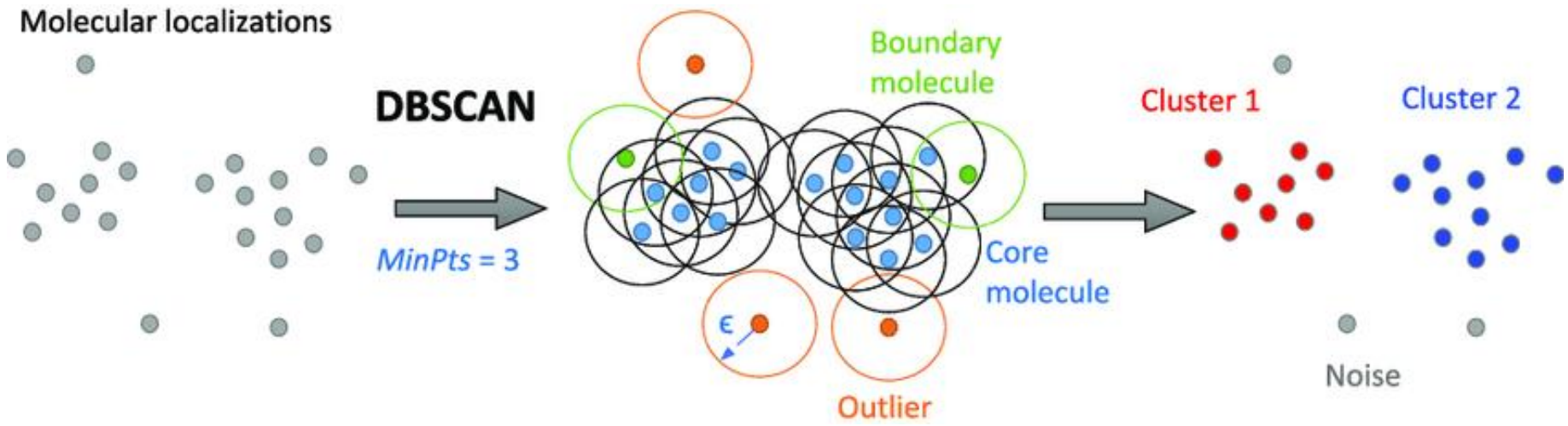
Сгенерировано на https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Методы кластеризации: DBSCAN

Алгоритм:

1. Выбрать точку без метки
2. Если в окрестности меньше, чем min_pts точек, пометить ее как шумовую
3. Если точка не шумовая, создать кластер, поместить в него текущую точку
4. Для всех точек из окрестности S:
   a. Если точка шумовая, отнести к этому кластеру, но не использовать для расширения
   b. Если точка основная, отнести к данному кластеру, а ее окрестность добавить к S
5. Повторить шаги 1—4

Методы кластеризации: DBSCAN

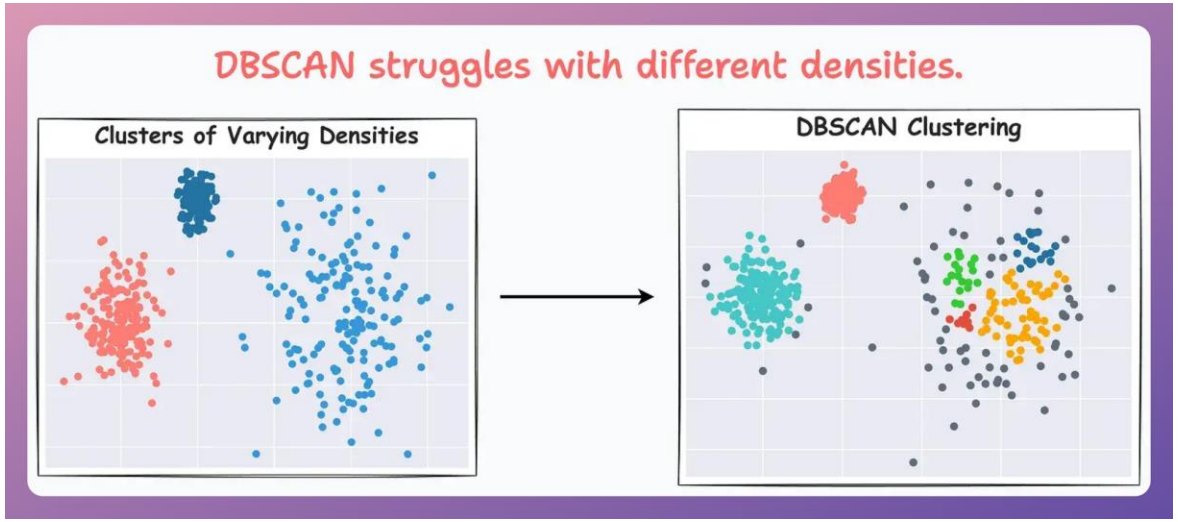


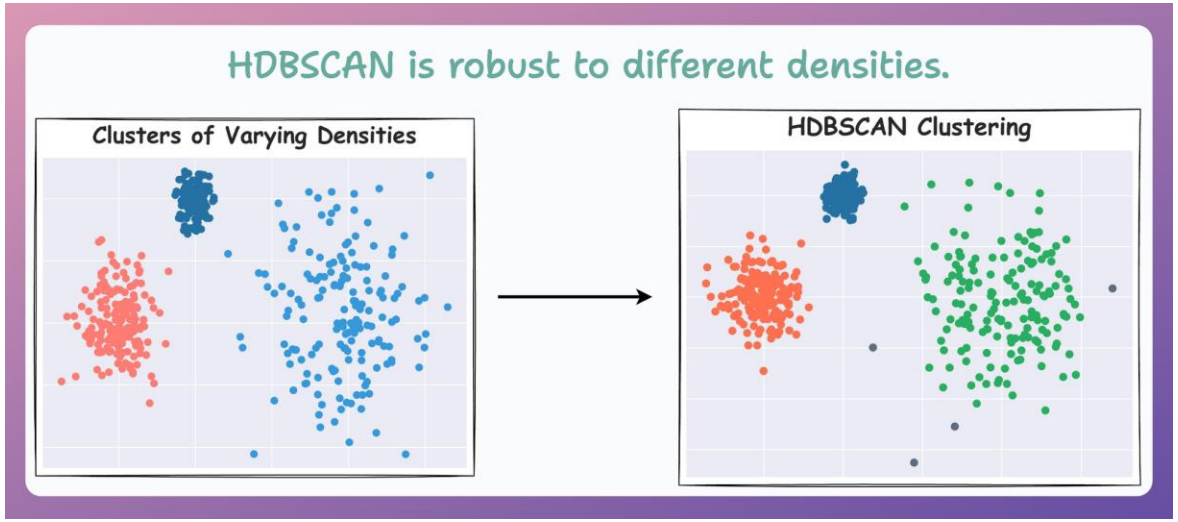
Источник: https://www.researchgate.net/publication/342141592_A_Review_of_Super-Resolution_Single-Molecule_Localization_Microscopy_Cluster_Analysis_and_Quantification_Methods

Методы кластеризации: HDBSCAN

Улучшенная версия DBSCAN.

Использует иерархический подход в объединении кластеров, найденных алгоритмом DBSCAN





Источник: https://www.dailydoseofds.com/hdbscan-the-supercharged-version-of-dbscan-an-algorithmic-deep-dive/

Методы кластеризации

Метод	Параметры	Масштабируемость	Сценарии использования
K-Means	Число кластеров	Очень большой n_samples, среднее n_clusters (в MiniBatch-версии)	Основное назначение — кластеры одинакового размера, плоская геометрия, не слишком много кластеров, индуктивный
Ward hierarchical clustering	Количество кластеров или порог расстояния	Большой n_samples и n_clusters	Множество кластеров, возможно с ограничениями связности, трансдуктивный
Agglomerative Clustering	Число кластеров или порог, тип связи, метрика	Большой n_samples и n_clusters	Множество кластеров, возможно с ограничениями связности, трансдуктивный
DBSCAN	eps, minPts	Очень большой n_samples, среднее n_clusters	Неплоская геометрия, кластеры неравномерного размера, удаление выбросов, трансдуктивный
HDBSCAN	Минимальное количество элементов в кластере, минимальное количество соседей по точкам	большой n_samples, среднее n_clusters	Неплоская геометрия, неравномерные размеры кластеров, удаление выбросов, трансдуктивный, иерархическая, переменная плотность кластеров

Источник: https://scikit-learn.ru/stable/modules/clustering.html

Методы кластеризации

Важно помнить:

- Алгоритм кластеризации не знает, чего вы хотите
- Невозможно гарантировать, что, например, при кластеризации текстов вы получите разбиение именно по темам
- Нередко кластеры оказываются неинтерпретируемыми